J. R. Statist. Soc. B (2007)
69, *Part* 3, *pp.* 429–446

Semiparametric estimators of functional measurement error models with unknown error

Peter Hall

Australian National University, Canberra, Australia

and Yanyuan Ma

Texas A&M University, College Station, USA

[Received January 2006. Revised December 2006]

Summary. We consider functional measurement error models where the measurement error distribution is estimated non-parametrically. We derive a locally efficient semiparametric estimator but propose not to implement it owing to its numerical complexity. Instead, a plug-in estimator is proposed, where the measurement error distribution is estimated through non-parametric kernel methods based on multiple measurements. The root *n* consistency and asymptotic normality of the plug-in estimator are derived. Despite the theoretical inefficiency of the plug-in estimator, simulations demonstrate its near optimal performance. Computational advantages relative to the theoretically efficient estimator make the plug-in estimator practically appealing. Application of the estimator is illustrated by using the Framingham data example.

Keywords: Errors in variables; Latent variables; Measurement error; Semiparametric efficiency; Semiparametric methods

1. Introduction

Measurement error models frequently occur in practice when the measurements on covariates contain error. It is well known that ignoring the measurement error can lead to biased estimation. Various methods have been proposed to take this problem into account. A comprehensive study of these methods has been given by Fuller (1987) for linear models, and by Carroll *et al.* (2006) for non-linear models.

In general, a measurement error model can be summarized in two parts. Part 1 consists of a classical model, which is most often a regression of the form $p_{Y|X,Z}(Y|X,Z,\beta)$, where Y is the response variable, X and Z are unobservable and observable covariates respectively and β is a parameter of interest. For example, in a linear regression case with normal model error, we could have

$$p_{Y|X,Z}(Y|X,Z,\beta)=\sigma^{-1}\,\phi\left(\frac{Y-\beta_0-X^{\mathrm{T}}\beta_1-Z^{\mathrm{T}}\beta_2}{\sigma}\right),$$

where ϕ is the normal probability density function and $\beta=(\beta_0,\beta_1^{\mathrm{T}},\beta_2^{\mathrm{T}})^{\mathrm{T}}$. In a linear logistic regression we could have

$$p_{Y|X,Z}(Y|X,Z,\beta)=\frac{\exp\{Y(\beta_0+X^{\mathrm{T}}\beta_1+Z^{\mathrm{T}}\beta_2)\}}{1+\exp(\beta_0+X^{\mathrm{T}}\beta_1+Z^{\mathrm{T}}\beta_2)},$$

Address for correspondence: Yanyuan Ma, Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX, 77843-3143, USA.
E-mail: ma@stat.tamu.edu

 1369–7412/07/69429

where $\beta=(\beta_0,\beta_1^{\mathrm{T}},\beta_2^{\mathrm{T}})^{\mathrm{T}}$; and in a linear Poisson regression we could have

$$p_{Y|X,Z}(Y|X,Z,\beta)=\frac{\exp\{-(\beta_0+X^{\mathrm{T}}\beta_1+Z^{\mathrm{T}}\beta_2)\}(\beta_0+X^{\mathrm{T}}\beta_1+Z^{\mathrm{T}}\beta_2)^Y}{Y!},$$

where $\beta=(\beta_0,\beta_1^{\mathrm{T}},\beta_2^{\mathrm{T}})^{\mathrm{T}}$. Traditionally, except for some special cases, e.g. linear regression models (Bickel and Ritov, 1987), polynomial regression models (Chan and Mak, 1985; Cheng and Schneeweiss, 1998), partially linear models (Liang *et al.*, 1999) and models containing an unspecified function of the observable covariate Z (Ma and Carroll, 2006), most research has focused on parametric models. There, β is the only unknown parameter and is finite dimensional. To keep this paper focused we shall also consider parametric models, although the method is applicable in a larger class.

Part 2 of a measurement error model consists of a model describing the relationship between a surrogate variable W and the unobservable covariate X. That is, instead of observing X, a measurement is made of a variable W, and the relationship between X, Z and W is described by $p_{W|X,Z}(W|X,Z)$. In addition, the surrogacy assumption implies that

$$p_{W,Y|X,Z}(W,Y|X,Z)=p_{W|X,Z}(W|X,Z)\;p_{Y|X,Z}(Y|X,Z).$$

Typically, conditional on X, W is independent of Z, and it is usually desired that the measurement error be additive and symmetric, in particular $W=X+U$. In practice, various transformations have been proposed to achieve the additivity and symmetry; see Nusser *et al.* (1996) and Eckert *et al.* (1997). Hence, modelling either the observed data or the transformed data, we have

$$p_{W|X,Z}(W|X,Z)=p_U(W-X),$$

where p_U is symmetric. This is the model that is considered in the present paper.

The likelihood of a single observation (W,Y,Z) is therefore

$$L(\beta,W,Y,Z)=\int p_{Y|X,Z}(Y|X,Z,\beta)\;p_U(W-X)\;p_{X,Z}(X,Z)\;\mathrm{d}\mu(X),$$

where $\mu(X)$ is the probability measure that is associated with X. In the functional measurement error model setting with which we work, no distributional assumption is made about $p_{X,Z}(X,Z)$. However, the identifiability of β requires that a model for p_U be assumed. The most frequently seen model for p_U is normal with mean 0 and either a known or unknown variance–covariance matrix Ω_U. When Ω_U is not known, the likelihood becomes

$$L(\beta,\Omega_U,W_i,Y_i,Z_i)=\int p_{Y|X,Z}(Y|X,Z,\beta)\;p_U(W-X,\Omega_U)\;p_{X,Z}(X,Z)\;\mathrm{d}\mu(X),$$

and Ω_U is estimated along with β.

The model for p_U is often adopted on grounds of convenience. Indeed, when X is not observed and, instead, only a single surrogate W is available, it is impossible to justify from the data what model for p_U is reasonable in a particular problem. Even when additional information is available for estimating p_U, the tendency is always to adopt an approximate p_U in a certain parametric family. In other words, the part 2 model of a measurement error problem is often subjective and susceptible to misspecification. Obviously, a misspecified model for the measurement error can lead to bias in estimating β.

The purpose of this paper is to incorporate an objective part 2 model in the functional measurement error problems when additional information is available. Such information could be additional instrumental variables, or simply multiple measurements. For example, in the case of a single unobservable X_i, multiple measurements $W_{i1}, \ldots, W_{im}$ might be available. Often, instruments or multiple measurements are obtained to estimate the error variance Ω_U in a parametric model for the measurement error. Here, we propose utilizing the multiple observations, to obtain instead a non-parametric kernel estimator of p_U. Because the resulting estimator $\hat{p}_U$ does not converge at a typical parametric rate, a very different treatment from that in the parametric case is required to achieve semiparametric efficiency in estimation of β. In fact, although the theoretical form of the semiparametric efficient estimator can be derived, its computational complexity makes it difficult to implement.

Therefore we propose a plug-in estimator, i.e. we treat $\hat{p}_U$ as a known distribution in the part 2 model, and proceed with an existing estimator. Besides estimating p_U by using kernel methods, it is possible to use deconvolution. However, a disadvantage of deconvolution is a potentially slow rate of convergence; see, for example, work of Carroll and Hall (1988), Stefanski and Carroll (1990) and Fan (1991).

Because the part 2 model no longer has an explicit form, computational advantages that rely on the specific form of normality are lost. For example, estimators in generalized linear models that take advantage of the existence of a complete and sufficient statistic, as studied by Stefanski and Carroll (1985, 1987) and Ma and Tsiatis (2006), become much more computationally challenging. General regular asymptotically linear estimators still apply, although the asymptotic properties of these estimators are different. In fact, owing to the relationship between the regular asymptotically linear estimator and p_U, general results on plug-in estimation cannot be applied straightforwardly, and so an independent asymptotic analysis must be carried out. We emphasize that our work is in the setting of a functional model and, in particular, no distributional assumption is made about $p_{X,Z}(X, Z)$.

As far as we are aware, this is the first attempt at relaxing parametric assumptions about the error distribution in general functional measurement error problems. Previous work in functional models concerns only estimation and inference in the absence of the distributional assumption of unobserved variables. In such contexts, Stefanski and Carroll (1985, 1987) derived consistent estimators in generalized linear models; Tsiatis and Ma (2004) presented consistent and locally efficient estimators in a general model; approximately consistent estimators were given by Cook and Stefanski (1995), via simulation–extrapolation, and by Carroll and Stefanski (1990) and Gleser (1990), using regression calibration. Carroll and Wand (1991) considered relaxing the parametric error distribution assumption in a special logistic model with validation data, where they used non-parametric regression to estimate a quantity that depends directly on the error distribution. In contrast, our method requires no distributional assumption for the measurement error, and, at the same time, no distributional assumption on the latent variable. In addition, the method yields an asymptotically root n consistent estimator for completely general measurement error model problems.

The rest of the paper is organized as follows. We derive the locally efficient semiparametric estimator and describe its complexity of implementation in Section 2. Section 3 describes the plug-in estimator in a very simple hypothetical situation where measurement errors are directly observable. In Section 4 the estimator is generalized to practical situations where multiple measurements are available. Simulation results are given in Section 5, and in Section 6 we apply our method in a real data set. Asymptotic properties of the plug-in estimator are demonstrated in Section 7, and concluding remarks are given in Section 8. Some technical details are collected in Appendix A; others are available from the second author.

2. Locally efficient estimator

Consider a hypothetical situation, where we observe $V_1, \ldots, V_N$ which have the measurement error distribution with density p_U. For the case that is relevant to this paper, we shall assume that $N = n$, although this is not required for the method. Thus, combining with the observed (W_i, Y_i, Z_i)s, we assume that we have independent and identically distributed observations (W_i, Y_i, Z_i, V_i), where $W_i = X_i + U_i$, and U_i and V_i have the same distribution for $i = 1, \ldots, n$. In addition, we assume that V_i is independent of Y_i, U_i, X_i and Z_i. We also suppose that the density $p_U(\cdot)$ is symmetric around zero.

For notational simplicity we drop the subscript i. The probability density function of a single observation (W, Y, Z, V) is given by

$$p_{W,Y,Z,V}(W, Y, Z, V, \beta, p_U, p_Z, p_{X|Z}) = \int p_Z(Z)\, p_{X|Z}(X|Z)\, p_U(W - X) \times p_{Y|X,Z}(Y|X, Z, \beta)\, \mathrm{d}\mu(X)\, p_U(V).$$

Recall that β is the parameter of interest, and that p_U, p_Z and $p_{X|Z}$ are the three infinite dimensional nuisance parameters. In what follows, we treat the semiparametric model by using a geometric approach, described by Bickel *et al.* (1993). We consider a Hilbert space $\mathcal{H}$ that consists of all the q-dimensional mean 0 functions of (W, Y, Z, V), with the inner product defined by covariance. Here, q denotes the dimension of β. Two subspaces of $\mathcal{H}$ are considered: the nuisance tangent space Λ and its orthogonal complement $\Lambda^{\perp}$. The definition of Λ requires two preliminary definitions. First, a parametric submodel is defined as a parametric model that is contained in the semiparametric model. In addition, the parametric submodel also contains the truth. Note that this is typically only a concept, since if we can actually find such a parametric submodel it is no longer necessary to adopt a semiparametric model. Say that the parametric submodel is $p_{W,Y,Z,V}(W, Y, Z, V, \beta, \eta)$, where η is a finite dimensional parameter. Second, a nuisance tangent space of a parametric submodel $p_{W,Y,Z,V}(W, Y, Z, V, \beta, \eta)$ is defined as the mean-squared closure of all $A\, \partial[\log\{p_{W,Y,Z,V}(W, Y, Z, V, \beta, \eta)\}]/\partial\eta$, where A is an arbitrary $q \times r$ matrix, and r denotes the dimension of η. The nuisance tangent space of the semiparametric model, Λ, is subsequently defined as the mean-squared closure of the sum of the nuisance tangent spaces of each parametric submodel. The orthogonal complement of Λ in $\mathcal{H}$ is defined as $\Lambda^{\perp}$. The efficient semiparametric estimator is then defined as the solution of the estimating equation

$$\sum_{i=1}^{n} S_{\text{eff}}(W_i, Y_i, Z_i, V_i, \beta) = 0,$$

where S_{eff} is the projection of the score vector S_β onto $\Lambda^{\perp}$: $S_{\text{eff}} = \Pi(S_\beta|\Lambda^{\perp}) = S_\beta - \Pi(S_\beta|\Lambda)$. Here, the score vector is given by

$$S_\beta = \frac{\partial[\log\{p_{W,Y,Z,V}(W, Y, Z, V, \beta, p_U, p_Z, p_{X|Z})\}]}{\partial\beta}.$$

Following this description, we derive S_{eff} in Appendix A and obtain the semiparametric efficient estimator to be the solution of

$$\sum_{i=1}^{n} [S_1(w_i, y_i, z_i) - E\{f(W - X)|w_i, y_i, z_i\} - f(v_i)] = 0. \tag{1}$$

Here,

$$S_1 = S_\beta(W, Y, Z) - E\{a(X, Z)|W, Y, Z\},$$

$a(X,Z)$ satisfies

$$E(S_\beta|X,Z)=E[E\{a(X,Z)|W,Y,Z\}],$$

and f satisfies

$$\begin{aligned} E[E\{f(W-X)|W,Y,Z\}|X,Z]&=0,\\ E[S_1-E\{f(W-X)|W,Y,Z\}|W-X=u]-f(u)&=0. \end{aligned} \tag{2}$$

Calculation of the expectations in equation (1) requires a known form of $p_{X|Z}$ and p_U. A rich parametric model can be used for p_U. As long as we can estimate p_U at a faster rate than $n^{1/4}$, i.e. $\|p_U-\hat{p}_U\|=o_p(n^{-1/4})$, we can use $\hat{p}_U$ in equation (1) and achieve the same asymptotic estimation variance for β (see Newey (1990)). Often, non-parametric estimation of p_U at the rate $o_p(n^{-1/4})$ will guarantee this property as well.

The situation for $p_{X|Z}$ is even more optimistic; since $E(S_{\text{eff}}|X,Z)=0$, we can posit an arbitrary density $p_{X|Z}$ and proceed with the calculation in equation (1). If the posited model $p^*_{X|Z}$ happens to be correct, we obtain the efficient estimator; in other cases we still obtain a consistent estimator. Hence, the estimator is the so-called locally efficient estimator.

Implementing the efficient estimator in equation (1) also requires obtaining S_1 and f. The main method for calculating S_1 is by solving for a from the integral equation before expression (2). The integral equation can be solved by using, for example, discretization, which converts an integral equation problem into a linear algebra problem. Although the same method could be used to obtain f, the calculation there is very complex because of the multiple integrations that are involved. In addition, typically, the purpose of using discretization is so that a small number of discretization points will suffice for the calculation, whereas here at least the values of f at each observed v_i will be needed to implement the estimation equation (1). This will complicate the numerical treatment significantly. Certainly other methods are possible, e.g. approximating f by using a linear combination of truncated basis functions. However, expression (2) indicates that f relies on S_1, which itself is approximated; hence the resulting numerical error may overwhelm the purported efficiency. Thus, instead of the locally efficient estimator, we propose a simpler plug-in estimator.

3. Plug-in estimator

The estimator that we propose is derived in two steps. First, we estimate p_U non-parametrically, using the observations $V_1,\ldots,V_n$. Specifically, we employ a standard kernel estimator to obtain

$$\hat{p}_U(u)=n^{-1}\sum_{i=1}^{n}K_h(V_i-u),$$

where K is a kernel function and $K_h(u)=h^{-1}K(u/h)$. Next, we implement the estimator that was developed by Tsiatis and Ma (2004) to construct a plug-in estimator, where the p_U that is required by the original estimator is replaced by the estimator $\hat{p}_U$. Note that this is a typical plug-in estimator. Although the construction of our estimator is indeed a special case of general plug-in estimators, the property that was derived by Tsiatis and Ma (2004) does not apply here.

The implementational details of the estimator are as follows.

(a) Estimate $p_U(u)$ by using

$$n^{-1}\sum_{i=1}^{n}K_h(V_i-u),$$

and denote the result by $\hat{p}_U(u)$.

(b) Posit a model for $p_{X|Z}(x|z)$, and denote it by $p^*_{X|Z}(x|z)$.
(c) Solve for $\hat{\beta}$ from

$$\sum_{i=1}^{n} \hat{S}^*_{\beta}(w_i, y_i, z_i, \beta) - \hat{E}^*\{\hat{a}(X, Z)|w_i, y_i, z_i\} = 0, \tag{3}$$

where

$$\hat{S}^*_{\beta} = \hat{E}^*\left(\frac{\partial[\log\{p_{Y|X,Z}(Y|X, Z, \beta)\}]}{\partial\beta}\middle| W, Y, Z\right) \tag{4}$$

is the estimated score vector and $\hat{a}(X, Z)$ is a function that satisfies

$$\hat{E}[\hat{E}^*\{\hat{a}(X, Z)|W, Y, Z\}|X, Z] = \hat{E}\{\hat{S}^*_{\beta}(W, Y, Z)|X, Z\}. \tag{5}$$

Here and throughout the paper, a circumflex denotes an operation that is performed under the estimated $\hat{p}_U$, and the superscript asterisk denotes an operation that is calculated under the posited $p^*_{X|Z}$. We solve for $\hat{a}(X, Z)$ from equation (5) through the following device. For each observed value z_i, $i = 1, \ldots, n$, we discretize the function $p^*_{X|Z}(X|z_i)$ and convert the integral equation (5) into a linear system, and then solve for the function $\hat{a}(X, z_i)$ on a set of discretization points of X.

It is worth emphasizing that the plug-in estimator preserves the robustness property of the original estimator in equation (1), which means that the estimators in equation (3) are consistent even if the posited model $p^*_{X|Z}(x|z)$ is wrong. At the same time, the plug-in estimator provides an opportunity for achieving optimality with respect to $p_{X|Z}$, i.e. if the posited model $p^*_{X|Z}(x|z)$ happens to be correct, the resulting estimator will be efficient in dealing with the nuisance parameter $p_{X|Z}(x|z)$. Formula (3) is in the form of a traditional estimating equation; hence the estimation variance can be computed by using the usual sandwich matrix $A^{-1}B(A^{-1})^{\mathrm{T}}$, where A is the average of the first derivative of each term on the left-hand side of equation (3) with respect to β at its estimated value and B is the average of the square of each term on the left-hand side; see Section 7 for a rigorous proof. Here, β enters each term in a complex fashion; hence we recommend calculating the derivatives by using numerical difference approximations.

4. Multiple measurements

Unless instrumental variables are available, it is only hypothetical that we could observe the measurement errors v_i directly. In the absence of any instruments, a more realistic situation is that multiple observations $w_{i1}, \ldots, w_{im}$ are available for each unobservable X_i. For simplicity, we first assume the same replicates m for all the X_is. Although deconvolution is a common method for extracting the information about p_U, we can in fact form W_is and V_is from the W_{ij}s. Note that, contrary to the common notation, W_i here does not represent $(W_{i1}, \ldots, W_{im})^{\mathrm{T}}$. The simplest case is that where $m = 2$. We form $W_i = (W_{i1} + W_{i2})/2$ and $V_i = (W_{i1} - W_{i2})/2$. Because of the symmetry of the measurement error distribution, it can be verified easily that V_i and $U_i = W_i - X_i$ have the same distribution. Thus, the plug-in estimator in Section 3 is directly applicable. More generally, defining $k = \lfloor m/2 \rfloor$, we construct

$$W_i = \sum_{j=1}^{k} \frac{W_{ij}}{2k} + \sum_{j=k+1}^{m} \frac{W_{ij}}{2m - 2k},$$
$$V_i = \sum_{j=1}^{k} \frac{W_{ij}}{2k} - \sum_{j=k+1}^{m} \frac{W_{ij}}{2m - 2k}.$$

Our construction is such that the resulting W_i has the same dimension as X_i and has the smallest error variance among all linear combinations of the W_{ij}s, $j=1,\ldots,m$, that enable the construction of V_i that has the same distribution as $U_i = W_i - X_i$. Keeping the dimension of the W_is low has computational benefits, as the estimator involves calculating multiple integrals with respect to W. A minimum error variance ensures that the information which is contained in the multiple observations is retained maximally. We can certainly choose to manipulate the multiple observations in other ways. As long as the measurement error distributions can be appropriately estimated from data, the plug-in estimator can be implemented.

In practice, it might happen that the number of multiplications for each X_i is different from each other, e.g. that m_i measurements are available for each X_i in a subset $\mathcal{S}_{m_i}$ of size s_{m_i}. In this case, the above constructions of V_i and W_i are specific to each subset where the same multiplication number m_i is shared within the subset. Specifically, in the subset $\mathcal{S}_{m_i}$ with s_{m_i} members, where each member has m_i measurements available, we take $k_i = \lfloor m_i/2 \rfloor$, construct

$$W_i = \sum_{j=1}^{k_i} \frac{W_{ij}}{2k_i} + \sum_{j=k_i+1}^{m_i} \frac{W_{ij}}{2m_i - 2k_i},$$
$$V_i = \sum_{j=1}^{k_i} \frac{W_{ij}}{2k_i} - \sum_{j=k_i+1}^{m_i} \frac{W_{ij}}{2m_i - 2k_i}$$

and estimate p_{U,m_i} by using the constructed V_is. When a subset $\mathcal{S}_1$ of the X_is exists such that only a single measurement W_i is available for each X_i, the deconvolution cannot be spared to estimate the corresponding $p_{U,1}$. The simplest way to carry out the estimation is to use only two measurements for each $X_i \notin \mathcal{S}_1$, and to form

$$\hat{p}_{U,1}(u) = \frac{1}{2\pi} \int \Big[(n-s_1)^{-1} \sum_{i \notin \mathcal{S}_1} \exp\{\mathrm{i}t(W_{i1} - W_{i2})\} \Big]^{1/2} \exp(-\mathrm{i}tu) \, \mathrm{d}t. \tag{6}$$

As a result, the estimation procedure in Section 3 will be changed to the following steps.

(a) For $m_i \geqslant 1$, estimate $p_{U,m_i}(u)$ by using

$$s_{m_i}^{-1} \sum_{i \in \mathcal{S}_{m_i}} K_h(V_i - u)$$

if $m_i > 1$, and estimate $p_{U,1}(u)$ by using equation (6) if $m_i = 1$. Denote the result by $\hat{p}_{U,m_i}(u)$.

(b) Posit a model for $p_{X|Z}(x|z)$, and denote it by $p^*_{X|Z}(x|z)$.

(c) Solve for $\hat{\beta}$ from

$$\sum_{m_i \geqslant 1} \sum_{i \in \mathcal{S}_{m_i}} \hat{S}^*_{\beta}(w_i, y_i, z_i, \beta) - \hat{E}^*\{\hat{a}(X, Z)|w_i, y_i, z_i\} = 0,$$

where the estimated score vector and the integral equation for $\hat{a}(X_i, Z_i)$ are the same as in equations (4) and (5), except that the expectations are calculated under the corresponding $\hat{p}_{U,m_i}$, where m_i is the number of available measurements for X_i.

5. Simulations

To illustrate the method proposed we carried out a small simulation study. In the first simulation, we generated data $\{(W_i, Y_i), i = 1, \ldots, n\}$ from a quadratic logistic model

$$\operatorname{logit}\{p(Y_i = 1|X_i)\} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2,$$

with measurement error $W_i = X_i + U_i$, $U_i \sim N(0, \sigma^2)$. We set $\sigma = 0.4$, which represents a significant amount of measurement error. Here, the underlying X_is are generated from a normal distribution with mean -1 and variance 1, and the values for β are $\beta_0 = -1$, $\beta_1 = 1$ and $\beta_2 = 1$. In addition, we also generated $\{V_i, i = 1, \ldots, n\}$ from the same distribution of U_i. In implementing the method we adopted two different models for p_X^*: a normal model, which corresponds to the truth, and a uniform model on $[-4, 2]$. To study the performance of the method proposed we adopted three different models for p_U: the normal distribution, corresponding to the truth, the non-parametrically estimated distribution and a Laplace distribution, corresponding to a misspecified model. In implementing the Laplace distribution, we chose the parameter so that the 90% quantile matched the truth. Sample sizes $n = 500$ and $n = 1000$ were used in the simulation, and 1000 data sets were generated. The results of the simulation are included in Table 1. Note that the reported 95% confidence interval was constructed by using ± 1.96 times the estimated standard error.

The second simulation was almost identical to the first, except that the true measurement error for the data was generated from a Laplace distribution with standard error 0.4. Hence, in this study, the Laplace measurement error distribution model was the truth, whereas the normal model was misspecified. The results of the simulation are included in Table 2.

From these results it is clear that posing an improper model for the measurement error structure can lead to biasedness, whereas estimating this distribution provides a protection against biasedness. The consistency of the plug-in estimator proposed is preserved, whether or not the model for p_X is correctly specified.

In these examples, the efficiency of the plug-in estimator is almost as good as it would be if we had known the error distribution precisely. Note that, owing to our inability to calculate the

Table 1. Simulation 1: true measurement error distribution is normal†

p_U	*Parameter*	p_X^*	*Results for n = 500*				*Results for n = 1000*			
			Mean	*estvar*	*empvar*	*95% interval*	*Mean*	*estvar*	*empvar*	*95% interval*
Normal	$\beta_1(-1)$	Normal	−1.012	0.031	0.032	0.950	−1.007	0.015	0.015	0.958
		Uniform	−1.012	0.031	0.033	0.953	−1.007	0.015	0.015	0.958
	$\beta_2(1)$	Normal	1.057	0.116	0.117	0.949	1.029	0.051	0.052	0.944
		Uniform	1.059	0.125	0.121	0.955	1.030	0.053	0.052	0.951
	$\beta_3(1)$	Normal	1.048	0.045	0.045	0.959	1.026	0.020	0.020	0.953
		Uniform	1.049	0.050	0.046	0.970	1.026	0.021	0.020	0.956
Non-parametric	$\beta_1(-1)$	Normal	−1.013	0.032	0.033	0.949	−1.008	0.015	0.015	0.959
		Uniform	−0.997	0.032	0.034	0.951	−0.995	0.015	0.015	0.952
	$\beta_2(1)$	Normal	1.061	0.122	0.124	0.950	1.030	0.052	0.053	0.944
		Uniform	1.089	0.135	0.131	0.959	1.051	0.056	0.055	0.954
	$\beta_3(1)$	Normal	1.051	0.048	0.048	0.956	1.027	0.020	0.021	0.953
		Uniform	1.070	0.054	0.050	0.970	1.041	0.022	0.021	0.957
Laplace	$\beta_1(-1)$	Normal	−1.008	0.035	0.036	0.953	−1.005	0.017	0.016	0.955
		Uniform	−1.007	0.034	0.037	0.945	−1.005	0.016	0.017	0.950
	$\beta_2(1)$	Normal	1.217	0.156	0.158	0.953	1.182	0.067	0.066	0.937
		Uniform	1.260	0.173	0.164	0.959	1.224	0.075	0.070	0.921
	$\beta_3(1)$	Normal	1.158	0.063	0.064	0.958	1.131	0.026	0.026	0.935
		Uniform	1.186	0.070	0.064	0.956	1.159	0.030	0.027	0.923

†Normal, non-parametric and Laplace distributions of p_U are used, and the proposed models of p_X^* are normal and uniform. The mean, average of the estimated variances (estvar), empirical variance (empvar) and 95% confidence interval are reported.

Table 2. Simulation 2: true measurement error distribution is Laplace†

p_U	*Parameter*	p_X^*	*Results for n = 500*				*Results for n = 1000*			
			Mean	*estvar*	*empvar*	*95% interval*	*Mean*	*estvar*	*empvar*	*95% interval*
Laplace	$\beta_1(-1)$	Normal	−1.004	0.031	0.032	0.950	−1.004	0.015	0.016	0.943
		Uniform	−1.009	0.031	0.032	0.948	−1.009	0.015	0.016	0.945
	$\beta_2(1)$	Normal	1.049	0.117	0.115	0.953	1.023	0.054	0.054	0.938
		Uniform	1.040	0.117	0.112	0.951	1.013	0.055	0.053	0.941
	$\beta_3(1)$	Normal	1.036	0.043	0.042	0.953	1.020	0.020	0.020	0.945
		Uniform	1.032	0.045	0.042	0.951	1.015	0.021	0.020	0.945
Non-parametric	$\beta_1(-1)$	Normal	−1.004	0.031	0.032	0.950	−1.004	0.015	0.016	0.944
		Uniform	−1.002	0.030	0.032	0.947	−1.003	0.015	0.016	0.942
	$\beta_2(1)$	Normal	1.046	0.114	0.112	0.954	1.022	0.053	0.053	0.941
		Uniform	1.046	0.117	0.113	0.950	1.019	0.055	0.054	0.941
	$\beta_3(1)$	Normal	1.033	0.043	0.042	0.954	1.017	0.020	0.020	0.946
		Uniform	1.033	0.045	0.043	0.954	1.016	0.021	0.020	0.950
Normal	$\beta_1(-1)$	Normal	−0.995	0.028	0.028	0.949	−0.992	0.014	0.014	0.944
		Uniform	−0.995	0.028	0.028	0.945	−0.993	0.014	0.014	0.943
	$\beta_2(1)$	Normal	0.891	0.084	0.089	0.887	0.867	0.040	0.041	0.862
		Uniform	0.894	0.086	0.089	0.890	0.869	0.041	0.041	0.861
	$\beta_3(1)$	Normal	0.920	0.031	0.033	0.873	0.904	0.015	0.015	0.822
		Uniform	0.922	0.032	0.033	0.876	0.905	0.015	0.015	0.832

†Normal, non-parametric and Laplace distributions of p_U are used, and the proposed models of p_X^* are normal and uniform respectively. The mean, average of the estimated variances (estvar), empirical variance (empvar) and 95% confidence interval are reported.

efficient estimator, a direct analysis of the loss of efficiency that is caused by using the plug-in estimator instead of the fully efficient estimator is not available. However, the loss of efficiency is no greater than the difference in efficiency between the plug-in estimator and the estimator under a known, correct model for p_U. In our simulations, the latter difference is hardly detectable. Therefore, we conclude that the loss of efficiency that is caused by using the plug-in estimator is hardly detectable.

As is common in semiparametric modelling, the estimates are rather insensitive to the bandwidth that is used for non-parametric estimation of p_U. In our experiments we tried various bandwidths, and the resulting estimates, as well as the variance, changed only slightly. For example, in the first simulation study where the true p_U was normal, we employed bandwidths in the range from 0.144 to 0.289 for $n=500$ and from 0.126 to 0.251 for $n=1000$. Using bandwidths in this range, $\hat{\beta}$ changed by less than 2.5% for $n=500$, and by less than 2% for $n=1000$. This is explained by the asymptotic result that is summarized in theorem 2. In practice, various bandwidth selection procedures can be used. For example, we could use cross-validation to obtain an optimal bandwidth $\tilde{h}$ for estimating p_U, and then scale it to obtain $h=\tilde{h}n^{-1/10}$ for final implementation.

6. Example

In the setting of the Framingham heart study data (Kannel *et al.*, 1986) we consider a logistic regression of coronary heart disease Y on the true long-term average of systolic blood pressure T, age Z_1, smoking status Z_2 and serum cholesterol Z_3:

$$\mathrm{pr}(Y=1|T, Z_1, Z_2, Z_3) = H(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 T),$$

where $H(\cdot)$ is the logistic link function. Of course, in actuality the true long-term average of systolic blood pressure T is not observable, and instead we observe T_{M}, measured blood pressure. As described by Eckert *et al.* (1997), it is reasonable to assume a model

$$\frac{\log(T_M - 50)^\theta - 1}{\theta} = \frac{\log(T - 50)^\theta - 1}{\theta} + U,$$

where $\theta = 1.726$ and U is symmetrically distributed with mean 0. Hence, we leave the distribution for U unspecified and estimate p_U non-parametrically.

In the Framingham data set, two measurements, T_{M1} and T_{M2}, are available for each unobservable T. Thus, we form

$$V = \{\log(T_{\mathrm{M1}} - 50)^\theta - 1\}/\theta - \tfrac{1}{2}\{\log(T_{\mathrm{M1}} - 50)^\theta - 1\}/\theta$$

for estimating the distribution of the measurement error and

$$W = \{\log(T_{\mathrm{M1}} - 50)^\theta - 1\}/\theta + \tfrac{1}{2}\{\log(T_{\mathrm{M1}} - 50)^\theta - 1\}/\theta$$

as a new measurement. Reparameterizing $X = \{\log(T - 50)^\theta - 1\}/\theta$, the model becomes

$$\mathrm{pr}(Y=1|X, Z_1, Z_2, Z_3) = H(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4[\exp\{(\theta X + 1)^{1/\theta}\} + 50]),$$
$$W = X + U,$$

where $v_1, \ldots, v_n$ can be used to estimate p_U.

The problem is now in a format that can be treated by using the method proposed. For comparison, we also carried out estimation under the assumption that the measurement error distribution is normal. Results are presented in Table 3.

Differences between estimates under the normal distribution measurement error assumption and those derived by using a non-parametrically estimated distribution are detectable although quite small. From Fig. 1, it is clear that the normal model fits the data v_i reasonably well, and in particular that the non-parametric density estimate is close to the estimated normal density. This explains the similarities that we see in the results. (Eckert *et al.* (1997) noted that U does not have exactly a normal distribution.) Considering the similarity of the density curves for the v_is, and the difference that is obtained in the estimation of β, it is perhaps surprising that the distribution of p_U plays a rather important role in estimating β.

Table 3. Framingham heart study†

Model		β_0	β_1	β_2	β_3	β_4
Naïve	Estimate	−9.1846	0.0531	0.7084	0.0063	0.0167
Normal	Estimate	−9.5868	0.0504	0.7275	0.0062	0.0208
	Standard error	0.8356	0.0104	0.2957	0.0019	0.0046
Non-parametric	Estimate	−9.7585	0.0523	0.7170	0.0062	0.0213
	Standard error	0.8497	0.0103	0.2982	0.0019	0.0047

†The estimates and their standard errors of β, under a normal measurement error distribution model and a non-parametrically estimated measurement error distribution.

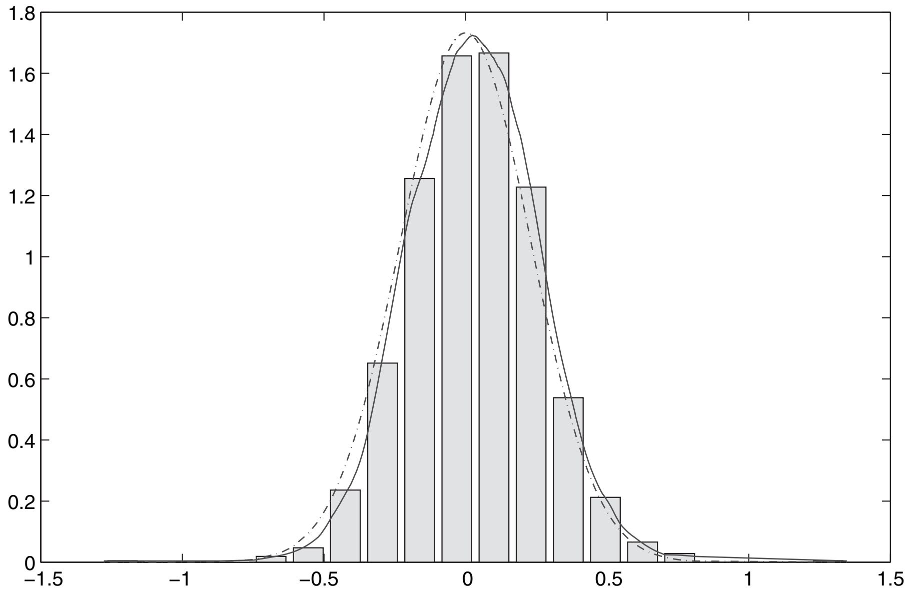

Fig. 1. Framingham heart study: histogram, normal density (· – · – ·) and non-parametrically estimated probability density function (———) for V from the Framingham data

7. Asymptotic properties

Before analysing the asymptotic properties of the estimator proposed, we first justify the need to derive a 'correct' model for p_U. Theorem 1 states that a misspecified model for p_U generally leads to biasedness of the estimator of β.

Theorem 1. Assume that a misspecified model p_U^{m} is adopted to construct an estimator of the form $\Sigma_{i=1}^n S_1(W_i, Y_i, Z_i) = 0$, which is consistent under p_U^{m}. Then the estimator is generally biased.

The proof of theorem 1 is given in Appendix A.

The estimating equation (3) builds on the orthogonal projection of the score vector S_β^*, calculated under the posited $p_{X|Z}^*$, onto the nuisance tangent space with respect to $p_{X|Z}^*$, which is denoted by Λ_1^*. The other nuisance parameter p_U is estimated and 'plugged in'. More detailed explanation of these concepts is given in Section 2. Although it can be true that S_β^*, Λ_1^* and the orthogonal projection operator Π depend continuously on the error distribution p_U, and so the plug-in estimator is in general a continuous function of p_U, the estimator may not be smooth.

In fact, since calculation of the projection involves an inverse problem step in terms of solving for $\hat{a}(X, Z)$, there is no evidence that the dependence of the estimator on p_U is also differentiable. This possible lack of smoothness invalidates the direct application of general plug-in results. However, under mild regularity conditions, root n consistency and asymptotic normality of $\hat{\beta}$ still hold. We summarize these asymptotic results in theorem 2.

Theorem 2. Let $M = E\{\partial(S_1^* - b_1)/\partial\beta\}^{-1}$ and $\Sigma = \mathrm{var}(S_1^* - b_1)$. Under the regularity conditions, when $n \to \infty$ the estimator $\hat{\beta}$ that is given in equation (3) satisfies

$$n^{1/2}(\hat{\beta} - \beta_0) \to N(0, M\Sigma M^{\mathrm{T}})$$

in distribution, provided that h satisfies $nh^2 \to \infty$ and $nh^4 \to 0$. Here,

$$S_1^*(W,Y,Z) = S_\beta^*(W,Y,Z) - E^*\{a(X,Z)|W,Y,Z\},$$

where $a(X,Z)$ satisfies

$$E(S_\beta^*|X,Z) = E[E^*\{a(X,Z)|W,Y,Z\}|X,Z],$$

i.e. S_1^* is the locally efficient score function of the Tsiatis–Ma estimator, $b_1(W,Y,Z) = E\{\hat{a}(X,Z) - a(X,Z)|W,Y,Z\}$, and $\hat{a}(X,Z)$ satisfies equation (5).

Theorem 2 indicates that the estimator of β is still root n consistent, even when non-parametric estimation of p_U is involved. The regularity conditions in the theorem are mild, in that they concern primarily the second-order differentiability and boundedness of several functions and variances.

A proof of theorem 2 is rather tedious and lengthy, and so we present only a sketch in Appendix A. A complete proof, along with a list of regularity conditions, is available from the second author.

The plug-in estimator has the advantage of being computationally simple. However, reduced asymptotic efficiency is a price to be paid for computational simplicity. In what follows, we first give an intuitive explanation of why the estimator is inefficient, and then we provide a formal proof of inefficiency of the estimator.

Consider a slightly different situation, where we have a parametric model for p_U, say $p_U(U,\gamma)$, and we use the same plug-in estimator, i.e. we first estimate γ by using data $v_1, \ldots, v_n$, then plug $p_U(U,\hat{\gamma})$ into the estimator for parametric model p_U to obtain the final estimator. Thus, the two estimating equations are

$$\begin{aligned} \sum_{i=1}^n \tilde{S}_1^*(w_i, y_i, z_i, \beta, \gamma) &= 0, \\ \sum_{i=1}^n S_v(v_i, \gamma) &= 0. \end{aligned} \tag{7}$$

Here, S_v is a mean 0 function of V. For example, if we use maximum likelihood to estimate γ, then $S_v(V,\gamma) = \partial[\log\{p_U(V,\gamma)\}]/\partial\gamma$; $\tilde{S}_1^* = \tilde{S}_\beta^* - \tilde{E}^*(\tilde{a}|W,Y,Z)$, where

$$\tilde{S}_\beta^* = \tilde{E}^*\left(\frac{\partial[\log\{p_{Y|X,Z}(Y|X,Z,\beta)\}]}{\partial\beta} \middle| W,Y,Z \right),$$

and $\tilde{a}(X,Z,\gamma)$ solves

$$\tilde{E}[\tilde{E}^*\{\tilde{a}(X,Z,\gamma)|W,Y,Z\}|X,Z] = \tilde{E}\{\tilde{S}_\beta^*(W,Y,Z,\gamma)|X,Z\}.$$

Here, tildes denote an operation that is performed using the estimated $p_U(U,\hat{\gamma})$.

Standard calculation shows that the variance of the estimator is of the form $M\Sigma M^{\mathrm{T}}$, where

$$M = \begin{pmatrix} E(S_1^* S_1^{\mathrm{T}}) & E(S_1^* S_2^{\mathrm{T}}) \\ 0 & E(S_v S_v^{\mathrm{T}}) \end{pmatrix}^{-1},$$

$$\Sigma = \begin{pmatrix} E(S_1^* S_1^{*\mathrm{T}}) & 0 \\ 0 & E(S_v^{\mathrm{T}} S_v) \end{pmatrix},$$

and S_1 and $S_2 + S_v$ are the two components of the efficient score function corresponding to β and γ. Note that, for simplicity, we wrote the component corresponding to γ as $S_2 + S_v$. Here S_v

is defined as in expression (7), and S_2+S_v is the efficient score function of γ. The score vector with respect to γ is

$$S_\gamma = E\left(\frac{\partial[\log\{p_U(W-X,\gamma)\}]}{\partial\gamma}\bigg| W,Y,Z\right)+S_v.$$

Using the same approach to obtain S_1, we find that the efficient score vector for γ is $S_\gamma - E(b|W,Y,Z)$, where $b(X,Z)$ solves

$$E[E\{b(X,Z)|W,Y,Z\}|X,Z]=E\{S_\gamma(W,Y,Z)|X,Z\}.$$

Hence, we have $S_2=S_\gamma - E(b|W,Y,Z)-S_v$.

It is worth noting that, because V is independent of X, W, Y and Z, $E(S_v|X,Z)=0$ and $E(S_vS_1^{\mathrm{T}})=0$, $E(S_1^*S_v)=0$ and $E(S_vS_2^{\mathrm{T}})=0$. To assess local efficiency, we consider the case when $p^*_{X|Z}$ is the truth, i.e. $p^*_{X|Z}=p_{X|Z}$; therefore, all the asterisks in the following derivations are eliminated. We are interested only in the variance that is associated with β; hence we consider only the $(1,1)$ block of the $M\Sigma M^{\mathrm{T}}$, which equals

$$E(S_1S_1^{\mathrm{T}})^{-1}+E(S_1S_1^{\mathrm{T}})^{-1}\,E(S_1S_2^{\mathrm{T}})\,E(S_vS_v^{\mathrm{T}})^{-1}\,E(S_1S_2^{\mathrm{T}})^{\mathrm{T}}\,E(S_1S_1^{\mathrm{T}})^{-1}. \tag{8}$$

Since the efficient estimator has its variance equal to the inverse of the variance of the efficient score function,

$$\begin{pmatrix} E(S_1S_1^{\mathrm{T}}) & E(S_1S_2^{\mathrm{T}}) \\ E(S_2S_1^{\mathrm{T}}) & E(S_2S_2^{\mathrm{T}})+E(S_vS_v^{\mathrm{T}}) \end{pmatrix}^{-1},$$

its $(1,1)$ block can be verified to be

$$\begin{aligned} &E(S_1S_1^{\mathrm{T}})^{-1}+E(S_1S_1^{\mathrm{T}})^{-1}\,E(S_1S_2^{\mathrm{T}})\{E(S_2S_2^{\mathrm{T}})+E(S_vS_v^{\mathrm{T}}) \\ &\qquad -E(S_1S_2^{\mathrm{T}})^{\mathrm{T}}\,E(S_1S_1^{\mathrm{T}})^{-1}\,E(S_1S_2^{\mathrm{T}})\}^{-1}\,E(S_1S_2^{\mathrm{T}})^{\mathrm{T}}\,E(S_1S_1^{\mathrm{T}})^{-1}. \end{aligned} \tag{9}$$

A comparison between expressions (8) and (9) reveals that, generally, the plug-in estimator is not locally efficient in this situation. Note also that, when p_U is completely known, the variance of the efficient estimator is $E(S_1S_1^{\mathrm{T}})$, which is strictly smaller than both expression (8) and expression (9). This agrees with the intuition that extra unknown parameters in the model generally cause a loss of efficiency.

The inefficiency of the plug-in estimator also holds for the non-parametric measurement error model. We state this result in the following remark.

Remark 1. The plug-in estimator that was proposed in Section 3 is not necessarily locally semiparametric efficient. That is, even if a correct model is posited for $p_{X|Z}$, i.e. $p^*_{X|Z}=p_{X|Z}$, the resulting estimator is not efficient.

Remark 1 can be shown by using the result that was obtained from the parametric measurement error distribution case; we give an illustration in Appendix A. Although the plug-in estimator is, in theory, not efficient, in practice its performance is very close to that of the optimal estimator; see the simulation results in Section 5. Considering the complexity of implementation of the efficient estimator that was presented in Section 2, we expect that the plug-in estimator will be favoured in practice.

8. Conclusion

We derived a plug-in estimator for measurement error models, when the measurement error distribution needs to be estimated from the data. This estimation can be either parametric or non-parametric, depending on the assumptions. We gave one way of non-parametrically estimating this error distribution, using multiple measurements.

We also established the asymptotic normality and root n consistency of the resulting plug-in estimator. We pointed out that the estimator is not efficient, although in practice the degree of suboptimality is often so small that it is hardly noticeable.

The method was demonstrated in a simulation study as well as for the Framingham data. Note that the method requires the availability of v_is, either directly or through manipulating the observations. Therefore, in practice, one could verify whether a parametric model for the distribution of v_is was applicable. Only when no suitable parametric model is found do we need to resort to the non-parametric estimation procedure.

Implementation of the method is almost identical to that for its parametric counterpart, i.e. when a parametric model p_U is available. The non-parametric density estimation step hardly increases computational complexity. However, computing the locally efficient estimator is substantially more complex. Hence, in practice, the plug-in estimator would be preferred.

Finally, construction of the v_is and w_is in the multiple-measurements case relies on the assumption that the measurement error density p_U is symmetric. Transformations to achieve symmetry have been given by Eckert *et al.* (1997). When no transformation is available, and only a general additive measurement error model can be obtained, p_U is not identifiable on the basis of the multiple measurements w_{ij}. In such cases a subjective model for p_U seems unavoidable.

Acknowledgement

The second author was supported by a grant from the National Cancer Institute (CA74552).

Appendix A

A.1. Characterization of the efficient estimator

It can be verified easily that the nuisance tangent space Λ can be written as $\Lambda = \Lambda_1 + \Lambda_2$, where

$$\Lambda_1 = [E\{f(X, Z)|W, Y, Z\} : E(f|Z) = 0]$$

$$\Lambda_2 = [E\{f(W - X)|W, Y, Z\} + f(V) : E(f) = 0].$$

Hence, the nuisance tangent space orthogonal complement $\Lambda^{\perp}$ can be written as $\Lambda^{\perp} = \Lambda_1^{\perp} \cap \Lambda_2^{\perp}$, where

$$\Lambda_1^{\perp} = \{g(W, Y, Z, V) : E(g|X, Z) = 0\},$$

$$\Lambda_2^{\perp} = \{g(W, Y, Z, V) : E(g|U = u) + E(g|V = u) = 0\}.$$

Using the results of Tsiatis and Ma (2004), we show that the projection of S_β onto $\Lambda_1^{\perp}$ is S_1, i.e. $\Pi(S_\beta|\Lambda_1^{\perp}) = S_1$. The projection of S_β onto $\Lambda^{\perp}$ can be obtained by further projecting S_1 onto $\Lambda^{\perp}$. Since $\Lambda_1^{\perp} \cap \Lambda = \Lambda_1^{\perp} \cap \Lambda_2$, we assume that

$$\Pi(S_1|\Lambda_1^{\perp} \cap \Lambda) = E\{f(W - X)|W, Y, Z\} + f(V),$$

where $E(f) = 0$. Then,

$$S_1 - E\{f(W - X)|W, Y, Z\} - f(V) \in \Lambda^{\perp}.$$

Hence, defining $U = W - X$, we have

$$E[S_1 - E\{f(U)|W, Y, Z\} - f(V)|X, Z] = 0,$$

$$E[S_1 - E\{f(U)|W, Y, Z\} - f(V)|U = u] + E[S_1 - E\{f(U)|W, Y, Z\} - f(V)|V = u] = 0.$$

In view of the independence between V and X, W, Y and Z, and the property $E(S_1|X, Z) = 0$, this simplifies to expression (2). Therefore,

$$S_{\text{eff}} = S_1(W, Y, Z) - E\{f(W-X)|W, Y, Z\} - f(V),$$

where f satisfies the conditions in expression (2).

A.2. Outline proof of theorem 1

Define $U = W - X$, and let p_U^{t} be the true probability density function of U. Under the conditions of theorem 1,

$$\int S_1(X+U, Y, Z)\; p_U^{\mathrm{m}}(U)\; p_{Y,X,Z}(Y, X, Z)\; \mathrm{d}\mu(Y)\; \mu(X)\; \mu(Z)\; \mu(U) = 0.$$

If the estimator is also consistent under the true model p_U^{t}, then

$$\int S_1(X+U, Y, Z)\; p_U^{\mathrm{t}}(U)\; p_{Y,X,Z}(Y, X, Z)\; \mathrm{d}\mu(Y)\; \mu(X)\; \mu(Z)\; \mu(U) = 0.$$

Putting

$$\delta(U) = \int S_1(X+U, Y, Z)\; p_{Y,X,Z}(Y, X, Z)\; \mathrm{d}\mu(Y)\; \mu(X)\; \mu(Z),$$

we have $\int \delta(U)\; p_U^{\mathrm{m}}(U)\; \mathrm{d}\mu(U) = 0$ and $\int \delta(U)\; p_U^{\mathrm{t}}(U)\; \mathrm{d}\mu(U) = 0$. Because no special relationship between p_U^{m} and p_I^{t} is assumed other than that both are symmetric, then the only case in which both equalities hold is when $\delta(U)$ is an odd function. The construction of S_1 does not ensure that $\delta(U)$ is odd, and so the estimator is generally biased.

A.3. Outline of proof of theorem 2

We give only an outline of the proof of theorem 2. The full proof, as well as a list of regularity conditions, can be obtained from the second author.

The proof of theorem 2 splits into seven steps. The first step analyses $\hat{p}_U$:

$$\hat{p}_U(u) - p_U(u) = h^2\; p_U''(u) C + n^{-1} \sum_{i=1}^{n} f(V_i, u, h) + o(h^2),$$

where $E\{f(V, u, h)\} = 0$ and

$$\operatorname{var}\{f(V, u, h)\} = h^{-1}\; p_U(u) D + o(h^{-1}).$$

Let $r(\cdot) = \hat{p}_U(\cdot) - p_U(\cdot)$ and

$$\tilde{r}(\cdot) = n^{-1} \sum_{i=1}^{n} f(V_i, \cdot, h).$$

Then,

$$\hat{S}_\beta^* = S_\beta^* + R_1(W, Y, Z) + O(r^2 + h^2),$$
$$\hat{E}(\hat{S}_\beta^*|X, Z) = E(S_\beta^*|X, Z) + R_2(X, Z) + O(h^2 + r^2),$$

where

$$R_1(W, Y, Z) = \sum_{i=1}^{n} [E^*\{S_\beta^{\mathrm{F}*}\; f(v_i, U, h)/p_U(U)|W, Y, Z\} - S_\beta^*\; E^*\{f(v_i, U, h)/p_U(U)|W, Y, Z\}],$$
$$R_2(X, Z) = \int S_\beta^*\, \tilde{r}(U)\; p(Y|X, Z)\; \mathrm{d}\mu(W)\; \mathrm{d}\mu(Y) + E(R_1|X, Z).$$

Next we study projection onto the nuisance tangent space. Assume that $\hat{a}(X, Z)$ and $a(X, Z)$ satisfy respectively

$$\hat{E}[\hat{E}^*\{\hat{a}(X, Z)|W, Y, Z\}|X, Z] = \hat{E}(\hat{S}_\beta^*|X, Z),$$

$$E[E^*\{a(X, Z)|W, Y, Z\}|X, Z] = E(S_\beta^*|X, Z).$$

Let

$$\hat{a}(X,Z)=a(X,Z)+b_0(X,Z),$$
$$b_1(W,Y,Z)=E^*(b_0|W,Y,Z),$$
$$b_2=E^*\{b_0\,\tilde{r}(U)/p_U(U)|W,Y,Z\}$$

and $b=b_1+b_2$. Then it can be shown that

$$\hat{E}^*\{\hat{a}(X,Z)|W,Y,Z\}=E^*(a|W,Y,Z)+b+B_1(W,Y,Z)+O(h^2+r^2),$$
$$\hat{E}[\hat{E}^*\{\hat{a}(X,Z)|W,Y,Z\}|X,Z]=E\{E^*(a|W,Y,Z)|X,Z\}+E(b|X,Z)+B_2(X,Z)+O(h^2+r^2),$$

where

$$B_1(W,Y,Z)=\sum_{i=1}^{n}[E^*\{a\,f(v_i,U,h)/p_U|W,Y,Z\}-\{E^*(a|W,Y,Z)+b_1\}E^*\{f(v_i,U,h)/p_U|W,Y,Z\}],$$
$$B_2(X,Z)=E(B_1|X,Z)+\int\{E^*(a|W,Y,Z)+b\}\,\tilde{r}(U)\;p(Y|X,Z)\,\mathrm{d}\mu(W)\,\mathrm{d}\mu(Y).$$

Next we address the estimating equation. Define $\hat{S}=\hat{S}^*_\beta-\hat{E}^*(\hat{a}|W,Y,Z)$, $S_1^*=S_\beta^*-E^*(a|W,Y,Z)$ and $S_2^*=B_1-R_1+b$. The estimator is given by

$$n^{-1}\sum_{i=1}^{n}\hat{S}(W_i,Y_i,Z_i,\hat{\beta})=0,$$

where $\hat{S}=S_1^*-S_2^*+O(h^2+r^2)$.

The proof is completed by showing, in succession, that

$$n^{1/2}(\beta^*-\beta_0)=O(n^{1/2}h^2)+O_p\{(n^{1/2}h)^{-1}\}+O_p(h^{3/2}),$$
$$\operatorname{var}(n^{1/2}\hat{\beta})=Q\operatorname{var}\{(S_1^*-S_2^*)(W,Y,Z,\beta_0)\}Q^{\mathrm{T}}+o(1),$$
$$n^{1/2}(\hat{\beta}-\beta_0)\rightarrow N(0,[E\{\partial(S_1^*-b_1)/\partial\beta\}]^{-1}\operatorname{var}(S_1^*-b_1)(E\{\partial(S_1^*-b_1)/\partial\beta\}^{-1})^{\mathrm{T}})$$

where the convergence is in distribution and holds provided that $nh^4\rightarrow 0$ and $nh^2\rightarrow\infty$.

A.4. Outline proof of remark 1

Assume the contrary, i.e. that the estimator is locally semiparametric efficient. Since theorem 2 already establishes root n consistency of the estimator, this means that the estimator is efficient if $p^*_{X|Z}$ is correctly specified. To emphasize that all the relevant calculations are under the true $p_{X|Z}$, in the proof we eliminate all asterisks.

Let $c=E(S_2)$. Borrowing from the proof of theorem 2, we first show that $c=o_p(1)$. We have already proved that $E(S_2)=E(b_0)+o_p(1)$. Now, $b_0=\hat{a}-a$,

$$\hat{E}\{\hat{E}(\hat{a}|W,Y,Z)|X,Z\}=\hat{E}(\hat{S}_\beta|X,Z)$$

and

$$E\{E(a|W,Y,Z)|X,Z\}=E(S_\beta|X,Z),$$

so

$$\begin{aligned}&\hat{E}\{\hat{E}(\hat{a}|W,Y,Z)|X,Z\}-E\{E(a|W,Y,Z)|X,Z\}\\&\quad=[\hat{E}\{\hat{E}(\hat{a}|W,Y,Z)|X,Z\}-E\{E(\hat{a}|W,Y,Z)|X,Z\}]+E\{E(b_0|W,Y,Z)|X,Z\}\\&\quad=E\{E(b_0|W,Y,Z)|X,Z\}+o_p(1)\\&\quad=\hat{E}(\hat{S}_\beta|X,Z)-E(S_\beta|X,Z)\\&\quad=\hat{E}(\hat{S}_\beta-S_\beta|X,Z)+\hat{E}(S_\beta|X,Z)-E(S_\beta|X,Z)=o_p(1).\end{aligned}$$

Hence, $E(b_0)=o_p(1)$, i.e. $c=o_p(1)$.

Thus, theorem 2 implies that the first-order asymptotic properties of the proposed estimator are the same as those for the estimator that is obtained by solving

$$\sum_{i=1}^{n}\{S_1(W_i,Y_i,Z_i,V_i)-S_2(W_i,Y_i,Z_i,V_i)+c\}=0.$$

The result in the proof of theorem 2 also shows that $S_1(W_i,Y_i,Z_i,V_i)-S_2(W_i,Y_i,Z_i,V_i)+c$ is proportional to a valid influence function. Hence, the efficiency results imply that it is proportional to the efficient score function.

Using the results in the efficient estimator derivation, we obtain $\Pi(S_1|\Lambda^{\perp})=d(S_1-S_2+c)$, where d is a constant. This is equivalent to $d(S_1-S_2+c)\in\Lambda^{\perp}$ and $(1-d)S_1+d(S_2-c)\in\Lambda\cap\Lambda_1^{\perp}$. So, we can write

$$(1-d)S_1+d(S_2-c)=E\{f_1(X,Z)|W,Y,Z\}+E\{f_2(W-X)|W,Y,Z\}+f_2(V),$$

where $E(f_1|Z)=0$, $E(f_2)=0$ and

$$E[E\{f_1(X,Z)|W,Y,Z\}+E\{f_2(W-X)|W,Y,Z\}+f_2(V)|X,Z]=0.$$

But S_1 and S_2-c are not functions of V, so $f_2=0$. This means that

$$(1-d)S_1+d(S_2-c)\in\Lambda_1\cap\Lambda_1^{\perp};$$

hence

$$(1-d)S_1+d(S_2-c)=0.$$

We have therefore shown that the efficient score function is S_1, the same as when the measurement error distribution is known; hence the optimal efficiency is the same as well. However, even when the measurement error distribution is known parametrically, we have shown that, generally, the optimal efficiency of estimation of β is less than when the measurement error distribution is completely known. This contradiction means that the plug-in estimator is not necessarily efficient.

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993) *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.
Bickel, P. J. and Ritov, A. J. C. (1987) Efficient estimation in the errors-in-variables model. *Ann. Statist.*, **15**, 513–540.
Carroll, R. J. and Hall, P. (1988) Optimal rates of convergence for deconvolving a density. *J. Am. Statist. Ass.*, **83**, 1184–1186.
Carroll, R. J., Ruppert, A., Stefanski, L. A. and Crainiceanu, C. (2006) *Measurement Error in Nonlinear Models: a Modern Perspective*, 2nd edn. London: CRC Press.
Carroll, R. J. and Stefanski, L. A. (1990) Approximate quasilikelihood estimation in models with surrogate predictors. *J. Am. Statist. Ass.*, **85**, 652–663.
Carroll, R. J. and Wand, M. P. (1991) Semiparametric estimation in logistic measurement error models. *J. R. Statist. Soc.* B, **53**, 573–585.
Chan, L. K. and Mak, T. K. (1985) On the polynomial functional relationship. *J. R. Statist. Soc.* B, **47**, 510–518.
Cheng, C.-L. and Schneeweiss, H. (1998) Polynomial regression with errors in the variables. *J. R. Statist. Soc.* B, **60**, 189–199.
Cook, J. and Stefanski, L. A. (1995) A simulation extrapolation method for parametric measurement error models. *J. Am. Statist. Ass.*, **89**, 1314–1328.
Eckert, R. S., Carroll, R. J. and Wang, N. (1997) Transformations to additivity in measurement error models. *Biometrics*, **53**, 262–272.
Fan, J. (1991) On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, **19**, 1257–1272.
Fuller, W. A. (1987) *Measurement Error Models*. New York: Wiley.
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* (eds P. J. Brown and W. A. Fuller). Providence: American Mathematics Society.
Kannel, W. B., Neaton, J. D., Wentworth, D., Thomas, H. E., Stamler, J., Hulley, S. B. and Kjelsberg, M. O. (1986) Overall and coronary heart disease mortality rates in relation to major risk factors in 325,348 men screened for MRFIT. *Am. Hrt J.*, **112**, 825–836.

Liang, H., Härdle, W. and Carroll, R. J. (1999) Estimation in a semiparametric partially linear errors-in-variables model. *Ann. Statist.*, **27**, 1519–1535.
Ma, Y. and Carroll, R. J. (2006) Locally efficient estimators for semiparametric models with measurement error. *J. Am. Statist. Ass.*, **101**, 1465–1474.
Ma, Y. and Tsiatis, A. A. (2006) On closed form semiparametric estimators for measurement error models. *Statist. Sin.*, **16**, 183–193.
Newey, W. K. (1990) Semiparametric efficiency bounds. *J. Appl. Econometr.*, **5**, 99–135.
Nusser, S. M., Carriquiry, A. L., Dodd, K. W. and Fuller, W. A. (1996) A semiparametric transformation approach to estimating usual daily intake distributions. *J. Am. Statist. Ass.*, **91**, 1440–1449.
Stefanski, L. A. and Carroll, R. J. (1985) Covariate measurement error in logistic regression. *Ann. Statist.*, **13**, 1335–1351.
Stefanski, L. A. and Carroll, R. J. (1987) Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika*, **74**, 703–716.
Stefanski, L. A. and Carroll, R. J. (1990) Deconvoluting kernel density estimators. *Statistics*, **21**, 169–184.
Tsiatis, A. A. and Ma, Y. (2004) Locally efficient semiparametric estimators for functional measurement error models. *Biometrika*, **91**, 835–848.