Statistica Sinica **21** (2011), 1639-1663
doi:http://dx.doi.org/10.5705/ss.2010.065

PSEUDO-LIKELIHOOD INFERENCE FOR REGRESSION MODELS WITH MISCLASSIFIED AND MISMEASURED VARIABLES

Annamaria Guolo

University of Verona

Abstract: This paper investigates the use of a pseudo-likelihood approach for inference in regression models with covariates affected by measurement errors. The maximum pseudo-likelihood estimator is obtained through a Monte Carlo expectation-maximization type algorithm and its asymptotic properties are described. The finite sample performance of the pseudo-likelihood approach is investigated through simulation studies, and compared to a full likelihood approach and to regression calibration under different measurement error structures, as well as continuous or discrete covariates. In contrast to the full likelihood approach, our method is computationally fast while remaining competitive from an inferential perspective. Satisfactory results are also provided over regression calibration. Pseudo-likelihood and the competing methods are finally applied to the analysis of two data sets.

Key words and phrases: Differential error, maximum likelihood estimator, measurement error, Monte Carlo expectation-maximization algorithm, nondifferential error, regression calibration.

1. Introduction

The problem of errors affecting the measure of variables is common in such areas of research as epidemiology, biology, health economics, and econometrics. Erroneous measures of a variable may be a consequence of instrument characteristics, self-reported data, or simply associated with costs of accurate observations. A large literature has been developed which emphasizes the impact of the mismeasurement on statistical analyses. The well-known consequence is the bias that may be induced on the parameter estimators (e.g., Armstrong (2003)). Many correction techniques have been proposed since the '80s to alleviate this problem. See Carroll et al. (2006) for a detailed review focused on mismeasured continuous variables, and Gustafson (2004) with emphasis on the misclassification of discrete variables.

The likelihood approach to correct for measurement error affecting covariates has the advantage over alternatives solutions of providing parameter estimators with optimality properties (e.g., Schafer (2002) and references therein). Nevertheless, the application of likelihood analysis is still limited in the literature

(Messer and Natarajan (2008)). The reason is ascribable to the complexity of the likelihood function, a feature which implies nonneglectable computational efforts for inferential purposes, as, for example, likelihood maximization.

To alleviate this problem, we explore an alternative approach that maintains the likelihood flavour and retains a high degree of efficiency while being computationally more convenient than maximum likelihood. Ours is a pseudo-likelihood approach that simplifies the likelihood as a function of the interest parameters only, following Gong and Samaniego (1981). Such an approach has not been fully explored in the measurement error context. In the paper, we illustrate how the pseudo-likelihood analysis can be applied in a broad range of problems, including mismeasured continuous and misclassified categorical covariates as well as accommodating different types of error structure. We maximize the pseudo-likelihood through a Monte Carlo expectation-maximization (MCEM) type algorithm. The algorithm makes use of an importance sampling procedure in the E-step to overcome many of the computational difficulties related to the specification of the conditional distribution. Moreover, we derive the asymptotic properties of the maximum pseudo-likelihood estimator by exploiting the results in Louis (1982). The problem of the sensitivity of the likelihood-based analysis to violations of the assumptions on the unobserved variables is also taken into account. We carry out an extensive simulation study in order to evaluate the performance of the pseudo-likelihood approach with respect to the full likelihood analysis, under different scenarios of interest. The likelihood and pseudo-likelihood approaches are also compared to regression calibration in case of continuous mismeasured covariates and to a modified version of regression calibration for misclassification problems.

The paper is structured as follows. The likelihood and the pseudo-likelihood approaches to measurement error correction are described in Section 2, while Section 3 is devoted to the illustration of the MCEM methodology for parameters estimation. The asymptotic distribution of the maximum pseudo-likelihood estimator is derived in Section 4. The simulation study is described in Section 5. Robustness issues with respect to violations of the model assumptions are discussed in Section 6. Section 7 is focused on the application of the correction methods to the analysis of two data examples, the first with two continuous mismeasured covariates, the second with a misclassified covariate derived from a dichotomization process. Discussion and final remarks are given in Section 8. Mathematical details about the pseudo-likelihood analysis and the distribution of the maximum pseudo-likelihood estimator are moved to two appendices.

2. Measurement Error Correction

2.1. Notation

Let Y be a response variable, either discrete or continuous, related to a set

of p covariates $\boldsymbol{X}$ through a parametric regression model, with density function $f_{Y|X}(y|\boldsymbol{x};\boldsymbol{\beta})$. The inferential interest is usually on the parameter vector $\boldsymbol{\beta}$. In place of $\boldsymbol{X}$, observations from a set of covariates $\boldsymbol{W}$ are available. The situation is known as the error-in-variables problem, also as the measurement error problem in case of continuous $\boldsymbol{X}$, or the misclassification problem in case of discrete $\boldsymbol{X}$. If a statistical analysis is performed ignoring the presence of measurement errors, then inference may be affected. The most relevant effect is the bias induced in the estimators of the parameters, and its severity may be more substantial in misclassification than in errors-in-variables problems (Gustafson and Le (2002)).

Suppose that the measurement error or the misclassification can be modeled by specifying the conditional distribution of $\boldsymbol{W}$ given $\boldsymbol{X}$ and Y. Let $f_{W|XY}(\boldsymbol{w}|\boldsymbol{x},y;\boldsymbol{\delta})$ be the density function of the corresponding model, depending on a set of parameters $\boldsymbol{\delta}$. In case of nondifferential error, the distribution of $\boldsymbol{W}$ is independent of Y, so that $f_{W|XY}(\boldsymbol{w}|\boldsymbol{x},y;\boldsymbol{\delta}) = f_{W|X}(\boldsymbol{w}|\boldsymbol{x};\boldsymbol{\delta})$. This situation is typical in case of instrumental measures of $\boldsymbol{X}$. Otherwise, the error is differential, which usually occurs in the case of self-reported questionnaires. A situation of particular interest has differential misclassification as a consequence of dichotomization of nondifferential mismeasured continuous covariates (Gustafson and Le (2002)). According to a similar view, the differential error implies the distribution of Y being dependent on $(\boldsymbol{X},\boldsymbol{W})$, while the dependence on $\boldsymbol{W}$ decays for nondifferential errors.

2.2. Likelihood and pseudo-likelihood approaches

Suppose that observations from Y and $\boldsymbol{W}$ are available for a sample of size n. Given the joint density function of $(Y,\boldsymbol{W},\boldsymbol{X})$, $f_{YWX}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x};\boldsymbol{\theta})$, depending on a set of parameters $\boldsymbol{\theta}$, the likelihood function for $\boldsymbol{\theta}$ is obtained by integrating out the unknown $\boldsymbol{X}$

$$L(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{w}) = \prod_{i=1}^{n} L(\boldsymbol{\theta};y_i,\boldsymbol{w}_i) = \prod_{i=1}^{n} \int f_{YWX}(y_i,\boldsymbol{w}_i,\boldsymbol{x}_i;\boldsymbol{\theta})d\boldsymbol{x}_i.$$

The integral is replaced by a sum in case of discrete $\boldsymbol{X}$. The likelihood function can be re-written as follows

$$L(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{w}) = \prod_{i=1}^{n} \int f_{Y|XW}(y_i|\boldsymbol{x}_i,\boldsymbol{w}_i;\boldsymbol{\beta}) f_{W|X}(\boldsymbol{w}_i|\boldsymbol{x}_i;\boldsymbol{\delta}) f_X(\boldsymbol{x}_i;\boldsymbol{\gamma})d\boldsymbol{x}_i, \qquad (2.1)$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathcal{T}},\boldsymbol{\delta}^{\mathcal{T}},\boldsymbol{\gamma}^{\mathcal{T}})^{\mathcal{T}}$, $f_{Y|XW}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w};\boldsymbol{\beta})$, $f_{W|X}(\boldsymbol{w}|\boldsymbol{x};\boldsymbol{\delta})$ have been defined above and $f_X(\boldsymbol{x};\boldsymbol{\gamma})$ is the density function of $\boldsymbol{X}$, depending on a set of parameters $\boldsymbol{\gamma}$. The likelihood specification allows for some additional error-free covariates $\boldsymbol{Z}$ by rewriting the density functions as $f_{Y|XWZ}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w},\boldsymbol{z};\boldsymbol{\beta})$, and so on.

Several studies in the literature have shown the advantages of the likelihood approach for measurement error or misclassification correction with respect to alternatives based on weaker assumptions. In particular, the advantages are substantial in terms of large sample optimality properties of the estimators (e.g., Schafer and Purdy (1996) and Küchenhoff and Carroll (1997)). Despite this, the likelihood approach has not had a considerable amount of application compared to alternative proposals. The main reason is the computational difficulty related to the likelihood evaluation and maximization, as the integral in (2.1) cannot usually be expressed in closed form. Here one usually relies on quadrature methods, at least for $\boldsymbol{X}$ of low dimension (Carroll et al. (2006, Sec. 8.3)). However, quadrature methods are less attractive because of the computational burden as well as the curse of dimensionality that can arise when increasing the dimension of $\boldsymbol{X}$.

In this paper we suggest alleviating the problem through an alternative likelihood-based estimation method. To this aim, we express the likelihood (2.1) as a function of the interest parameter $\boldsymbol{\beta}$ only. Let $\boldsymbol{\lambda} = (\boldsymbol{\delta}^{\mathcal{T}}, \boldsymbol{\gamma}^{\mathcal{T}})^{\mathcal{T}}$ denote the vector of nuisance parameters. When it is not feasible to eliminate $\boldsymbol{\lambda}$ through conditioning or factorization, the analysis can rely on a pseudo-likelihood approach, following Gong and Samaniego (1981). Thus, the likelihood maximization is carried out in two steps. In the first step, the nuisance parameter $\boldsymbol{\lambda}$ is conveniently estimated; then $\boldsymbol{\beta}$ is estimated by maximizing the pseudo-likelihood obtained with $\boldsymbol{\lambda}$ held fixed at the previous value. Let $\widehat{\boldsymbol{\lambda}} = (\widehat{\boldsymbol{\delta}}^{\mathcal{T}}, \widehat{\boldsymbol{\gamma}}^{\mathcal{T}})^{\mathcal{T}}$ denote the estimates of $\boldsymbol{\lambda}$ from the first step. Then the estimate of $\boldsymbol{\beta}$ maximizes the pseudo-likelihood

$$\begin{aligned} pL(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{w}, \widehat{\boldsymbol{\lambda}}) &= \prod_{i=1}^{n} pL(\boldsymbol{\beta}; y_i, \boldsymbol{w}_i, \widehat{\boldsymbol{\lambda}}) \\ &= \prod_{i=1}^{n} \int f_{Y|XW}(y_i|\boldsymbol{x}_i, \boldsymbol{w}_i; \boldsymbol{\beta}) f_{W|X}(\boldsymbol{w}_i|\boldsymbol{x}_i; \widehat{\boldsymbol{\delta}}) f_X(\boldsymbol{x}_i; \widehat{\boldsymbol{\gamma}}) d\boldsymbol{x}_i. \quad (2.2) \end{aligned}$$

The estimate of the nuisance parameter $\boldsymbol{\lambda}$ may be obtained from additional information, such as validation data. However a better solution, in terms of efficiency of the parameter estimators, is to maximize the reduced likelihood function

$$rL(\boldsymbol{\lambda}; \boldsymbol{w}) = \prod_{i=1}^{n} rL(\boldsymbol{\lambda}; \boldsymbol{w}_i) = \prod_{i=1}^{n} \int f_{W|X}(\boldsymbol{w}_i|\boldsymbol{x}_i; \boldsymbol{\delta}) f_X(\boldsymbol{x}_i; \boldsymbol{\gamma}) d\boldsymbol{x}_i, \quad (2.3)$$

thus exploiting the information included in *all* the data, and not only in a portion of the observations.

The properties of the resulting pseudo-likelihood $\boldsymbol{\beta}$ estimator go back to Gong and Samaniego (1981), who derived the asymptotic distribution of the estimator under some regularity conditions. In particular, the asymptotic variance-covariance matrix is provided to properly account for the uncertainty about the nuisance parameters estimation.

3. Monte Carlo EM Methodology

We propose a computationally convenient approach to maximizing the likelihood function (2.1) as well as the pseudo-likelihood function (2.2). If we consider $\boldsymbol{X}$ as missing data, then a MCEM strategy would provide a reasonable approach to inference. The computational approach described in the following section exploits an importance sampling technique within the E-step. The MCEM-type algorithm is first illustrated by referring to the likelihood analysis and then the details for applying it to the pseudo-likelihood analysis are provided.

3.1. Monte Carlo EM algorithm

Let $\boldsymbol{y}$ and $\boldsymbol{w}$ be the n-dimensional vectors of sample observations from Y and $\boldsymbol{W}$, and let $\ell(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{w}) = \log L(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{w})$ be the log-likelihood function derived from (2.1). Let also $\boldsymbol{x}$ be the n-dimensional vector of values from the unobserved $\boldsymbol{X}$. Following the idea underlying the EM algorithm, we take $\ell(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{w},\boldsymbol{x})$ to be the joint log-likelihood of the augmented data $(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x})$. Each iteration of the EM algorithm alternates an E-step and a M-step. Let $\boldsymbol{\theta}_r$ be the current value of $\boldsymbol{\theta}$ in the r-th iteration of the algorithm. The $(r+1)$-th E-step entails the calculation of

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_r) = E\left\{\ell(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{w},\boldsymbol{x})|\boldsymbol{y},\boldsymbol{w};\boldsymbol{\theta}_r\right\}, \tag{3.1}$$

where the expectation is with respect to the conditional distribution of the unobserved variable $\boldsymbol{X}$ given $(Y,\boldsymbol{W})$. In particular, we can decompose $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_r)$ as a sum of three components, the first being the log-likelihood related to the main model for the response, the second being the log-likelihood for the error model, and the third being the log-likelihood for the unobserved $\boldsymbol{X}$, namely

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_r) = E\left\{\ell(\boldsymbol{\theta};\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w})|\boldsymbol{y},\boldsymbol{w};\boldsymbol{\theta}_r\right\} + E\left\{\ell(\boldsymbol{\theta};\boldsymbol{w}|\boldsymbol{x})|\boldsymbol{y},\boldsymbol{w};\boldsymbol{\theta}_r\right\} + E\left\{\ell(\boldsymbol{\theta};\boldsymbol{x})|\boldsymbol{y},\boldsymbol{w};\boldsymbol{\theta}_r\right\}. \tag{3.2}$$

Then, the M-step performs the maximization of (3.1) with respect to $\boldsymbol{\theta}$, resulting in a new estimate $\boldsymbol{\theta}_{r+1}$. Given a starting point $\boldsymbol{\theta}_0$, the iteration between the E-step and the M-step is repeated until convergence.

Since the analytical evaluation of the expectation in (3.1) is usually not practical, we estimate it by means of a Monte Carlo approximation (Wei and Tanner (1990)). Let $f(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{w};\boldsymbol{\theta})$ be the density function of the model relating

$\boldsymbol{X}$ to $(Y, \boldsymbol{W})$. Suppose that M random samples $\boldsymbol{x}^*_{r,1}, \ldots, \boldsymbol{x}^*_{r,M}$, are simulated from $f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\theta}_r)$. Then, the Monte Carlo approximation of Q is

$$Q_m(\boldsymbol{\theta}|\boldsymbol{\theta}_r) = \frac{1}{M} \sum_{m=1}^{M} \left\{ \ell(\boldsymbol{\theta}; \boldsymbol{y}|\boldsymbol{x}^*_{r,m}, \boldsymbol{w}) + \ell(\boldsymbol{\theta}; \boldsymbol{w}|\boldsymbol{x}^*_{r,m}) + \ell(\boldsymbol{\theta}; \boldsymbol{x}^*_{r,m}) \right\}. \tag{3.3}$$

The specification of $f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\theta}_r)$ is usually difficult or even impractical in measurement error problems. We propose to overcome this difficulty by means of importance sampling. We can draw random samples of $\boldsymbol{X}$ from the density function $f(\boldsymbol{x}; \boldsymbol{\theta}_r)$ of $f(\boldsymbol{x}|\boldsymbol{w}; \boldsymbol{\theta}_r)$, which we assume has the same support as $f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\theta}_r)$. The importance density $f(\boldsymbol{x}|\boldsymbol{w}; \boldsymbol{\theta}_r)$ or $f(\boldsymbol{x}; \boldsymbol{\theta}_r)$ can be known, as for example from previous studies, or it can be estimated on validation data. Alternative flexible choices of the importance density are discussed in Section 6. The importance sampling Monte Carlo version of (3.3) is

$$Q_m(\boldsymbol{\theta}|\boldsymbol{\theta}_r) = \frac{1}{M} \sum_{m=1}^{M} k_{r,m} \left\{ \ell(\boldsymbol{\theta}; \boldsymbol{y}|\boldsymbol{x}^*_{r,m}, \boldsymbol{w}) + \ell(\boldsymbol{\theta}; \boldsymbol{w}|\boldsymbol{x}^*_{r,m}) + \ell(\boldsymbol{\theta}; \boldsymbol{x}^*_{r,m}) \right\}, \tag{3.4}$$

where $k_{r,m}$ are importance weights. If the importance density is $f(\boldsymbol{x}; \boldsymbol{\theta}_r)$, then $k_{r,m} = f(\boldsymbol{x}^*_{r,m}|\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\theta}_r)/f(\boldsymbol{x}^*_{r,m}; \boldsymbol{\theta}_r)$; else, if the importance density is $f(\boldsymbol{x}|\boldsymbol{w}; \boldsymbol{\theta}_r)$, then $k_{r,m} = f(\boldsymbol{x}^*_{r,m}|\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\theta}_r)/f(\boldsymbol{x}^*_{r,m}|\boldsymbol{w}; \boldsymbol{\theta}_r)$. The expression of the importance weights can be simplified. If we focus for example, on the importance density $f(\boldsymbol{x}|\boldsymbol{w}; \boldsymbol{\theta}_r)$, then

$$k_{r,m} = \frac{f(\boldsymbol{x}^*_{r,m}|\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\theta}_r)}{f(\boldsymbol{x}^*_{r,m}|\boldsymbol{w}; \boldsymbol{\theta}_r)} = \frac{f(\boldsymbol{y}|\boldsymbol{x}^*_{r,m}, \boldsymbol{w}; \boldsymbol{\theta}_r)}{f(\boldsymbol{y}|\boldsymbol{w}; \boldsymbol{\theta}_r)}.$$

The value of the weights can be approximated by exploiting the available Monte Carlo random samples $\boldsymbol{x}^*_{r,m}$, $m = 1, \ldots, M$,

$$k_{r,m} \approx \frac{f(\boldsymbol{y}|\boldsymbol{x}^*_{r,m}, \boldsymbol{w}; \boldsymbol{\theta}_r)}{M^{-1} \sum_{m=1}^{M} f(\boldsymbol{y}|\boldsymbol{x}^*_{r,m}, \boldsymbol{w}; \boldsymbol{\theta}_r)},$$

thus involving only the distribution of Y given $\boldsymbol{X}$, which is usually known. A similar expression can be obtained when using the importance weights from $f(\boldsymbol{x}; \boldsymbol{\theta}_r)$.

The same MCEM procedure is straightforwardly adaptable to maximize the logarithm of the pseudo-likelihood $p\ell(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{w}, \widehat{\boldsymbol{\lambda}}) = \log\ pL(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{w}, \widehat{\boldsymbol{\lambda}})$. Suppose that an estimate $\widehat{\boldsymbol{\lambda}}$ of the nuisance parameter $\boldsymbol{\lambda}$ is available. Then, given a starting point $\boldsymbol{\beta}_r$, the maximization of $p\ell(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{w}, \widehat{\boldsymbol{\lambda}})$ through an EM-type algorithm entails the maximization of

$$\begin{aligned} Q(\boldsymbol{\beta}|\boldsymbol{\beta}_r; \widehat{\boldsymbol{\lambda}}) = {} & E\left\{ p\ell(\boldsymbol{\beta}; \boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})|\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\beta}_r, \widehat{\boldsymbol{\lambda}} \right\} + E\left\{ p\ell(\boldsymbol{\lambda}; \boldsymbol{w}|\boldsymbol{x})|\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\beta}_r, \widehat{\boldsymbol{\lambda}} \right\} \\ & + E\left\{ p\ell(\boldsymbol{\lambda}; \boldsymbol{x})|\boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\beta}_r, \widehat{\boldsymbol{\lambda}} \right\}. \end{aligned} \tag{3.5}$$

The expectation in (3.5) can be simplified by removing the last two terms, since they do not depend on the interest parameter $\boldsymbol{\beta}$. It follows that the Monte Carlo approximation of (3.5) is

$$Q_m(\boldsymbol{\beta}|\boldsymbol{\beta}_r;\widehat{\boldsymbol{\lambda}}) = \frac{1}{M}\sum_{m=1}^{M} p\ell(\boldsymbol{\beta};\boldsymbol{y}|\boldsymbol{x}^*_{r,m},\boldsymbol{w},\widehat{\boldsymbol{\lambda}}),$$

where $\boldsymbol{x}^*_{r,1},\ldots,\boldsymbol{x}^*_{r,M}$, are M random samples from $f(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{w};\boldsymbol{\beta}_r,\widehat{\boldsymbol{\lambda}})$. As before, an importance sampling technique can be useful when simulating from $f(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{w};\boldsymbol{\beta}_r,\widehat{\boldsymbol{\lambda}})$ is not feasible. The maximization of $Q_m(\boldsymbol{\beta}|\boldsymbol{\beta}_r;\widehat{\boldsymbol{\lambda}})$ with respect to $\boldsymbol{\beta}$ can be performed by means of familiar routines available in standard softwares.

The maximization of the pseudo-likelihood (2.2) relies on assuming that an estimate $\widehat{\boldsymbol{\lambda}}$ of the nuisance parameter $\boldsymbol{\lambda}$ is available. A simple solution is to maximize the logarithm of the reduced likelihood $r\ell(\boldsymbol{\lambda};\boldsymbol{w}) = \log\ rL(\boldsymbol{\lambda};\boldsymbol{w})$. When it is not practical to express the integral in (2.3) in closed form, a MCEM approach can still be applied. In this case, given a starting point $\boldsymbol{\lambda}_r$, the maximization of $r\ell(\boldsymbol{\lambda})$ through an EM algorithm entails the maximization of

$$Q(\boldsymbol{\lambda}|\boldsymbol{\lambda}_r) = E\left\{r\ell(\boldsymbol{\lambda};\boldsymbol{w}|\boldsymbol{x})|\boldsymbol{w};\boldsymbol{\lambda}_r\right\} + E\left\{r\ell(\boldsymbol{\lambda};\boldsymbol{x})|\boldsymbol{w};\boldsymbol{\lambda}_r\right\},$$

or of its approximation

$$Q_m(\boldsymbol{\lambda}|\boldsymbol{\lambda}_r) = \frac{1}{M}\sum_{m=1}^{M} r\ell(\boldsymbol{\lambda};\boldsymbol{w}|\boldsymbol{x}^*_{r,m}),$$

where $\boldsymbol{x}^*_{r,1},\ldots,\boldsymbol{x}^*_{r,M}$, are M random samples from $f(\boldsymbol{x}|\boldsymbol{w};\boldsymbol{\lambda}_r)$. Again, an importance sampling technique can be useful when simulating from $f(\boldsymbol{x}|\boldsymbol{w};\boldsymbol{\lambda}_r)$ is not feasible.

The MCEM-type algorithm described above allows likelihood and pseudo-likelihood analysis to be suitable for problems with high-dimensional unobserved covariates $\boldsymbol{X}$, thus overcoming the curse of dimensionality affecting the quadrature approximations of integrals. Furthermore, both nondifferential and differential measurement error or misclassification can be taken into account. No restrictions are assumed on the error structure, and this is appealing because it allows broad applicability of the method.

3.2. Remarks

In case of discrete $\boldsymbol{X}$, the specification of the density $f(\boldsymbol{x}|\boldsymbol{w};\boldsymbol{\theta}_r)$ involves the reclassification probabilities $\mathrm{pr}(\boldsymbol{X}=\boldsymbol{x}|\boldsymbol{W}=\boldsymbol{w};\boldsymbol{\theta}_r)$, the marginal probabilities of $\boldsymbol{W}=\boldsymbol{w}$, $\mathrm{pr}(\boldsymbol{W}=\boldsymbol{w};\boldsymbol{\theta}_r)$, and of $\boldsymbol{X}=\boldsymbol{x}$, $\mathrm{pr}(\boldsymbol{X}=\boldsymbol{x};\boldsymbol{\theta}_r)$, so that

$$\mathrm{pr}(\boldsymbol{X}=\boldsymbol{x}|\boldsymbol{W}=\boldsymbol{w};\boldsymbol{\theta}_r) = \frac{\mathrm{pr}(\boldsymbol{W}=\boldsymbol{w}|\boldsymbol{X}=\boldsymbol{x};\boldsymbol{\theta}_r)\mathrm{pr}(\boldsymbol{X}=\boldsymbol{x};\boldsymbol{\theta}_r)}{\mathrm{pr}(\boldsymbol{W}=\boldsymbol{w};\boldsymbol{\theta}_r)}. \tag{3.6}$$

In case of univariate dichotomous X, the misclassification probability $\text{pr}(W = w|X = x;\boldsymbol{\theta}_r)$ required in (3.6) is related to the sensitivity $SN = \text{pr}(W = 1|X = 1;\boldsymbol{\theta}_r)$, and to the specificity $SP = \text{pr}(W = 0|X = 0;\boldsymbol{\theta}_r)$, which can be both known, for example, from the characteristics of the instruments used. Alternatively, SN and SP can be estimated from additional data. Similarly, the marginal distribution of W can be estimated from the data, while the marginal distribution of X, after some simple algebra, is

$$\text{pr}(X=0) = \frac{\text{pr}(W = w;\boldsymbol{\theta}_r)^w \left\{1-\text{pr}(W=w;\boldsymbol{\theta}_r)\right\}^{1-w} - SN^w(1-SN)^{1-w}}{(1-SP)^w SP^{1-w} - SN^w(1-SN)^{1-w}},$$

and $\text{pr}(X = 1) = 1 - \text{pr}(X = 0)$. A similar result holds in case of univariate categorical X, with r categories. Let $\boldsymbol{\Pi}= \Pi_{wx} = \text{pr}(W = w|X = x;\boldsymbol{\beta}_r)$ be the $r \times r$ misclassification matrix, $w \in \{1,\dots,r\}$, $x \in \{1,\dots,r\}$. Then, the marginal distribution of X can be derived as the solution of the system of linear equations

$$\begin{pmatrix} \text{pr}(W=1) \\ \text{pr}(W=2) \\ \vdots \\ \text{pr}(W=r) \end{pmatrix} = \begin{pmatrix} \Pi_{11} & \Pi_{12} & \cdots & \Pi_{1r} \\ \Pi_{21} & \Pi_{22} & \cdots & \Pi_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ \Pi_{r1} & \Pi_{r2} & \cdots & \Pi_{rr} \end{pmatrix} \cdot \begin{pmatrix} \text{pr}(X=1) \\ \text{pr}(X=2) \\ \vdots \\ \text{pr}(X=r) \end{pmatrix} = \boldsymbol{\Pi} \cdot \begin{pmatrix} \text{pr}(X=1) \\ \text{pr}(X=2) \\ \vdots \\ \text{pr}(X=r) \end{pmatrix}.$$

4. Asymptotics

4.1. Maximum likelihood estimator distribution

The asymptotic distribution of the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ can be approximated by a multivariate normal distribution with mean $\boldsymbol{\theta}$ and variance-covariance matrix $\boldsymbol{I}_\theta(\widehat{\boldsymbol{\theta}})^{-1}$, where $\boldsymbol{I}_\theta(\widehat{\boldsymbol{\theta}})$ is the observed information matrix for $\boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, namely

$$n^{-1/2}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{\mathcal{L}} MVN\left\{0, \boldsymbol{I}_\theta(\widehat{\boldsymbol{\theta}})^{-1}\right\},$$

with $\xrightarrow{\mathcal{L}}$ indicating the convergence in law. The quantity $I_\theta(\widehat{\boldsymbol{\theta}})$ is provided by the well-known result of Louis (1982) as the sum of

$$\boldsymbol{I}_1(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathcal{T}}} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) \quad \text{and} \quad \boldsymbol{I}_2(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = -var\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\ell(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{x}|\boldsymbol{w})|\boldsymbol{y};\widehat{\boldsymbol{\theta}}\right\},$$

both evaluated at $\widehat{\boldsymbol{\theta}}$. The matrix $\boldsymbol{I}_1(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ can be estimated by

$$-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathcal{T}}} Q_m(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}) = -\frac{1}{M}\sum_{i=1}^{n}\sum_{m=1}^{M} k_{m,i}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathcal{T}}}\ell(\boldsymbol{\theta};y_i,\boldsymbol{w}_i,\boldsymbol{x}^*_{m,i}),$$

with $\boldsymbol{x}^*_{m,i}, m = 1, \ldots, M$, being a random importance sample from the final iteration of the EM algorithm for the $i-$th individual, and $k_{m,i}$ being the corresponding importance weights. Similarly, an estimate of $\boldsymbol{I}_2(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})$ results from the sum of

$$\sum_{i=1}^{n} \left\{ \frac{1}{M} \sum_{m=1}^{M} k_{m,i} \, \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; y_i, \boldsymbol{w}_i, \boldsymbol{x}^*_{m,i}) \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \right\} \left\{ \frac{1}{M} \sum_{m=1}^{M} k_{m,i} \, \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; y_i, \boldsymbol{w}_i, \boldsymbol{x}^*_{m,i}) \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \right\}^{\mathcal{T}}$$

and

$$-\frac{1}{M} \sum_{i=1}^{n} \sum_{m=1}^{M} k_{m,i} \, \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; y_i, \boldsymbol{w}_i, \boldsymbol{x}^*_{m,i}) \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; y_i, \boldsymbol{w}_i, \boldsymbol{x}^*_{m,i}) \right\}^{\mathcal{T}} \Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}.$$

4.2. Maximum pseudo-likelihood estimator distribution

Following Gong and Samaniego (1981), the distribution of the maximum pseudo-likelihood estimator $\widehat{\boldsymbol{\beta}}$ of the interest parameter $\boldsymbol{\beta}$ can be still approximated by a multivariate normal distribution. However, in this case, although the mean is still equal to $\boldsymbol{\beta}$, the covariance matrix must be derived in order to take into account the uncertainty in estimating the nuisance parameter $\boldsymbol{\lambda}$ in the first step of the algorithm. For ease of notation, let $p\ell(\boldsymbol{\beta}; y_i, \boldsymbol{w}_i, \boldsymbol{\lambda}) = p\ell_i(\boldsymbol{\beta}; \boldsymbol{\lambda})$ and $r\ell(\boldsymbol{\lambda}; \boldsymbol{w}_i) = r\ell_i(\boldsymbol{\lambda})$. Then, it can be shown that, under some regularity conditions (Gong and Samaniego (1981), Liang and Self (1996)),

$$n^{-1/2} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right) \xrightarrow{\mathcal{L}} MVN\left(0, \boldsymbol{\Sigma}\right),$$

where the covariance matrix $\boldsymbol{\Sigma}$ can be expressed as

$$\boldsymbol{\Sigma} = \boldsymbol{I}^{-1}_{\beta\beta} \left(\boldsymbol{\Sigma}_{\beta\beta} - \boldsymbol{I}_{\beta\lambda} \boldsymbol{I}^{-1}_{\lambda\lambda} \boldsymbol{\Sigma}_{\lambda\beta} - \boldsymbol{\Sigma}^{\mathcal{T}}_{\lambda\beta} \boldsymbol{I}^{-1}_{\lambda\lambda} \boldsymbol{I}^{\mathcal{T}}_{\beta\lambda} + \boldsymbol{I}_{\beta\lambda} \boldsymbol{I}^{-1}_{\lambda\lambda} \boldsymbol{\Sigma}_{\lambda\lambda} \boldsymbol{I}^{-1}_{\lambda\lambda} \boldsymbol{I}^{\mathcal{T}}_{\beta\lambda} \right) \boldsymbol{I}^{-1}_{\beta\beta}. \quad (4.1)$$

See also Carroll et al. (2006, Sec. A.6.6). The sandwhich matrix $\boldsymbol{\Sigma}$ is composed of the term $\boldsymbol{I}_{\beta\beta}$, that is the covariance of the score function for $\boldsymbol{\beta}$ evaluated at the true value $(\boldsymbol{\beta}, \boldsymbol{\lambda})$, and the middle term, the covariance of the score function for $\boldsymbol{\beta}$ with $\boldsymbol{\lambda} = \widehat{\boldsymbol{\lambda}}$, accounting for the uncertainty of $\widehat{\boldsymbol{\lambda}}$ in estimating $\boldsymbol{\lambda}$. The

components of the information matrix $\boldsymbol{\Sigma}$ are

$$\boldsymbol{I}_{\beta\beta} = -\sum_{i=1}^{n} E\left\{\frac{\partial^2 p\ell_i(\boldsymbol{\beta};\boldsymbol{\lambda})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathcal{T}}}\right\}\Bigg|_{\boldsymbol{\lambda}=\widehat{\boldsymbol{\lambda}}}, \quad \boldsymbol{I}_{\lambda\lambda} = -\sum_{i=1}^{n} E\left\{\frac{\partial^2 r\ell_i(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}^{\mathcal{T}}}\right\},$$

$$\boldsymbol{I}_{\beta\lambda} = -\sum_{i=1}^{n} E\left\{\frac{\partial^2 p\ell_i(\boldsymbol{\beta};\boldsymbol{\lambda})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\lambda}^{\mathcal{T}}}\right\}\Bigg|_{\boldsymbol{\lambda}=\widehat{\boldsymbol{\lambda}}},$$

$$\boldsymbol{\Sigma}_{\beta\beta} = \sum_{i=1}^{n} \frac{\partial p\ell_i(\boldsymbol{\beta};\boldsymbol{\lambda})}{\partial\boldsymbol{\beta}}\Bigg|_{\boldsymbol{\lambda}=\widehat{\boldsymbol{\lambda}}} \left\{\frac{\partial p\ell_i(\boldsymbol{\beta};\boldsymbol{\lambda})}{\partial\boldsymbol{\beta}}\Bigg|_{\boldsymbol{\lambda}=\widehat{\boldsymbol{\lambda}}}\right\}^{\mathcal{T}},$$

$$\boldsymbol{\Sigma}_{\lambda\beta} = \sum_{i=1}^{n} \frac{\partial r\ell_i(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}} \left\{\frac{\partial p\ell_i(\boldsymbol{\beta};\boldsymbol{\lambda})}{\partial\boldsymbol{\beta}}\Bigg|_{\boldsymbol{\lambda}=\widehat{\boldsymbol{\lambda}}}\right\}^{\mathcal{T}}, \quad \boldsymbol{\Sigma}_{\lambda\lambda} = \sum_{i=1}^{n} \frac{\partial r\ell_i(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}} \left\{\frac{\partial r\ell_i(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}}\right\}^{\mathcal{T}}.$$

The matrices $\boldsymbol{I}_{\beta\beta}$ and $\boldsymbol{I}_{\lambda\lambda}$ can be computed by exploiting Louis (1982) formula and a similar approach provides an expression also for $\boldsymbol{I}_{\beta\lambda}$. Matrix $\boldsymbol{\Sigma}_{\beta\beta}$ can be approximated by

$$\sum_{i=1}^{n} \left\{\frac{1}{M}\sum_{m=1}^{M} k_{m,i}\, \frac{\partial}{\partial\boldsymbol{\beta}} p\ell(\boldsymbol{\beta}; y_i, \boldsymbol{w}_i, \boldsymbol{x}^*_{m,i}, \boldsymbol{\lambda})\Big|_{\boldsymbol{\lambda}=\widehat{\boldsymbol{\lambda}}}\right\} \left\{\frac{1}{M}\sum_{m=1}^{M} k_{m,i}\, \frac{\partial}{\partial\boldsymbol{\beta}} p\ell(\boldsymbol{\beta}; y_i, \boldsymbol{w}_i, \boldsymbol{x}^*_{m,i}, \boldsymbol{\lambda})\Big|_{\boldsymbol{\lambda}=\widehat{\boldsymbol{\lambda}}}\right\}^{\mathcal{T}},$$

where $\boldsymbol{x}^*_{m,i}$ is a random importance sample from the final iteration of the importance sampling EM algorithm with associated importance weights $k_{m,i}$. Matrices $\boldsymbol{\Sigma}_{\lambda\beta}$ and $\boldsymbol{\Sigma}_{\lambda\lambda}$ can be approximated similarly. Details are reported in Appendix A.1.

5. Simulation Studies

Extensive simulation studies have been performed with the aim of investigating the behavior of the likelihood and the pseudo-likelihood methods under different scenarios. The results are compared to those from the *naive* analysis and from regression calibration, which is one of the most common methods for measurement error correction, as described in the following section.

5.1. Regression calibration

Regression calibration (henceforth RC) (Rosner, Willett, and Spiegelman (1989)) is a frequently used method originally developed to correct for continuous mismeasured covariates $\boldsymbol{X}$. The method develops in two steps. In the calibration step, the unknown values of $\boldsymbol{X}$ are estimated by the conditional expectation of $\boldsymbol{X}$ given $\boldsymbol{W}$, that is $\boldsymbol{X}^*_{RC} = E(\boldsymbol{X}|\boldsymbol{W})$; in the regression step, a standard analysis is

performed with $\boldsymbol{X}$ replaced by $\boldsymbol{X}^*_{RC}$. Standard errors can be computed through resampling techniques, such as the bootstrap. The method yields consistent estimates of the parameters in linear regression but only approximately consistent estimates in nonlinear regression models (Carroll et al. (2006, Chap. 4)). A drawback of the method is that it is not suited to deal with differential errors. Moreover, aside from additive homoschedastic measurement error structures the results from RC may be misleading.

In the misclassification framework, an attempt at transposing the substitution idea underlying RC is due to Dalen et al. (2006). They propose to calibrate the values of $\boldsymbol{X}$, as in the first step of RC, and then to perform the regression step with $\boldsymbol{X}$ replaced by the categorized version of $\boldsymbol{X}^*_{RC}$. However, the simulation studies performed by Dalen et al. (2006) show that the method is not successful in improving the *naive* results, since high levels of bias are still retained in the estimators.

5.2. Implementation of the MCEM algorithm

We implement the MCEM algorithm described in the previous paragraph by using the `R` programming language (R Development Core Team (2009)), version 2.10.1. The convergence of the MCEM algorithm is speeded up by substituting the maximization of the $Q_m(\boldsymbol{\theta})$ function in the M-step with a one-step Newton-Raphson procedure. Let $\boldsymbol{\Delta}(\boldsymbol{\theta})$ and $\boldsymbol{H}(\boldsymbol{\theta})$ be, respectively, the gradient and the Hessian of $Q_m(\boldsymbol{\theta}|\boldsymbol{\theta}_r)$, both evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_r$. Their expression can be obtained by exploiting the results of Louis (1982), see Appendix A.1. Thus, according to the one-step Newton-Raphson method, the M-step provides the updated vector $\boldsymbol{\theta}_{r+1}$ as $\boldsymbol{\theta}_{r+1} = \boldsymbol{\theta}_r - \boldsymbol{H}(\boldsymbol{\theta}_r)^{-1}\boldsymbol{\Delta}(\boldsymbol{\theta}_r)^{\mathcal{T}}$.

A standard stopping rule for the deterministic EM algorithm is to stop when the relative difference between estimates in successive iterations is such that

$$\max_i \left(\frac{|\boldsymbol{\theta}_{r+1,i} - \boldsymbol{\theta}_{r,i}|}{|\boldsymbol{\theta}_{r,i} + \varepsilon_1|} \right) < \varepsilon_2, \tag{5.1}$$

where ε_1 and ε_2 are predetermined constants. We adopt the same stopping rule, choosing $\varepsilon_1 = 0.001$ and $\varepsilon_2 = 0.005$, as suggested by Booth and Hobert (1999). To reduce the risk of a premature stop, the algorithm is applied until the stopping rule (5.1) is satisfied for three consecutive iterations.

5.3 Simulation details

We focus on regression models with a continuous or a discrete response variable, namely a linear and a logistic regression model. Scenarios with continuous or categorical covariates are taken into account. Either differential or nondifferential errors are considered. The simulation settings are specified as follows.

(a) *Misclassification model.* We consider a logistic regression model $\text{logit}\{\text{pr}(Y = 1|X, Z)\} = \beta_0 + \beta_X X + \beta_Z Z$, with a discrete covariate X and an additional continuous covariate Z correlated with X. Let X be a binary covariate, with $\text{pr}(X = 1) = 0.8$. Values of Z are simulated from $\text{Normal}(0.5 - x, 1.0)$. The misclassification model is assumed to be nondifferential, with sensitivity SN and specificity SP. The parameters of interest are set equal to $(\beta_0, \beta_X, \beta_Z)^{\mathcal{T}} = (0.0, 1.0, 1.0)^{\mathcal{T}}$, while the misclassification parameters (SN, SP) take values in $\{(0.9, 0.8), (0.8, 0.8), (0.8, 0.7)\}$.

(b) *Multidimensional continuous* $\boldsymbol{X}$*.* We consider a logistic regression model $\text{logit}\{\text{pr}(Y = 1|X_1, X_2)\} = \beta_0 + \beta_{X_1} X_1 + \beta_{X_2} X_2$, with continuous correlated covariates X_1 and X_2. Value of X_1 and X_2 are simulated from $\text{Normal}(\mu_1, \sigma_1^2)$ and $\text{Normal}(\mu_2, \sigma_2^2)$, respectively, with correlation $\rho_{X_1X_2}$. The measurement error is assumed to be nondifferential, $W_1 = X_1 + \varepsilon$ and $W_2 = X_2 + \varepsilon$, with W_1 and W_2 independent of each other given (X_1, X_2), and ε being $\text{Normal}(0.0, \sigma_U^2)$. The parameters of interest are set equal to $(\beta_0, \beta_{X_1}, \beta_{X_2})^{\mathcal{T}} = (0.0, 1.0, 1.0)^{\mathcal{T}}$, the measurement error components are set equal to $(\sigma_U^2, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)^{\mathcal{T}} = (1.0, 0.0, 0.8, 0.0, 1.0)^{\mathcal{T}}$, and the correlation parameter $\rho_{X_1X_2}$ takes value in $\{0.0, 0.2, 0.5\}$.

(c) *Replicates of* X*.* We consider a logistic regression model $\text{logit}\{\text{pr}(Y = 1|X)\} = \beta_0 + \beta_X X$, with a continuous covariate X. Values of X are simulated from $\text{Normal}(\mu_X, \sigma_X^2)$. Two replicated nondifferential mismeasured versions of X are taken into account, $W_1 = X + \varepsilon_1$, $W_2 = X + \varepsilon_2$, where W_1 and W_2 independent given X, ε_1 is $\text{Normal}(0.0, \sigma_1^2)$, and ε_2 is $\text{Normal}(0.0, \sigma_2^2)$. The availability of multiple surrogates for one unobserved X is typical in occupational exposure studies (Weller et al. (2007)). The parameters of interest are set equal to $(\beta_0, \beta_X)^{\mathcal{T}} = (0.0, 1.0)^{\mathcal{T}}$, while those of the measurement error component are chosen equal to $(\sigma_1^2, \sigma_2^2, \mu_X, \sigma_X^2)^{\mathcal{T}} = (1.0, 0.7, 0.8, 1.0)^{\mathcal{T}}$.

(d) *Differential misclassification due to dichotomization.* We consider a linear regression model $Y = \beta_0 + \beta_V V + \epsilon$, with ϵ being a standard normal variable. The predictor of interest, V, is the dichotomization of the continuous X at threshold c, namely

$$V = \begin{cases} 1, & \text{if } X > c \\ 0, & \text{if } X \leq c \end{cases},$$

as often arises in medical studies. We consider values of X generated from $\text{Normal}(\mu_X, \sigma_X^2)$ and c assuming values in $\{-1.0, -0.5, 0.0\}$. Suppose that a nondifferential mismeasured version of X, W, is available together with its dichotomization, V^*, at threshold c. Gustafson and Le (2002), and previously Flegal, Keyl, and Nieto (1991), show that the misclassification induced by V^* is differential. According to their notation, we refer to this kind of misclassification as being differential due to dichotomization. In our simulation study,

we consider the nondifferential measurement error relating W to X being classical, $W = X + U$, with U being Normal$(0.0, \sigma_U^2)$. The parameters of interest are set equal to $(\beta_0, \beta_V)^{\mathcal{T}} = (0.0, 2.0)^{\mathcal{T}}$, while those of the measurement error components are chosen equal to $(\sigma_U^2, \mu_X, \sigma_X^2)^{\mathcal{T}} = (1.0, 0.0, 1.0)^{\mathcal{T}}$.

In each scenario, the inferential interest is on the parameter vector $\boldsymbol{\beta}$ of the model relating Y to $\boldsymbol{X}$ and to Z, if present. The parameter identifiability in each scenario is guaranteed, provided that additional information is available. In simulation settings a) and b), we suppose that an internal validation dataset of size equal to 10% of the main data is available (see Appendix A.2). Instead, in simulation setting c) the additional information is represented by the replicates of X, while in simulation setting d) by the threshold c being fixed. The distribution of $\boldsymbol{X}$ is supposed to be correctly specified for all the examined scenarios. The problem of possible misspecification of the model is discussed in Section 6. We performed 500 replications of the simulation experiment, each of them with a sample size of $n = 600$. As starting points for the MCEM procedure, we took the *naive* estimate for $\boldsymbol{\beta}$ and the estimate of $(\boldsymbol{\delta}, \boldsymbol{\gamma})$ provided by the method of moments on the observations from W. Standard errors of the maximum likelihood and pseudo-likelihood estimators were obtained by means of the procedure described in Section 4, while the standard error for the RC estimators derived from a bootstrap approach, with 100 bootstrap samples. The Monte Carlo sample size m was chosen equal to 500 in case of continuous response and to $m = 1{,}000$ in case of discrete response.

5.4. Simulation results

Results of the simulation studies performed under scenarios a)-d) are reported in Tables 1−4. They refer to the *naive* analysis (*Naive*), the regression calibration approach (*RC*), the likelihood (*Lik*), and the pseudo-likelihood analysis (*pLik*). In case of misclassified variables, regression calibration follows Dalen et al. (2006) (*Round-RC*).

Correction methods are compared in terms of bias of the estimators of the parameters (Bias) and associated standard errors (SE). The empirical standard error of the estimates is also reported (Sim-SE). Finally, the efficiency of the methods with respect to the likelihood approach is evaluated by computing the ratio of the mean squared error of the estimators to that of the maximum likelihood estimator (Relative MSE).

The simulation results highlight that the *naive* approach provides estimators, which are notably more biased than alternatives, usually underestimating the value of the parameters of interest. The poor performance of the method persists under all the examined scenarios, and it is worse for large measurement error.

Table 1. Bias, estimated standard error (SE), empirical standard error (Sim-SE), and relative mean squared error (Relative MSE) of the estimators of β_X and β_Z, for model a) in Section 5.3, with sensitivity SN and specificity SP. Parameter estimators: *naive* analysis ($Naive$), regression calibration in the version of Dalen et al. (2006) ($Round-RC$), likelihood analysis (Lik), pseudo-likelihood analysis ($pLik$).

	$\beta_X = 1.0$				$\beta_Z = 1.0$			
	Bias	SE	Sim-SE	Relative MSE	Bias	SE	Sim-SE	Relative MSE
				(SN, SP) = (0.9, 0.8)				
Naive	-0.45	0.24	0.23	0.49	-0.09	0.11	0.10	0.72
Round-RC	-0.45	0.24	0.25	0.49	-0.09	0.11	0.12	0.70
Lik	0.03	0.35	0.41	1.00	-0.01	0.12	0.13	1.00
pLik	0.01	0.36	0.39	0.96	0.01	0.12	0.13	0.97
				(SN, SP) = (0.8, 0.8)				
Naive	-0.62	0.21	0.20	0.38	-0.11	0.11	0.11	0.67
Round-RC	-0.62	0.21	0.21	0.38	-0.11	0.11	0.11	0.69
Lik	-0.01	0.40	0.44	1.00	0.01	0.13	0.14	1.00
pLik	-0.02	0.44	0.49	0.86	0.01	0.14	0.13	0.82
				(SN, SP) = (0.8, 0.7)				
Naive	-0.69	0.21	0.22	0.37	-0.12	0.11	0.11	0.65
Round-RC	-0.68	0.21	0.21	0.38	-0.12	0.11	0.11	0.69
Lik	0.04	0.43	0.52	1.00	0.02	0.13	0.14	1.00
pLik	-0.07	0.48	0.52	0.80	-0.01	0.14	0.14	0.83

See, for example, Table 1, where the bias increases with the reduction of the sensitivity and/or the specificity values. In the threshold model d), as Gustafson and Le (2002) point out, the bias is stronger when the threshold c is larger in magnitude. Despite a small variance with respect to other approaches, the substantial bias of the *naive* estimators produces a high relative MSE.

The RC approach in case of continuous mismeasured covariates follows a pattern similar to that of the *naive* analysis, although the consequences of measurement error are less pronounced. Estimators of the parameters of interest still retain some bias, as for example in Table 2. When focussing on discrete covariates, the modified version of RC, suggested by Dalen et al. (2006), is dramatically unsuccessful in correcting for misclassified X. The method maintains much of the misclassification of the data and provides results close to the *naive* ones, see Tables 1−4. This poor performance globally substantiates the findings by Dalen et al. (2006).

Conversely, relying on a likelihood approach provides noticeable advantages in correcting for measurement error or misclassification. Most of the measurement error and misclassification is detected, thus giving rise to estimators with a low bias, whatever scenario is of interest. The satisfactory performance is

Table 2. Bias, estimated standard error (SE), empirical standard error (Sim-SE), and relative mean squared error (Relative MSE) of the estimators of β_{X_1} and β_{X_2}, for model b) in Section 5.3, with correlation $\rho_{X_1X_2}$. Parameter estimators: *naive* analysis (*Naive*), regression calibration (*RC*), likelihood analysis (*Lik*), pseudo-likelihood analysis (*pLik*).

	$\beta_{X_1} = 1.0$				$\beta_{X_2} = 1.0$			
	Bias	SE	Sim-SE	Relative MSE	Bias	SE	Sim-SE	Relative MSE
				$\rho_{X_1X_2} = 0.0$				
Naive	-0.62	0.07	0.07	0.16	-0.58	0.07	0.07	0.16
RC	-0.13	0.31	0.27	0.56	-0.13	0.25	0.22	0.68
Lik	0.06	0.25	0.25	1.00	0.07	0.23	0.23	1.00
pLik	0.06	0.27	0.27	0.82	0.07	0.26	0.23	0.76
				$\rho_{X_1X_2} = 0.2$				
Naive	-0.59	0.08	0.08	0.20	-0.54	0.07	0.07	0.20
RC	-0.14	0.32	0.28	0.58	-0.14	0.26	0.25	0.67
Lik	0.06	0.26	0.27	1.00	0.07	0.23	0.25	1.00
pLik	0.06	0.26	0.28	0.98	0.07	0.29	0.25	0.69
				$\rho_{X_1X_2} = 0.5$				
Naive	-0.55	0.08	0.08	0.31	-0.50	0.08	0.08	0.30
RC	-0.15	0.44	0.36	0.43	-0.15	0.38	0.32	0.44
Lik	0.06	0.30	0.31	1.00	0.05	0.27	0.27	1.00
pLik	0.06	0.35	0.32	0.75	0.06	0.34	0.28	0.63

Table 3. Bias, estimated standard error (SE), empirical standard error (Sim-SE), and relative mean squared error (Relative MSE) of the estimators of β_0 and β_X, for model c) in Section 5.3. Parameter estimators: *naive* analysis (*Naive*), regression calibration (*RC*), likelihood analysis (*Lik*), pseudo-likelihood analysis (*pLik*).

	$\beta_0 = 0.0$				$\beta_X = 1.0$			
	Bias	SE	Sim-SE	Relative MSE	Bias	SE	Sim-SE	Relative MSE
Naive	0.20	0.11	0.11	0.03	-0.31	0.09	0.10	0.14
RC	-0.01	0.15	0.16	0.60	-0.05	0.15	0.15	0.77
Lik	-0.01	0.12	0.13	1.00	0.01	0.14	0.15	1.00
pLik	-0.01	0.14	0.13	0.67	0.02	0.15	0.14	0.78

maintained also under increasing measurement error and reducing sensitivity or specificity. See, for example, Table 1 and Table 4. Under all the examined error structures, there is a close agreement between the theoretically calculated standard errors and the simulated standard errors. Usually, they are both higher than the *naive* standard errors as a consequence of taking into account the measurement error variability.

Results remain satisfactory when applying the pseudo-likelihood approach. The bias of the estimators is low and close to that of the maximum likelihood estimators for all the examined scenarios. As in the likelihood approach, the

Table 4. Bias, estimated standard error (SE), empirical standard error (Sim-SE), and relative mean squared error (Relative MSE) of the estimators of β_V, for model d) in Section 5.3, with threshold c. Parameter estimators: *naive* analysis ($Naive$), regression calibration in the version of Dalen et al. (2006) ($Round - RC$), likelihood analysis (Lik), pseudo-likelihood analysis ($pLik$).

	$\beta_V = 2.0$			
	Bias	SE	Sim-SE	Relative MSE
		$c = -1.0$		
Naive	-1.24	0.12	0.13	0.02
Round-RC	-1.25	0.13	0.13	0.02
Lik	0.02	0.17	0.17	1.00
pLik	0.10	0.18	0.24	0.68
		$c = -0.5$		
Naive	-1.06	0.12	0.12	0.02
Round-RC	-1.06	0.12	0.12	0.02
Lik	-0.01	0.14	0.16	1.00
pLik	-0.02	0.14	0.14	0.95
		$c = 0.0$		
Naive	-1.01	0.11	0.11	0.02
Round-RC	-1.00	0.11	0.12	0.02
Lik	-0.03	0.13	0.14	1.00
pLik	-0.07	0.13	0.15	0.83

standard errors theoretically calculated as described in Section 4.2 are close to the simulated standard errors. As expected, the standard errors associated to the maximum pseudo-likelihood estimators are slightly higher than those of maximum likelihood estimators, as a consequence of the separate estimation of the nuisance parameters. Globally, the pseudo-likelihood approach provides estimators with high levels of relative MSE, as for example in Table 1, thus gaining high advantages in efficiency. In the meantime, the application of the method is computationally convenient. In particular, time consumption for pseudo-likelihood versus likelihood analysis is reduced about four times in models a) and c), and slightly less than three times times in models b) and d). The computational time is evaluated on the basis of ten independent replications of the simulation experiment.

6. Sensitivity to Model Assumptions

The construction of the likelihood or pseudo-likelihood function can be prone to model misspecification of the different components, with the subsequent risk of unreliable inferential results. The problem mainly affects the unobserved $\boldsymbol{X}$, since the lack of additional information or of knowledge about the phenomenon

precludes a proper specification of the model. Furthermore, as Guolo (2008) points out, the situation is exacerbated when handling case-control data, since the distribution of $\boldsymbol{X}$ in the case-control sampling scheme can notably differ from that at the population level.

The issue has been addressed in the literature by suggesting a flexible modeling of the distribution of $\boldsymbol{X}$. For example, Carroll, Roeder, and Wasserman (1999) and Richardson et al. (2002) focus on a mixture of normal variables. Alternatively, Guolo (2008) considers the skew-normal distribution (Azzalini (1985)) and shows that this distribution of $\boldsymbol{X}$ in the case-control sampling scheme results in likelihood estimation and inferences that are asymptotically correct, thus adding robustness to the approach.

Following Guolo (2008), we specify a skew-normal distribution for $\boldsymbol{X}$ both as a component of the pseudo-likelihood (2.2) specification, and as the importance density in the E-step of the MCEM-type algorithm (Section 3.1).

A simulation study has been carried out to evaluate the performance of the pseudo-likelihood approach and the competing methods to correct for mismeasured $\boldsymbol{X}$ when the distribution of $\boldsymbol{X}$ is modeled through the skew-normal. We focus for simplicity on a univariate $X \sim SN(\mu_X, \sigma_X, \alpha_X)$, with density function

$$f_X(x;\boldsymbol{\gamma}) = f_X(x;\mu_X,\sigma_X,\alpha_X) = \frac{2}{\sigma_X}\phi\left(\frac{x-\mu_X}{\sigma_X}\right)\Phi\left\{\frac{\alpha_X(x-\mu_X)}{\sigma_X}\right\}, \qquad (6.1)$$

where $\boldsymbol{\gamma} = (\mu_X, \sigma_X, \alpha_X)^{\mathcal{T}}$, μ_X, σ_X, α_X are the location, the scale, and the shape parameter, respectively, and $\phi(\cdot)$ and $\Phi(\cdot)$ represent the standard normal density and distribution functions. We consider a logistic regression model $\text{logit}\{\text{pr}(Y = 1|X)\} = \beta_0 + \beta_X X$, with continuous X, with values simulated from a *Weibull* distribution with shape and scale parameters equal to 2.0 and 0.5, respectively. The measurement error is assumed to be nondifferential and additive on the log-scale, that is, $\log W = \log X + \varepsilon$, with ε being Normal$(0, \sigma_\varepsilon^2)$. The parameters are set equal to $(\beta_0, \beta_X, \sigma_\varepsilon^2)^{\mathcal{T}} = (0.0, 0.1, 0.3^2)^{\mathcal{T}}$. We perform 500 replications of the simulation experiment, each of them with a sample size of $n = 600$. The parameter identifiability is guaranteed by an internal validation dataset of size equal to 10% of the main data (see Appendix A.2). As starting points for the MCEM procedure we take the *naive* estimate for $\boldsymbol{\beta}$ and the estimate σ_ϵ^2 provided by the method of moments based on the observations from W. An initial estimate of $\boldsymbol{\gamma}$ can be obtained by fitting the skew-normal distribution (6.1) on the observations of W. Simulation results are reported in Table 5.

Overall, simulation results recover the performance of the correction methods experienced in previous studies reported in Section 5. The presence of measurement error affects the *naive* estimator, as expected, by inducing a notable bias

Table 5. Bias, estimated standard error (SE), empirical standard error (Sim-SE), and relative mean squared error (Relative MSE) of the estimators of β_X, for the model described in Section 5.5. Parameter estimators: *naive* analysis ($Naive$), regression calibration (RC), likelihood analysis (Lik) under the actual specification of the distribution of X, and pseudo-likelihood ($pLik$) analysis under the skew-normal specification of the distribution of X.

	$\beta_X = 1.0$			
	Bias	SE	Sim-SE	Relative MSE
Naive	-0.30	0.35	0.32	0.94
RC	-0.67	0.16	0.16	0.42
Lik	0.02	0.45	0.42	1.00
pLik	0.05	0.45	0.44	0.97

of the estimator of β_X. In the meantime, a substantial bias affects also the RC estimator; the reason is that the method is well suited for classical additive measurement error, not for a multiplicative structure. As in the previous simulation studies, likelihood analysis is the preferable solution to correct for measurement error, with low levels of bias of the β_X estimator. Also the pseudo-likelihood approach experiences a satisfactory performance. In fact, the bias of the estimator of β_X is low and close to that of the maximum likelihood estimator and the relative MSE is very high, thus resulting in efficiency advantages. Resorting to the skew-normal distribution of X in pseudo-likelihood analysis is encouraging, since the approach provides results comparable to likelihood analysis based on the correct distribution of X with only a slight increase of standard errors. Time consumption for pseudo-likelihood versus likelihood analysis is reduced about two times, on the basis of ten independent replications of the simulation experiment.

7. Examples

7.1. Tonga trench earthquakes data

(Fuller, 1987, Sec. 3.1) listed the data about the depth and location of 43 earthquakes occurring near the Tonga trench in the Pacific Ocean near Fiji, between January 1965 and January 1966. Data are constituted by the depth Y of the earthquakes, the perpendicular distance X_1 from a line approximately parallel to the Tonga trench, and the distance X_2 from an arbitrary line perpendicular to the Tonga trench. All the variables are measured in hundred of kilometers. Following Sykes, Isacks, and Oliver, J. (1969), who previously analyzed the data, Fuller (1987) suggests the regression of the depth on the locations have the form $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2$, since the depths reasonably occur in a pattern that curves away from the earth's surface. Moreover, the measured distances of the earthquake, W_1 and W_2, are subject to an error with variance of about

100 kilometers squared. Schafer (2001) analyzes the same data according to a semiparametric likelihood approach in which the distribution of the unobserved Xs is left unspecified, then estimated by nonparametric maximum likelihood. According to his suggestion about the measurement error distribution, we take $W_i = X_i + \varepsilon$, $i = 1, 2$, with ε distributed as Normal$(0, 100)$, and also suppose that the measurement errors in the two directions are independent of one another, given the true location.

The semiparametric approach of Schafer leads to the fitted model $\widehat{E}(Y) = -20.4 + 0.51x_1 + 0.00124x_1^2 + 0.071x_2$, with standard error of the parameter estimators equal to $(13.2, 0.19, 0.00051, 0.041)^{\mathcal{T}}$, respectively. We analyze the data through the likelihood and the pseudo-likelihood approaches described in the previous sections, with the distribution of X_1 and X_2 specified according to a multivariate normal distribution. In both approaches, the correlation between X_1 and X_2 is found negligible. The maximum pseudo-likelihood estimator of $(\beta_0, \beta_1, \beta_2, \beta_3)^{\mathcal{T}}$ is $(-19.3, 0.505, 0.00121, 0.073)^{\mathcal{T}}$, with standard error $(7.6, 0.105, 0.00031, 0.031)^{\mathcal{T}}$. The maximum likelihood estimator is almost identical to the maximum pseudo-likelihood one, with about the same standard error. The Schafer (2001) results, as well as those obtained from our likelihood and pseudo-likelihood approaches, are close to the *naive* ones. This is not a surprise, since the measurement error variance is small when compared to the total variance of the measures, about 0.01 for each measure.

7.2. A cholesterol study

We consider data extracted from the Lipids Research Clinics study, on the risk of coronary heart disease as a function of blood cholesterol level. We focus on a portion of the data involving men aged $60 - 70$ who do not smoke, a total of 256 records. The case status Y occurs if a subject has had a heart attack, an abnormal exercise electrocardiogram, a history of angina pectoris, or the like. Covariates are low-density lipoprotein (LDL), cholesterol level, and total cholesterol level (TCL). TCL may be considered as a surrogate of LDL, whose direct measure is expensive and time consuming. Measurement error arises since TCL provides a measure of LDL plus unknown quantities of other components as triglycerides and high density lipoprotein. Thus, according to the notation used above, $X = LDL/100$ and $W = TCL/100$. We assume a nondifferential lognormal measurement error structure, that seems to be well supported by the data (Roeder, Carroll, and Lindsay (1996)). In examining the data, we are interested in modifying the predictor W in order to derive a discrete variable that is commonly adopted to discriminate between optimal and non-optimal total cholesterol levels. To this aim, we construct the predictor V^* by dichotomizing W with respect to the threshold 2 (200 in the TCL scale).

Using X as the covariate in the logistic regression model $\text{logit}\{\text{pr}(Y = 1)\} = \beta_0 + \beta_1 X$ provides an estimate of β_1 equal to 0.656, with a standard error of 0.336. The *naive* analysis with W substituting X, instead, provides an estimate $\beta_1 = 0.549$, with a standard error of 0.313. The *naive* analysis based on V^* provides an estimate of β_1 equal to 0.488, with standard error 0.275. The maximum likelihood estimate of β_1 is 0.649, with an associate standard error of 0.405, while the maximum pseudo-likelihood estimate is 0.651, with a standard error of 0.486.

8. Discussion and Final Remarks

In this paper we explored a pseudo-likelihood approach to correct for mismeasured covariates in regression models. The method has been proposed as an alternative to a full likelihood analysis, whose application can be cumbersome mainly because of computational difficulties. The pseudo-likelihood we focused on is a simplification of the likelihood function, expressed as a function of the interest parameters only, while the nuisance parameters related to the measurement error structure are fixed at pre-determined values. The asymptotic distribution of the maximum pseudo-likelihood estimator and, in particular, its asymptotic variance-covariance matrix are provided by exploiting the results of Gong and Samaniego (1981).

We illustrated a convenient approach for parameter estimation based on a MCEM-type algorithm. The procedure can accommodate both mismeasured continuous and misclassified categorical covariates with no restrictions on the measurement error structure. The algorithm is developed similarly for the likelihood and the pseudo-likelihood case, the latter in two steps. Extensive simulation studies show that the pseudo-likelihood approach provides satisfactory results with small bias comparable to that of the maximum likelihood estimator, while the price paid for avoiding the contemporary estimation of interest and nuisance parameters is a modest increase of the standard error. Advantages over the standard regression calibration are substantial. In particular, dramatic improvements are obtained in case of categorical misclassified covariates, with respect to regression calibration according to Dalen et al. (2006).

Given the difficulty of a correct specification of the distribution of the unobserved $\boldsymbol{X}$, we proposed a flexible modeling through the skew-normal family of distributions, resorting to the results in Guolo (2008). Simulation studies suggest that this specification within the pseudo-likelihood analysis leads to well-behaved inferential conclusions.

The MCEM-type algorithm takes advantage of an importance sampling procedure in the E-step to simulate from the target conditional distribution of the unobserved $\boldsymbol{X}$ given $(Y, \boldsymbol{W})$. Monte Carlo Markov Chain methods can be a

convenient alternative in case of high-dimensional problems or target densities of non-standard forms. The price to pay is a possibly slow convergence of the algorithm to the stationary distribution and a difficult estimation of the standard error of the parameter estimators (see Caffo, Jank, and Jones (2005)).

Care should be taken when using stochastic versions of the EM algorithm, since the method can be prone to some challenges. A question of major interest when applying a MCEM procedure is its convergence. In our study, we considered the application of a deterministic rule, namely (5.1), for three consecutive times in order to reduce the risk of a premature stop, as suggested by Booth and Hobert (1999). Actually, alternative criteria can be applied which are based on consecutive differences of the Q function (Caffo, Jank, and Jones (2005)) or on monitoring the likelihood gradient (Gu and Zhu (2001)). Furthermore, the MCEM algorithm is known not to guarantee the convergence to a global maximum; a common practice is to initialize the algorithm from different starting points, although the procedure can be cumbersome as the parameter dimension increases. The development of MCEM versions that overcome the problem of convergence to a global maximum is a topic of increasing interest in the literature, see Jank (2006a). We refer the reader to Jank (2006b) for a detailed review of challenges related to stochastic EM algorithms and recent possible solutions.

In this paper, we followed a frequentist approach to inference, although Bayesian analysis can also handle the models we focused on. We refer the reader to Carroll et al. (2006, Chap. 6) for a detailed illustration of Bayesian methods for measurement error problems, and to Gustafson (2004) for a Bayesian perspective on misclassification. Standard softwares, such as WinBUGS, are powerful instruments for Bayesian analysis. In the measurement error context, the implementation of WinBUGS routines is illustrated in Carroll et al. (2006, Sec. 9.8.4).

Acknowledgements

This work was partially supported by the Italian Ministry for Instruction, University and Research. The author greatly acknowledges Cristiano Varin for his valuable suggestions on a preprint version of the paper and an associate editor and two referees for helpful comments that led to improvements in the paper.

A. Appendix

A.1. Asymptotic variance-covariance matrix of the maximum pseudo-likelihood estimator

Consider the logarithm of the pseudo-likelihood for $\boldsymbol{\beta}$, $p\ell(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{\lambda})$, as a function of $\boldsymbol{\beta}$ with $\boldsymbol{\lambda}$ fixed at a predetermined estimate $\boldsymbol{\lambda} = \widehat{\boldsymbol{\lambda}}$. For simplicity, we write $p\ell(\boldsymbol{\beta}; \boldsymbol{y}, \boldsymbol{w}, \boldsymbol{\lambda}) = p\ell(\boldsymbol{\beta}; \boldsymbol{\lambda}) = \log \int f_{YWX}(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{\lambda}) d\boldsymbol{x}$. An estimate

of $\boldsymbol{\lambda}$ can be obtained by the reduced log-likelihood $r\ell(\boldsymbol{\lambda};\boldsymbol{w}) = r\ell(\boldsymbol{\lambda})$. Denote by ${}_{\beta}f'_{YWX}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\lambda})$ and ${}_{\beta}f''_{YWX}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\lambda})$ the first and second derivative of $f_{YWX}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\lambda})$ with respect to $\boldsymbol{\beta}$, respectively. A similar notation is used when deriving $f_{WX}(\boldsymbol{w},\boldsymbol{x};\boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$. Following Louis (1982), the gradient of $p\ell(\boldsymbol{\beta};\boldsymbol{\lambda})$ with respect to $\boldsymbol{\beta}$ is

$$\frac{\partial p\ell(\boldsymbol{\beta};\boldsymbol{\lambda})}{\partial \boldsymbol{\beta}} = \frac{\int {}_{\beta}f'_{YWX}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\lambda})d\boldsymbol{x}}{\int f_{YWX}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\lambda})d\boldsymbol{x}},$$

and, by multiplying and dividing the integrand in the numerator by $f_{YWX}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\lambda})$, we obtain

$$\begin{aligned}\frac{\partial p\ell(\boldsymbol{\beta};\boldsymbol{\lambda})}{\partial \boldsymbol{\beta}} &= \int \frac{\partial \log f_{YWX}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\lambda})}{\partial \boldsymbol{\beta}} f_{X|YW}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{w};\boldsymbol{\beta},\boldsymbol{\lambda})d\boldsymbol{x} \\ &= E\left\{ \frac{\partial \log f_{YWX}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\lambda})}{\partial \boldsymbol{\beta}} \middle| \boldsymbol{y},\boldsymbol{w};\boldsymbol{\beta},\boldsymbol{\lambda} \right\} = {}_1\boldsymbol{S}_{\beta}(\boldsymbol{\beta};\boldsymbol{\lambda}).\end{aligned}$$

The Hessian of $p\ell(\boldsymbol{\beta};\boldsymbol{\lambda})$ with respect to $\boldsymbol{\beta}$ is

$$\begin{aligned}&\frac{\partial^2 p\ell(\boldsymbol{\beta};\boldsymbol{\lambda})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathcal{T}}} \\ &= \frac{\int {}_{\beta}f''_{YWX}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\lambda})d\boldsymbol{x}}{\int f_{YWX}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\lambda})d\boldsymbol{x}} \\ &\quad - \frac{\int {}_{\beta}f'_{YWX}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\lambda})d\boldsymbol{x}\{\int {}_{\beta}f'_{YWX}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\lambda})d\boldsymbol{x}\}^{\mathcal{T}}}{\{\int f_{YWX}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\lambda})d\boldsymbol{x}\}^2} \\ &= \frac{\int {}_{\beta}f''_{YWX}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\lambda})d\boldsymbol{x}}{\int f_{YWX}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\lambda})d\boldsymbol{x}} - {}_1\boldsymbol{S}_{\beta}(\boldsymbol{\beta};\boldsymbol{\lambda}){}_1\boldsymbol{S}^{\mathcal{T}}_{\beta}(\boldsymbol{\beta};\boldsymbol{\lambda}) \\ &= E\left\{ \frac{{}_{\beta}f''_{YWX}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\lambda})}{f_{YWX}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\lambda})} \middle| \boldsymbol{y},\boldsymbol{w};\boldsymbol{\beta},\boldsymbol{\lambda} \right\} - {}_1\boldsymbol{S}_{\beta}(\boldsymbol{\beta};\boldsymbol{\lambda}){}_1\boldsymbol{S}^{\mathcal{T}}_{\beta}(\boldsymbol{\beta};\boldsymbol{\lambda}) \\ &= E\left\{ \frac{\partial^2 \log f_{YWX}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\lambda})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^{\mathcal{T}}} \middle| \boldsymbol{y},\boldsymbol{w};\boldsymbol{\beta},\boldsymbol{\lambda} \right\} \\ &\quad + E\left[\frac{\partial \log f_{YWX}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\lambda})}{\partial \boldsymbol{\beta}} \left\{ \frac{\partial \log f_{YWX}(\boldsymbol{y},\boldsymbol{w},\boldsymbol{x};\boldsymbol{\beta},\boldsymbol{\lambda})}{\partial \boldsymbol{\beta}} \right\}^{\mathcal{T}} \middle| \boldsymbol{y},\boldsymbol{w};\boldsymbol{\beta},\boldsymbol{\lambda} \right] \\ &\quad - {}_1\boldsymbol{S}_{\beta}(\boldsymbol{\beta};\boldsymbol{\lambda}){}_1\boldsymbol{S}^{\mathcal{T}}_{\beta}(\boldsymbol{\beta};\boldsymbol{\lambda}).\end{aligned}$$

The Monte Carlo estimate of the negative of $\partial^2 p\ell(\boldsymbol{\beta};\boldsymbol{\lambda})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\mathcal{T}}$ evaluated at $\boldsymbol{\lambda} = \widehat{\boldsymbol{\lambda}}$ is $\boldsymbol{I}_{\beta\beta}$, as provided in Section 4.2.

In the same way, consider $r\ell(\boldsymbol{\lambda}) = \log \int f_{WX}(\boldsymbol{w},\boldsymbol{x};\boldsymbol{\lambda})d\boldsymbol{x}$. The gradient of $r\ell(\boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$ is

$$\frac{\partial r\ell(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \frac{\int {}_{\lambda}f'_{WX}(\boldsymbol{w},\boldsymbol{x};\boldsymbol{\lambda})d\boldsymbol{x}}{\int f_{WX}(\boldsymbol{w},\boldsymbol{x};\boldsymbol{\lambda})d\boldsymbol{x}}.$$

By multiplying and dividing the integrand in the numerator by $f_{WX}(\boldsymbol{w}, \boldsymbol{x}; \boldsymbol{\lambda})$, we obtain

$$\frac{\partial r\ell(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = E\left\{ \frac{\partial \log f_{WX}(\boldsymbol{w}, \boldsymbol{x}; \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} \middle| \boldsymbol{w}; \boldsymbol{\lambda} \right\} = {}_1\boldsymbol{S}_\lambda(\boldsymbol{\lambda}).$$

Similarly,

$$\begin{aligned}\frac{\partial^2 r\ell(\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}^{\mathcal{T}}} &= E\left\{ \frac{\partial^2 \log f_{WX}(\boldsymbol{w}, \boldsymbol{x}; \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}^{\mathcal{T}}} \middle| \boldsymbol{w}; \boldsymbol{\lambda} \right\} \\ &\quad + E\left[\frac{\partial \log f_{WX}(\boldsymbol{w}, \boldsymbol{x}; \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} \left\{ \frac{\partial \log f_{WX}(\boldsymbol{w}, \boldsymbol{x}; \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} \right\}^{\mathcal{T}} \middle| \boldsymbol{w}; \boldsymbol{\lambda} \right] \\ &\quad - {}_1\boldsymbol{S}_\lambda(\boldsymbol{\lambda}) {}_1\boldsymbol{S}_\lambda^{\mathcal{T}}(\boldsymbol{\lambda}).\end{aligned}$$

The Monte Carlo estimate of the negative of $\partial^2 r\ell(\boldsymbol{\lambda})/\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}^{\mathcal{T}}$ is $\boldsymbol{I}_{\lambda\lambda}$, as provided in Section 4.2.

The quantity $\boldsymbol{I}_{\beta\lambda}$ is obtained from the negative of the Monte Carlo estimate of

$$\begin{aligned}&\frac{\partial^2 p\ell(\boldsymbol{\beta}; \boldsymbol{\lambda})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\lambda}^{\mathcal{T}}} \\ &= E\left\{ \frac{\partial^2 \log f_{YWX}(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\lambda}^{\mathcal{T}}} \middle| \boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\beta}, \boldsymbol{\lambda} \right\} \\ &\quad + E\left[\frac{\partial \log f_{YWX}(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\beta}} \left\{ \frac{\partial \log f_{YWX}(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} \right\}^{\mathcal{T}} \middle| \boldsymbol{y}, \boldsymbol{w}; \boldsymbol{\beta}, \boldsymbol{\lambda} \right] \\ &\quad - {}_1\boldsymbol{S}_\beta(\boldsymbol{\beta}; \boldsymbol{\lambda}) {}_1\boldsymbol{S}_\lambda^{\mathcal{T}}(\boldsymbol{\lambda}).\end{aligned}$$

The Monte Carlo estimate of $\partial p\ell(\boldsymbol{\beta}; \boldsymbol{\lambda})/\partial \boldsymbol{\beta}$, $\partial p\ell(\boldsymbol{\beta}; \boldsymbol{\lambda})/\partial \boldsymbol{\lambda}$, and $\partial r\ell(\boldsymbol{\lambda})/\partial \boldsymbol{\lambda}$ are necessary in the estimate of, respectively, $\boldsymbol{\Sigma}_{\beta\beta}$, $\boldsymbol{\Sigma}_{\lambda\beta}$, and $\boldsymbol{\Sigma}_{\lambda\lambda}$, as reported in Section 4.2.

A.2. Extra Information

In applications, additional information may be available about the mismeasured covariates $\boldsymbol{X}$. The additional information included in the inferential process ensures parameter identifiability and it is helpful in specifying a distribution for $\boldsymbol{X}$ and/or $\boldsymbol{X}|\boldsymbol{W}$. A common source of additional information is represented by validation data. Suppose that n observations are available for $(Y, \boldsymbol{W})$. If we focus on internal validation data of dimension n_1, values of $\boldsymbol{X}$ are also available, while they are not for the remaining $n_2 = n - n_1$ observations. Thus, the likelihood function (2.1) for $\boldsymbol{\theta}$ based on the observed $n_1 + n_2 = n$ data is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n_1} f_{YWX}(y_i, \boldsymbol{w}_i, \boldsymbol{x}_i; \boldsymbol{\theta}) \prod_{i=n_1+1}^{n} \int f_{YWX}(y_i, \boldsymbol{w}_i, \boldsymbol{x}_i; \boldsymbol{\theta}) d\boldsymbol{x}_i. \tag{A.1}$$

The maximization of the likelihood function (A.1) can exploit the MCEM strategy described in Section 3 only with respect to the second component involving the integral. The gradient and the Hessian of the likelihood function (A.1) can be obtained as the sum of the gradient and of the Hessian for the first component of the likelihood, provided by standard routines, and those for the second component derived starting from the Louis (1982) results.

References

Armstrong, B. (2003). Exposure measurement error: consequences and design issues. In *Exposure Assessment in Occupational and Environmental Epidemiology* (Edited by M. J. Nieuwenhuijsen). Oxford University Press, Oxford.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist.* **12**, 171-178.

Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. Roy. Statist. Soc. Ser. B* **61**, 265-285.

Caffo, B. S., Jank, W. S. and Jones, G. L. (2005). Ascent- based Monte Carlo EM. *J. Roy. Statist. Soc. Ser. B* **67**, 235-252.

Carroll, R. J., Roeder, K. and Wasserman, L. (1999). Flexible parametric measurement error models. *Biometrics* **55**, 44-54.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall, CRC Press, Boca Raton.

Dalen, I., Buonaccorsi, J. P., Laake, P., Hjartåker, A. and Thoresen, M. (2006). Regression analysis with categorized regression calibrated exposure: some interesting findings. *Emerging Themes in Epidemiology* **3**.

Flegal, K. M., Keyl, P. M., and Nieto, F. J. (1991). Differential misclassification arising from nondifferential errors in exposure measurement. *Amer. J. Epidemiology* **134**, 1233-1244.

Fuller, W. A. (1987). *Measurement Error Models*. Wiley, New York.

Gong, G. and Samaniego, F.J. (1981). Pseudo maximum likelihood estimation: theory and applications. *Ann. Statist.* **9**, 861-869.

Gu, M. G. and Zhu, H.-T. (2001). Maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation. *J. Roy. Statist. Soc. Ser. B* **63**, 339-355.

Guolo, A. (2008). A flexible approach to measurement error correction in case-control studies. *Biometrics* **64**, 1207-1214.

Gustafson, P. (2004). *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Chapman & Hall, CRC Press, Boca Raton.

Gustafson, P. and Le, N. D. (2002). Comparing the effects of continuous and discrete covariate mismeasurement, with emphasis on the dichotomization of mismeasured predictors. *Biometrics* **58**, 878-887.

Küchenhoff, H. and Carroll, R. J. (1997). Segmented regression with errors in predictors: semi-parametric and parametric methods. *Statist. Medicine* **16**, 169-188.

Jank, W. S. (2006a). Ascent EM for fast and global solutions to finite mixtures: An application to curve-clustering of online auctions. *Comput. Statist. Data Anal.* **51**, 747-761.

Jank, W. S. (2006b). The EM algorithm, its stochastic implementation and global optimization: some challenges and opportunities for OR. In *Topics in Modeling, Optimization, and Decision Technologies: Honoring Saul Gass' Contributions to Operations Research* (Edited by Alt, Fu, and Golden), 367-392. Springer Verlag, NY.

Liang, K-Y. and Self, S. G. (1996). On the asymptotic behaviour of the pseudolikelihood ratio test statistic. *J. Roy. Statist. Soc. Ser. B* **58**, 785-796.

Louis, T. A. (1982). Finding the observed information matrix using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44**, 226-233.

Messer, K. and Natarajan, L. (2008). Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment. *Statist. Medicine* **27**, 6332-6350.

R Development Core Team (2009). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available at `http://www.R-project.org`.

Richardson, S., Leblond, L., Jaussent, I. and Green, P.J. (2002). Mixture models in measurement error problems, with reference to epidemiological studies. *J. Roy. Statist. Soc. Ser. A* **165**, 549-566.

Roeder, K., Carroll, R. J. and Lindsay, B. G. (1996). A semiparametric mixture approach to case-control studies with errors in covariables. *J. Amer. Statist. Assoc.* **91**, 722-732.

Rosner, B., Willett, W. C. and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and condence intervals for systematic within-person measurement error. *Statist. Medicine* **8**, 1051-1070.

Schafer, D. W. (2001). Semiparametric maximum likelihood for measurement error model regression. *Biometrics* **57**, 53-61.

Schafer, D. W. (2002). Likelihood analysis and flexible structural modeling for measurement error model regression. *J. Statist. Comput. Simul.* **72**, 33-45.

Schafer, D. W. and Purdy, K. G. (1996). Likelihood analysis for errors-in-variables regression with replicate measurements. *Biometrika* **83**, 813-824.

Sykes, L. R., Isacks, B. L. and Oliver, J. (1969). Spatial distribution of deep and shallow earthquakes of small magnitudes in the Fiji-Tonga region. *Bull. Seismological Soc. Amer.* **59**, 1093-1113.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85**, 699-704.

Weller, E. A., Milton, D. K., Eisen, E. and Spiegelman, D. (2007). Regression calibration for logistic regression with multiple surrogates for one exposure. *J. Statist. Plann. Inference* **137**, 449-461.

Department of Economics, University of Verona, via dell'Artigliere, 19, I-37129 Verona, Italy.

E-mail: annamaria.guolo@univr.it

(Received March 2010; accepted August 2010)