Working Paper Series, N. 2, January 2007

Robust Techniques for Measurement Error Correction in Case-Control Studies: A Review

A. Guolo
Department of Statistical Sciences
University of Padua
Italy

Abstract: Measurement error affecting the independent variables in regression models is a common problem in many scientific areas. It is well known that the implications of ignoring measurement errors in inferential procedures may be substantial, often turning out in unreliable results. Many different measurement error correction techniques have been suggested in literature since the 80's. Most of them require many assumptions on the involved variables to be satisfied. However, it may be usually very hard to check whether these assumptions are satisfied, mainly because of the lack of information about the unobservable and mismeasured phenomenon. Thus, alternatives based on weaker assumptions on the variables may be preferable, in that they offer a gain in robustness of results. In this paper, we provide a review of robust techniques to correct for measurement errors affecting the covariates. Attention is paid to methods which share properties of robustness against misspecifications of relationships between variables. Techniques are grouped according to the kind of underlying modeling assumptions and inferential methods. Details about the techniques are given and their applicability is discussed. The basic framework is the epidemiological setting, where literature about the measurement error phenomenon is very substantial. The focus will be mainly on case-control studies.

Keywords: case-control study, empirical likelihood, estimating equation, kernel regression, logistic regression, measurement error, normal mixture, quasi-likelihood

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Annamaria Guolo
tel: +39 049 827 4192
guolo@stat.unipd.it
http://www.stat.unipd.it/~guolo

Robust Techniques for Measurement Error Correction in Case-Control Studies: A Review

A. Guolo

Department of Statistical Sciences
University of Padua
Italy

Abstract: Measurement error affecting the independent variables in regression models is a common problem in many scientific areas. It is well known that the implications of ignoring measurement errors in inferential procedures may be substantial, often turning out in unreliable results. Many different measurement error correction techniques have been suggested in literature since the 80's. Most of them require many assumptions on the involved variables to be satisfied. However, it may be usually very hard to check whether these assumptions are satisfied, mainly because of the lack of information about the unobservable and mismeasured phenomenon. Thus, alternatives based on weaker assumptions on the variables may be preferable, in that they offer a gain in robustness of results. In this paper, we provide a review of robust techniques to correct for measurement errors affecting the covariates. Attention is paid to methods which share properties of robustness against misspecifications of relationships between variables. Techniques are grouped according to the kind of underlying modeling assumptions and inferential methods. Details about the techniques are given and their applicability is discussed. The basic framework is the epidemiological setting, where literature about the measurement error phenomenon is very substantial. The focus will be mainly on case-control studies.

Keywords: case-control study, empirical likelihood, estimating equation, kernel regression, logistic regression, measurement error, normal mixture, quasi-likelihood

1 Introduction

Measurement error is a widely present problem in many scientific areas. In particular, it is a commonplace in observational studies, such as those carried out in environmental epidemiology (Zeger *et al.*, 2000). Erroneous measurements are due to different reasons, the most obvious being the inaccuracy of the instruments. Other examples include high costs of exact measures, the subjective nature of some variables, such as self-reported information and intrinsic biological variability. Measurement error is responsible for non-negligible inference problems if it is not corrected for (Armstrong, 2003). In particular, it has been long recognized that measurement error can bias the estimates. Further effects are unreliable coverage level of confidence intervals and reduced power of tests.

A large number of methods aiming to correct for measurement error have been proposed in literature since the 80's. They differ according to the assumptions about the distribution of the unobserved variable, to the availability of additional data about the unobserved variable and to the theoretical background of the approach, which may be parametric or nonparametric. A detailed review is Carroll *et al.* (2006). Previously, a review of measurement error correction techniques in case-control studies, when extra information is available, has been proposed by Thürigen *et al.* (2000). The review of techniques we provide here differs from the one by Thürigen *et al.* (2000) in that the focus is on methods which share the property of being robust against misspecifications of the relationships between variables. Most of these techniques have been proposed in literature during the last few years and a comprehensive overview of them is not available yet, to the best of our knowledge. The performance of these techniques in correcting for measurement errors has not

been deeply investigated in applications, although situations where the availability of robust methods would be preferable arise very often. The most common situation is avoiding estimators of parameters to be inconsistent, as it may happen when the assumptions underlying nonrobust methods are not satisfied, at least approximately. To stimulate the use and the development of robust techniques to correct for measurement error affecting the covariates, we provide a review of the methods, through a classification made up on their underlying theory. We do not consider results about robustness against leverage points or outliers, which both are rare in this literature. We mainly refer to the epidemiological setting and to case-control studies.

The paper is organized as follows. In Section 2 we define the framework which we focus on and the corresponding notation we will adopt thereafter. Robust measurement error correction techniques are described in Section 3, following a classification into groups which share a similar theoretical approach. A discussion about the applicability of the methods is given in Section 4.

2 Notation

Suppose that case-control data are available. Let Y be the response variable. In the case-control setting we focus on, this is the case-control status, or the disease status, indicator. Let X be the covariates which may be not directly observed. In epidemiological studies, they typically represent risk factors contributing to the presence of the disease. Instead of X, the mismeasured variables W are observed. These are usually called *proxy* variables. It may be assumed that other variables, Z, can be measured without error.

In measurement error literature, we distinguish different models relating the variables. The model relating the variable Y to the unobserved variables

X and the error-free variables Z is referred to as the *disease model*. Its density is indicated by $f_{Y|XZ}(y|x, z; \beta)$. In case-control studies this model is typically the logistic regression model. The interest usually focuses on the vector of parameters β, which is the vector of relative risks associated with a unit change in the exposure to the risk factors X.

The measurement error process is specified by modelling the relationship between X and W, possibly depending on Z. It is called *measurement error model*. The simplest measurement error model is the classical error model $W = X + U$, where U has mean zero and variance equal to σ_U^2 and is independent of X. The classical measurement error model is an unbiased and additive error model, such that $E[W|X] = X$. An alternative model is the *Berkson error model*, which typically arises in laboratory studies and experimental situations in which the observed variable is controlled for. The model relates X and W as $X = W + U$, where U has mean zero and variance equal to σ_U^2 and is independent of W. In the Berkson model $E[X|W] = W$ and W is said to be an unbiased predictor of X.

Different types of measurement error can arise in practice. An important distinction is made between differential and nondifferential measurement errors. The error in W is *nondifferential* if no additional information on Y is contained in (W, X, Z) with respect to (X, Z). This means that the conditional distribution of Y given (W, X, Z), $f_{Y|WXZ}(y|w, x, z; \beta)$, is the same than the distribution of Y given (X, Z), $f_{Y|XZ}(y|x, z; \beta)$, that is, $f_{Y|WXZ}(y|w, x, z; \beta) = f_{Y|XZ}(y|x, z; \beta)$. In this case, W is said to be a surrogate for X. When, instead, $f_{Y|WXZ}(y|w, x, z; \beta) \neq f_{Y|XZ}(y|x, z; \beta)$, the error is said to be *differential*.

In applications, many different error sources can be encountered. This im-

plies that both nondifferential and differential errors, with classical or Berkson components, can be defined. An accurate specification of the error model, distinguishing between differential and nondifferential errors with classical or Berkson components, is crucial because of the different impacts of the errors on the inferential results and the different available correction techniques. Therefore, a good identification of the error model is important for the successful application of measurement error correction techniques (Heid *et al.*, 2004).

These techniques can be roughly classified into two groups, according to their interpretation of the unobserved variables X. We define a method to be *functional* if it makes no assumption on the unobserved variables X, that is, they are modeled as unknown, nonrandom constants (parameters). On the contrary, we define a method to be *structural* if it considers the X's to be random variables. In this case, the specification of the distribution for the X's is needed, possibly depending on Z. This gives rise to the *exposure model*, whose density is indicated by $f_{X|Z}(x|z;\delta)$.

The simplest way to correct for measurement error is by adopting the so-called regression calibration (RC, for short) method (Rosner *et al.*, 1989, 1990; Carroll and Stefanski, 1990; Gleser, 1990). This is the most commonly adopted method to correct for measurement error in covariates, mainly because of the simplicity of its applicability with existing softwares. The idea underlying the method is the estimation of the regression of X on W and, possibly, Z on additional data, that is, further data than the main study sample. Additional information can be available in different forms. For example, a subsample of observations from X can be recorded for a small group of subjects of the main study sample. It originates the internal validation data set, from which the so-called gold standard measures of X are available. A common alternative

is collecting replication data, which are replicates of the observations from X. They can be obtained by the same process which provides observations from W.

According to the idea underlying RC, the resulting predictions of X obtained by the regression of X on (W, Z) in the additional data set are then substituted to the unknown values of X in the disease model. After that, standard analysis can be run. RC often leads to consistent or approximately consistent estimators of the parameter of interest. However, it requires some hypotheses to be satisfied, first of all that a linear homoscedastic relationship between X and W and, possibly, Z, holds. If this is not the case, RC results could be quite misleading.

Thus, alternative techniques to correct for measurement error may be preferable. An example is given by likelihood-based correction techniques, which have the advantage of ensuring good properties of the corresponding estimators, as, for example, efficiency and optimality, although at the notable price of a bigger computational burden. The application of likelihood techniques requires the parametric specification of the distribution for the unobserved variable X, that is, the exposure model, together with the specification of the disease model and of the measurement error model previously defined.

Let a classical structure for measurement error hold and let $f_{W|XZ}(w|x, z; \gamma)$ be the density associated with this model. If n_1 is the number of subjects on which observations (y_i, w_i, z_i), $i = 1, \ldots, n_1$, from the variables (Y, W, Z) are recorded, the likelihood is given by integrating over the true and unobserved X

$$L(\theta; y, w, z) = \prod_{i=1}^{n_1} \int f_{Y|XZ}(y_i|x_i, z_i; \beta) f_{W|XZ}(w_i|x_i, z_i; \gamma) f_{X|Z}(x_i|z_i; \delta) dx_i, \quad (1)$$

where $\theta = (\beta, \gamma, \delta)^{\mathrm{T}}$. If the Berkson error model holds in place of the classical

one, then the likelihood function for θ is given by

$$L(\theta; y, w, z) = \prod_{i=1}^{n_1} \int f_{Y|XZ}(y_i|x_i, z_i; \beta) f_{X|WZ}(x_i|w_i, z_i; \gamma) f_{W|Z}(w_i|z_i; \delta) dx_i, \quad (2)$$

which can be simplified to

$$L(\theta; y, w, z) = \prod_{i=1}^{n_1} \int f_{Y|XZ}(y_i|x_i, z_i; \beta) f_{X|WZ}(x_i|w_i, z_i; \gamma) dx_i, \quad (3)$$

if we consider that $f_{W|Z}(w|z; \delta)$ carries no information about the interest parameter β and does not depend on X. The integrals in (1) and (3) are replaced by a sum if X is a discrete random variable.

Often additional information about the measurement error distribution is necessary for parameters in (1) and (3) to be identifiable. Such additional information may be in the form of validation data or replicates. Suppose that internal validation data are available. Let n_2 be the dimension of the internal validation data set, in which we observe (y_i, x_i, z_i), $i = 1, \ldots, n_2$, from (Y, X, Z). To take account of this, the likelihood in (1) is re-expressed as follows

$$\begin{aligned} L(\theta; y, w, z) \quad = \quad & \prod_{i=1}^{n_1} \int f_{Y|XZ}(y_i|x_i, z_i; \beta) f_{W|XZ}(w_i|x_i, z_i; \gamma) f_{X|Z}(x_i|z_i; \delta) dx_i \\ & \prod_{i=1}^{n_2} f_{Y|XZ}(y_i|x_i, z_i; \beta) f_{W|XZ}(w_i|x_i, z_i; \gamma) f_{X|Z}(x_i|z_i; \delta), \end{aligned}$$

while the one in (3) is re-expressed as follows

$$L(\theta; y, w, z) = \prod_{i=1}^{n_1} \int f_{Y|XZ}(y_i|x_i, z_i; \beta) f_{X|WZ}(x_i|w_i, z_i; \gamma) dx_i \prod_{i=1}^{n_2} f_{Y|XZ}(y_i|x_i, z_i; \beta) f_{X|WZ}(x_i|w_i, z_i;$$

Similar modifications of the likelihood are defined to take account of other additional data as, for example, external validation data or replicates(Higdon and Schafer, 2001),(Schafer, 2002).

3 Robust techniques

As outlined in Section 2, a parametric approach requires the specification of some models for all the involved variables. In particular, the likelihood-based approach requires the exposure model to be specified, which is often difficult because of the lack of observations from X. This implies that issues of model misspecification naturally arise. It is well known that model misspecification can result in inconsistent estimators of the model parameters (Carroll *et al.*, 1998). Recently, Huang *et al.* (2006) suggest methods for diagnosing the effects of model misspecification of the exposure distribution, by checking both formally and empirically robustness properties. Alternatives to parametric modeling which retain good properties of efficiency with respect to parametric inference while reducing sensitivity to modeling assumptions on the variables may be preferable. Examples are flexible-parametric modeling and semiparametric modeling, which are illustrated in Section 3.1 and Section 3.2. Moreover, other solutions are provided by different techniques. We classified them in quasi-likelihood, estimating equations and empirical likelihood. Details about these techniques are given, respectevely, in Section 3.3, Section 3.4 and Section 3.5. Robust techniques which cannot be included in one of the previous groups are illustrated in Section 3.6.

3.1 Flexible-parametric modeling methods

The use of a parametric model with a high flexibility in defining some components of the problem, such as, for example, the exposure model, has the advantage of being easily defined and making inference retaining a high degree of efficiency if compared to parametric inference. The method is suggested by Carroll *et al.* (1999b). These Authors propose to use a mixture of normal

distributions as a flexible specification for a component of the problem. In particular, they focus on linear models and change-point Berkson models, with nondifferential errors and use a mixture of normal distributions to model the unobservable covariate X and the measurement error, respectively. The mixture distribution is incorporated into the likelihood function, thus summarizing data contribution for inferential procedures performed through a frequentist or a Bayesian approach. A Bayesian approach is adopted to obtain point estimates and confidence intervals for all parameters of interest, using Markov chain Monte Carlo (MCMC) for simulating from the posterior distribution of the parameters. The number of components in the normal mixture, indicated by k, is also considered an unknown parameter. According to Carroll *et al.* (1999b), it can be estimated like the other parameters or it can be chosen through a sensitivity analysis, by evaluating how inferential results vary as a function of k. The first solution is adopted in the linear model, while the second is used in the change-point Berkson model. Simulation studies are performed to compare the behaviour of the likelihood based on the mixture of normals to the method of moments and the likelihood based on the normal distribution, in terms of properties of the resulting estimators. Several sampling distributions for the unobservable covariate X are assumed, as, for example, the log χ^2 distribution, the normal distribution and the skew normal distribution. Results indicate that the mixture method can outperform the one based on the normal distribution in terms of bias of the estimators, except in situations where the distribution of the unobservable covariate is highly skewed, as, for example, when a $\log \chi^2$ distribution is assumed. As expected, the method of moments is the less satisfactory solution for a large class of the assumed distributions, both in terms of bias and variance of the estimators.

As Carroll *et al.* (1999b), also Carroll *et al.* (1999a) use a mixture of normal distributions to model the exposure, with the aim of increasing robustness to model misspecification. The difference is that the proposal by Carroll *et al.* (1999a) considers regression splines as a way to correct for measurement errors. The type of regression splines the Authors focus on depends on the conditional distribution of X given W. Moreover, the conditional distribution of X given W is shown to depend on the marginal distribution of X, under the assumption of additive and normally distributed measurement error. The Authors propose to model the distribution of X by a mixture of normal distributions, with an unknown number of components. The distribution of X is estimated by a modified version of the Gibbs Sampling algorithm (Wasserman and Roeder, 1997). To ensure parameter identifiability, the measurement error variance is assumed to be known. If this is not the case, as it usually happens in practice, additional information is needed.

The idea of using a mixture distribution is also adopted by Richardson *et al.* (2002), within a Bayesian framework. The Authors focus on mixture models with a variable number of components for flexibly modeling the distribution of X in Bayesian hierarchical models. This suggestion was given before in Richardson and Green (1997), who use MCMC methods based on the reversible jump algorithm proposed by Green (1995). Richardson *et al.* (2002) refer to epidemiological case-control studies, which involve validation data. The focus is mainly on the logistic disease model, where covariates are affected by normal or lognormal classical measurement errors. A key assumption is measurement error to be nondifferential. The proposed method is a functional one, thus assuming that the X's are unknown parameters for which a prior is needed. This prior is given by a mixture of univariate normals with an unknown

number of components, k. Treating k as being unknown and integrating over its posterior distribution when estimating regression parameters of interest enhances the adaptivity of the mixture to heterogeneity in the underlying distribution of X. The prior distribution for k is chosen to be vague. In particular, a uniform distribution over the range $1 - 30$ is used. However, the Authors suggest that in practice the mixture rarely uses more than ten components, so that k could be defined on a smaller range without any loss of flexibility. Several simulation studies are performed to evaluate the influence of misspecifications of the prior distribution for X and to show the improvement of using a flexible mixture distribution for X instead of a normal one.

In all the papers we focused on, the advantage of using flexible parametric models is well outlined. It relies upon their simple applicability and the robustness added to the analysis. However, a crucial point is the choice of the number of mixture components. It can be fixed as suggested by Carroll *et al.* (1999b), although this is obviously a matter of subjectiveness, or it can be left undefined, with the consequent risk of overparametrising the model. If k is allowed to increase too much, so as, for example, when it grows with the sample size (Roeder and Wasserman, 1997) the corresponding model may become useless in practice, making inference results unreliable. In fact, usually there is not information enough to allow the estimation of a large number of components. Thus, a modest value of k is more convenient. Moreover, also under a small number of mixture components, if the resulting mixture distribution is not a good approximation of the real one, the estimators can be biased. In all these cases a different approach, such as, for example, a semiparametric approach, may be preferable.

3.2 Semiparametric analysis

An alternative to the flexible parametric modeling is the semiparametric approach. It represents a response to the sensitivity of modeling assumptions, although it can be sometimes challenging to implement. The semiparametric approach has the advantage of robustness, in that it does not require the specification of the distribution of X and/or of W. However, it may lack efficiency with respect to a full likelihood approach, if the parametric specification of the model is approximately correct. This loss of efficiency may be substantial even for moderate sample sizes[Carroll *et al.* (1998)]. Different proposals in literature suggest to perform a semiparametric analysis by allowing a nonparametric specification of one or more components of the model, that is, the disease, the measurement error and/or the exposure component.

One of the first proposals of semiparametric analysis in measurement error problems is the paper by Carroll and Wand (1991). It concerns logistic regression models, with nondifferential errors on covariates. A validation data set is supposed to be available. No parametric assumption is made for the distribution of the true and unobservable covariate X or its surrogate W. The Authors develop an estimating algorithm, which is based on a kernel regression to approximate the likelihood, without modeling the distribution of X given W. Their method provides a semiparametric estimate of the parameters of the disease model, together with an asymptotically normal limit distribution of the estimators and an estimated bandwidth of the kernel regression. Independently, Pepe and Fleming (1991) consider a similar problem in the case of a discrete random variable X.

The assumption underlying the proposal by Carroll and Wand (1991) and by Pepe and Fleming (1991) is that missingness of observations from X does

not depend on the response Y. Robins *et al.* (1995) suggest a new class of estimators for the parameters of the disease model that remains consistent and asymptotically normally distributed even when the probability that X is missing depend on the observations from Y. The procedure requires a validation data consisting on observations from the X, the response variable Y and the error-free variable Z, to be available. They are needed to nonparametrically estimate the distribution of X, conditionally on Y and Z. In situation when a nonparametric estimation of the distribution of X given Z may be not practible because of the curse of dimensionality (Huber, 1985), that is, when the vector of error-free covariates Z includes more than two covariates, the estimators remains asymptotically unbiased and are computationally simple. Moreover, under certain conditions on Y and Z, the proposed class of estimators contains estimators of the parameters which are semiparametric efficient in the sense of Begun *et al.* (1983). Simulation studies performed with reference to a logistic disease model indicate that the estimators by Robins *et al.* (1995) is preferable to the one by Pepe and Fleming (1991), in terms of absolute relative efficiency.

Wang and Wang (1997) suggest a semiparametric correction technique again based on kernel regression. The focus is on logistic regression models with validation data available. The observations from X are thought to be missing data in the main study sample, with a path of missingness which depends on (Y, W) but not on X, that is, X is assumed to be missing at random (MAR). No distributional assumption is made on components such as the selection probabilities of the validation data set or the probability density of X conditionally on the other variables. The paper investigates two kernel estimation methods which extend the proposals by Breslow and Cain (1988) and by Reilly and Pepe (1995) when (W, Z) are continuous. The proposal by

Breslow and Cain (1988) suggests the use of a pseudo-conditional likelihood function in a two-stage case-control study, so that at the second stage some X's are observed in each stratum classified by (Y, W), where W is a categorical variable. The proposal by Reilly and Pepe (1995), instead, is a modified pseudo-likelihood approach for the case that (Y, Z, W) are all discrete variables and X is MAR. It extends the previous works by Carroll and Wand (1991) and Pepe and Fleming (1991). They both propose semiparametric estimators of the parameters of interest, without modeling the conditional distribution of X given (W, Z). Their solutions may lead to inconsistent estimators if the missingness process of X is not independent of Y. Reilly and Pepe (1995) extend this proposal by allowing the selection probabilities of X to depend on Y and (W, Z), when (W, Z) are discrete.

Wang and Wang (1997) extend the previous works by allowing the covariates and the surrogates to be continuous. The extension of the proposal by Breslow and Cain (1988) is obtained by using a nonparametric kernel estimation of the selection probabilities of X in the validation data. The extension of the estimator by Reilly and Pepe (1995) is based on the nonparametric kernel estimation of the conditionally expected estimating score of X given (Y, W, Z). The asymptotic properties of the two estimators are given. The simulation studies carried out by Wang and Wang (1997) to evaluate the performance of their proposals, under additive and non-normal measurement error, show a high relative efficiency of the estimators of the parameters if compared to the maximum likelihood estimator, when the modeling assumptions are incorrect.

Another semiparametric approach to correct for measurement error when validation data are available is the pseudo-likelihood analysis suggested by Carroll *et al.* (1993). It is defined for handling nondifferential errors and mod-

ified so as to include also differential errors. The method requires a parametric formulation of the disease model and the measurement error model, which can be checked in the validation subsample, while the exposure model is left unspecified. The marginal distribution of X is estimated by using a weighted average of the empirical distribution of $X|Y = y$ obtained from the complete data. This estimate is plugged into the likelihood, from which the maximum pseudo-likelihood estimates of the remaining parameters can be obtained. Simulation studies indicate that the approach gives satisfactory results with respect to the maximum likelihood approach, in terms of bias and standard errors of the estimators. However, small sample sizes can affect the estimation process with numerical instability problems due to the empirical distribution functions which are used. Moreover, modeling the relationship between Y and W by using the estimates of X may only partially recover the information about the parameters of interest which is contained in the validation data. In other words, some information about the distribution of X in the reduced data might be lost. As a consequence, maximizing the full likelihood turns out to yield more information about the parameters than a pseudo-likelihood approach, which is, of course, less efficient.

Roeder *et al.* (1996) propose an alternative to the pseudo-likelihood method by Carroll *et al.* (1993), when validation data are available. Both differential and nondifferential errors are allowed. A parametric formulation is given for the disease model and for the measurement error model, which can be checked in the validation subsample. Instead, the empirical distribution function of X, calculated on the same validation subsample, is used as a first estimate of the marginal distribution of X. The estimate is then updated by the EM algorithm or the gradient method within the estimation process of the disease model

parameters. The idea comes from Kiefer and Wolfowitz (1956), who treat the nuisance parameters x as random variables from an unspecified distribution. The estimation of the parameters is carried out via nonparametric maximum likelihood (NPML), as suggested by Laird (1978). Simulation experiments show that the proposal by Roeder *et al.* (1996) performs at least as well as or even better than the pseudo-likelihood method by Carroll *et al.* (1993), with the amount of improvement depending on the sample size and the type of measurement error.

A similar idea is followed by Schafer (2001). The Author generalizes the use of nonparametric maximum likelihood proposed by Laird (1978) for semiparametric likelihood analysis of linear, generalized linear and nonlinear regression models, where the covariates are affected by nondifferential errors. Moreover, a convenient computational form for the data analysis is provided. The approach is illustrated under a variety of structures and types of extra information about the measurement error distribution. The integral of the full likelihood (1) is approximated by a k-node quadrature

$$L(\theta; y, w, z) = \sum_{k=1}^{K} \pi_k f_{Y|XZ}(y_k|x_k^*, z_k; \beta) f_{W|X,Z}(w_k|x_k^*, z_k; \gamma), \qquad (4)$$

where π_k is $\alpha_k f_{X|Z}(x_k^*|z_k; \delta)$, the α_k's and x_k^*'s are known quadrature masses and nodes and $\theta = (\beta, \gamma, \delta)^{\mathrm{T}}$. Laird$^{Laird\ (1978)}$'s algorithm can be applied for simultaneous maximum likelihood estimation of the parameters of the disease and the measurement error model and for the estimation of $f_{X|Z}(x_k^*|z_k; \delta)$. This amounts to the estimation of the quadrature masses α_k and of the nodes x_k^*. The EM algorithm is suggested to this aim. Simulation studies indicate that this semiparametric approach retains a high degree of efficiency with respect to the full maximum likelihood inference based on correct distributional assump-

tions and can outperform maximum likelihood methods based on incorrect distributional assumptions.

Schafer (2002) follows an approach similar to Schafer (2001) for the semi-parametric analysis of linear, generalized linear and nonlinear regression models, where covariates are affected by nondifferential errors. Different types of extra information about the measurement error distribution are considered. The underlying idea is the evaluation of the integral (4) by a k-node Gauss-Hermite quadrature. It is evident that expression (4) has the form of a finite mixture of densities with mixing proportions given by π_k. Applying the EM algorithm to estimate the parameters requires the introduction of k-dimensional multinomial random variables to identify the relevant mixture component for each i, which are treated as missing data. The main difference with respect to the previously mentioned approach by Schafer (2001) is that here the number of nodes at which the integrand is evaluated is treated as a fixed quantity. That is, the approach can be thought of as an attempt of flexible structural modeling of the exposure. This implies that the only parameters to be estimated are the parameters of the disease model and the measurement error model. However, this approach bears some issues which require further investigation. First of all, there is no guarantee of numerical stability of the EM algorithm. Secondly, there is no clear indication about the number of nodes required in any situation, although 20 seems to be sufficient at least in the examples analyzed by the Author. Finally, the approach has been proposed in situations with a single unobservable covariate. While its extension to several X's measured with error is theoretically possible, the application may be unrealistic because of computational difficulties.

Within a Bayesian framework, Müller *et al.* (1997) propose to correct for

measurement error in covariates by a semiparametric approach wihch is especially designed for handling case-control data. The method focuses on semiparametrically modeling the distribution of X. This is obtained by using a mixture of normal models with a Dirichlet process prior on the mixing measure (Antoniak, 1974; Escobar and West, 1995). Using multivariate normal kernels in the mixture implicitly assumes that covariates are continuous. However, the Authors indicate that the application of the method to categorical covariates is possible by using different distributions in place of a mixture of normals. The procedure to estimate the parameters of the disease model is based on a combination of Markov chain Monte Carlo techniques. The method by Müller *et al.* (1997) is developed under the assumption of nondifferential errors and the availability of validation data. Simulation studies performed assuming a logistic disease model show that the method is robust against heteroschedasticity over cases and controls, whereas it is sensitive to differential error. When nondifferential measurement errors hold, the method is preferable in terms of bias and mean squared error to the proposal by Carroll *et al.* (1993). Under differential measurement error, instead, the situation reverses, the method by Carroll *et al.* (1993) having the advantage of exhibiting a smaller bias.

Later, Mallick *et al.* (2002) develop semiparametric Bayesian methods for regression models where measurement errors follow a classical structure, a Berkson structure or a combination of both of them. The method suggested by the Authors is semiparametric in the specification of both the disease model and the exposure model. The disease model is supposed to be monotone in the unobserved variable X and thus it is specified through a semiparametric monotone form. In particular, a mixture of beta cumulative distribution functions is used. The distribution of the unobserved X is also semipara-

metrically modeled, by using a Pólya tree distribution (Lavine, 1992; Walker and Mallick, 1999). However, as the Authors suggest, flexible semiparametric alternatives to the Pólya distribution could be used. Simulation studies performed under a logistic disease model and a combination of classical and Berkson measurement error components indicate a satisfactory behaviour of the proposed method with respect to the *naive* analysis and the one based on the true simulated data for X.

In econometric research, Li and Hsiao (2004) recently proposed a semiparametric approach to correct for classical errors in covariates in generalized linear models. The hypothesis of nondifferential error is relaxed by assuming only that $E[U|Y] = 0$. Additional data as replicated measures of X are considered to be available. The proposal by Li and Hsiao (2004) does not make distributional assumptions on the unobservable variable X or the measurement errors. The method is based on maximizing an asymptotically corrected likelihood (ACL) function. It is a two-stage method. At the first stage, the distribution of X is nonparametrically identified. This is done by using the empirical characteristic functions and truncated inverse Fourier transform, as suggested by Li (2002). At the second stage, a semiparametric estimator of the parameters of interest is derived by maximizing the ACL function using the estimated distribution of X obtained at the first stage. The Authors show that the ACL converges to the same likelihood function one would obtain with observed X. However, some future lines of research are pointed out. First of all, the need of evaluating the asymptotic distribution and the rate of convergence of the ACL estimator. Simulation studies compare the proposed ACL estimator to the *naive* maximum likelihood estimator and to the corrected score estimator by Nakamura (1990), which is based on the normality assumption of errors

(see Section 3.4). The comparison is in terms of bias and standard error of the estimators. Results show that the ACL method outperforms the corrected score when the measurement error distribution is misspecified as a normal. Standard errors are larger than those of alternative methods, as a consequence of the first stage nonparametric estimation, while bias reduction is substantial. This leads to a notable reduction in mean squared error. As expected, *naive* analysis yields worse results.

3.3 Quasi-likelihood methods

Quasi-likelihood is a promising alternative to the full likelihood approach for the analysis of measurement error data. It has the advantage of combining higher flexibility with a smaller computational effort. Quasi-likelihood requires the specification of the first two moments, that is, of the mean and the variance, of the conditional distribution of Y given X and Z and not of its entire distribution (see (Carroll *et al.*, 2006), Section 8.8). That is, one needs only to specify

$$E[Y|X,Z] = m_Y(x,z;\beta_1) \qquad \text{and} \qquad Var[Y|X,Z] = g_Y(x,z;\beta_1,\beta_2). \quad (5)$$

The approach includes quasi-likelihood methods for generalized linear models as special cases. Quasi-likelihood methods require that the mean and variance functions be evaluated on the observed data and not on the unobservable ones. These are given by

$$E[Y|W,Z] = E[m_Y(\cdot)|W,Z] \text{ and } Var[Y|W,Z] = E[g_Y(\cdot)|W,Z] + Var[m_Y(\cdot)|W,Z]. \quad (6)$$

An example is given in Carroll and Stefanski (1990). The Authors consider the application of the quasi-likelihood method in case-control studies, where

data are affected by nondifferential measurement errors, which can be classical as well as Berkson errors. Validation data, in the form of gold standard measurements of X, are required. No assumption is made on the distribution of X given W, but only on the first two moments of the measurement error given W. The Authors propose M-estimators for the parameters of interest, starting from estimating equations based on Taylor series expansions of the mean and variance functions. Their asymptotic distribution is provided under different additional data scenarios.

Wang *et al.* (1996) consider quasi-likelihood estimation under the hypothesis that correlated replicates of the proxy variable W are available. A nondifferential and classical additive measurement error on the covariate is assumed. The Authors perform a quasi-likelihood analysis by computing the mean and variance functions through Monte Carlo methods. The distribution of X is suggested to be flexibly modeled by using a mixture of normals. The application of the method is illustrated on a real data set. The results show the improvement with respect to a RC approach which ignores the correlation structure of replicates, both in terms of bias ans standard error of the parameter estimators.

3.4 Estimating equations

The use of estimating equations in measurement error problems has been mainly studied in two variants which are referred to as corrected score and conditional score methods, although alternatives have been recently suggested.

The corrected and conditional score methods were developed starting from the estimating equations for regression parameters in the absence of measurement error. An estimating equation is unbiased if it has expectation zero. An

example is the score function, that is, the first derivative of the log-likelihood function with respect to the parameters. Measurement error induces bias in estimating equations, which in turn gives rise to biased estimators for the parameters. Thus, the purpose is to modify the estimating equations so as to obtain unbiased estimating equations.

The *corrected score method* specifies corrected score functions, which are unbiased estimators of the score function yielding the estimator one would use if there was no measurement error. The method of corrected score functions was studied by Stefanski (1989) and Nakamura (1990). In the absence of measurement error, consider the estimate of β which solves $\sum_{i=1}^{n} \psi(y_i, x_i, z_i; \beta) = 0$, where n is the sample size and $\psi(\cdot)$ is the estimating function. The function $\psi(\cdot)$ is typically a likelihood score function from the model for the data without error. It is unbiased if its expectation is zero, that is, $E[\psi(Y, X, Z; \beta)] = 0$. Generally, it is no longer unbiased when W replaces X. Corrected score functions instead, say $\psi^*(y, w, z; \beta)$, have the property that $E[\psi^*(Y, W, Z; \beta)] = \psi(Y, X, Z; \beta)$, where the expectation is with respect to the distribution of W given (Y, X, Z). The corrected scores are unbiased whenever the original scores are. Unbiasedness is a major requirement for consistency of the estimators obtained from corrected score functions.

The corrected score method applies to generalized linear models, as, for example, the gamma regression model with logarithmic link. It requires that a measurement error distribution be specified. The normal distribution is typically used for this purpose (Stefanski, 1989). Corrected score functions do not always exist and finding them when they do is not always as easy as in the linear case. A typical example is logistic regression which does not admit a corrected score function, except under restrictions (Buzas and Stefanski, 1996).

Stefanski (1989) derived corrected score functions for some common models and generally applicable approximate corrected score functions. Recently, a method for obtaining corrected score functions via computer simulation was studied (Novick and Stefanski, 2002).

The *conditional score method* was introduced by Stefanski and Carroll (1985) and developed into the usually applied formulation by Stefanski and Carroll (1987) within an important class of generalized linear models. The most important example is logistic regression. Carroll *et al.* (2006), Section 7, describe extensions of the method to Poisson-loglinear, gamma-inverse and other models.

The conditional score is a functional method based on the theory of sufficient statistics, on which we can condition to eliminate the nuisance parameters x. Stefanski and Carroll (1987) assumed that the measurement errors are normally distributed. However, the estimator can reduce bias also for small departures from this assumption (Huang and Wang, 2001). Stefanski and Carroll (1987) focus on logistic regression with classical measurement error, although the method applies to other generalized linear models, provided the measurement errors are normal and the models are in the canonical form (see (Carroll *et al.*, 2006), Section 7). They provide the conditional score estimator for logistic regression and show that it behaves satisfactorily in terms of efficiency with respect to the full maximum likelihood estimator which, however, requires the specification of an exposure model (Stefanski and Carroll, 1990).

For models other than logistic regression, the conditional score estimating equations are far more complicated (see (Carroll *et al.*, 2006), Section 6.4) and typically computed by means of numerical integration.

Outside the conditional score and the corrected score formulation, other

proposals to correct for measurement error have been suggested which are based on the theory of estimating equations. An example is the paper by Iturria *et al.* (1999). The Authors derive estimators of parameters of the disease model and their asymptotic standard errors in the polynomial regression model, by referring to corrected estimating equations which do not necessarily come from the score function. Additive and multiplicative measurement errors are considered. Conditions under which it is possible to estimate parameters are given. These conditions do not rely on distributional assumptions about the X's, but use ratios of the W's, thus making the method be a robust solution. The method may be easily extended to general estimating functions. The basic idea is that an estimating function can be expanded as a polynomial, thus allowing the proposal by Iturria *et al.* (1999) to be applied. Simulation studies carried out to compare the method and the likelihood approach show that the first provides more reliable results whenever models for measurement errors are misspecified. This is mainly the case for skewed errors.

Recently, Wang and Pepe (2000) focused on the use of estimating equations to correct for measurement error in marginal or partly conditional regression models for longitudinal data. Measurement errors are assumed to be nondifferential. Estimating equations are considered which are not necessarily likelihood score equations. They have to be unbiased when evaluated on the complete data, that is, on observations from (Y, X, Z). The Authors propose to base the estimation of the parameters of the disease model on the expectation of the estimating equation for the complete data conditioned on the available data. The estimates are derived as solutions of the resulting estimating equations. The expected estimating equations (EEE, for short), as they are called, yield an estimator which has the property of being equal to the maximum likelihood

estimator if the complete data scores are likelihood scores and conditioning is with respect to all the available data. The asymptotic distribution of the estimator is derived. Its behaviour is compared to the RC estimator through simulations studies of a logistic disease model, under an order one autoregressive model for the error process. Simulation results indicate that for moderate sample sizes, with large relative risk, the EEE estimator is more efficient than the RC estimator, while it can suffer from both a large bias and a large standard error in small samples. This agrees with the behaviour of the maximum likelihood estimator which suffers from bias in the presence of small sample sizes.

As Wang and Pepe (2000), also Pan *et al.* (2006) focus on longitudinal data, where a single covariate X is assumed to be affected by measurement error. The error is supposed to be additive and nondifferential. The Authors mainly refer to the transition models, that is, models where the conditional mean of the response variable at the current time point is modeled as a function of its value at the previous time and covariates (see (Diggle *et al.*, 2002), Chapter 10). Within this setting, an estimating equation approach is proposed by modifying the conditional score method by Stefanski and Carroll (1987). This gives rise to the so-called pseudo conditional score estimators of the disease model parameters. They are shown to be consistent and asymptotically normally distributed. Moreover, an alternative to the pseudo conditional score estimator is proposed, namely a semiparametric efficient one-step estimator. It improves the efficiency of the pseudo conditional score estimator, by taking advantage of the explicit expression of the efficient score function for the parameters of interests. Moreover, the one-step estimator reaches the semiparametric efficiency bound in the presence of validation data. However, the

explicit formulation of the efficient score function which the one-step estimator relies on does not exist for models more complicated than the linear model, as, for example, the logistic transition model.

3.5 Empirical likelihood

The paper by Wang and Rao (2002) is the first example of application of empirical likelihood in measurement error problems. The empirical likelihood, introduced by Owen (1988), is useful to construct confidence regions under a nonparametric model. It has some advantages with respect to classical methods, in that it does not require the definition of pivotal quantities for inferential purposes and provides confidence regions which are range-preserving and transformation-respecting (Hall and La Scala, 1990).

Wang and Rao (2002) focus on linear regression model, when validation data are available. Measurement errors are assumed to be nondifferential. The regression model is re-written in an equivalent form where unobserved covariates X are substituted by $E[X|W]$. The empirical log-likelihood function is then evaluated starting from this formulation. To estimate the parameters of interest, the quantity $E[X|W]$ has to be replaced by known values derived from the validation data. The idea is similar to the one underlying the RC approach. This substitution leads to an estimated empirical log-likelihood. The Authors show that the resulting estimated empirical log-likelihood follows asymptotically a χ^2 distribution and use it to define confidence regions for the parameters of interest. However, such an approach can suffer from the curse of dimensionality when the dimension of X and hence of W is large, because of the required estimation of $E[X|W]$. In this case, dimension-reduction models may be preferable for estimating $E[X|W]$. However, the corresponding asymptotic

theory has still to be developed and is an interesting field of future research.

Cui and Chen (2003) suggest a different approach based on empirical likelihood to derive confidence regions for the parameter of the disease model. The focus is on linear regression models, where covariates are assumed to be affected by classical and nondifferential measurement errors. The Authors illustrate how to construct empirical likelihood confidence regions by starting from a modification of the score function. This adds up squared orthogonal distances for each data point to a hyperplane in the parameter space. Such a score function differs from the one from an ordinary linear model in that the former has more than two solutions, of which only one is genuine. This solution is found by constraining the empirical likelihood to a restricted region of the parameter space. The Authors evaluate the coverage accuracy and Bartlett correctability of the confidence regions derived from this approach. Simulation studies are performed to compare the behaviour of the proposed empirical likelihood confidence region to that based on the asymptotic normal distribution of the estimators of the parameters. The results show that the empirical likelihood-based method provides confidence regions with better coverage and shorter lenghts than the normal approximation counterpart. This improvement is already notable for small or moderate sample sizes.

3.6 Further techniques

Further approaches which make no distributional assumption on the involved variables were proposed in literature, although they can not be classified into one of the previous groups. They are summarized below.

- Cook and Stefanski (1994) develop a simulation-extrapolation method (SIMEX, for short), which is a functional simulation-based method to

correct for measurement error affecting the covariates. It has been further developed by Stefanski and Cook (1995), Carroll *et al.* (1996) and Wang *et al.* (1998). The method is robust in that it does not make distributional assumptions on the unobserved variables X. The idea underlying SIMEX is that the effect of the measurement error can be determined by simulation. The method develops in two steps. The first one is a resampling-like stage, in which data sets with additional measurement error are generated starting from the original one. For each data set the *naive* estimate of the parameters is obtained, so that the trend of the estimates versus the variance of the extra error terms can be established. The corrected estimators of the parameters are obtained in the second stage by extrapolating this trend back to the case of no measurement error. Carroll *et al.* (1996) investigate the asymptotic distribution of the SIMEX estimator. They show that it is asymptotically normally distributed and provide methods to consistently estimate the variance. Later, Fung and Krewski (1999) propose a comparison between RC and SIMEX estimators, by means of a computer simulation in a logistic regression framework. Their study shows that RC and SIMEX estimators generally exhibit a satisfactory and similar performance in terms of bias, mean squared error and coverage of confidence intervals. When a Berkson measurement error model in highly correlated predictors holds, however, the SIMEX method seems to be preferable. On the other hand, RC has the nontrivial advantage to be a simpler and less computationally intensive method.

- Haukka (1995) suggests to correct for covariate measurement error in generalized linear models by using bootstrap techniques. The method

is referred to as two-stage bootstrap, because both the primary and the validation data are resampled. It requires validation data to be available. At the first stage, a bootstrap sample is taken from the validation data set. It is used to estimate the parameters of the measurement error model relating X to the *proxy* variables W and to the error-free covariates Z. A bootstrap sample is then taken from the primary data. This sample is used to estimate the parameters of interest, with X replaced by the predicted values obtained in the regression at the previous stage. Bootstrap sampling generally involves $50 - 100$ replications. The empirical distribution of the estimator is used for making inference on the parameters. The method is illustrated under the assumption of continuous linear measurement errors, although extensions to other measurement functions require only slight modifications of the procedure. The nonparametric nature of the method turns out in a nontrivial gain in robustness, if compared to simpler approaches as, for example, RC. In fact, simulation studies, performed in the logistic regression framework, showed that the method is a valid alternative to the RC, although it can lead to larger confidence intervals, especially in situations where the distribution of the errors is asymmetric. Despite of this, the principal disadvantage of Haukka (1995)'s method relies in its computational burden connected with the intensive application of the bootstrap technique.

- Lee and Sepanski (1995) propose an estimation method which is computationally simpler than semiparametric and nonparametric methods described in Section 3.2, both for linear and nonlinear disease models. The method relies neither on distributional assumptions nor on specifications of the equations relating the measured variable W to the true variable

X, thus obtaining considerable gain in robustness. Additive measurement errors are considered and they are allowed to affect the covariates as well as the dependent variable. The method is based on replacing the regression function of Y on (X, Z) by a wide-sense conditional expectation, or least squares projection, of the regression function on functions of W (Chamberlain, 1982). The underlying idea is that the original regression function can be projected onto a finite-order polynomial of W. This wide-sense conditional expectation can be estimated from validation data using the ordinary least squares method. After replacement of the original regression function by this conditional expectation, nonlinear least squares can be used to estimate the parameters. The choice of the polynomial for the projection space is arbitrary. Simulation studies performed by the Authors suggest that few polynomials of low degree are good enough even for highly nonlinear functions.

- In econometrics, Chesher (2000) notes that, to the first order of approximation, the bias implied by measurement errors can be determined by a functional of the marginal distribution of the mismeasured variable W. The suggested correction technique, which follows Chesher (1991), is based on the construction of a nonparametric estimate of the functional of the distribution of W. The assumptions of independence between X and the errors U and of nondifferential errors are needed. Monte Carlo experiments, performed both when the measurement errors are normally and non-normally distributed, indicate that the proposed method can substantially reduce bias in estimators, if compared to a *naive* approach. Moreover, in linear and polynomial models, the method can be combined with the classical instrumental variables procedure, thus improving the

efficiency of both approaches.

- Sepanski (1994) suggest to correct for measurement error in a class of models including the generalized linear models by an approach strictly related to RC. It is a semiparametric RC method, requiring a validation data set consisting of exact measures of X. It applies to nondifferential measurement error which are not necessarily classical and additive. The underlying idea is the substitution of the unobserved X's by the estimates of $E[X|W]$ obtained from a nonparametric kernel regression in the validation data. Once the unknown X's are replaced by these estimates, a standard analysis can be performed. The parallelism with RC is evident. However, this method gives rise to a gain in robustness against deviations from the linear relationship between X and W underlying the original RC idea, when this relationship does not hold. Moreover, the Authors provide the asymptotic distribution of the regression parameter estimators and discuss the choice of the bandwidth parameter, involving higher-order expansions for the covariance matrix of the corrected estimators. Although the focus is on nonparametric kernel regression to estimate $E[X|W]$, the Authors suggest that other smoothing techniques could be used, including local linear kernel smoothing, lowess, spline smoothing and generalized additive models. Simulation studies carried out to compare the method against parametric alternatives, as, for example, RC, indicate that it has a comparable performance, which in some cases is also better, mainly under multiplicative error structures.

- Another approach which can be related to RC is suggested by Pierce and Kellerer (2004). The Authors propose to adjust for errors in covariates

by using a nonparametric assessment of the true covariate distribution. Their proposal can be used within the RC approach. In fact, the expected value of X given W, which is needed in the RC procedure, can be nonparametrically derived, although it involves a deconvolution which is difficult to carry out directly. However, with multiplicative and log-normal measurement errors, the Authors derive simple but accurate approximations for the k-th order moment of X given W, with $k = 1, 2, \ldots$. These approximations depend only on the first and second derivatives of the logarithm of the density of W and the coefficient of variation of W given X. Both classical and Berkson errors are considered.

- Berry *et al.* (2002) suggest a robust approach to the analysis of measurement error data, where robustness is related to misspecification of the disease model and not on the exposure model, as commonly adopted. The Authors propose a flexible nonparametric estimation of the regression function, by using smoothing splines or regression P-splines, within a Bayesian framework. The posterior distribution of the parameters of interest may be obtained from two algorithms. The first one, the so-called iterative conditional modes, uses a componentwise maximization routine to find the mode of the posterior distribution, while the second is a fully Bayesian method based on Monte Carlo Markov Chain techniques to generate observations for the posterior distribution. Although the last is computationally more difficult than the first one, it is preferable in that it allows exploring the posterior distribution, rather then only finding the mode. Simulation studies performed to evaluate the potential of the correction technique by Berry *et al.* (2002) with respect to alternatives show that it is competitive in efficiency with similar approaches performed in

the frequentist framework, as, for example, the method by Carroll *et al.* (1999a). The normal distribution for the additive measurement error and for the exposure variable is assumed, although simulation studies show that small departures from this assumption only slightly modify the results.

- Jiang and Turnbull (2004) base statistical inference in measurement error models on the so-called *indirect method*. This is an approach to inference which has been exploited in econometrics (Gouriéroux *et al.*, 1993) as a robust alternative to likelihood-based procedures. The indirect method is based on the search of an intermediate statistic as a functional of the empirical distributon function. The intermediate statistic typically follows an asymptotic normal distribution, but it is not necessarily a consistent estimator of the parameter of interest. An example is the *naive* estimator. Jiang and Turnbull (2004) focus on the indirect method to suggest a consistent estimator of the disease model parameter without requiring parametric assumptions on the distribution of (X, W), thus obtaining a notable gain in robustness of results. Moreover, only the first moment is specified for the disease model. The assumption of nondifferential errors and the availability of validation data is needed. A consistent estimator of the parameter of interest is found starting from the *naive* solution and its asymptotic distribution is derived. The application of the method is evaluated within a logistic framework and compared to that of RC. Results outline the improvement of the indirect method in estimating the parameter of interest, mainly in situations where assumptions required by RC are not satisfied.

- Tsiatis and Ma (2004) propose a class of semiparametric estimators, which are called locally efficient semiparametric estimators, within the functional measurement error setting. This class is derived by defining estimating equations for the parameters of the disease model. The estimating equations are obtained from the efficient score derived as the residual after projecting the score vector with respect to the disease model parameters onto the tangent space for the distribution of X. Tsiatis and Ma (2004) show that the residual has mean zero even under misspecified distributions for X. This allows one to form estimating equations for the parameters of the disease model which yield to consistent and asymptotically normally distributed estimators. Moreover, if the model for X is correctly specified, the resulting estimator is semiparametric efficient. The assumption underlying the method of known measurement error distribution may be relaxed if additional data are available to estimate the unknown parameters of the distribution. Simulation studies are performed to evaluate the behaviour of the proposed estimator, under a quadratic logistic disease model and two measurement error structures, the first having normally distributed errors while the second having exponentially distributed errors. Results show that, in both of the cases, the proposed locally efficient estimator is robust against misspecification of the distribution of X. If compared to the RC estimator, the locally efficient estimator is preferable in terms of bias and empirical coverage of confidence intervals.

4 Discussion

We have provided a review of techniques to correct for measurement error in covariates which represent solutions to the sensitivity to assumptions typical of parametric approaches. Different solutions have been proposed in literature, which may be more or less challenging to implement. Some of them combine a parametric and a nonparametric spefication of relationships between variables, while other methods face the problem by adopting a totally nonparametric approach. Although solutions are variously developed, they share characteristics of robustness against model misspecifications, the principal being the misspecification of the exposure model. However, in all cases, this advantage does not come without costs. The higher price to pay for it is the possible loss in efficiency relative to parametric models if they are approximately correct.

Furthermore, some computational problems related to the difficulties in implementing most of the suggested methods are non-negligible. Focus, for example, on semiparametric techniques. The proposed methods in this group share the common approach of nonparametrically estimating one of the relationships between variables, that is, the disease, the measurement error or the exposure model. Although these modifications are applied to the likelihood function given in (1), problems related to a full likelihood approach may still be present, like difficulties in the maximization procedure and in the evaluation of the involved integrals. Usually numerical methods or analytical approximations are required and the associated computational effort tends to increase in case of high-dimensional models. If this is the case, alternative solutions may be preferable. From a strictly practical point of view, the most feasible solutions seem to be those based on the idea underlying RC, the so-called semiparametric RC methods. Starting from the simplest technique to correct for

measurement error, i.e. regression calibration, a nonparametric modification yields a gain in robustness, without affecting the feasibility of the approach.

Other solutions, as for example estimating equations, in spite of a well known underlying theory, may be less attractive because of difficulties in application, which are not necessarily computational difficulties. As it can be seen from the paper by Wang and Pepe (2000), deriving unbiased estimating equations for the parameters is very often a nontrivial problem, mainly in situations with matched or unmatched case-control data. In this case, in fact, if one starts from a formulation like the one in (1), for example, it is not possible to obtain estimating equations and estimators of parameters in an explicit form. Moreover, bias correction can be hardly achieved. Further investigation in this area seems to be needed.

Empirical likelihood is a powerful tool for inference in nonparametric settings. The methods suggested by Wang and Rao (2002) and Cui and Chen (2003), which apply empirical likelihood in measurement error problems, seem to be promising in terms of robustness properties, nevertheless studies are restricted to linear regression models at the moment. Although the attention of this review has been mainly focused on models appropriate to handle case-control data, we have mentioned the previous works on empirical likelihood in order to highlight the fact that, on the basis of the promising results, extensions to more general models may be an interesting field of further investigations.

Most of the proposals reviewed here heve been developed under the assumption of nondifferential measurement errors. The possibility for differential measurement errors, instead, has been rarely examined. Although a nondifferential assumption is appropriate in many situations, mainly through a good experimental design, sometimes it may not be appropriate in case-control stud-

ies. In fact, when the possibility of select or recall bias arises, as it is typical in case-control studies, thus measurement error can depend on the disease status, that is, it can be differential. In this situation, many of the existing techniques to correct for measurement errors are not applicable. This suggests the need for further research to extend correction methods developed under the assumption of nondifferential errors to the situation of differential errors.

A common feature of methods examined here is their application to problems where just a single covariate is affected by measurement error. Additional error-free covariates may be considered. The main reason relies on the computational effort required by a more extensive analysis, which may become quite cumbersome. As the dimension of X increases, the extension of most of the procedures is not straightforward and their application may become less attractive. An example is the augmented complexity of integrals which have to be evaluated in semiparametric methods. Thus, further investigations are needed in this area. The research for extension of the existing methods to higher dimensions of unobserved covariates and/or their surrogates is required so as to make them suitable for more realistic problems. These may involve more than one covariate affected by measurement error, with the possibility of some correlation patterns among errors.

Acknowledgments

This research was supported by *Associazione Italiana per la Ricerca sul Cancro*, with additional support provided by the Italian Ministry for Education, University and Research. The author is grateful to Prof. Alessandra Salvan and Dr. Alessandra R. Brazzale for helpful comments on the preprint version of the paper. She also acknowledges Prof. Raymond J. Carroll for his

suggestions.

References

Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to non-parametric problems. *The Annals of Statistics*, **2**, 1152–74.

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

Begun, J.M., Hall, W.J., Huang, W.M. and Wellner, J.A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *The Annals of Statistics*, **11**, 432–452.

Berry, M., Carroll, R.J. and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, **97**, 160–169.

Breslow, N.E. and Cain, K.C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, **75**, 11–20.

Buzas, J.S. and Stefanski, L.A. (1996). A note on corrected score estimation. *Statistics & Probability Letters*, **28**, 1–8.

Carroll, R.J., Freedman, L.S. and Pee, D. (1998). Design aspects of calibration studies in nutrition, with analysis of missing datat in linear measurement error models. *Biometrics*, **53**, 1440–1457.

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

Carroll, R.J., Küchenhoff, H., Lombard, F. and Stefanski, L.A. (1996). Asymptotics for the SIMEX estimator in structural measurement error models. *Journal of the American Statistical Association*, **91**, 242–250.

Carroll, R.J., Maca, J.D. and Ruppert, D. (1999a). Nonparametric regression in the presence of measurement error. *Biometrika*, **86**, 541–554.

Carroll, R.J., Roeder, K. and Wasserman, L. (1999b). 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.

Carroll, R.J. and Stefanski, L.A. (1990). Approximate quasilikelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, **85**, 652–663.

Carroll, R.J. and Wand, M.P. (1991). Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society, Series B*, **53**, 573–585.

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics*, **18**, 5–46.

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

Chesher, A. (2000). Measurement error bias reduction. *Unpublished Manuscript*, University College of London.

Cook, J. and Stefanski, L.A. (1994). A simulation extrapolation method for parametric measurement error models. *Journal of the American Statistical Association*, **89**, 1314–1328.

Cui, H. and Chen, S.X. (2003). Empirical likelihood confidence region for parameters in the errors-in-variables models. *Journal of Multivariate Analysis*, **84**, 101–115.

Diggle, P.J., Heagerty, P., Liang, K.Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data*. Second Edition, Oxford University Press, Oxford.

Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. J*ournal of the American Statistical Association*, **90**, 577-588.

Fung, K.Y. and Krewski, D. (1999). Evaluation of regression calibration and SIMEX methods in logistic regression when one of the predictors is subject to additive measurement error. *Journal of Epidemiological Biostatistic*, **4**, 65–74.

Gleser, L.J. (1990). Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models. In *Statistical Analysis of Measurement Error Models and Application* (P. J. Brown and W. A. Fuller, Eds). American Mathematics Society, Providence.

Gouriéroux, C., Monfort, A. and Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics*, **8**, 85–118.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Hall, P. and La Scala, B. (1990). Methodology and algorithms of empirical likelihood. *International Statistical Review*, **58**, 109–127.

Haukka, J.K. (1995). Correction for covariate measurement error in generalized linear models — a bootstrap approach. *Biometrics*, **26**, 1127–1132.

Heid, I.M., Küchenhoff, H., Miles, J., Kreienbrock, L. and Wichmann, H.E. (2004). Two dimensions of measurement errors: classical and Berkson error in residential radon exposure assessment. *Journal of Exposure Analysis and Environmental Epidemiology*, **14**, 365–377.

Higdon, R. and Schafer, D.W. (2001). Maximum likelihood computations for regression with measurement error. *Computational Statistics & Data Analysis*, **35**, 283–299.

Huang, X., Stefanski, L.A. and Davidian, M. (2006). Latent-model robustness in structural measurement error models. *Biometrika*, **93**, 53–64.

Huang, Y. and Wang, C. (2001). Consistent functional methods for logistic regression with errors in covariates. *Journal of the American Statistical Association*, **96**, 1469–1482.

Huber, P. (1985). Projection pursuit. *The Annals of Statistics*, **13**, 435–474.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, **27**, 886–906.

Iturria, S.J., Carroll, R.J. and Firth, D. (1999). Polynomial regression and estimating functions in the presence of multiplicative measurement error. *Journal of the Royal Statistical Society, Series B*, **61**, 547–561.

Jiang, W. and Turnbull, B. (2004). The indirect method: inference based on intermediate statistics – A synthesis and examples. *Statistical Science*, **19**, 239–263.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, **73**, 805–811.

Lavine, M. (1992). Some aspects of Pólya tree distributions for statistical modeling. *The Annals of Statistics*, **20**, 1222–1235.

Lee, L-F. and Sepanski, J.H. (1995). Estimation of linear and nonlinear errors-in-variables models using validation data. *Journal of the American Statistical Association*, **90**, 130–140.

Li, T. (2002). Robust and consistent estimation in nonlinear errors-in-variables models. *Journal of Econometrics*, **110**, 1–26.

Li, T. and Hsiao, C. (2004). Robust estimation of generalized linear models with measurement errors. *Journal of Econometrics*, **118**, 51–65.

Mallick, B., Hoffman, F.O. and Carroll, R.J. (2002). Semiparametric regression modeling with mixtures of Berkson and classical error, with application to fallout from the Nevada Test Site. *Biometrics*, **58**, 13–20.

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

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

Novick, S.J. and Stefanski, L.A. (2002). Corrected score estimation via complex variable simulation extrapolation. *Journal of the American Statistical Association*, **97**, 472–481.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–249.

Pan, W., Zeng, D. and Lin, X. (2006). Estimation in semiparametric transition measurement error models for longitudinal data. *Harvard University Biostatistics Working Paper Series*, **52**. `http://www.bepress.com/harvardbiostat/paper52`

Pepe, M.S. and Fleming, T.R. (1991). A general nonparametric method for dealing with errors in missing or surrogate data. *Journal of the American Statistical Association*, **86**, 108–113.

Pierce, D.A. and Kellerer, A.M. (2004). Adjusting for covariate errors with nonparametric assessment of the true covariate distribution. *Biometrika*, **91**, 863–876.

Reilly, M. and Pepe, M.S. (1995). A mean-score method for missing and auxiliary covariate data in regression models. *Biometrika*, **82**, 299–314.

Richardson, S., Leblond, L., Jaussent, I. and Green, P.J. (2002). Mixture models in measurement error problems, with reference to epidemiological studies. *Journal of the Royal Statistical Society, Series A*, **165**, 549–566.

Richardson, S. and Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B*, **59**, 731–792.

Robins, J.M., Hsieh, F. and Newey, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *Journal of the Royal Statistical Society, Series B*, **57**, 409–424.

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

Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, **92**, 894–902.

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

Rosner, B., Spiegelman, D. and Willett, W.C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *American Journal of Epidemiology*, **132**, 734–745.

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

Schafer, D. (2002). Likelihood analysis and flexible structural modeling for measurement error model regression. *Journal of Statistical Computation and Simulation*, **72**, 33–45.

Sepanski, J.H., Knickerbocker, R. and Carroll, R.J. (1994). A semiparametric correction for attenuation. *Journal of the American Statistical Association*, **89**, 1366–1373.

Stefanski, L.A. (1989). Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. *Communications in Statistics, Series A*, **18**, 4335–4358.

Stefanski, L.A. and Carroll, R.J. (1985). Covariate measurement error in logistic regression. *The Annals of Statistics*, **13**, 1335–1351.

Stefanski, L.A. and Carroll, R.J. (1987). Conditional scores and optimal scores in generalized linear measurement error models. *Biometrika*, **74**, 703–716.

Stefanski, L.A. and Carroll, R.J. (1990). Score tests in generalized linear measurement error models. *Journal of the Royal Statistical Society, Series B*, **52**, 345–359.

Stefanski, L.A. and Cook, J. (1995). Simulation extrapolation: the measurement error jackknife. *Journal of the American Statistical Association*, **90**, 1247–1256.

Thürigen, D., Spiegelman, D., Blettner, M., Heuer, C. and Brenner, H. (2000). Measurement error correction using validation data: a review of methods and their applicability in case-control studies. *Statistical Methods in Medical Research*, **9**, 447–474.

Tsiatis, A.A. and Ma, Y. (2004). Locally efficient semiparametric estimators for functional measurement error models. *Biometrika*, **91**, 835–848.

Walker, S. and Mallick, B.K. (1999). Semiparametric accelerated life time models. *Biometrics*, **55**, 477–483.

Wang, C.-Y. and Pepe, M.S. (2000). Expected estimating equations to accomodate covariate measurement error. *Journal of the Royal Statistical Society, Series B*, **62**, 509–524.

Wang, C.Y. and Wang, S. (1997). Semiparametric methods in logistic regression with measurement error. *Statistica Sinica*, **7**, 1103–1120.

Wang, N., Carroll, R.J. and Liang, K.-Y. (1996). Quasi-likelihood estimation in measurement error models with correlated replicates. *Biometrics*, **52**, 401–411.

Wang, N., Lin, X., Gutierrez, R.G. and Carroll, R.J. (1998). Bias analysis and SIMEX inference in generalized linear mixed measurement error models. *Journal of the American Statistical Association*, **93**, 249–262.

Wang, Q. and Rao, J.N.K. (2002). Empirical likelihood-based inference in linear errors-in-covariables models with validation data. *Biometrika*, **89**, 345–358.

Wasserman, L.A. and Roeder, K. (1997). Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, **90**, 1247–1256.

Zeger, S.L., Thomas, D., Dominici, F., Samet, J.M., Schwartz, J., Dockery, D. and Cohen, A. (2000). Exposure measurement error in time-series studies of air pollution: concepts and consequences. *Environmental Health Perspective*, **108**, 419–426.

Working Paper Series
Department of Statistical Sciences, University of Padua

You may order paper copies of the working papers by emailing wp@stat.unipd.it

Most of the working papers can also be found at the following url: http://wp.stat.unipd.it

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

Department of Statistical Sciences
University of Padua
Italy