UNIVERSIDADE DE ÉVORA

DEPARTAMENTO DE ECONOMIA

DOCUMENTO DE TRABALHO Nº **2004/02**

March

Covariate Measurement Error in Endogenous Stratified Samples

Esmeralda A. Ramalho
Universidade de Évora, Departamento de Economia

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 the presence of covariate mea-surement error in endogenous stratifield samples. Using Chesher's (2000) methodology, we develop approximately consistent estimators for the parameters of the structural model, in the sense that their inconsistency is of smaller order than that of the conventional estimators which ignore the existence of covariate measurement error. The approximate bias corrected estimators are obtained by applying the generalized method of moments (GMM) to a modifeld version of the moment indicators suggested by Imbens and Lancaster (1996) for endogenous stratified samples. Only the specification of the conditional distribution of the response vari-able given the latent covariates and the classical additive measurement error model assumption are required, the availability of information on both the marginal probability of the strata in the population and the variance of the measurement error not being essential. A score test to detect the presence of covariate measurement error arises as a by-product of this approach.
Monte Carlo evidence is presented which suggests that, in endogenous stratified samples of moderate sizes, the modified GMM estimators perform well.

Palavras-chave/Keyword: endogenous stratified samples, covariate measurement error, generalized method of moments estimation, score tests

Classificação JEL/JEL Classification: C51, C52

1 Introduction

In many research settings, empirical researchers are often faced with the problem of making inferences from endogenous stratified (ES) samples. In this nonrandom sampling scheme, different subsets of the underlying population of interest are sampled with different frequencies, the selection being based on the variable of interest and, possibly, other variables. In contrast to random sampling (RS), where each unit of the population has the same probability of being sampled, with ES sampling individuals are not equally likely to be included in the sample, which is particularly convenient to deal with situations where a random sample of the target population would only include a few sampling units associated to some values of the variable of interest. For example, if the aim is modeling travel demand, it is very common for one or more modes of travel to have a very low market share, which would require the collection of a very large random sample to assure a reasonable number of individuals making each choice. Instead, to reduce data collection costs, often the sample is stratified on mode choice.

Despite the substantial development in inference methods for ES samples, see *inter alia* Manski and Lerman (1977), Manski and McFadden (1981), Cosslett (1981a,b), Imbens (1992), Imbens and Lancaster (1996) and Wooldridge (1999, 2001), little attention seems to have been paid to the possible presence of covariate measurement error (CME), an issue which affects many econometric data. To the best of our knowledge, the few existing approaches to this problem rely on very strong assumptions, requiring, for example, the specification of a precise form for the relation between the error-free variables and their error-prone measures and the availability of a validation sample, in which both measures are available for some sampling units. Furthermore, only a particular class of ES samples has been considered, namely the case of binary logistic choice-based (CB) samples, where inference procedures are specially simple since, after correcting for CME, the stratification can be ignored and estimation techniques used under RS may be employed; see *inter alia* Carroll, Gail and Lubin (1993), Wang and Carroll (1996), Roeder, Carroll and Lindsay (1996), Muller and Roeder (1997), and Wang, Wang and Carroll (1997).

The main aim of this paper is the development of appropriate estimation procedures to deal with the presence of CME in ES samples without making such strong assumptions. In fact, we merely require two of the assumptions made in the papers cited before: the specification of a structural model, characterized by the error-free conditional distribution of the variable of interest given the covariates, and the existence of CME of the classical additive kind, such that the measurement error and the true covariates are independent. Using Chesher's (1991) methodology, we obtain an approximate form of the contaminated distributions for a small error variance, which allows us to accommodate CME in the standard model for ES samples. The approximations employed

do not require the specification of the exact form of the error distribution, being only dependent on the variance of the measurement error and on the log-density derivatives of both the error-free distributions of the covariates and the response variable given the covariates. However, the specification of the distribution of the latent covariates can be avoided by nonparametrically estimating the derivatives of its log-density as in Chesher (1998, 2000, 2001). Furthermore, knowledge of the variance of the measurement error is not essential, although the availability of this information allows more efficient estimators to be obtained.

This flexible setting, in which only the formulation of the structural model is required, permits the development of approximately consistent estimators for the parameters of the structural model, in the sense that their inconsistency is of smaller order than that of the conventional estimators which ignore the existence of CME. The approximate bias corrected estimators are extension of Imbens and Lancaster's (1996) efficient GMM estimators for ES samples. Following Chesher's (2000) procedures to deal with CME in random samples, the moment indicators suggested by those authors are modified in such a way that, when evaluated at the observable error-prone variables, their expected value, taken under the approximation for the joint contaminated sampling distribution of the variable of interest, the error-prone covariates and the stratum indicator, is approximately zero. This base set of corrected moment indicators can be utilized when both the marginal probability of the strata in the population and the variance of the measurement error are unknown, in which case they are jointly estimated with the other parameters of interest. In case one or both of these quantities are known, the available information can be incorporated in the estimation procedure, allowing more efficient estimators to be obtained.

The most closely related work to ours, Santos Silva (1999), uses small parameter approximations to address the more general problem of unobservables, which may be due not only to CME but also to neglected heterogeneity, in endogenous samples, which include not only ES samples but also truncated and length biased samples, for example. It provides an extensive analysis of the effects of unobservables in maximum likelihood (ML) estimators based on the sampling conditional distribution of the response variable given the contaminated covariates, as well as a score test for the detection of the presence of unobservables. However, in our paper, the analysis of ES samples is not undertaken conditional on the covariates, which allow us to obtain more efficient estimators than the ones that would arise from the formulation developed by Santos Silva (1999).

Special attention is given to the CB binary logit model, where the practice of employing the same estimation techniques as in RS, only correcting for the CME, is widely spread. In this paper we show that a similar strategy, which only requires a simple modification of the RS procedure, can be successfully implemented using our method in both the cases where the variance of the

measurement error is known and unknown. For this reason, and also because the use of corrected GMM estimators in RS is very recent, we frequently address the RS case studied by Chesher (2000), where, in contrast to the first papers dealing with the correction of estimating equations for CME, due to Nakamura (1990) and Buonaccorsi (1996), the availability of replicate measurements for the observed error-prone covariates to obtain a prior estimate of the variance of the measurement error is not required.

The remainder of the paper is organized as follows. Section 2 formalizes the likelihood functions which take into account the presence of CME in ES samples. Section 3 develops GMM estimation procedures appropriate for this framework. Simplified versions of these procedures to deal with the particular cases of RS and CB logistic samples are presented in section 4. Section 5 reports some Monte Carlo evidence on the performance in practice of some of the proposed estimators. Finally, section 6 concludes. The appendix contains some cumbersome calculations which were suppressed from the main text.

2 Model specification

This section develops an extended version of the standard model for ES samples, based on small error variance approximations, which accommodates CME. The approximations derived show how the error-prone and the error-free models are related, which provides a very convenient framework to investigate the impact of CME in this sampling design.

2.1 Background

Consider a sample of $i = 1, ..., N$ individuals and let Y be the response variable of interest, continuous or discrete, and X a vector of k exogenous variables. Both Y and X are random variables defined on $\mathcal{Y} \times \mathcal{X}$ with population joint density function

$$f_{YX}(y,x) = f_{Y|X}(y|x,\theta)\, f_X(x), \tag{1}$$

where the conditional density function $f_{Y|X}(y|x,\theta)$ is known up to the parameter vector of interest θ and the marginal density function $f_X(x)$ is unknown.

ES 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 non-empty and possibly overlapping strata, which are subsets of $\mathcal{Y} \times \mathcal{X}$. Each stratum is designated as $\mathcal{C}_s = \mathcal{Y}_s \times \mathcal{X}$, for $s \in \mathcal{S}$, $\mathcal{S} = \{1, ..., J\}$, and $\mathcal{Y}_s$ is a subset of $\mathcal{Y}$. The probability of a randomly drawn observation lying in stratum $\mathcal{C}_s$ is

$$Q_s = \int_{\mathcal{X}} \int_{\mathcal{Y}_s} f_{Y|X}(y|x,\theta)\, f_X(x)\, dydx. \tag{2}$$

Assuming that the sample is drawn according to the multinomial sampling scheme, the agent who collects the sample defines the probability H_s of observing an unit from stratum s.[1] In this setting, the sampling density function of $Z = (Y, X, S)$ is given by

$$h_Z(z) = \frac{H_s}{Q_s} f_{Y|X}(y|x, \theta) f_X(x), \tag{3}$$

$(y, x) \in \mathcal{C}_s$, $s \in \mathcal{S}$. On the other hand, the marginal density function of X induced by this sampling scheme is given by

$$h_X(x) = \sum_{s \in \mathcal{S}} \int_{\mathcal{Y}_s} h_Z(z)\, dy = f_X(x)\, b_X(x), \tag{4}$$

where

$$b_X(x) = \sum_{s \in \mathcal{S}} \frac{H_s}{Q_s} \int_{\mathcal{Y}_s} f_{Y|X}(y|x, \theta)\, dy \tag{5}$$

reflects the bias induced by the nonrandom sampling design over the population density function of X, $f_X(x)$. Only when the sample is self-weighted, in which case H_s equals Q_s, does ES sampling become equivalent to RS because both the sampling densities (3) and (4) are reduced to, respectively, the population versions $f_{YX}(y,x)$ and $f_X(x)$.

Throughout this paper we give special attention to the case where the response variable takes values on a set of $(C+1)$ mutually exclusive alternatives, $Y \in \{0, 1, ..., C\}$, in which ES sampling takes the designation of CB sampling because the strata are determined by the alternative chosen. Actually, most papers dealing with ES sampling address the case of CB sampling; see, for example, Manski and Lerman (1977), Manski and McFadden (1981), Cosslett (1981a,b) and Imbens (1992).

2.2 Model incorporating covariate measurement error

Denote the observable covariates, possibly mismeasured, with the superscript $*$. Assume that, instead of the latent covariates X, we observe X^* according to

$$X^* = X + U, \tag{6}$$

where X and U are k-dimensional vectors of, respectively, error-free variates and unobservable measurement errors, which have an absolute continuous joint distribution. Assume also that U is defined on $\mathcal{U}$, the third absolute moments of U are finite, X and U are independently distributed, $E(U) = 0$, and $E(UU') = \Sigma = [\sigma_{jk}]$, where Σ is a positive semi-definite $k \times k$ matrix. If part of X is measured without error, the appropriate terms in Σ are set to zero. Furthermore, assume

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

that the density function of the unobservable measurement error U, $f_U(u)$, is unknown to the econometrician.

As only the covariates are contaminated and the strata are only defined in terms of the variable of interest, which is assumed to be error-free, the design of the strata is not affected by the mismeasurement. Thus, for each individual, one observes $Z^* = (Y, X^*, S)$, i.e. the error-free variable of interest, the mismeasured covariates and the error-free stratum indicator.

To proceed with likelihood-based inference, one needs to specify the likelihood function which describes the observed data Z^*. However, the simple evaluation of the joint sampling density of Z in (3) at the observable Z^*, $h_Z(z^*) = \frac{H_s}{Q_s} f_{Y|X}(y|x^*, \theta) f_X(x^*)$, does not provide a valid likelihood function because, in general, in presence of CME, the shape of the distributions of the observable variables is distorted when compared to that of its error-free version; see, for example, Chesher (1991). In fact, to model the contaminated data, we have to consider the contaminated joint density function of Z^*, which is denoted here as $h_{Z^*}(z^*)$. By writing the contaminated sampling joint density of the observable Z^* and the measurement error U,

$$h_{Z^*U}(z^*, u) = \frac{H_s}{Q_s} f_{Y|X}(y|x^* - u, \theta) f_X(x^* - u) f_U(u), \tag{7}$$

it becomes obvious that, unless $f_U(u)$ is specified, in which case the integration of (7) over $\mathcal{U}$ yields

$$h_{Z^*}(z^*) = \frac{H_s}{Q_s} \int_{\mathcal{U}} f_{Y|X}(y|x^* - u, \theta) f_X(x^* - u) f_U(u)\, du, \tag{8}$$

the obtention of $h_{Z^*}(z^*)$ is not straightforward.[2] However, by employing Chesher's (1991) method, we may obtain an asymptotic approximation for (8) that does not depend on $f_U(u)$. This approach, which uses an approximate likelihood function to describe the contaminated data, has already been used for endogenous sampling [Santos Silva (1999)], in the analysis of duration response measurement error [Dumangane (2000), Dumangane and Chesher (2001) and Chesher, Dumangane and Smith (2002)], in the study of the impact of CME in quantile regression [Chesher (2001)], and in the analysis of the effect of measurement error on measures of welfare inequality and poverty [Chesher and Schluter (2002)].

The approximation for (8) results from a second order Taylor series expansion of (7) around $\Sigma = 0$, followed by a marginalization of the resulting approximation with respect to U,

$$\begin{aligned} h_{Z^*}(z^*) &= \frac{H_s}{Q_s} f_{Y|X}(y|x^*, \theta) f_X(x^*) \left[1 + \sigma_{jk} m_{YX}^{jk}(y, x^*)\right] + o(\Sigma) \\ &= h_Z(z^*) \left[1 + \sigma_{jk} m_{YX}^{jk}(y, x^*)\right] + o(\Sigma), \end{aligned} \tag{9}$$

[2] Note that even if $f_U(u)$ were specified, often $h_{Z^*}(z^*)$ would have a very complicated form.

with

$$m_{YX}^{jk}(y,x^*) = 0.5\left[l_{Y|X}^{jk}(y|x^*,\theta) + l_{Y|X}^{j}(y|x^*,\theta)\, l_{Y|X}^{k}(y|x^*,\theta) + 2l_{Y|X}^{j}(y|x^*,\theta)\, l_{X}^{k}(x^*) + l_{X}^{jk}(x^*) + l_{X}^{j}(x^*)\, l_{X}^{k}(x^*)\right], \quad (10)$$

where superscripts denote derivatives with respect to the latent covariates which are mismeasured, $l_{Y|X}(y|x^*,\theta) = \ln f_{Y|X}(y|x^*,\theta)$, $l_X(x^*) = \ln f_X(x^*)$, $o(\Sigma)$ is such that $\lim_{\max(\sigma_{jj})\to 0} \frac{o(\Sigma)}{\max(\sigma_{jj})} = 0$, and the Einstein summation convention is employed with summation over repeated subscripts and superscripts.

The $O(\Sigma)$ approximation in (9), denoted as $h_{Z^*}^{a}(z^*)$, does not depend on $f_U(u)$. It is written in terms of the latent likelihood function $h_Z(z)$ evaluated at Z^* and a distortion term $\sigma_{jk} m_{YX}^{jk}(y,x^*)$ which is function of the variance of U and the derivatives of the error-free log-densities $f_{Y|X}(y|x,\theta)$ and $f_X(x)$ evaluated at the observable variables. This distortion is only eliminated when the covariates are correctly measured, in which case we observe Z and, as $\Sigma = 0$, (9) becomes identical to the error-free sampling joint density $h_Z(z)$ given in (3).

By integrating (9) over $\mathcal{Y}_s$ and summing over $\mathcal{S}$, we obtain the contaminated marginal density of the error-prone covariates in the sample,

$$\begin{aligned} h_{X^*}(x^*) &= \sum_{s\in\mathcal{S}} \int_{\mathcal{Y}_s} \frac{H_s}{Q_s} f_{Y|X}(y|x^*,\theta)\, f_X(x^*)\, dy + \sigma_{jk} \sum_{s\in\mathcal{S}} \int_{\mathcal{Y}_s} \frac{H_s}{Q_s} f_{Y|X}(y|x^*,\theta)\, f_X(x^*)\, m_{YX}^{jk}(y,x^*)\, dy \\ &\quad + o(\Sigma) \\ &= f_X(x^*)\, b_X(x^*) + 0.5\sigma_{jk} \sum_{s\in\mathcal{S}} \frac{H_s}{Q_s} \int_{\mathcal{Y}_s} \left\{ f_{Y|X}^{jk}(y|x^*,\theta)\, f_X(x^*) + 2 f_{Y|X}^{j}(y|x^*,\theta)\, f_X^{k}(x^*) \right. \\ &\quad \left. + f_{Y|X}(y|x^*,\theta)\, f_X^{jk}(x^*) \right\} dy + o(\Sigma), \end{aligned} \quad (11)$$

which now presents two sources of distortions relative to the underlying marginal density of X in the population, $f_X(x)$. One source of bias, $b_X(x)$ given in equation (5), is only due to the sampling design and is also present when all the variables are properly measured; see the latent sampling density $h_X(x)$ in (4). The other source of deformation, given by the second term in (11), reflects the combined effects of the ES sampling design and the CME.

As widely discussed [see, for example, Chesher (1991, 1998) and Dumangane (2000)], additive approximations of the type of (9) may not produce a proper density function, in the sense that they may not be positive and integrate to one. Thus, they may not be used directly for ML estimation. However, this problem can be circumvented by re-expressing (9) as an augmented density in the class defined by Chesher and Smith (1997)

$$h_{Z^*}^{aug}(z^*) = h_Z(z^*)\, \Psi\left[\sigma_{jk} m_{YX}^{jk}(y,x^*)\right] q(H_s,Q_s,\theta,\Sigma)^{-1} + o(\Sigma), \quad (12)$$

where $\Psi(w)$ is a positive valued function with finite derivatives of all orders, $\nabla_w \Psi(0) \neq 0$, and $q(H_s,Q_s,\theta,\Sigma) = \sum_{s\in\mathcal{S}} \int_{\mathcal{Y}_s} \int_{\mathcal{X}^*} h_Z(z^*)\, \Psi\left[\sigma_{jk} m_{YX}^{jk}(y,x^*)\right] dx^* dy$, which is assumed to exist.

In the next section, rather than maximizing the log-likelihood function obtained from the approximation in (12), we correct the moment conditions employed by Imbens and Lancaster (1996), which, in the absence of measurement error, result from the maximization of the log-likelihood function based on (3). This correction consists of subtracting their expectation taken with respect to $h^a_{Z^*}(z^*)$ in (9) from them. However, as we show later, (12) will be very useful to specify the quantities required for the efficient version of the score test sensitive to CME derived in subsection 3.3.

When the design is self-weighting, as $H_s = Q_s$, $h_{Z^*}(z^*)$ and $h^{aug}_{Z^*}(z^*)$ in, respectively, (9) and (12) give the approximations for the contaminated joint density function $f_{YX^*}(y, x^*)$, the error-prone version of $f_{YX}(y, x)$ in (1). As the weighting nature of the sampling scheme is eliminated, the analysis can then be conducted conditional on the contaminated covariates, as is usual in RS. Thus, in this setting, the features of the contaminated dataset may be simply described by the error-prone conditional density of the variable of interest given the contaminated covariates,

$$f_{Y|X^*}(y|x^*) = f_{Y|X}(y|x^*, \theta)\left[1 + \sigma_{jk} m_{YX}^{RS^{jk}}(y, x^*)\right] + o(\Sigma) \tag{13}$$

or

$$f^{aug}_{Y|X^*}(y|x^*) = f_{Y|X}(y|x^*, \theta)\,\Psi\left[\sigma_{jk} m_{YX}^{RS^{jk}}(y, x^*)\right] q(\theta, \Sigma)^{-1} + o(\Sigma), \tag{14}$$

with

$$m_{YX}^{RS^{jk}}(y, x^*) = 0.5\left[l^{jk}_{Y|X}(y|x^*, \theta) + l^{j}_{Y|X}(y|x^*, \theta)\, l^{k}_{Y|X}(y|x^*, \theta) + 2 l^{j}_{Y|X}(y|x^*)\, l^{k}_{X}(x^*)\right] \tag{15}$$

and $q(\theta, \Sigma) = \int_{\mathcal{Y}} f_{Y|X}(y|x^*, \theta)\,\Psi\left[\sigma_{jk} m_{YX}^{RS^{jk}}(y, x^*)\right] dy$. Both (13) and (14) are embedded in, respectively, (9) and (12). The self-weighting eliminates the ratio $\frac{H_s}{Q_s}$, while conditioning on the covariates suppresses $f_X(x^*)$ as well as the terms $l^{jk}_X(x^*)$ and $l^{j}_X(x^*)\, l^{k}_X(x^*)$ contained in $m^{jk}_{YX}(y, x^*)$; for a detailed discussion of inference based on small error variance approximations under RS see, for example, Chesher (1991, 1998)

3 Generalized method of moments estimation

In the previous section we showed that the presence of CME in ES samples distorts the joint sampling distribution $h_Z(z)$ in (3) as a consequence of the bias induced in both $f_{Y|X}(y|x, \theta)$ and $f_X(x)$. Hence, it is expected that, in general, due to the failure of distributional assumptions, all the conventional likelihood-based estimators for ES samples, for example, those proposed by Manski and Lerman (1977), Manski and McFadden (1981), Cosslett (1981a,b), Imbens (1992), and Imbens and Lancaster (1996), are inconsistent for the parameters of interest. Moreover, as CME affects the shape of both the conditional expected value and the conditional median of the

variable of interest given the latent regressors [see, for example, Chesher (1991, 1998)], both the weighted least squares and the weighted least absolute deviation estimator, which are encompassed by Wooldridge's (1999, 2001) weighted M-estimators, are also inconsistent.

This section addresses the problem of CME in the efficient GMM estimation setting proposed by Imbens and Lancaster (1996) for ES samples. The motivation for modifying these estimators instead of the other cited alternatives is twofold. First, in contrast to all the other estimators, with the exception of those proposed by Wooldridge (1999, 2001), they are appropriate for any ES sample, not only for those where the variable of interest is discrete. Second, they are asymptotically efficient, a property only shared by Cosslett's (1981a,b) estimators.

The estimators proposed by Imbens and Lancaster (1996) for ES samples without measurement error maximize the log-likelihood based on (3) assuming that the covariates follow a discrete distribution which is jointly estimated with the parameters of interest. After some transformations, the dependence on the discrete distribution is removed from the score functions, which are then used as moment indicators in GMM estimation. The resulting set of moment indicators is

$$g_{H_t}(z) = H_t - I_{(s=t)} \tag{16}$$

$$g_\theta(z) = \nabla_\theta \ln f_{Y|X}(y|x,\theta) - \nabla_\theta \ln b_X(x) \tag{17}$$

$$g_{Q_t}(z) = Q_t - \frac{\int_{\mathcal{Y}_t} f_{Y|X}(y|x,\theta)\, dy}{b_X(x)} \tag{18}$$

$$g_{Q_J}(z) = 1 - b_X(x)^{-1}, \tag{19}$$

where $I_{(s=t)}$ takes the value 1 for $s = t$ and 0 for $s \neq t$, $t = 1, ..., J-1$, ∇_θ denotes derivative with respect to θ, and the vector of parameters of interest is $\gamma = (H, \theta, Q)$, with $H = (H_1, ..., H_{J-1})$ and $Q = (Q_1, ..., Q_J)$.[3] The objective function to be minimized is

$$\Upsilon_N(\gamma) = g_N(\gamma)' W_N g_N(\gamma), \tag{20}$$

where $g_N(\gamma) = \frac{1}{N}\sum_{i=1}^N g_\gamma(z_i)$ is the sample counterpart of the moment conditions $E_{h_Z}[g_\gamma(z_i)] = 0$, the expectation being taken with respect to the sampling joint density (3), which is henceforth denoted by $h_Z(z,\gamma)$ to emphasise the dependence on γ, the moment indicators $g_\gamma(z_i)$ are given in (16)-(19), and W_N is a positive semi-definite weighting matrix. Imbens and Lancaster (1996) prove that the resulting optimal estimator, $\hat{\gamma}$, obtained from the use of the weighting matrix $W_N = \Omega_N^{-1}$ in (20), where Ω_N is a consistent estimator of $\Omega = E_{h_Z}\left[g_\gamma(z)\, g_\gamma(z)'\right]$, converges almost surely to the true value γ^0 and is asymptotically normal,

$$\sqrt{N}\left(\hat{\gamma} - \gamma^0\right) \xrightarrow{d} N\left[0, \left(G'\Omega^{-1}G\right)^{-1}\right], \tag{21}$$

[3] Note that $H_J = 1 - \sum_{s=1}^{J-1} H_s$. With non-overlapping strata $Q_J = 1 - \sum_{s=1}^{J-1} Q_s$, in which case Q has dimension $(J-1)$ and the moment indicator (19) can be suppressed from the system (16)-(19).

where $G = E_{h_Z}\left[\nabla_\gamma g_\gamma(z)'\right]$.

The set of moment indicators (16)-(19) is valid both when the marginal probability of each stratum in the population, contained in vector Q, is known or unknown. In the former case, these probabilities are substituted in the moment indicators and the vector Q is suppressed from the vector of parameters of interest γ, which becomes $\gamma = (H, \theta)$, generating a case of overidentifying moment conditions. In the latter, the parameters to be estimated are $\gamma = (H, \theta, Q)$, which generates a just-identified problem.

In this section, following Dumangane and Chesher (2001), we first derive a Kiefer and Skoog-type (1984) measure for the inconsistency of Imbens and Lancaster's (1996) GMM estimators when the presence of CME is ignored. Then, subsection 3.2 extends these GMM estimators to deal with contaminated data by correcting the original moment conditions so that their expectation taken under the contaminated distribution of Z^* is approximately zero. Subsection 3.3 suggests a score test for the detection of CME. Finally, subsection 3.4 describes a nonparametric procedure for the estimation of the derivatives of the log-density of the latent covariates required for GMM estimation and for the score test.

3.1 Inconsistency of Imbens and Lancaster's (1996) generalized method of moments estimators

In presence of CME, Imbens and Lancaster's (1996) GMM estimators $\hat{\gamma}$, which merely replace X by X^* in moment indicators (16)-(19), do not converge to the true value γ^0. Below we use small parameter approximations to obtain an expression for the bias suffered by these estimators when the presence of CME is not acknowledged. To the best of our knowledge, Kiefer and Skoog (1984), in the ML context, were the first to use this methodology to measure the effects of model misspecification; see also Chesher, Lancaster and Irish (1983) and Levine (1985), as well as Stefanski's (1985) proposal for M-estimators. Recently, Dumangane and Chesher (2001) extended Kiefer and Skoog's (1984) approach to obtain the distortions caused by response measurement error in the GMM framework, which we now adapt for the ES setting.

Let $\gamma\left(\phi^0\right)$ denote the probability limit to which Imbens and Lancaster's (1996) estimators $\hat{\gamma}$ converge, where $\phi^0 = \left(\gamma^0, \sigma^0\right)$ is the true value of $\phi = (\gamma, \sigma)$, with σ defined as a vector of dimension D containing all the different nonzero elements of matrix Σ.[4] The vector of parameters ϕ is present in $h^a_{Z^*}(z^*)$ of (9), which is henceforth denoted as $h^a_{Z^*}(z^*, \phi)$ to stress this dependence. Naturally, $\gamma\left(\gamma^0, 0\right) = \gamma^0$. Moreover, since the proportion of the strata in the sample is only defined

[4]Note that $\sigma_{jk} = \sigma_{kj}$ for $k \neq j$. Thus, $D = \frac{\left(k^*+1\right)k^*}{2}$, $0 \leq k^* \leq k$, for k^* defined as the number of mismeasured covariates.

in terms of the error-free variable of interest Y, the vector H contained in γ is still consistently estimated with CME. In fact, the moment indicators (16) do not depend on the mismeasured variable, not being, thus, affected by measurement error.

Using small parameter approximations, $\gamma\left(\phi^0\right)$ may be written as

$$\gamma\left(\phi^0\right) = \gamma^0 + \sigma_{jk} \left.\frac{d\gamma\left(\phi^0\right)}{d\sigma_{jk}}\right|_{\gamma\left(\phi^0\right)=\gamma^0} + o\left(\Sigma\right), \tag{22}$$

where the second term is the inconsistency measure suggested by Kiefer and Skoog (1984). To obtain this measure, we need to consider the implicit equations for $\gamma\left(\phi^0\right)$,

$$E_{h_{Z^*}}\left[g_{\gamma\left(\phi^0\right)}\left(z^*\right)\right] = 0, \tag{23}$$

where $g_{\gamma\left(\phi^0\right)}\left(z^*\right)$ are the moment indicators given in (16)-(19) evaluated at Z^* and $E_{h_{Z^*}}\left[.\right]$ denotes expectation taken with respect to $h_{Z^*}\left(z^*\right)$. Calculating this expectation using the approximation $h^a_{Z^*}\left(z^*,\phi\right)$ in (9), we obtain

$$\sum_{s\in\mathcal{S}} \int_{\mathcal{Y}_s} \int_{\mathcal{X}^*} g_{\gamma\left(\phi^0\right)}\left(z^*\right) h_Z\left(z^*,\gamma\right)\left[1+\sigma_{jk} m^{jk}_{YX}\left(y,x^*\right)\right] dx^* dy = 0. \tag{24}$$

Totally differentiating (24) with respect to $\gamma\left(\phi^0\right)$ and σ_{jk} and evaluating the resulting expression at $\gamma\left(\phi^0\right) = \gamma\left(\gamma^0,0\right) = \gamma^0$, yields

$$\sum_{s\in\mathcal{S}} \int_{\mathcal{Y}_s} \int_{\mathcal{X}^*} \left[\nabla_{\gamma^0} g_{\gamma^0}\left(z^*\right) h_Z\left(z^*,\gamma^0\right) d\gamma\left(\phi^0\right) + g_{\gamma^0}\left(z^*\right) m^{jk}_{YX}\left(y,x^*\right) h_Z\left(z^*,\gamma^0\right) d\sigma_{jk}\right] dx^* dy = 0, \tag{25}$$

which may be re-expressed as

$$G d\gamma\left(\phi^0\right) = -E_{h_Z}\left[g_{\gamma^0}\left(z^*\right) m^{jk}_{YX}\left(y,x^*\right)\right] d\sigma_{jk}, \tag{26}$$

where G, defined below (21), is evaluated at Z^* and γ^0. Pre-multiplying both sides of (26) by $G'W$, where W is the probability limit of the weighting matrix W_N employed in the GMM objective function (20), and solving for $\left.\frac{d\gamma\left(\phi^0\right)}{d\sigma_{jk}}\right|_{\gamma\left(\phi^0\right)=\gamma^0}$, it follows that

$$\left.\frac{d\gamma\left(\phi^0\right)}{d\sigma_{jk}}\right|_{\gamma\left(\phi^0\right)=\gamma^0} = -\left(G'WG\right)^{-1} G'W E_{h_Z}\left[g_{\gamma^0}\left(z^*\right) m^{jk}_{YX}\left(y,x^*\right)\right]. \tag{27}$$

Thus, the Kiefer and Skoog's (1984) type inconsistency measure of Imbens and Lancaster's (1996) estimators, given by (27) multiplied by σ_{jk}, is function of the variance of the measurement error and of expectations taken under the latent joint density $h_z\left(z,\gamma\right)$ in (3) of the error-free quantities $g_{\gamma^0}\left(z^*\right) m^{jk}_{YX}\left(y,x^*\right)$ and $\nabla_\gamma g_{\gamma^0}\left(z^*\right)$ evaluated at Z^*. Note that in case the vector $\gamma = (H,\theta,Q)$ is estimated, the just-identified nature of the GMM problem allows us to reduce

(27) to $-G^{-1}E_{h_Z}\left[g_{\gamma^0}(z^*)\, m^{jk}_{YX}(y,x^*)\right]$, while when Q is known, optimal GMM estimation of $\gamma=(H,\theta)$ is performed and W is replaced by $\Omega^{-1}=E_{h_Z}\left[g_{\gamma^0}(z^*)\; g_{\gamma^0}(z^*)'\right]^{-1}$ in (27). In this setup, the probability limits for $\hat{\gamma}=\left(\hat{H},\hat{\theta},\hat{Q}\right)$ and $\hat{\gamma}=\left(\hat{H},\hat{\theta}\right)$ are given by, respectively,

$$\gamma\left(\phi^0\right)=\gamma^0-\sigma_{jk}G^{-1}E_{h_Z}\left[g_{\gamma^0}(z^*)\, m^{jk}_{YX}(y,x^*)\right]+o(\Sigma). \tag{28}$$

and

$$\gamma\left(\phi^0\right)=\gamma^0-\sigma_{jk}\left(G'\Omega^{-1}G\right)^{-1}G'\Omega^{-1}E_{h_Z}\left[g_{\gamma^0}(z^*)\, m^{jk}_{YX}(y,x^*)\right]+o(\Sigma). \tag{29}$$

The term $\sigma_{jk}E_{h_Z}\left[g_{\gamma}(z^*)\, m^{jk}_{YX}(y,x^*)\right]$ present in both the inconsistency measures in (28) and (29), from now on denoted as $b_{\phi}(z^*)$, may be seen as an approximation for the expectation of the original moment indicators $g_{\gamma}(z)$ evaluated at Z^*, $E_{h_{Z^*}}[g_{\gamma}(z^*)]$. In effect, using approximation $h^a_{Z^*}(z^*,\phi)$ in (9) to calculate this expectation, we find

$$\begin{aligned}
E_{h_{Z^*}}[g_{\gamma}(z^*)] &= \sum_{s\in\mathcal{S}}\int_{\mathcal{Y}_s}\int_{\mathcal{X}^*} g_{\gamma}(z^*)\, h(z^*,\gamma)\left[1+\sigma_{jk}m^{jk}_{YX}(y,x^*)\right]dx^*dy+o(\Sigma)\\
&= E_{h_Z}[g_{\gamma}(z^*)]+\sigma_{jk}E_{h_Z}\left[g_{\gamma}(z^*)\, m^{jk}_{YX}(y,x^*)\right]+o(\Sigma)\\
&= \sigma_{jk}E_{h_Z}\left[g_{\gamma}(z^*)\, m^{jk}_{YX}(y,x^*)\right]+o(\Sigma)\\
&= b_{\phi}(z^*)+o(\Sigma).
\end{aligned} \tag{30}$$

Thus, $b_{\phi}(z^*)$ may be interpreted as the approximate bias in the original moment indicators incurred by the presence of measurement error.[5] The approximate biases in moment indicators (16)-(19), derived in appendix 7.1, are given by, respectively,

$$b_{H_t}(z^*)=0 \tag{31}$$

$$\begin{aligned}
b_{\theta}(z^*) &= \sigma_{jk}E_{f_X}\left\{\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\int_{\mathcal{Y}_s} m^{RS^{jk}}_{YX}(y,x^*)\left[\nabla_{\theta}f_{Y|X}(y|x^*,\theta)-\frac{f_{Y|X}(y|x^*,\theta)}{b_X(x^*)}\right.\right.\\
&\left.\left.\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\nabla_{\theta}f_{Y|X}(y|x^*,\theta)\right]dy\right\}
\end{aligned} \tag{32}$$

$$b_{Q_t}(z^*)=-\sigma_{jk}E_{f_X}\left[\frac{\int_{\mathcal{Y}_t}f_{Y|X}(y|x^*,\theta)\,dy\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\int_{\mathcal{Y}_s}f_{Y|X}(y|x^*,\theta)\, m^{jk}_{YX}(y,x^*)\,dy}{b_X(x^*)}\right] \tag{33}$$

$$b_{Q_J}(z^*)=-\sigma_{jk}E_{f_X}\left[\frac{\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\int_{\mathcal{Y}_s}f_{Y|X}(y|x^*,\theta)\, m^{jk}_{YX}(y,x^*)\,dy}{b_X(x^*)}\right], \tag{34}$$

where $E_{f_X}[.]$ denotes expectation taken with respect to $f_X(x)$. The distortion in $g_{H_t}(z^*)$ is zero because these moment indicators are not a function of the mismeasured variable. As far as

[5] Recall that, previously, we had already defined another bias function, $b_X(x)$ [see equation (5)], which has a very different nature from that in (30) because it concerns only the distortion imposed by endogenous sampling over the marginal density of X, $f_X(x)$.

the other moment indicators are concerned, the bias $b_\phi(z^*)$ is eliminated only when there is no mismeasurement, in which case $\sigma = 0$. In these conditions, as $E_{h_{Z^*}}[g_\gamma(z^*)] = 0$, the distortion terms in the probability limits (28) and (29) are suppressed and $\gamma(\phi^0) = \gamma^0$.

In this subsection we demonstrated that, in presence of CME, the bias in the original moment indicators, $b_\phi(z^*)$, causes the inconsistency of conventional GMM estimators. In the subsequent subsections it will be shown that the approximate bias functions in (31)-(34) are a crucial element not only in the modification of moment indicators (16)-(19) to handle CME, but also in the implementation of an efficient version of a score test sensitive to the presence of this form of measurement error.

3.2 Correction of Imbens and Lancaster's (1996) moment indicators

A direct adaptation of Imbens and Lancaster's (1996) method to handle CME would require the calculation of the set of first order conditions resulting from the maximization of a log-likelihood function based on the $O(\Sigma)$ approximation in (12) with respect to the vector of parameters (H, θ, π, σ), where π is a vector containing the probability mass parameters associated with $f_X(x^*)$ at the given set of support points. Then, the resulting set of score functions would have to be transformed in order to remove their dependence on π and, together with two extra sets of moments associated with the estimation of Q and the elements of σ, they would be used as moment indicators for GMM estimation.

Alternatively, as we do here, we may employ Chesher's (2000) method, which avoids dealing with the complicated function (12); see also Dumangane (2000) and Dumangane and Chesher (2001), who follow the same approach to handle response measurement error in duration models, correcting the score functions of models commonly employed in that area. The idea is very simple. As shown in (30), the expectation of the original moment indicators evaluated at Z^* taken with respect to the approximate contaminated density $h^a_{z^*}(z^*, \phi)$ is not zero but $b_\phi(z^*) + o(\Sigma)$. Hence, if we subtract $b_\phi(z^*)$ from the original moment indicators, the resulting modified moment indicators,

$$g^*_\phi(z^*) = g_\gamma(z^*) - b_\phi(z^*), \tag{35}$$

have expectation $E_{h_{Z^*}}\left[g^*_\phi(z^*)\right] = o(\Sigma)$. Although this expectation is not zero with CME, (35) may be used to obtain approximately consistent estimators.

To implement this approach, we need to calculate both the expectations and the quantities $l^k_X(x^*)$ and $l^{jk}_X(x^*)$ present in $b_\phi(z^*)$, which involve the marginal distribution of the covariates. In order to avoid the specification of $f_X(x)$, one may estimate the expectations by simple averages or, following Cosslett (1993), take averages with the weight $\frac{Q_{s_i}}{H_{s_i}}$. Moreover, $l^k_X(x^*)$ and $l^{jk}_X(x^*)$ may be estimated nonparametrically as described in subsection 3.4. On the other hand, the modified

moment indicators (35) depend also on the variance of the measurement error, which often is unknown in practical situations. In order to make possible its estimation simultaneously with the parameters of interest γ, we introduce a further set of moment indicators, denoted $g^*_{\sigma_{jk}}(z^*)$, which corresponds to the set of score functions for σ obtained from the log-likelihood function based on $h^a_{Z^*}(z^*,\phi)$ in (9). Thus, in presence of CME, we suggest the utilization of the base set of modified moment indicators given by

$$g^*_{H_t}(z^*) = H_t - I_{(s=t)} \tag{36}$$

$$\begin{aligned} g^*_\theta(z^*) &= \nabla_\theta \ln f_{Y|X}(y|x^*,\theta) - \frac{1}{b_X(x^*)}\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\int_{\mathcal{Y}_s}\left\{\nabla_\theta f_{Y|X}(y|x^*,\theta) + \sigma_{jk} m^{RS^{jk}}_{YX}(y,x^*)\right. \\ &\quad \left.\left[\nabla_\theta f_{Y|X}(y|x^*,\theta)\, b_X(x^*) - \nabla_\theta b_X(x^*)\, f_{Y|X}(y|x^*,\theta)\right]\right\} dy + o(\Sigma) \end{aligned} \tag{37}$$

$$g^*_{Q_t}(z^*) = Q_t - \frac{1}{b_X(x^*)}\int_{\mathcal{Y}_t} f_{Y|X}(y|x^*,\theta)\, dy\left[1 - \sigma_{jk}\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\int_{\mathcal{Y}_s} f_{Y|X}(y|x^*,\theta)\, m^{jk}_{YX}(y,x^*)\, dy\right] + o(\Sigma) \tag{38}$$

$$g^*_{Q_J}(z^*) = 1 - \frac{1}{b_X(x^*)}\left[1 - \sigma_{jk}\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\int_{\mathcal{Y}_s} f_{Y|X}(y|x^*,\theta)\, m^{jk}_{YX}(y,x^*)\, dy\right] + o(\Sigma) \tag{39}$$

$$g^*_{\sigma_{jk}}(z^*) = \frac{m^{jk}_{YX}(y,x^*)}{1+\sigma_{jk} m^{jk}_{YX}(y,x^*)} + o(\Sigma), \tag{40}$$

which is composed of (16), modified versions of (17)-(19) calculated in appendix 7.2, and the additional moment indicators (40) concerned with the estimation of the variance of U. Naturally, with correct measurement, as $\Sigma = 0$, moment indicators (36)-(39) coincide with their original counterparts (16)-(19) proposed by Imbens and Lancaster (1996).

The system (36)-(40) can be solved using standard GMM procedures. The modified GMM (MGMM) estimators $\hat{\phi}$ are obtained by minimizing an objective function analogous to that in (20), with $g_N(\gamma)$ replaced by $g^*_N(\phi) = \frac{1}{N}\sum_{i=1}^N g^*_\phi(z^*_i)$, which is the sample counterpart of $E_{h_{Z^*}}\left[g^*_\phi(z^*)\right]$, the moment indicators $g^*_\phi(z^*)$ being given in (36)-(40). As $E_{h_{Z^*}}\left[g^*_\phi(z^*)\right] = o(\Sigma)$, only when CME is absent the MGMM estimators will be consistent for the parameters of interest. With CME, the probability limit of the MGMM estimators $\hat{\phi}$ is ϕ^* and not the true value ϕ^0. In fact, following Dumangane and Chesher (2001), it is straightforward to show that the probability limit of the MGMM estimators is[6]

$$p\lim\hat{\phi} = \phi^* = \phi^0 + o(\Sigma). \tag{41}$$

[6] These authors extend for the GMM framework a result derived by Chesher and Santos Silva (2002), who obtained the order of inconsistency of a quasi-ML estimator for the parameters of interest of a logit model for taste variation. The demonstration for our case of CME is similar to that of Dumangane and Chesher (2001) for response measurement error because the operations we propose in the next subsection to circumvent the specification of the derivatives of the log-density of X, $l^k_X(x^*)$ and $l^{jk}_X(x^*)$, do not change the order of the approximation $h^a_{Z^*}(z^*,\phi)$ in (9); see also appendix 7.3.

Obviously, $\hat{\phi}$ suffer from some inconsistency, which depends on the magnitude of the variance of the measurement error Σ. However, the asymptotic bias of these approximate estimators is less than that of the GMM estimators which ignore the presence of CME; compare (41) with the probability limits $\gamma\left(\phi^0\right)$ of (28) and (29) obtained previously.

Analogously to ES sampling with no measurement error, if Q or σ, or both quantities, are known, their values are substituted in (36)-(40), and the vector ϕ of estimated parameters is reduced, respectively, to (H, θ, σ), (H, θ, Q), or (H, θ). The resulting overidentifying system imposes restrictions concerning the known quantities, allowing more efficient estimators to be obtained relative to the case where all the parameters would have to be estimated. When neither Q or σ are known, a just-identified GMM estimator for $\phi = (H, \theta, Q, \sigma)$ needs to be calculated.

3.3 A score test to detect covariate measurement error

This subsection outlines a score test sensitive to CME for the GMM estimation framework proposed previously.[7] The idea is testing if the D elements of vector σ are zero. The null hypothesis is $H_0 : \sigma = 0$, for which the score test statistic [see Newey and McFadden (1994, Theorem 9.2.)] is given by

$$T = N g_N^{*\prime} \Omega_N^{*-1} G_N^* V_N^* G_N^{*\prime} \Omega_N^{*-1} g_N^*, \tag{42}$$

where $g_N^* \equiv g_N^*(\phi)$ and Ω_N^*, G_N^* and V_N^* are consistent estimators of, respectively, $\Omega^* = E_{h_{Z^*}}\left[g_\phi^*(z^*) g_\phi^*(z^*)'\right]$, $G^* = E_{h_{Z^*}}\left[\nabla_\phi g_\phi^*(z^*)'\right]$ and $V^* = \left(G^{*\prime}\Omega^{*-1}G^*\right)^{-1}$, all of them evaluated at consistent estimators of the parameters of the restricted model, $\hat{\phi} = (\hat{\gamma}, 0)$. Under the null hypothesis, T converges in distribution to a chi-square random variable with D degrees of freedom. Note that, under H_0, the moment indicators g_N^* in (36)-(40) are reduced to, respectively, (16)-(19) and

$$g_{\sigma_{jk}}^*(z^*)\Big|_{\sigma_{jk}=0} = m_{YX}^{jk}(y, x^*), \tag{43}$$

which may also be obtained from the maximization of the log-likelihood based on (12), the augmented density defined by Chesher and Smith (1997). Hence, the implementation of the efficient version of the test is very simple since, under H_0, the covariance between (16)-(19) and (43) is given by the approximate bias functions (31)-(34) with σ_{jk} suppressed.

Following Dumangane and Chesher's (2001) approach, we could use the moment indicators $g_{\sigma_{jk}}^*(z^*)\Big|_{\sigma_{jk}=0}$ in (43) to obtain alternative estimating functions for σ. In fact, $g_{\sigma_{jk}}^*(z^*)\Big|_{\sigma_{jk}=0}$ may be modified to produce moment indicators for the estimation of σ_{jk} according to the same

[7] This type of test to detect measurement error was proposed in the ML framework by Chesher (1990) and applied, for example, in Santos Silva (1999) and Chesher, Dumangane and Smith (2002) in the context of, respectively, endogenous samples and duration models.

principle utilised to modify the original moment indicators by Imbens and Lancaster (1996), which is based on equation (35). The bias function for this case is $b_{\sigma_{jk}}(z^*) = \sigma_{jk} E_{h_Z}\left[m_{YX}^{jk}(y,x^*)^2\right] = \sigma_{jk} E_{f_X}\left[\sum_{s\in\mathcal{S}} \frac{H_s}{Q_s} \int_{\mathcal{Y}_s} m_{YX}^{jk}(y,x^*)^2 f_{Y|X}(y|x,\theta)\,dy\right]$ and the resulting moment indicators are $g^*_{\sigma_{jk}}(z^*) = m_{YX}^{jk}(y,x^*) - \sigma_{jk}\sum_{s\in\mathcal{S}} \frac{H_s}{Q_s} \int_{\mathcal{Y}_s} m_{YX}^{jk}(y,x^*)^2 f_{Y|X}(y|x,\theta)\,dy + o(\Sigma)$. Note that if we had considered taking a simple average of $m_{YX}^{jk}(y,x^*)^2$ to estimate $E_{h_Z}\left[m_{YX}^{jk}(y,x^*)^2\right]$, instead of averaging $\sum_{s\in\mathcal{S}} \frac{H_s}{Q_s} \int_{\mathcal{Y}_s} m_{YX}^{jk}(y,x^*)^2 f_{Y|X}(y|x,\theta)\,dy$ to estimate $E_{f_X}\left[\sum_{s\in\mathcal{S}} \frac{H_s}{Q_s} \int_{\mathcal{Y}_s} m_{YX}^{jk}\left((y,x^*)^2 f_{Y|X}(y|x,\theta)\,dy\right]\right.$, the moment indicators for σ would be given by $g^*_{\sigma_{jk}}(z^*) = m_{YX}^{jk}(y,x^*) - \sigma_{jk} m_{YX}^{jk}(y,x^*)^2 + o(\Sigma)$, which contains an $O(\Sigma)$ approximation for the formulation we propose in (40). So, both approaches may be considered approximately equivalent.

Both the score test and the estimators suggested require the derivatives of the log-density of the error-free covariates evaluated at the observed covariates, $l_X^k(x^*)$ and $l_X^{jk}(x^*)$. As the error-free marginal distribution of the covariates $f_X(x)$ is unknown to the researcher, the next subsection suggests a nonparametric procedure to estimate these quantities.

3.4 Nonparametric estimation of the features of $f_X(x)$

Any regression model incorporating CME based on asymptotic approximations for a small error variance is a function of the derivatives of the log-density of the error-free covariates. Hence, unless the econometrician is prepared to specify $f_X(x)$, all estimators and specification tests require the estimation of these derivatives in a first stage; see Chesher (1998, 2000, 2001). Following these papers, we adopt Barron and Sheu's (1991) nonparametric method based on sequences of exponential families to estimate $l_X^k(x^*)$ and $l_X^{jk}(x^*)$. However, our problem is more complicated since, while under RS the features of $f_X(x^*)$ can be estimated using error-prone data described by $f_{X^*}(x^*) = f_X(x^*) + 0.5\sigma_{jk} f_X^{jk}(x^*) + o(\Sigma)$, under ES sampling the available data conforms with the more complex sampling density $h_{X^*}(x^*)$ of (11), which prevents direct estimation of the derivatives of interest as in RS.

Our approach consists of writing the aimed features, $l_X^k(x^*)$ and $l_X^{jk}(x^*)$, in terms of estimable or known quantities which may be substituted in either the moment indicators (37)-(40) or the test statistics (42), in such a way that the order of the approximation error in $h_{Z^*}(z^*)$ of (9) is not increased. The resulting expressions for $l_X^k(x^*)$ and $l_X^{jk}(x^*)$, derived in appendix 7.3, are

$$l_X^k(x^*) = l_{h_{X^*}}^k(x^*) - lb_X^k(x^*) \tag{44}$$

and

$$l_X^{jk}(x^*) = l_{h_{X^*}}^{jk}(x^*) - lb_X^{jk}(x^*), \tag{45}$$

where $l_{h_{X^*}}^k(x^*) = [\ln h_{X^*}(x^*)]^k$, $l_{h_{X^*}}^{jk}(x^*) = [\ln h_{X^*}(x^*)]^{jk}$, $lb_X^k(x) = [\ln b_X(x)]^k$, and $lb_X^{jk}(x) = [\ln b_X(x)]^{jk}$. Both (44) and (45) are functions of the conditional density function $f_{Y|X}(y|x,\theta)$,

which is assumed known, the strata marginal probabilities in the sample, H_s, and in the population, Q_s, which may be either known or estimated, and the derivatives $l^k_{h_{X^*}}(x^*)$ and $l^{jk}_{h_{X^*}}(x^*)$, which may by estimated nonparametrically by Barron and Sheu's (1991) method, as described next.

When X is a scalar random variable, we may write the unknown density $h_{X^*}(x^*)$ as

$$r_{X^*}(x^*) = h^0_{X^*}(x^*) \exp\left\{\sum_{l=1}^{M} \beta_l \omega_l(x^*) - \ln \int_0^1 h^0_{X^*}(x^*) \exp\left[\sum_{l=1}^{M} \beta_l \omega_l(x^*)\right] dx^*\right\}, \tag{46}$$

where $h^0_{X^*}(x^*)$ is a reference probability density with support on $[0,1]$, M defines the length of the exponential series, $\omega_l(x^*)$, $l = 1...\ M$, are bounded and linearly independent functions spanning a linear space of functions, $\beta = (\beta_1,...,\ \beta_M)$ is a vector of unknown parameters, and $M/N \to 0$. Omitting irrelevant terms, the log-likelihood based on (46) may be rewritten as

$$\ln r_{X^*}(x^*) = \sum_{l=1}^{M} \beta_l \left[\omega_l(x^*) - \bar{\omega}_l\right] - \ln \int_0^1 \exp\left\{\sum_{l=1}^{M} \beta_l \left[\omega_l(x^*) - \bar{\omega}_l\right]\right\} dx^*, \tag{47}$$

where $\bar{\omega}_l$ is the sample mean of the lth Legendre polynomial. Note that maximizing (47) is identical to minimizing

$$R(\beta) = \int_0^1 \exp\left\{\sum_{l=1}^{M} \beta_l \left[\omega_l(x^*) - \bar{\omega}_l\right]\right\} dx^*.$$

The calculation of the integral in $R(\beta)$ may be avoided by using the approximation

$$R(\beta)^a = \frac{1}{T+1} \sum_{t=1}^{T+1} \exp\left\{\sum_{l=1}^{M} \beta_l \left[\omega_l\left(\frac{t-1}{T}\right) - \bar{\omega}_l\right]\right\}. \tag{48}$$

Using the uniform density on $[0,1]$ as reference density and Legendre polynomials in $\omega_l(x^*)$, the log-density derivatives are simply

$$l^k_{h_{X^*}}(x^*) = \sum_{l=1}^{M} \beta_l \omega^k_l(x^*) \tag{49}$$

and

$$l^{jk}_{h_{X^*}}(x^*) = \sum_{l=1}^{M} \beta_l \omega^{jk}_l(x^*). \tag{50}$$

Thus, our procedure consists of estimating nonparametrically $l^k_{h_{X^*}}(x^*)$ and $l^{jk}_{h_{X^*}}(x^*)$ in a first stage, which are then substituted into (44) and (45), respectively. Next, we may replace $l^k_X(x^*)$ and $l^{jk}_X(x^*)$ by, respectively, (44) and (45) in $m^{jk}_{YX}(y, x^*)$ contained in both the moment indicators (37)-(40) and the test statistics in (42). In our Monte Carlo experiments, similarly to Chesher (1998), we set $T = 100$ and $M = 6$.

4 Particular cases

4.1 Random sampling

In the end of subsection 2.2, we showed that the RS model specification can be seen as a particular case of that suggested for ES samples by merely setting $\frac{H_s}{Q_s} = 1$ and conditioning the analysis on the covariates. Thus, it is natural that a simplified version of the GMM procedures described in section 3 is appropriate to deal with RS. Namely, the vector of parameters of interest is reduced to $\phi = (\theta, \sigma)$, all expectations are now taken over $f_{Y|X^*}(y|x^*)$ instead of $h_{Z^*}(z^*, \phi)$, and $m_{YX}^{jk}(y, x^*)$ of (10) can be replaced by $m_{YX}^{RS^{jk}}(y, x^*)$ of (15) in all formulas.

In this RS setup, GMM estimation of θ ignoring the presence of CME uses the moment indicator $g_\theta(z^*)$ of (17) with the second term suppressed, $g_\theta^{RS}(z^*) = \nabla_\theta \ln f_{Y|X}(y|x^*, \theta)$, which corresponds to standard ML estimation based on the likelihood $f_{Y|X}(y|x, \theta)$ evaluated at (Y, X^*). The probability limit $\gamma(\phi^0) = \theta(\theta^0, \sigma^0)$ to which this naive RS ML estimator for θ converges can be written from (28) as

$$\gamma(\phi^0) = \theta^0 - \sigma_{jk} G^{-1} E_{f_{Y|X}}\left[\nabla_\theta \ln f_{Y|X}(y|x^*, \theta)\, m_{YX}^{RS^{jk}}(y, x^*)\,|X\right] + o(\Sigma). \tag{51}$$

where $G = E_{f_{Y|X}}[-\nabla_{\theta\theta'} \ln f_{Y|X}(y|\ x^*, \theta)\,|X]$. Hence, the approximate bias function for θ, corresponding to $b_\theta(z^*)$ of (32), is now reduced to:

$$b_\theta^{RS}(z^*) = \sigma_{jk} \int_{\mathcal{Y}} \nabla_\theta f_{Y|X}(y|x^*, \theta)\, m_{YX}^{RS^{jk}}(y, x^*)\, dy. \tag{52}$$

When CME is acknowledged, the estimation of the vector of parameters of interest $\phi = (\theta, \sigma)$ (or $\phi = \theta$ if σ is known) is based upon a reduced version of (37) and (40):

$$g_\theta^{RS^*}(z^*) = \nabla_\theta \ln f(y|x, \theta) - \sigma_{jk} \int_{\mathcal{Y}} \nabla_\theta f_{Y|X}(y|x^*, \theta)\, m_{YX}^{RS^{jk}}(y, x^*)\, dy + o(\Sigma) \tag{53}$$

and

$$g_{\sigma_{jk}}^{RS^*}(z^*) = \frac{m_{YX}^{RS^{jk}}(y, x^*)}{1 + \sigma_{jk} m_{YX}^{RS^{jk}}(y, x^*)} + o(\Sigma). \tag{54}$$

The MGMM estimators based on (53)-(54) coincide with those proposed by Chesher (2000), being an alternative to the estimator suggested by Buonaccorsi (1996), who, assuming the existence of repeated observations of the error-prone covariates to estimate the variance of the measurement error, proposes a modification for estimating functions of the class $g_\theta^B(y, x) = \left[y - E_{f_{Y|X}}(y|x, \theta)\right] d(x, \theta)$, where $d(x, \theta)$ is a weighting function. The method suggested by Chesher (2000), implemented in this paper, allows the modification of any estimation function, including those of Buonaccorsi (1996), and presents the additional advantage of not requiring the availability of repeated measurements on X^*, since the variance of U can be estimated from the moment indicators

(54); for a comparison of the performance in practice of both Buonaccorsi's (1996) and Chesher's (2000) estimators, see the Monte Carlo simulation study of subsection 5.2. It is also relevant to note that in the next subsection we show that a GMM estimator for the slope parameters of CB logit models with CME may be based on the system (53)-(54) instead of the more extended version (36)-(40) proposed previously for ES samples.

The adaptation of the score test proposed in subsection 3.3 to detect CME in random datasets is straightforward by using the simplifications suggested before. Obviously, now the moment indicators of interest for the implementation of the test are (53) and (54) evaluated at $\sigma = 0$, respectively, $g_\theta^{RS^*}(z^*)\big|_{\sigma_{jk}=0} = g_\theta^{RS}(z^*) = \nabla_\theta \ln f_{Y|X}(y|x^*,\theta)$ and

$$g^*_{\sigma_{jk}}(z^*)\Big|_{\sigma_{jk}=0} = m_{YX}^{RS^{jk}}(y,x^*)\,. \tag{55}$$

The covariance between these two sets of moment indicators required by the efficient version of the test is given by the approximate bias function in (52) with σ_{jk} suppressed.

Finally, notice that, also in this setup, the implementation of both the MGMM estimators and the score test requires the nonparametric estimation of $l_X^k(x^*)$ contained in the term $m_{YX}^{RS^{jk}}(y,x^*)$ of (53)-(55). In this case Barron and Sheu's (1991) method described at the end of subsection 3.4 provides directly a nonparametric estimator for $l_{X^*}^k(x^*)$, which is then replaced in $l_X^k(x^*)$.[8]

4.2 Choice-based binary logistic samples

In CB samples, when the variable of interest conditional on the error-free covariates is described by a binary logit model and a validation sample is available, Carroll, Gail and Lubin (1993), Wang and Carroll (1996), Roeder, Carroll and Lindsay (1996), Muller and Roeder (1997), and Wang, Wang and Carroll (1997), based on the results of Prentice and Pyke (1979) for CB samples with correct measurement, propose a range of ML-based estimators where the sampling scheme is ignored and estimation proceeds as in RS, only accounting for the existence of CME.

This section investigates the estimation of this class of models in our framework, in which the regression model is written in terms of small parameter asymptotic approximations and the econometrician does not possess a validation sample. In absence of measurement error, RS estimation of logit models with CB samples is justified by the fact that the conditional probability of Y given X is coincident in the population and in the sample, apart from a distortion in the intercept term. Thus, the idea here is examining whether with CME, for a sufficiently small Σ, an analogous property holds, i.e. whether both the approximations of the contaminated version

[8] Note that this method now considers $f_{X^*}(x^*)$ as the unknown density of interest, from which $l_{X^*}^k(x^*)$ is estimated according to (49).

of the conditional probability of Y given X^* coincide in the population and the sample. As in Chesher (1991), the former approximation may be expressed by

$$P_1^* = P\left[1 + \sigma_{jk}\Lambda^{jk}\left(1 - P\right)\right] + o\left(\Sigma\right) \tag{56}$$

and

$$P_0^* = \left(1 - P\right)\left(1 - \sigma_{jk}\Lambda^{jk}P\right) + o\left(\Sigma\right), \tag{57}$$

where $P_1^* = \Pr_{Y|X^*}\left(1|x^*,\theta,\sigma\right)$, $P_0^* = \Pr_{Y|X^*}\left(0|x^*,\theta,\sigma\right)$, $P = \Pr_{Y|X}\left(1|x^*,\theta\right) = \left(1 + e^{-x^{*\prime}\theta}\right)^{-1}$, with x^* containing a constant term, and $\Lambda^{jk} = 0.5\theta^j\theta^k\left[1 - 2P + \frac{2}{\theta^j}l_X^k\left(x^*\right)\right]$. Denoting the probability of observing $Y = 1$ in the sample and in the population as, respectively, H and Q, the approximate conditional probability of observing response $Y = 1$ in the sample given X^*, $P_1^{CB^*} = \Pr_{Y|X^*}^{CB}\left(1|x^*,\theta,\Sigma\right)$, is

$$\begin{aligned} P_1^{CB^*} &= \frac{\frac{H}{Q}P_1^*}{\frac{H}{Q}P_1^* + \frac{1-H}{1-Q}P_0^*} + o\left(\Sigma\right) \\ &= \left(1 + \frac{Q}{H}\frac{1-H}{1-Q}\frac{1-P_1^*}{P_1^*}\right)^{-1} + o\left(\Sigma\right). \end{aligned} \tag{58}$$

When Σ is sufficiently small, to order $o\left(\Sigma\right)$, in (56) $P_1^* \simeq P$ and in (58) $P_1^{CB^*} \simeq \left(1 + \frac{Q}{H}\frac{1-H}{1-Q}\frac{1-P}{P}\right)^{-1} = \left(1 + \frac{Q}{H}\frac{1-H}{1-Q}e^{-x'\theta}\right)^{-1}$. Thus, it is clear that, apart from the shift of $-\ln\left(\frac{Q}{H}\frac{1-H}{1-Q}\right)$ in the constant term, $P_1^{CB^*}$ approximately coincides with P, describing a logit model, which allow us to use the RS moment indicators (53) and (54) given in subsection 4.1 to estimate the slope parameters of interest as well as the intercept terms displaced by $-\ln\left(\frac{Q}{H}\frac{1-H}{1-Q}\right)$. Note, however, that (53) and (54) contain $l_X^k\left(x^*\right)$, which, as in CB samples the sampling density of the covariates, $h_{X^*}\left(x^*\right)$, deviates from $f_{X^*}\left(x^*\right)$, has to be substituted for (44), instead of being directly estimated as in RS; see subsection 3.4. Moreover, as $lb_X\left(x^*\right)^k$ in (44) is a function of Q_y, this method may only be used when this marginal probability is known.

In this setting, although the RS ML estimator corrected for CME may not be applied, the general estimation procedures for ES samples are substantially simplified, since the use of the extended system of moment indicators in (36)-(40) is circumvented. Relative to previous papers on CB logistic samples, though our estimator is slightly more complicated, since it involves the nonparametric estimation of $l_X^k\left(x^*\right)$, it offers the advantage of not requiring a validation sample, relying only on the assumption that the logit formulation adopted for the structural model is correct. Furthermore, note that most of the estimators cited at the beginning of this subsection require not only the specification of the structural model, but also the formulation of a conditional distribution or a conditional expected value describing the relation between the observable and the error-free covariates.

5 Performance in practice

In this section we undertake three Monte Carlo simulation studies to investigate the small sample behaviour of some of the MGMM estimators described previously. Subsection 5.1 considers binary logit and probit models with CB sampling, while subsection 5.2 studies binary logit models based on RS.

5.1 Binary models with choice-based sampling

The main aim of this subsection is to assess the performance in practice of some of our MGMM estimators for ES samples with CME. First, we consider binary logit models, where the simplified estimation procedures described in subsection 4.2 may be employed. Then, we simulate binary probit models, where the complete methodology proposed in subsection 3.2 must be utilized. In both cases the binary CB samples generated involve two strata, stratum 1 and stratum 0, with individuals choosing, respectively, alternative $Y = 1$ and $Y = 0$. The probability of observing an unit from the former (latter) stratum in the sample and in the population is, respectively, H $(1 - H)$ and Q $(1 - Q)$. Q was set equal to 0.9 and, for each experiment, two sampling designs were considered, characterized by $H = \{0.5, 0.7\}$. The sampling scheme where $H = 0.5$, usually termed the equal shares design because the proportion of each strata in the sample is identical, is claimed to be close to an optimal design, in the sense that minimizes the asymptotic variance of the estimators; see Cosslett (1981a), Lancaster and Imbens (1991) and Imbens (1992). All experiments, implemented in S-Plus, are based on 1000 replications for a sample size of 500.

5.1.1 Logit model

In this first set of experiments the variable of interest Y, conditional on X, is distributed as logit with $\Pr_{Y|X}(1|x,\theta) = \left(1 + e^{-\theta_0 - x\theta_1}\right)^{-1}$ and the marginal choice probability Q is assumed known. The error-free covariate was generated with mean 3 and variance 4, either as a mixture of normal distributions, where the variate is $N(2, 1.2915)$ with probability 0.7 and $N(5.333, 1.2915)$ with probability 0.3, or Student $\sqrt{\frac{4}{3}}t(3)$. In order to produce $Q = 0.9$, θ_0 was set equal to 0, while θ_1 was fixed to 1.3 and 1.0 with, respectively, the former and latter distribution assumed for X. The error-prone observed covariate was generated from $X^* = X + U$, where U is distributed independently of both X and Y. In all experiments, the variance of U, denoted as σ, was set equal to 0.25. In designs a and c, U follows a $N(0, 0.25)$ distribution while in b and d, U is $\sqrt{\frac{0.25}{3}}t(3)$. Table 1 summarizes the experimental designs just described.

Table 1 about here

Three different estimators were calculated: the naive GMM estimator (which in this case is a ML estimator), denoted NE, and the MGMM estimators for known and unknown σ, respectively termed MEa and MEb. For the two MGMM estimators, the derivatives of the log-density of the covariates evaluated at the observed values of X^*, denoted as l_X^1, were nonparametrically estimated in a first step by following the procedures described in subsection 3.4. Both the MGMM estimators are based on the following individual moment indicators, which were written from equations (53) and (54),

$$g_{\theta_0}^*(z^*) = \frac{p}{P(1-P)}\left[y - P - 0.5\sigma\theta_1^2 p\left(\frac{p^1}{p} + \frac{2}{\theta_1}l_X^1\right)\right] + o(\Sigma) \tag{59}$$

$$g_{\theta_1}^*(z^*) = \frac{xp}{P(1-P)}\left[y - P - 0.5\sigma\theta_1^2 p\left(\frac{p^1}{p} + \frac{2}{\theta_1}l_X^1\right)\right] + o(\Sigma) \tag{60}$$

$$g_{\sigma}^*(z^*) = \frac{\frac{\theta_1^2 p(y-P)}{P(1-P)}\left(\frac{p^1}{p} + \frac{2}{\theta_1}l_X^1\right)}{2 + \sigma\frac{\theta_1^2 p(y-P)}{P(1-P)}\left(\frac{p^1}{p} + \frac{2}{\theta_1}l_X^1\right)} + o(\Sigma), \tag{61}$$

where $P = \Pr_{Y|X}(1|x^*, \theta)$, $p = \nabla_{\theta_0} P$ and $p^1 = \nabla_{\theta_0'} p$. In the MEa case, σ was replaced by its known value, while for the MEb case it was estimated simultaneously with the other parameters of interest. With regard to the NE, the moment indicators employed are (59) and (60) with $\sigma = 0$.

Table 2 reports the mean and the median bias in percentage terms along with the standard deviation across the replications for the estimates of the slope coefficient θ_1. Figure 1 shows the estimated sampling distributions of NE, MEa, and MEb. In all cases, the naive estimators display considerable mean and median downward biases, always greater than 9.4%. These two statistics are substantially less for our two modified estimators. In fact, the smallest reduction in the mean and median biases of NE occurs in experiments c for $H = 0.5$ where, even so, these statistics are reduced to, respectively, 46% and 50% in MEa and 50% and 65% in MEb. These conclusions are also illustrated in Figure 1, where the sampling distributions of both MEa and MEb are always more centrally located around the true value of θ_1 than that of NE, which lies substantially beneath this value in all cases. As expected, MEa shows, in general, a better performance in terms of mean and median biases than MEb, since it incorporates information on σ.

Table 2 about here

Figure 1 about here

As for the standard deviations of both the MGMM estimators, as usual in estimators accounting for measurement error [see, for example, the simulation experiments in Chesher (1998) and Hausman, Abravaya and Scott-Morton (1998)], they appear inflated when compared with those of the inconsistent naive estimators. This occurs because the former estimators reflect the additional variability in the data induced by CME. Moreover, once again, the favorable influence of including

additional information on σ is evident, the sampling variability of MEa being always smaller than that of MEb. Note also that in all the estimators considered, the standard deviations are lower for $H = 0.5$ than for $H = 0.7$, which certainly is a result of the close to optimality characteristic of the former sampling scheme.

5.1.2 Probit model

In this framework, we assume that $Y|X$ is described by a probit model with no intercept, such that $\Pr_{Y|X}(1\,|x\,,\theta) = \Phi(x\theta)$.[9] The generation of the contaminated covariate follows the design previously coded as a (see Table 1) and, to obtain $Q = 0.9$, θ was set equal to 0.75. Furthermore, we assume that σ is unknown to the researcher, the situation which exhibited the worst results in the previous Monte Carlo experiments because, in the two MGMM estimators, σ needs to be estimated simultaneously with the other parameters of interest.

As the endogeneity of the sample has to be taken into account in probit models, we calculated the estimators for ES samples for both the cases where there is information on Q and when this parameter has to be estimated. When Q is known (unknown), we considered GMM estimators ignoring the presence of CME and correcting for this problem, denoted, respectively, as NEa (NEb) and MEa (MEb). Thus, in these experiments, Q is the source of additional information. The derivatives of the log-density of X evaluated at X^*, denoted l_X^1 and l_X^2, were estimated as described in subsection 3.4 and the base set of individual moment indicators corresponding to (36)-(40) is

$$g_H^*(z^*) = H - y \tag{62}$$

$$g_\theta^*(z^*) = \frac{xp}{P(1-P)}\left\{y - \frac{H}{Q}\frac{1}{b_X(x)}\left[P + 0.5\sigma\theta^2 p\left(\frac{p^1}{p} + \frac{2}{\theta}l_X^1\right)\frac{1-H}{1-Q}\right]\right\} + o(\Sigma) \tag{63}$$

$$g_Q^*(z^*) = Q - \frac{P}{b_X(x)}\left\{1 - 0.5\sigma\left[\theta^2 p\left(\frac{H}{Q} - \frac{1-H}{1-Q}\right)\left(\frac{p^1}{p} + \frac{2}{\theta}l^1\right) + \left[l_X^2 + \left(l_X^1\right)^2\right]b(x)\right]\right\} + o(\Sigma) \tag{64}$$

$$g_\sigma^*(z^*) = \frac{\frac{\theta^2 p(y-P)}{P(1-P)}\left(\frac{p^1}{p} + \frac{2}{\theta}l_X^1\right) + l_X^2 + \left(l_X^1\right)^2}{2 + \sigma\left[\frac{\theta^2 p(y-P)}{P(1-P)}\left(\frac{p^1}{p} + \frac{2}{\theta}l_X^1\right) + l_X^2 + \left(l_X^1\right)^2\right]} + o(\Sigma)\,, \tag{65}$$

where $P = \Pr_{Y|X}(1|x^*,\theta)$, $p = \nabla_{x\theta}P$ and $p^1 = \nabla_{x\theta}p$. Note that in CB sampling designs where each choice defines one stratum $\sum_{s\in\mathcal{S}} Q_s = 1$. Thus, the moment indicator (39) was suppressed. Moreover, to obtain NEa and NEb, only the moment indicators (62)-(64) need to be considered with $\sigma = 0$, which thus coincide with those of Imbens' (1992) simulation study concerning GMM

[9] We did not use an intercept term in these experiments in order to reduce the computational time. Obviously, in the experiments concerning logit models, discussed in the previous subsection, an intercept was considered because only in that case estimation could be undertaken as if the sampling were random.

estimators for CB samples. For both NEa and MEa, estimation was performed with Q replaced by its known value in (62)-(65).

Table 3 and Figure 2 contain, respectively, the mean and the median bias in percentage terms and the standard deviation of the estimates of θ across replications, and the estimated sampling distributions of NEa, NEb, MEa, and MEb. They suggest very different comments for the cases where Q is known and unknown. In the former case, NEa presents relatively small mean and median biases, which, even so, were substantially reduced by our MEa at a cost of a small increment in the dispersion. In the latter situation, NEb is seriously downward biased. The MEb eliminate part of this bias, which in the worst case (see MEb for $H = 0.70$), is reduced to approximately 65% in the mean and 34.1% in the median. Despite this improvement, note that, for $H = 0.5$ the mean and median biases of MEb are approximately three times superior than that of the naive estimator which combines information on Q, NEa. Moreover, MEb also exhibits very large standard deviations across the replications. Thus, some care must be taken when applying them in samples of the size considered here.

Table 3 about here

Figure 2 about here

In these experiments, the benefits of including additional information concerning the marginal choice probabilities Q are apparent. On the one hand, the naive estimators become clearly more robust to the presence of CME. On the other hand, MEa presents a very promising performance, which is specially encouraging if we take into account that σ is estimated, a situation, which, in general, leads to a degradation of the Monte Carlo simulation results.

5.2 Binary logit models with random sampling

The goal of this subsection is twofold. Firstly, we intend to briefly examine the performance in practice of the MGMM estimators in binary logit models using RS, in which case they coincide with Chesher's (2000) estimators. To the best of our knowledge, in this framework, no simulation study concerning discrete choice models had been undertaken. Secondly, we compare the performance of these MGMM with that of Buonaccorsi's (1996) estimator, which is simpler to implement, since it does not require the employment of nonparametric estimation in a first stage to obtain the derivatives of the log-density of the latent covariates, but always needs prior information on σ.

Again, the logit model for Y given X is characterized by $\Pr_{Y|X}(1|x,\theta) = \left(1 + e^{-\theta_0 - x\theta_1}\right)^{-1}$ and the observed covariates are defined as $X^* = X + U$, where U is distributed independently of both X and Y. X and U were generated in three different ways, summarized in Table 4. Firstly

we considered the design previously coded as a. Then, we adopted one of Buonaccorsi's (1996) designs, which assumes that X is a $N(0, 0.1)$ variate and U is $N(0, 0.1/3)$.[10] In the third type of experiments, relative to the previous design, we just altered the variance of X to 1 and that of U to 1/3. In experiment a the parameters of interest were fixed as $\theta_0 = 0$ and $\theta_1 = 1.3$, as before, while in the others, as in Buonaccorsi (1996), $\theta_0 = -1.4$ and $\theta_1 = 1.4$.

Table 4 about here

In this setting, CME is the only sampling problem that has to be taken into account in inference. In addition to the three estimators for RS previously considered in subsection 5.1.1 for the logit model, NE, MEa, and MEb, we also employ Buonaccorsi's (1996) estimator for CME, denoted as BE. The estimation procedures for obtaining NE are the same as described in that subsection. Concerning the two MGMM estimators, MEa and MEb, now the nonparametric estimates of l_X^1 are used directly in the moment indicators (59)-(61); see the directions for estimation with RS in subsection 3.4. Finally, BE was implemented by using the moment indicators (59) and (60) with l_X^1 replaced by, respectively, 1 and $\frac{1}{X^*}$. Experiments were conducted in S-Plus, involving samples sizes of 300 and 1000 repetitions for each setting.

The statistics exhibited in Table 5, as well as the graphs in Figure 3, concerning estimates of θ_1, show that the behaviour of the naive estimators is again unacceptable in terms of bias. As for the estimators accounting for measurement error, BE show a smaller variability across replications, probably due to not requiring the nonparametric estimation of l_X^1. Comparing the results for the mean and median bias of the two estimators where σ is assumed known, BE and MEa, the former does better only in experiment b, showing larger mean and median biases in the other two cases, where the variance of the error-free covariate is larger. Moreover, BE has the disadvantage that in scheme a all statistics are slightly worse than those of NE, which creates serious doubts about its usefulness. Thus, in this example, the global behaviour of MEa is better, even with the larger σ of design c, which is encouraging, as we employ approximations based on small σ.

Table 5 about here

Figure 3 about here

On the other hand, the results for MEb, which involve the estimation of σ, are also promising and very similar to those of MEa in designs a and b. However, its performance decays substantially in design c, due to the increase in σ. Even so, relative to NE, mean and median biases are reduced by more than 50% in all cases.

[10] Note that, in contrast to that paper, we did not consider repeated measurements of X^*.

6 Conclusion

In this paper we have proposed a general framework to deal with the presence of CME in ES samples. First, a regression model to describe the observed data was specified by using Chesher's (1991) asymptotic approximations for a small error variance. Then, we considered Imbens and Lancaster's (1996) efficient GMM estimators originally proposed for ES samples properly measured. After identifying the sources of bias of these estimators in the presence of CME, we suggested a modification to them in order to obtain approximately consistent estimators and outlined a score test sensitive to CME.

We found that the inconsistency of Imbens and Lancaster's (1996) estimators when the covariates' contamination is not acknowledged, obtained by a Kiefer and Skoog-type (1984) measure adapted for the GMM framework, is a function of the approximate expectation of the moment indicators proposed by the former authors, taken with respect to the contaminated sampling joint distribution of the variable of interest, the error-prone covariates, and the stratum indicator. This approximation may be interpreted as the approximate bias induced by CME in the original moment indicators. Using Chesher's (2000) method, by subtracting this approximate bias function from the original moment indicators, we obtained modified moment indicators for which the expectation taken under the approximate distribution of the observed data is approximately zero. Thus, we suggest the use of the traditional GMM techniques based on them to obtain approximately consistent estimators for the parameters of interest. A component of the approximate bias function is also employed in the efficient version of the score test to detect the presence of contamination.

All the major contributions of this paper require the calculation of the referred to approximate bias function. Though these calculations are often complicated, as they involve derivatives of the structural model and the nonparametric estimation of features of the error-free distribution of the covariates, once these functions are obtained, the score test for the presence of measurement error is easily implemented and, when the null hypothesis of absence of contamination is rejected, the employment of the MGMM estimators proposed here is straightforward.

The flexibility of this approach is especially visible in two levels. On the one hand, relative to the model specified for ES samples, it merely employs one additional mild assumption on the measurement error model, which requires that the measurement error is independently distributed from the latent covariates. On the other hand, neither the population marginal probability of each stratum nor the variance of the measurement error need to be known, although when this kind of information is available, it may be easily incorporated in the estimation procedure, allowing more efficient estimators to be obtained.

Monte Carlo evidence was presented which suggests that, in ES sampling designs of moderate sizes, the MGMM estimators perform well. In these experiments, the bias reduction is substantial, in particular in situations where available information on either the strata marginal probabilities or the variance of the measurement error is incorporated in the estimation procedure.

7 Appendix

7.1 Calculation of the approximate bias functions

The approximate bias function (32) is calculated from the general formula $\sigma_{jk}E_{h_Z}\left[g_\gamma(z^*)\, m_{YX}^{jk}(y,x^*)\right]$ as

$$\begin{aligned}
b_\theta(z^*) &= \sigma_{jk}E_{h_Z}\left[g_\theta(z^*)\, m_{YX}^{jk}(y,x^*)\right] \\
&= \sigma_{jk}E_{f_X}\left[\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\int_{\mathcal{Y}_s} g_\theta(z^*)\, m_{YX}^{jk}(y,x^*)\, f_{Y|X}(y|x^*,\theta)\, dy\right] \\
&= \sigma_{jk}E_{f_X}\left\{\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\int_{\mathcal{Y}_s} g_\theta(z^*)\, f_{Y|X}(y|x^*,\theta)\left[m_{YX}^{RS^{jk}}(y,x^*) + l_X^{jk}(x^*)\right.\right. \\
&\quad \left.\left. + l_X^j(x^*)\, l_X^k(x^*)\right] dy\right\} \\
&= \sigma_{jk}E_{f_X}\left[\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\int_{\mathcal{Y}_s} g_\theta(z^*)\, f_{Y|X}(y|x^*,\theta)\, m_{YX}^{RS^{jk}}(y,x^*)\, dy\right] \\
&= \sigma_{jk}E_{f_X}\left\{\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\int_{\mathcal{Y}_s} m_{YX}^{RS^{jk}}(y,x^*)\left[\nabla_\theta f_{Y|X}(y|x^*,\theta) - \frac{f_{Y|X}(y|x^*,\theta)}{b_X(x^*)}\right.\right. \\
&\quad \left.\left.\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\nabla_\theta f_{Y|X}(y|x^*,\theta)\right] dy\right\},
\end{aligned}$$

where the suppression of the terms $l_X^{jk}(x^*)$ and $l_X^j(x^*)\, l_X^k(x^*)$ in $m_{YX}^{jk}(y,x^*)$ exploits the fact that $\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\int_{\mathcal{Y}_s} g_\gamma(z^*)\, f_{Y|X}(y|x^*,\theta)\, dy = 0$.

The calculation of both (33) and (34) is similar, using the fact that $E_{h_Z}\left[m_{YX}^{jk}(y,x^*)\right] = 0$:

$$\begin{aligned}
b_{Q_t}(z^*) &= \sigma_{jk}E_{h_Z}\left[g_{Q_t}(z^*)\, m_{YX}^{jk}(y,x^*)\right] \\
&= \sigma_{jk}E_{f_X}\left[\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\int_{\mathcal{Y}_s} g_{Q_t}(z^*)\, f_{Y|X}(y|x^*,\theta)\, m_{YX}^{jk}(y,x^*)\, dy\right] \\
&= \sigma_{jk}Q_t E_{h_Z}\left[m_{YX}^{jk}(y,x^*)\right] - \sigma_{jk}E_{f_X}\left[\frac{1}{b_X(x^*)}\int_{\mathcal{Y}_t} f_{Y|X}(y|x^*,\theta)\, dy\right. \\
&\quad \left.\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\int_{\mathcal{Y}_s} f_{Y|X}(y|x^*,\theta)\, m_{YX}^{jk}(y,x^*)\, dy\right] \\
&= -\sigma_{jk}E_{f_X}\left[\frac{1}{b_X(x^*)}\int_{\mathcal{Y}_t} f_{Y|X}(y|x^*,\theta)\, dy\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\int_{\mathcal{Y}_s} f_{Y|X}(y|x^*,\theta)\, m_{YX}^{jk}(y,x^*)\, dy\right]
\end{aligned}$$

$$
\begin{aligned}
b_{Q_J}(z^*) &= \sigma_{jk} E_{h_Z}\left[g_{Q_J}(z^*)\, m_{YX}^{jk}(y,x^*)\right] \\
&= \sigma_{jk} E_{f_X}\left[\sum_{s\in\mathcal{S}} \frac{H_s}{Q_s}\int_{\mathcal{Y}_s} g_{Q_J}(z^*)\, f_{Y|X}(y|x^*,\theta)\, m_{YX}^{jk}(y,x^*)\, dy\right] \\
&= \sigma_{jk} E_{h_Z}\left[m_{YX}^{jk}(y,x^*)\right] - \sigma_{jk} E_{f_X}\left[\frac{1}{b_X(x^*)}\sum_{s\in\mathcal{S}} \frac{H_s}{Q_s}\int_{\mathcal{Y}_s} f_{Y|X}(y|x^*,\theta)\, m_{YX}^{jk}(y,x^*)\, dy\right] \\
&= -\sigma_{jk} E_{f_X}\left[\frac{1}{b_X(x^*)}\sum_{s\in\mathcal{S}} \frac{H_s}{Q_s}\int_{\mathcal{Y}_s} f_{Y|X}(y|x^*,\theta)\, m_{YX}^{jk}(y,x^*)\, dy\right].
\end{aligned}
$$

7.2 Calculation of the modified moment indicators

The corrected moment indicators (36)-(39) are obtained from (35) and employ the bias functions in (32)-(34) with $E_{f_X}[.]$ estimated by simple averages:

$$
\begin{aligned}
g_\theta(z^*) &= \nabla_\theta \ln f_{Y|X}(y|x^*,\theta) - \frac{\nabla_\theta b_X(x^*)}{b_X(x^*)} - \sigma_{jk}\left\{\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\int_{\mathcal{Y}_s} m_{RS}^{jk}(y,x^*)\right. \\
&\quad \left.\left[\nabla_\theta f_{Y|X}(y|x^*,\theta) - \frac{f_{Y|X}(y|x^*,\theta)}{b_X(x^*)}\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\nabla_\theta f_{Y|X}(y|x^*,\theta)\right] dy\right\} + o(\Sigma) \\
&= \nabla_\theta \ln f_{Y|X}(y|x^*,\theta) - \frac{1}{b_X(x^*)}\left\{\nabla_\theta b_X(x^*) + \sigma_{jk}\left\{\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\int_{\mathcal{Y}_s} m_{RS}^{jk}(y,x^*)\right.\right. \\
&\quad \left.\left.\left[\nabla_\theta f_{Y|X}(y|x^*,\theta)\, b_X(x^*) - \nabla_\theta b_X(x^*)\, f_{Y|X}(y|x^*,\theta)\right]\right\}\right\} + o(\Sigma) \\
&= \nabla_\theta \ln f_{Y|X}(y|x^*,\theta) - \frac{1}{b_X(x^*)}\left\{\nabla_\theta b_X(x^*) + \sigma_{jk}\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\right. \\
&\quad \left.\int_{\mathcal{Y}_s} m_{RS}^{jk}(y,x^*)\left[\nabla_\theta f_{Y|X}(y|x^*,\theta)\, b_X(x^*) - \nabla_\theta b_X(x^*)\, f_{Y|X}(y|x^*,\theta)\right]\right\} dy + o(\Sigma),
\end{aligned}
$$

$$
\begin{aligned}
g_{Q_t}(z^*) &= Q_t - \frac{\int_{\mathcal{Y}_t} f_{Y|X}(y|x^*,\theta)\, dy}{b_X(x^*)} + \\
&\quad \sigma_{jk}\frac{1}{b_X(x^*)}\int_{\mathcal{Y}_t} f_{Y|X}(y|x^*,\theta)\, dy \sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\int_{\mathcal{Y}_s} f_{Y|X}(y|x^*,\theta)\, m_{YX}^{jk}(y,x^*)\, dy + o(\Sigma) \\
&= Q_t - \frac{\int_{\mathcal{Y}_t} f_{Y|X}(y|x^*,\theta)\, dy}{b_X(x^*)}\left[1 - \sigma_{jk}\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\int_{\mathcal{Y}_s} f_{Y|X}(y|x^*,\theta)\, m_{YX}^{jk}(y,x^*)\, dy\right] + o(\Sigma),
\end{aligned}
$$

and

$$
\begin{aligned}
g_{Q_J}(z^*) &= 1 - \frac{1}{b_X(x^*)} + \sigma_{jk}\frac{1}{b_X(x^*)}\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\int_{\mathcal{Y}_s} f_{Y|X}(y|x^*,\theta)\, m_{YX}^{jk}(y,x^*)\, dy + o(\Sigma) \\
&= 1 - \frac{1}{b_X(x^*)}\left[1 - \sigma_{jk}\sum_{s\in\mathcal{S}}\frac{H_s}{Q_s}\int_{\mathcal{Y}_s} f_{Y|X}(y|x^*,\theta)\, m_{YX}^{jk}(y,x^*)\, dy\right] + o(\Sigma).
\end{aligned}
$$

7.3 Using the derivatives of the log-density of X^* instead of X in approximation $h^a_{Z^*}(z^*)$

Consider the sampling density of X^*, $h_{X^*}(x^*)$, given in (11). The respective log-density is

$$\begin{aligned}\ln h_{X^*}(x^*) &= \ln h_X(x^*) + 0.5\sigma_{jk}\frac{h^{jk}_X(x^*)}{h_X(x^*)} + o(\Sigma) \\ &= l_X(x^*) + \ln b_X(x^*) + 0.5\sigma_{jk}\frac{h^{jk}_X(x^*)}{h_X(x^*)} + o(\Sigma)\end{aligned}$$

and its derivatives are

$$l^t_{h_{X^*}}(x^*) = l^t_X(x^*) + lb^t_X(x^*) + 0.5\sigma_{jk}\left[\frac{h^{tjk}_X(x^*)}{h_X(x^*)} - \frac{h^{jk}_X(x^*)\,h^t_X(x^*)}{h_X(x^*)^2}\right] + o(\Sigma) \tag{66}$$

and

$$l^{vt}_{h_{X^*}}(x^*) = l^{vt}_X(x^*) + lb^{vt}_X(x^*) + 0.5\sigma_{jk}\left[\frac{h^{tjk}_X(x^*)}{h_X(x^*)} - \frac{h^{jk}_X(x^*)\,h^t_X(x^*)}{h_X(x^*)^2}\right]^v + o(\Sigma), \tag{67}$$

where $t = 1, ..., k$, $v = 1, ..., k$, $lb^k_X(x) = [\ln b_X(x)]^k$ and $lb^{jk}_X(x) = [\ln b_X(x)]^{jk}$ are evaluated at X^*

Writing (66) and (67) as, respectively,

$$l^t_X(x^*) = l^t_{h_{X^*}}(x^*) - lb^t_X(x^*) - 0.5\sigma_{jk}\left[\frac{h^{tjk}_X(x^*)}{h_X(x^*)} - \frac{h^{jk}_X(x^*)\,h^t_X(x^*)}{h_X(x^*)^2}\right] - o(\Sigma)$$

and

$$l^{vt}_X(x^*) = l^{vt}_{h_{X^*}}(x^*) - lb^{vt}_X(x^*) - 0.5\sigma_{jk}\left[\frac{h^{tjk}_X(x^*)}{h_X(x^*)} - \frac{h^{jk}_X(x^*)\,h^t_X(x^*)}{h_X(x^*)^2}\right]^v - o(\Sigma)$$

it follows that, in $h^a_{Z^*}(z^*)$ of (9),

$$\sigma_{jk}l^j_{Y|X}(y|x^*,\theta)\,l^k_X(x^*) - \sigma_{jk}l^j_{Y|X}(y|x^*,\theta)\left[l^k_{h_{X^*}}(x^*) - lb^k_X(x^*)\right] = o(\Sigma),$$

$$\sigma_{jk}l^{jk}_X(x^*) - \sigma_{jk}\left[l^{jk}_{h_{X^*}}(x^*) - lb^{jk}_X(x^*)\right] = o(\Sigma),$$

and

$$\sigma_{jk}l^j_X(x^*)\,l^k_X(x^*) - \sigma_{jk}\left[l^j_{h_{X^*}}(x^*) - lb^j_X(x^*)\right]\left[l^k_{h_{X^*}}(x^*) - lb^k_X(x^*)\right] = o(\Sigma).$$

Thus, the order of approximation $h^a_{Z^*}(z^*)$ is not increased by the replacement of $l^k_X(x^*)$ and $l^{jk}_X(x^*)$ by, respectively, $\left[l^t_{h_{X^*}}(x^*) - lb^t_X(x^*)\right]$ and $\left[l^{vt}_{h_{X^*}}(x^*) - lb^{vt}_X(x^*)\right]$.

References

Barron, A. & Sheu C. (1991) Approximation of density functions by sequences of exponential families. The Annals of Statistics **19**, 1347-1369.

Buonaccorsi, J.P. (1996) A modified estimating equation approach to correcting for measurement error in regression. Biometrika **83**, 433-440.

Carroll, R.J., Gail, M. & Lubin, J. (1993) Case-control studies with errors in covariates. Journal of the American Statistical Association **88**, 185-199.

Chesher, A. (1990) The effect of measurement error and a measurement error sensitive specification test. Discussion Paper 90/274, University of Bristol, Department of Economics.

Chesher, A. (1991) The effect of measurement error. Biometrika **78**, 451-462.

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

Chesher, A. (2000) Improved GMM estimation under covariate measurement error. Presented at the Eighth World Congress of the Econometric Society, Seattle.

Chesher (2001) Parameter approximations for quantile regressions with measurement error. CEMMAP Working Paper 02/01, The Institute for Fiscal Studies, UCL, Departement of Economics..

Chesher, A., Dumangane, M. & Smith, R.J. (2002) Duration response measurement error. Journal of Econometrics **111**, 169-194.

Chesher, A., Lancaster, T. & Irish, M. (1985) On detecting the failure of distributional assumptions. Annales de l'INSEE **59/60**, 7-45.

Chesher, A. & Santos Silva, J.M.C. (2002) Taste variation in discrete choice models. Review of Economic Studies **69**, 147-168.

Chesher, A. & Schluter, C. (2002) Welfare measurement and measurement error. Review of Economic Studies **69**, 357-378.

Chesher, A. & Smith, R.J. (1997) Likelihood ratio specification tests. Econometrica **65**, 627-646.

Cosslett, S. (1981a), Efficient estimation of discrete-choice models. In C. Manski and D. McFadden (eds.) Structural Analysis of discrete data with econometric applications. Massachusetts: The MIT Press, pp. 51-111.

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

Cosslett, S. (1993), Endogenous stratification, semiparametric and non-parametric estimation. In G. Maddala, C. Rao & H. Vinod (eds.) Handbook of Statistics *11*. Amesterdam: North-Holland, pp. 1-43.

Dumangane, M. (2000) Essays on duration response measurement error. Ph.D. dissertation, University of Bristol.

Dumangane, M. & Chesher, A. (2001) GMM-estimation with measurement error contaminated duration data. Presented at the 56th European Meeting of the Econometric Society, Lausanne.

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.

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

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

Kiefer, N. & Skoog, G. (1984) Local asymptotic specification error analysis. Econometrica **52**, 873-885.

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

Levine, D. (1985) The sensitivity of the MLE to measurement error. Journal of Econometrics **28**, 223-230.

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

Manski, C. & McFadden, D. (1981), Alternative estimators and sample designs for discrete choice analysis. In C. Manski & D. McFadden (eds.) Structural Analysis of discrete data with econometric applications. Massachusetts: The MIT Press, pp. 2-50.

Muller, P. & Roeder, K. (1997) A Bayesian semiparametric model for case-control studies with errors in variables. Biometrika **84**, 523–537.

Nakamura, T. (1990) Corrected score function for errors-in-variables models: methodology and application to generalized linear models. Biometrika **77**, 127-137.

Newey, W.K. & McFadden, D. (1994), Large sample estimation and hypothesis testing. In R. F. Engle and D. L. McFadden (eds.) Handbook of Econometrics, vol. IV. Amsterdam: North Holland, pp. 2111-2245.

Prentice, R. & Pyke, R. (1979) Logistic disease incidence and case-control studies. Biometrika **66**, 403-411.

Roeder, K., Carroll, R.J. & Lindsay, B.G. (1996) A semiparametric mixture approach to case-control studies with errors in covariables. Journal of the American Statistical Association **91**, 722-732.

Santos Silva, J.M.C. (1999) Unobservables under endogenous sampling. Mimeo, Universidade Tecnica de Lisboa, ISEG.

Stefanski, L.A. (1985) The effects of measurement error on parameter estimation. Biometrika

72, 583-592.

Wang, C.Y. & Carroll, R.J. (1996) On robust estimation in case-control studies with errors in covariates. Mimeo, Texas A&M University, Department of Statistics.

Wang, C., Wang, S. & Carroll, R. (1997) Estimation in choice-based sampling with measurement error and bootstrap analysis. Journal of Econometrics **77**, 65-86.

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

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

Table 1: Experimental designs employed with binary CB samples

Experiment designation	X (mean=3,variance=4)	U (mean=0,variance=0.25)
a	Mixed normal	Normal
b	Mixed normal	Scaled Student t(3)
c	Scaled Student t(3)	Normal
d	Scaled Student t(3)	Scaled Student t(3)

Table 2: Logit model with CB sampling - summary statistics for the slope parameter from 1000 replications

Q=0.9, σ=0.25					
Experiment	H	Estimator	Bias		St. D.
			Mean	Median	
a	0.70	NE	-0.133	-0.140	0.109
		MEa	-0.003	-0.014	0.142
		MEb	-0.036	-0.044	0.164
	0.50	NE	-0.128	-0.132	0.105
		MEa	-0.019	-0.020	0.120
		MEb	-0.020	-0.015	0.130
b	0.70	NE	-0.118	-0.119	0.118
		MEa	0.018	0.012	0.158
		MEb	-0.034	-0.045	0.164
	0.50	NE	-0.112	-0.115	0.112
		MEa	0.005	0.005	0.131
		MEb	-0.014	-0.013	0.140
c	0.70	NE	-0.106	-0.110	0.092
		MEa	-0.022	-0.030	0.112
		MEb	-0.045	-0.062	0.138
	0.50	NE	-0.115	-0.120	0.087
		MEa	-0.053	-0.060	0.101
		MEb	-0.058	-0.078	0.125
d	0.70	NE	-0.094	-0.099	0.096
		MEa	-0.008	-0.016	0.120
		MEb	-0.041	-0.063	0.131
	0.50	NE	-0.101	-0.102	0.091
		MEa	-0.037	-0.039	0.105
		MEb	-0.046	-0.063	0.123

Table 3: Probit model with CB sampling - summary statistics for the parameter of interest from 1000 replications

Experiment a, θ=.75, Q=0.9, σ=0.25				
H	Estimator	Bias		St. D.
		Mean	Median	
0.70	NEa	-0.033	-0.035	0.031
	NEb	-0.124	-0.129	0.052
	MEa	-0.008	-0.011	0.036
	MEb	-0.043	-0.085	0.107
0.50	NEa	-0.047	-0.049	0.027
	NEb	-0.121	-0.123	0.049
	MEa	-0.027	-0.030	0.029
	MEb	-0.008	-0.003	0.094

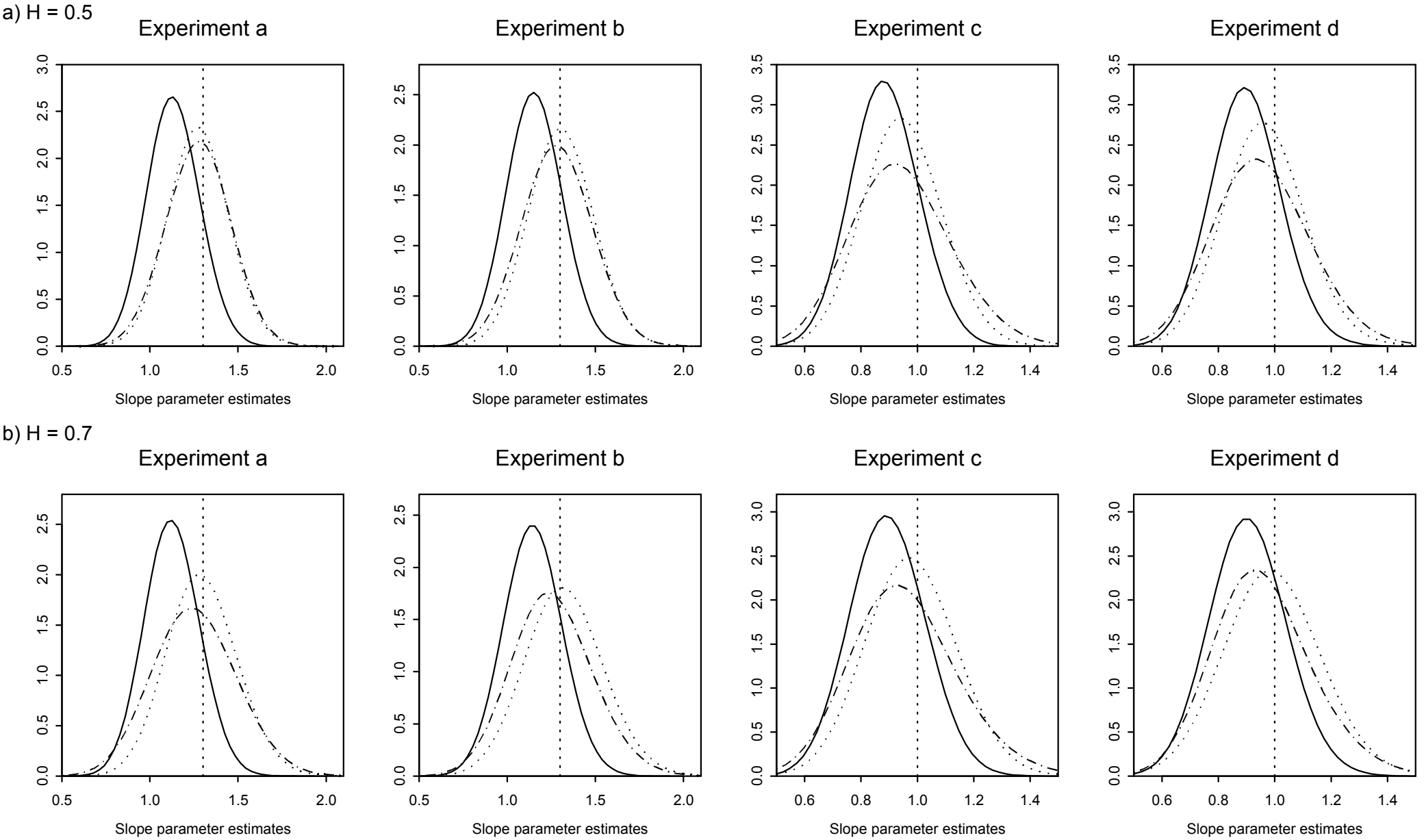
Table 4: Experimental designs employed with binary logistic random samples

Experiment designation	X	U
a	Mixed normal (mean=3,variance=4)	Normal (mean=0,variance=0.25)
b	Normal (mean=0,variance=0.1)	Normal (mean=0,variance=0.1/3)
c	Normal (mean=0,variance=1)	Normal (mean=0,variance=1/3)

Table 5: Logit model with RS - summary statistics for slope parameter from 1000 replications

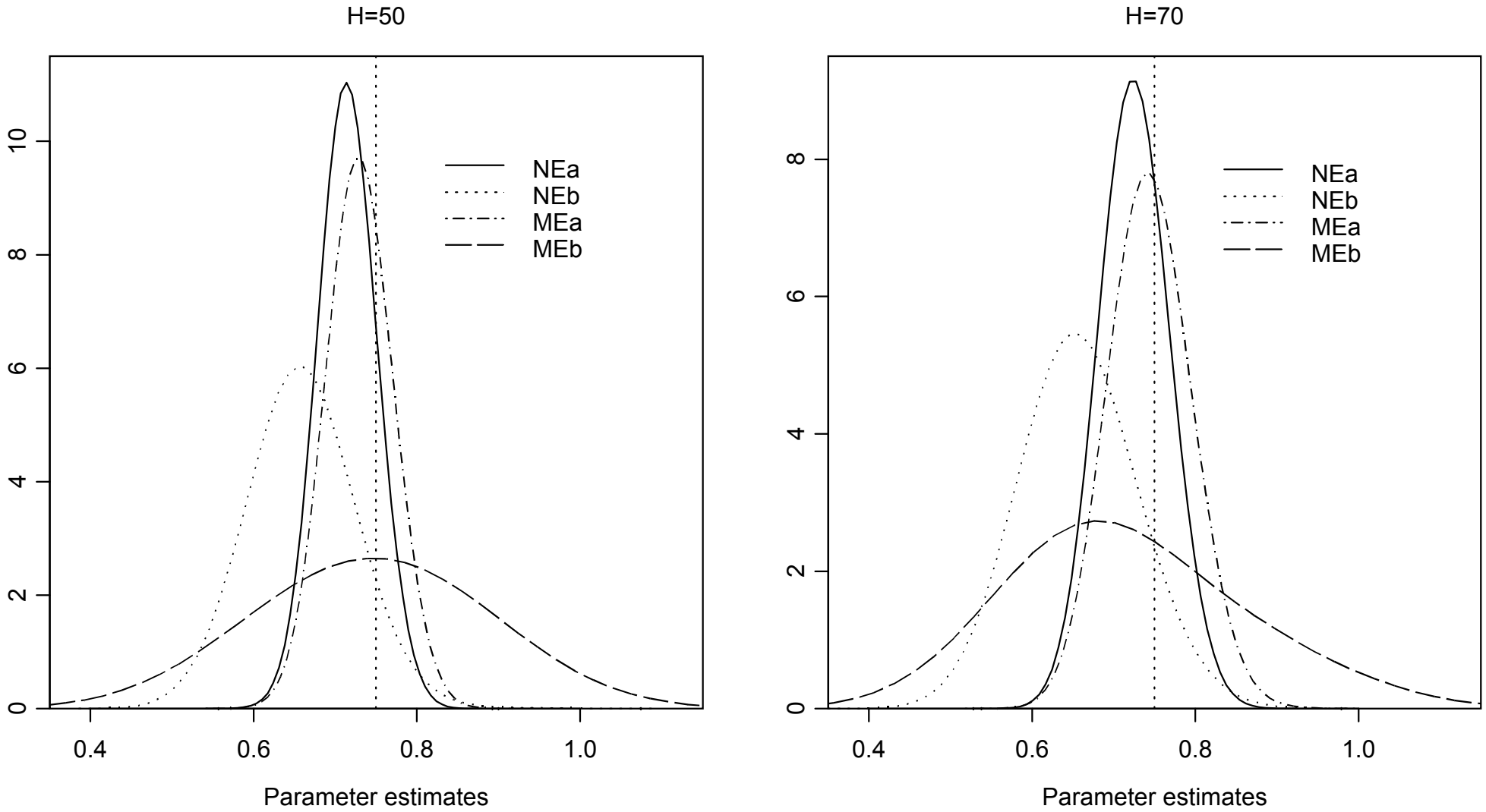
Experiment	Estimator	Bias		St. D.
		Mean	Median	
a	NE	-0.127	-0.137	0.213
	BE	-0.155	-0.170	0.228
	MEa	0.038	0.007	0.315
	MEb	0.022	-0.006	0.309
b	NE	-0.245	-0.251	0.432
	BE	-0.007	-0.012	0.588
	MEa	-0.038	-0.046	0.622
	MEb	0.016	0.049	0.670
c	NE	-0.298	-0.304	0.156
	BE	-0.160	-0.166	0.187
	MEa	-0.029	-0.041	0.262
	MEb	-0.115	-0.149	0.298

Figure 1: Logit model with CB sampling - estimated sampling distributions for the slope parameter estimates



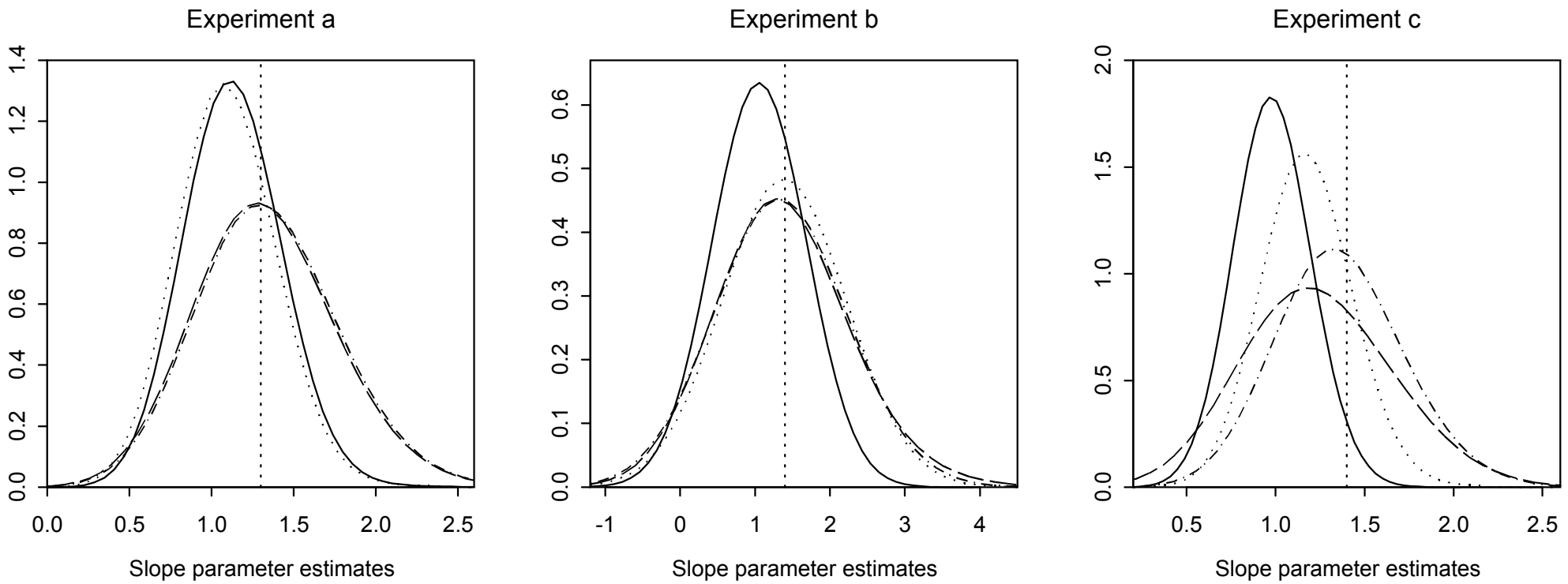
Notes: NE (solid line), MEa (dashed line) and MEb (dot-dashed line). The vertical dotted line indicates the true value of the slope parameter.

Figure 2: Probit model with CB sampling - estimated sampling distributions for the parameter estimates



Notes: The vertical dotted line indicates the true value of the parameter.

Figure 3: Logit model with RS - estimated sampling distributions for the slope parameter estimates



Notes: NE (solid line), BE (dotted line), MEa (dot-dashed line) and MEb (dashed line). The vertical dotted line indicates the true value of the slope parameter.