CEFAGE
universidade de évora
centro de estudos e formação avançada em gestão e economia

CEFAGE-UE Working Paper
2009/15

Covariate Measurement Error: Bias Reduction under Response-based Sampling

Esmeralda Ramalho

Departamento de Economia, Universidade de Évora and CEFAGE-UE

CEFAGE-UE, Universidade de Évora, Largo dos Colegiais 2, 7000-803 Évora - Portugal
Tel.: (+351) 266 740 869, E-mail: cefage@uevora.pt, Web page: http://www.cefage.uevora.pt

Covariate Measurement Error: Bias Reduction under Response-based Sampling

Esmeralda A. Ramalho*

Departamento de Economia, Universidade de Évora and CEFAGE-UE

November 3, 2009

Abstract

In this paper we propose a general framework to deal with the presence of covariate measurement error (CME) in response-based (RB) samples. Using Chesher's (1991) methodology, we obtain a small error variance approximation for the contaminated sampling distributions that characterise RB samples with CME. Then, following Chesher (2000), we develop generalised method of moments (GMM) estimators that reduce the bias of the most well known likelihood-based estimators for RB samples which ignore the existence of CME and derive a score test to detect the presence of this type of measurement error. Our approach only requires the specification of the conditional distribution of the response variable given the latent covariates and the classical additive measurement error model assumption, 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. Monte Carlo evidence is presented which suggests that, in RB samples of moderate sizes, the bias-reduced GMM estimators perform well.

JEL: C51, C52.

Keywords: response-based samples, covariate measurement error, generalized method of moments estimation, score tests.

*Financial support from Fundação para a Ciência e a Tecnologia is gratefully acknowledged (grant PTDC/ECO/64693/2006). The author is grateful to Andrew Chesher, Montezuma Dumangane, Joaquim Ramalho, João Santos Silva, and Richard Smith for comments on previous versions of this paper. Address for correspondence: Esmeralda A. Ramalho, Department of Economics, Universidade de Évora, Largo dos Colegiais, 7000-803 ÉVORA, Portugal (e-mail: ela@uevora.pt).

1 Introduction

Response-based (RB) samples arise in many research settings. In fact, either by imposition of the sampling design or due to the presence of missing data, often the sampling probability depends on the response variable in such a way that the structure of the underlying population of interest is not reflected by the sample. Given that RB samples require, in general, specific inference procedures, they have been widely studied in econometrics; see *inter alia*, Manski and Lerman (1977), Manski and McFadden (1981), Cosslett (1981a,b), Imbens (1992), Imbens and Lancaster (1996), Wooldridge (1999, 2001), and the survey by Ramalho and Ramalho (2006).

On the other hand, in recent years, there has been an increasing interest in inference methods to deal with covariate measurement error (CME) in nonlinear models; see, for example, Chesher (1991, 2000, 2001), Schennach (2004a, 2004b, 2007), Hu and Ridder (2007), and Hu and Schennach (2008). However, the analysis of RB samples in this framework has only been addressed by a few papers in the statistical literature that focus on the particular case of binary logistic choice-based (CB) samples, where inference procedures are specially simple since, after correcting for CME, the endogeneity of the sampling can be ignored and estimation techniques used under random sampling (RS) may be employed; see *inter alia*, Carroll, Gail and Lubin (1993), Roeder, Carroll and Lindsay (1996), Muller and Roeder (1997), and Wang, Wang and Carroll (1997). Moreover, these approaches rely on very strong assumptions, requiring, for example, the specification of an exact form for the relationship between the error-free variables and their error-prone measures and the availability of a validation sample.

In this paper we propose an approach to deal with CME in general RB samples that circumvent those strong assumptions. Specifically, we use Chesher's (1991) methodology to obtain an approximate form for the contaminated sampling distributions for a small error variance, which allows us to accommodate CME in standard models for RB samples. This approach merely requires the specification of the structural model, characterized by the error-free conditional distribution of the variable of interest given the covariates, which is a standard assumption in all the likelihood-based estimators for RB samples cited before, and the existence of CME of the classical additive kind, such that the measurement error and the true covariates are independent, which was also required in all the previous approaches on CME on RB samples. In fact, the approximations employed are only dependent on features of the structural model and on the variance of the measurement error and the derivatives of the log-density of the error-free distributions of the covariates, which can be estimated from the data.

This flexible setting is used to derive estimators that reduce the bias of the naive estimators for RB samples that ignore CME and to derive a score test to detect this type of measurement error. We consider the generalized method of moment (GMM) estimation framework proposed by Ramalho and Ramalho (2006) to deal with error-free RB samples, which nests all the previous likelihood-based estimators for RB samples as particular cases. First, we derive an inconsistency measure for naive GMM estimators. Then, following Chesher (2000), we derive modified GMM estimators where the bias is reduced when compared with that of the naive GMM estimators that ignore CME. Basically, the moment indicators of Ramalho and Ramalho (2006) 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. The score test is a straightforward by-product of this approach.

The remainder of the paper is organized as follows. Section 2 formalizes the likelihood functions which take into account the presence of CME in RB samples. Section 3 develops GMM inference procedures appropriate for this framework. Section 4 reports some Monte Carlo evidence on the performance in practice of some of the proposed estimators. Finally, section 5 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 RB 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, providing a very convenient framework to investigate the impact of CME in this sampling scheme.

2.1 Response-based sampling

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.

RB sampling may be seen as the result of a stratification mechanism where the probability of observing a sampling unit depends of the response variable. The stratification may be deliberate, in cases where the sampling design defines that stratification, or not, in cases where the aim is collecting a random sample of the population of interest but some sampling units are not observed (because they decided not to participate in the survey, for example). Independently of the motivation for the stratification, the model to describe RB sampling assumes that the population of interest is partitioned into J non-empty and possibly overlapping strata, which are subsets of $\mathcal{Y} \times \mathcal{X}$, from each of which a random sample is drawn. For simplicity, suppose that these strata are defined only in terms of the response variable and designate each stratum 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

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

Assume that the sample is drawn according to the multinomial sampling scheme and define the sampling probability of each stratum in the sample as H_s.[1] In this setting, the sampling density function of $Z = (Y, X, S)$ is given by

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

where $(y, x) \in \mathcal{C}_s$, $s \in \mathcal{S}$, and $b_S(s) = \frac{H_s}{Q_s}$. 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

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

Both $b_S(s)$ and $b_X(x)$ reflect the bias induced by the nonrandom sampling design over the population densities $f_{YX}(y,x) = f_{Y|X}(y|x, \theta) f_X(x)$ and $f_X(x)$. Only when the sample is self-weighted, in which case the proportion of the strata in the population is preserved in the sample, that is H_s equals Q_s, as $b_S(s) = b_X(x) = 1$, does RB sampling become equivalent to RS; see Ramalho and Ramalho (2006) for more details.

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 RB

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

sampling takes the designation of CB sampling. Actually, most papers on RB sampling focus on this particular case; 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 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^*, $h_Z(z^*) = b_S(s) f_{Y|X}(y|x^*, \theta) f_X(x^*)$, 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, using (3) and (6),

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

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

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

the derivation of $h_{Z^*}(z^*)$ is not straightforward.[2] However, by employing Chesher's (1991) method, we may obtain an approximation for (8) for a small error variance that does not depend on $f_U(u)$ and for which the validity depends essentially on the fact that higher moments of the measurement error distribution are small relatively to its variance. This approach, which uses an approximate likelihood function to describe the contaminated data, has already been used in the analysis of duration response measurement error (Dumangane 2000, Chesher, Dumangane and Smith 2002, and Dumangane 2006), 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,

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

with

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

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

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

of the error-prone covariates in the sample,

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

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 RB 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 it may not be positive for all Z^* and integrate to one. Thus, they may not be used directly for maximum likelihood (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)

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

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 considered by Ramalho and Ramalho (2006) using expectations taken with respect to approximation $h_{Z^*}^{a}(z^*)$ in (9). However, (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.[3]

[3] Note that the maximization of a log likelihood function based on (12) to obtain estimators for RB samples which are an extension of those of Imbens and Lancaster (1996) would be a much more complex approach, since it would involve the derivation of ML estimators for a set of support points of the marginal distribution of X, which would be replaced in the first order conditions of the remaining parameters of interest.

3 Generalized method of moments estimation

Under the assumption of correct measurement of all variables, Ramalho and Ramalho (2006) show that the most well known likelihood-based estimators for RB samples can be seen as GMM estimators resulting from the use of a set of moment indicators $g_\gamma(z)$, where γ is the vector of parameters to be estimated, which is some combination of

$$g_{H_t}(z) = H_t - 1(s = t), \qquad t = 1, ..., J-1, \tag{13}$$

$$g_\theta^1(z) = b_S(s)^{-1} \nabla_\theta \ln f_{Y|X}(y|x, \theta) \qquad \text{or} \qquad g_\theta^2(z) = \nabla_\theta \ln f_{Y|X}(y|x, \theta) - b_x^{-1} \nabla_\theta b_x \tag{14}$$

$$g_{Q_t}(z) = Q_t - b_x^{-1} \int_{\mathcal{Y}_t} f(y|x; \theta)\, dy, \qquad t = 1, ..., J, \tag{15}$$

where $I_{(s=t)}$ takes the value 1 for $s = t$ and 0 for $s \neq t$, $t = 1$, ..., $J-1$, and ∇_θ denotes derivative with respect to θ.

In fact, Manski and Lerman's (1977) weighted maximum likelihood (ML) and Manski and McFadden's (1981) conditional ML estimators use, respectively, $g_\gamma(z) = g_\theta^1(z)$ and $g_\gamma(z) = g_\theta^2(z)$ with $H = (H_1, ..., H_{J-1})$ and $Q = (Q_1, ..., Q_J)$ replaced by their known values, while Imbens and Lancaster's (1996) GMM estimators, which are generalizations of those of Cosslett (1981a,b) and Imbens (1992) for a continuous response variable, use the combination $g_\gamma(z) = [g_{H_t}(z), g_\theta^2(z), g_{Q_t}(z)]$. Imbens and Lancaster's (1996) estimators, and also the alternative estimator that employs $g_\gamma(z) = [g_{H_t}(z), g_\theta^1(z), g_{Q_t}(z)]$ are 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 of parameters of interest 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. Note that only the estimators by Imbens and Lancaster (1996) (and, consequently those of Cosslett 1981a,b and Imbens 1992) are asymptotically efficient.

The objective function to be minimized is

$$\Upsilon_N(\gamma) = g_N(\gamma)' W_N g_N(\gamma), \tag{16}$$

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), the moment indicators $g_\gamma(z_i)$ are given in (13)-(15), and W_N is a positive semi-definite weighting matrix. The resulting optimal estimator, $\hat{\gamma}$, obtained from the use of the weighting matrix $W_N = \Omega_N^{-1}$ in (16), 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, $\sqrt{N}(\hat{\gamma}-\gamma^0) \xrightarrow{d} N\left[0,(G'\Omega^{-1}G)^{-1}\right]$, where $G=E_{h_Z}\left[\nabla_\gamma g_\gamma(z)'\right]$.

In the remainder of this section, we first derive a measure for the inconsistency of the GMM estimators based on (13)-(15) 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 indicators in (13)-(15) 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. 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. Finally, subsection 3.5 discusses the particular case of CB logistic samples, where the estimation procedure may be simplified.

3.1 Inconsistency of naive GMM estimators

In presence of CME, the naive GMM estimators $\hat{\gamma}$, which merely replace X by X^* in combinations of the moment indicators (13)-(15), 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. This bias is a consequence of the fact that the expected value of these moment indicators taken under the distribution of the observed data, $h_{Z^*}(z^*)$, is not zero. In effect, using approximation $h^a_{Z^*}(z^*)$ in (9) to calculate an approximation for this expectation, we find

$$\begin{aligned}
E_{h_{Z^*}}\left[g_\gamma(z^*)\right] &= \sum_{s\in\mathcal{S}}\int_{\mathcal{Y}_s}\int_{\mathcal{X}^*} g_\gamma(z^*)\, h(z^*,\gamma)\left[1+\sigma_{jk}m^{jk}_{YX}(y,x^*)\right]dx^*dy+o(\Sigma)\\
&= E_{h_Z}\left[g_\gamma(z^*)\right]+\sigma_{jk}E_{h_Z}\left[g_\gamma(z^*)\, m^{jk}_{YX}(y,x^*)\right]+o(\Sigma)\\
&= \sigma_{jk}E_{h_Z}\left[g_\gamma(z^*)\, m^{jk}_{YX}(y,x^*)\right]+o(\Sigma)\\
&= b_\phi+o(\Sigma), \qquad (17)
\end{aligned}$$

where $\phi=(\gamma,\sigma)$, with σ defined as a vector of dimension D containing all the different nonzero elements of matrix Σ.[4] Thus, b_ϕ may be interpreted as the approximate bias in the original moment indicators incurred by the presence of measurement error.[5] The approximate biases in

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

[5] Recall that, previously, we had already defined two bias functions, $b_S(s)$ and $b_X(x)$, with a very different nature from that in (17), because they reflect distortions imposed by the endogenous sampling scheme.

moment indicators (13)-(15), derived in appendix 6.1, are given by, respectively,

$$b_{H_t} = 0 \tag{18}$$

$$b_{\theta}^{1} = \sigma_{jk} E_{f_X} \left[\sum_{s \in \mathcal{S}} \int_{\mathcal{Y}_s} \nabla_{\theta} f_{Y|X} (y|x^*, \theta)\, m_{YX}^{RS^{jk}} (y, x^*)\, dy \right] \tag{19}$$

$$\begin{aligned} b_{\theta}^{2} &= \sigma_{jk} E_{f_X} \left\{ \sum_{s \in \mathcal{S}} b_S (s) \int_{\mathcal{Y}_s} m_{YX}^{RS^{jk}} (y, x^*) \left[\nabla_{\theta} f_{Y|X} (y|x^*, \theta) - b_X (x^*)^{-1} f_{Y|X} (y|x^*, \theta) \right. \right. \\ & \left. \left. \nabla_{\theta} b_X (x^*) \right] dy \right\} \end{aligned} \tag{20}$$

$$b_{Q_t} = -\sigma_{jk} E_{f_X} \left[b_X (x^*)^{-1} \int_{\mathcal{Y}_t} f_{Y|X} (y|x^*, \theta)\, dy \sum_{s \in \mathcal{S}} b_S (s) \int_{\mathcal{Y}_s} f_{Y|X} (y|x^*, \theta)\, m_{YX}^{jk} (y, x^*)\, dy \right] \tag{21}$$

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 the other moment indicators are concerned, the bias b_{ϕ} is eliminated only when there is no mismeasurement, in which case $\sigma_{jk} = 0$.

The bias of the original moment indicators causes the inconsistency of conventional estimators for RB samples. Using the framework of Hall (2005), p. 121-122, concerning the asymptotic theory for first step GMM estimators based on misspecified models, and the approximate bias in the moment indicators given in (17), the bias of the first step naive GMM estimators, $\hat{\gamma}_{1s}$, may be written as

$$\sqrt{N} \left(\hat{\gamma}_{1s} - \gamma^0 \right) = -\sigma_{jk} \left(G' W_N G \right)^{-1} G' W_N \sqrt{N} E_{h_Z} \left[g_{\gamma^0} (z^*)\, m_{YX}^{jk} (y, x^*) \right] + o(\Sigma). \tag{22}$$

This bias is then transmitted to the estimators of the subsequent steps via the weighting matrix W_N.

In the subsequent subsections it will be shown that the approximate bias functions in (18)-(21) may be used not only to modify the original moment indicators (13)-(15) to handle CME, but also in the implementation of an efficient version of a score test sensitive to the presence of CME.

3.2 Derivation of corrected moment indicators

In this subsection we employ Chesher's (2000) method to obtain a modified set of moment indicators where the bias is reduced when compared with that of the original ones; see also Dumangane (2000, 2006) who follow the same approach to handle response measurement error in duration models, correcting the score functions of the models commonly employed in that area. The idea is very simple. As shown in (17), the expectation of the original moment indicators evaluated at Z^* taken with respect to the approximate contaminated density $h^a_{z^*}(z^*)$ is not zero but $b_\phi + o(\Sigma)$. Hence, if we subtract b_ϕ from the original moment indicators, the resulting modified moment indicators,

$$g^*_\phi(z^*) = g_\gamma(z^*) - b_\phi, \tag{23}$$

have expectation $E_{h_{Z^*}}\left[g^*_\phi(z^*)\right] = o(\Sigma)$. Although this expectation is not zero with CME, (23) may be used to eliminate a substantial part of the bias of the naive GMM 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_ϕ, 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 $b_S(s_i)^{-1}$ or $b_X(x)^{-1}$. 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 (23) 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 $h^a_{Z^*}(z^*)$ in (9). Thus, in presence of CME, we suggest the utilization of the base set of modified moment indicators given by

$$g^*_{H_t}(z^*) = H_t - I_{(s=t)} \tag{24}$$

$$g^{*^1}_\theta(z^*) = b_S(s)^{-1}\nabla_\theta \ln f_{Y|X}(y|x,\theta) - \sigma_{jk} b_X(x)^{-1} \sum_{s\in\mathcal{S}} \int_{\mathcal{Y}_s} \nabla_\theta f_{Y|X}(y|x^*,\theta)\, m^{RS^{jk}}_{YX}(y,x^*)\,dy \tag{25}$$

$$\begin{aligned} g^{*^2}_\theta(z^*) &= \nabla_\theta \ln f_{Y|X}(y|x^*,\theta) - b_X(x^*)^{-1}\sum_{s\in\mathcal{S}} b_S(s) \int_{\mathcal{Y}_s} \left\{\nabla_\theta f_{Y|X}(y|x^*,\theta) + \sigma_{jk} b_X(x)^{-1} m^{RS^{jk}}_{YX}(y,x^*)\right. \\ &\quad \left.\left[\nabla_\theta f_{Y|X}(y|x^*,\theta)\, b_X(x^*) - \nabla_\theta b_X(x^*)\, f_{Y|X}(y|x^*,\theta)\right]\right\} dy + o(\Sigma) \end{aligned} \tag{26}$$

$$\begin{aligned} g^*_{Q_t}(z^*) &= Q_t - b_X(x^*)^{-1}\int_{\mathcal{Y}_t} f_{Y|X}(y|x^*,\theta)\,dy \left[1 - \sigma_{jk} b_X(x)^{-1}\sum_{s\in\mathcal{S}} b_S(s)\int_{\mathcal{Y}_s} f_{Y|X}(y|x^*,\theta)\, m^{jk}_{YX}(y,x^*)\,dy \right. \\ &\quad + o(\Sigma) \end{aligned} \qquad (2$$

$$g^*_{\sigma_{jk}}(z^*) = \frac{m^{jk}_{YX}(y,x^*)}{1+\sigma_{jk} m^{jk}_{YX}(y,x^*)} + o(\Sigma), \tag{28}$$

which is composed of (13), modified versions of (14)-(15) calculated in appendix 6.2, and the additional moment indicators (28) concerned with the estimation of the variance of U. Naturally, with correct measurement, as $\Sigma = 0$, moment indicators (24)-(27) coincide with their original counterparts in (13)-(15). It is also apparent how this system is simplified in case of self-weighting or RS: only the moment indicators (25) and (28) are employed with $b_S(s) = b_X(x^*) = 1$, which yields $g_\theta^{RS^*}(z^*) = \nabla_\theta \ln f(y|x,\theta) - \sigma_{jk} \int_{\mathcal{Y}} \nabla_\theta f_{Y|X}(y|x^*,\theta)\, m_{YX}^{RS^{jk}}(y,x^*)\, dy + o(\Sigma)$ and $g_{\sigma_{jk}}^{RS^*}(z^*) = \frac{m_{YX}^{RS^{jk}}(y,x^*)}{1+\sigma_{jk} m_{YX}^{RS^{jk}}(y,x^*)} + o(\Sigma)$.[6]

The system (24)-(28) 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 (16), 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 by combinations of (24)-(28). Analogously to RB sampling with no measurement error, if Q or σ, or both quantities, are known, their values are substituted in (24)-(28), 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 nor σ are known, a just-identified GMM estimator for $\phi = (H,\theta,Q,\sigma)$ needs to be calculated.

As $E_{h_{Z^*}}\left[g_\phi^*(z^*)\right] = o(\Sigma)$, only when CME is absent will the MGMM estimators be consistent for the parameters of interest. Otherwise, with CME, the probability limit of the MGMM estimators $\hat{\phi}$ is ϕ^* and not the true value ϕ^0. However, part of the bias induced by CME is removed. The magnitude of this bias-reduction depends naturally on the variance of the measurement error: the smaller are the components of Σ, the closer are the moment conditions $E_{h_{Z^*}}\left[g_\phi^*(z^*)\right]$ to zero and, consequently, the closer is ϕ^* to the true value ϕ^0. For examples of estimators where the bias is reduced by using small parameters approximations, see Chesher and Santos Silva (2002), who proposed a quasi-ML estimator for logit models with taste variation, and Dumangane (2006), who derived bias-reduced GMM estimators for duration response measurement error.

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. This type of test was suggested in the ML framework by Chesher (1990) and applied, for example, by Chesher, Dumangane and Smith (2002) in the context of duration models. The idea is testing if the D elements of vector σ are zero. The null hypothesis is $H_0 : \sigma = 0$,

[6] Note that both (25) and (26) are reduced to the same result under self-weighting or RS.

for which the score test statistic (see Newey and McFadden 1994, Theorem 9.2.) is given by

$$T = N g_N^{*\prime} \Omega_N^{*-1} G_N^* V_N^* G_N^{*\prime} \Omega_N^{*-1} g_N^*, \tag{29}$$

where $g_N^* \equiv g_N^*(\phi)$ and Ω_N^*, G_N^* and V_N^* are consistent estimators of, respectively, $\Omega^* = E_{h_{Z^*}}\left[g_\phi^*(z^*)\right.$ $\left.g_\phi^*(z^*)'\right]$, $G^* = E_{h_{Z^*}}\left[\nabla_\phi g_\phi^*(z^*)'\right]$ 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 (24)-(28) are reduced to, respectively, (13)-(15) and

$$g_{\sigma_{jk}}^*(z^*)\Big|_{\sigma_{jk}=0} = m_{YX}^{jk}(y, x^*), \tag{30}$$

which may also be obtained from the maximization of the log-likelihood based on (12), the augmented density of the type 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 (13)-(15) and (30) is given by the approximate bias functions (18)-(21) with σ_{jk} suppressed.

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 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 RB 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 (25)-(28) or the test statistics (29), in such a way that the order of the approximation error in $h_{Z^*}(z^*)$ of (9) is not increased. In fact, the log-density of $h_{X^*}(x^*)$ given in (11) can be written as

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

and its derivatives are

$$l_{h_{X^*}}^t(x^*) = l_X^t(x^*) + lb_X^t(x^*) + 0.5\sigma_{jk}\left[\frac{h_X^{tjk}(x^*)}{h_X(x^*)} - \frac{h_X^{jk}(x^*)\,h_X^t(x^*)}{h_X(x^*)^2}\right] + o(\Sigma) \tag{31}$$

and

$$l_{h_{X^*}}^{vt}(x^*) = l_X^{vt}(x^*) + lb_X^{vt}(x^*) + 0.5\sigma_{jk}\left[\frac{h_X^{tjk}(x^*)}{h_X(x^*)} - \frac{h_X^{jk}(x^*)\,h_X^t(x^*)}{h_X(x^*)^2}\right]^v + o(\Sigma), \tag{32}$$

where $t = 1, ..., k$, $v = 1, ..., k$, $lb_X^k(x) = [\ln b_X(x)]^k$ and $lb_X^{jk}(x) = [\ln b_X(x)]^{jk}$ are evaluated at X^*. Because in the previous subsections $l_X^k(x^*)$ and $l_X^{jk}(x^*)$ appear in terms of first order in σ_{jk}, they can be simply written from (31) and (32) as

$$l_X^k(x^*) = l_{h_{X^*}}^k(x^*) - lb_X^k(x^*) \tag{33}$$

$$l_X^{jk}(x^*) = l_{h_{X^*}}^{jk}(x^*) - lb_X^{jk}(x^*), \tag{34}$$

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 (33) and (34) 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 $l_{h_{X^*}}^k(x^*)$ and $l_{h_{X^*}}^{jk}(x^*)$, 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

$$L(\beta) = h_{X^*}^0(x^*)\exp\left\{\sum_{l=1}^{M}\beta_l\omega_l(x^*) - \ln\int_0^1 h_{X^*}^0(x^*)\exp\left[\sum_{l=1}^{M}\beta_l\omega_l(x^*)\right]dx^*\right\}, \tag{35}$$

where $h_{X^*}^0(x^*)$ 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 (35) may be rewritten as

$$L(\beta) = \sum_{l=1}^{M}\beta_l\left[\omega_l(x^*) - \bar{\omega}_l\right] - \ln\int_0^1 \exp\left\{\sum_{l=1}^{M}\beta_l\left[\omega_l(x^*) - \bar{\omega}_l\right]\right\}dx^*, \tag{36}$$

where $\bar{\omega}_l$ is the sample mean of the lth Legendre polynomial. Note that maximizing (36) 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

$$R(\beta)^a = \frac{1}{T+1}\sum_{t=1}^{T+1} \exp\left\{\sum_{l=1}^{M} \beta_l \left[\omega_l\left(\frac{t-1}{T}\right) - \bar{\omega}_l\right]\right\}. \tag{37}$$

The resulting log-density derivatives are simply

$$l^k_{h_{X^*}}(x^*) = \sum_{l=1}^{M} \beta_l \omega_l^k (x^*) \tag{38}$$

and

$$l^{jk}_{h_{X^*}}(x^*) = \sum_{l=1}^{M} \beta_l \omega_l^{jk} (x^*). \tag{39}$$

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 (33) and (34), respectively. Next, we may replace $l^k_X(x^*)$ and $l^{jk}_X(x^*)$ by, respectively, (33) and (34) in $m^{jk}_{YX}(y, x^*)$ contained in both the moment indicators (25)-(28) and the test statistics in (29). In our Monte Carlo experiments, similarly to Chesher (1998), we set $T = 100$ and $M = 6$. Moreover, although an uniform density on $[0, 1]$ is assumed in (35), X is mapped onto the interval $[0.1, 0.9]$ to avoid an unreasonable tail behaviour.

In practice, regression models typically include several covariates, which complicates the estimation of (38) and (39). However, following Chesher (1998), the problem can be simplified by taking into account that the interest is in the estimation of derivatives with respect to a covariate measured with error, say X_1, of the conditional log density of X_2 given X_1, where X_2 excludes X_1 from X. The problem is specially simplified in cases where this conditional distribution is assumed to be a member of a location scale family with location determined by a linear index which is a function of X_1; see section 4.3 of Chesher (1998). This approach was followed in the simulation experiments of section 4 involving multiple covariates.

3.5 The particular case of 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, Carroll, Gail and Lubin (1993), 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 approximations. In absence of CME, 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 of the conditional probability of Y given X^* present the same structure in the population and the sample. As in Chesher (1991), the former approximation may be expressed by

$$P_1^* = P\left[1 + \sigma_{jk}\Lambda^{jk}(1-P)\right] + o(\Sigma), \tag{40}$$

where $P_1^* = \Pr_{Y|X^*}(1|x^*, \theta, \sigma)$, $P = \Pr_{Y|X}(1|x^*, \theta) = \left(1 + e^{-x^{*\prime}\theta}\right)^{-1}$, with x^* containing a constant term, and $\Lambda^{jk} = 0.5\theta^j\theta^k\left[1 - 2P + \frac{2}{\theta^j} l_X^k(x^*)\right]$. Denoting the marginal 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^*, $P_1^{CB^*} = \Pr_{Y|X^*}^{CB}(1|x^*, \theta, \Sigma)$, is

$$\begin{aligned}
P_1^{CB^*} &= \frac{\frac{H}{Q}P_1^*}{\frac{H}{Q}P_1^* + \frac{1-H}{1-Q}P_0^*} + o(\Sigma) \\
&= \left(1 + \frac{Q}{H}\frac{1-H}{1-Q}\frac{1-P_1^*}{P_1^*}\right)^{-1} + o(\Sigma) \\
&= \left(1 + \frac{Q}{H}\frac{1-H}{1-Q}\frac{1-P}{P}\frac{1-\sigma_{jk}\Lambda^{jk}P}{1+\sigma_{jk}\Lambda^{jk}(1-P)}\right)^{-1} + o(\Sigma),
\end{aligned} \tag{41}$$

where $P_0^* = \Pr_{Y|X^*}(0|x^*, \theta, \sigma) = (1-P)\left(1 - \sigma_{jk}\Lambda^{jk}P\right) + o(\Sigma)$. Because $\frac{1-P}{P}$ and $\frac{1-\sigma_{jk}\Lambda^{jk}P}{1+\sigma_{jk}\Lambda^{jk}(1-P)}$ can be written, respectively, as $e^{-x'\theta}$ and $1 - \sigma_{jk}\Lambda^{jk} + o(\Sigma)$

$$\begin{aligned}
P_1^{CB^*} &= \left(1 + \frac{Q}{H}\frac{1-H}{1-Q}e^{-x'\theta} - \sigma_{jk}\Lambda^{jk}\frac{Q}{H}\frac{1-H}{1-Q}e^{-x'\theta}\right)^{-1} + o(\Sigma) \\
&= \left(1 + \frac{Q}{H}\frac{1-H}{1-Q}e^{-x'\theta}\right)^{-1}\left[1 - \sigma_{jk}\Lambda^{jk}\left(1 + \frac{Q}{H}\frac{1-H}{1-Q}\right)^{-1}\frac{Q}{H}\frac{1-H}{1-Q}e^{-x'\theta}e^{-x'\theta}\right]^{-1} + o(\Sigma) \\
&= \left(1 + \frac{Q}{H}\frac{1-H}{1-Q}e^{-x'\theta}\right)^{-1}\left[1 + \sigma_{jk}\Lambda^{jk}\left(1 + \frac{Q}{H}\frac{1-H}{1-Q}\right)^{-1}\frac{Q}{H}\frac{1-H}{1-Q}e^{-2x'\theta}\right] + o(\Sigma),
\end{aligned} \tag{42}$$

where, in the last step, we use the fact that $\left[1 - \sigma_{jk}\Lambda^{jk}\left(1 + \frac{Q}{H}\frac{1-H}{1-Q}\right)^{-1}\frac{Q}{H}\frac{1-H}{1-Q}e^{-2x'\theta}\right]^{-1} = 1 + \sigma_{jk}\Lambda^{jk}\left(1 + \frac{Q}{H}\frac{1-H}{1-Q}\right)^{-1}\frac{Q}{H}\frac{1-H}{1-Q}e^{-2x'\theta} + o(\Sigma)$.

Because the first term of (42) corresponds to P_1^{CB} and the term multiplying $\sigma_{jk}\Lambda^{jk}$ inside the squared brackets is $1-P_1^{CB}$, the relationship between $P_1^{CB^*}$ and P_1^{CB} presents the same structure of that of P_1^* and P. Thus, we may obtain bias-reduced estimators for the slope parameters of interest as well as the intercept terms displaced by $-\ln\left(\frac{Q}{H}\frac{1-H}{1-Q}\right)$ by using the reduced system employed under RS, described below (24)-(28), which is composed only by two classes of moment indicators. Note, however, that those moment indicators contain $l_X^k(x^*)$, which, as in CB samples the sampling density of the covariates, $h_{X^*}(x^*)$, deviates from $f_{X^*}(x^*)$, has to be substituted for (33), instead of being directly estimated as in RS; see subsection 3.4. Moreover, as $lb_X(x^*)^k$ in (33) is a function of Q_y, this method may only be used when this marginal probability is known.

In this setting, the use of the extended system of moment indicators in (24)-(28) is circumvented. Relative to previous papers on CB logistic samples, our GMM estimator offers the advantage of avoiding both the availability of a validation sample and the formulation of a conditional distribution or a conditional expected value describing the relationship between the observable and the error-free covariates.

4 Performance in practice

In this section we undertake two small Monte Carlo simulation studies to investigate the finite sample behaviour of some of the MGMM estimators described previously. We focus on cases of binary data subject to CB sampling. First, subsection 4.1 considers cases where the structural model is logit and, thus, the simplified estimation procedures described in subsection 3.5 may be employed. Then, in subsection 4.2 we simulate loglog and probit models, which require the use of the complete methodology proposed in subsection 3.2. In this case, we focus on the correction of Imbens and Lancaster's (1996) efficient estimators. In both subsections, 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 $Q\ (1-Q)$. In most of the experiments Q was set equal to 0.9 and, for each experiment, two sampling designs were considered, characterized by $H=\{0.5, 0.7\}$. Note that the sampling scheme where $H=0.5$ is claimed to be close to an optimal design, in the sense that minimizes the asymptotic variance of the estimators; see, for example, Cosslett (1981a) and Imbens (1992). All experiments, implemented in S-Plus, are based on 1000 replications for a sample size of 500.

4.1 Logit model with CB sampling

In this first set of experiments the variable of interest Y, conditional on X, is distributed as logit with $\Pr_{Y|X}(1|x,\theta) = \left(1+e^{-x\theta}\right)^{-1}$ and the marginal choice probability Q is assumed known. In most of the experiments we consider $\theta = (\theta_0, \theta_1)$ and generate an error-free covariate with mean 3 and variance 4, either as a mixture of normal distributions, where the variate is $N(2, 1.2915^2)$ with probability 0.7 and $N(5.333, 1.2915^2)$ 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 1.048 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. Two values are considered for the variance of U, denoted as σ: $\sigma = \{0.16, 0.25\}$. In designs a and c, U follows a $N(0,\sigma)$ distribution, while in b and d, U is a scaled chi-square variate with 4 degrees of freedom. Table 1 summarizes the experimental designs just described.

Table 1 about here

Additionally, we perform a restricted set of experiments where logit models with multiple covariates are considered. The experimental design follows closely one of Chesher's (1998) designs. We consider models in which two error-free covariates contained in X_2 are generated from a multivariate normal distribution $N(0, I_2)$ and a variable X_1, generated as $X_1 = X_2'\xi + W$, where $\xi = \left(\frac{1}{3}, -\frac{1}{3}\right)$ and W follows the mixed normal distribution described previously, is subject to measurement error following a $N(0,\sigma)$ or a scaled χ_4^2 distribution for $\sigma = 0.25$. We set $\theta = (\theta_0, \theta_1, \theta_2, \theta_3) = (1, -1, 1, 0)$, which yields $Q = 0.8$, and focus on cases where $H = 0.5$.

Three different estimators were calculated: the naive (N) GMM estimator (which in this case is a ML estimator) and the MGMM estimators for known and unknown σ, respectively designated Mσ and M. 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,

$$g_{\theta_0}^*(z^*) = \frac{p}{P(1-P)}\left[y - P - 0.5\sigma\theta_1^2 p\left(\frac{p^1}{p} + \frac{2}{\theta_1}l_X^1\right)\right] + o(\Sigma) \quad (43)$$

$$g_{\theta_1}^*(z^*) = \frac{xp}{P(1-P)}\left[y - P - 0.5\sigma\theta_1^2 p\left(\frac{p^1}{p} + \frac{2}{\theta_1}l_X^1\right)\right] + o(\Sigma) \quad (44)$$

$$g_{\sigma}^*(z^*) = \frac{\frac{\theta_1^2 p(y-P)}{P(1-P)}\left(\frac{p^1}{p} + \frac{2}{\theta_1}l_X^1\right)}{2 + \sigma\frac{\theta_1^2 p(y-P)}{P(1-P)}\left(\frac{p^1}{p} + \frac{2}{\theta_1}l_X^1\right)} + o(\Sigma), \quad (45)$$

where $P = \Pr_{Y|X}(1|x^*, \theta)$, $p = \nabla_{\theta_0} P$ and $p^1 = \nabla_{\theta_0'} p$. In the M$\sigma$ case, σ was replaced by its known value, while for the M case σ was estimated simultaneously with the other parameters of interest. With regard to the NE, the moment indicators employed are (43) and (44) 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 of N, Mσ, and M estimators for $\sigma = 0.16$. Figure 1 shows the estimated sampling densities of the estimators of Table 2 in the first two rows, in the third row displays results for N and M estimators for $\sigma = \{0.16, 0.25\}$, and in the fourth row illustrates the sampling densities for the estimates of θ_1 and θ_2 in multiple covariate cases where $H = 0.5$ and $\sigma = 0.25$.

Table 2 about here

Figure 1 about here

The analysis of Table 2 suggests that in all cases the naive estimators display considerable mean and median downward biases, always greater than 6.1%. 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, $41, 2\%$ and 47.1% in Mσ and 36.8% and 54.3% in M estimators. As for the standard deviations of both the MGMM estimators, as usual in estimators accounting for measurement error, they appear inflated when compared with those of the inconsistent naive estimators; see, for example, the simulation experiments in Chesher (1998) and Hausman, Abrevaya and Scott-Morton (1998). This occurs because the former estimators reflect the additional variability in the data induced by CME. However, notice that the inclusion of additional information on σ clearly attenuates this increase in the dispersion, since the variability of Mσ is always smaller than that of M estimators. Note also that in all cases 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.

On the other hand, the first two rows of Figure 1 show clearly that the sampling distributions of both Mσ and M estimators are always more centrally located around the true value of θ_1 than that of N estimators, which lies substantially beneath this value in all cases. This Figure also illustrates that the increase in the variance of the measurement error induces more bias in N estimators but produces only a small decay in the performance of the MGMM estimators, which are based on the use of small error variance approximations; see the third row. Finally, the sampling densities for the case of multiple covariates displayed in the last row of Figure 1 suggest that the consequences of CME are more severe on the parameter associated to the error-prone covariate, θ_1, and show that the bias-reduction ability of MGMM estimators is apparent in both the coefficients associated to error-prone and error-free covariates.

4.2 Loglog and probit models with CB sampling

In this framework, we assume that $Y|X$ is described by either a loglog or a probit model with no intercept, such that $\Pr_{Y|X}(1|x, \theta) = e^{-e^{-x\theta}}$ or $\Pr_{Y|X}(1|x, \theta) = \Phi(x\theta)$, respectively.[7] The generation of the contaminated covariate follows the designs previously coded as a and b (see Table 1) and, to obtain $Q = 0.9$, θ was set equal to 1.551 and 0.75 in loglog and probit models, respectively. Furthermore, we assume that σ is unknown to the researcher, the case which exhibited the worst results in the previous Monte Carlo experiments.

As the endogeneity of the sample has to be taken into account in loglog and probit models, we calculated the estimators for RB 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 Nq (N) and Mq (M). Thus, in these experiments, Q is the only 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 is

$$g_H^*(z^*) = H - y \tag{46}$$

$$g_\theta^*(z^*) = \frac{xp}{P(1-P)}\left\{y - b_1 b_X^{-1}\left[P + 0.5\sigma\theta^2 p\left(\frac{p^1}{p} + \frac{2}{\theta}l_X^1\right)b_0\right]\right\} + o(\Sigma) \tag{47}$$

$$g_Q^*(z^*) = Q - Pb_X^{-1}\left\{1 - 0.5\sigma\left[\theta^2 p(b_1 - b_0)\left(\frac{p^1}{p} + \frac{2}{\theta}l^1\right) + \left[l_X^2 + \left(l_X^1\right)^2\right]b_X\right]\right\} + o(\Sigma) \tag{48}$$

$$g_\sigma^*(z^*) = \frac{\frac{\theta^2 p(y-P)}{P(1-P)}\left(\frac{p^1}{p} + \frac{2}{\theta}l_X^1\right) + l_X^2 + \left(l_X^1\right)^2}{2 + \sigma\left[\frac{\theta^2 p(y-P)}{P(1-P)}\left(\frac{p^1}{p} + \frac{2}{\theta}l_X^1\right) + l_X^2 + \left(l_X^1\right)^2\right]} + o(\Sigma), \tag{49}$$

where $P = \Pr_{Y|X}(1|x^*, \theta)$, $p = \nabla_{x\theta}P$, $p^1 = \nabla_{x\theta}p$, $b_0 = b_S(s=0) = \frac{1-H}{1-Q}$, $b_1 = b_S(s=1) = \frac{H}{Q}$, and $b_X = b_X(x) = b_0 + P(b_1 - b_0)$. Note that to obtain Nq and N estimators only the moment indicators (46)-(48) need to be considered with $\sigma = 0$, which thus coincide with those of Imbens' (1992) simulation study concerning GMM estimators for CB samples. For both Nq and Mq estimators, estimation was performed with Q replaced by its known value in (46)-(49).

[7] 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.

Table 3 contains results for N, Nq, M, and Mq estimators for loglog models in cases where $\sigma = 0.16$. Figure 2 shows the estimated sampling densities of the Nq and Mq estimators contained in Table 3 and includes results for these estimators for $\sigma = 0.5$ in the first row, represents the estimates of σ obtained in MQ estimators for $\sigma = \{0.16, 0.5\}$ in the second row, and in the last two rows displays results for probit models for N, Nq, M, and Mq estimators in cases where $\sigma = 0.16$ and for N and M estimators for $\sigma = \{0.16, 0.25\}$.

The results of Table 3 suggest very different comments for the cases where Q is known and unknown. In the former case, Nq estimators present relatively small mean and median biases (the maximum bias is 3.3%), which, even so, were substantially reduced by our Mq estimators at a cost of a small increment in the dispersion. In the latter situation, the N estimator is seriously downward biased, presenting a minimum mean bias of 12.6%. M estimators eliminate part of this bias, which in the worst case (see experiment b), is reduced to approximately 40.3% in the mean and 53.0% in the median. Despite this improvement, note that these biases are often superior to those of the naive estimator which combines information on Q, Nq. Moreover, M estimators also exhibit very large standard deviations across the replications. Thus, they do not appear to be reliable for samples of the size considered here.

Table 3 about here

Figure 2 about here

The first row of Figure 2 confirms that the bias of Nq can be substantially increased as the variance of the measurement error grows. Our Mq estimators are more centrally located around the true value of θ_1, presenting a only slight decay in the performance, especially in terms of dispersion, when σ is increased. Note that this promissing performance is not affected by the large variability displayed by the estimates of σ, illustrated in second row of Figure 2. The conclusions for probit models are similar to those of the loglog, except that in these models the dispersion of M estimators relative to their naive version is not so pronounced. In fact, the performance of the M estimators in probit models is more similar to that observed for logit models, which is not surprising given the well known similarity of these two functional forms.

In these experiments, the benefits of including additional information concerning the marginal choice probabilities Q are also apparent. On the one hand, the naive estimators become clearly more robust to the presence of CME. On the other hand, the Mq estimator 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 Conclusion

In this paper we have proposed a general framework to deal with the presence of CME in RB 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 the GMM estimation framework of Ramalho and Ramalho (2006), which encompasses the most well known likelihood-based estimators for RB 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 bias-reduced estimators and outlined a score test sensitive to CME.

We found that the inconsistency of naive GMM estimators for RB sampling when the covariates' contamination is not acknowledged is a function of the approximate expectation of the naive moment indicators, 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, the use of the traditional GMM techniques based on them gives rise to bias-reduced 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 functions. Though these calculations are often complicated, as they involve derivatives of the structural model and 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 RB samples, it merely requires the presence of CME of the classical type. 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, producing gains in terms of bias-reduction and dispersion.

Monte Carlo evidence was presented which suggests that, in RB sampling designs of moderate size, 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.

6 Appendix

6.1 Calculation of the approximate bias functions

The approximate bias functions are calculated from the general formula $\sigma_{jk} E_{h_Z}\left[g_\gamma\left(z^*\right) m_{YX}^{jk}\left(y, x^*\right)\right]$. The bias of $g_\theta^1\left(z^*\right)$ and $g_\theta^2\left(z^*\right)$ involves, initially, the same calculations:

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

where $m_{YX}^{RS^{jk}}\left(y, x^*\right) = 0.5\left[l^{jk}\left(y|x^*\right)+l^j\left(y|x^*\right) l^k\left(y|x^*\right)+2 l^j\left(y|x^*\right) l^k\left(x^*\right)\right]$ is the simplified version of $m_{YX}^{jk}\left(y, x^*\right)$ for cases where the sampling is random and results from the suppression of the terms $l_X^{jk}\left(x^*\right)$ and $l_X^j\left(x^*\right) l_X^k\left(x^*\right)$. The suppression of those two terms in $m_{YX}^{jk}\left(y, x^*\right)$ exploits the fact that $\sum_{s \in \mathcal{S}} b_S(s) \int_{\mathcal{Y}_s} g_\gamma\left(z^*\right) f_{Y|X}\left(y|x^*, \theta\right) dy=0$. Then, $g_\theta\left(z^*\right)$ is replaced by $g_\theta^1\left(z^*\right)$ and $g_\theta^2\left(z^*\right)$ to obtain, respectively, $b_\theta^1=\sigma_{jk} E_{f_X}\left[\sum_{s \in \mathcal{S}} \int_{\mathcal{Y}_s} \nabla_\theta f_{Y|X}\left(y|x^*, \theta\right) m_{YX}^{RS^{jk}}\left(y, x^*\right)\right] dy$ and $b_\theta^2= \sigma_{jk} E_{f_X}\left\{\sum_{s \in \mathcal{S}} b_S(s) \int_{\mathcal{Y}_s} m_{YX}^{RS^{jk}}\left(y, x^*\right)\left[\nabla_\theta f_{Y|X}\left(y|x^*, \theta\right)-\ b_X\left(x^*\right)^{-1} f_{Y|X}\left(y|x^*, \theta\right)\ \nabla_\theta b_X\left(x^*\right)\right] dy\right\}$. On the other hand, the calculation of the bias of $g_Q\left(z^*\right)$ uses the fact that $E_{h_Z}\left[m_{YX}^{jk}\left(y, x^*\right)\right]=0$:

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

6.2 Calculation of the modified moment indicators

The corrected moment indicators (24)-(28) are obtained from (23) and employ the bias functions in (19)-(21) with $E_{f_X}[.]$ estimated by averages weighted by $b_X(x)^{-1}$:

$$g_\theta^1(z^*) = b_S(s)^{-1}\nabla_\theta \ln f_{Y|X}(y|x,\theta) - \sigma_{jk}b_X(x)^{-1}\sum_{s\in\mathcal{S}}\int_{\mathcal{Y}_s}\nabla_\theta f_{Y|X}(y|x^*,\theta)\, m_{YX}^{RS^{jk}}(y,x^*)\,dy,$$

$$\begin{aligned} g_\theta^2(z^*) &= \nabla_\theta \ln f_{Y|X}(y|x^*,\theta) - b_X(x^*)^{-1}\nabla_\theta b_X(x^*) - \sigma_{jk}b_X(x)^{-1}\left\{\sum_{s\in\mathcal{S}} b_S(s)\int_{\mathcal{Y}_s} m_{RS}^{jk}(y,x^*)\right. \\ &\quad \left.\left[\nabla_\theta f_{Y|X}(y|x^*,\theta) - b_X(x^*)^{-1} f_{Y|X}(y|x^*,\theta)\sum_{s\in\mathcal{S}} b_S(s)\nabla_\theta f_{Y|X}(y|x^*,\theta)\right]dy\right\} + o(\Sigma) \\ &= \nabla_\theta \ln f_{Y|X}(y|x^*,\theta) - b_X(x^*)^{-1}\left\{\nabla_\theta b_X(x^*) + \sigma_{jk}b_X(x)^{-1}\left\{\sum_{s\in\mathcal{S}} b_S(s)\int_{\mathcal{Y}_s} m_{RS}^{jk}(y,x^*)\right.\right. \\ &\quad \left.\left.\left[\nabla_\theta f_{Y|X}(y|x^*,\theta)\, b_X(x^*) - \nabla_\theta b_X(x^*)\, f_{Y|X}(y|x^*,\theta)\right]\right\}\right\} + o(\Sigma) \end{aligned}$$

and

$$\begin{aligned} g_{Q_t}(z^*) &= Q_t - b_X(x^*)^{-1}\int_{\mathcal{Y}_t} f_{Y|X}(y|x^*,\theta)\,dy + \\ &\quad \sigma_{jk}b_X(x^*)^{-2}\int_{\mathcal{Y}_t} f_{Y|X}(y|x^*,\theta)\,dy\sum_{s\in\mathcal{S}} b_S(s)\int_{\mathcal{Y}_s} f_{Y|X}(y|x^*,\theta)\, m_{YX}^{jk}(y,x^*)\,dy + o(\Sigma) \\ &= Q_t - b_X(x^*)^{-1}\int_{\mathcal{Y}_t} f_{Y|X}(y|x^*,\theta)\,dy\left[1 - \sigma_{jk}b_X(x)^{-1}\sum_{s\in\mathcal{S}} b_S(s)\int_{\mathcal{Y}_s} f_{Y|X}(y|x^*,\theta)\, m_{YX}^{jk}(y,x^*)\,dy\right. \\ &\quad + o(\Sigma). \end{aligned}$$

References

Barron, A., and C. Sheu (1991): "Approximation of density functions by sequences of exponential families," Annals of Statistics, 19, 1347-1369.

Carroll, R.J., Gail, M., and J. Lubin (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, Department of Economics, University of Bristol.

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

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

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

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

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

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

Chesher, A., and R.J. Smith (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, 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 and H. Vinod, eds., Handbook of Statistics *11*, Amesterdam: North-Holland, 1-43.

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

Dumangane, M. (2006): "Measurement error bias reduction in unemployment duration data," CEMMAP Working Paper 03/06, The Institute for Fiscal Studies, UCL.

Hall, A.R. (2005): Generalized Method of Moments. New York: Oxford University Press.

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

Hu, Y., and S.M. Schennach (2008): "Instrumental variable treatment of nonclassical measurement error models," Econometrica, 76, 195-216.

Hu, Y., and G. Ridder (2007): "Estimation of nonlinear models with measurement error using marginal information," Mimeo.

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

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

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

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

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

Newey, W.K., and D. McFadden (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., and R. Pyke (1979): "Logistic disease incidence and case-control studies," Biometrika, 66, 403-411.

Ramalho, E.A., and J.J.S. Ramalho (2006): "Bias-corrected moment-based estimators for parametric models under endogenous stratified sampling," Econometric Reviews, 25, 475-496.

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

Schennach, S.M. (2004a): "Estimation of nonlinear models with measurement error," Econometrica, 72, 33-75.

Schennach, S.M. (2004b): "Nonparametric estimation in the presence of measurement error," Econometric Theory, 20, 1046-1093.

Schennach, S.M. (2007): "Instrumental variable estimation of nonlinear errors-in-variables models," Econometrica, 75, 201-239.

Stefanski, L.A. (1985): "The effects of measurement error on parameter estimation," Biometrika, 72, 583-592.

Wang, C., Wang, S., and R. Carroll (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=σ)
a	Mixed normal	Normal
b	Mixed normal	Scaled χ_4^2
c	Scaled Student t(3)	Normal
d	Scaled Student t(3)	Scaled χ_4^2

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

Q=0.9, σ=0.16					
Experiment	H	Estimator	Bias		St. D.
			Mean	Median	
a	0.70	N	-0.092	-0.096	0.114
		Mσ	0.003	-0.004	0.137
		M	-0.003	-0.015	0.161
	0.50	N	-0.081	-0.088	0.110
		Mσ	0.006	0.003	0.123
		M	0.019	0.020	0.131
b	0.70	N	-0.087	-0.090	0.111
		Mσ	0.013	0.006	0.140
		M	0.003	-0.005	0.154
	0.50	N	-0.086	-0.092	0.109
		Mσ	0.002	-0.003	0.121
		M	0.018	0.018	0.136
c	0.70	N	-0.064	-0.067	0.099
		Mσ	-0.009	-0.013	0.117
		M	-0.023	-0.034	0.127
	0.50	N	-0.068	-0.070	0.100
		Mσ	-0.028	-0.033	0.111
		M	-0.025	-0.038	0.128
d	0.70	N	-0.061	-0.065	0.101
		Mσ	-0.002	-0.010	0.118
		M	-0.011	-0.029	0.135
	0.50	N	-0.070	-0.076	0.093
		Mσ	-0.029	-0.033	0.103
		M	-0.025	-0.034	0.122

Table 3: Loglog models with CB sampling - summary statistics for the slope parameter from 1000 replications

Q=0.9, σ=0.16					
Experiment	H	Estimator	Bias		St. D.
			Mean	Median	
a	0.70	N	-0.145	-0.151	0.149
		NQ	-0.024	-0.026	0.076
		M	0.038	-0.028	0.559
		MQ	-0.015	-0.020	0.083
	0.50	N	-0.126	-0.132	0.144
		NQ	-0.033	-0.033	0.064
		M	0.024	0.063	0.575
		MQ	-0.017	-0.018	0.071
b	0.70	N	-0.129	-0.142	0.150
		NQ	-0.016	-0.018	0.077
		M	0.052	-0.004	0.593
		MQ	-0.009	-0.013	0.081
	0.50	N	-0.128	-0.134	0.139
		NQ	-0.030	-0.032	0.067
		M	-0.036	-0.071	0.601
		MQ	-0.014	-0.015	0.072

Figure 1: Logit model − estimated sampling densities for slope parameters estimates

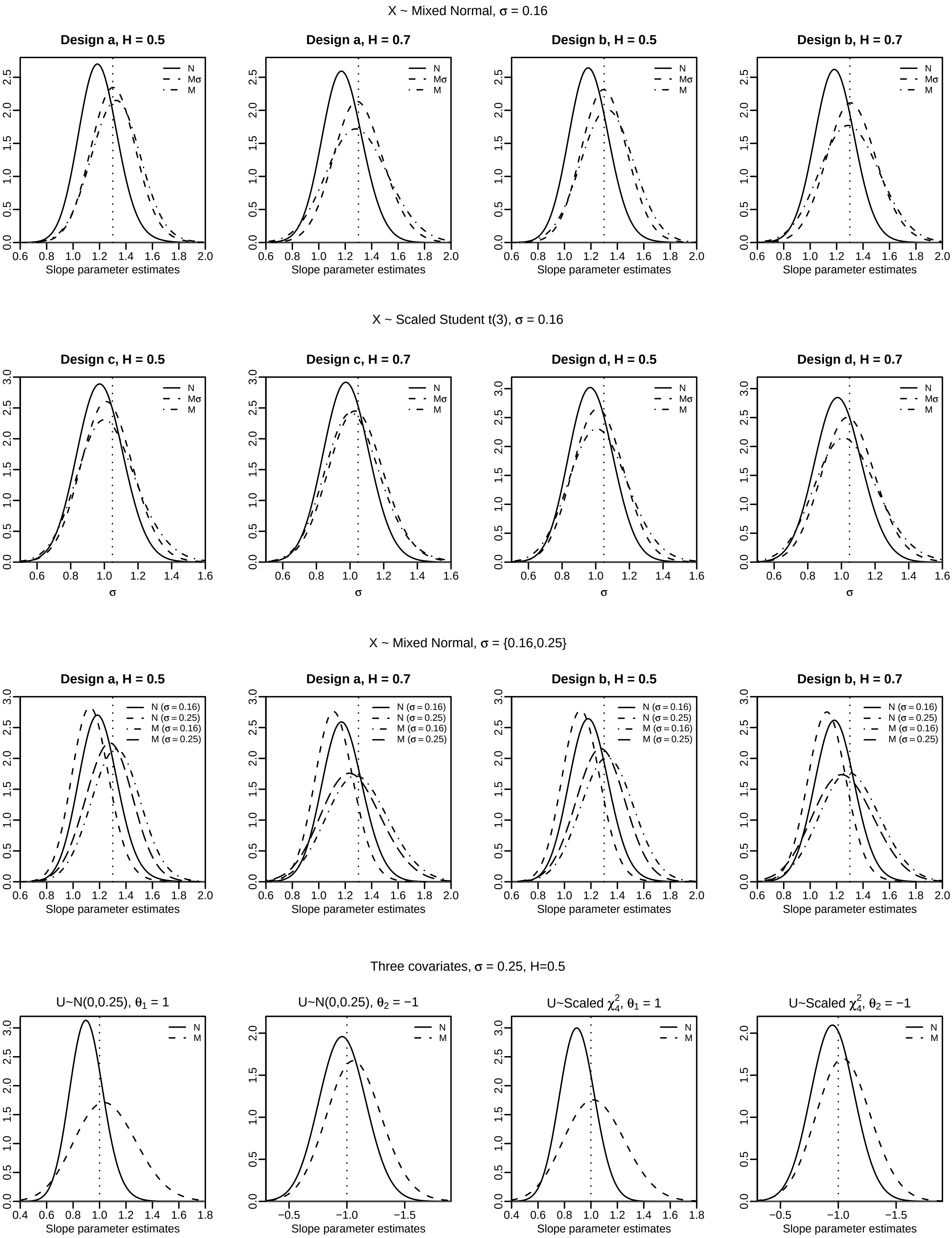

Figure 2: Loglog and probit models − estimated sampling densities