New Weighted Geometric Mean Method to Estimate the Slope of Measurement Error Model

Anwar Saqr[1*] and Shahjahan Khan[2]

[1] Al-Buraimi University College, Sultanate of OMAN

Email: *saqr.anwer@yahoo.com*

[2]School of Agricultural, Computational and Environmental Sciences

International Centre for Applied Climate Sciences

University of Southern Queensland, Toowoomba, AUSTRALIA

Email: *Shahjahan.Khan@usq.edu.au*

Abstract

This paper introduces a new weighted geometric mean (WG) estimator to fit regression line when both the response and explanatory variables are subject to measurement errors. The proposed estimator is based on the mathematical relationship between the vertical and orthogonal distances of the observed points and the regression line (cf. Saqr and Khan, 2012). It minimizes the orthogonal distance of the observed points from the unfitted line. The WG estimator is less sensitive to the ratio of error variances (λ). It is a better alternative than the currently used geometric mean (GM) and OLS-bisector estimators. Extensive simulation results show that the proposed WG estimator is much more stable than the geometric mean and OLS-bisector estimators. The mean absolute error of the WG estimator is consistently smaller than the geometric mean and OLS-bisector estimators.

Key Words: Linear regression models, Measurement error models, Reflection of points; Ratio of error variances; Geometric mean estimator, OLS-bisector.

2010 Mathematical Subject Classification: Primary 62J05, Secondary 62F10.

1 Introduction

The geometric mean estimator is applied in many disciplines to estimate regression parameters when both variables are subject to errors. This technique has been introduced many times under different

*On leave from Department of Statistics, Faculty of Sciences, University of Al-Jabal Al-Gharbi, Gharyan, LIBYA.

name such as reduced major axis, or least products regression (cf Ludbrook, 2010). In spite of its popular use there are criticisms about its over sensitively on the ratio of error variances λ.

Dent (1935) suggested the geometric mean functional relationship estimator to be a solution of the likelihood equations when there is no additional information in the case of the normal functional model (cf Cheng and Ness, 1999, p. 43). This estimator is called geometric mean (GM) estimator, because it is the geometric mean of the least squares coefficients for the regression of the observed (manifest) response (y) variable on the observed explanatory variable (m) and the reciprocal of that for m on y.

Halfon (1985) and Draper and Yang (1997) pointed out that the geometric mean estimator minimizes the vertical and horizontal distances between the observed points and the regression line. Jolicoeur (1975) stated that it is difficult to interpret the meaning of the slope of the geometric mean regression. Isobe et al. (1990) examined five different methods, and pointed out that the OLS bisector (OLS-b) estimator is the best method to use, when there is no basis to distinguish between the explanatory and response variables.

The problem of measurement error or error-in-variable has a long history in statistics (see Adcock, 1877; Fuller, 2006; and Cheng and Ness, 1999, and the references therein) and has received growing attention from many statisticians and econometricians (see Johnston, 1972; and Maddala, 2001). Measurement error (ME), as its name implies, is the result of recording values that are randomly different from the actual values. The basic theory of regression analysis assumes that the explanatory variable is measured without error. Unfortunately, real data are seldom observed directly, especially in economics, finance, agriculture, medical and physical sciences, and social sciences without any errors. It is well known that the presence of measurement error in the explanatory variable makes the ordinary least squares (OLS) method inappropriate in large as well as small samples. Measurement error can seriously distort inference when they are not taken into account explicitly. Simple OLS estimates indicate substantial decreasing returns to scale, but are subject to the usual attenuation bias. In general, presence of measurement error produces biased and inconsistent OLS estimators (cf Cheng and Ness, 1999, p. 3).

There are many researchers such as Wald (1940), Bartlett (1949), Durbin (1954), Riggs et al., (1978) and Saqr and Khan (2011, 2012) considered fitting regression line when both variables are subject to error. 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). Geary (1942, 1943, 1948 and 1949) wrote a series of papers on the method of moments. The method of moments has been done by Pal (1980), van Montfort et al., (1987), van Montfort (1989) and Cragg (1997). Their work centres on how to find the optimal estimators based on higher moments. Reiersol (1950) pointed out that the parameters of the measurement error model are identifiable if the cumulant $k(r, s)$ of the joint

distribution function of the manifest response variable y and the manifest explanatory variable m is exist a nonzero, and finite or infinite with $r = 1$ and $s > 1$ or opposite.

The purpose of this paper is provide a new estimator to fit regression line when both variables are subject to measurement errors. The proposed weighted geometric mean estimator is a better alternative than the geometric mean estimator and OLS-bisector estimator. The WG estimator is based on the mathematical relationship between the vertical and orthogonal distances of the observed points and the regression line. It minimizes the orthogonal distance from unfitted line, and is less sensitive to the ratio of error variances (λ). The simulation results show that the proposed estimator is much closer to the true slope than the geometric mean and OLS-bisector estimators. Here we use the mean absolute error (MAE) instead of the root mean squared error (RMSE) because MAE is regularly employed in model evaluation studies. Willmott and Matsuura (2005) pointed out that the RMSE is not a good indicator of average model performance and might be a misleading indicator of average error, whereas the MAE is a better indicator to describe average model-performance error and inter-comparisons of average model performance error should be based on MAE.

In the next section the measurement error regression model is introduced. Section 3 presents the mathematical relationship between the vertical and orthogonal distances of the observed points from both fitted and unfitted regression line. The geometric mean estimator and deriving this estimator are provided in Sections 4. The proposed weighted geometric mean estimator is introduced in Section 5. The simulation studies, and the concluding remarks are included in Sections 6 and 7.

2 Measurement error models

In the conventional notation, let x be the true measurement on the explanatory variable which is otherwise known as the *latent* variable. In the presence of measurement error the observed value of the latent variable is different from x. Let m be the observable or *manifest* variable of the explanatory variable. Similarly let η be the true value of the response variable and y be the manifest response variable.

If the *latent* variables x_j and η_j are measured without error then their linear relationship without the equation error is expressed as

$$\eta_j = \beta_0 + \beta_1 x_j, \quad j = 1, 2, \ldots, n. \tag{2.1}$$

If there is error in both response and explanatory variables, the actual observed values of m and y are not the true values, and we define

$$m_j = x_j + u_j, \text{ and } y_j = \eta_j + e_j \qquad j = 1, 2, \ldots, n, \tag{2.2}$$

where η_j is the jth realisation of the *latent* response variable, x_j is the jth value of the *latent* explanatory variable, e_j is the measurement error in the response variable and u_j is the measurement

error in the explanatory variable. It is assumed that,

$$e_j \sim N(0, \sigma_e^2),\ u_j \sim N(0, \sigma_u^2),\ \text{and } cov(u, e) = 0. \tag{2.3}$$

Note that 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*, whereas, the model with independent and identically distributed random variable x is called *structural model*. The later is considered in this paper.

The simple regression model with measurement error in both variables and without equation error is known as the standard measurement error model, which can be expressed as

$$y_j = \beta_0 + \beta_1 m_j + v_j, \quad j = 1, 2, \ldots, n, \tag{2.4}$$

where $v_j = e_j - \beta_1 u_j$, and

$$\sigma_v^2 = \sigma_e^2 + \beta_1^2 \sigma_u^2. \tag{2.5}$$

Note in equation (2.4) m_j and v_j are not independent, and hence least squares method is not valid for the above model. The ordinary least squares (OLS) estimator of the regression parameters is inappropriate (biased and inconsistent) in the presence of measurement error (see Johnston, 1972, p. 284).

3 Relationship between the vertical and orthogonal distances

It is well known that there are different approaches to minimize the vertical, horizontal, orthogonal, or both orthogonal and horizontal distances in regression analysis. The ordinary least squares method works on the basis of minimizing the vertical distance when there are no measurement errors. Inverse least squares method minimizes the horizontal distance when there is measurement error only in the explanatory variable. The orthogonal regression approach suggests to minimize the orthogonal distance under the assumption that the ratio of error variances is equal to one, that is, $\lambda = \sigma_e^2 \sigma_u^{-2} = 1$. The maximum likelihood estimator minimizes both the horizontal and orthogonal distances when λ is known.

It is crucial to note the difference between the distance of the observed point from the fitted line, unfitted line, and unobserved point. Although, many authors use distance between the observed point and regression line without being specific. This issue is crucial when there are measurement errors in both variables. This section introduces the mathematical relationship between the vertical and orthogonal distances of the observed points and the fitted regression line.

Let (m_j, y_j) be the observed point and (x_j, η_j) be the corresponding unobserved point. Then the fitted line is given by

$$\eta_j = \beta_0 + \beta_1 x_j, \quad j = 1, 2, \ldots, n. \tag{3.1}$$

Note that all the true points (x_j, η_j) are on the fitted line (3.1), because there is no equation error in the model.

Now we define the reflection point (m_j^*, y_j^*) of the observed point (m_j, y_j) about the fitted line (3.1) as follows:

$$m_j^* = m_j \cos 2\psi + (y_j - \beta_0) \sin 2\psi, \tag{3.2}$$

$$y_j^* = m_j \sin 2\psi - (y_j - \beta_0) \cos 2\psi + \beta_0, \tag{3.3}$$

where $\psi = tan^{-1}\beta_1$, and β_0, and β_1 are the regression parameters. For details on reflection of points please see Vaisman (1997, p. 164-169). Details on the reflection method is found in Saqr and Khan (2012a, 2016). However, the formulas (3.2) and (3.3) can be rewritten for sample statistics as follows:

$$m_j^* = m_j \cos 2\hat{\psi} + (y_j - \hat{\beta}_0) \sin 2\hat{\psi}, \tag{3.4}$$

$$y_j^* = m_j \sin 2\hat{\psi} - (y_j - \hat{\beta}_0) \cos 2\psi + \hat{\beta}_0, \tag{3.5}$$

where $\hat{\psi} = tan^{-1}\hat{\beta}_1$, and $\hat{\beta}_0$, and $\hat{\beta}_1$ are the coefficients of estimated regression model without measurement error i.e $\hat{\eta}_j = \hat{\beta}_0 + \hat{\beta}_1 x_j, \quad j = 1, 2, \ldots, n.$

Theorem 3.1 *The* reflection *variable m_j^* is an unbiased measure of both manifest m_j and* latent *x_j explanatory variables, that is, under expectation*

$$\bar{m}^* = \bar{m} = \bar{x}.$$

Proof: From (3.4) taking sum over j, we get

$$\begin{aligned}
\sum_{j=1}^{n} m_j^* &= \sum_{j=1}^{n} m_j \cos 2\hat{\psi} + \sum_{j=1}^{n} y_j \sin 2\hat{\psi} - n\hat{\beta}_0 \sin 2\hat{\psi} \\
&= \sum_{j=1}^{n} m_j \cos 2\hat{\psi} + \sum_{j=1}^{n} y_j \sin 2\hat{\psi} - \sum_{j=1}^{n} y_j \sin 2\hat{\psi} + \hat{\beta}_1 \sum_{j=1}^{n} m_j \sin 2\hat{\psi} \\
&= \sum_{j=1}^{n} m_j \cos 2\hat{\psi} + \hat{\beta}_1 \sum_{j=1}^{n} m_j \sin 2\hat{\psi} = \sum_{j=1}^{n} m_j (\cos 2\hat{\psi} + \hat{\beta}_1 \sin 2\hat{\psi}) \\
&= \sum_{j=1}^{n} m_j (\cos^2 \hat{\psi} - \sin^2 \hat{\psi} + \frac{\sin \hat{\psi}}{\cos \hat{\psi}} (\sin \hat{\psi} \cos \hat{\psi})) = \sum_{j=1}^{n} m_j (\cos^2 \hat{\psi} + \sin^2 \hat{\psi}) \\
&= \sum_{j=1}^{n} m_j.
\end{aligned}$$

Multiplying both sides by $\frac{1}{n}$, we get

$$\sum_{j=1}^{n} \frac{m_j^*}{n} = \sum_{j=1}^{n} \frac{m_j}{n},$$

where $m_j = x_j + u_j$, and $u_j \sim N(0, \sigma_u^2)$, hence

$$\bar{m}^* = \bar{m} = \bar{x}. \tag{3.6}$$

Finally $E(\bar{m}^*) = E(\bar{m}) = E(\bar{x})$. Similarly, it can be shown that $E(\bar{y}^*) = E(\bar{y}) = E(\bar{\eta})$.

For simplicity, we consider that the relationships between the orthogonal and vertical distance of the observed point (m_j, y_j) and the fitted line $(\eta_j = \beta_0 + \beta_1 x_j)$ as a first case. While the second case is related to the relationship between the orthogonal and vertical distance of the observed point (m_j, y_j) and unfitted line $(\hat{y}_j = \hat{\beta}_{0m} + \hat{\beta}_{1m} m_j)$.

3.1 Fitted line case (True model)

It is well known from the properties of the reflection process that the reflection line (i.e. the fitted line) is a bisector and orthogonal on the distance between the observed point (m_j, y_j) and its reflection point (m_j^*, y_j^*). Then the half of the square distance between the observed point (m_j, y_j) and its reflection point (m_j^*, y_j^*) will equal the orthogonal distance square (Od_j^2) between the observed point (m_j, y_j) and the fitted line. It is given by

$$Od_j = \frac{1}{2}\sqrt{(m*_j - m_j)^2 + (y_j^* - y_j)^2}. \tag{3.7}$$

Then from (3.2) and (3.3) the orthogonal distance square (Od_j^2) is given by

$$Od_j^2 = \frac{1}{4}\left[(2m_j sin^2\psi + y_j sin2\psi - \beta_0 sin2\psi)^2 + (m_j sin2\psi - 2y_j cos^2\psi + 2\beta_0 cos^2\psi)^2\ \right],$$

from (2.1) and (2.2) we get $m_j = x_j + u_j$, $y_j = \eta_j + e_j = \beta_0 + \beta_1 x_j + e_j$, and $\beta_1 = \dfrac{sin\psi}{cos\psi}$ so

$$\begin{aligned}
Od_j^2 &= \frac{1}{4}\left[(2(x_j + u_j)sin^2\psi + (\beta_0 + \beta_1 x_j + e_j)sin2\psi - \beta_0 sin2\psi)^2\right] \\
&+ \frac{1}{4}\left[((x_j + u_j)sin2\psi - 2(\beta_0 + \beta_1 x_j + e_j)cos^2\psi + 2\beta_0 cos^2\psi)^2\ \right] \\
&= \frac{1}{4}\left[(-2m_j sin^2\psi + \beta_1 x_j sin2\psi + e_j sin2\psi)^2 + (m_j sin2\psi - 2\beta_1 x_j cos^2\psi - 2e_j cos^2\psi)^2\right] \\
&= \frac{1}{4}\left[(2u_j sin^2\psi - e_j sin2\psi)^2 + (u_j sin2\psi - 2e_j cos^2\psi)^2\right] \\
&= \frac{1}{4}\left[u_j^2(4sin^4\psi + sin^2 2\psi) - 4u_j e_j(sin2\psi sin^2\psi + sin2\psi cos^2\psi) + e_j^2(4cos^4\psi + sin^2 2\psi)\right] \\
&= u_j^2 sin^2\psi - u_j e_j sin2\psi + e_j^2 cos^2\psi.
\end{aligned}$$

From (2.3) $E(u_j) = E(e_j) = 0$ and $E(u_j e_j) = 0$ then

$$E(Od_j^2) = E(u_j^2)sin^2\psi + E(e_j^2)cos^2\psi.$$

From (4.1) that $E(m_j^* - m_j) = E(y_j^* - y) = 0$ then the variance of Od_j is

$$\sigma_{Od}^2 = \sigma_u^2 sin^2\psi + \sigma_e^2 cos^2\psi.$$

From (2.5), and $\beta_1^2 = \sin^2\psi\ cos^{-2}\psi$, the above variance becomes

$$\sigma_{Od}^2 = (\sigma_e^2 + \sigma_u^2\frac{sin^2\psi}{cos^2\psi})cos^2\psi = (\sigma_e^2 + \beta_1^2\sigma_u^2)cos^2\psi.$$

Then the relationship between the variance of the orthogonal distance and the variance of vertical distance is given by

$$\sigma_{Od}^2 = \sigma_v^2 cos^2\psi = \frac{\sigma_v^2}{1 + \beta_1^2}, \tag{3.8}$$

where $\sigma_v^2 = \sigma_e^2 + \beta_1^2\sigma_u^2$, and $cos^2\psi = \frac{1}{\frac{1}{cos^2\psi}} = \frac{1}{\frac{cos^2\psi + sin^2\psi}{cos^2\psi}} = \frac{1}{1 + \frac{sin^2\psi}{cos^2\psi}} = \frac{1}{1+\beta_1^2}$.

Note that both vertical and orthogonal distances are measured as the distance between the observed point (m_j, y_j) and the fitted line, but it does not measure the distance between the observed point (m_j, y_j) and the unobserved point (x_j, η_j). Under certain assumptions such as $\lambda = 1$ or $\beta_1 = 1$ the distance between the observed point and the unobserved point is equal to the double of the orthogonal distance, where the distance between the observed point and the unobserved point is given by

$$Pd = \sqrt{(m_j - x_j)^2 + (y_j - \eta_j)^2} = \sqrt{u_j^2 + e_j^2},$$

where u_j, and e_j are the measurement errors in the explanatory and response variables respectively. From (2.3) the variance of the (Pd) distance is

$$\sigma_{Pd}^2 = \sigma_e^2 + \sigma_u^2.$$

From (2.5) and when $\lambda = 1$,

$$\sigma_{Pd}^2 = 2\sigma_e^2.$$

3.2 Unfitted line case (ME model)

In order to find the relationship between the orthogonal (Om) and vertical (v) distances of the observed point (m_j, y_j) and the unfitted line $(\hat{y}_j = \hat{\beta}_{0m} + \hat{\beta}_{1m} m_j)$ for the model (2.4), let (m_j^{**}, y_j^{**}) be the reflection point of (m_j, y_j) about the unfitted line as following:

$$m_j^{**} = m_j cos2\hat{\theta} + (y_j - \hat{\beta}_{0m}) sin2\hat{\theta}, \tag{3.9}$$

$$y_j^{**} = m_j sin2\hat{\theta} - (y_j - \hat{\beta}_{0m}) cos2\hat{\theta} + \hat{\beta}_{0m}, \tag{3.10}$$

where $\hat{\theta} = tan^{-1}\hat{\beta}_{1m}$, $\hat{\beta}_{0m}$, and $\hat{\beta}_{1m}$. The relationship between the sample variance of the orthogonal distance (Om) and vertical distance (v), as similar to the first case it is given by

$$Om_j = \frac{1}{2}\sqrt{(m_j^{**} - m_j)^2 + (y_j^{**} - y_j)^2}. \tag{3.11}$$

Then from (3.9), (3.10) and (3.11) the orthogonal distance square (Om_j^2) is given by

$$\begin{aligned}
Om_j^2 &= \frac{1}{4}\left[(m_j cos2\hat{\theta} + (y_j - \hat{\beta}_{0m}) sin2\hat{\theta} - m_j)^2 + (m_j sin2\hat{\theta} - (y_j - \hat{\beta}_{0m}) cos2\hat{\theta} + \hat{\beta}_{0m} - y_j)^2 \right] \\
&= \frac{1}{4}\left[(-2m_j sin^2\hat{\theta} + y_j sin2\hat{\theta} - \hat{\beta}_{0m} sin2\hat{\theta})^2 + (m_j sin2\hat{\theta} - 2y_j cos^2\hat{\theta} - 2\hat{\beta}_{0m} cos^2\hat{\theta})^2, \right]
\end{aligned}$$

where $\hat{\beta}_{0m} = \bar{y} - \hat{\beta}_{1m}\bar{m}$, then

$$\begin{aligned}
Om_j^2 &= \frac{1}{4}\left[\Big((y_j - \bar{y}) sin2\hat{\theta} - 2(m_j - \bar{m}) sin^2\hat{\theta}\Big)^2 + \Big((m_j - \bar{m}) sin2\hat{\theta} - 2(y_j - \bar{y}) cos^2\hat{\theta}\Big)^2\right] \\
&= \frac{1}{4}\left[(y_j - \bar{y})^2 sin^2 2\hat{\theta} - 4(y_j - \bar{y})(m_j - \bar{m}) sin2\hat{\theta} sin^2\hat{\theta} + 4(m_j - \bar{m})^2 sin^4\hat{\theta}\right] \\
&+ \frac{1}{4}\left[(m_j - \bar{m})^2 sin^2 2\hat{\theta} - 4(y_j - \bar{y})(m_j - \bar{m}) sin2\hat{\theta} cos^2\hat{\theta} + 4(y_j - \bar{y})^2 cos^4\hat{\theta}\right] \\
&= \frac{1}{4}\left[4(m_j - \bar{m})^2 sin^2\hat{\theta} - 4(y_j - \bar{y})(m_j - \bar{m}) sin2\hat{\theta} + 4(y_j - \bar{y})^2 cos^2\hat{\theta}\right].
\end{aligned}$$

By taking sum over j, we get

$$\sum_{j=1}^{n} Om_j^2 = \sum_{j=1}^{n}(m_j-\bar{m})^2 sin^2\hat{\theta} - \sum_{j=1}^{n}(y_j-\bar{y})(m_j-\bar{m}) sin2\hat{\theta} + \sum_{j=1}^{n}(y_j-\bar{y})^2 cos^2\hat{\theta},$$

Then from Theorem 3.1 and (3.11) the mean of Om equals zero, and hence

$$\begin{aligned} \frac{S_{Om}^2}{cos^2\hat{\theta}} &= \hat{\beta}_{1m}^2 S_m^2 - 2\hat{\beta}_{1m} S_{xy} + S_y^2 = S_y^2 - \hat{\beta}_{1m} S_{xy} = S_v^2, \\ S_{Om}^2 &= S_v^2 cos^2\hat{\theta} = \frac{S_v^2}{1+\hat{\beta}_{1m}^2}, \end{aligned} \tag{3.12}$$

where S_v^2 is estimator of $\sigma_v^2 = \sigma_e^2 + \beta_1^2\sigma_u^2$. So in general (3.12) could be rewritten as

$$\sigma_{Om}^2 = \sigma_v^2 cos^2\theta = \frac{\sigma_v^2}{1+\beta_{1m}^2}. \tag{3.13}$$

From (3.8) and (3.13) the relationship between the orthogonal distances for the two cases becomes

$$\sigma_{Od}^2 = \sigma_{Om}^2 \frac{cos^2\psi}{cos^2\theta} = \sigma_{Om}^2 \left(\frac{1+\beta_{1m}^2}{1+\beta_1^2}\right). \tag{3.14}$$

Note that in general, $\sigma_{Od}^2 < \sigma_{Om}^2$, and they are equal if and only if there is no measurement error. Therefore, any method to minimize σ_{Om}^2, will not work well, and that is what is happening with the geometric mean method. The next section will show that the GM method is minimizing σ_{Om}^2, rather than σ_{Od}^2.

4 The geometric mean estimator

One of the simple approaches to handle the measurement error in the regression analysis is the geometric mean (GM) functional relationship, initially proposed by Teissier (1948) and later by Barker et al. (1988) (cf Draper and Yang, 1997). This estimator has frequently been mentioned in the literature for two reasons. First, when there is no basis for distinguishing between the response and explanatory variables. Second, to handle the measurement error when no prior information is available. The geometric mean method has received much attention from the experts, and some have suggested that it is more useful than the ordinary least squares method (see Sprent and Dolby, 1980).

The geometric mean estimator of the slope is the geometric mean of the slope of y on m regression line, and the reciprocal of the slope of m on y regression line, where m and y both are random (see Leng et al. 2007). It is given by

$$\hat{\beta}_{1G} = sgn(SP_{my})\sqrt{\frac{SS_y}{SS_m}} = sgn(Sp_{my})\left(\frac{S_y}{S_m}\right),$$

where $SS_m = \sum_{j=1}^{n}(m_j-\bar{m})^2$, $SS_y = \sum_{j=1}^{n}(y_j-\bar{y})^2$, $SP_{my} = \sum_{j=1}^{n}(m_j-\bar{m})(y_j-\bar{y})$, and S_y and S_m are the standard deviation of y and m respectively.

In the literature, the geometric mean regression is also known as the standardized major axis (MA) (cf. Warton et al., (2006)). It is also known as reduced major axis (RMA), or the line of organic correlation (cf Tessier, 1948, Kermack and Haldane, 1950, Ricker, 1973). In physics it is known as a type of standard weighting model (Machonald and Thompson, 1992), while the astronomers call it Strömberg's impartial line (Feigelson and Babu, 1992).

A host of recent publications indicate that using the GM or RMA is necessary and sufficient to fit the straight line when both the response and explanatory variables are subject to errors (see Levinton and Allen, 2005, Zimmerman et al. 2005, Sladek et al. 2006, and Vincent and Lailvaux, 2006). While Jolicoeur (1975) and Spernt and Dolby (1980) pointed out that the GM estimator is unbiased if and only if

$$\lambda = \frac{\sigma_y^2}{\sigma_m^2} \quad \text{or} \quad \lambda = \beta_1^2 \ .$$

But several other studies indicate that this assumption is unrealistic (cf Sprent and Dolby, 1980).

It is commonly recommended to use the geometric mean estimator without mentioning the justifications (Smith, 2009). Jolicoeur (1975) stated that it is difficult to interpret the meaning of the slope of the geometric mean regression. However, the common believe is the geometric mean regression minimizes the vertical and horizontal distances between the observed points and the fitted line [Halfon (1985) and Draper and Yang (1997)].

4.1 Derivation of the geometric mean estimator

This section demonstrates that the current geometric mean (GM) estimator is define based on the principle of minimizing the orthogonal (Om) distance of the observed point (m_j, y_j) and the unfitted line ($\hat{y}_j = \hat{\beta}_{0m} + \hat{\beta}_{1m} m_j$), while it was intended that the GM estimator was derived based on the minimization of the orthogonal distance (Od) of the observed point (m_j, y_j) and the fitted line ($\eta_j = \beta_0 + \beta_1 x_j$). From (3.12) the geometric mean estimator can be derived as

$$\begin{aligned} Min\, SS_{Om} &= SS_v cos^2\hat{\theta} = \sum_{j=1}^{n} (y_j - \hat{\beta}_{0m} - \hat{\beta}_{1m} m_j)^2 cos^2\hat{\theta}, \\ &\text{where } SS_{Om} = (n-1)S_{Om}^2, \text{ and } SS_v = (n-1)S_v^2, \\ &= \sum_{j=1}^{n} ((y_j - \bar{y}) - \hat{\beta}_{1m}(m_j - \bar{m}))^2 cos^2\hat{\theta} \\ &= \sum_{j=1}^{n} ((y_j - \bar{y}) cos\hat{\theta} - (m_j - \bar{m}) sin\hat{\theta})^2. \end{aligned} \tag{4.1}$$

Let $L_1 = sin\hat{\theta}$, and $L_2 = cos\hat{\theta}$. Then

$$SS_{Om} = \sum_{j=1}^{n} ((y_j - \bar{y})L_2 - (m_j - \bar{m})L_1)^2.$$

Figure 1: Graph of distances between the observed point, fitted regression line, unobserved point, and unfitted regression line

Differentiation of SS_{Om} w.r.t. L_1, and L_2 and setting them equal to zero, we get

$$\begin{aligned}
\frac{\partial SS_{Om}}{\partial L_1} &= 2\sum_{j=1}^{n}((y_j-\bar{y})L_2-(m_j-\bar{m})L_1)(-(m_j-\bar{m}))=0, \\
L_1S_m^2 &= L_2S_{ym}, \text{ and} \qquad (4.2) \\
\frac{\partial SS_{Om}}{\partial L_2} &= 2\sum_{j=1}^{n}((y_j-\bar{y})L_2-(m_j-\bar{m})L_1)(y_j-\bar{y})=0, \\
L_2S_y^2 &= L_1S_{ym}. \qquad (4.3)
\end{aligned}$$

From (4.1), (4.2), and $\hat{\beta}_{1m}=\dfrac{L_1}{L_2}$ we get two estimators of the slope

$$\hat{\beta}_1=\frac{S_{ym}}{S_m^2} \text{ and } \hat{\beta}_2=\frac{S_y^2}{S_{ym}} \qquad (4.4)$$

Then the geometric mean of the estimators in (4.4) is the GM estimator, that is,

$$\hat{\beta}_{1G}=sgn\{S_{ym}\}\sqrt{\frac{S_y^2}{S_m^2}}.$$

Obviously, the above GM estimator is derived by minimizing the orthogonal distance between the observed point (m_j, y_j) and unfitted line. Therefore, it does not minimize the distance between the observed point (m_j, y_j) and the fitted regression line.

5 Proposed weighted geometric mean estimator

The proposed weighted geometric mean (WG) estimator minimizes the orthogonal distance between the observed point (m_j, y_j) and the unfitted regression line. This estimator is based on the relationship (3.13) between the vertical and orthogonal distances of the observed points and the unfitted regression line. The WG estimator is derived from equations (4.3) and (4.4).
Multiply equation (4.2) by S_y^2, and equation (4.3) by S_{ym}, we get

$$L_1 S_m^2 S_y^2 = L_2 S_{ym} S_y^2 \tag{5.1}$$
$$L_1 S_{ym}^2 = L_2 S_{ym} S_y^2, \tag{5.2}$$

from equation (5.1) plus equation (5.2) we get

$$\begin{aligned} L_1(S_m^2 S_y^2 + S_{ym}^2) &= L_2 2 S_{ym} S_y^2 \\ (S_m^2 S_y^2 + S_{ym}^2) sin\hat{\theta} &= 2 S_{ym} S_y^2 cos\hat{\theta}. \end{aligned}$$

Hence the proposed estimator which considered it as a weighted geometric mean (WG) estimator is given by

$$\hat{\beta}_{1WG} = \frac{sin\hat{\theta}}{cos\hat{\theta}} = \frac{2 S_{ym} S_y^2}{S_y^2 S_m^2 + S_{ym}^2}. \tag{5.3}$$

This estimator could be simplified as follow

$$\begin{aligned} \hat{\beta}_{1WG} &= \frac{2 S_y^2 S_m^{-2}}{S_y^2 S_{ym}^{-1} + S_{ym} S_m^{-2}} \\ &= \frac{2\hat{\beta}_{1G}^2}{(\hat{\beta}_1 + \hat{\beta}_2)} = \mathcal{W}\ \hat{\beta}_{1G}, \end{aligned} \tag{5.4}$$

where $\mathcal{W} = \dfrac{\hat{\beta}_{1G}}{\hat{\beta}_{OLS-mean}}$, $\hat{\beta}_{OLS-mean}$ is obtained by taking the arithmetic mean of the slopes of the two ordinary least squares regression lines of OLS(y/m) and OLS(m/y). Note if the geometric mean estimator (GME) is equal to OLS-mean estimator, then the proposed weighted geometric mean estimator (WGE) is equal to both the geometric mean and OLS-mean estimators, because $\mathcal{W}$ is equal to one.

The reasons for suggesting weighted geometric mean estimator (WGE) instead of the geometric mean estimator, and OLS-bisector estimator will be apparent from the results of the next section.

6 Simulation studies

In this section we compare the proposed WG estimator with the GM and OLS-bisector estimators for a wide range of values of λ $(0.08 \leq \lambda \leq 100)$.

Figure 2: Graph of the mean slope of estimators, and the mean absolute error when the parameters $\beta_0 = 20, \beta_1 = 0.55$ and $0.08 \leq \lambda \leq 100$.

We perform large scale simulations to illustrate that the proposed estimator is asymptotically unbiased and consistent compared to the geometric mean estimator, and OLS-bisector estimator. The latter estimator is given by

$$\hat{\beta}_{1OLS-B} = (\hat{\beta}_1 + \hat{\beta}_2)^{-1} \left[\hat{\beta}_1 \hat{\beta}_2 - 1 + \sqrt{(1 + \hat{\beta}_1^2)(1 + \hat{\beta}_2^2)} \right],$$

where $\hat{\beta}_1 = \dfrac{S_{ym}}{S_m^2}$, and $\hat{\beta}_2 = \dfrac{S_y^2}{S_{my}}$.

For the simulation, the data set is generated based on 1000 replications of samples size 100 of normal structural model as follows:

1. Generate 100 independent values $x_1, \ldots, x_{100}$ of $x \sim N(0,\, 8)$.
2. Generate 100 independent values $u_1, \ldots, u_{100}$ of $u \sim N(0,\, 7)$.
3. Generate 100 independent values $e_1, \ldots, e_{100}$ of $e \sim N(0,\, \sigma_e)$, where $2 \leq \sigma_e \leq 71$, for each 1000 replications it is increased by 1.
4. Specify the values of β_0 and β_1 for the regression line.
5. Calculate the values of the three estimators and their mean absolute errors.

From Figures 1a-3a and under $0.08 \leq \lambda \leq 100$, the values of the OLS-bisector estimator are away from the true values of β_1, but it is much closer than those of the geometric mean estimator. The values of the geometric mean estimator (GME) are far above the true value of β_1. The GME appears to be an over estimate of the slope, and it is more so far larger values of β_1. Clearly the proposed

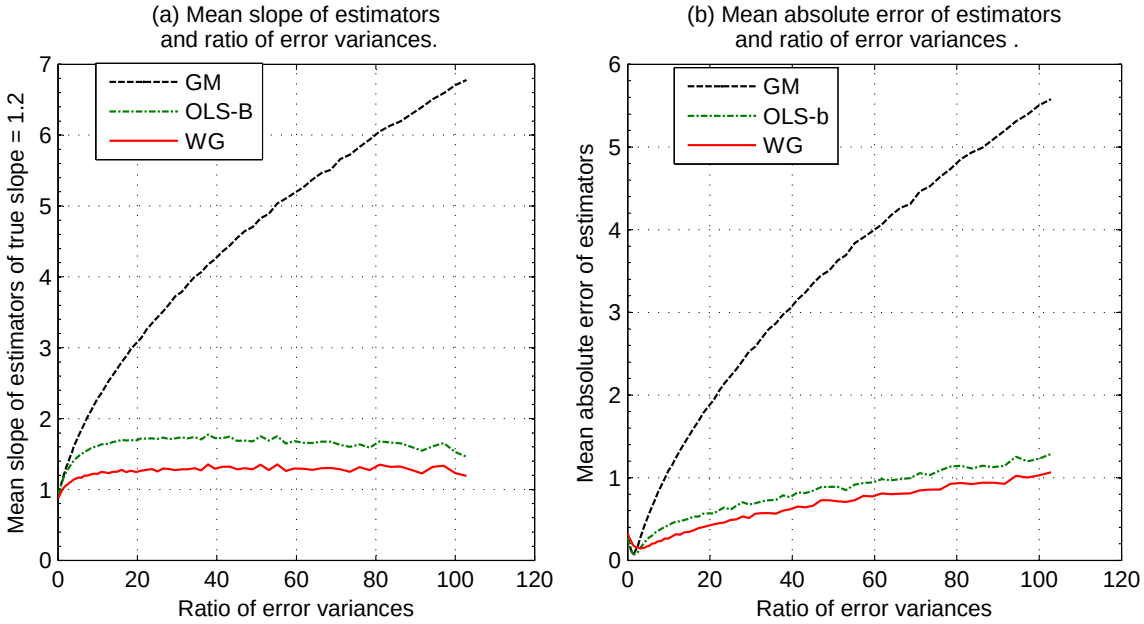

Figure 3: Graph of the mean slope of estimators, and the mean absolute error when the parameters $\beta_0 = 27, \beta_1 = -0.75$ and $0.08 \leq \lambda \leq 100$.

Figure 4: Graph of the mean slope of estimators, and the mean absolute error when the parameters $\beta_0 = -15, \beta_1 = 1.2$ and $0.08 \leq \lambda \leq 100$.

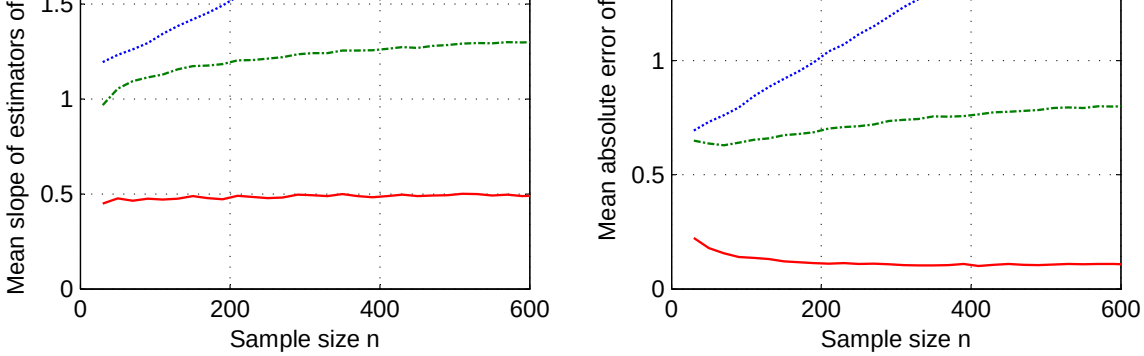

Table 1: Simulated mean values of the estimated slope and the mean absolute error for various selected values of the true intercept and slope when $0.08 \leq \lambda \leq 100$.

True slope	GM	OLS-B	WG	True model
0.55	3.4981	0.9340	0.5904	$\eta_j = 20 + 0.55x_j$
(MAE)	(2.9527)	(0.9989)	(0.5341)	
-0.75	-3.5299	-1.1455	-0.7857	$\eta_j = 27 - 0.75x_j$
(MAE)	(2.7910)	(0.8780)	(0.5328)	
1.2	3.6321	1.5622	1.2213	$\eta_j = -15 + 1.2x_j$
(MAE)	(2.4676)	(0.6548)	(0.5302)	

Figure 5: Graph of consistency evaluation of the mean slope of estimators, and the mean absolute error when the true $\beta_1 = 0.5$ and large λ.

Figure 6: Graph of consistency evaluation of the mean slope of estimators, and the mean absolute error when the true $\beta_1 = 1$ and large λ.

WGE is much closer to the true values of β_1 than other two estimators. It is clear, from Figures 1b-3b that the measurement error makes the mean absolute error of the geometric mean estimator the highest. While the mean absolute error of the OLS-bisector estimator appears to be smaller than those of the geometric mean estimator, they are not small. Obviously, the mean absolute error of the WGE is better and the smallest compared to the other estimators, and it seems to be stable over the range of selected ratio of error variances $0.08 \leq \lambda \leq 100$. Table 1 summarizes the results of the simulation studies which indicate that the proposed estimator is more precise than the other competing estimators. Sarach and Celik (2011) discussed eight different regression techniques, and pointed out that the OLS-bisector estimator is near to the real value than all other regression techniques, and the mean squares error of OLS-bisector is smaller than all other techniques. The current study reveals that the proposed WGE is consistently better than the OLS-bisector estimator in term of the closeness of $\hat{\beta}_{1WG}$ to β_1, and the mean absolute error as shown in Figures 5 and 6.

7 Concluding Remarks

This paper proposes a new estimator based on the mathematical relationship between the vertical and orthogonal distances of the observed points and the regression line. This estimator is appropriate to fitting a straight line when both variables are subject to measurement errors, especially when there is no basis for distinguishing between response and explanatory variables. This method is straightforward, and easy to implement.

Extensive simulation studies confirm that the values of the proposed WG estimator are always nearer to the true value of the slope more than the OLS-bisector, and the mean absolute error of WGE is consistently smaller than that of the OLS-bisector. Therefore, the proposed estimator possesses better statistical proprieties than the geometric mean and OLS-bisector estimators. The new method is stable and works well for different sample sizes and for different values of λ.

References

Adcock, R J (1877). Note on the method of least squares. Analyst., 4, 183-184.

Barker, F., Soh, Y C. and Evans, R J. (1988). Properties of the geometric mean functional relationship, Biometrics, 44(1), 279281.

Bartlett, M S. (1949). Fitting a stright line when both variables are subject to error. Biometrics 5, 207-212.

Burr, D. (1988). On Errors-in-Variables in Binary Regression-Berkson Case. J. Am. Statist. Assoc 83, 739-743.

Cheng, C L, and Ness, J W. (1999). *Kendall's Library Of Statistics 6, Statistical Regression With Measurement Error*. New York: Wiley.

Cragg, J G. (1997). Using higher moments to estimate the simple errors-in-variables model. The RAND Journal of Economics 28, 71-91.

Dent B M. (1935). On observation of points connected by a linear relation. Proc. Physical Soc. London, 47, 92-108.

Draper, N R. and Yang, Y. (1997). Generalization of the geometric mean functional relationship. Computational Statistics and Data Analysis 23 355-372.

Durbin, J. (1954). Errors-in-variables. Int. Statist. Rev 22, 23-32.

Feigelson E D, Babu G J. (1992). Linear regression in astronomy. II. Astrophys J 397, 5562.

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.

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

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

Geary, R C. (1943). Relations between statistics: The general and the sampling problem when the samples are large. Proceedings of the Royal Irish Academy 49, 177-196.

Geary, R C. (1948). Studies in relations between economics time series. Journal of the Royal Statistical Society, 10,158-172.

Geary, R C. (1949). Determination of linear relations between systematic parts of variables with errors of observation the variances of which are unknown. Econometrica 17, 30-58.

Halfon, E. (1085). Regression method in ecotoxicology: a better formulation using the geometric mean functional regression. Notes. Environ. Sci. Technol 19, 747-749.

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

Isobe T, Feigelson ED, Akritas MG, Babu GJ. (1990). Linear regression in astronomy I. Astrophys. J. 364, 10413.

Johnson, J. (1972). *Econometric Methods*. New York: McGraw Hill Book Company.

Jolicouer, P. (1975). Linear regressions in fishery research: some comments. J. Fish. Res. Board Can. 32(8), 14911494.

Kermack KA, Haldane JBS. (1950). Organic correlation and allometry. Biometrika 37, 3041.

Leng L, T. Zhang, L Kleinman, and W. Zhu (2007). Ordinary least square regression, orthogonal regression, geometric mean regression and their applications in aerosol science. Journal of Physics Conference Series 78 012084-012088.

Levinton JS, Allen BJ. (2005). The paradox of the weakening combatant: trade-off between closing force and gripping speed in a sexually selected combat structure. Funct Ecol 19, 159165.

Ludbrook, J. (2010). Linear regression analysis for comparing two measurers or methods of measurement: But which regression?. Clinical and Experimental Pharmacology and Physiology 37, 692699.

Macdonald J R, Thompson W J. (1992). Least-squares fitting when both variables contain errors: pitfalls and possibilities. Am J Physiol 60:6673.

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

Maddala, G.S. (2001). *Introduction to Econometrics*. Prentice Hall International, Inc, Second edition.

Pal, M. (1980). Consistent moment estimators of regression coecients in the presence of errors in variables. J. Econometrics 14, 349-364.

Reiersøl, O. (1950). Identifiability of a linear relation between variables which are subject to error. Econometrica 18, 375-89.

Ricker W. E. (1973). Linear regressions in Fishery research. J Fish. Res. Board Can 30, 409-434

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.

Saqr, A. and Khan, S. (2011) Instrumental variable estimator of the slope parameter when the explanatory variable is subject to measurement error. In: 11th Islamic Countries Conference on Statistical Sciences (ICCSS-11), 19-22 Dec 2011, Lahore, Pakistan, pp. 39-53.

Saqr, A. and Khan, S. (2012) Reflection method of estimation for measurement error models. Journal of Applied Probability and Statistics, 7 (2). pp. 71-88. ISSN 1930-6792

Saqr, A. and Khan, S. (2012a) Slope estimator for the linear error-in-variables model. In: 12th Islamic Countries Conference on Statistical Sciences (ICCS 2012): Statistics for Everyone and Everywhere, 19-22 Dec 2012, Doha, Qatar, pp. 61-70.

Saqr, A. and Khan, S. (2016) Mathematical reflection approach to instrumental variable estimation method for simple regression model. Pakistan Journal of Statistics, 32 (1). pp. 37-48. ISSN 1012-9367

Sladek V, Berner M and Sailer R. (2006). Mobility in central european late neolithic and early bronze age: femoral cross-sectional geometry. Am J Phys Anthropol 130, 320332.

Sarach S and Celik H (2011). Performance of OLS-bisector regression in method comparison studies. World Applied Sciences Journal 12(10):1860-1865.

Smith, R J. (2009). Use and Misuse of the Reduced Major Axis for Line-Fitting. American Journal Of Physical Anthropology 140, 476486.

Sprent, P and Dolby, G R. (1980). Query: the geometric mean functional relationship. Biometrics 36(3), 547-550

Teissier, G. (1948). La relation d'allometrie sa signification statistique et biologique. Biometrics 4, 14-53.

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

van Montfort, K. (1989). *Estimating in Structural Models with Non-Normal Distributed Varibles: Some Alternative Approaches*. Leiden. DSWO Press.

van Montfort, K., Mooijaart, A., and de Leeuw, J. (1987). Regression with errors in variables: estimators based on third order moments. Statist Neerlandica 41, 223-237.

Vincent SE, and Lailvaux SP. (2006). Female morphology, web design, and the potential for multiple mating in Nephila clavipes: do fat-bottomed girls make the spider world go round? Biol J Linn Soc 87, 95102.

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

Warton, D I and Wright I J, Falster D S, Westoby M. (2006). Bivariate line-fitting methods for allometry. Biol Rev 81, 259291.

Willmott, J., and Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research, 30(1), 79-82.

Zimmerman F, Breitenmoser-Wu rsten C, Breitenmoser U. (2005). Natal dispersal of Eurasian lynx (Lynx lynx) in Switzerland. J Zool 267, 381395.