Journal of Statistical Research *ISSN 0256 - 422 X*
2012, Vol. 46, No. 2, pp. 75-90

ROBUST REGRESSION ESTIMATOR FOR A SEMIPARAMETRIC MEASUREMENT ERROR MODEL WITH MULTIPLE COVARIATES USING MONTE-CARLO METHODS

ANASTASIOS A. TSIATIS

Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203
Email: tsiatis@ncsu.edu

SUMMARY

Previous methods for deriving a locally efficient semiparametric estimator for the parameters in a regression model when some of the covariates are measured with error and no additional assumptions are made on the distribution of covariates, i.e., the so called functional measurement error model, involved solving a difficult ill-posed integral equation which limits the utility of these methods to problems with only a few covariates. In this paper we propose using the Landweber-Fridman regularization scheme for approximating the solution to the integral equation. We show how Monte-Carlo methods can be used to estimate the elements in the Landweber-Fridman regularization algorithm and how stochastic approximation can be implemented together with the Monte-Carlo methods to find the locally efficient estimator. This methodology allows the application of the semiparametric theory to problems that were previously infeasible.

Keywords and phrases: Functional measurement error model; Ill-posed integral equation; Landweber-Fridman regularization; Locally efficient semiparametric estimator; Stochastic approximation.

1 Introduction

Consider the problem of estimating the parameter β in a parametric regression model where the conditional distribution of the response variable Y given covariates X and Z is given by the model

$$p_{Y|X,Z}(y|x,z;\beta), \tag{1.1}$$

and β is, say q-dimensional. With a sample of complete data $(Y_i, X_i, Z_i),\ i = 1, \dots, n$, assumed independent and identically distributed, a consistent, asymptotically normal, and efficient estimator for β can be obtained readily using the standard maximum likelihood estimator. In measurement error problems, the variable X is not available, but rather, a surrogate W for X is. For example, the covariate X may be measured with error in which case we would observe $W = X + \epsilon$, where ϵ denotes the measurement error. We will assume that the covariate Z is measured without error and available for analysis. For the purpose of this discussion we will assume that the distribution of W given X and Z is known to us with density

$$p_{W|X,Z}(w|x,z),$$

however, this assumption can be weakened to allow this density to be defined through a model with a finite number of unknown parameters. For example, in a classical measurement error model it is often assumed that the measurement error ϵ follows a normal distribution with mean zero and variance σ_e^2 independent of X and Z, in which case, the conditional distribution of W given X and Z would be a normal distribution with mean X and variance σ_e^2. The measurement error variance may be known in some cases or may be left as an unknown parameter which could be estimated if we had replicate measurements of W.

In addition, we make the usual surrogacy assumption that $Y \perp\!\!\!\perp W|(X,Z)$, where "$\perp\!\!\!\perp$" denotes independence or conditional independence. In our previous illustration, the surrogacy assumption would hold if the measurement error ϵ was independent of (Y,X,Z). Specifically, the surrogacy assumption is given as

$$p_{Y|W,X,Z}(y|w,x,z) = p_{Y|X,Z}(y|x,z). \tag{1.2}$$

In order to be as robust as possible we will not make any additional assumptions regarding the joint distribution of X and Z. Consequently, this is a semiparametric model which is also referred to as a functional measurement error model by Carroll et al. (1995, Chap. 7). The statistical problem is to estimate the parameter β in (1.1) from a sample of independent and identically distributed data $(Y_i, W_i, Z_i),\ i = 1,\ldots,n$ for this semiparametric model. In Tsiatis and Ma (2004), semiparametric theory was used to derive estimators for such models. Specifically, they considered estimating β using the efficient score derived by computing the residual of the score vector with respect to β after projecting it onto the nuisance tangent space. However, in order to derive the projection one needs to posit some distribution for the conditional density of X given Z, which, of course, is unknown to us. Because of the curse of dimensionality, estimating this conditional distribution nonparametrically is infeasible. Rather, a lower dimensional, possibly incorrect, parametric model is posited. Tsiatis and Ma showed that using the resulting efficient score as the basis for an estimating equation resulted in a consistent, asymptotically normal estimator for β whether the posited distribution for X given Z was correctly specified or not. Thus this estimator is a locally efficient semiparametric estimator.

The major difficulty with this method is that deriving the projection onto the nuisance tangent space involves solving an ill-posed integral equation. In the paper by Tsiatis and Ma (2004), the projection was derived by discretizing X which then reduced the problem to solving a finite linear system of equations. This method, however, breaks down quickly as the dimensions of X and Z increase. In this paper we will use Monte-Carlo methods to approximate the elements of the Landweber-Fridman regularization scheme for solving ill-posed integral equations (see Kress 1989, Chap. 15). This methodology will allow us to derive locally efficient semiparametric estimators for β in more complex settings that would not be feasible using discretization.

2 Model Framework and Notation

We begin by introducing the notation and reviewing some of the theory that is necessary to obtain the efficient score. If all the data (Y, X, Z, W) were available, then the score vector with respect to β is equal to $S^*_\beta(Y, X, Z, W; \beta)$, where

$$S^*_\beta(y, x, z, w; \beta) = \frac{\partial \log p_{Y|X,Z,W}(y|x, z, w; \beta)}{\partial \beta}.$$

Because of the surrogacy assumption (1.2), it follows readily that

$$S^*_\beta(Y, X, Z, W; \beta) = S^*_\beta(Y, X, Z; \beta),$$

where

$$S^*_\beta(y, x, z; \beta) = \frac{\partial \log p_{Y|X,Z}(y|x, z; \beta)}{\partial \beta}.$$

Also, because (Y, W, Z) is a many to one transformation of (Y, X, Z, W), we use the results of Rao (1973, p. 330) to show that the resulting score vector for β with respect to the probability model for (Y, W, Z) is given by

$$S_\beta(Y, W, Z; \beta) = E\{S^*_\beta(Y, X, Z; \beta)|Y, W, Z\}. \tag{2.1}$$

Semiparametric theory, as described by Bickel et al. (1993), is used to derive the efficient score. As such, we consider the Hilbert space $\mathcal{H}$ consisting of all q-dimensional (where q denotes the dimension of the parameter β) mean-zero, finite variance functions of the observed data (Y, W, Z) equipped with the covariance inner product $\langle h_1, h_2 \rangle = E\{h_1^T(Y, W, Z)h_2(Y, W, Z)\}$, where $h_1, h_2 \in \mathcal{H}$ and superscript "^{T}" denotes transpose. The nuisance parameters in the statistical model define the joint density of (X, Z) which is left arbitrary (i.e., a nonparametric model for the joint distribution of (X, Z)). The nuisance tangent space is defined as the mean-square closure of all the nuisance score vectors for parametric submodels, and using arguments in Tsiatis and Ma (2004), the nuisance tangent space Λ is shown to be equal to all q-dimensional, mean-zero measurable functions of (Y, W, Z) (i.e., elements of $\mathcal{H}$) such that

$$\Lambda = \big[E\{\alpha(X, Z)|Y, W, Z\} : E\{\alpha(X, Z)\} = 0\big]. \tag{2.2}$$

We denote the score vector with respect to β by $S_\beta(Y, W, Z) = S_\beta(Y, W, Z; \beta_0)$; i.e., the score vector given by (2.1) evaluated at the true value of β which we denote by β_0. The projection of the score vector with respect to β onto the nuisance tangent space Λ, $\Pi\{S_\beta(Y, W, Z)|\Lambda\}$, where $\Pi\{\cdot|\cdot\}$ denotes the projection of an element onto a linear subspace of the Hilbert space, is defined as the element $E\{\alpha^0(X, Z)|Y, W, Z\} \in \Lambda$, where $E\{\alpha^0(X, Z)\} = 0$, that satisfies the relationship

$$E\Big([S_\beta(Y, W, Z) - E\{\alpha^0(X, Z)|Y, W, Z\}]^T E\{\alpha(X, Z)|Y, W, Z\}\Big) = 0 \tag{2.3}$$

for all $[\alpha(X, Z) : E\{\alpha(X, Z)\} = 0]$.

Remark 1. Because of the projection theorem for Hilbert spaces (Luenberger, 1969, p. 51) together with the fact that Λ is a closed linear subspace we know that $\alpha^0(X,Z)$ must exist and that the projection $E\{\alpha^0(X,Z)|Y,W,Z\}$ satisfying (2.3) must be unique although the function $\alpha^0(X,Z)$ may not be unique. □

In Tsiatis and Ma (2004), the projection was derived using iterated conditional expectation arguments. For this exposition, it will be more natural to derive the projection using linear operators and their adjoint as we now demonstrate.

3 Deriving the Efficient Score using Linear Operators

We first remind the reader of some basic results for linear operators and their adjoint. Let A be some bounded linear operator mapping elements from a Hilbert space $\mathcal{G}$ to a Hilbert space $\mathcal{H}$; i.e., $A : \mathcal{G} \to \mathcal{H}$. The adjoint of a linear operator A is defined as the linear operator $A^* : \mathcal{H} \to \mathcal{G}$ such that for any $g \in \mathcal{G}$ and any $h \in \mathcal{H}$

$$\langle A(g), h\rangle = \langle g, A^*(h)\rangle,$$

where $\langle\cdot,\cdot\rangle$ denotes inner product defined for the corresponding Hilbert space.

For our problem we have already defined the Hilbert space $\mathcal{H}$ and we define the Hilbert space $\mathcal{G}$ as the set of all q-dimensional mean-zero finite variance measurable functions of (X,Z) equipped with the usual covariance inner product. We consider the bounded linear operator $A : \mathcal{G} \to \mathcal{H}$ to be defined as $A\{g(X,Z)\} = E\{g(X,Z)|Y,W,Z\}$ for any $g \in \mathcal{G}$. The adjoint A^* of A is easily shown to be

$$A^*\{h(Y,W,Z)\} = E\{h(Y,W,Z)|X,Z\}$$

for any $h \in \mathcal{H}$. For completeness, we give a proof of this in the appendix.

We note that the linear subspace Λ defined by (2.2) is equal to $A(\mathcal{G})$. To derive the projection onto this linear subspace we use the following theorem given in Luenberger (1969, p. 160).

Theorem 1. For a fixed $h \in \mathcal{H}$, the element $g \in \mathcal{G}$ minimizes $||h - A(g)||$ if and only if

$$A^*A(g) = A^*h,$$

where A^* denotes the adjoint of A and the norm $||h - A(g)|| = \{\langle h - A(g), h - A(g)\rangle\}^{1/2}$. □

As a consequence of Theorem 1, the projection of $S_\beta(Y,W,Z)$, an element of $\mathcal{H}$, onto the closed linear subspace $\Lambda = A(\mathcal{G})$ is the solution to the normal equation

$$A^*A\{\alpha^0(X,Z)\} = A^*\{S_\beta(Y,W,Z)\}, \tag{3.1}$$

or, equivalently,

$$E\Big[E\{\alpha^0(X,Z)|Y,W,Z\}|X,Z\Big] = E\{S_\beta(Y,W,Z)|X,Z\}. \tag{3.2}$$

Equation (3.2) is exactly the solution that was given by Tsiatis and Ma (2004) in their equation (9) for finding the projection.

4 Issues in Deriving the Efficient Score

The definition of the efficient score is given as

$$S_{\text{eff}}(Y, W, Z) = S_\beta(Y, W, Z) - \Pi\{S_\beta(Y, W, Z)|\Lambda\}.$$

As we just demonstrated,

$$S_{\text{eff}}(Y, W, Z) = S_\beta(Y, W, Z) - A\{\alpha^0(X, Z)\},$$

where $\alpha^0(X, Z)$ is a solution to (3.1). Because $S_\beta(Y, W, Z) = E\{S_\beta^*(Y, X, Z)|Y, W, Z\}$, we can also write the efficient score as

$$S_{\text{eff}}(Y, W, Z) = A\{S_\beta^*(Y, X, Z) - \alpha^0(X, Z)\}.$$

Two issues arise immediately when deriving the efficient score:

(i) The solution to equation (3.1) depends on knowing the joint density of X and Z. However, since both A and A^* are conditional expectations which also condition on Z, the solution to equation (3.1) just depends on the conditional distribution of X given Z and not additionally on the marginal distribution of Z. This is important because no additional assumptions are needed regarding the marginal distribution of Z for any of the asymptotic results regarding consistency and asymptotic normality to hold. Nonetheless, the solution to equation (3.1) still depends on knowing the conditional distribution of X given Z which our functional measurement error model leaves unspecified.

(ii) Even if the conditional distribution of X given Z were known, equation (3.1) is an ill-posed integral equation which is difficult to solve.

To address issue (i), we posit a model for the distribution of X given Z; namely, $p_{X|Z}(x|z; \psi)$ in terms of a finite-dimensional parameter ψ, which may be misspecified, and define

$$\begin{aligned} A\{g(Y, X, Z); \beta, \psi\} &= E_{\beta,\psi}\{g(Y, X, Z)|Y, W, Z\}, & (4.1)\\ A^*\{h(Y, W, Z); \beta\} &= E_\beta\{h(Y, W, Z)|X, Z\}, & (4.2)\end{aligned}$$

where the conditional expectation in (4.2) is with respect to the conditional density of (Y, W) given (X, Z), which, by the surrogacy assumption (1.2), is equal to the product of $p_{Y|X,Z}(y|x, z; \beta)$, (i.e., the correctly specified model with the true density being at $\beta = \beta_0$) and $p_{W|X,Z}(w|x, z)$, which is assumed known. The conditional expectation in (4.1) is with respect to the conditional density of X given (Y, W, Z) which is obtained using Baye's rule; namely,

$$p_{X|Y,W,Z}(x|y, w, z; \beta, \psi) = \frac{p_{Y|X,Z}(y|x, z; \beta) p_{W|X,Z}(w|x, z) p_{X|Z}(x|z; \psi)}{\int p_{Y|X,Z}(y|x, z; \beta) p_{W|X,Z}(w|x, z) p_{X|Z}(x|z; \psi) dx}. \quad (4.3)$$

In terms of this notation we define the projected score vector as

$$S_{\text{eff}}(Y, W, Z; \beta, \psi) = A\big[\{S_\beta^*(Y, X, Z; \beta) - \alpha^0(X, Z; \beta, \psi)\}; \beta, \psi\big], \quad (4.4)$$

where $\alpha^0(X, Z; \beta, \psi)$ is the solution to the normal equation

$$A^*\Big[A\{\alpha^0(X, Z; \beta, \psi); \beta, \psi\}; \beta\Big] = A^*\Big[A\{S^*_\beta(Y, X, Z; \beta); \beta, \psi\}; \beta\Big]. \tag{4.5}$$

The estimator for β is obtained as the solution to the estimating equation

$$\sum_{i=1}^{n} S_{\text{eff}}(Y_i, W_i, Z_i; \beta, \hat{\psi}_n) = 0, \tag{4.6}$$

where $\hat{\psi}_n$ is an estimator that is root-n consistent. That is, even if the model for the conditional distribution of X given Z is misspecified, we will assume that there exists a constant ψ^* which is the limit in probability of $\hat{\psi}_n$ such that $n^{1/2}(\hat{\psi}_n - \psi^*)$ is bounded in probability. It was shown in Tsiatis and Ma (2004) that the estimator which solves (4.6) is asymptotically equivalent to the estimator which solves the equation

$$\sum_{i=1}^{n} S_{\text{eff}}(Y_i, W_i, Z_i; \beta, \psi^*) = 0. \tag{4.7}$$

Under suitable regularity conditions, the estimator in (4.7) will be a consistent, asymptotically normal estimator if the estimating function, evaluated at $\beta = \beta_0$, has expectation zero. We will now give a short argument to show that

$$E\{S_{\text{eff}}(Y, W, Z; \beta_0, \psi^*)\} = 0 \tag{4.8}$$

even if the posited model for $p_{X|Z}(x|z; \psi)$ is misspecified. By using the law of iterated conditional expectations, the expectation on the left hand side of (4.8) is also equal to

$$E\Big[A^*\{S_{\text{eff}}(Y, W, Z; \beta_0, \psi^*)\}; \beta_0\Big], \tag{4.9}$$

which is true because A^* involves a conditional expectation with respect to a correctly specified conditional distribution. By the definition of the efficient score given in (4.4), the term inside the brackets of (4.9) is equal to

$$\begin{aligned} &A^*\Big(A\Big[\{S^*_\beta(Y, X, Z; \beta_0) - \alpha^0(X, Z; \beta_0, \psi^*)\}; \beta_0, \psi^*\Big]; \beta_0\Big) = \\ &A^*\Big[A\{S^*_\beta(Y, X, Z; \beta_0); \beta_0, \psi^*\}; \beta_0\Big] - A^*\Big[A\{\alpha^0(X, Z; \beta_0, \psi^*); \beta_0, \psi^*\}; \beta_0\Big], \end{aligned}$$

which equals zero as a consequence of (4.5). Consequently, (4.9) is also equal to zero, which implies that the estimating function is an unbiased estimator of zero and that the resulting estimator for β, given by the solution to (4.7), will be consistent and asymptotically normal even if the model for the conditional distribution of X given Z, in terms of ψ, was misspecified.

Item (ii) regarding the computations involved in solving the integral equation is critical. In the case where X was a single covariate, Tsiatis and Ma (2004) showed that a reasonable solution could be obtained by discretizing X. However, if X is multivariate or with additional Z, the problem becomes prohibitive and other methods for obtaining approximate solutions to the integral equation are necessary to make this approach feasible. Toward that end, we consider the Landweber-Fridman regularization scheme for solving ill-posed integral equations as discussed by Kress (1989, p. 239) which we now describe.

Because the linear operator A is itself a projection operator, this implies that the norm of A is less than or equal to one; i.e., $||A|| \leq 1$, where the norm of a linear operator is defined, say, in Kress (1989, p. 13). For such a case, the Landweber-Fridman regularization scheme would approximate the solution $\alpha^0(X,Z)$ in (3.1) by

$$\begin{aligned} \alpha^0(X,Z) &= \theta\sum_{k=0}^{t}(I-\theta A^*A)^k A^*\{S_\beta(Y,W,Z)\} \\ &= \theta\sum_{k=0}^{t}(I-\theta A^*A)^k A^*A\{S^*_\beta(Y,X,Z)\}, \end{aligned} \tag{4.10}$$

where $\theta < 1$ is a scalar constant and t is an integer that serves as a regularization parameter where the accuracy of the approximation becomes better as t increases but the stability becomes worse. To be clear about the notation

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

$$(A^*A)^2\{S^*_\beta(Y,X,Z)\} = E\{E(E[E\{S^*_\beta(Y,X,Z)|Y,W,Z\}|X,Z]|Y,W,Z)|X,Z\},$$

and so on.

As as consequence of (4.4) and (4.10), we can approximate the estimating equation (4.7) by

$$\sum_{i=1}^{n} M(Y_i,W_i,Z_i;\beta,\psi^*) = 0, \tag{4.11}$$

where

$$\begin{aligned} &M(Y_i,W_i,Z_i;\beta,\psi^*) \\ &= A_i\Big[S^*_\beta(Y,X,Z;\beta) - \theta\sum_{k=0}^{t}(1-\theta A^*A)^k A^*A\{S^*_\beta(Y,X,Z;\beta);\beta,\psi^*\};\beta,\psi^*\Big] \\ &= A_i\Big[S^*_\beta(Y,X,Z;\beta) + \sum_{k=1}^{t+1} c(k,t)\theta^k (A^*A)^k\{S^*_\beta(Y,X,Z;\beta);\beta,\psi^*\};\beta,\psi^*\Big], \end{aligned}$$

$A_i\{g(Y,X,Z);\beta,\psi^*\} = E_{\beta,\psi^*}\{g(Y,X,Z)|Y_i,W_i,Z_i\}$, and

$$c(k,t) = (-1)^k \sum_{\ell=k-1}^{t} [\ell!/\{(k-1)!(\ell-k+1)!\}], k = 1,\ldots,t+1. \tag{4.12}$$

Of course, to implement this methodology we need to evaluate, or, at least approximate $(A^*A)^k A^*\{S_\beta(Y,W,Z)\}$ for different k as well as choose appropriate values for the regularization parameters θ and t.

5 Implementation using Monte-Carlo Methods

Computing terms such as $(A^*A)^k A^*\{S_\beta(Y,W,Z)\}$ involves repeated conditional expectations alternating between conditioning on (X,Z) which involves the conditional density of (Y,W) given (X,Z) and conditioning on (Y,W,Z) which involves the conditional density of X given (Y,W,Z). For that matter even $S_\beta(Y,W,Z) = A\{S_\beta^*(Y,X,Z)\}$ involves computing a conditional expectation. Computing such repeated conditional expectations using numerical integration is infeasible. We therefore propose using Monte-Carlo simulations to approximate these iterated conditional expectations. In order to proceed we need to be able to generate random (Y,W)'s from the conditional density of $p_{Y,W|X,Z}(y,w|x,z;\beta)$ and random X's from the conditional density of $p_{X|Y,W,Z}(x|y,w,z;\beta,\psi)$. Since the latter distribution depends on the conditional density of X given Z, which may be misspecified, we have some flexibility in choosing this conditional density to facilitate the Monte-Carlo data generation. Assuming we can generate random data from these conditional distributions, we will now describe how Monte-Carlo methods can be used to find approximations to the integral equations and the locally efficient estimator. Later we will illustrate in greater detail how these methods can be used when modeling binary data with a logistic regression model.

We begin by showing how to evaluate, for the i-th individual in the sample, the contribution to the observed score vector with respect to β;

$$S_\beta(Y_i,W_i,Z_i;\beta,\psi^*) = E_{\beta,\psi^*}\{S_\beta^*(Y,X,Z;\beta)|Y_i,W_i,Z_i\} = A_i\{S_\beta^*(Y,X,Z;\beta);\beta,\psi^*\}. \quad (5.1)$$

We propose generating m random X's from the conditional distribution of $X|Y = Y_i, W = W_i, Z = Z_i$ and denoting these as $x_1^{(0)},\ldots,x_m^{(0)}$. We then approximate (5.1) by

$$\hat{A}_i\{S_\beta^*(Y,X,Z;\beta);\beta,\psi^*\} = m^{-1}\sum_{j=1}^{m} S_\beta^*(Y_i,x_j^{(0)},Z_i;\beta).$$

We also note that by construction

$$E_{\beta,\psi^*}^*\left[\hat{A}_i\{S_\beta^*(Y,X,Z;\beta);\beta,\psi^*\}|Y_i,W_i,Z_i\right] = S_\beta(Y_i,W_i,Z_i;\beta,\psi^*), \quad (5.2)$$

where, we emphasize by using E^* that, this expectation is taken with respect to the Monte-Carlo data generating distribution. Because the data generating process always begins conditional on the observed data (Y_i,W_i,Z_i), expectations E^* are necessarily conditional expectations with respect to the observed data. In equation (5.2) we made this conditional expectation explicit, but, from here on, we will suppress this notation and it will be assumed that any reference to E^* is a conditional expectation with respect to the observed data. Equation (5.2) is true for any choice of $m \geq 1$.

In order to compute $E\{\alpha^0(X,Z)|Y_i,W_i,Z_i\}$, which is part of the estimating equation (4.7), using the Landweber-Fridman regularization scheme given by (4.10), we need to approximate quantities such as

$$A_i(A^*A)^k A^*\{S_\beta(Y,W,Z;\beta,\psi^*);\beta,\psi^*\} = A_i(A^*A)^{k+1}S_\beta^*(Y,X,Z;\beta)\beta,\psi^*\}, \quad (5.3)$$

for different $k = 0, \ldots, t$. We proceed as follows: As before, we generate $x_j^{(0)}, j = 1, \ldots, m$ from the conditional distribution of $(X|Y = Y_i, W = W_i, Z = Z_i)$. Next, we generate $(y_j^{(1)}, w_j^{(1)}), j = 1, \ldots, m$ from the conditional distribution of $(Y, W|X = x_j^{(0)}, Z = Z_i)$ and $x_j^{(1)}, j = 1, \ldots, m$ from the conditional distribution of $(X|Y = y_j^{(1)}, W = w_j^{(1)}, Z = Z_i)$. We then approximate (5.3), for $k = 0$ by

$$m^{-1} \sum_{j=1}^{m} S_\beta^*(y_j^{(1)}, x_j^{(1)}, Z_i; \beta).$$

Continuing in this iterative fashion, we generate $(y_j^{(\ell)}, w_j^{(\ell)})$ from the conditional distribution of $(Y, W|X = x_j^{(\ell-1)}, Z = Z_i)$ and $x_j^{(\ell)}$ from the conditional distribution of $(X|Y = y_j^{(\ell)}, W = w_j^{(\ell)}, Z = Z_i)$ and approximate (5.3) by

$$m^{-1} \sum_{j=1}^{m} S_\beta^*(y_j^{(k+1)}, x_j^{(k+1)}, Z_i; \beta) \tag{5.4}$$

for $k = 0, \ldots, t$. We again note that the conditional expectation $E_{\beta,\psi*}^*\{(5.4)\}$ is equal to the statistic (5.3).

Recall that our goal is to get an estimator for β by solving the approximating estimating equation (4.11). If we denote by $O = \{O_i = (Y_i, W_i, Z_i), i = 1, \ldots, n\}$ the sample of observed data, then from the development above we know that

$$E_{\beta,\psi^*}^* \left\{ \sum_{i=1}^{n} m^{-1} \sum_{j=1}^{m} \sum_{k=0}^{t+1} c(k,t)\theta^k S_\beta^*(y_{ji}^{(k)}, x_{ji}^{(k)}, Z_i, \beta)|O \right\} = \sum_{i=1}^{n} M(Y_i, W_i, Z_i; \beta, \psi^*),$$

where $c(0, t) = 1$ for all t and $c(k, t)$ for $k \geq 1$ was defined by (4.12). Therefore, we want to find the value β that solves the estimating equation

$$\sum_{i=1}^{n} M(O_i; \beta) = 0.$$

Note that we are suppressing the parameter ψ^* since this remains fixed when solving the estimating equation for β.

Using Monte-Carlo methods we showed how to generate unbiased estimators for $\sum_{i=1}^{n} M(O_i; \beta)$; namely, $\sum_{i=1}^{n} \hat{M}(O_i; \beta)$, where

$$\hat{M}(O_i; \beta) = m^{-1} \sum_{j=1}^{m} \sum_{k=0}^{t+1} c(k,t)\theta^k S_\beta^*(y_{ji}^{(k)}, x_{ji}^{(k)}, Z_i, \beta), \tag{5.5}$$

such that $E_\beta^*\{\sum_{i=1}^{n} \hat{M}(O_i; \beta)\} = \sum_{i=1}^{n} M(O_i; \beta)$. With these Monte-Carlo generated unbiased estimators for the elements in the estimating equation, we will now show how a modified version of stochastic approximation that was proposed by Yin and Wu (1997) can be used to obtain an approximation to the root of the estimating equation $\sum_{i=1}^{n} M(O_i; \beta) = 0$.

Stochastic approximation, first proposed by Robbins and Monro (1951), is a recursive procedure for finding the root of an equation $H(\beta) = 0$, where $E(\hat{H}|\beta) = H(\beta)$, using data $(\hat{H}_\ell, \beta_\ell),\ \ell = 1, 2, \ldots$. For our purposes $H(\beta) = \sum_{i=1}^n M(O_i, \beta)$, $\hat{H} = \sum_{i=1}^n \hat{M}(O_i; \beta)$, which, as we showed above, has the property that $E(\hat{H}|\beta) = E^*_\beta(\hat{H}) = \sum_{i=1}^n M(O_i; \beta)$. Specifically, we will consider two variants of stochastic approximation. At the initial stage we will use a recursive scheme where the next iterate in the recursion is given by

$$\hat{\beta}^{(\ell+1)} = \bar{\beta}^{(\ell)} - J^{-1}\bar{M}^{(\ell)}, \tag{5.6}$$

where $\bar{\beta}^{(\ell)} = \ell^{-1}\sum_{g=1}^{\ell} \hat{\beta}^{(g)}$, $\bar{M}^{(\ell)} = \ell^{-1}\sum_{g=1}^{\ell}\sum_{i=1}^{n} \hat{M}(O_i; \hat{\beta}^{(g)})$, and J is a $q \times q$ matrix which roughly approximates the gradient matrix $\partial H(\beta)/\partial\beta^T$. Under some mild regularity conditions the sequence $\hat{\beta}^{(\ell)}$ will converge to $\hat{\beta}_n$, the solution to the estimating equation $\sum_{i=1}^n M(O_i, \beta) = 0$.

Stochastic approximation is an efficient way of finding the root of the estimating equation even if J is obtained in an ad hoc fashion and not necessarily a "good" (i.e., consistent) estimator of $\partial H(\beta)/\partial\beta^T$. To implement the proposed methodology we recommend using the naive estimator for β which ignores measurement error as the initial value for the stochastic approximation recursion in (5.6). That is, we take $\hat{\beta}^{(1)}$ to be the solution to the naive score equation that would have been used if there was no measurement error; namely,

$$\sum_{i=1}^{n} S^*_\beta(Y_i, W_i, Z_i; \beta) = 0. \tag{5.7}$$

We also recommend that the matrix J in (5.6) be approximated by

$$J = \sum_{i=1}^{n} \frac{\partial S^*_\beta(Y_i, W_i, Z_i; \hat{\beta}^{(1)})}{\partial\beta^T},$$

i.e., minus the naive observed information matrix had there been no measurement error.

However, stochastic approximation as described by (5.6) does not provide for a good estimate of the gradient $\partial H(\beta)/\partial\beta^T$, which, as we will see in the next section, is important for deriving an estimator for the asymptotic variance of $\hat{\beta}_n$. Because the sequence $\hat{\beta}^{(\ell)}$ in (5.6) converges sufficiently fast, there is not enough variation in this sequence to get a good estimator of the derivative. Consequently, we will consider additional perturbation of the β's in the stochastic approximation recursions in order to also obtain reasonable estimators for the derivative of the estimating equation.

Toward that end, we will now consider how to estimate the gradient matrix $\partial H(\beta)/\partial\beta^T$ while introducing the second variant of stochastic approximation. This is based on the premise that the initial algorithm (5.6) will give us a sequence $\hat{\beta}^{(\ell)}$ which will eventually be in a small neighborhood of $\hat{\beta}_n$. Consequently, the function $M(O_i, \beta)$ will be approximately linear in β in a neighborhood about $\hat{\beta}_n$. That is, in a neighborhood of $\hat{\beta}_n$,

$$M(O_i, \beta) \approx F_i^{q\times(q+1)}(1, \beta^T)^T, \tag{5.8}$$

where the elements of the j-th row of the $q \times (q+1)$ matrix F_i are denoted by $f_{0ji}, f_{1ji}, \ldots, f_{qji}$, and the j-th element of $M(O_i, \beta)$ is

$$M_j(O_i, \beta) \approx f_{0ji} + f_{1ji}\beta_1 + \ldots + f_{qji}\beta_q,$$

where β_r denotes the r-th element of β for $r = 1, \ldots, q$.

Because $E^*_\beta\{\hat{M}(O_i, \beta)\} = M(O_i, \beta)$, we propose estimating the matrix F_i by least squares using all the available data $\{\hat{M}(O_i, \hat{\beta}^{(g)} + \epsilon^{(g)}), \hat{\beta}^{(g)} + \epsilon^{(g)}\}$ for $g = 1, \ldots, \ell$, where $\epsilon^{(g)}$ is a q-dimensional vector of additional random perturbations that we may add to obtain a better estimate for the gradient. Specifically, the least squares estimate for F_i after the ℓ-th iteration in the recursion is given by

$$\hat{F}_i^{(\ell)} = \mathcal{M}_i^{(\ell)} B^{(\ell)} (B^{(\ell)T} B^{(\ell)})^{-1},$$

where $\mathcal{M}_i^{(\ell)}$ is a $q \times \ell$ matrix made up of the ℓ q-dimensional column vectors $\hat{M}(O_i, \hat{\beta}^{(g)} + \epsilon^{(g)})$ for $g = 1, \ldots, \ell$ and $B^{(\ell)}$ is the $\ell \times (q+1)$ matrix made of the ℓ $(q+1)$-dimensional row vectors $\{1, (\hat{\beta}^{(g)} + \epsilon^{(g)})^T\}$ for $g = 1, \ldots, \ell$. If we partition the matrix $\hat{F}_i^{(\ell)}$ as $[\hat{F}_i^{(\ell)0} | \hat{F}_i^{(\ell)1}]$, where $\hat{F}_i^{(\ell)0}$ is the first column of $\hat{F}_i^{(\ell)}$ and $\hat{F}_i^{(\ell)1}$ is the $q \times q$ matrix corresponding to the last q columns of $\hat{F}_i^{(\ell)}$, then a reasonable estimator for $\hat{\beta}^{(\ell+1)}$ would be the solution to the linear equation

$$\sum_{i=1}^n (\hat{F}_i^{(\ell)0} + \hat{F}_i^{(\ell)1} \beta) = 0;$$

that is,

$$\hat{\beta}^{(\ell+1)} = -(\sum_{i=1}^n \hat{F}_i^{(\ell)1})^{-1} \sum_{i=1}^n \hat{F}_i^{(\ell)0}. \tag{5.9}$$

Because of (5.8), a natural estimator for the gradient matrix is given by

$$\hat{D}^{(\ell)} = n^{-1} \sum_{i=1}^n \hat{F}_i^{(\ell)1}.$$

Therefore our proposal for stochastic approximation is to use the recursion (5.6) with no additional random perturbation initially until the estimator stabilizes and then to switch to the recursion given by (5.9). At the end of this process, say, after a total of ℓ^* iterations, the final estimator for β would be

$$\hat{\beta}_n = \hat{\beta}^{(\ell^*+1)}, \tag{5.10}$$

and the final estimator for the gradient matrix would be

$$\hat{D} = \hat{D}^{(\ell^*)}. \tag{5.11}$$

6 Estimator for the Asymptotic Variance

Using standard results for M-estimators, see, for example, Stefanski and Boos (2002), we know that, under suitable regularity conditions, the estimator $\hat{\beta}_n$ that solves the estimating equation

$$\sum_{i=1}^n M(O_i, \beta) = 0$$

will be asymptotically normal with mean zero and variance matrix

$$\{D(\beta_0)\}^{-1}V(\beta_0)\{D^T(\beta_0)\}^{-1}, \tag{6.1}$$

where $D(\beta) = E\{\partial M(O,\beta)/\partial\beta^T\}$ and $V(\beta) = E\{M(O,\beta)M^T(O,\beta)\}$.

In the previous section we discussed how to use the results of our stochastic approximation to obtain an estimator for the gradient matrix. If the estimating function $M(O,\beta)$ was known, then we can estimate $V(\beta_0)$ by $n^{-1}\sum_{i=1}^n M(O_i,\hat\beta_n)M^T(O_i,\hat\beta_n)$. However, since $M(O_i,\beta)$ is not known exactly, we propose using the linear approximation to estimate $M(O_i,\hat\beta_n)$ by $\hat F_i^{(\ell^*)0} + \hat F_i^{(\ell^*)1}\hat\beta_n$ and hence to estimate the variance matrix by

$$\hat V = n^{-1}\sum_{i=1}^n (\hat F_i^{(\ell^*)0} + \hat F_i^{(\ell^*)1}\hat\beta_n)(\hat F_i^{(\ell^*)0} + \hat F_i^{(\ell^*)1}\hat\beta_n)^T. \tag{6.2}$$

The asymptotic variance for $\hat\beta_n$ is then estimated using the sandwich variance estimator

$$(\hat D)^{-1}\hat V(\hat D^T)^{-1}, \tag{6.3}$$

where $\hat D$ is obtained using (5.11) and $\hat V$ is obtained using (6.2).

Even with these recommendations, there are still additional issues that need to be considered. There is the choice of θ and t in (3.1) for the Landweber-Fridman regularization algorithm, the number of Monte-Carlo replicates m in equation (5.5) for the unbiased estimator of $M(O_i,\beta)$ at each iteration, the number of iterations for the first stage of the stochastic approximation, for the second stage of the stochastic approximation we have to decide how large the random perturbations $\epsilon^{(g)}$ should be, and the total number of iterations ℓ^*. Although, we will not be able to give general results for the best choice of these values, we will give some recommendations based on empirical results from several examples which we used for illustration.

7 Example and Simulation Results

To illustrate these methods we conducted a simulation experiment similar to that in Tsiatis and Ma (2004) except that we considered two covariates measured with error instead of one which is a scenario where the methods of Tsiatis and Ma would not be feasible. Although we only used two covariates we could have easily applied this methodology to more than two covariates without any additional difficulty. Specifically, we let the response variable Y be a binary indicator and considered the quadratic logistic regression model

$$\text{logit}\{P(Y=1|X_1,X_2)\} = \beta_1 + \beta_2X_1 + \beta_3X_1^2 + \beta_4X_2,$$

where $\text{logit}(p) = \log\{p/(1-p)\}$ and $(\beta_1,\beta_2,\beta_3,\beta_4) = (-1.0, 0.7, 0.7, 0.5)$. To allow for measurement error we only observe $W_1 = X_1 + \epsilon_1$ and $W_2 = X_2 + \epsilon_2$, where ϵ_1 and ϵ_2 are independently normally distributed with mean zero and standard deviation $\sigma_{\epsilon 1}$ and $\sigma_{\epsilon 2}$ respectively, and independent of (Y, X_1, X_2). In this illustration we took (X_1, X_2) to be a bivariate normal both with variance

1, with means -1 and 0 respectively, and correlation $.71$. We allowed for a substantial amount of measurement error by taking $\sigma_{\epsilon 1} = \sigma_{\epsilon 2} = .4$.

The estimator for β and its asymptotic variance were obtained using the stochastic approximation methods described in sections 5 and 6. The key to implementing these methods is to derive the approximate estimating function $\hat{M}(O_i;\beta)$ given by (5.5). This entails generating random deviates from the conditional distribution of (Y, W_1, W_2) given (X_1, X_2) and random deviates from the conditional distribution of (X_1, X_2) given (Y, W_1, W_2) in order to derive $S^*_\beta(y_j^{(k)}, x_j^{(k)}, Z_i;\beta)$ in (5.5). Specifically, given $(X_1 = x_1, X_2 = x_2)$ we can generate random deviates (Y, W_1, W_2) easily as independent variables from a Bernoulli with probability $\exp(\beta_1 + \beta_2 x_1 + \beta_3 x_1^2 + \beta_4 x_2)/\{1 + \exp(\beta_1 + \beta_2 x_1 + \beta_3 x_1^2 + \beta_4 x_2)\}$, $N(x_1, \sigma^2_{\epsilon 1})$, and $N(x_2, \sigma^2_{\epsilon 2})$, respectively. In order to generate random deviates from the conditional distribution of (X_1, X_2) given (Y, W_1, W_2) we use a rejection sampling scheme. Using Baye's rule the conditional density is derived as

$$p(x_1, x_2|y, w_1, w_2) = \frac{\theta(x_1, x_2, y)p(x_1, x_2|w_1, w_2)}{\text{normalizing constant}}, y = 0, 1,$$

where

$$\theta(x_1, x_2, y) = \frac{\exp\{(\beta_1 + \beta_2 x_1 + \beta_3 x_1^2 + \beta_4 x_2)y\}}{\{1 + \exp(\beta_1 + \beta_2 x_1 + \beta_3 x_1^2 + \beta_4 x_2)\}}.$$

We note that this is a different but equivalent representation for the conditional density of $X = (X_1, X_2)$ given Y and $W = (W_1, W_2)$ than that given by equation (4.3) that lends itself more easily to rejection sampling. Consequently, if we could generate a random (X_1, X_2) from the conditional density $p(x_1, x_2|w_1, w_2)$, then we could generate from the conditional density $p(x_1, x_2|y, w_1, w_2)$ by keeping this (X_1, X_2) if another randomly generated uniform random variable is less than $\theta(x_1, x_2, y)$, otherwise, repeating this process until we keep such an (X_1, X_2).

Even though the true underlying distribution of (X_1, X_2) was a bivariate normal with positive correlation, for the purposes of Monte-Carlo data generation, we took X_1 and X_2 to be independent normal random variables. This not only facilitates the ease in generating random draws but also serves to show the robustness of the method to misspecification. Specifically, we took $X_j \sim N(\mu_j, \sigma_j^2), j = 1, 2$ where μ_j and σ_j^2 were estimated using the sample average and sample variance of the W_j's. Clearly, not allowing for correlation of X_1 and X_2 and overestimating the variance by using the sample variance of the measured with error W's leads to misspecification. For such a scenario random deviates from the distribution of (X_1, X_2) given (W_1, W_2) can be generated by taking independent

$$X_j \sim N\left(\frac{\sigma_j^2 W_j + \sigma^2_{\epsilon j}\mu_j}{\sigma_j^2 + \sigma^2_{\epsilon j}}, \frac{\sigma_j^2 \sigma^2_{\epsilon j}}{\sigma_j^2 + \sigma^2_{\epsilon j}}\right).$$

We conducted 1000 simulations, each with sample size $n = 500$. For the Landweber-Fridman regularization scheme we found that the regularization parameter $\theta = .6$ and the regularization parameter integer $t = 4$ worked well, although we found the results to be insensitive to slight deviations from these values. (We illustrate by also giving results for $t = 3$). For the first stage of the stochastic approximation using (5.6) we used 50 iterations with $m = 100$. We found that this was sufficient to give us a reasonable approximation to the desired estimator but would not give us a good estimator for the gradient matrix. Therefore, we also used an additional 20,000 iterations for the second stage of the stochastic approximation using (5.9) with $m = 1$ and random perturbations generated from a normal distribution with mean zero and standard deviation of .05. The estimator was obtained using (5.10), the asymptotic variance matrix of the estimator was estimated using the sandwich variance (6.3), and 95% confidence intervals were constructed using the estimate ± 1.96 estimated standard errors. For comparison, we also considered the naive estimator (5.7), where the parameters were estimated using standard logistic regression maximum likelihood with W instead of X in the quadratic logistic regression model.

The results of the simulations are summarized in Table 1. As expected, the naive estimators for β are severely biased whereas the locally efficient estimators all give good results. These estimators exhibit little bias, the average of the estimated variances closely approximates the Monte-Carlo variance, and the proportion of times that the estimated 95% confidence interval covers the true value is close to the nominal level.

Table 1: Bias, variance and coverage probabilities of the naive, and locally efficient semiparametric estimators for the quadratic logistic regression model with normal measurement error

Estimator		$\beta_1(-1)$	$\beta_2(0.7)$	$\beta_3(0.7)$	$\beta_4(0.5)$
naive	mean	-0.85	0.53	0.49	0.28
	emp sd	0.17	0.19	0.08	0.12
	est sd	0.16	0.19	0.09	0.12
	emp cov	0.83	0.83	0.31	0.55
semipar ($t = 4$)	mean	-1.01	0.74	0.73	0.51
	emp sd	0.23	0.30	0.14	0.19
	est sd	.0.23	0.30	0.14	0.19
	emp cov	0.94	0.94	0.95	0.95
semipar ($t = 3$)	mean	-1.00	0.74	0.73	0.50
	emp sd	.0.23	0.29	0.13	0.19
	est sd	0.23	0.30	0.15	0.19
	emp cov	0.95	0.94	0.96	0.94

semipar ($t = 4$) and semipar ($t = 3$) denote the locally efficient semiparametric estimators derived using the regularization parameter t equal to 4 and 3, respectively; emp sd is the empirical Monte-Carlo standard deviation of the estimators; est sd is the average of the estimated standard deviations; emp cov is the proportion of the simulations whose estimated 95% confidence intervals cover the true value of the parameters.

8 Discussion

Using the Landweber-Fridman regularization scheme for solving ill-posed integral equations together with Monte-Carlo methods for estimating the elements in this regularization scheme and stochastic approximation to obtain solutions to the estimating equation we showed how to obtain a locally efficient semiparametric estimator for the regression parameters in a functional measurement error model. By adding additional perturbation in the stochastic approximation we were also able to estimate the asymptotic variance of the proposed estimator. These methods are general enough to be applied to a wide variety of regression models with multiple covariates measured with error that were too difficult to solve using existing methodology.

We illustrated our method by considering a quadratic logistic regression model with two covariates measured with error. The algorithm was programmed in FORTRAN and it took about 5 minutes on average on a Pentium Dual Core 2.13 GHz 2GB RAM computer to obtain results for a single run with a sample size of 500. Because of the manner in which the algorithm works an increase in the sample size or the number of covariates measured with error would increase the running time linearly rather than exponentially which would be necessary for existing methods.

There is still the issue of how to best choose the tuning parameters that are used for the Landweber-Fridman regularization scheme and the stochastic approximation. The most time consuming part of this process was obtaining a good estimator for the gradient matrix. In our example, we experimented with different values of these tuning parameters until we obtained answers that were stable. For the range of measurement error we were considering, the choice of $\theta = .6$ and $t = 3$ as the regularization parameters seemed to work reasonable well. At this point we cannot give firm recommendations on the number of iterations or the magnitude of the perturbations to use in general and suggest that one experiments with these until the results stabilize.

A Deriving the Adjoint

We need to show that for any $g(X,Z) \in \mathcal{G}$ and $h(Y,W,Z) \in \mathcal{H}$ that

$$\langle A\{g(X,Z)\}, h(Y,W,Z)\rangle = \langle g(X,Z), A^*\{h(Y,W,Z)\}\rangle,$$

where $\langle g_1, g_2\rangle = E(g_1^T g_2)$. This follows because

$$\begin{aligned} E\Big[E\{g^T(X,Z)|Y,W,Z\}h(Y,W,Z)\Big] &= E\Big[E\{g^T(X,Z)h(Y,W,Z)|Y,W,Z\}\Big] \\ = E\{g^T(X,Z)h(Y,W,Z)\} &= E\Big[E\{g^T(X,Z)h(Y,W,Z)|X,Z\}\Big] \\ &= E\Big[g^T(X,Z)E\{h(Y,W,Z)|X,Z\}\Big]. \end{aligned}$$

References

[1] Bickel, P.J., Klaassen, C.A., Ritov, Y., and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models,* Baltimore: Johns Hopkins.

[2] Carroll, R.J., Ruppert, D., and Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models,* New York: Chapman and Hall.

[3] Kress, R. (1989), *Linear Integral Equations,* Berlin: Springer-Verlag.

[4] Luenberger, D.G. (1969), *Optimization by Vector Space Methods,* New York: Wiley.

[5] Rao, C.R. (1973). *Linear Statistical Inference and its Applications,* New York: Wiley.

[6] Robbins, H., and Monro, S. (1951), A Stochastic Approximation Method, *Ann. Math. Statist,* **22**, 400-407.

[7] Stefanski, L.A., and Boos, D.D. (2002), The Calculus of M-Estimation, *American Statistician,* **56**, 29-38.

[8] Tsiatis, A.A., and Ma, Y. (2004), Locally Efficient Semiparametric Estimators for Functional Measurement Error Models, *Biometrika,* **91**, 835-848.

[9] Ying, Z., and Wu, C.F. (1997), An Asymptotic Theory of Sequential Designs Based in Maximum Likelihood Recursions, *Statistic Sinica,* **7**, 75-91.