ELSEVIER Journal of Econometrics 106 (2002) 171–201

JOURNAL OF Econometrics

www.elsevier.com/locate/econbase

Regression models for choice-based samples with misclassification in the response variable

Esmeralda A. Ramalho[a,b,*]

[a] *Department of Economics, University of Bristol, 8 Woodland Road,*
Bristol BS8 1 TN, UK

[b] *Departamento de Economia, Universidade de Évora, Largo dos Colegias 2,*
7000-803 Évora, Portugal

Received 5 May 2000; revised 31 January 2001; accepted 30 April 2001

Abstract

In this paper, we provide a general framework to deal with the presence of misclassification in the response variable in choice-based samples. The contaminated data sampling distribution is written as a function of the error-free conditional distribution of the dependent variable given the covariates and the conditional misclassification probabilities of the observable variable of interest given its latent values. We propose an extension of Imbens' (Econometrica 60 (1992) 1187) efficient generalized method of moments to estimate this model and outline a specification test to detect the presence of this sort of measurement error. The performance of both the estimators and the test is investigated in a Monte Carlo simulation study, which shows very encouraging results. © 2002 Elsevier Science S.A. All rights reserved.

JEL classification: C25; C51; C52

Keywords: Choice-based sampling; Misclassification; Discrete choice models; Generalized method of moments estimation; Score tests

1. Introduction

In this paper, we provide a general framework to deal with the presence of misclassification in the discrete response variable in choice-based

* Corresponding author. Tel.: +44-117-954-6996.
E-mail address: ela@uevora.pt (E.A. Ramalho).

PII: S0304-4076(01)00094-X

samples. A choice-based (CB) sample (also commonly known as response-based, case-control, or retrospective sample) results from an endogenous stratified sampling scheme where each stratum is defined according to the individual responses, described by the discrete values taken by the response variable. The main motivation for this sampling scheme is the possibility of oversampling rare alternatives, which not only may improve the accuracy of the econometric analysis but also reduce survey costs. However, this sampling structure requires unconventional estimation methods, which mostly assume the availability of prior information on the marginal response probabilities. The extensive literature on CB samples includes the seminal papers of Manski and Lerman (1977), Manski and McFadden (1981), Cosslett (1981a, b) and Imbens (1992). Imbens and Lancaster (1996), Wooldridge (1998, 1999) and the literature review by Cosslett (1993) discuss the more general area of endogenous stratification.

With a discrete response variable, measurement error induces a transposition of its integer values, that is misclassification or miscategorization. This problem may arise because of a variety of reasons. For example, the respondent might not understand the question or may be reluctant to provide the correct response, or the agent collecting the data may record an incorrect code for the answer. Seminal works addressing misclassification problems in regression models include inter alia Espeland and Odoroff (1985), Palmgren and Ekholm (1987), Ekholm and Palmgren (1987), Copas (1988), Wang and Carroll (1993), Hausman et al. (1998) and Abrevaya and Hausman (1999). In general, the regression model is written as a function of the error-free conditional distribution of the response variable given the covariates and the conditional misclassification probabilities of the observable variable of interest given its latent values. Sometimes these misclassification probabilities are assumed known, having been estimated previously using data from a so-called double sampling scheme.[1] Otherwise, these probabilities are jointly estimated with the other parameters of interest. Recently, the random sampling (RS) estimation problem has been addressed in a semi-parametric framework, see Hausman et al. (1998) and Abrevaya and Hausman (1999), which requires neither the specification of an error model nor the assumption of a conditional distribution for the latent response variable given the covariates.

Although both questions of misclassification and CB sampling have already been intensely analysed, the former literature focuses on cases where sampling is random, while the latter assumes that the response variable is correctly measured. As far as we know, there is no work dealing with both problems simultaneously, in which case misclassification is potentially more damaging, since the sampling design is based on the contaminated values of the response

[1] This sampling scheme merges the error-prone data with an additional subsample in which the study participants provide the correct responses.

variable. The closest approach to ours is due to Wang and Carroll (1993), who consider the case of binary logistic CB samples. They propose the use of the same estimation procedures as in a misclassified RS, arguing that, except for a shift in the intercept term, all the parameters of interest are consistently estimated. However, we show that for consistent estimation both misclassification and CB sampling must be taken into account.

This paper formulates a model which combines the standard models for CB sampling and misclassification, including both of them as particular cases. We identify which components of the CB sampling model are affected by misclassification and, analogously to misclassification in RS, we write these components as a function of their error-free versions and the misclassification probabilities. To estimate the parameters of interest, we extend Imbens' (1992) efficient generalized method of moments (GMM) estimator for CB samples. We exploit the similarity between the availability or otherwise of information on the misclassification probabilities and that of knowledge or otherwise of the marginal choice probabilities in stratified samples. In the absence of additional information, those features are jointly estimated with the parameters of interest. Otherwise, available information is incorporated, enabling more efficient estimators to be obtained. Therefore, our estimation method for models of misclassified CB samples permits, simultaneously, the unification of the two standard approaches for estimation of models with misclassification in RS.

These estimators are useful if misclassification in the response variable is suspected. Hence, we provide a score test in a GMM setting for the detection of this type of measurement error. Moreover, with some simplifications, this test can be specialised for RS, for which, to the best of our knowledge, excepting a score test in binary logit models (Copas, 1988), no specification tests for misclassification exist.

In this paper, we focus mainly on discrete choice models with several alternatives, the leading models of interest when the response variable is discrete. However, merely by changing the interpretation of some features, our framework is directly applicable to other models, for example, for count data and ordered responses, where the importance of accounting for improper measurement in the response variable has been recognized under RS, respectively, by Whittemore and Gong (1991) and Cameron and Trivedi (1998, pp. 310–312) and Abrevaya and Hausman (1999).

The structure of the paper is as follows. Section 2 formulates a model for CB sampling which incorporates misclassification. Section 3 analyses the main consequences of sample contamination on the probability distributions underpinning the latent model. Section 4 presents GMM estimation methods appropriate for misclassification in the response variable. Section 5 outlines a score test to detect the presence of this sort of measurement error. Section 6 concludes. Sketches of various proofs are given in Appendices A–C.

2. Model specification

The model for CB samples with misclassification presented in this section extends the standard formulation for correctly measured CB samples. Although most of the literature on misclassification is concentrated on binary models (see, for example, Copas, 1988; Carroll and Pederson, 1993; Wang and Carroll, 1993; Hausman et al., 1998), we consider the general case of more than two discrete responses. Section 2.1 presents the model for CB sampling and Section 2.2 adapts this formulation to handle misclassification in the response variable.

2.1. Standard regression models for CB samples

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

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

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

CB sampling involves the partition of the population into strata, from each of which a random sample is drawn. For simplicity, suppose that the strata are defined only in terms of the response variable. Assume the existence of J nonempty and possibly overlapping strata, which are subsets of $\mathscr{Y}\times\mathscr{X}$. Each stratum is designated as $\mathscr{C}_s=\mathscr{Y}_s\times\mathscr{X}$, for $S\in\mathscr{S}$, $\mathscr{S}=\{1,\ldots,J\}$, and $\mathscr{Y}_s$ defined as a subset of $\mathscr{Y}$. The probability of a randomly drawn observation lying in stratum $\mathscr{C}_s$ is

$$Q_s=\int_{\mathscr{X}}\sum_{y\in\mathscr{Y}_s}\Pr(y|x,\theta)f(x)\,\mathrm{d}x. \tag{2}$$

This probability is the sum over all the values of the response variable included in $\mathscr{Y}_s$ of the marginal probability of observing an individual choosing $Y=y$, Q_y, defined by

$$Q_y=\int_{\mathscr{X}}\Pr(y|x,\theta)f(x)\,\mathrm{d}x. \tag{3}$$

Assuming that the sample is drawn according to the multinomial sampling scheme, the agent who collects the sample defines the probability H_s that an observation falls into stratum s in the sample.[2] In this setting the sampling density function of (Y,X,S) is given by

$$h(y,x,s) = \frac{H_s}{Q_s}\Pr(y|x,\theta)f(x), \tag{4}$$

$(y,x) \in \mathscr{C}_s$, $s \in \mathscr{S}$. Only when the sample is self-weighted, in which case H_s equals Q_s, does this scheme becomes equivalent to RS because the sampling and the population joint densities coincide.

2.2. An extended model for CB samples incorporating misclassification

In any CB sampling design, the presence of response measurement error affects not only the response variable, Y, as in RS, but also the design of the strata, which is based on the mismeasured variable. Let Y^* and S^* represent, respectively, the observable response and the indicator of the observable strata, with $Y^* \in \mathscr{Y}^*$ and $S^* \in \mathscr{S}^*$. Denote each observable stratum as $\mathscr{C}^*_{s^*} = \mathscr{Y}^*_{s^*} \times \mathscr{X}$, where $\mathscr{Y}^*_{s^*}$ is a subset of $\mathscr{Y}^*$. Assume that Y and S are, respectively, the latent true response and the indicator of the strata defined in terms of Y. Employing the superscript "$*$" to denote the contaminated versions of the density and probability functions, the consequent sampling joint density for the observable (Y^*,X,S^*) is written as

$$h^*(y^*,x,s^*) = \frac{H^*_{s^*}}{Q^*_{s^*}}\Pr{}^*(y^*|x,\theta)f(x). \tag{5}$$

All components, except the marginal distribution of X, are distorted. Probability H_s is contaminated because, as only Y^* is observed, the analyst establishes a sampling design based on this observable error-prone variable, while the sampling structure in terms of latent variables, characterized by H_s, is unknown. On the other hand, the distortion in the choice marginal probability, Q_y, and the stratum occupancy probability, Q_s, results from the fact that these probabilities are now a function of the conditional distribution of Y^* given X [see Eqs. (3) and (2)].

The error model we employ relies on the assumption that, conditional on the latent response, the reported outcome is independent of the individual characteristics X. Hence, Y^* is a surrogate measure for Y, according to the definition provided by Carroll et al. (1995, p. 235). We define, thus, the conditional probability of observing the response Y^* given the latent outcome

[2] For a detailed discussion on the three sampling shemes for CB samples, multinomial sampling, standard stratified sampling and variable probability sampling, see, for example, Cosslett (1993) and Imbens and Lancaster (1996).

Y as

$$\Pr(Y^* = y^*|Y = y, x) = \Pr(Y^* = y^*|Y = y) = \alpha_{y^*y}, \tag{6}$$

where $0 \leqslant \alpha_{y^*y} \leqslant 1$ and $\sum_{y^* \in \mathscr{Y}^*} \alpha_{y^*y} = 1$ for all y^*, y.[3] This type of approach is wide spread in misclassification models, see for example, Poterba and Summers (1995), Hausman et al. (1998) and Abrevaya and Hausman (1999), which concentrate mainly on the case where Y^* is a surrogate response, although the last two papers also suggest extensions for when Y^* conditional on Y and X is not independent of X.

For RS binary data with misclassification error, Hausman et al. (1998) gave the following identification condition: for a given choice $Y^* = j^*$, the conditional misclassification probability must be smaller than the conditional probability of correct classification, that is, $\alpha_{j^*y} < \alpha_{j^*j^*}$ for $y \neq j^*$, which, as $\sum_{y^* \in \mathscr{Y}^*} \alpha_{y^*y} = 1$, implies that the sum of the two misclassification probabilities is smaller than one: $\alpha_{10} + \alpha_{01} < 1$. This condition implies that there is no point in trying to model data when their quality is so poor that the probability of observing $Y^* \neq Y$ is larger than that of $Y^* = Y$.

For the multiple choice case, this identification condition is straightforwardly extended: for a given error-prone response $Y^* = j^*$, the sum of the conditional misclassification probabilities must be smaller than the conditional probability of correct classification, that is, $\sum_{y \in \mathscr{Y} | y \neq j^*} \alpha_{j^*y} < \alpha_{j^*j^*}$, which implies that $\sum_{y \in \mathscr{Y} | y \neq j^*} \alpha_{j^*y} + \sum_{y^* \in \mathscr{Y}^* | y^* \neq j^*} \alpha_{y^*j^*} < 1$. For example, for $C = 2$ if misclassification occurs only between adjacent responses, $\alpha_{y^*y} = 0$ if $|y^* - y| > 1$, this condition requires that the sum of the four possible misclassification probabilities is smaller than one: $\alpha_{10} + \alpha_{01} + \alpha_{12} + \alpha_{21} < 1$.

A consequence of the assumption of relatively small misclassification probabilities is that no choice or stratum may become unobservable due to misclassification. Assuming, in addition, that we do not observe any response which is not contained in $\mathscr{Y}$, then $\mathscr{Y}^* = \mathscr{Y}$ and $\mathscr{S}^* = \mathscr{S}$ and, simultaneously, $\mathscr{Y}^*_{s^*} = \mathscr{Y}_s$.

The contaminated population features of (5) may be obtained straightforwardly. The conditional probability of $Y^* = y^*$ given $X = x$ is

$$\Pr{}^*(y^*|x, \theta, \alpha) = \sum_{y \in \mathscr{Y}} \alpha_{y^*y} \Pr(y|x, \theta), \tag{7}$$

where α is a vector with dimension $D = (C+1)C$, composed of the misclassification probabilities α_{y^*y}, for $y^* \neq y$.[4] The marginal probability of observing

[3] Note that the conditional probabilities of misclassification (correct classification) are given by α_{y^*y} for $y^* \neq y$ $(y^* = y)$.

[4] Note that the $(C+1)$ conditional probabilities of correct classification need not be included in α as $\alpha_{yy} = 1 - \sum_{y^* \in \mathscr{Y}^* | y^* \neq y} \alpha_{y^*y}$.

an unit from stratum $\mathscr{C}^*_{s^*}$ is

$$Q^*_{s^*} = \sum_{y^* \in \mathscr{Y}^*_{s^*}} Q^*_{y^*}, \tag{8}$$

where $Q^*_{y^*}$ is the distorted marginal probability of choice $Y^* = y^*$:

$$\begin{aligned} Q^*_{y^*} &= \int_{\mathscr{X}} \sum_{y \in \mathscr{Y}} \alpha_{y^* y} \Pr(y|x,\theta) f(x)\,\mathrm{d}x \\ &= \sum_{y \in \mathscr{Y}} \alpha_{y^* y} Q_y. \end{aligned} \tag{9}$$

It is clear that the distributions of the observable variables Y^* and S^* are a weighted version of those of the latent Y and S. Note also that in the misclassification model considered here, in which the mismeasured outcome depends only on the latent response, both $\Pr(y|x,\theta)$ and Q_y are affected in the same way. However, if Y^* conditional on Y and X was not independent of X, probabilities $Q^*_{y^*}$, and consequently $Q^*_{s^*}$, could no longer be written as a function only of the error free marginal choice probabilities Q_y.

Substituting $\Pr^*(y^*|x,\theta,\alpha)$ and $Q^*_{s^*}$ in (5), we obtain

$$h^*(y^*,x,s^*) = \frac{H^*_{s^*}}{\sum_{y^* \in \mathscr{Y}^*_{s^*}} \sum_{y \in \mathscr{Y}} \alpha_{y^* y} Q_y} \sum_{y \in \mathscr{Y}} \alpha_{y^* y} \Pr(y|x,\theta) f(x). \tag{10}$$

Note that when all the responses are correctly classified, that is, $\alpha_{y^* y} = 0$ for $y^* \neq y$ and $\alpha_{y^* y} = 1$ for $y^* = y$, the model assumes the usual CB sampling form. In case of self-weighting, when $H^*_{s^*} = Q^*_{s^*}$, this model conforms with the formulation for misclassification under RS.

Given the distortions suffered by the sampling density function, all parametric and semi-parametric estimation procedures which ignore the measurement error are likely to lead to inconsistent estimators for the parameters of interest, since they rely on a misspecified distribution for the data. This inconsistency is documented for RS by Hausman et al. (1998) who show that, in binary models, even a small amount of miscategorization generates inconsistent maximum likelihood estimators. Clearly, the models for CB samples behave in a similar fashion, being a generalization of the RS specification.

As in RS, our specification may be reinterpreted in two ways. On the one hand, it can be seen as a robust model for data contaminated by not only measurement error but also outliers.[5] On the other hand, as in Hausman et al. (1998), our model may describe a situation where a fraction $\alpha_{y^* y}$ for $y^* \neq y$ of respondents always choose Y^*, independent of the characteristics X, while

[5] Note that Copas (1988) presents his model as an alternative approach to the traditional techniques for dealing with outliers in logit models proposed by Pregibon (1982).

the responses of the remaining individuals conform with the structural model $\Pr(y|x,\theta)$.[6]

3. Main consequences of misclassification in the response variable

In this section, we analyse the main consequences of mismeasurement of the response variable. After a brief discussion of the general effects of misclassification, we focus on the particular case of multiplicative intercept models.

3.1. General effects of misclassification

In general, misclassification generates an additional weighting relative to that already present in CB samples. In effect, in a CB sample the joint density function of Y and X in the population is weighted by the ratio H_s/Q_s [see Eq. (4)]. If misclassification is added, see (7)–(9), even in the population the features of the observable variables involve a weighting scheme, via α_{y^*y}.

Naturally, the double weighting structure of the sample is reflected by all the sampling densities and probabilities. Even the sampling density of the covariates, which are assumed to be error-free,

$$h^*(x) = f(x) \sum_{s^* \in \mathscr{S}^*} \frac{H^*_{s^*}}{\sum_{y^* \in \mathscr{Y}^*_{s^*}} \sum_{y \in \mathscr{Y}} \alpha_{y^*y} Q_y} \sum_{y^* \in \mathscr{Y}^*_{s^*}} \sum_{y \in \mathscr{Y}} \Pr(y|x,\theta)\alpha_{y^*y} \qquad (11)$$

displays this property, becoming informative not only about θ, as usual in endogenous samples, but also about α_{y^*y}. Hence, X is not ancillary for any of these parameters.

Moreover, this sampling pattern implies that the conditional misclassification probabilities differ between the sample and population. As the joint sampling probability of observing Y^* and the latent response Y, $\Pr^*_{\mathrm{CBS}}(Y^* = y^*, Y = y)$, can be written as

$$\begin{aligned}\Pr^*_{\mathrm{CBS}}(Y^* = y^*, Y = y) &= H^*_{y^*} \Pr^*_{\mathrm{CBS}}(Y = y | Y^* = y^*, \theta, \alpha) \\ &= \frac{H^*_{y^*}}{Q^*_{y^*}} \alpha_{y^*y} Q_y\end{aligned}$$

and the error-free marginal probability of alternative y in the sample, H_y, is given by

$$H_y = Q_y \sum_{y^* \in \mathscr{Y}^*} \frac{H^*_{y^*}}{Q^*_{y^*}} \alpha_{y^*y},$$

[6] This situation is closely related to the one underlying dogit models for discrete choice data and zero inflated models for count data (for details see, respectively, Gaudry and Dagenais, 1979; Lambert, 1992).

the probability of observing the outcome Y^* in the sample, given that the true response is Y, is

$$\delta_{y^*y} = \Pr^*_{\text{CBS}}(Y^* = y^* | Y = y, \theta, \alpha) = \frac{(H^*_{y^*}/Q^*_{y^*})\alpha_{y^*y}}{\sum_{y^* \in \mathscr{Y}^*} (H^*_{y^*}/Q^*_{y^*})\alpha_{y^*y}}. \tag{12}$$

Eq. (12) emphasizes that only in case of self-weighting, $H^*_{y^*} = Q^*_{y^*}$, are the sampling and the population probabilities of observing response $Y^* = y^*$ given choice $Y = y$ (respectively, δ_{y^*y} and α_{y^*y}) equal. In this situation, the sampling scheme has a single weighting due to the misclassification. The intuition behind (12) is that if a given response $Y^* = y^*$ is oversampled (undersampled), such that $H^*_{y^*} > Q^*_{y^*}$ ($H^*_{y^*} < Q^*_{y^*}$), the proportion of individuals misclassifying responses $Y = y$ as $Y^* = y^*$ appears inflated (depressed) in the sample relative to the population.

3.2. Multiplicative intercept models

When the response variable conditional on the covariates is described by a multiplicative intercept model (Hsieh et al., 1985), which includes the logit model with a full set of choice dummies as particular case, the common practice in maximum likelihood-based techniques for correctly measured CB samples is to ignore the stratification. The resulting estimators of the covariates' coefficients are still consistent because the sampling and the population conditional distribution of Y given X coincide, except for a distortion in the intercept term. This property has been extensively analysed, in particular for logit models, following Manski and Lerman (1977) (see, for example, Prentice and Pyke, 1979; Manski and McFadden, 1981; Cosslett, 1981a, 1993; Hsieh et al., 1985).

When the response variable is subject to error, Wang and Carroll (1993) propose an analog approach for CB samples. Working with binary logit models with constant probability of misclassification, they argue that estimation using a case control sample can be conducted as if the sample were random, producing consistent estimators for the slope parameters.[7] Basically, their idea consists of correcting for the misclassification effect and, supposing that the logit property of preserving the distributional shape under CB sampling holds, ignoring the sampling design. We show next that this approach is incorrect, not only under the particular conditions considered by these authors, but in general for multiplicative intercept models with any pattern of misclassification.

[7] Note that the assumption of constant probability of misclassification requires the equality of all the conditional misclassification probabilities. Hence, in binary models we have $\alpha_{10} = \alpha_{01}$.

Under the assumption that Y is correctly measured, multiplicative intercept models enjoy the convenient property that the sampling and the population conditional probability of Y given X have the same shape. Defining the multiplicative intercept model as in Hsieh et al. (1985),

$$\Pr(y|x,\lambda_y,\theta_1)=\frac{\lambda_y V_y(\theta_1)}{\sum_{y\in\mathscr{Y}}\lambda_y V_y(\theta_1)}, \tag{13}$$

where $\lambda_1=1$ and $V_y(\theta_1)>0$ for all y, the sampling conditional probability of observing $Y=y$ given $X=x$ induced by the CB design is

$$\Pr{}_{\mathrm{CBS}}(y|x,\lambda_y,\theta_1)=\frac{(H_y/Q_y)\lambda_y V_y(\theta_1)}{\sum_{y\in\mathscr{Y}}(H_y/Q_y)\lambda_y V_y(\theta_1)} \tag{14}$$

which coincides with (13) apart from the constant term $(H_y/Q_y)\lambda_y$.[8] Thus, (13) can be used to estimate θ_1 consistently, although the estimator for λ_y is inconsistent.

By analogy, comparing the contaminated versions of (13) and (14), respectively,

$$\Pr{}^*(y^*|x,\lambda_y,\theta_1,\alpha)=\frac{\sum_{y\in\mathscr{Y}}\alpha_{y^*y}\lambda_y V_y(\theta_1)}{\sum_{y\in\mathscr{Y}}\lambda_y V_y(\theta_1)} \tag{15}$$

and

$$\Pr{}^*_{\mathrm{CBS}}(y^*|x,\lambda_y,\theta_1,\alpha)=\frac{(H^*_{y^*}/Q^*_{y^*})\sum_{y\in\mathscr{Y}}\alpha_{y^*y}\lambda_y V_y(\theta_1)}{\sum_{y^*\in\mathscr{Y}^*}(H^*_{y^*}/Q^*_{y^*})\sum_{y\in\mathscr{Y}}\alpha_{y^*y}\lambda_y V_y(\theta_1)} \tag{16}$$

clearly the shape of (15) is not preserved in (16). Hence, the sampling design cannot be ignored, because correcting only for miscategorization produces inconsistent estimators for all parameters of interest. Consequently, even in this particular class of models, estimation must account for both mismeasurement and stratification. The next section describes appropriate procedures for the estimation of any model involving with an error-prone discrete response variable using a CB sample.

4. GMM estimation

Models accounting for misclassification in RS are commonly estimated by maximum likelihood, requiring both the specification of the conditional distribution of Y given X and the definition of the error model. In CB samples, due to the nonancillarity of the covariates, the marginal distribution of X in the population, $f(x)$, would be required. However, the specification of $f(x)$ can be avoided by applying Imbens' (1992) GMM methodology. This section

[8] The multinomial logit model arises if $\lambda_y=\mathrm{e}^{\theta_{0y}}$, $V_y(\theta_1)=\mathrm{e}^{x'\theta_{1y}}$, and $\theta_{01}=\theta_{11}=0$.

extends the GMM estimation method for CB samples correctly measured to deal with misclassification.

Our approach provides a range of estimators for CB samples with misclassification which are appropriate in the presence or otherwise of additional information on either the marginal choice probabilities or the conditional misclassification probabilities, or both, and includes cases when this information is exact or subject to sampling variation. This is especially important because it provides an unified framework for two lines of investigation for models appropriate for miscategorized samples, one which estimates the misclassification probabilities simultaneously with the other parameters of interest and another which uses estimates for those probabilities obtained in a first stage as exact information. Moreover, our method can account for sampling variation resulting from the estimation of the misclassification probabilities in a first stage. Section 4.1 defines the moment conditions employed in GMM estimation, Section 4.2 presents the GMM estimators, and Section 4.3 reports a Monte Carlo experiment conducted to evaluate the performance of some of the proposed estimators.

4.1. Derivation of the moment conditions

The efficient GMM estimator proposed by Imbens (1992) for regression models with correctly measured CB samples involves the estimation of the vector of parameters of interest θ as well as other features related to stratified sampling: the proportion of the strata in the sample, H_s, and the marginal probability of each choice, Q_y, from which, for each stratum $\mathscr{C}_s$, summing over $\mathscr{Y}_s$, Q_s can also be estimated. Although H_s is known in the multinomial sampling scheme, its estimation is justified by the fact that it enables the analysis to be undertaken conditional on $\hat{H}_s = N_s/N$, where N_s is the number of individuals drawned from stratum $\mathscr{C}_s$, an ancillary statistic for θ. This procedure conforms with the principle of conditionality (see Cox and Hinkley, 1974, p. 38), according to which estimation should be conducted conditional on ancillary statistics.[9] As for Q_y, its value may or may not be known. In the former case, this aggregate information is combined with the sample data, while in the latter Q_y is estimated.

The estimation procedures proposed to deal with the model incorporating misclassification defined in Section 2.2 enable us to estimate a similar vector of parameters of interest. Besides θ, we estimate the observed sampling probability of each stratum, $H^*_{s^*}$, the error free aggregate choice probabilities, Q_y, and the conditional probabilities of Y^* given Y, α_{y^*y}. $H^*_{s^*}$ is estimated instead of H_s because now the proportion of the strata we observe in the sample is distorted. Conversely, we still estimate the error-free marginal choice

[9] For a detailed discussion about the importance of conditioning on $\hat{H}_s$, see Lancaster (1991).

probabilities and we also allow these probabilities to be known, since they may be estimated from a correctly measured auxiliary sample. However, as we will show later, when this aggregate information is subject to the same error structure as our sample and, thus, we know $Q^*_{y^*}$ instead of Q_y, the adaptation of our method is straightforward. Finally, the probabilities α_{y^*y} are treated similarly to Q_y. If some aggregate information is available about the proportion of the population classifying alternative Y as Y^*, or it can be estimated from an auxiliary sample, this information on α_{y^*y} is incorporated in the estimation procedure. Otherwise, these probabilities are jointly estimated with the other parameters of interest.

In general terms, Imbens' (1992) approach is based on maximum likelihood estimation of a parametric model assuming that the covariates follow a discrete distribution, which is jointly estimated with the parameters of interest. After some transformations, which involve the concentration of the log-likelihood function with respect to the estimators of the mass point probabilities of X, the dependence on the discrete distribution assumption is removed, allowing the generalization of the procedure to any distribution for X. The score functions are then used as moment conditions in GMM estimation, producing, in general, efficient estimators in the semi-parametric sense, attaining full efficiency if the distribution of the covariates is indeed discrete.

Similarly to Imbens (1992), let the unknown marginal distribution of X, $f(x)$, be discrete with L points of support x^l, $l=1,2,\ldots,L$, and associated probability mass parameters $\Pr(X=x^l)=\pi_l$, $\pi_l>0$, $L>J$, $l=1,2,\ldots,L$. The resultant log-likelihood function, based on the contaminated sampling joint density (10), is given by

$$LL(H^*,\theta,\pi,\alpha)=\sum_{i=1}^{N}\Bigg[\ln H^*_{s_i^*}+\ln\Pr{}^*(y_i^*|x^{l_i},\theta,\alpha)+\ln\pi_{l_i} \\ -\ln\sum_{l=1}^{L}\pi_l\sum_{y_i^*\in\mathscr{Y}^*_{s^*}}\Pr{}^*(y_i^*|x^l,\theta,\alpha)\Bigg]. \tag{17}$$

Maximization is performed with respect to the $(J+k+D+L-2)$ dimensional vector of parameters (H^*,θ,α,π), where H^* and π denote, respectively, the $(J-1)$- and $(L-1)$-dimensional vectors $(H_1^*,H_2^*,\ldots,H_{J-1}^*)$ and $(\pi_1,\pi_2,\ldots,\pi_{L-1})$, subject to the restriction $\sum_{l=1}^{L}\pi_l=1$.[10] The resulting first order conditions are

$$S_{H_{t^*}}(\hat H,\hat\theta,\hat\alpha,\hat\pi)=\sum_{i=1}^{N}\left[\frac{I_{(s_i^*=t^*)}}{\hat H^*_{t^*}}-\frac{I_{(s_i^*=J)}}{\hat H^*_{J^*}}\right]=0, \tag{18}$$

[10] Note that $H^*_{J^*}=1-\sum_{s^*=1}^{J-1}H^*_{s^*}$ and $\pi_L=1-\sum_{l=1}^{L-1}\pi_l$.

$$S_{\pi_m}(\hat{H},\hat{\theta},\hat{\alpha},\hat{\pi}) = \sum_{i=1}^{N}\left[\frac{I_{(l_i=m)}}{\hat{\pi}_m} - \hat{\mu} - \frac{\sum_{y_i^*\in\mathcal{Y}_{s^*}^*}\Pr^*(y_i^*|x^m,\hat{\theta},\hat{\alpha})}{\sum_{l=1}^{L}\hat{\pi}_l\sum_{y_i^*\in\mathcal{Y}_{s^*}^*}\Pr^*(y_i^*|x^l,\hat{\theta},\hat{\alpha})}\right]$$

$$= 0, S_\theta(\hat{H},\hat{\theta},\hat{\alpha},\hat{\pi}) \tag{19}$$

$$= \sum_{i=1}^{N}\left[\nabla_\theta \ln \Pr^*(y_i^*|x^{l_i},\hat{\theta},\hat{\alpha})\right.$$

$$\left. -\nabla_\theta \ln \sum_{l=1}^{L}\hat{\pi}_l \sum_{y_i^*\in\mathcal{Y}_{s^*}^*}\Pr^*(y_i^*|x^l,\hat{\theta},\hat{\alpha})\right] = 0, \tag{20}$$

$$S_{\alpha_{t^*t}}(\hat{H},\hat{\theta},\hat{\alpha},\hat{\pi})$$

$$= \sum_{i=1}^{N}\left[\nabla_\alpha \ln \Pr^*(y_i^*|x^{l_i},\hat{\theta},\hat{\alpha}) - \nabla_{\alpha_{t^*t}} \ln \sum_{l=1}^{L}\hat{\pi}_l \sum_{y_i^*\in\mathcal{Y}_{s^*}^*}\Pr^*(y_i^*|x^l,\hat{\theta},\hat{\alpha})\right]$$

$$= 0, \tag{21}$$

$$S_\mu(\hat{H},\hat{\theta},\hat{\alpha},\hat{\pi}) = \sum_{l=1}^{L}\hat{\pi}_l - 1 = 0, \tag{22}$$

where, for any parameter β, $\nabla_\beta = \partial f(\beta)/\partial\beta$ and μ is the Lagrange multiplier associated with the restriction $\sum_{l=1}^{L}\pi_l = 1$.

This system is similar to that of Imbens (1992). Simply, $\Pr(y|x,\theta)$ and H_s are replaced by their contaminated versions, $\Pr^*(y^*|x,\theta,\alpha)$ and $H_{s^*}^*$, and another set of first order conditions, (21), corresponding to the score function for the misclassification probabilities, is included. Performing operations similar to those in Imbens (1992), presented in Appendix A, the dependence on the discrete distribution assumed for X is removed and the parameter Q_y is included in the first order conditions. The maximum likelihood estimator for γ, the full vector of parameters of interest to be defined in the next subsection, is characterized by the system

$$g_{H^*}(\gamma) = H_{t^*}^* - I_{(s^*=t)}, \tag{23}$$

$$g_\theta(\gamma) = \nabla_\theta \ln \Pr^*(y^*|x,\theta,\alpha)$$

$$-\nabla_\theta \ln \sum_{s^*\in\mathcal{S}^*}\frac{H_{s^*}^*}{\sum_{y^*\in\mathcal{Y}_{s^*}^*}\sum_{y\in\mathcal{Y}}\alpha_{y^*y}Q_y}\sum_{y^*\in\mathcal{Y}_{s^*}^*}\Pr^*(y^*|x,\theta,\alpha), \tag{24}$$

$$g_\alpha(\gamma) = \nabla_{\alpha_{t^*t}} \ln \Pr^*(y^*|x,\theta,\alpha)$$

$$-\nabla_\alpha \ln \sum_{s^*\in\mathcal{S}^*}\frac{H_{s^*}^*}{\sum_{y^*\in\mathcal{Y}_{s^*}^*}\sum_{y\in\mathcal{Y}}\alpha_{y^*y}Q_y}\sum_{y^*\in\mathcal{Y}_{s^*}^*}\Pr^*(y^*|x,\theta,\alpha), \tag{25}$$

$$g_Q(\gamma) = \sum_{y \in \mathscr{Y}} \alpha_{y^* y} Q_y$$
$$- \frac{\Pr^*(y^*|x,\theta,\alpha)}{\sum_{s^* \in \mathscr{S}^*} \frac{H^*_{s^*}}{\sum_{y^* \in \mathscr{Y}^*_{s^*}} \sum_{y \in \mathscr{Y}} \alpha_{y^* y} Q_y} \sum_{y^* \in \mathscr{Y}^*_{s^*}} \Pr^*(y^*|x,\theta,\alpha)}. \tag{26}$$

Eqs. (23), (24), and (26) coincide with those obtained by Imbens (1992) if $\alpha_{y^* y} = 0$ for $y^* \neq y$.

Using this system as moment indicators, we may employ traditional GMM techniques. The objective function maximized is

$$\Upsilon_N(\gamma) = g_N(\gamma)' W_N g_N(\gamma)', \tag{27}$$

where $g_N(\gamma) = 1/N \sum_{i=1}^N g(\gamma, y_i, x_i, s_i)$ is the sample counterpart of the moment conditions $\mathrm{E}[g(\gamma, y_i, x_i, s_i)] = 0$, the moment indicators $g(\gamma, y_i, x_i, s_i)$ are given in (23)–(26), and W_N is a positive semi-definite weighting matrix.

4.2. Alternative estimators

The system of Eqs. (23)–(26) is easily adapted to incorporate available information on Q_y and $\alpha_{y^* y}$, producing several estimators with different degrees of efficiency, ranging from the case where there is no information on either, to the situation where there is exact information on both, including also the intermediate case where information is uncertain. Treatment of the first two cases follows closely that of Imbens (1992), while the analysis for the last is based mainly on Imbens and Lancaster (1994). The possibility of employing additional information on $Q^*_{y^*}$ instead of Q_y is not analysed separately because it follows straightforwardly from the case where we know the latter probabilities.

The main drawback of this approach is its requirement of large samples. This characteristic is inherited from the misclassification model for RS, a special case of our formulation for CB sampling. For the binary RS model with constant probability of misclassification, Copas (1988) found that the contribution of $\alpha_{y^* y}$ is mainly in the tails of the response function, which makes the estimation of $\alpha_{y^* y}$ very difficult in samples of reduced dimension.[11] To solve this problem he suggests a bias-corrected version of the maximum likelihood estimator appropriate for small misclassification probabilities to be employed in small samples. An important extension of our work would be the determination of the adjustment required in CB sampling.

[11] See, also, Cox and Snell (1989, p. 123).

4.2.1. No additional information on both α_{y^*y} and Q_y

When there is no supplementary information on both α_{y^*y} and Q_y, $\gamma = (H^*, \theta, \alpha, Q)$, where Q contains the C marginal choice probabilities Q_y.[12] Hence, because the number of estimated parameters equals the number of moment conditions, estimation consists in solving the system $g_N(\hat{\gamma}) = 0$, rendering the choice of the weighting matrix W_N irrelevant. Under the usual regularity conditions required for GMM, see Newey and McFadden (1994, Theorems 2.6, 3.4), the resulting estimator, $\hat{\gamma}$, converges almost surely to the true value γ_0 and satisfies

$$\sqrt{N}(\hat{\gamma} - \gamma_0) \xrightarrow{\mathrm{d}} \mathrm{N}(0, G^{-1}\Omega G'^{-1}), \tag{28}$$

where both Ω and G are square matrices of dimension $(J - 1 + k + D + C)$, defined by $\Omega = \mathrm{E}[g(\gamma, y, x, s) g(\gamma, y, x, s)']$ and $G = \mathrm{E}[\nabla_\gamma g(\gamma, y, x, s)']$.[13] When X is discrete this estimator coincides with the maximum likelihood estimator, being, thus, efficient. Otherwise, asymptotic efficiency, in the semi-parametric sense, can be proved by an analogous demonstration to that of Imbens (1992, Theorem 3.3); see Appendix B.

4.2.2. Exact information on either α_{y^*y} or Q_y or both α_{y^*y} and Q_y

There are situations in which we may assume we have exact information about either α_{y^*y} or Q_y, or both probabilities. When Q_y is estimated in a previous stage using a large auxiliary random sample, it is common practice in the literature on stratified samples to consider these estimates as exact (see, for example, Manski and Lerman, 1977; Manski and McFadden, 1981; Cosslett, 1981a, b; Imbens, 1992; Wooldridge, 1998, 1999). Moreover, in the misclassification literature, when α_{y^*y} is estimated in a previous stage using a validation sample, the sampling variability of the estimator obtained is not considered (Poterba and Summers, 1995).

Hence, estimation concerns $\gamma_\alpha = (H^*, \theta, Q)$, $\gamma_Q = (H^*, \theta, \alpha)$, or $\gamma_{Q\alpha} = (H^*, \theta)$, according to the information available on, respectively, α_{y^*y}, Q_y, or both α_{y^*y} and Q_y. As in Imbens' (1992) CB sampling GMM estimation framework, when Q_y is exactly known, the true quantities are substituted in the moment indicators, which produces an overidentified system of moment restrictions.[14] Thus, the optimal estimator, $\hat{\eta}$, respectively, $\hat{\gamma}_Q$, $\hat{\gamma}_\alpha$ or $\hat{\gamma}_{Q\alpha}$, obtained from using the weighting matrix $W_N = \Omega_N^{-1}$ in (27), where Ω_N is a consistent estimator

[12] Notice that, similarly to $H^*_{J^*}$, $Q_{C+1} = 1 - \sum_{y=0}^{C} Q_y$.

[13] Notice that these expectations are taken with respect to the sampling joint density incorporating the misclassification error, given in (10).

[14] As suggested by a referee and an associate editor, a Bayesian approach would be an alternative way to handle the additional information on both the marginal choice probabilities and the conditional misclassification probabilities, by including this information through a prior density for these parameters (see, for example, Poirier, 1998).

of Ω, converges almost surely to η_0 and satisfies

$$\sqrt{N}(\hat{\eta}-\eta_0)\overset{\mathrm{d}}{\rightarrow}\mathrm{N}[0,(G'\Omega^{-1}G)^{-1}] \tag{29}$$

with Ω and G defined below (28) but with γ replaced by η and obvious adaptations in terms of dimension. Similarly to $\hat{\gamma}$, asymptotic efficiency can be proved as in Imbens (1992, Theorem 3.3) (see Appendix B). Moreover, when the distribution of the covariates is discrete, by an analogous demonstration to that of Lemma 3.1 in Imbens (1992), it can be proved that this estimator is asymptotically as efficient as the constrained maximum likelihood estimator, attaining, thus, the Cramér–Rao lower bound. The incorporation of new information allows a reduction in the asymptotic variance when compared with (28) (see Appendix C).

Note that, if instead of knowing Q_y, we had information about its contaminated version, $Q^*_{y^*}$, we would merely replace $\sum_{y\in\mathscr{Y}}\alpha_{y^*y}Q_y$ by $Q^*_{y^*}$ in the moment conditions and take into account that Q contains probabilities $Q^*_{y^*}$ instead of Q_y. All other cases follow in a similar manner.

4.2.3. *Stochastic information on either α_{y^*y} or Q_y or both α_{y^*y} and Q_y*

If knowledge on α_{y^*y} or Q_y is inexact/stochastic, an estimation procedure based on that of Imbens and Lancaster (1994) can be adopted. Focussing mainly on a RS setting, they analyse the combination of information from two different data sets concerned with the same population and consider the particular situation where the information from an auxiliary data set is subject to sampling variation. The same kind of approach is also adopted by Lancaster and Imbens (1991) and Imbens and Hellerstein (1999), which deal with uncertain information on Q_y in models for pure CB samples.

Applying this framework here provides a new approach to integrating information on α_{y^*y} with the estimation procedure, since, as far as we know, sampling variation in estimators of α_{y^*y} obtained in a previous stage has been overlooked in the literature on misclassification in the response variable. The main consequence of this practice is that the variance of all parameters is underestimated.

Imbens and Lancaster (1994) assume that the analyst knows $\tilde{\rho}$, an estimator of a given feature of interest ρ_0, obtained from an auxiliary RS of dimension M. The analysis relies on the premise that both N and M go to infinity at the same rate, hence $w=M/N$ is a constant. Also, they require that $(\tilde{\rho}-\rho_0)$ is independent of the main sample and assume that all that is known about the auxiliary sample is M and $\tilde{\rho}$, which satisfies $\sqrt{M}(\tilde{\rho}-\rho_0)\overset{\mathrm{d}}{\rightarrow}\mathrm{N}(0,\Delta)$. This additional information is incorporated in the estimation process by introducing a new set of moment conditions

$$g_\rho(\rho)=\tilde{\rho}-\rho. \tag{30}$$

Here, $\tilde{\rho}$ is composed of $\tilde{\alpha}_{y^*y}$ and/or $\tilde{Q}_y$ and γ is to be estimated. As (30) represents further overidentifying moment conditions, the limiting distribution of this estimator of γ is described by (29). The re-definition of the matrices Ω and G is straightforward. The former, due to the independence of the two samples, is block diagonal, with the first block corresponding to the definition of this matrix in the case where no information on α_{y^*y} and Q_y is available and the second defined as Δ/w. The latter, as compared to the case where no information on α_{y^*y} and Q_y is available, has C, D or $(C+D)$ additional rows (corresponding to information on $\tilde{Q}_y$, $\tilde{\alpha}_{y^*y}$, or $\tilde{Q}_y$ and $\tilde{\alpha}_{y^*y}$, respectively), which are zero except for $\mathrm{E}[\nabla_\rho g_\rho(\rho)']$, which is the symmetric of an identity matrix of appropriate dimension.

As in Imbens and Lancaster (1994), it can be straightforwardly proved that the variance matrix of the resultant GMM estimator is bounded between those obtained when information is certain and when all parameters are estimated using the main sample (see Appendix C).

4.3. Monte Carlo experiment

This subsection reports a Monte Carlo simulation study designed to investigate the performance of some of the estimators proposed above for a situation characterized by different information on the aggregate choice probabilities, magnitudes of misclassification, and stratification designs. The experimental scheme is described in the next subsection and the simulation results are discussed in Section 4.3.2.

4.3.1. Experimental design

In all experiments we deal with CB binary data where each of the two responses defines one stratum, with stratum 0 and stratum 1 containing individuals choosing, respectively, alternatives $Y^*=0$ and $Y^*=1$. Constant misclassification probabilities between the two alternatives are assumed, such that $\alpha_{10}=\alpha_{01}\equiv\bar{\alpha}$, which is unknown in all experiments.

The contaminated proportions of each stratum in the population and in the sample are represented, respectively, by Q^* and H^* for stratum 1, and $1-Q^*$ and $1-H^*$ for stratum 0. The latent variable of interest, Y, conditional on the scalar $X=x$, is assumed to be generated by a logit model with no intercept, that is, $\Pr(1|x,\theta)=(1+\mathrm{e}^{-x\theta})^{-1}$. In order to produce a choice probability for $Y=1$, denoted by Q, equal to 0.9, the parameter θ was set equal to 1.46 and the covariate X was generated as a normal variate with mean 3 and variance 4.

Writing $P^*=\Pr^*(1|x,\theta,\bar{\alpha})$ and $P=\Pr(1|x,\theta)$, the conditional probability of observing $Y^*=1$ given $X=x$ and the marginal probability of $Y^*=1$ are, respectively,

$$P^*=\bar{\alpha}+(1-2\bar{\alpha})P \tag{31}$$

and

$$Q^* = \bar{\alpha} + (1 - 2\bar{\alpha})Q. \tag{32}$$

The base set of individual moment indicators employed to estimate all the models is written from (23)–(26) as[15]

$$g_{H^*}(H^*, \theta, \bar{\alpha}, Q) = H^* - y^*, \tag{33}$$

$$\begin{aligned} &g_\theta(H^*, \theta, \bar{\alpha}, Q) \\ &\quad = (1 - 2\bar{\alpha})\nabla_\theta P \left[\frac{y^* - P^*}{(1 - P^*)P^*} - \frac{H^*/Q^* - (1 - H^*)/(1 - Q^*)}{B^*} \right], \end{aligned} \tag{34}$$

$$\begin{aligned} &g_{\bar{\alpha}}(H^*, \theta, \bar{\alpha}, Q) \\ &\quad = (1 - 2P) \left[\frac{y^* - P^*}{(1 - P^*)P^*} - \frac{H^*/Q^* - (1 - H^*)/(1 - Q^*)}{B^*} \right], \end{aligned} \tag{35}$$

$$g_{Q^*}(H^*, \theta, \bar{\alpha}, Q) = \bar{\alpha} + Q(1 - 2\bar{\alpha}) - \frac{P^*}{B^*}, \tag{36}$$

where

$$B^* = \frac{1 - H^*}{1 - Q^*} + \left(\frac{H^*}{Q^*} - \frac{1 - H^*}{1 - Q^*} \right) P^*.$$

Four GMM estimators were compared in the Monte Carlo experiment. Two of them correspond to those employed with the ordinary logit model for CB sampling, assuming Q known, GMME1, or unknown, GMME2. The other two are their extended versions for misclassification developed in this paper, which are denoted MGMME1 and MGMME2, respectively.[16] For each estimator six different outcomes are presented, combining three levels of misclassification with two different stratification schemes. As in the simulation study of Hausman et al. (1998), we consider two cases of reduced error probabilities, $\bar{\alpha} = 0.02$ and 0.05, and a situation where the amount of error is substantial, $\bar{\alpha} = 0.20$. The two possibilities of stratification admitted are characterized by $H^* = 0.75$ and 0.50. The latter produces a sampling structure termed equal shares (Lancaster and Imbens, 1991), which is claimed to be close to an optimal design (Cosslett, 1981a; Lancaster and Imbens, 1991; Imbens, 1992).

All computations were done using the package S-Plus. Each experiment employed 1000 replications, with the sample sizes of 5000, as in Hausman et al. (1998), and 1000.

[15] Omitting the correction for CB sampling from $g_\theta(H^*, \theta, \bar{\alpha}, Q)$ and $g_\alpha(H^*, \theta, \bar{\alpha}, Q)$ yields the score functions given by Hausman et al. (1998) for a misclassified binary RS.

[16] To obtain GMME1 and GMME2 moment conditions (33), (34) and (36) are employed with $\alpha = 0$, which thus coincide with those of Imbens' (1992) simulation study concerning GMM estimators for CB sampling. For both GMME1 and MGMME1, estimation was performed with Q replaced by its known value in the moment conditions.

4.3.2. Results

In this subsection we examine the results of the Monte Carlo experiments just described. The analysis is based on Table 1 which contains the mean and the median bias in percentage terms and the standard deviation across replications of the estimates of the parameters of interest. In all cases, the mean and median of our estimators are substantially closer to the true values of parameters θ and Q than those of the uncorrected estimators.

Both the uncorrected estimators present significantly downward biased results for the mean and the median across replications. This distortion is more moderate in GMME1, due to the inclusion of additional information on Q. Also, while GMME1 always converges, GMME2 did not in several replications when $\bar{\alpha}=0.05$ or 0.20. In any case, even with a modest amount of misclassification, $\bar{\alpha}=0.02$, the difference between the replications mean and the true value of θ is 21.2% and 36.8% (24.2% and 44.0%) for, respectively, GMME1 and GMME2 with $H^*=0.75$ ($H^*=0.50$) and $N=5000$. Moreover, the results reveal that, while sample size does not seem to have a significant influence on the behaviour of the uncorrected estimators, sampling design appears to play an important role, since the equal shares experiment ($H^*=0.50$), where the level of stratification is higher, displays worse results not only in terms of bias but also in convergence failures. This arises because, in this experimental design, the total number of misclassified observations is larger. For example, for $\bar{\alpha}=0.05$, we have sampling misclassification probabilities, defined in Eq. (12), $\delta_{01}=0.32$, $\delta_{10}=0.01$ for $H^*=0.50$ and $\delta_{01}=0.14$, $\delta_{10}=0.02$ for $H^*=0.75$.

The modified estimators, in global terms, perform well, especially for the larger sample size, where there is little bias, either with Q known or otherwise. For $N=5000$ and $\bar{\alpha}\in\{0.02,0.05\}$, both MGMME1 and MGMME2 possess mean and median biases for $\hat{\theta}$ of less than 1%. Moreover, despite the bias increases when $\bar{\alpha}=0.20$, the maximum mean bias is only 2.2% (for MGMME1 with $H^*=0.50$), which is a substantial improvement over the corresponding uncorrected estimator which has a negative mean bias of around 50%. Naturally, for the smaller sample size ($N=1000$), their performance decays significantly, in particular, for $\bar{\alpha}=0.20$, where the corrected estimators achieve a mean bias of 22.3%. For this smaller sample size, the behaviour of the two modified estimators for θ is somewhat different, with MGMME1, in general, substantially less biased than MGMME2.[17] For example, for $\bar{\alpha}\in\{0.02,0.05\}$ the larger bias in mean and median of $\hat{\theta}$ for

[17] The only exception being when $\bar{\alpha}=0.20$ and $H=0.75$. For this case, MGMME1 displays a large dispersion due to six outliers where the estimate of θ is larger than 20, which result in its poor behaviour in terms of the mean across replications. However, the median of MGMME1 is closer to $\theta_0=1.46$ than that of MGMME2.

Table 1
Summary statistics for GMM estimators from 1000 replications

$\theta_0 = 1.46$, $Q_0 = 0.90$												
				$\hat{\theta}$			$\hat{\alpha}$			$\hat{Q}$		
N	$\bar{\alpha}$	H^*	Estimator	Bias		SD	Bias		SD	Bias		SD
				Mean	Med.		Mean	Med.		Mean	Med.	
5000	0.02	0.75	GMME1	−0.212	−0.213	0.014	—	—	—	—	—	—
			GMME2	−0.368	−0.368	0.022	—	—	—	−0.058	−0.058	0.005
			MGMME1	0.004	0.003	0.036	−0.015	−0.015	0.002	—	—	—
			MGMME2	0.000	0.000	0.064	−0.015	−0.015	0.002	−0.001	−0.001	0.006
		0.50	GMME1	−0.242	−0.242	0.013	—	—	—	—	—	—
			GMME2	−0.440	−0.441	0.016	—	—	—	−0.089	−0.089	0.005
			MGMME1	0.004	0.005	0.033	0.002	−0.005	0.002	—	—	—
			MGMME2	−0.006	−0.009	0.058	0.013	0.015	0.002	−0.002	−0.002	0.006
	0.05	0.75	GMME1	−0.331	−0.332	0.008	—	—	—	—	—	—
			GMME2	−0.565	−0.565	0.016	—	—	—	−0.112	−0.112	0.006
			MGMME1	0.009	0.009	0.049	0.007	0.014	0.004	—	—	—
			MGMME2	0.006	0.005	0.083	0.005	0.010	0.004	−0.001	−0.000	0.006
		0.50	GMME1	−0.363	−0.362	0.011	—	—	—	—	—	—
			GMME2[a]	−0.622	−0.622	0.016	—	—	—	−0.146	−0.147	0.006
			MGMME1	0.007	0.008	0.042	−0.003	0.000	0.003	—	—	—
			MGMME2	−0.002	−0.006	0.079	0.003	−0.002	0.004	−0.002	−0.002	0.007
	0.20	0.75	GMME1	−0.466	−0.465	0.010	—	—	—	—	—	—
			GMME2[b]	−0.833	−0.835	0.016	—	—	—	−0.263	−0.265	0.008
			MGMME1	0.009	0.005	0.070	0.005	0.003	0.010	—	—	—
			MGMME2	0.012	0.005	0.160	−0.000	−0.002	0.010	−0.000	0.000	0.007
		0.50	GMME1	−0.501	−0.502	0.007	—	—	—	—	—	—
			GMME2[c]	−0.848	−0.848	0.015	—	—	—	−0.279	−0.280	0.010
			MGMME1	0.022	0.017	0.074	0.007	0.007	0.011	—	—	—
			MGMME2	−0.009	−0.018	0.171	0.000	−0.003	0.010	−0.003	−0.003	0.008
1000	0.02	0.75	GMME1	−0.219	−0.219	0.020	—	—	—	—	—	—
			GMME2	−0.374	−0.373	0.039	—	—	—	−0.059	−0.059	0.009
			MGMME1	−0.010	−0.012	0.075	−0.271	−0.033	0.133	—	—	—
			MGMME2	−0.011	−0.028	0.123	−0.012	−0.030	0.003	−0.004	−0.003	0.011
		0.50	GMME1	−0.264	−0.265	0.014	—	—	—	—	—	—
			GMME2	−0.461	−0.461	0.026	—	—	—	−0.091	−0.092	0.008
			MGMME1	0.011	0.012	0.055	−0.100	0.071	0.111	—	—	—
			MGMME2	0.065	0.062	1.117	−0.001	−0.002	0.003	0.010	0.010	0.011

Table 1 (*Continued.*)

$\theta_0 = 1.46,\ Q_0 = 0.90$												
				$\hat{\theta}$			$\hat{\alpha}$			$\hat{Q}$		
N	$\bar{\alpha}$	H^*	Estimator	Bias		SD	Bias		SD	Bias		SD
				Mean	Med.		Mean	Med.		Mean	Med.	
	0.05	0.75	GMME1	−0.319	−0.320	0.013	—	—	—	—	—	—
			GMME2	−0.526	−0.526	0.039	—	—	—	−0.095	−0.095	0.009
			MGMME1	−0.020	−0.024	0.073	−0.099	−0.106	0.006	—	—	—
			MGMME2	−0.047	−0.057	0.158	−0.101	−0.109	0.006	−0.005	−0.005	0.012
		0.50	GMME1	−0.369	−0.369	0.008	—	—	—	—	—	—
			GMME2[a]	−0.604	−0.608	0.098	—	—	—	−0.134	−0.135	0.015
			MGMME1	−0.030	−0.029	0.046	−0.075	−0.080	0.003	—	—	—
			MGMME2	−0.031	−0.036	0.119	−0.073	−0.073	0.006	−0.001	−0.001	0.012
	0.20	0.75	GMME1	−0.457	−0.457	0.007	—	—	—	—	—	—
			GMME2[b]	−0.801	−0.803	0.095	—	—	—	−0.237	−0.238	0.018
			MGMME1	0.121	−0.016	2.59	−0.118	−0.135	0.023	—	—	—
			MGMME2	−0.047	−0.087	0.411	−0.142	−0.150	0.016	−0.006	−0.006	0.013
		0.50	GMME1	−0.505	−0.503	0.123	—	—	—	—	—	—
			GMME2[c]	−0.810	−0.859	0.362	—	—	—	−0.272	−0.290	0.086
			MGMME1	0.048	0.039	0.160	0.038	0.026	0.021	—	—	—
			MGMME2	0.223	0.176	0.378	0.008	0.003	0.024	0.009	0.010	0.010

[a] Convergence in 969 and 977 replications for, respectively, $N = 5000$ and 1000.
[b] Convergence in 946 and 978 replications for, respectively, $N = 5000$ and 1000.
[c] Convergence in 769 and 810 replications for, respectively, $N = 5000$ and 1000.

MGMME2 (6.5% for $\bar{\alpha} = 0.02$ and $H^* = 0.50$) is more than double that of MGMME1 (3.0% for $\bar{\alpha} = 0.05$ and $H^* = 0.50$). Hence, the use of additional information appears to be especially important when the sample is smaller. Conversely, the stratification design does not seem to affect MGMME1 and MGMME2, as both means and medians of $\hat{\theta}$, $\hat{\alpha}$ and $\hat{Q}$ behave in a quite similar way for both $H^* = 0.75$ and 0.50. To summarize, although the performance of the corrected estimators is negatively affected by an increase in the probability of misclassification and, mainly, by a reduction in sample size, these estimators are clearly preferable to the uncorrected alternatives.

However, examining the column SD containing the standard deviations of $\hat{\theta}$ across the replications, analogously to the simulation study of Hausman et al. (1998), it becomes clear that the variability of the estimates is larger in the corrected estimators. This loss of precision relative to their uncorrected versions reflects the fact that misclassification is taken into account in the estimation procedure. Concentrating attention on the results for MGMME1 and MGMME2, the conclusions suggested are those one could intuitively expect. Firstly, an increase in the misclassification probability and a reduction in sample size negatively affects the accuracy of these estimators. Secondly, in conformity to when misclassification is absent, as documented in the simulation study of Imbens (1992), the inclusion of information on Q permits considerable gains in efficiency, as standard deviations of MGMME1 are almost one half of those of MGMME2, with one exception only (see footnote 17). Furthermore, MGMME1 is more robust to contamination with, for example, the replications standard deviation of this estimator for $\bar{\alpha}=0.20$ smaller than that of MGMME2 with $\bar{\alpha}=0.05$.

5. A score test to detect the presence of misclassification

Although the presence of misclassification, even when insubstantial, generates inconsistent estimators when standard estimation methods are employed, to the best of our knowledge, there are no specification tests for the detection of this problem, with the exception of Copas (1988), who suggests a score test for a binary logit model under the assumption of RS.

In this section, we suggest an appropriate test statistic in the context of CB sampling. Following Newey and McFadden (1994), we outline a score test in the GMM framework. This test is useful for situations in which misrecording error is suspected and there is no additional information on misclassification probabilities. We formulate the test in Section 5.1 and present a small Monte Carlo simulation study in Section 5.2.

5.1. General form

The basic idea concerns testing whether the parameters α_{y^*y} are zero for $y \neq y^*$. As usual for score tests, only estimation of the restricted model is required and the features of the unrestricted model are evaluated at the restricted estimator. Thus, using conventional estimators for CB samples (Imbens, 1992), we merely estimate the vector $\phi_Q=(H,\theta,0)$ or $\phi=(H,\theta,Q,0)$, according to the availability of information on Q_y or otherwise.

The null hypothesis is H$_0$: $\alpha_{y^*y}=0$ for $y \neq y^*$, for which the score test statistic (see Newey and McFadden, 1994, Theorem 9.2.) is given by

$$T=Ng_N'\Omega_N^{-1}G_N V_N G_N'\Omega_N^{-1}g_N, \tag{37}$$

where Ω_N, G_N and V_N are consistent estimators of, respectively, Ω, G and $V=(G'\Omega^{-1}G)^{-1}$, all of them, as well as g_N, the sample counterpart of (23) –(26), evaluated at consistent estimators of the parameters of the restricted model, $\hat{\phi}$ or $\hat{\phi}_Q$. Under the null hypothesis T converges in distribution to a chi-square random variable with D degrees of freedom.

In the just identified case, Eq. (37) simplifies to

$$T = Ng'_{\alpha N}\Omega_N^{\alpha\alpha}g_{\alpha N} \tag{38}$$

with $g_{\alpha N}$ being the sample counterpart of (25) and $\Omega_N^{\alpha\alpha}$ a consistent estimator of $\Omega^{\alpha\alpha} = [\mathrm{E}(g_\alpha g'_\alpha) - \mathrm{E}(g_\alpha g'_\varphi)\mathrm{E}(g_\varphi g'_\varphi)^{-1}\mathrm{E}(g_\alpha g'_\varphi)]^{-1}$, where $\varphi = (H^*, \theta, Q)$ is a $(J-1+k+C)$-dimensioned vector, obtained by excluding α from the vector γ.

A calculation of V_N and Ω_N should require numerical methods, as the expectations present in their definitions involve integration over $\mathscr{X}$ and summation over $\mathscr{Y}^*_{s^*}$ and $\mathscr{S}^*$. In addition, the marginal distribution of the covariates would need to be known, which is unlikely in practice. To circumvent this difficulty, these quantities may be estimated by simple averages or, as $\Pr(y|x,\theta)$ is required for GMM estimation, we can perform the summation over $\mathscr{Y}^*_{s^*}$ and $\mathscr{S}^*$ of both the cross products contained in Ω and the derivatives of the moment conditions in G multiplied by $(H^*_{s^*}/Q^*_{s^*})\Pr^*(y^*|x,\theta,\alpha)$. Then, instead of integrating over $\mathscr{X}$, we can either calculate simple averages or, following Cosslett (1993), weight those features either by

$$\frac{1}{N}\sum_{i=1}^{N}\frac{Q^*_{s^*_i}}{H^*_{s^*_i}} \quad \text{or} \quad \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{s^*_i\in\mathscr{S}^*}\frac{Q^*_{s^*_i}}{H^*_{s^*_i}}\sum_{y^*_i\in\mathscr{Y}^*_{s^*}}\Pr{}^*(y^*_i|x_i,\theta,\alpha)\right]^{-1}.$$

Thus, despite the lack of an explicit analytical expression for T, which would greatly simplify its implementation, this test can be easily applied because we have analytical expressions for the moment conditions and their derivatives for the unrestricted model. Moreover, the score statistic T can be straightforwardly adapted to detect misclassification in RS.[18] As the number of moment conditions equals the number of estimated parameters, the test statistic is given in (38). Note that, for RS, calculation of $\Omega_N^{\alpha\alpha}$ does not require knowledge of $f(x)$, since in RS analysis can be conducted conditional on the covariates.

5.2. Monte Carlo experiment

The object of this subsection is a brief examination of the finite sample properties of the score test T of (37). We focus on an analysis of its ability to detect the presence of small and moderate amounts of misclassification under different conditions, namely, with two stratification designs and two sample sizes. In the power analysis we employed two different misclassification probabilities.

[18] In this case, the moment conditions of the unrestricted model are given by (24) and (25) with $H^*_y = Q^*_y$. To estimate the restricted model, we use (24) with $H^*_y = Q^*_y$ and $\alpha_{y^*y} = 0$ for $y \neq y^*$.

Table 2
Score test: empirical size and power (nominal size of 5%)

N	H^*	Size	Power	
			$\bar{\alpha}=0.02$	$\bar{\alpha}=0.05$
250	0.75	3.3	20.3	33.8
	0.50	4.7	32.5	47.7
750	0.75	4.1	41.7	67.1
	0.50	5.3	61.9	84.3

All the experiments use a similar experimental structure to that defined in Section 4.3.1, with the difference that we only consider the two smallest misclassification probabilities, $\bar{\alpha}=0.02$, and 0.05, work with two smaller samples sizes, $N=250$ and 750, and replicate each experiment 10,000 times. Moreover, the analysis concentrates on the case where the researcher has no information about the error-free marginal choice probabilities, Q. Thus, the appropriate restricted estimator is GMME2 defined in Section 4.3.1.

As we are dealing with a constant probability of misclassification for both values of the response variable, we test $\bar{\alpha}=0$. The test statistic is computed according to (38), since number of estimated parameters and moment conditions is the same. In (38), $g_{\alpha N}$ is the mean of (35) and the matrix $\Omega_N^{\alpha\alpha}$ is calculated by averaging the summation over the two values of Y^* of the cross-products of the moment conditions (33)–(36) multiplied by $(H_{y^*}^*/Q_{y^*}^*)\Pr^*(y^*|x,\theta,\bar{\alpha})$.

The results are summarized in Table 2, which reports the estimated size and power of the test based on the 5% nominal $\chi^2_{(1)}$ critical value of 3.84. The performance in terms of power is as expected, given the small amounts of misclassification considered, improving substantially for the larger sample size $N=750$. Clearly, the best performances occur under conditions which are more destructive for the uncorrected estimator. Thus, the favourable effects on power of larger levels of misclassification are clear, even though the results are reasonable for $\bar{\alpha}=0.02$. The superior performance of T for the equal shares design is underpinned by results in Section 4.3.2, where we concluded that estimators which do not account for misrecording are more severely affected for $H^*=0.50$ (see GMME1 and GMME2 in Table 1 for both values of H^*).

More surprisingly, the results show that the major determinant of the actual size is the sampling design. While for the equal shares structure the estimated size is close to the nominal level of 5% for both sample sizes, with $H^*=0.75$ the test is always slightly undersized, though it exhibits some improvement for $N=750$, which is probably related to the approximate optimality of the former design.

6. Conclusion

The presence of misclassification in the response variable in CB samples creates a complex structure where the observed data is subject to a double weighting due to the joint effects of stratification and misclassification. On the one hand, misclassification destroys the shape of all sampling distributions, even that of the covariates, which are assumed error-free variables relative to the population distribution. On the other hand, the sampling design makes the misclassification probabilities differ between the population and the sample. The former distributional distortions cause the failure of all parametric and semi-parametric estimation methods usually employed in CB sampling, resulting in inconsistent estimators for the parameters of interest, while the latter produces similar effects if the analog of the common practice of ignoring the CB design in multiplicative intercept models is applied and the model adopted merely incorporates measurement error.

Under these conditions, econometric inference must account for both the sampling design employed and the mismeasurement of the variable of interest. Thus, in this paper we extent Imbens' (1992) efficient GMM estimation for CB samples to incorporate misclassification, providing a framework where the inclusion of additional information on either the misclassification probabilities, the marginal error-free choice probabilities, or both quantities is straightforward. Using this setting, we also outline a score test sensitive to this form of measurement error in order to provide a practical basis for the assessment of whether to employ the usual estimation procedures in CB samples or the estimators accounting for misclassification proposed in this paper. A Monte Carlo investigation documents the good performance of the proposed estimation procedures in large samples and a moderate probability of misrecording. Also, the score test possesses encouraging size and power properties in samples of moderate size.

Acknowledgements

I am grateful to two referees, an associate editor, Andrew Chesher, João Santos Silva, and especially Richard J. Smith, for helpful comments and suggestions. Financial support from Fundação para a Ciência e Tecnologia is gratefully acknowledged.

Appendix A. Derivation of the GMM moment conditions

The transformation of (18)–(22) into (23)–(26) involves the derivation of the maximum likelihood estimator for π and its substitution into (20) and (21), and the incorporation of the definition of $Q^*_{y^*}$, given in (9).

Define the maximum likelihood estimator of $Q^*_{y^*}$, according to Eq. (9), as

$$\hat{Q}^*_{y^*} = \sum_{l=1}^{L} \hat{\pi}_l \operatorname{Pr}^*(y^* | x^l, \hat{\theta}, \hat{\alpha}) \tag{A.1}$$

so that $Q^*_{s^*}$ can be estimated by $\hat{Q}^*_{s^*} = \sum_{y^* \in \mathcal{Y}^*_{s^*}} \hat{Q}^*_{y^*}$. Moreover, as $\hat{\mu} = 0$, using Eq. (19) we obtain[19]

$$\hat{\pi}_m = \frac{1}{N} \sum_{i=1}^{N} I_{(l_i = m)} \left[\frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{y^*_i \in \mathcal{Y}^*_{s^*}} \operatorname{Pr}^*(y^*_i | x^m, \hat{\theta}, \hat{\alpha})}{\hat{Q}^*_{s^*_i}} \right]^{-1}$$

$$= \frac{1}{N} \sum_{i=1}^{N} I_{(l_i = m)} \left[\sum_{s^* \in \mathcal{S}^*} \frac{\hat{H}_{s^*}}{\hat{Q}^*_{s^*}} \sum_{y^* \in \mathcal{Y}^*_{s^*}} \operatorname{Pr}^*(y^* | x^m, \hat{\theta}, \hat{\alpha}) \right]^{-1}.$$

Hence, the dependence of (20) and (21) on π is removed by replacing $\hat{\pi}_l$ in the last term of both expressions. As these terms are quite similar, only the calculations for (20) are presented

$$\sum_{i=1}^{N} \frac{1}{\hat{Q}^*_{s^*_i}} \sum_{l=1}^{L} \hat{\pi}_l \sum_{y^*_i \in \mathcal{Y}^*_{s^*}} \nabla_\theta \operatorname{Pr}^*(y^*_i | x^l, \hat{\theta}, \hat{\alpha})$$

$$= \sum_{i=1}^{N} \frac{1}{\hat{Q}^*_{s^*_i}} \sum_{l=1}^{L} \frac{1}{N} \sum_{i'=1}^{N} I_{(l_{i'} = l)} \left[\sum_{s^* \in \mathcal{S}^*} \frac{\hat{H}^*_{s^*}}{\hat{Q}^*_{s^*}} \sum_{y^* \in \mathcal{Y}^*_{s^*}} \operatorname{Pr}^*(y^* | x^l, \hat{\theta}, \hat{\alpha}) \right]^{-1}$$

$$\times \sum_{y^*_i \in \mathcal{Y}^*_{s^*}} \nabla_\theta \operatorname{Pr}^*(y^*_i | x^l, \hat{\theta}, \hat{\alpha})$$

$$= \sum_{i=1}^{N} \frac{1}{\hat{Q}^*_{s^*_i}} \frac{1}{N} \sum_{i'=1}^{N} \left[\sum_{s^* \in \mathcal{S}^*} \frac{\hat{H}^*_{s^*}}{\hat{Q}^*_{s^*}} \sum_{y^* \in \mathcal{Y}^*_{s^*}} \operatorname{Pr}^*(y^* | x^{l_{i'}}, \hat{\theta}, \hat{\alpha}) \right]^{-1}$$

$$\times \sum_{y^*_i \in \mathcal{Y}^*_{s^*}} \nabla_\theta \operatorname{Pr}^*(y^*_i | x^{l_{i'}}, \hat{\theta}, \hat{\alpha})$$

[19] This is a result of multiplying (19) by $\hat{\pi}_m$ and summing over m.

$$
\begin{aligned}
&= \sum_{i'=1}^{N} \left[\sum_{s^* \in \mathscr{S}^*} \frac{\hat{H}^*_{s^*}}{\hat{Q}^*_{s^*}} \sum_{y^* \in \mathscr{Y}^*_{s^*}} \Pr{}^*(y^*|x^{l_{i'}}, \hat{\theta}, \hat{\alpha}) \right]^{-1} \\
&\quad \times \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\hat{Q}^*_{s^*_i}} \sum_{y^*_i \in \mathscr{Y}^*_{s^*}} \nabla_\theta \Pr{}^*(y^*_i|x^{l_{i'}}, \hat{\theta}, \hat{\alpha}) \\
&= \sum_{i'=1}^{N} \left[\sum_{s^* \in \mathscr{S}^*} \frac{\hat{H}^*_{s^*}}{\hat{Q}^*_{s^*}} \sum_{y^* \in \mathscr{Y}^*_{s^*}} \Pr{}^*(y^*|x^{l_{i'}}, \hat{\theta}, \hat{\alpha}) \right]^{-1} \\
&\quad \times \sum_{s^* \in \mathscr{S}^*} \frac{\hat{H}_{s^*}}{\hat{Q}_{s^*}} \sum_{y^* \in \mathscr{Y}^*_{s^*}} \nabla_\theta \Pr{}^*(y^*|x^{l_{i'}}, \hat{\theta}, \hat{\alpha}) \\
&= \sum_{i=1}^{N} \nabla_\theta \ln \sum_{s^* \in \mathscr{S}^*} \frac{\hat{H}^*_{s^*}}{\hat{Q}^*_{s^*}} \sum_{y^* \in \mathscr{Y}^*_{s^*}} \Pr{}^*(y^*|x^{l_i}, \hat{\theta}, \hat{\alpha}).
\end{aligned}
$$

On the other hand, the additional set of moment conditions associated with the definition of Q_y, given in (26), results from the replacement of $\hat{\pi}_l$ in Eq. (A.1):

$$
\begin{aligned}
\hat{Q}^*_{y^*} &= \sum_{l=1}^{L} \frac{1}{N} \sum_{i=1}^{N} I_{(l_i=l)} \left[\sum_{s^* \in \mathscr{S}^*} \frac{\hat{H}_{s^*}}{\hat{Q}^*_{s^*}} \sum_{y^* \in \mathscr{Y}^*_{s^*}} \Pr{}^*(y^*|x^{l}, \hat{\theta}, \hat{\alpha}) \right]^{-1} \\
&\quad \times \Pr{}^*(y^*|x^{l}, \hat{\theta}, \hat{\alpha}) \\
&= \Pr{}^*(y^*|x^{l_i}, \hat{\theta}, \hat{\alpha}) \left[\sum_{s^* \in \mathscr{S}^*} \frac{\hat{H}_{s^*}}{\hat{Q}^*_{s^*}} \sum_{y^* \in \mathscr{Y}^*_{s^*}} \Pr{}^*(y^*|x^{l_i}, \hat{\theta}, \hat{\alpha}) \right]^{-1}.
\end{aligned}
$$

As we are interested in either estimating Q_y or including its known value in the estimation procedure, we substitute $\hat{Q}^*_{y^*}$ by $\sum_{y \in \mathscr{Y}} \hat{\alpha}_{y^*y} \hat{Q}_y$ (and, obviously, $\hat{Q}^*_{s^*}$ by $\sum_{y^* \in \mathscr{Y}^*_{s^*}} \sum_{y \in \mathscr{Y}} \hat{\alpha}_{y^*y} \hat{Q}_y$) in all the first order conditions. If the information on Q_y is contaminated, this substitution is omitted.

Appendix B. Efficiency of GMM estimators for CB samples with misclassification in the response variable

Following Imbens (1992), the efficiency of our estimators for both when the exact values of α_{y^*y} or/and Q_y are known and when no additional information on these quantities is available, can be proved by showing that the Cramér–Rao lower bounds associated with a sequence of parametric models which satisfy the same regularity conditions as our model, converges to the asymptotic covariance matrix of our semi-parametric estimator.

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

Parametric models, indexed by ε, result from substituting $f(x,\varpi)$ in (10):

$$h_\varepsilon^*(y^*, x, s^*) = H_{s^*}^* \frac{\Pr^*(y^*|x,\theta,\alpha) f_\varepsilon(x) \sum_{l=1}^{L_\varepsilon} \phi_{lx}\varpi_l}{\sum_{y^* \in \mathscr{Y}_{s^*}^*} \sum_{l=1}^{L_\varepsilon} \varpi_l \int_{\mathscr{X}_l} \Pr^*(y^*|x,\theta,\alpha) f_\varepsilon(x) \phi_{lx}\,\mathrm{d}x}$$

which, as $f_\varepsilon(x)$ is a known function, depend on $(J - 1 + k + L_\varepsilon - 1)$ unknown parameters $(H_{s^*}^*, \theta, \phi_{lx})$.

Constructing the log-likelihood function, taking the first order derivatives and noting that the maximum likelihood estimator for $Q_{y^*}^*$ is

$$\hat{Q}_{y^*}^* = \sum_{l=1}^{L_\varepsilon} \varpi_l \int_{\mathscr{X}_l} \Pr^*(y^*|x,\hat{\theta},\hat{\alpha}) f_\varepsilon(x) \phi_{lx}\,\mathrm{d}x$$

allows the dependence on ϖ_l to be removed by the same procedure employed to remove dependence on $\hat{\pi}_l$ in systems (18)–(22), described in Appendix A. The resultant moment indicators are

$$g_{H^*\varepsilon}(\gamma) = H_{t^*}^* - I_{(s^*=t)}, \tag{B.1}$$

$$g_{\theta\varepsilon}(\gamma) = \nabla_\theta \ln \Pr^*(y^*|x,\theta,\alpha) - \nabla_\theta \ln \sum_{s^* \in \mathscr{S}^*} \frac{H_{s^*}^*}{\sum_{y^* \in \mathscr{Y}_{s^*}^*} \sum_{y \in \mathscr{Y}} \alpha_{y^* y} Q_y} \times \sum_{l=1}^{L_\varepsilon} \phi_{lx} \int_{\mathscr{X}_l} f_\varepsilon(x) \sum_{y^* \in \mathscr{Y}_{s^*}^*} \Pr^*(y^*|x,\theta,\alpha)\,\mathrm{d}x, \tag{B.2}$$

$$g_{\alpha\varepsilon}(\gamma) = \nabla_{\alpha_{t^*t}} \ln \Pr^*(y^*|x,\theta,\alpha) - \nabla_\alpha \ln \sum_{s^* \in \mathscr{S}^*} \frac{H_{s^*}^*}{\sum_{y^* \in \mathscr{Y}_{s^*}^*} \sum_{y \in \mathscr{Y}} \alpha_{y^* y} Q_y} \times \sum_{l=1}^{L_\varepsilon} \phi_{lx} \int_{\mathscr{X}_l} f_\varepsilon(x) \sum_{y^* \in \mathscr{Y}_{s^*}^*} \Pr^*(y^*|x,\theta,\alpha)\,\mathrm{d}x, \tag{B.3}$$

$$g_{Q\varepsilon}(\gamma) = \sum_{y \in \mathscr{Y}} \alpha_{y^* y} Q_y - \frac{\sum_{l=1}^{L_\varepsilon} \phi_{lx} \int_{\mathscr{X}_l} \Pr^*(y^*|x,\theta,\alpha) f_\varepsilon(x)\,\mathrm{d}x}{\sum_{s^* \in \mathscr{S}^*} \frac{H_{s^*}^*}{\sum_{y^* \in \mathscr{Y}_{s^*}^*} \sum_{y \in \mathscr{Y}} \alpha_{y^* y} Q_y} \sum_{l=1}^{L_\varepsilon} \phi_{lx} \int_{\mathscr{X}_l} f_\varepsilon(x) \sum_{y^* \in \mathscr{Y}_{s^*}^*} \Pr^*(y^*|x,\theta,\alpha)\,\mathrm{d}x}. \tag{B.4}$$

To compare the asymptotic covariance matrix of this parametric estimator with that of our semi-parametric estimators, define $\mathrm{E}_\varepsilon[\Pr^*(y^*|x,\theta,\alpha)] = \sum_{l=1}^{L_\varepsilon} \phi_{lx} \int_{\mathscr{X}_l} \Pr^*(y^*|x,\theta,\alpha) f_\varepsilon(x)\,\mathrm{d}x$ and $\mathrm{E}_\varepsilon[\nabla_\theta \Pr^*(y^*|x,\theta,\alpha)]$, $\mathrm{E}_\varepsilon[\nabla_\alpha \Pr^*(y^*|x,\theta,\alpha)]$, $\mathrm{E}_\varepsilon[\nabla_{\theta\theta'} \Pr^*(y^*|x,\theta,\alpha)]$ and $\mathrm{E}_\varepsilon[\nabla_{\alpha\alpha'} \Pr^*(y^*|x,\theta,\alpha)]$ similarly. Hence, it is clear that this system corresponds to (23)–(26) with $\Pr^*(y^*|x,\theta,\alpha)$, $\nabla_\theta \Pr^*(y^*|x,\theta,\alpha)$ and $\nabla_\alpha \Pr^*(y^*|x,\theta,\alpha)$ replaced by their expectations.

Assuming that $\Pr^*(y^*|x,\theta,\alpha)$, $\nabla_\theta \Pr^*(y^*|x,\theta,\alpha)$, $\nabla_\alpha \Pr^*(y^*|x,\theta,\alpha)$, $\nabla_{\theta\theta'} \Pr^*(y^*|x,\theta,\alpha)$ and $\nabla_{\alpha\alpha'} \Pr^*(y^*|x,\theta,\alpha)$ are continuously differentiable with respect to x, there is uniform convergence of $\mathrm{E}_\varepsilon[\Pr^*(y^*|x,\theta,\alpha)]$, $\mathrm{E}_\varepsilon[\nabla_\theta \Pr^*(y^*|x,\theta,\alpha)]$, $\mathrm{E}_\varepsilon[\nabla_\alpha \Pr^*(y^*|x,\theta,\alpha)]$, $\mathrm{E}_\varepsilon[\nabla_{\theta\theta'} \Pr^*(y^*|x,\theta,\alpha)]$ and $\mathrm{E}_\varepsilon[\nabla_{\alpha\alpha'} \Pr^*(y^*|x,\theta,\alpha)]$ to $\Pr^*(y^*|x,\theta,\alpha)$, $\nabla_\theta \Pr^*(y^*|x,\theta,\alpha)$, $\nabla_\alpha \Pr^*(y^*|x,\theta,\alpha)$, $\nabla_{\theta\theta'} \Pr^*(y^*|x,\theta,\alpha)$ and $\nabla_{\alpha\alpha'} \Pr^*(y^*|x,\theta,\alpha)$, respectively. Thus, in the case when there is no information on α_{y^*y} and Q_y, the limits of $\Omega_\varepsilon = \mathrm{E}_\varepsilon[g_\varepsilon(\gamma,y,x,s)g_\varepsilon(\gamma,y,x,s)']$ and $G_\varepsilon = \mathrm{E}_\varepsilon[\nabla_\gamma g_\varepsilon(\gamma,y,x,s)']$ equal those of Ω and G and the covariance matrix, $G_\varepsilon^{-1}\Omega_\varepsilon G_\varepsilon'^{-1}$, the Cramér–Rao bound, converges to $G^{-1}\Omega G'^{-1}$, which implies that our semi-parametric estimator is efficient. Analogously, in presence of exact information on α_{y^*y} or/and Q_y, merely by re-defining $\Omega_\varepsilon = \mathrm{E}_\varepsilon[g_\varepsilon(\eta,y,x,s)g_\varepsilon(\eta,y,x,s)']$ and $G_\varepsilon = \mathrm{E}_\varepsilon[\nabla_\eta g_\varepsilon(\eta,y,x,s)']$ the same conclusion is reached, since the covariance matrix of the optimal parametric-based GMM estimator $(G_\varepsilon \Omega_\varepsilon^{-1} G_\varepsilon')^{-1}$ converges to $(G\Omega^{-1}G')^{-1}$.

Appendix C. Comparison of the asymptotic covariance matrices of the alternative GMM estimators

In order to compare the asymptotic covariance matrix of the GMM estimator incorporating additional exact information on either α_{y^*y} or Q_y or both α_{y^*y} and Q_y and that where no additional information is available, partition the vector of parameters of interest $\gamma = (H,\theta,\alpha,Q)$ into $\gamma = (\gamma_1,\gamma_2)$, where γ_1 equals γ_Q, γ_α or $\gamma_{Q\alpha}$, and γ_2 contains, respectively, Q, α or Q and α. Partition the base set of moment indicators accordingly, $g(\gamma,y,x,s)' = [g_{\gamma_1}(\gamma,y,x,s)', g_{\gamma_2}(\gamma,y,x,s)']$, and define the asymptotic variance matrix of $\sqrt{N}(\hat{\gamma} - \gamma_0)$ as $V = (G'\Omega^{-1}G)^{-1}$, partitioned for γ_1, conformably as

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}^{-1}.$$

With no additional information, if $\hat{\gamma}$ is the resultant GMM estimator, $V[\sqrt{N}(\hat{\gamma}_1 - \gamma_{1_0})] = (V_{11} - V_{12}V_{22}^{-1}V_{21})^{-1}$, otherwise, with exact information on γ_2, if $\tilde{\gamma}_1$ is the GMM estimator for γ_1, $V[\sqrt{N}(\tilde{\gamma}_1 - \gamma_{1_0})] = V_{11}^{-1} \leqslant V[\sqrt{N}(\hat{\gamma}_1 - \gamma_{1_0})]$.

When information on either α_{y^*y} or Q_y or both α_{y^*y} and Q_y is stochastic, a set of moment indicators is added to $g(\gamma,y,x,s)$, yielding the moment indicators $k(\gamma, y, x, s)' = [g_{\gamma_1}(\gamma, y, x, s), g_{\gamma_2}(\gamma, y, x, s), g_\rho(\rho)]$. If $\breve{\gamma} = (\breve{H}^*, \breve{\theta}, \breve{\alpha}, \breve{Q})$ is the resultant GMM estimator, $V[\sqrt{N}(\breve{\gamma}_1 - \gamma_{1_0})] = [V_{11} - V_{12}(V_{22} + w\Delta^{-1})^{-1}V_{21}]^{-1}$ which is larger than the asymptotic variance matrix in the case when information is exact, $V[\sqrt{N}(\tilde{\gamma}_1 - \gamma_{1_0})]$, and smaller than that obtained when there is no additional information, $V[\sqrt{N}(\hat{\gamma}_1 - \gamma_{1_0})]$. In effect, $V[\sqrt{N}(\breve{\gamma}_1 - \gamma_{1_0})]$ is expected to be similar to the former (latter) asymptotic variance matrix when w is large (small), that is, when the auxiliary sample size, M, is large (small) relative to the main sample size, N.

References

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

Cameron, A.C., Trivedi, P.K., 1998. Regression Analysis of Count Data. Cambridge University Press, Cambridge.

Carroll, R.J., Pederson, S., 1993. On robustness in the logistic regression model. Journal of the Royal Statistical Society B 55, 693–706.

Carroll, R.J., Ruppert, D., Stefanski, L.A., 1995. Measurement Error in Nonlinear Models. Chapman & Hall, London.

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

Cosslett, S.R., 1981a. Efficient estimation of discrete-choice models. In: Manski, C.F., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometric Applications. The MIT Press, Cambridge, MA.

Cosslett, S.R., 1981b. Maximum likelihood estimator for choice-based samples. Econometrica 49, 1289–1316.

Cosslett, S.R., 1993. Endogenous stratification, semiparametric and non-parametric estimation. In: Maddala, G.S., Rao, C.R., Vinod, H.D. (Eds.), Handbook of Statistics 11. Elsevier, Amsterdam.

Cox, D.R., Hinkley, D.V., 1974. Theoretical Statistics. Chapman & Hall, London.

Cox, D.R., Snell, E.J., 1989. Analysis of Binary Data, 2nd Edition. Chapman & Hall, London.

Ekholm, A., Palmgren, J., 1987. Correction for misclassification using double sampled data. Journal of Official Statistics 3, 419–429.

Espeland, M.A., Odoroff, C.L., 1985. Log-linear models for doubly sampled categorical data fitted by the EM algorithm. Journal of the American Statistical Association 80, 663–670.

Gaudry, M., Dagenais, M., 1979. The dogit model. Transportation Research B13B, 105–111.

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

Hsieh, D.A., Manski, C.F., McFadden, D., 1985. Estimation of response probabilities from augmented retrospective observations. Journal of the American Statistical Association 80, 651–662.

Imbens, G.W., 1992. An efficient method of moments estimator for discrete choice models with choice-based sampling. Econometrica 60, 1187–1214.

Imbens, G.W., Hellerstein, J.K., 1999. Imposing moment restrictions from auxiliary data by weighting. Review of Economics and Statistics 81, 1–14.

Imbens, G.W., Lancaster, T., 1994. Combining micro and macro data in microeconometric models. Review of Economic Studies 61, 655–680.

Imbens, G.W., Lancaster, T., 1996. Efficient estimation and stratified sampling. Journal of Econometrics 74, 289–318.

Lambert, D., 1992. Zero-inflated Poisson regression with an application to defects in manufacturing. Technometrics 34, 1–14.

Lancaster, T., 1991. A paradox in choice-based sampling. Working paper, Department of Economics, Brown University.

Lancaster, T., Imbens, G., 1991. Choice-based sampling—inference and optimality. Discussion paper No. 91/304, Department of Economics, University of Bristol.

Manski, C.F., Lerman, S.R., 1977. The estimation of choice probabilities from choice based samples. Econometrica 45, 1977–1988.

Manski, C.F., McFadden, D., 1981. Alternative estimators and sample designs for discrete choice analysis. In: Manski, C.F., McFadden, D. (Eds.), Structural Analysis of Discrete Data with Econometric Applications. The MIT Press, Cambridge, MA.

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

Palmgren, J., Ekholm, A., 1987. Exponential family non-linear models for categorical data with errors of observation. Applied Stochastic Models and Data Analysis 3, 111–124.

Poirier, D.J., 1998. Revising beliefs in nonidentified models. Econometric Theory 14, 483–509.

Poterba, J.M., Summers, L.H., 1995. Unemployment benefits and labor market transitions: a multinomial logit model with errors in classification. Review of Economics and Statistics 77, 207–216.

Pregibon, D., 1982. Resistant fits for some commonly used logistic models with medical applications. Biometrics 38, 485–498.

Prentice, R.L., Pyke, R., 1979. Logistic disease incidence and case-control studies. Biometrika 66, 403–411.

Wang, C.Y., Carroll, R.J., 1993. On robust estimation in logistic case-control studies. Biometrika 80, 237–241.

Whittemore, A.S., Gong, G., 1991. Poisson regression with misclassified counts: application to cervical cancer mortality rates. Applied Statistics 40, 81–93.

Wooldridge, J.M., 1998. Asymptotic properties of weighted *m*-estimators for standard stratified samples. Working paper, Department of Economics, Michigan State University.

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