Article electronically published on March 18, 2017

Commun.Fac.Sci.Univ.Ank.Series A1
Volume 66, Number 2, Pages 311–322 (2017)
DOI: 10.1501/Commua1_0000000821
ISSN 1303–5991

http://communications.science.ankara.edu.tr/index.php?series=A1

ESTIMATION METHODS FOR SIMPLE LINEAR REGRESSION WITH MEASUREMENT ERROR: A REAL DATA APPLICATION

RUKİYE E. DAĞALP, İHSAN KARABULUT, AND FİKRİ ÖZTÜRK

Abstract. The classical measurement error model is discussed in the context of parameter estimation of the simple linear regression. The attenuation effect of measurement error on the parameter estimation is eliminated using the regression calibration and simulation extrapolation methods. The mass density of pebbles population is investigated as a real data application. The mass and volume of a pebble are regarded an error-free and error-prone variables, respectively. The population mass density is considered to be the slope parameter of the simple linear regression without intercept.

1. Introduction

The classical simple linear regression model is making inferences in the functional relationship between the explanatory or independent variable X and the response or dependent variable Y from the observations (x, y). Sometimes, the explanatory variable cannot be directly observable or difficult to observe for some situations. In these situations, a substitute variable W, generally called error-prone predictor, is observed instead of X that is, the random variable X is observed with measurement error U. The substitution of W for X leads to estimates that are sometimes seriously biased. The goal of the measurement error modeling is to obtain unbiased estimates with observed data (w, y).

Consider the classical linear regression model with one explanatory variable as

$$Y = \alpha + \beta_X X + \varepsilon \tag{1.1}$$

when experimental error ε with mean 0, variance σ_ε^2 and the additive measurement error model as

$$W = X + U \tag{1.2}$$

when measurement error U with mean 0, variance σ_U^2. When the explanatory variable is error-prone predictor the models given in (1.1) and (1.2) together is called

Received by the editors: November 08, 2016, Accepted: January 24, 2017.
2010 *Mathematics Subject Classification.* Primary 05C38, 15A15; Secondary 05A15, 15A18.
Key words and phrases. Classical measurement error model, consistent estimator, error in variables, linear regression, mass density, regression calibration, SIMEX method, M-estimation,

classical measurement error model Carroll, Ruppert, Stefanski and Crainiceanu (2006). (ε, U, X) is an independent triplet with the distribution

$$\begin{pmatrix} \varepsilon \\ U \\ X \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 & 0 & 0 \\ 0 & \sigma_U^2 & 0 \\ 0 & 0 & \sigma_X^2 \end{pmatrix} \right\} \tag{1.3}$$

For the error-free data , the usual bivariate normal regression model given in (1.1) the normal estimating equations for α, β_X and σ_ε^2 can be derived from the conditional distribution of $Y_1, Y_2, \cdots, Y_n$ given $X_1, X_2, \cdots, X_n$ for the random sample $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ of size n as in Casella and Berger (1990).

$$\sum_{i=1}^{n} (Y_i - \alpha - \beta_X X_i) \begin{pmatrix} 1 \\ X_i \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

$$\sum_{i=1}^{n} \left\{ \frac{(n-p)}{n} \sigma_\varepsilon^2 - \{Y_i - \alpha - \beta_X X_i\}^2 \right\} = 0.$$

For error-prone data, models in (1.1) and (1.2) can be rewritten as

$$\begin{aligned} Y_i &= \alpha + \beta_X X_i + \varepsilon_i; \quad i = 1, 2, \cdots, n \\ W_i &= X_i + U_i; \quad i = 1, 2, \cdots, n \end{aligned}$$

for the random sample $(W_1, Y_1), (W_2, Y_2), \cdots, (W_n, Y_n)$ of size n.

The situation is that the explanatory variable X is measured as W, i.e. ignoring the measurement error, and modeling the regression of Y on W using the model in (1.1) causes impairments of statistical inferences such as biased estimation. To be specific the effect of the measurement error on the estimating equations is to bias on the slope estimate in the direction of 0. This type bias is commonly referred to as the attenuation in the context of the simple linear regression. The amount of the attenuation is called reliability ratio as in Fuller (1987), Carroll, Ruppert, Stefanski and Crainiceanu (2006) and denoted by λ.

The ordinary least squares (OLS) slope estimator $\hat{\beta}_W$ for the regression of Y on W is called the naive estimator and the OLS slope estimator $\hat{\beta}_X$ for the regression of Y on X is called the true estimator. Let us define S_{YW}, S_{YU} and S_{XU} are the sample covariances of Y and W, Y and U, and X and Urespectively. Similarly, $S_{UU} = S_U^2$ and $S_{XX} = S_X^2$ the sample variances of U and X, respectively. The ordinary least square estimator on the observed data (Y, W) is written as $\hat{\beta}_W = \hat{\lambda}\hat{\beta}_X + o_p(1)$, where the estimator of the reliability ratio is $\hat{\lambda} = S_X^2/(S_X^2 + S_U^2)$ and $o_p(1)$ indicates that the remainder term converges in probability to zero. In order to show that firstly, consider the naive ordinary least square estimator of slope parameter as

$$\hat{\beta}_W = \frac{S_{YW}}{S_{WW}} = \frac{S_{YX} + S_{YU}}{S_{XX} + S_{UU}}.$$

Secondly, by the Law of Large numbers, S_{XU} and S_{YU} converge in probability to zero under the independence assumption, and likewise $S_{XX} \xrightarrow{p} \sigma_X^2$, $S_{UU} \xrightarrow{p} \sigma_U^2$, $S_{YX} \xrightarrow{p} \sigma_{YX}$ as $n \to \infty$. From these results, $\hat{\lambda} \to \lambda = \sigma_X^2/(\sigma_X^2 + \sigma_U^2)$ as $n \to \infty$ and the regression slope parameter is $\beta_X = \sigma_{YX}/\sigma_X^2$, therefore $\hat{\beta}_W \xrightarrow{p} \lambda\beta_X$ as $n \to \infty$ (Serfling, 1980). The attenuation factor λ is a real number in the range [0,1] since σ_X^2, σ_U^2 are finite. If $\sigma_X^2 > 0$ and finite, then $\sigma_U^2 = 0 \Leftrightarrow \lambda = 1$. In this situation, there is no measurement error, say, $X = W$. If $\sigma_U^2 = \infty \Leftrightarrow \lambda = 0$, then the data is all error.

To illustrate the attenuation induced by the measurement error, the data for the true explanatory variable X, the regression model error ε, and the measurement error U were generated from the trivariate normal distribution as

$$\Big(X, U, \varepsilon\Big)^T \sim N\Big((0,0,0)^T, \text{diag}\,\{1, 0.5, 0.5\}\Big).$$

The data for the response variable Y were generated with the regression model, $Y = \alpha + \beta_X X + \varepsilon$ with $\alpha = 0$, $\beta_X = 3$ and the observed data W were obtained from $W = X + U$. Notice that how to the true data (X_i, Y_i)'s are more tightly grouped around a well-delineated line, while the error-prone data (W_i, Y_i) have much variability about the dashed line in Figure 1.

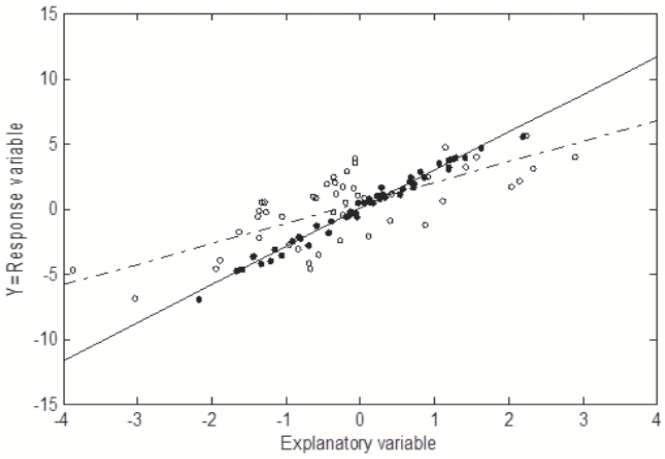

FIGURE 1. Illustration of the simple linear regression with measurement error. The filled circles are the plot of the true data (X, Y), and the dashed line is the least squares fit of these data. The empty circles are the plot of the observed data (W, Y) and the solid line, which is the attenuated line, is the least squares fit of the measurement error data. For these data $\sigma_X^2 = 1$, $\sigma_U^2 = \sigma_\varepsilon^2 = 0.5$ and $(\alpha, \beta_X) = (0, 3)$.

The results of fitting the model with measurement error can be summarized roughly as fallows. The regression model for $E(Y|X)$ depends on unobservable explanatory variable X instead of observable substitute variable W to X. As a consequence, the estimators of the parameters of interest $\theta = (\alpha, \beta_X)^T$ appear in the the model are functions of the observed substitute variable W. The data have also additional variability because of the measurement error. Thus, it is difficult to find unbiased estimator of the parameters of interest with the substitute variable W instead of X. On the other hand, one of the feature of the measurement model has a lack of identifiability problem (Fuller (1987), pp. 9-10). Because, the estimation of measurement error variance, $\hat{\sigma}_U^2$, can not be obtained with the data (W, Y) at hand, so for this estimation it is required replicated data unless it is known. As a results, not only is the regression slope estimator biased and the fitted line attenuated, but also the data are noisier with increased error about the fitted line. In this manuscript, two methods of correcting the attenuation, the regression calibration and simulation of extrapolation called SIMEX, are explained in more detail and compared in terms of the attenuation and variability. For the illustration a simulation study is presented for different sample sizes and different values of $\sigma_U^2, \sigma_\varepsilon^2$. Moreover, a real data example is given for an application to linear regression without intercept.

2. The Methods of Estimates

2.1. **Regression Calibration.** The regression calibration (RC) is a straightforward method for fitting the regression models in the presence of measurement error and was derived and recommended by Carroll and Stefanski (1990) and Gleser (1990). The RC is one of the most useful methods to reduce the effect of measurement error and correcting the attenuation in regression model. The basis of RC is the replacement of the true explanatory variable X by the estimation of $E(X|W)$, which is denoted as $\widehat{E(X|W)}$ and also will be called as RC function. After this replacement, the regression analysis is performed on $\left(\widehat{E(X|W)}, Y\right)$. RC is simple, widely used, effective, reasonably well investigated and potentially applicable in addition to correcting the attenuation (Carroll and Stefanski (1990, 1994)).

Carroll and Stefanski (1990) suggested an algorithm yielding a linear approximation to the RC estimate to eliminate bias in the estimated regression coefficients for measurement error analysis. To operate the algorithm, measurement error variance σ_U^2 has to be known or estimable. If the data are replicated externally or internally to estimate the error variance, then the algorithm is applicable. RC estimate of β_X can be derived in two steps:

- The mean squares of model error (MSE) for fitting the regression Y on W is taken $\hat{\sigma}_W^2$. If there is only one explanatory variable in the analysis like this article, $\hat{\sigma}_W^2$ is the sample variance of W.
- RC estimate is $\hat{\beta}_X = \hat{\beta}_W \hat{\sigma}_W^2 / (\hat{\sigma}_W^2 - \hat{\sigma}_U^2)$.

With the replicated data, it is possible to estimate the measurement error variance σ_U^2. Replicate data means that measurement of X is replicated measurements W measuring the same X. Suppose $W_{i1}, W_{i2}, \ldots, W_{ik_i}$ are k_i replicated measurements of X_i, and their mean is $\overline{W}_{i.}$. The replication provides to obtain the estimate of measurement error variance as

$$\widehat{\sigma}_U^2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k_i}(W_{ij} - \overline{W}_{i.})^2}{\sum_{i=1}^{n}(k_i - 1)}. \tag{2.1}$$

Given $\overline{W}$, the best linear approximation to X is

$$E(X\,|\overline{W}\,) \approx \mu_X + \sigma_X^2\left[\left(\sigma_X^2 + \sigma_U^2\right)/k\right]^{-1}\left(\overline{W} - \mu_W\right),$$

where k is the number of the replication of X, μ_X and μ_W are the means of X and W, σ_X^2 and σ_U^2 are the covariance matrices of X and U, respectively (Carroll, Ruppert, Stefanski and Crainiceanu (2006)).

For the RC estimate, the sample mean of the replicated data of X_i is $\overline{W}_{i.} = \sum_{j=1}^{k_i} W_{ij}/k_i$ and the pooled sample variance of the replicated data is given in (2.1). Similarly, the other estimates are defined as

$$\widehat{\mu}_X = \widehat{\mu}_X = \frac{\sum_{i=1}^{n} k_i\overline{W}_{i.}}{\sum_{i=1}^{n} k_i},\ v = \sum_{i=1}^{n} k_i - \frac{\sum_{i=1}^{n} k_i^2}{\sum_{i=1}^{n} k_i},$$

$$\widehat{\sigma}_X^2 = \left[\left\{\sum_{i=1}^{n} k_i\left(\overline{W}_{i.} - \widehat{\mu}_W\right)^2\right\} - (n-1)\widehat{\sigma}_U^2\right]/v.$$

resulting RC estimate is

$$\widehat{X}_i = \widehat{E(X_i\,|\overline{W}_{i.}\,)} \approx \hat{\mu}_W + \widehat{\sigma}_X^2\left[\frac{\widehat{\sigma}_X^2 + \widehat{\sigma}_U^2}{k_i}\right]^{-1}\left(\overline{W}_{i.} - \hat{\mu}_W\right),\ i = 1, 2, \ldots, n. \tag{2.2}$$

The estimated RC in (2.2) is reproduced by replacing the unknown parameters by their classical method of moments estimators in the best linear approximation to X given above. To derive RC estimate it is required to have the estimates $\widehat{\sigma}_U^2$ and $\widehat{\sigma}_W^2$ from observed data. If the data are not replicated or unavailable to replicate, and but there is an estimate σ_U^2, gotten from an another study, still the estimate $\widehat{X}_i$ can be obtained from equation (2.2). Even if there are exactly two replicates of W, then the sample variance of U is derived from the half of the sample variance of difference $W_{i1} - W_{i2}$. Thus, the estimated RC is attained as in (2.1). When the

estimate of variance σ_U^2 is derived from replicated data, the covariance $\sigma_{\varepsilon U}$ of ε and U is assumed to be zero since the independence of ε and U. When the replication is not available for each observation, the algorithm for RC estimator produces consistent estimates for linear regression. After reducing the measurement error in the explanatory variable, then the regression parameters are estimated and the statistical inference proceeds with a standard analysis. SIMEX, given in the next section has the same advantages with RC, but it is more computationally intensive than RC.

2.2. **Simulation Extrapolation.** Simulation Extrapolation (SIMEX) is a simulation based method of estimating and correcting the attenuation due to the measurement error. SIMEX method was proposed and developed by Cook and Stefanski (1994), Stefanski and Cook (1995) as an alternative method to reducing bias. SIMEX estimation is a computational, graphical method and depends on a computer algorithm that determines parameter estimates.The essential idea is to determine the bias for an estimate caused by the measurement error by implementing a virtual experiment via simulation then the unbiased estimation is found on the graph for the no measurement error case.

SIMEX estimates are obtained by adding additional measurement errors to the observed values of the explanatory variable W in a resampling-like stage and recalculating the naive estimators from the contaminated data. For each additional measurement error, the naive estimator is obtained and the trend is then extrapolated back to the case of no measurement error. The details of the algorithm are given in this section for simple linear regression.

The key features of SIMEX method are described easily for the simple linear regression. For this, the notation is adopted from Carroll, Ruppert, Stefanski and Crainiceanu (2006). Suppose that simple linear regression $Y = \alpha + \beta_x X + \varepsilon$, with additive measurement error model $W = X + U$, where U is independent of (Y, X) and has mean zero and variance σ_U^2. For the calculation purpose, assume $U = \sigma_U Z$, where Z is a standard normal random variable. σ_X^2 denotes the variance of the explanatory variable X, the measurement error variance σ_U^2 assumed to be known or to be estimated. Let now the additive measurement error model be $W = X + \sigma_U Z$. When the measurement error variance is ignored, it is well known that the ordinary least square estimate $\hat{\beta}_W$, denotes the naive estimator, of β_X converges in probability to $\dfrac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2}\beta_X$ as $n \to \infty$, but not to β_X.

The key idea of SIMEX method is to obtain the ordinary least square estimate of slope from the original data that is the naive estimate $\hat{\beta}_W$. There are $M-1$ additional data sets with successively added measurement error that each set has the variance $\sigma_U^2(1+\nu_m)$, $m = 1, 2, \cdots, M$ where $0 = \nu_1 < \nu_2 < \cdots < \nu_M$. In the following the set of ν_ms is denoted by Λ. For any mth data set, the ordinary least square estimate of slope is calculated and the estimator $\hat{\beta}_{W,m}$ consistently estimates

$\frac{\sigma_X^2}{(\sigma_X^2+(1+\nu)\sigma_U^2)}\beta_X$. Note that, for $\nu=-1$ the naive estimator turns out to be an unbiased estimator of β_X. This suggest that the relationship between $\hat{\beta}_W$ and ν can be formulated as a nonlinear regression model. Therefore, ν is taken as if an independent variable and $\hat{\beta}_W$ is taken as if a dependent variable. The model has a mean function of the form

$$E(\hat{\beta}_W|\nu)=g(\nu)=\frac{\sigma_X^2}{\sigma_X^2+\sigma_U^2(1+\nu)}\beta_X \ ,\nu\geqslant 0. \tag{2.3}$$

A generic plot of ν versus $g(\nu)$ is obtained as in Figure 2. The parameter of interest, β_X is achieved from the function $g(\nu)$ by extrapolation to $\nu=-1$ (Carroll, Ruppert, Stefanski and Crainiceanu (2006)). Cook and Stefanski (1994)

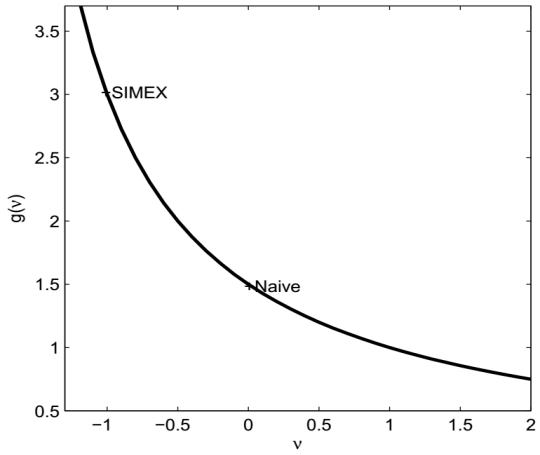

FIGURE 2. A generic SIMEX plot of the effect of measurement error of size $\sigma_U^2(1+\nu)$ on parameter estimates. The SIMEX estimate is an extrapolation to $\nu=-1$ and the naive estimate occurs at $\nu=0$.

showed that equation (2.3) fits the nonlinear function $g(\nu)=\gamma_0+\gamma_1(\gamma_2+\nu)^{-1}$ of ν named as the rational linear extrapolant that generates consistent estimator of the parameter of interest. For the unbiased parameter estimation let us recall a commonly used M-estimation method via score function ψ which satisfies the

$$E\left[\psi(Y,X;\theta)\,|X\right]=0$$

where $Y=g(\nu)$, $X=\nu$ and $\theta=(\gamma_0,\gamma_1,\gamma_2)^T$ are considered as response, explanatory variables and the vector of parameters, respectively.

The conditionally unbiased ψ function can be devised using the M-estimation methods described by Carroll, Ruppert, Stefanski and Crainiceanu (2006, Sec.7.3).

The parameter θ relating Y and X is consistently estimated by $\hat{\theta}$ satisfying the estimating equation

$$\sum_{i=1}^{n} \psi(Y_i, X_i; \theta) = 0. \tag{2.4}$$

The SIMEX algorithm suggested by Cook and Stefanski (1994) can be summarized in the following steps:

- Fix the set Λ and choose $\nu_m \in \Lambda$.
- Take a constant $B > 0$ and generally considered as $B = 50, 100, 500$.
- Generate the random variables $Z_{ib} \sim N(0,1)$ via computer for $b = 1, \cdots, B$ and $i = 1, \cdots, n$.
- Define the variance $Var(W_i + \sigma_U \nu_m^{1/2} Z_{ib} \,|X_i\,) = (1 + \nu_m)\sigma_U^2, i = 1, \cdots, n$.
- Define the estimate $\hat{\theta}_b(\nu_m)$ be a solution of $\sum\limits_{i=1}^{n} \psi(Y_i, W_i + \sigma_U \nu_m^{1/2} Z_{ib}; \theta) = 0$ for each b.
- Average these estimations as

$$\hat{\theta}_S(\nu_m) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b(\nu_m),$$

  where the subscript S refers to the simulation nature of the estimator.
- Repeat the steps for $m = 1, 2, \cdots, M$ and find $\hat{\theta}_S(\nu_m)$ for each m.
- Plot the generated data of pairs $(\nu_m, \hat{\theta}_S(\nu_m))$.
- Fit the parametric model $g(\theta, \nu) = \gamma_0 + \gamma_1(\gamma_2 + \nu)^{-1}$ by $(\nu_m, \hat{\theta}_S(\nu_m))$ and estimate $\theta = (\gamma_0, \gamma_1, \gamma_2)^T$.
- Find the SIMEX estimator as $\hat{\theta}_{\text{SIMEX}} = g(\hat{\theta}, -1)$

When $\nu = 0$, the SIMEX algorithm produces the estimator $\hat{\theta}_S(0)$, which denotes the naive estimator the same as the method of moments estimator. The estimating equation in (2.4) satisfies

$$E\Big[\psi\left(Y, W + \sigma_U \nu^{1/2} Z; \theta\right)\Big] = 0.$$

Therefore, the parameter estimator $\hat{\theta}_S(\nu)$ converges in probability to $\theta(\nu)$ by the standard estimating equation theory (Serfling, 1980).

3. Simulation Study

In this section, the performance of regression calibration and SIMEX methods are illustrated by a simulation study to eliminate the effects of measurement error on the parameter estimation. Throughout the simulation study measurement error variance, σ_U^2 is assumed to be known. The data $\{X_i, U_i, \varepsilon_i\}_{i=1}^{n}$ are generated from trivariate normal distribution as in (1.3) with $\mu_X = 0$ and $\sigma_X^2 = 1$ for the selected the measurement error variances $\sigma_U^2 = \{0.25, 0.5, 0.75, 1.0, 1.5, 2.0\}$, the model error variances $\sigma_\varepsilon^2 = \{0.5, 1.0\}$, sample sizes $n = 50, 100, 200$, and $B = 100$ simulation

runs. The data for the response variable Y and the observed variable W were created by using the generated data for the models given in (1.1),(1.2) with $\alpha = 0$, $\beta_X = 3$. The simulation results are given only for the estimated parameter β_X in Table 1 and Table 2. The slope estimations depending on the methods are listed below with the associated data:

- True estimation calculated from the true data $\{Y_i, X_i\}_{i=1}^n$,
- Naive estimation calculated from the observed data $\{Y_i, W_i\}_{i=1}^n$,
- RC estimation calculated from the observed data $\{Y_i, W_i\}_{i=1}^n$,
- SIMEX estimation calculated from the observed data $\{Y_i, W_i\}_{i=1}^n$.

TABLE 1. Simulation study results for the true, naive, RC and SIMEX estimators for $\sigma_U^2 = \{0.25, 0.5, 0.75, 1.0, 1.5, 2.0\}$, $\sigma_\varepsilon^2 = 0.5$, $n = \{50, 100, 200\}$, and $(\alpha, \beta_X) = \{0, 3\}$.
The table entries are means of 100 simulation runs.

$\sigma_\varepsilon^2 = 0.5$		σ_u^2=0.25	σ_u^2=0.5	σ_u^2=0.75	σ_u^2=1	σ_u^2=1.5	σ_u^2=2
n	Estimator	$\beta_X = 3$	$\beta_X = 3$	$\beta_X = 3$	$\beta_X = 3$	$\beta_X = 3$	$\beta_X = 3$
50	True	2.9988	3.0000	3.0005	3.0008	3.0006	3.0002
	Naive	2.9832	2.9752	2.9518	2.9405	2.9080	2.8914
	RC	2.9986	3.0055	2.9976	3.0015	2.9992	3.0094
	SIMEX	2.9982	3.0049	2.9969	3.0010	2.9982	3.0071
100	True	3.0011	3.0002	3.0010	2.9996	3.0007	3.0006
	Naive	2.9857	2.9744	2.9543	2.9391	2.9077	2.8717
	RC	3.0012	3.0057	3.0003	2.9992	2.9977	2.9890
	SIMEX	3.0010	3.0053	3.0001	2.9993	2.9968	2.9888
200	True	2.9998	2.9997	2.9995	2.9996	3.0000	2.9999
	Naive	2.9827	2.9675	2.9499	2.9392	2.9084	2.8535
	RC	2.9983	2.9984	2.9962	3.0005	2.9988	3.0062
	SIMEX	2.9982	2.9983	2.9960	2.9999	2.9987	3.0055

TABLE 2. Simulation study results for the true, naive, RC and SIMEX estimators for $\sigma_U^2 = \{0.25, 0.5, 0.75, 1.0, 1.5, 2.0\}$, $\sigma_\varepsilon^2 = 1$, $n = \{50, 100, 200\}$, and $(\alpha, \beta_X) = \{0, 3\}$. The table entries are means of 100 simulation runs.

$\sigma_\varepsilon^2 = 0.5$		σ_u^2=0.25	σ_u^2=0.5	σ_u^2=0.75	σ_u^2=1	σ_u^2=1.5	σ_u^2=2
n	Estimator	$\beta_X = 3$	$\beta_X = 3$	$\beta_X = 3$	$\beta_X = 3$	$\beta_X = 3$	$\beta_X = 3$
50	True	2.9999	2.9987	3.0048	3.0001	3.0003	3.0014
	Naive	2.9834	2.9662	2.9608	2.9391	2.9124	2.8887
	RC	2.9988	2.9972	3.0072	3.0008	3.0018	3.0087
	SIMEX	2.9986	2.9967	3.0064	2.9998	3.0004	3.0076
100	True	3.0010	2.9995	3.0004	3.0018	3.0032	2.9997
	Naive	2.9836	2.9714	2.9518	2.9363	2.9177	2.8816
	RC	2.9994	3.0022	2.9981	2.9964	3.0087	2.9985
	SIMEX	2.9993	3.0022	2.9977	2.9958	3.0084	2.9990
200	True	3.0005	2.9997	3.0002	3.0001	2.9998	2.9997
	Naive	2.9845	2.9680	2.9532	2.9375	2.9094	2.8808
	RC	2.9999	2.9989	2.9994	2.9982	3.0006	3.0005
	SIMEX	2.9998	2.9989	2.9991	2.9981	3.0001	3.0001

The true, naive, RC and SIMEX estimations were compared in terms of bias. As seen in tables, RC and SIMEX methods eliminate the attenuation due to measurement error. That means, both methods correct the bias of the estimates of regression parameters as well as true estimation. When the sample size increases, SIMEX method gives slightly better estimates than RC.

4. An Application

The interest is to find the density of pebble population in one of the coasts of Antalya. Pebbles are different colors (granite or white, etc.) which reflect their texture and density. For the application purpose the sample pebbles are randomly selected from the population of pebbles which have the same color and texture, namely granite colored. So, the density of a granite pebble can be obtained as

$$\rho = \frac{m}{V} \;\Rightarrow\; m = \rho V$$

where ρ=density, V=volume(cm^3) and m= mass(g). The volume and mass of each pebble are measured with measuring cylinder and a very sensitive weight scale, and then their densities can be obtained from the density formula given above. However, the volume of a pebble is not easily measured even though it is measured with a very sensitive instrument. The volume measurement can be considered as an error-prone variable; therefore it is possible to be modeled as in (1.2). Throughout the application the volume of the pebbles are assumed to be never measured accurately. On the other hand the population density ρ can be estimated as a slope parameter of a simple linear regression model without intercept.

For the analysis purposes, suppose that the measurement error of volume is $U \sim N(0, \sigma_U^2)$. If the measurement error variance is unknown, it has to be estimated by

a replicated sample. For this aim, a metal sphere with known volume $V = 9.20\ cm^3$ is measured several times to calculate the measurement error variance using this replicated data given in Table 3. For this data the measuring cylinder filled with water randomly and measured the level of the water referred as "Before" then the metal sphere put inside the measuring cylinder and measured the level of the water referred as "After". The difference After and Before denotes the volume of the metal sphere for each measurement replication. Note that the each value of "Difference" varies even if it is measured the same metal sphere which indicates that volume measurand has a measurement error.

TABLE 3. Measurements of a sphere has volume 9.20 cm^3, measured with a measuring cylinder.

Before	After	Difference
53.00	61.50	8.50
64.00	73.75	9.75
45.00	53.75	8.75
26.00	44.75	8.75
41.25	49.75	8.50

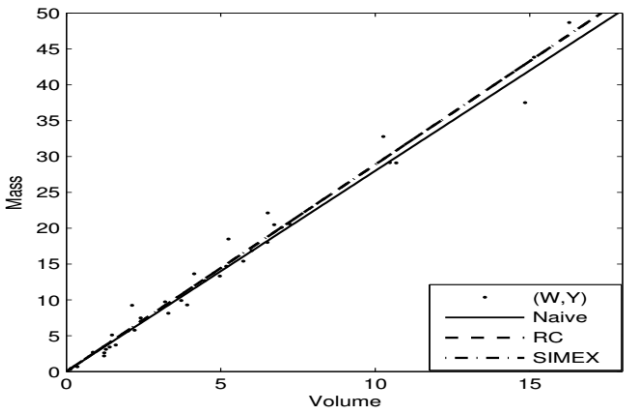

FIGURE 3. The estimated regression lines of the pebbles data for naive, RC and SIMEX estimator

From the replicated data given in Table 3, the measurement error variance σ_U^2 is estimated as 0.52. The volume and the mass of pebbles are called as $W = V$ and $Y = m$, respectively to be compatible with the models (1.1) and (1.2). To fit the data, the simple linear regression model and the additive measurement error model

are

$$\begin{aligned} Y_i &= \rho X_i + \varepsilon_i,\ i = 1, 2, \ldots, 38 \\ W_i &= X_i + U_i,\ i = 1, 2, \ldots, 38 \end{aligned}$$

The estimated regression lines using the pebbles data are $\hat{y}_i = 2.8575 w_i$ with the OLS estimator, $\hat{y}_i = 2.8868 w_i$ with the RC estimator and $\hat{y}_i = 2.8869 w_i$ with the SIMEX estimator.

The three estimated regression lines of OLS, RC and SIMEX appear to be very close in Figure 3. The slope estimations of RC and SIMEX produce relatively close to each other than the OLS. There are some possible reasons for indistinctiveness of the estimated lines such as small measurement error variance, small pebble sizes, small sample size etc. The estimated reliability ratio is $\hat{\lambda} = 0.9898$ for the current application. It seems that the effect of measurement error will be more apparent as the pebble sizes increase.

References

[1] Carroll, R.J & Stefanski, L.A. , Approximate quasilikelihood estimation in models with surrogate predictors, *Journal of the American Statistical Association*, (1990), 85, pp. 652-663.
[2] Carroll, R.J., Ruppert, D., Stefanski, L.A.& Crainiceanu, C.M., Measurement Error in Nonlinear Models, 2nd edn.Chapman & Hall/CRC 2006.
[3] Casella, G. & Berger, R. L. Statistical Inference, Duxbury Press, Belmont, 1990.
[4] Cook, J.R. & Stefanski, L.A., Simulation Extrapolation Estimation in Parametric Measurement Error Models, *Journal of the American Statistical Association*, (1994), 89, pp. 1314-1328.
[5] Fuller, W.A., Measurement Error Models, John Wiley and Sons, New York, 1987.
[6] Gleser, L.J., Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models. In Statistical Analysis of Error Measurement Models and Application, P. J. Brown and W. A. Fuller, ed., Providence: American Mathematics Society, 1990.
[7] Serfling, R.J., Approximation Theorems of Mathematical Statistics, John Wiley and Sons, Singapore, 1980.
[8] Stefanski, L.A. & Cook, J.R., Simulation-Extrapolation: The Measurement Error, *Journal of the American Statistical Association*, (1995), 90, pp. 1247-1256.

Current address: Rukiye E. Dağalp (Corresponding author): Ankara University, Faculty of Sciences, Department of Statistics, 06100 Tandoğan-Ankara/Turkey.

E-mail address: rdagalp@ankara.edu.tr

Current address: İhsan Karabulut:Ankara University, Faculty of Sciences, Department of Statistics, 06100 Tandoğan-Ankara/Turkey.

E-mail address: kbulut@science.ankara.edu.tr

Current address: Fikri Öztürk:Ankara University, Faculty of Sciences, Department of Statistics, 06100 Tandoğan-Ankara/Turkey.

E-mail address: ozturk@science.ankara.edu.tr