Commun.Fac.Sci.Univ.Ank.Series A1
Volume 60, Number 1, Pages 71–84 (2011)
ISSN 1303–5991

ASYMPTOTIC PROPERTIES OF SIMPLE LINEAR MEASUREMENT ERROR MODELS

RUKIYE E. DAGALP

ABSTRACT. The main objective of this paper is to study estimators of regression models on the independent variable X which is not directly observed for some reasons. In such a situation, a substitute variable W is observed instead. This substitution complicates the statistical analysis of the observed data when the purpose of the analysis is inference about a model defined in terms of X. The substitution causes a inconsistent estimator; this is defined as a measurement error problem. To correct this problem, the conditional score and corrected score methods are proposed by Stefanski&Carroll (1985) and Nakamura (1990), respectively. In this study, large sample distribution theory for both the conditional score and corrected score estimators are derived and the performance of the estimators and the adequacy of the large sample distribution theory are obtained via Monte Carlo simulation.

1. INTRODUCTION

The regression analysis is a statistical methodology for studying the functional relationship between two or more quantitative variables so that one can be explained from the other variables. Y named as the response variable is a dependent variable whose variation can be explained by an explanatory or named independent variable X. The independent variable in the regression analysis cannot be observed for some reasons either because it is too expensive to obtain, unavailable, or mismeasured. In this kind of situations, a substitute variable W is observed instead of the true variable X. Therefore, the statistical inference of regression coefficient involves the additive measurement error model, $W = X + U$, where U is a measurement error with 0 mean and a constant variance. The effect of measurement error on fitting a regression model causes inconsistent parameter estimation and also its statistical inferences. The statistical analysis of inaccurately measured data or data

Received by the editors May 13 2011, Accepted: June 29, 2011.
2000 *Mathematics Subject Classification.* Primary 05C38, 15A15; Secondary 05A15, 15A18.
Key words and phrases. Conditional score function, Corrected score function, functional model, generalized linear models, measurement error models, error-free and error-prone predictors, error in variables, M-estimator .

measured with a substitute variable causes a common problem that is called attenuation, in other words measurement error problem. When measurement error is in presence, the statistical models and methods for analyzing the such data is studied by Fuller and Hidiroglou (1978), Moran (1971) and recently Prentice (1982), Wolter and Fuller (1982a, 1982b), Carroll et al.(1984), Stefanski (1985) and Stefanski and Carroll (1985). The model fitting has been studied for functional case in which the X is regarded as an unknown fixed constants, and structural case in which the X is regarded as a random variable. For fixed X, several approaches are given to correct bias due to measurement error. One is regression calibration method studied by Carroll and Stefanski (1990) and Glesjer (1990)). The conditional score function suggested by Stefanski and Carroll (1987) and another one is corrected score function given by Stefanski (1989) and Nakamura (1990). Also, the conditional-score and corrected-score are defined and derived for the interactions between error-free and error-prone regressors in the models by Dagalp (2001). The SIMEX (simulation extrapolation) method is proposed by Cook and Stefanski (1994). Actually, the analysis of data with measurement error is well described by Fuller (1987) and Carroll et al. (1995). Hanfelt and Liang (1997) focused more directly on hypothesis test which is an alternative to Wald's test for the regression parameters. In this article, basic theory of the conditional-score and the corrected-score methods are explained. Asymptotic properties of the conditional-score and corrected-score estimators are derived and results of a simulation studied are presented.

Consider the usual simple linear regression model $Y_i = \alpha + \beta_x X_i + \epsilon_i$, $i = 1, ..., n$ where ϵ_i represents experimental error, and the additive measurement error model for the observed $W_i = X_i + U_i$, where U_i represents measurement error. In this article, the simple linear regression model is regarded as a form of the exponential family given in McCullagh and Nelder (1989, Chap. 2). Given a covariate $X = x$, the response variable Y has density function as a generalized linear model in canonical form

$$f_{Y|X}(y|x;\theta) = \exp\left\{\frac{y\eta - b(\eta)}{\sigma_\epsilon^2} + c(y, \sigma_\epsilon^2)\right\}, \tag{1.1}$$

with respect to σ-finite measure $m(\cdot)$. where $\eta = \alpha + \beta_x x$ is called the natural parameter, $b(\eta) = \frac{\eta^2}{2}$, $c(y, \sigma_\epsilon^2) = -\frac{y^2}{2\sigma_\epsilon^2} - \log(\sqrt{2\pi}\sigma_\epsilon^2)$, and $\theta = (\alpha, \beta_x, \sigma_\epsilon^2)^T$ is the unknown parameter. The mean and variance of Y given X are $b'(\eta)$ and $\sigma_\epsilon^2 b''(\eta)$, where b' and b'' are the first and second derivatives of $b(\eta)$ with respect to η, respectively. When the measurement error U is distributed as a normal random variable with mean zero and variance σ_u^2, the density of W given $X = x$ is

$$f_{W|X}(w|x, \sigma_u^2) = (2\pi)^{-\frac{1}{2}}(\sigma_u^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma_u^2}(w - x)^2\right\}. \tag{1.2}$$

The models in (1.1) and (1.2) together define a generalized linear measurement error model, but in this paper, simple linear model case is taken into consideration. In the simple linear case, it is emphasized that the parameters of interest (α, β_x)

appear in regression model for $b'(\eta) = E(Y \mid X) = \eta$, which depends on the unobserved true variable X. For some cases, the variable W is observed instead of X with a measurement error variable. Hence, the second component of the model relates the observed substitute variable W to X. If the model fits the observed data, prediction of the parameter of interest results in biased estimators. If X is directly observed, then the $(\alpha, \beta_x, \sigma_\epsilon^2)$ is estimated by solving the normal equations given in matrix form as

$$\sum_{i=1}^{n} \{Y_i - b'(\eta_i)\} \begin{pmatrix} 1 \\ X_i \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{1.3}$$

$$\sum_{i=1}^{n} \left[\left(\frac{n-p}{n} \right) \sigma_\epsilon^2 - \frac{\{Y_i - b'(\eta_i)\}^2}{b''(\eta_i)} \right] = 0 \tag{1.4}$$

where the natural parameter, $\eta_i = \alpha + \beta_x X_i$ and p is the number of parameters in the model. Note that the estimating equations (1.3) and (1.4) result in maximum likelihood estimators when $n - p$ is replaced by n.The equations in (1.3) yield the ordinary least squares estimate of slope on the true data given by

$$\widehat{\beta}_{Y|X} = \frac{S_{XY}}{S_{XX}},$$

and substituting W for X in (1.3) yields to the so-called *naive* slope estimator that is given by

$$\widehat{\beta}_{Y|W} = \frac{S_{WY}}{S_{WW}} = \frac{S_{XY} + S_{YU}}{S_{XX} + 2S_{XU} + S_{UU}},$$

where S_{XX} is the sample variance of $X_1, ..., X_n$, S_{UU} is the sample variance of $U_1, ..., U_n$, S_{XU} is the sample covariance of (X_i, U_i) and S_{YU} is the sample covariance of (Y_i, U_i), $i = 1, 2, ..., n$. By the Law of Large Numbers, both S_{YU} and S_{XU} converge in probability to zero, $S_{XX} \xrightarrow{P} \sigma_X^2$, and $S_{UU} \xrightarrow{P} \sigma_U^2$, as $n \longrightarrow \infty$. Thus, by Slutky's theorem,

$$\widehat{\beta}_{Y|W} \xrightarrow{P} \lambda \beta_X, \text{ as } n \longrightarrow \infty$$

where $\lambda = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2}$ so-called the attenuation factor (Fuller, 1987) is a real number in the range $[0, 1]$ for σ_X^2 positive and finite. The extreme case $\lambda = 1$ is obtained when there is no measurement error $\left(\sigma_U^2 = 0\right)$. The other extreme case is approached only in the limit as $\sigma_U^2 \longrightarrow \infty$ for fixed $\sigma_X^2 < \infty$. The *naive* slope estimator $\widehat{\beta}_{Y|W}$ is biased towards zero because of $0 \leq \lambda \leq 1$.

The purpose is to obtain sufficient and consistent estimators of the parameters on the observed data instead of the true data. Some methods are suggested to correct this attenuation to get statistical inference for the parameters of interest. The conditional-score and corrected-score are defined in the following sections for measurement error problem.

2. Conditional-Score Estimation

Consider the simple linear regression model of Y on X

$$Y = \alpha + \beta_X X + \epsilon,$$

and the additive measurement error model

$$W = X + U,$$

where the error-free regressor $X \sim N\left(\mu_X, \sigma_X^2\right)$, the experimental error $\epsilon \sim N\left(0, \sigma_\epsilon^2\right)$, and the measurement error $U \sim N\left(0, \sigma_U^2\right)$ are independent variables. A common strategy for handling models with unobserved $X_1, \ldots, X_n$, so-called nuisance parameters, is to base inference on conditional likelihoods since functional maximum likelihood estimation neither is computationally achievable nor the estimators are generally consistent due to the large number of nuisance parameters (Neyman and Scott, 1948; Stefanski and Carroll, 1987). Stefanski and Carroll (1987) showed how to derive conditional estimating equations for generalized linear measurement error models. When X is given, the conditional density of Y and the conditional density of W given as

$$Y \mid X \sim N\left(\eta, \sigma_\epsilon^2\right), \quad \text{and} \quad W \mid X \sim N\left(X, \sigma_U^2\right)$$

are independent. Hence, the joint density of (Y, W) given the unobserved predictor X, is

$$f_{Y,W|X}(y, w|x; \theta) = f_{Y|X}(y|x; \theta) f_{W|X}(w|x).$$

To find the estimator for error-free unobserved regressor X, functional maximum likelihood estimation maximizes the likelihood as a function of θ and the unobserved predictors $X_1, \ldots, X_n$, i.e.,

$$L(\theta; X_1, \ldots, X_n|(Y_1, W_1), \ldots, (Y_n, W_n)) = \sum_{i=1}^{n} \log\{f_{Y,W|X}(Y_i, W_i|x_i; \theta)\}. \quad (2.1)$$

Consider the joint density in (2.1) and define $\Omega = \frac{\sigma_U^2}{\sigma_\epsilon^2}$. The joint density of (Y, W) given $X = x$ can be written as

$$f_{Y,W|X}(y, w|x; \theta) = h_1(\delta, x) h_2(y, w; \theta),$$

where

$$\begin{aligned} h_1(\delta, x) &= \exp\left\{ \frac{x}{\sigma_U^2} \{y\Omega\beta_X + w\} - \frac{1}{2}\frac{x^2}{\sigma_U^2} \right\}, \\ h_2(y, w) &= \exp\left\{ \frac{y\alpha - b(\eta)}{\sigma_\epsilon^2} + c(y, \sigma_\epsilon^2) - \frac{1}{2}\frac{w^2}{\sigma_U^2} - \frac{1}{2}\log\left[(2\pi)\sigma_U^2\right] \right\} \end{aligned}$$

and

$$\delta = w + y\Omega\beta_X \ .$$

When X is regarded as a parameter and all other parameters as known in the density of $(Y, W|X)$, the statistic

$$\Delta = \Delta(Y, W; \theta) = W + Y\Omega\beta_X \tag{2.2}$$

is complete and sufficient for X by the Factorization Theorem (Casella and Berger 1990, p.250). Thus, the distribution of $Y|\Delta$ depends only on Y, W and θ, but not on the unobserved true regressor X. From the joint density function of (Y, Δ), the conditional density of Y given Δ is

$$f_{Y|\Delta}(y|\delta;\theta) = \exp\left\{y\varphi - \frac{1}{2}\frac{y^2\beta_X^2\Omega}{\sigma_\epsilon^2} + c(y,\sigma_\epsilon^2) - \log\left\{S(\varphi,\beta_X,\sigma_\epsilon^2)\right\}\right\}, \tag{2.3}$$

where

$$\varphi = \frac{\eta + (\delta - x)\beta_X}{\sigma_\epsilon^2} = \frac{\alpha + \beta_X\delta}{\sigma_\epsilon^2},$$

and $S(\cdot,\cdot,\cdot)$ is defined as

$$S(\varphi,\beta_X,\sigma_\epsilon^2) = \int \exp\left\{y\varphi - \frac{1}{2}\frac{y^2\beta_X^2\Omega}{\sigma_\epsilon^2} + c(y,\sigma_\epsilon^2)\right\} dy.$$

The moments of Y given $\Delta = \delta$ can be computed from the partial derivatives of $S(\varphi,\beta_X,\sigma_\epsilon^2)$ with respect to φ because (2.3) is an exponential family density in φ and Y is the natural sufficient statistic. The conditional distribution of Y given $\Delta = \delta$ is an exponential family with respect to the σ-finite measure $m(\cdot)$ which does not depend on θ. Thus

$$E\{f'_{Y|\Delta}(y|\delta;\theta)\} = \int f'_{Y|\Delta}(y|\delta;\theta)dy = 0, \tag{2.4}$$

where

$$f'_{Y|\Delta}(y|\delta;\theta) = \frac{\partial}{\partial\theta} f_{Y|\Delta}(y|\delta;\theta).$$

From this equation, consistent estimating equations for θ can be derived to estimate the regression parameters of interest. An alternative derivation of the conditional-score from the conditional distribution of Y given $\Delta = \delta$ is suggested by Stefanski and Carroll (1985) as defined and derived clearly by Dagalp (2001). The conditional-score function given by Stefanski and Carroll (1987) is

$$\Psi_C(Y, W; \theta) = l' - E\left[l'|\Delta\right], \tag{2.5}$$

where

$$l' = l'(Y, W; \theta) = E\left\{\frac{f'_{Y|X}(Y|X;\theta)}{f_{Y|X}(Y|X;\theta)}|Y, W\right\} = E\left\{\frac{f'_{Y|X}(Y|X;\theta)}{f_{Y|X}(Y|X;\theta)}|\Delta\right\}. \tag{2.6}$$

Using the fact that the σ-algebra generated by Δ is a sub-σ-algebra generated by Y and W , thus combining the results in (2.5) and (2.6), the conditional-score function

is driven as

$$\Psi_C(Y,W;\theta)=\begin{pmatrix} Y-E(Y|\Delta) \\ \{Y-E(Y|\Delta)\}\Delta \\ \{Y-E(Y|\Delta)\}^2-\frac{\sigma_\epsilon^2}{1+\Omega\beta_X} \end{pmatrix}. \tag{2.7}$$

It follows from (2.5) and (2.7) that

$$E\left\{\Psi_C(Y,W;\theta)\right\}=0, \tag{2.8}$$

and from (2.8) , the conditional-score estimators can be obtained as solutions of

$$\sum_{i=1}^{n}\Psi_C(Y_i,W_i;\hat{\theta})=0$$

named as a consistent estimating equations for θ.

3. Corrected-Score Estimation

The corrected-score method is proposed and defined by Stefanski (1989) and Nakamura (1990) and later studied by Novick (2000). It is a natural competitor to the conditional-score estimator and a technique for eliminating asymptotic bias caused by measurement error for the generalized measurement error models. In this paper, the corrected-score method is defined for simple linear measurement error models by using unbiased estimating equations for the parameters of interest. Assume that there exists an unbiased score function Ψ for the true data X as an error-free predictor. The unknown parameter θ in the absence of measurement error is estimated as the solution of the consistent estimating equations and is called the true estimator.

Suppose that $\Psi(Y,X;\theta)$ is a score function from the model for the true data such that the estimator $\hat{\theta}_{True}$, solving

$$\sum_{i=1}^{n}\Psi(Y_i,X_i;\hat{\theta}_{True})=0 \tag{3.1}$$

is consistent for θ. It follows

$$E\left\{\Psi\left(Y,X;\theta\right)\right\}=0.$$

Suppose that there exists a certain smooth Ψ functions such that

$$E\left\{\Psi_M(Y,W;\theta)\mid Y,X\right\}=\Psi(Y,X;\theta) \tag{3.2}$$

where $\Psi_M(Y,X;\theta)$ is a corrected-score function of the observed data. It follows from this property that

$$E\left\{\Psi_M(Y,W;\theta)\right\}=0,$$

so that $\Psi_M(.,.;.)$ is a Fisher-consistent score function (Carroll, Ruppert and Stefanski, 1995). The M-estimator $\hat{\theta}_M$ based on the observed data is defined as the

solution to

$$\sum_{i=1}^{n} \Psi(Y_i, X_i; \hat{\theta}_M) = 0.$$

The problem is that, the corrected-score function satisfying (3.2) does not always exist, and when it exists, it is not easily find. The corrected-score functions for some models have been studied and derived by Stefanski (1989) who provides a mean of determining corrected-score functions for a large class of models. Let $Z \sim N(0,1)$ be independent of (Y,W) and $i = \sqrt{-1}$. For the existence of a certain smooth Ψ functions a corrected-score can be found as

$$\Psi_M(Y,W;\theta) = E\left\{\Psi(Y, W + i\sigma_U Z; \theta) | Y, W\right\}. \tag{3.3}$$

When it exists, the corrected-score function in (3.3) can sometimes be found mathematically and can always be computed by Monte Carlo simulation (Novick, 2000).

Consider the linear regression model $Y = \alpha + \beta_X X + \epsilon$ and the measurement error model $W = X + U$ where the error-free regressor $X \sim N\left(\mu_X, \sigma_X^2\right)$, the experimental error $\epsilon \sim N\left(0, \sigma_\epsilon^2\right)$, and the measurement error $U \sim N\left(0, \sigma_U^2\right)$ are independent variables, the true-data likelihood score function is

$$\Psi(Y,X;\theta) = \begin{pmatrix} Y - \eta \\ (Y-\eta)X \\ (Y-\eta)^2 - \sigma_\epsilon^2 \end{pmatrix}, \tag{3.4}$$

where $\eta = \alpha + \beta_X X$, and $\theta = (\alpha, \beta_X, \sigma_\epsilon^2)^T$. To find the corrected-score function in (3.3) from certain smooth Ψ functions in (3.4), result shows that the corrected-score function $\Psi_M(\cdot,\cdot,\cdot,\cdot)$ is given as

$$\Psi_M(Y,W;\theta) = \begin{pmatrix} Y - \eta_w \\ (Y-\eta_w)W + \beta_X \sigma_U^2 \\ (Y-\eta_w)^2 - \beta_X^2\sigma_U^2 - \sigma_\epsilon^2 \end{pmatrix}, \tag{3.5}$$

where $\eta_w = \alpha + \beta_X W$. Taking the expectation of Ψ_M and routine calculations show that

$$E\left\{\Psi_M(Y,W;\theta) \mid Y, X\right\} = \Psi(Y,X;\theta),$$

and thus

$$E\left\{\Psi_M(Y,W;\theta) \mid X\right\} = E\left\{\Psi(Y,X;\theta) \mid X\right\} = 0.$$

The corrected-score estimating equations based on the observed data are then

$$\sum_{i=1}^{n} \Psi_M(Y_i, W_i; \hat{\theta}) = 0. \tag{3.6}$$

An estimator satisfying the equations in (3.6) is called a **corrected-score estimator** and $\Psi_M(\cdot,\cdot,\cdot,\cdot)$ in (3.5) is called the **corrected-score function** for linear regression.

4. Large-Sample Inference for Conditional-Score and Corrected-Score Estimates

Corrected-score and conditional-score defined in this paper provide M-estimators which are consistent in the absence of measurement error. These estimators are determined by solving the unbiased estimating equations of the form in (3.1). The conditional-score and corrected-score estimators are obtained by replacing Ψ function by Ψ_C in (2.7) and Ψ_M in (3.5), respectively. Provided that

$$E\left\{\Psi(Y,W;\theta)\right\}=0.$$

If it is known that the large sample distribution of M-estimators is multivariate normal with mean 0 and a covariance matrix that depends on Ψ and $\Psi^{'}=\left(\partial/\partial\theta^{T}\right)\Psi$ under some regularity and moments conditions (Stefanski, 1985). Asymptotic distribution of M-estimators are reviewed and applied to conditional-score and corrected-score estimators by a Taylor-series expansion,

$$\frac{1}{n}\sum_{i=1}^{n}\Psi(Y_i,W_i;\theta)+\widehat{A}_n(\theta)\left(\widehat{\theta}-\theta\right)+o_p(n^{-1})=0,$$

where $\widehat{A}_n(\theta)=\dfrac{1}{n}\sum\limits_{i=1}^{n}\dfrac{\partial}{\partial\theta^{T}}\Psi(Y_i,W_i;\theta)$. Under the assumption $\widehat{A}_n(\theta)$ is a non-singular matrix, then

$$\sqrt{n}\left(\widehat{\theta}-\theta\right)=-\left\{\widehat{A}_n(\theta)\right\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Psi(Y_i,W_i;\theta)+o_p(1).$$

The asymptotic normality follows from the Central Limit Theorem. Thus, the M-estimator $\widehat{\theta}$ is asymptotically normally distributed with mean θ and covariance matrix $n^{-1}A_n^{-1}(\theta)B_n(\theta)\left\{A_n^{-1}(\theta)\right\}^{T}$, where

$$\begin{aligned}A_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n}E\left\{\frac{\partial}{\partial\theta^{T}}\Psi(Yi,W_i;\theta)\right\},\\ B_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n}E\left\{\Psi(Y_i,W_i;\theta)\Psi^{T}(Y_i,W_i;\theta)\right\}.\end{aligned}$$

For the case of independent and identically distributed variables $\{Y_i,W_i\}$, $i=1,\ldots,n$, A and B are defined as

$$\begin{aligned}A_n(\theta)\overset{n\to\infty}{\longrightarrow}A(\theta) &= E\left\{\frac{\partial}{\partial\theta^{T}}\Psi(Y,W;\theta)\right\},\\ B_n(\theta)\overset{n\to\infty}{\longrightarrow}B(\theta) &= E\left\{\Psi(Y,W;\theta)\Psi^{T}(Y,W;\theta)\right\}.\end{aligned}$$

For the conditional-score function in (2.7), the matrices A and B are designated A_C and B_C, that are also given by the expressions, respectively;

$$\begin{aligned} A_C &= E\left\{E\left\{\frac{\partial}{\partial\theta^T}\Psi_C(Y,W;\theta)\mid\Delta\right\}\right\} = E\{a_C(\Delta;\theta)\}, \\ B_C &= E\left[E\left\{\Psi_C(Y,W;\theta)\Psi_C^T(Y,W;\theta)\mid\Delta\right\}\right] = E\{b_C(\Delta;\theta)\}. \end{aligned} \tag{4.1}$$

For the corrected-score function in (3.5), the A and B matrices are designated A_M and B_M that can be written as

$$\begin{aligned} A_M &= E\left\{E\left\{\frac{\partial}{\partial\theta^T}\Psi_M(Y,W;\theta)\mid X\right\}\right\} = E\{a_M(X;\theta)\}, \\ B_M &= E\left\{E\left\{\Psi_M(Y,W;\theta)\Psi_M^T(Y,W;\theta)\mid X\right\}\right\} = E\{b_M(X;\theta)\}. \end{aligned}$$

Replacing each term in the form $X^j,\ j=1,2$ by $E\{(W+i\sigma_U Z)^j|W\}$ in $a_M(X;\theta)$ results $a_M^*(W;\theta)$ having the property that

$$E\left\{a_M^*(W;\theta)|X\right\} = E\left\{\frac{\partial}{\partial\theta^T}\Psi_M(Y,W;\theta)\mid X\right\}.$$

and replacing the covariate X by W in $b_M(X;\theta)$, $b_M^*(W;\theta)$ is obtained as $b_M^*(W;\theta) = E\left\{b_M(W+i\sigma_U Z;\theta)|W,\right\}$ satisfying

$$E\left\{b_M^*(W;\theta)|X\right\} = b_M(X;\theta).$$

5. Empirical and Conditional Model Based Met-hods of Estimating Asymptotic Covariance Mat-rix

In this section, two methods of estimating the covariance matrix of θ are described. The first is often called the **sandwich variance estimator** which uses empirical expectation to estimate the A and B matrices appearing in the sandwich variance matrix. In this case, the estimators of $A_n(\theta)$ and $B_n(\theta)$ for the conditional-score and the corrected-score functions are

$$\begin{aligned} \widehat{A}_n(\widehat{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta^T}\Psi(Y_i,W_i;\theta)\,\big|_{\theta=\widehat{\theta}}\,, \\ \widehat{B}_n(\widehat{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\Psi(Y_i,W_i;\widehat{\theta})\Psi^T(Y_i,W_i;\widehat{\theta}). \end{aligned}$$

Thus, the sandwich variance estimator of the asymptotic covariance matrix of $\widehat{\theta}$ is

$$\widehat{V}(\widehat{\theta}) = n^{-1}\{\widehat{A}_n(\widehat{\theta})\}^{-1}\widehat{B}_n(\widehat{\theta})\left[\{\widehat{A}_n(\widehat{\theta})\}^{-1}\right]^T.$$

The second method is called the **conditional model-based expectation method**. For the conditional-score function, we have the distribution of Y given the sufficient statistic Δ for X in (2.2) and the observed covariate. It follows from (4.1) that

$$\widetilde{A}_{Cn}(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} a_C(\Delta_i, \widehat{\theta}),$$
$$\widetilde{B}_{Cn}(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} b_C(\Delta_i, \widehat{\theta}),$$

are consistent estimators of A_C and B_C, respectively.

When the true predictor X is measured without error, Ψ is the likelihood score function. Thus, the estimator of θ that satisfies the M-estimating equations in (3.1) is the maximum likelihood estimator (MLE) and denoted $\widehat{\theta}_{True}$. In this case, the matrices $A_n(\theta)$ and $B_n(\theta)$ are equal to the Fisher Information matrix, denoted by $I(\theta)$,

$$I(\theta) = B_n(\theta) = -A_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} E\left\{\frac{\partial}{\partial\theta^T}\Psi(Y_i, W_i; \theta)\right\},$$

and the asymptotic distribution of $\widehat{\theta}_{True}$ is

$$\sqrt{n}\left(\widehat{\theta}_{True} - \theta\right) \xrightarrow{D} N\left(0, I^{-1}(\theta)\right).$$

6. The Results of the Monte Carlo Simulation

In this section, results from a Monte Carlo simulation study are reported. The SAS `PROC IML` software was used to perform all simulations and analysis of the output for both the conditional-score and the corrected-score functions. Each simulation consisted of generating $B = 100$ data sets of size n. The data $\{X_j, \epsilon_j, U_j\}_{j=1}^n$ were generated from the normal distribution

$$\begin{pmatrix} X \\ \epsilon \\ U \end{pmatrix} \sim N\left\{\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sigma_\epsilon^2 & 0 \\ 0 & 0 & \sigma_U^2 \end{pmatrix}\right\}.$$

For each data set, the response variable Y_j is normally distributed with mean $\boldsymbol{\theta}^T\mathbf{X}_j$, where $\mathbf{X}_j = (1, X_j,)^T$. The measured data W_j was generated as $W_j = X_j + U_j$, $j = 1, \ldots, n$.

For the simulation study $\theta = (\alpha, \beta_X, \sigma_\epsilon^2)^T = (0, 0.5, 0.5)^T$, the measurement error variance $\sigma_U^2 = \{0.5, 0.75\}$ and sample sizes of $n = \{50, 100, 200, 500\}$ were investigated. Four estimators were studied:

- $\widehat{\boldsymbol{\theta}}_{True}$ calculated from the true data $\{Y_j, X_j,\}_{j=1}^n$;
- $\widehat{\boldsymbol{\theta}}_{Naive}$ calculated from the observed data $\{Y_j, W_j\}_{j=1}^n$;
- $\widehat{\boldsymbol{\theta}}_C$, the conditional-score estimator, calculated from the observed data $\{Y_j, W_j\}_{j=1}^n$;
- $\widehat{\boldsymbol{\theta}}_M$, the corrected-score estimator, calculated from the observed data $\{Y_j, W_j\}_{j=1}^n$.

For all of the estimators given above, Newton-Raphson iterations with an initial guess, say $\hat{\theta}^{(0)}$ are

$$\hat{\theta}^{(k+1)} \approx \hat{\theta}^{(k)} - \hat{A}_n^{-1}(\hat{\theta}^{(k)})G_n(\hat{\theta}^{(k)}),\ k = 1, 2, \cdots B,$$

where B is the iteration number, and

$$\begin{aligned} \hat{A}_n(\hat{\theta}^{(k)}) &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \theta^T}\Psi(Y_i, W_i; \theta)\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}^{(k)}}, \\ G_n(\hat{\theta}^{(k)}) &= \frac{1}{n}\sum_{i=1}^{n} \Psi(Y_i, W_i; \hat{\theta}^{(k)}). \end{aligned}$$

The iteration stops when two successive iterates differ by less than a specified tolerance or when the number of iterations exceeds an allowed maximum.

The true, naive, conditional-score, corrected-score estimators of α, β_X and σ_ϵ^2 were compared and Monte Carlo means are reported in Table 1 for sample size $50, 100, 200, 500$ and 1000 when measurement error variance σ_U^2 takes 0.5 and 0.75 values. The silent features of Table 1 include the attenuation in the naive estimators of β_X for two different values of σ_U^2. When the sample size increases, the naive estimator cannot obtain better estimate. The conditional-score and corrected-score estimates of slope using W as the measurement of X produce unbiased estimators under the assumption that the error variance is known and equal to 0.5 and 0.75. Both the conditional-score estimator and the corrected-score estimator of the slope completely correct for the attenuation due to measurement error, even if the sample size is small. When measurement error variance is 0.5, the naive estimator is drastically biased even though the sample is large. If the conditional-score and the corrected-score methods are compared for the slope estimator, both methods correct the attenuation and yield much better estimators for $\sigma_U^2 = 0.5$. When the true and naive slope estimators are obtained by the conditional-score and the corrrected-score methods, both estimators are the same. Also noteworthy is the fact that the conditional-score and corrected-score methods obtain the same true and naive estimator, therefore it shows both methods work skillfully for error-free and error-prone regressor at all sample sizes.

TABLE 1. Simulation study results of the true, naive, conditional-score and corrected-score estimators for the simple linear regression model with parameters $\boldsymbol{\theta} = (\alpha, \beta_X, \sigma_\epsilon^2)^T = (0, 0.5, 0.5)^T$ and measurement error variance $\sigma_U^2 = \{0.5, 0.75\}$. Table entries are means of 100 Monte Carlo runs for sample size $n = \{50, 100, 200, 500, 1000\}$.

		$\sigma_U^2 = 0.5$			$\sigma_U^2 = 0.75$		
n	Estimator	$\alpha = 0$	$\beta_X = 0.5$	$\sigma_\epsilon^2 = 0.5$	$\alpha = 0$	$\beta_X = 0.5$	$\sigma_\epsilon^2 = 0.5$
50	True	−0.0091	0.4949	0.4672	−0.0028	0.4803	0.4596
	Naive	−0.0114	0.3304	0.5602	0.0020	0.2794	0.5556
	Conditional	−0.0152	0.5509	0.4431	0.0065	0.4156	0.4752
	Corrected	−0.0198	0.5016	0.4557	0.0071	0.4187	0.4681
100	True	0.0032	0.4989	0.4930	−0.0048	0.5054	0.4823
	Naive	0.0014	0.3356	0.5890	−0.0101	0.2882	0.5964
	Conditional	0.0012	0.5297	0.4852	−0.0097	0.4243	0.5188
	Corrected	−0.0000	0.5301	0.4777	−0.0096	0.4224	0.5211
200	True	0.0056	0.5064	0.4829	0.0015	0.5000	0.4862
	Naive	0.0095	0.3351	0.5829	−0.0006	0.2883	0.6088
	Conditional	0.0098	0.5080	0.4809	−0.0017	0.4151	0.5323
	Corrected	0.0099	0.5083	0.4800	−0.0017	0.4099	0.5386
500	True	0.0003	0.4974	0.4941	−0.0045	0.5035	0.4933
	Naive	−0.0023	0.3312	0.5884	−0.0050	0.2887	0.6184
	Conditional	−0.0024	0.5026	0.4907	−0.0047	0.4096	0.5424
	Corrected	−0.0025	0.5025	0.4907	−0.0050	0.4045	0.5498
1000	True	−0.0014	0.4999	0.4987	−0.0025	0.5023	0.4950
	Naive	−0.0022	0.3315	0.5826	−0.0025	0.2884	0.6027
	Conditional	−0.0014	0.5008	0.4991	−0.0018	0.4047	0.5438
	Corrected	−0.0014	0.5009	0.4990	−0.0019	0.4047	0.5439

Two methods, the conditional score model-based expectation method and the corrected score model-based expectation method, for estimating the covariance matrix of θ are given in Table 2 via Monte Carlo simulation for $\{100, 1000, 10000, 100000\}$. The conditional-score yields smaller variance estimates then the corrected-score. ARE shows that the corrected-score produces smaller variances then conditional-score for 100 and 1000 Monte Carlo runs when $\sigma_U^2 = 0.5$. If the measurement error variance is equal or greater than 0.75, the variance of the conditional-score estimator is obtained smaller than the corrected score's.

TABLE 2. Simulation study results of variance estimations of the conditional-score and corrected-score estimators for the simple linear regression model with parameters $\boldsymbol{\theta} = (\alpha, \beta_X, \sigma_\epsilon^2)^T = (0, 0.5, 0.5)^T$ and measurement error variance $\sigma_U^2 = \{0.5, 0.75\}$. Table entries are means of $B = \{100, 1000, 10000, 100000\}$ Monte Carlo runs

Variance Estimation							
		$\sigma_U^2 = 0.5$			$\sigma_U^2 = 0.75$		
B	Estimator	$\alpha = 0$	$\beta_X = 0.5$	$\sigma_\epsilon^2 = 0.5$	$\alpha = 0$	$\beta_X = 0.5$	$\sigma_\epsilon^2 = 0.5$
100	$Var(\hat{\theta}_C)$	0.6834	0.8676	0.7647	0.7801	1.4140	0.8945
	$Var(\hat{\theta}_M)$	0.6893	0.7947	0.8261	0.8035	1.5077	1.0297
	ARE	0.9914	1.0917	0.9257	0.9708	0.9378	0.8687
1000	$Var(\hat{\theta}_C)$	0.6420	1.1921	0.7809	0.7171	1.6050	0.9218
	$Var(\hat{\theta}_M)$	0.6434	1.1620	0.8547	0.7220	1.8470	1.0930
	ARE	0.9978	1.0259	0.9136	0.9932	0.8689	0.8434
10000	$Var(\hat{\theta}_C)$	0.6437	1.0070	0.7803	0.7153	1.7303	0.9130
	$Var(\hat{\theta}_M)$	0.6445	1.0309	0.8566	0.7180	2.3505	1.0763
	ARE	0.9987	0.9768	0.9109	0.9962	0.7361	0.8483
100000	$Var(\hat{\theta}_C)$	0.6434	1.0009	0.7793	0.7148	1.6532	0.9141
	$Var(\hat{\theta}_M)$	0.6445	1.0375	0.8555	0.7171	2.1672	1.0796
	ARE	0.9983	0.9647	0.9109	0.9967	0.7628	0.8467

REFERENCES

[1] Carroll, R. J., Ruppert, D. and Stefanski, L. A., Measurement Error in Nonlinear Models, Chapman and Hall, London, 1995,1st.

[2] Casella, G. and Berger, R. L., Statistical Inference, Wadsworth Publishing Company, Belmont, 1990.

[3] Cook, J.R. and Stefanski, L.A. (1994), Simulation-Extrapolation Estimation in Parametric Measurement Error Models, Journal of American Statistical Association, 89, 1314-1327.

[4] Dagalp, R. D., (2001), Estimators for Generalized Linear Measurement Error Models with Interaction Terms, Ph.D. Thesis, North Carolina State University.

[5] Fuller, W. A., Measurement Error Models, John Willey and Sons, New York,1987.

[6] Fuller, A. W. and Hidiroglou, M. A., (1978), Regression Estimation After Correcting for Attenuation, Journal of the American Statistical Association, 73, 99-104.

[7] Glesjer, L.J. (1990), Improvements of the naive approach to estimation in nonlinear errors-in variables regression models. In Statistical Analysis of Measurement Error Models. Providence, RI: American Mathematics Society.

[8] Godambe, V. P., (1976), Conditional Likelihood and Unconditional Optimum Estimating Equations", Biometrika, 63, 277-84.

[9] Hanfelt, J. J. and Liang, K.,(1997), Approximate Likelihoods for Generalized Linear Errors-in-variables Models, J.R. Statist. Soc. B, 59, 627-637.

[10] Huang, L. and Wang, H., (2005), Assessing Interaction Effects in Linear Measurement Error Models, Appl. Statist., 54, 21-30.
[11] Kopka, H. and Daly, P. W.,A Guide to LaTeX,Addison Wesley Longman Limited, Harlow,1999,3rd.
[12] Lindsay, B.,(1982), Conditional Score Functions: Some Optimality Results, Biometrika, 69, 503-512.
[13] Lindsay, B., (1983), Efficiency of the Conditional Score in a Mixture Setting, The Annals of Statistics, 11, 246-497.
[14] Lindsay, B., (1985), Using Empirical Partially Bayes Inference for Increased Efficiency, The Annals of Statistics, 13, 914-931.
[15] McCullagh, P. and Nelder, J. A., Generalized Linear Models, Chapman and Hall, London, 1989,2nd.
[16] Moran, P., (1971), Estimating Structural and Functional Relationships, J. Mult. Anal. I, 232-55.
[17] Nakamura, T., (1990), Corrected Score Function for Errors-in-Variables Models: Methodology and Application to Generalized Linear Models, Biometrika, 77, 127-137.
[18] Neyman, J. and Scott, E. L., (1948), Consistent Estimates Based on Partially Consistent Observations, Econometrica, 16, 1-32.
[19] Novick, S., (2000), Parametric Modeling in the Presence of Measurement Error: Monte Carlo Corrected Scores, North Carolina State University.
[20] Prentice, R. L., (1982),Covariate Measurement Errors and Parameter Estimation in a Failure Time Regression Model, Biometrika, 69, 331-42.
[21] SAS/IML Software: Usage and Reference Version 6, SAS Institute Inc. Cary, 1990,1st.
[22] Searle, S.R., Linear Models,John Willey and Sons, New York, 1971, 1st.
[23] Seber, G. A. F. and Wild, C. J., Nonlinear Regression, John Willey and Sons, New York, 1989.
[24] Serfling, R. J.,Approximation Theorems of Mathematical Statistics, John Willey and Sons, Singapore, 1980.
[25] Stefanski, L. A., (1985), The effects of Measurement Error on Parameter Estimation, Biometrika, 72, 583-592.
[26] Stefanski, L. A. and Carroll, R. J., (1985), Covariate Measurement Error in Logistic Regression, The Annals of Statistics, 13, 1335-1351.
[27] Stefanski, L. A. and Carroll, R. J., (1987), Conditional Scores and Optimal Scores for Generalized Linear Measurement error models, Biometrika, 4, 703-716.
[28] Stefanski, L. A. (1989), Unbiased Estimation of a Nonlinear Function of a Normal Mean with Application to Measurement Error Models, Communication of Statistical Theory Methodology, 18,4335-4358.
[29] Stefanski, L. A. and Carroll, R. J., (1990), Score Tests in Generalized Linear Measurement Error Models, Journal of Royal Statistical Society Series B, 52, 345-359.
[30] Wolter, J.M. and Fuller, W.A., (1982a), Estimation of Nonlinear Error-in- variables Models, The Annals of Statistics, 10, 1335-48.
[31] Wolter, J.M. and Fuller, W.A., (1982b), Estimation of Quadratic Error-in- variables Models, Biometrika, 69, 174-82.

Current address: Department of Statistics, Faculty of Science, Ankara University
06100 Tandogan, Ankara - TURKEY
E-mail address: dagalp@science.ankara.edu.tr
URL: http:/communucations.science.ankara.edu.tr