Proc. ICCS-11, Lahore, Pakistan
December 19-22, 2011, Vol. 21, pp. 39-53

INSTRUMENTAL VARIABLE ESTIMATOR OF THE SLOPE PARAMETER WHEN THE EXPLANATORY VARIABLE IS SUBJECT TO MEASUREMENT ERROR

Anwar Saqr and **Shahjahan Khan**
Department of Mathematics and Computing
Australian Centre for Sustainable Catchments
University of Southern Queensland Toowoomba
Australia.
Email: anwar.saqr@usq.edu.au and khans@usq.edu.au

ABSTRACT

This paper proposes a new instrumental variable to estimate the parameters of a simple linear regression model where the explanatory variable is subject to measurement error. The new instrumental variable is defined using reflection of the observed values of the explanatory variable. Like other instrumental variable estimators it is unbiased and consistent, but over performs estimators proposed by Wald (1940),Bartlett (1949), and Durbin (1954) if the ratio of the error variances is equal or less than one $\lambda \leq 1$. The method is straightforward, easy to implement, and performs much better than the existing instrumental variable based estimators. The theoretical superiority of the proposed estimator over the existing instrumental variable based estimators is established by analytical results of simulation. Two illustrative examples for numerical comparisons of the results are also included.

Keywords and phrases: Regression parameters, Measurement error models, Instrumental variable, Reflection of point; Reliability ratio; Sum of square error.
2010 Mathematics Subject Classification: 62F10 and 62J05.

1 Introduction

The linear regression model is arguably the most frequently used statistical tool in various fields of scientific investigations, including bioassays and econometric studies. Commonly used bioassays where regression model can be used may include prediction of the body weight based on body fat, or the yield of a crop based on soil moisture level. However, measuring the explanatory variable, namely, the body fat or soil moisture level is likely to involve measurement errors. The ordinary least squares (OLS) estimator of the regression

[1]On leave from Department of Statistics, AlJabal AlGarby University, Gharian, LIBYA

parameters is inappropriate (biased and inconsistent) in the presence of measurement error or error in variable. As a result, in real life, measurement error pauses a serious problem, as it directly impacts on estimators of the parameters and their standard error. It is well known that the measurement error in the response variable is not as serious as it is in the explanatory variable. The errors in the response variable can be absorbed in the error term of the model. The error in the explanatory variable causes various problems, and requires to be handled appropriately.

The measurement error is a real problem and it has been considered by a host of authors since the third quarter of the nineteenth century. Adcock (1877, 1878) discussed the problem in the context of least squares method. Pearson (1901) suggested some estimators based on Adcock's work. The problem has been seriously considered by researchers from the later part of the first half of the last century. Wald (1940), Bartlet (1949), Durbin (1954), and Riggs et al. (1978) considered fitting regression line when both variables are subject to error. Berkson (1950) noted that if there is error in the explanatory variable the bias in the estimated regression line will be there regardless of the data being a random sample or the population. Burr (1988) considered error in explanatory variable for the binary responses model. Freedman et al. (2004) suggested a reconstructed moment base method to deal with error in the explanatory variable. The problem of error in both explanatory and response variables was considered by Geary (1942), Madansky (1959) and Halperin (1961).

Instrumental variable (IV) technique is a well known method to obtain unbiased and consistent estimators in the presence of measurement error in the explanatory variable. The method requires to define an IV that is uncorrelated with the model error but highly correlated with the explanatory variable. The grouping method was first suggested by Wald (1940), followed by Bartlett (1949) and then Durbin (1954). Maddala (1988) showed that Wald method is equivalent to using the instrumental variable Z equal -1 and +1 for values less than or greater than the median of the *manifest* variable, Bartlett proposed to divide the values in three equal groups and use the first and third groups, and Durbin used the ranks of the values to define the IV. In each of the method there is loss of information (for not using actual values and dropping some of the data points), and there are different formulae to find the sum of squares error, and hence lead to different mean sum of square error, making the analysis incomparable.

In this paper we propose a new way to define IV using the *reflection* of the explanatory variable. The estimator based on this method is unbiased and consistent. Moreover, it allows to define the sum of squares error uniquely, same way as in the case of no measurement error. In addition there is no loss of information in this method. Both analytical results and numerical illustrations confirms the superiority of the proposed (instrumental variable) estimator over the existing estimators.

Degracie and Fuller (1972) considered estimation of the slope and covariance when the concomitant variable is measured with error. Grubbs (1973) discussed errors of measurement, precision, accuracy and the statistical inference. Aigner (1973) considered regression with a binary variable subject to errors of observation. Florens et al. (1974) considered

Bayesian inference in error-in-variables models. Schneeweiss (1976) proposed consistent estimation of the regression model with errors in the variables. Bhargava (1977) introduced maximum likelihood estimation in a multivariate errors-in-variables regression model with unknown error covariance matrix. Garber and Klepper (1980) extended the classical normal errors-in-variables model. Prentice (1982) dealt with covariate measurement errors and parameter estimation in a failure time regression model. Amemiya et al. (1984) proposed estimation of the multivariate errors-in-variables model with estimated error covariacne matrix. Klepper and Leamer (1984) provided consistent sets of estimates for regression with errors in all variables. Stefanski and Carroll (1985) discussed covariate measurement error in logistic regression. Carroll et al (1985) proposed comparison of least squares and errors-in-variable regression with special reference to randomized analysis of covariance. Armstrong (1985) dealt with the measurement error in the generalized linear model. Bekker (1986) proved comments on identification in the linear errors in variables model. Schafer (1986) combined information on measurement error in errors-in-variables model. Recently Fuller (2006) covered various aspects of the measurement error models and related inferences.

The estimation methods suggested by the above studies make assumption that the variances of the explanatory variable without and with measurement errors or the *reliability ratio* of the two variances are known (cf Fuller 2006, p.5). An alternative assumption is that the variance of the measurement error is known. All these assumptions are unrealistic and the methods based on them are not free from the restrictions imposed by the assumptions.

In this paper we propose a new instrumental variable method to estimate the parameters of the simple regression model without making any of the above assumptions on the variances of the explanatory variable. The proposed method uses the reflection of the *manifest* values of the explanatory variable. The reflection points about the regression line are defined by using transformation formula involving sin and cos functions. The use of the reflections of the observed values of the explanatory variable in defining IV provide a much better estimator of the slope and intercept parameters. It also reduces the mean sum of squares error. The analysis of variance and regression inferences based on the reflections have much better statistical properties than that using the observed values of the explanatory variable, or any other IV estimator.

In the next Section the measurement error regression model is introduced. Section 3 covers the existing estimation methods for the measurement error model. The proposed new estimator based on the reflections of the observed values of the explanatory variable is provided in Section 4. The superior properties of the new estimators are discussed in Section 5. Two numerical illustrations are provided in Section 6, and some concluding remarks are included in Section 7.

2 Measurement error models

In the conventional notation, let X denote the true measurement on the explanatory variable. This is also called the *latent* variable. In the presence of measurement error the actual

observations are different from X. Let M be the observable, or *manifest* variable of the explanatory variable. When the true value of the *latent* variable X is observed, the commonly used classical simple linear regression model is represented by

$$Y_j = \beta_{0x} + \beta_{1x}X_j + e_j, \quad j = 1, 2, \ldots, n, \tag{2.1}$$

where Y_j is the jth realisation of the response variable, X_j is the fixed jth value of the explanatory variable, and e_j is the error variable for $j = 1, 2, \ldots, n$. It is assumed that the model errors are independently distributed according to normal law with constant but unknown variance, that is, $e_j \sim N(0, \sigma_{ee})$.

If there is error in the explanatory variable, the actual observed value, M, is not the 'true' value of the explanatory variable. When the observed value of the explanatory variable contains measurement error, we define

$$M_j = X_j + u_j, \quad j = 1, 2, \ldots, n \tag{2.2}$$

where u_j is the measurement error, and is assumed to be distributed as $N(0, \sigma_{uu})$. Note that, unlike X_j, M_j is a random variable which is assumed to be distributed as $N(\mu_m, \sigma_{mm})$. The model with the fixed X is called the *functional model*, and that with the random or stochastic M is called the *structural model*.

The simple regression model with measurement error in the explanatory variable can be expressed as

$$Y_j = \beta_{0x} + \beta_{1x}M_j + v_j, \quad j = 1, 2, \ldots, n, \tag{2.3}$$

where $v_j = e_j - \beta_{1x}u_j$. Note in equation (2.1) X and e are independent but in equation (2.3) M and v are not independent. So the application of least squares method is not valid for the models with measurement error. Thus, unlike for the model in (2.1), the validity of the estimator of the slope and intercept of the model in (2.3) is not obvious. However, Fuller (2006, p.3) assumes that u_j, X_j and e_j are mutually independent for the estimation of parameters. It also assumes that the *reliability ratio*, $\kappa_{xm} = \sigma_{mm}^{-1}\sigma_{xx}$, where σ_{mm} is the variance of the *manifest* variable M, and σ_{xx} is the variance of the *latent* variable X, is known.

3 Existing Estimators of parameters

The ordinary least squares (OLS) estimator of the regression parameters for the *functional model* are

$$\hat{\beta}_{1x} = \frac{S_{xy}}{S_{xx}}, \text{ and } \hat{\beta}_{0x} = \bar{Y} - \hat{\beta}_{1x}\bar{X}, \tag{3.1}$$

where

$$S_{xy} = \sum_{j=1}^{n}(X_j - \bar{X})(Y_j - \bar{Y}), \quad S_{xx} = \sum_{j=1}^{n}(X_j - \bar{X})^2, \tag{3.2}$$

in which $\bar{Y} = n^{-1}\sum_{j=1}^{n} Y_j$ and $\bar{X} = n^{-1}\sum_{j=1}^{n} X_j$. The estimators of slope and intercept parameters are linear functions of the responses, and they are well known best linear unbiased estimators.

The sampling distribution of the estimator of the regression parameters is given by

$$\begin{pmatrix} \hat{\beta}_{0x} \\ \hat{\beta}_{1x} \end{pmatrix} \sim N_2 \left[\begin{pmatrix} \beta_{0x} \\ \beta_{1x} \end{pmatrix}, \sigma_{ee} \begin{pmatrix} \frac{1}{n} + \frac{\bar{X}^2}{S_{xx}} & \frac{-\bar{X}}{S_{xx}} \\ \frac{-\bar{X}}{S_{xx}} & \frac{1}{S_{xx}} \end{pmatrix} \right]. \tag{3.3}$$

The unbiased estimator of the error variance σ_{ee} is given by $\hat{\sigma}_{ee} = (n-2)^{-1}SSE_e = s_{ee}$, where $SSE_e = \sum_{j=1}^{n}(Y_j - \hat{Y}_j)^2$ in which $\hat{Y}_j = \hat{\beta}_{0x} + \hat{\beta}_{1x}X$ is the estimated value of Y_j. Also, $\sigma_{ee}^{-1}SSE_e$ follows a χ^2 distribution with $(n-2)$ degrees of freedom.

In the presence of measurement error, the M values are observed instead of X.

The least squares method yields the fitted model to be

$$\hat{Y}_m = \hat{\beta}_{0m} + \hat{\beta}_{1m}M, \tag{3.4}$$

where $\hat{\beta}_{1m} = \frac{S_{my}}{S_{mm}} = \hat{\beta}_{1x}k_{xm}$ in which $k_{xm} = \frac{S_{xx}}{S_{mm}}$. It can be easily shown that $\hat{\beta}_{1m}$ is a biased estimator of β_{1x}. Also, the above estimator is not a consistent estimator of β_{1x}. But there are other estimators in the literature that are unbiased and consistent. The instrumental variable (IV) method provides such an unbiased and consistent estimator.

Note that the regression parameters are different for the model with the *manifest* variable than that with the *latent* variable. Even though the aim is to estimate and test β_{0x} and β_{1x}, but in reality one may end up estimating and testing β_{0m} and β_{1m} if we fully rely on M, and over look the presence of the measurement error. In the literature, the regression parameters are estimated, for observed X values, under certain assumptions. One of the assumptions is that the *reliability ratio*, κ_{mx} is known. Fuller (2006) used this assumption for the estimation of the regression parameters for the *functional model*.

3.1 Instrumental variable estimator

In the presence of measurement error in the explanatory variable the IV estimator for the regression parameters is defined as

$$\hat{\boldsymbol{\beta}} = (Z'M)^{-1}Z'\boldsymbol{y}, \tag{3.5}$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)'$ is the vector of estimator of the intercept and slope parameters of the model $M = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ m_1 & m_2 & \cdots & m_n \end{pmatrix}$ and $Z = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_n \end{pmatrix}$ in which z_j's are the values of the second row of the instrumental variable Z. The selection of the values of z_j's require that it is highly correlated with the explanatory variable but uncorrelated with the model errors. The above IV estimator is unbiased and consistent. The variance of the above estimator is given by

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma_{\mathrm{u}}^2(\mathrm{Z'M})^{-1}(\mathrm{Z'Z})(\mathrm{Z'M})^{-1}. \tag{3.6}$$

Obviously the value of the estimator and the variance depend on the choice of Z. For instance, Wald method, as suggested by Maddala (1988), defines Z by assigning z_j to be -1 or +1 depending on m_j being smaller or larger than the median value of the *manifest* variable. The estimator of slope under this choice of IV is $\hat{\beta}_{1W} = \{\bar{Y}_2 - \bar{Y}_1\}/\{\bar{M}_2 - \bar{M}_1\}$, where $\bar{Y}_1$ is the mean of Y-values associated with the values of M less than its median, and $\bar{Y}_2$ is that for the lager than median values of M. Bartlett followed the same selection criterion of z_j's but suggested exclusion of middle $1/3$ of the values, and his estimator is based on the lower and upper $1/3$ of the values of M and associated Y's. The estimator is expressed as $\hat{\beta}_{1B} = \{\bar{Y}_3 - \bar{Y}_1\}/\{\bar{M}_3 - \bar{M}_1\}$, where $\bar{Y}_1$ is the mean of Y-values associated with the smallest $1/3$ of the values of M, and $\bar{Y}_3$ is that for the largest $1/3$. Durbin proposed to use the rank of M as z_j's. His method yields the following estimator of the slope parameter $\hat{\beta}_{1D} = \left[\sum_{j=1}^{n} jY_j\right] / \left[\sum_{j=1}^{n} jm_j\right]$, but does not define the estimator of the intercept.

The IV method of estimation of the regression parameters does not require any unrealistic assumption on the *reliability ratio*. But the actual estimator depends on how the IV is defined, as definition of Z affects both the estimator and its variance. This paper proposes a new method of defining IV based on the *reflection* of the explanatory or manifest variable. In general, the available methods of defining IV causes a significant loss of sample information (data) either by replacing the observed values of the explanatory variable by -1 or +1, or exclusion of some data, or due to ranking of data. But the proposed definition of the IV does not loss any information. Furthermore, the method produces more precise estimator than those proposed by Wald, Bartlett, and Durbin. Moreover, the new estimator based on the reflection of manifest variable is unbiased and consistent.

4 Proposed new IV and estimator

To avoid the unwanted and troublesome influence of the measurement error in the explanatory variable, the idea of *reflection* of the manifest variable is used for all the values of explanatory variable. The *reflection* of the points is taken about the fitted regression line. This is essentially done by a transformation of the observed values of the explanatory variable to their reflection on the Euclidean plane. In the conventional notation, the reflection of the explanatory variable $M_j = X_j + u_j$ (with measurement error u_j) for $j = 1, 2, \ldots, n$, can be defined as

$$X^* = M\cos 2\psi + (Y - \hat{\beta}_{0m})\sin 2\psi, \tag{4.1}$$

where $\hat{\beta}_{0m}$ is the least square estimate of the intercept parameter, ψ is the angle measure defined as $\psi = \arctan \hat{\beta}_{1m}$ in which $\hat{\beta}_{1m}$ is the least square estimate of the slope parameter in the *manifest* model, and cos and sin are the usual trigonometric cosine and sine functions respectively. For the definition of *reflection* points on the Cartesian plane readers may see Vaisman (1997, p. 164-169).

The proposed method requires to compute the reflection of all the data points, and use the transformed values of M, say, X^* in defining the IV to fit the regression line of Y. The

estimator of the slope parameter under the proposed method is

$$\hat{\beta}_{1R} = (Z_r'M)^{-1}Z_r'\boldsymbol{y} = \frac{S_{x^*y}}{S_{mm}}, \tag{4.2}$$

where $Z_r = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1^* & x_2^* & \cdots & x_n^* \end{pmatrix}$ and $S_{x^*m} = S_{mm}$ in which $S_{x^*m} = \sum_{j=1}^{n}(x_j^* - \bar{x})(m_j - \bar{m})$. It can be shown that $\mathrm{Cov}(Z_r, u) = \mathrm{Cov}(Z_r, v) = 0$, that is, the proposed IV is independent of u and v, but very strongly correlated with M.Also it cab be easily shown that $E[M_j] = E[X_j] = E[X_j^*] = \mu_x$.

Theorem 4.1 The estimator of the slope parameter of the simple regression model using IV based on the *reflection* of M is the same as that produced by X, that is, $\hat{\beta}_{1X} = \hat{\beta}_{1R}$.

Proof: From the definition we get

$$\hat{\beta}_{1X} = \frac{S_{xy}}{S_{xx}} = \frac{S_{my}}{S_{xx}} \quad \text{and} \quad \hat{\beta}_{1R} = \frac{S_{x^*y}}{S_{mm}}. \tag{4.3}$$

From (2.2), it is easy to show that $S_{my} = S_{xy}$ and $S_{mm} = S_{xx} + S_{uu}$. But the main body of the proof is based on

$$S_{x^*y} - S_{my} = SSE_m \sin 2\psi, \tag{4.4}$$

where ψ is as defined in equation (4.1), and SSE_m is the sum of squares error for the *manifest* model. The above result follows from the fact that

$$\begin{aligned} x_j^* - m_j &= m_j \cos 2\psi + (y_j - \hat{\beta}_{0m}) \sin 2\psi - m_j \\ &= m_j(\cos 2\psi - 1) + y_j \sin 2\psi - \hat{\beta}_{0m} \sin 2\psi \\ &= -m_j(2\sin^2\psi) + y_j \sin 2\psi - \bar{y} \sin 2\psi + \bar{m} 2\sin^2\psi \\ &= (y_j - \bar{y}) \sin 2\psi - (m_j - \bar{m}) 2\sin^2\psi, \end{aligned} \tag{4.5}$$

where x_j^* is the reflection of m_j. Multiplying both sides of the above equation by y_j and taking sum over j, we get

$$\begin{aligned} \sum(x_j^* - m_j)y_j &= \sum(y_j - \bar{y})y_j \sin 2\psi - \sum(m_j - \bar{m})y_j 2\sin^2\psi \\ S_{x^*y} - S_{my} &= S_{yy} \sin 2\psi - S_{my} 2\sin^2\psi \\ \frac{S_{x^*y} - S_{my}}{\sin 2\psi} &= SST - SSR_m = SSE_m, \end{aligned} \tag{4.6}$$

where $S_{yy} = SST$ is the sum of squares total, SSR_m is the sum of squares regression, and SSE_m is the sum of squares error for the regression of Y on M. Note that $\frac{2\sin^2\psi}{\sin 2\psi} = \tan\psi = \hat{\beta}_{1m}$.

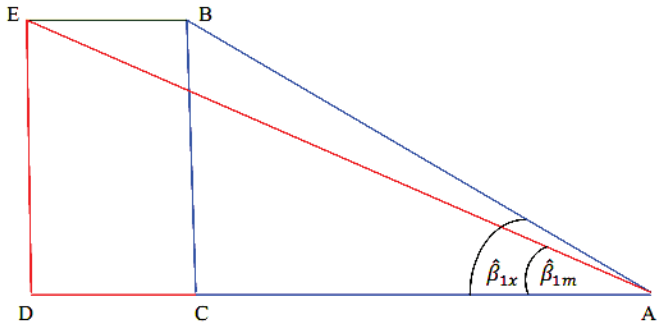

Figure 1: Graph representing the Sum of Squares and Products in the presence of measurement error in the explanatory variable.

Then using equation (4.6), we can write

$$\hat{\beta}_{1x} = \frac{S_{xy}}{S_{xx}} = \frac{S_{my}}{S_{xx}} = \frac{S_{x^*y} - SSE_m \sin 2\psi}{S_{mm} - S_{uu}}$$

$$\hat{\beta}_{1R} = \frac{S_{x^*y}}{S_{mm}} = \frac{S_{my} + SSE_m \sin 2\psi}{S_{xx} + S_{uu}} = \frac{S_{xy} + SSE_m \sin 2\psi}{S_{xx} + S_{uu}}$$

From Figure 2 $\angle FAD = \angle FBE$ then

$$\hat{\beta}_{1R} = \frac{S_{x^*y}}{S_{mm}} = \frac{S_{x^*y} - S_{xy}}{S_{mm} - S_{xx}} \tag{4.7}$$

which leads to

$$S_{x^*y} S_{xx} = S_{xy} S_{mm}, \tag{4.8}$$

and finally simplification yields

$$\frac{S_{x^*y}}{S_{mm}} = \frac{S_{xy}}{S_{xx}} \text{ or, } \hat{\beta}_{1\mathrm{R}} = \hat{\beta}_{1\mathrm{X}}. \tag{4.9}$$

Hence the proof

4.1 Geometric Explanation

The presence of measurement error in the explanatory variable and its impact on the estimator of the slope as well as how the proposed method 'treats' the measurement error can be explained by graphs. The graphical representation also explains how the actual estimator of the slope is recovered by the new method.

Figure 1 represents the sum of squares and sum of products associated with the definition of the estimators of slope both for the *latent* and *manifest* variables. This graph represents

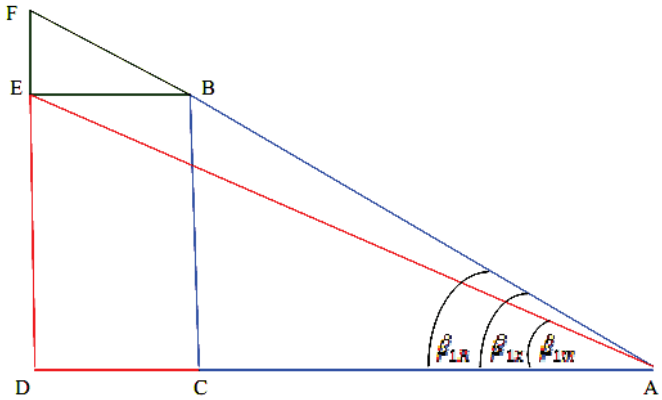

Figure 2: Graph representing the Sum of Squares and Products when the measurement error in the explanatory variable is 'treated' by reflection.

the presence of measurement error in the explanatory variable as well as the two estimators of the slope parameter. On the other hand Figure 2 displays the same along with that of the reflection of the *manifest* variable and three estimators of the slope parameter.

From Figure 1, the true estimator of the slope when the *latent* variable is available, that is, $\hat{\beta}_{1X}$ is represented by the tan of $\angle BAC$ of $\triangle ABC$. In the absence of the values of the *latent* variable this is unavailable. But for the *manifest* variable one can find the estimator of the slope to be $\hat{\beta}_{1m}$ which is represented by the tan of $\angle DAE$ of $\triangle ADE$. Note that here DC (or equivalently BE) represents the sum of squares of measurement error (S_{uu}). Furthermore, under the assumptions of $\mathrm{E}[Yu] = 0$ and $\mathrm{E}[Xu] = 0$, we have $BC = DE$ or $S_{xy} = S_{my}$. Finally, $\hat{\beta}_{1X} = \frac{S_{xy}}{S_{xx}} = \frac{BC}{AC}$, and $\hat{\beta}_{1m} = \frac{S_{my}}{S_{mm}} = \frac{ED}{AD}$.

The introduction of the reflection of the manifest variable changes $\triangle ADE$ of Figure 1 to $\triangle ADF$ in Figure 2. In fact the main difference between the two Figures is that Figure 2 has the small $\triangle BEF$ added to Figure 1. This triangle represents the effect of the reflection of the manifest variable. From Figure 2 the estimates of the slope are

$$\hat{\beta}_{1m} = \frac{S_{my}}{S_{mm}} \left(= \frac{DE}{DA} \right) \tag{4.10}$$

$$\hat{\beta}_{1X} = \frac{S_{xy}}{S_{xx}} \left(= \frac{BC}{AC} \right) \tag{4.11}$$

$$\hat{\beta}_{1R} = \frac{S_{x^*y}}{S_{mm}} \left(= \frac{FD}{AD} \right). \tag{4.12}$$

Since the tan of $\angle BAC$ represents the estimator $\hat{\beta}_{1X}$ and tan of $\angle DAF$ represents $\hat{\beta}_{1R}$, we conclude that $\hat{\beta}_{1X} = \hat{\beta}_{1R}$. Note that $\angle BAC = \angle DAF$.

5 Some properties and relationships

The estimated regression lines based on the OLS, and IV methods are summarised in the following way:

$$\hat{Y}_X = \hat{\beta}_{0X} + \hat{\beta}_{1X}X \quad (5.1)$$
$$\hat{Y}_R = \hat{\beta}_{0R} + \hat{\beta}_{1R}M \quad (5.2)$$
$$\hat{Y}_W = \hat{\beta}_{0W} + \hat{\beta}_{1W}M \quad (5.3)$$
$$\hat{Y}_B = \hat{\beta}_{0B} + \hat{\beta}_{1B}M. \quad (5.4)$$

Obviously, in the absence of X, the fitted model in (5.1) is unavailable. The other fitted lines are obtainable since the manifest variable M is always observed along with the response Y. Furthermore, even though the regression parameters are the same, the estimated models are different since observed M is different from the true value of the explanatory variable X. Thus

$$\hat{\beta}_{0x} + \hat{\beta}_{1x}X \neq \hat{\beta}_{0x} + \hat{\beta}_{1x}M.$$

Another useful fact is that the sum of squares total is the same for regression of Y on X and that on M. That is,

$$SS_{yy} = SSR_x + SSE_x = SSR_m + SSE_m.$$

Similarly, the following relationship of the regression sum of squares for models using X, M, and X^* are observed:

$$SSR_x = \hat{\beta}_{1x}SS_{yx} = \hat{\beta}_{1R}SS_{my} = SSR_R \neq \hat{\beta}^2_{1R}SS_{mm} = \hat{\beta}_{1R}SS_{x^*y}.$$

Finally, the coefficient of determination is noted to be

$$R^2_x = \frac{SSR_x}{SST} = \frac{SSR_R}{SST}.$$

6 Examples for Illustration

In this section, two illustrative examples based on two real life data sets are provided. Bot cases reveal the superiority of the proposed new IV estimator. The first data set has mea surement error in the explanatory variable only, but the second data set has measuremer error in both the response. For the second example assume that the ratio of error varianc $\lambda = \frac{\sigma_{ee}}{\sigma_{uu}} < 1$, where σ_{ee} is the error variance of the response variable and σ_{uu} is the erro variance of the explanatory variable..

6.1 Yield of Corn Data

The data set of the first example deals with the yield of corn (Y) for different levels soil nitrogen (M), and is taken from Fuller (2006, p.18). Here the explanatory variabl

soil nitrogen level, has been determined with measurement error. Fuller has analysed the data with the existing method with usual assumptions including known *reliability ratio*. We provide the regression analyses of the data for both with (a) the measurement error in the explanatory variable M, and (b) the instrumental variables including one defined by X^*, the reflection of the observed explanatory variable M. Comparison of the regression estimates and related results from different methods are provided below. The Table 1 below shows the fitted regression lines, mean sum of squares error, and the coefficient of determination based on the OLS and various IV methods including the *reflection* method.

Table 1: Fitted regression models for the corn yield data

Method	Fitted regression equation	MS Errorr	R^2
Least Squares	$\hat{Y}_M = 73.153 + 0.344M$	57.321	0.412
Wald	$\hat{Y}_W = 75.91 + 0.305M$	60.98	0.364
Bartlett	$\hat{Y}_B = 72.38 + 0.355M$	56.05	0.425
Reflection	$\hat{Y}_R = 65.8164 + 0.4479M$	45.224	0.536
σ_{uu} Known	$\hat{Y}_V = 67.561 + 0.423M$	48.125	0.506

The straightforward regression of Y on M produces the estimated (OLS) regression line, $\hat{Y}_m = 73.153 + 0.344M$ with mean sum of squares error, $MSE_m = 57.321$ (see the first regression line in Table 1) and $R^2_m = 0.421$. This analysis does not take into account the presence, and hence the effect, of the measurement errors in the explanatory variable. As such these results are not based on any sound statistical method and hence unacceptable.

Fuller (2006, p.18-19) assumes that $\sigma_{uu} = 57$, and that the *reliability ratio*, κ_{xm} is known. Under the above assumptions the estimated regression line reported to be $\hat{Y}_V = 67.561 + 0.423M$ with modified MSE, $MSE_V = \hat{\sigma}_{ee} = 48.125$, and $R^2_V = 0.506$. Clearly, there has been an improvement in the proportion of variability in Y that is explained by M under the method used by Fuller (2006). The MSE has also decreased (from $MSE_m = 57.321$ to $MSE_V = 48.125$) under the Fuller method. Thus the Fuller method is not only a better method than the OLS, but also provides a much better fit.

The use of the reflection of M in the specification of the instrumental variable leads to the fitted regression line, $\hat{Y}_R = 65.8164 + 0.4479M$ with mean sum of squares error, $MSE_R = 45.224$ (see second last row of Table 1) and $R^2_R = 0.536$. Unlike Fuller's method, these results are obtained without additional assumptions on any of the parameters of the model or the reliability ratio. However, the regression parameters obtained by using the reflection of M are fairly close to those obtained by Fuller under the previously stated assumptions. The regression line produced by the proposed method provides a much better fit than that obtained by Fuller. Obviously, the MSE_R under the IV is much smaller than $\hat{\sigma}_{ee}$ obtained by Fuller's method. Moreover, under the proposed method the value of the coefficient of determination is 53.6%, compared to only 50.60% under the Fuller's method.

The estimates of the regression parameters of the *manifest* model are $\hat{\beta}_{1m} = 0.344$ and $\hat{\beta}_{0m} = 73.153$, and that of the proposed *instrumental variable* model are $\hat{\beta}_{1R} = 0.4479$ and $\hat{\beta}_{0R} = 65.8164$. These figures support the results in Theorem (4.1), that is, $\hat{\beta}_{1R} = 0.4479 >$

$\hat{\beta}_{1m} = 0.344$, and $\hat{\beta}_{0R} = 65.8164 < \hat{\beta}_{0m} = 73.153$. Note here the correlation is positive.

It is important to compare the results of the new IV estimator with other IV estimators such as the Wald and Bartlett methods specified earlier. The results of Wald method yields, $\hat{Y}_W = 75.91 + 0.305M$ with $MSE_W = 60.98$ and $R^2_W = 0.364$. Moreover, using Bartlett's definition of the IV, we get $\hat{Y}_B = 72.38 + 0.355M$ with $MSE_B = 56.05$ and $R^2_B = 0.425$. Practically both methods are inefficient, although the Bartlett method produces better fit (larger R^2_B) than that of Wald R^2_W.

Clearly, the Wald's method produces the worst of the five fitted models in terms of the MSE (or R^2). The Bartlett's method is better than the Wald's method in terms of the value of R^2. The Fuller's method provides a much better fit than the OLS, Wald and Bartlett methods. However, the reflection based IV fitted model has the largest R^2. At the same time the regression estimates of the slope and intercept for the Fuller method is much close to that of the reflection based estimator. Thus the IV based on the reflection of M provibes the best model without making any additional assumptions on the error variance or reliability ratio.

6.2 Hen Pheasants Data

The data set for the second example is also taken from Fuller (2006, p.34). The data deal with the number of hen pheasants in Iowa at two different season/time of the year,and were collected by the Iowa Conservation Commission. These data are based on the average number of birds sighted by trained observers traveling a number of specific routes in late April and early May , and again in August. Both measures are subject to error for two reasons. First, the routes are a sample of all possible routes in Iowa. Second, observers cannot be expected to sight all pheasants along the route. The response variable Y is the average number of hens in August, and the explanatory variable M is the average number of hens in Spring, where the ratio of error variances $\lambda < 1$. On the basis of previous analyses, it has been estimated that the error variance for the Spring count is about six times larger than that in August. The fitted regression models and associated statistics are provided in the Table 2 below.

Table 2: Fitted regression models for the hen peasants data

Method	Fitted regression equation	MS Error	R^2
Least Squares	$\hat{Y}_M = 2.142 + 0.649M$	0.347	0.826
Wald	$\hat{Y}_W = 2.498 + 0.614M$	0.44	0.78
Bartlett	$\hat{Y}_B = 2.036 + 0.66M$	0.32	0.84
Reflection	$\hat{Y}_R = 1.323 + 0.731M$	0.14	0.93
Moments	$\hat{Y}_{MO} = 1.116 + 0.751M$	0.09	0.95

The first regression equation and the associated statistics in Table 2, $\hat{Y}_m = 2.142 + 0.649M$, $MSE_m = 0.347$ and $R^2_m = 0.826$, are obtained by the OLS method using M which is subject to the measurement error.

The method of moments (MOM) estimator, under the assumption that the ratio $\delta = \sigma_{uu}^{-1}\sigma_{ee}$ is known can be found. Following Fuller (2006, p.35), for $\delta = \frac{1}{6}$, the fitted regression equation becomes $\hat{Y}_{MO} = 1.1158 + 0.7516M$ with $MSE_{MO} = 0.09$ and $R^2_{MO} = 0.95$. This is a much better fitted model, with an increased value of R^2, than that obtained by the OLS method.

The second last row of Table 2 represents the regression line and other statistics produced by the proposed new instrumental variable method based on the reflection of M: $\hat{Y}_R = 1.323 + 0.731M$, $MSE_R = 0.139$ and $R^2_R = 0.93$.

The IV estimator based on Wald's method yields $\hat{Y}_W = 2.498 + 0.614M$ with $MSE_w = 0.44$ and $R^2_w = 0.78$. Similarly, Bartlett's IV method gives $\hat{Y}_B = 2.036 + 0.66M$, $MSE_B = 0.32$ and $R^2_B = 0.84$.

In terms of the R^2 value the Wald's method is the worst, followed by the OLS method. Thus Wald's IV method may produce worst fit than the OLS method. The Bartlett's method gives a similar R^2 as the OLS method. However, the MOM estimation produces the largest R^2, although it is not too far from that produced by the proposed reflection based method. It is important to note that the MOM is based on the assumption that the value of δ is known, whereas no assumption is required for the reflection method. Furthermore, due to the nature of the definition of the IV, we have only 'treated' the measurement error in the explanatory variable. It seems that similar treatment of the response variable would produce results better than the method of moments.

Among the IV estimators the proposed reflection based IV performs much better than the others in terms of providing the best fitted model with largest R^2. This is not surprising due to the fact that IVs proposed by Wald, Bartlett, and Durbin fails to use part of the information of the sample data to define the IV. Although the MOM estimation method provides slightly better fit than the proposed reflection based IV method, the former is dependent on the unrealistic assumption that δ is know. When no information is available on δ, which is normally the case, the new method ensures the best fitted model.

7 Concluding Remarks

The paper considers the simple regression model with measurement error in the explanatory variable. It proposes a new estimation procedure based on the idea of a new instrumental variable which is defined from reflection of the *manifest* variable. Also, it provides the theory of the available literature and compares the existing methods with proposed new method. Unlike, some of the existing methods it does not loss of information. Moreover, the statistical properties of the proposed estimator are much superior to those of the available methods in the literature.

The analytical results and the illustrative examples demonstrate the fact that the proposed method significantly reduces the mean sum of squares error than the currently used methods. As such, the coefficient of determination of the proposed method is higher than

that of the existing methods.

The proposed method in the paper is new, easy to implement, and performs much better than the existing method. We have demonstrated the superiority of our method both analytically and via numerical illustration.

Surprisingly, the proposed IV method recovers the true estimator of the slope, $\hat{\beta}_{1X}$, from the *manifest* variable and *stochastic* model even if the true values of the *latent* variable is unobservable. The Theorem 4.1 and the Figure 2 demonstrate this remarkable fact. The same comment would apply for the estimator of the intercept.

References

1. Adcock, R J (1877). Note on the method of least squares. *Analyst.*, **4**, 183-184.
2. Adcock, R J (1878). A problem in least squares. *Analyst.*, **5**, 53-54.
3. Aigner, D J (1973). Regression with a binary variable subject to errors of observation. *J. Econometrics* **1**, 49-60.
4. Amemiya, Y. and Fuller, W A (1984). Estimation of the multivariate errors-in-variables model with estimated error covariacne matrix. *Ann. Statist.* **12**, 497-509.
5. Armstrong, B (1985). Measruement error in the generlzed linear model. *Commun. Statist. Part B* **14** 529-544
6. Bartlet, M S (1949). Fitting a stright line when both variables are subject to error. *Biometrics* **5** 207-212.
7. Bekker, P A (1986). Comments on ident9ification in the linear errors in variables model. *Econometrica* **54** 215-217.
8. Berkson, J (1950). Are there two regressions? *J. Am. Statist. Assoc.* **45**, 164-180.
9. Bhargava, A K (1977). Maximum likelihood estimation in a multivariate errors-in-variables regression model with unknown error covariance matrix. *Commun. Statist. Part A* **6** 587-601.
10. Burr, D (1988). On Errors-in-Variables in Binary Regression-Berkson Case, *Journal of the American Statistical Association*, **83(403)**, 739-743
11. Carroll, R J, Gallo, P, and Gleser, I J (1985). Comparison of least squares and errors-in-variable regression with special reference to randomized analysis of covariance. *J. Am. Statist. Assoc.* **80** 929-932.
12. Degracie, J S and Fuller, W A (1972). Estimation of the slope and covariance when the concomitant variable is measured with error. *J. Am. Statist. Assoc.* **67** 930-937.
13. Durbin, J (1954). Errors-in-variables. *Int. Statist. Rev.* **22**, 23-32.

14. Florens, J P, Moucharrt, M and Richard, F (1974). Bayesian inference in error-in-variables models. *J. Multivariate Anal.* **4** 419-452.

15. Freedman, L S; Fainberg, V; Kipnis, V; Midthune, D; and Caroll, R J (2004). A New Method for Dealing with Measurement Error in Explanatory Variable of Regression Models. *Biometrics*, **60**, 172-181.

16. Fuller, W A (2006). Measurement Error Models. Wiley, New Jersey.

17. Garber, S and Klepper, S (1980). Extending the classical normal errors-in-variables model. *Econometrics* **48** 1541-1546.

18. Geary, R C (1942). Inherent relations between random variables. *Proc. R. Irish Acad. Sect. A* **47**, 36-67.

19. Grubbs, F E (1973). Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. *Technometrics* **15** 53-66.

20. Halperin, M (1961). Fitting of straight lines and prediction when both variables are subject to error. *Jou. Amer. Statist. Assoc*, **56**, 657-669.

21. Klepper, S and Leamer, E E (1984). Consistent sets of estimates for regression with errors in all variables. *Econometrica* 55 163-184.

22. Madansky, A (1959). The fitting of straight lines when both variables are subject to error. *Jou. Amer. Statist. Assoc*, **54**, 173-205

23. Maddala, G.S (1988) Introduction to Econometrics. Prentice Hall International, Second edition.

24. Pearson, K (1901). On lines and planes of closest fit to systems of points in space. Philos. Mag. **2**, 559-572.

25. Prentice, R L (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69** 331-342.

26. Riggs, D S, Guarnieri, J A and Addelman, S (1978). Fitting straight line when both variables are subject to error. *Life Sci.* **22**, 1305-1360.

27. Schafer, D W (1986). Combining information on measurement error in errors-in-variables model. *J. Am. Statist. Assoc.* **81** 181-185.

28. Schneeweiss H (1976). Consistent estimation of a regression with errors in the variables. *Metrika* 23 101-115.

29. Stefanski, L A and Carroll, R J (1985). Covariate measurement error in logistic regression. *Ann. Statist.* **12** 1335-1351.

30. Vaisman, I (1997). Analytical Geormetry. World Scientific, Singapore.

31. Wald, A (1940). Fitting of straight lines if both variables are subject to error. *Ann. Math. Statist.* 11 284-300.