Journal of the Korean
OR/ MS Society
Vol. 21, No. 1, April 1996

Applicability of the Ordinary Least Squares Procedure When Both Variables Are Subject to Error

Kil-Soo Kim*, Jai-Hyun Byun**, Bong-Jin Yum*

Abstract

An errors-in-variables model (EVM) differs from the classical regression model in that in the former the independent variable is also subject to error. This paper shows that to assess the applicability of the ordinary least squares (OLS) estimation procedure to the EVM, the relative dispersion of the independent variable to its error variance must be also considered in addition to Mandel's criterion. The effect of physically reducing the variance of errors in the independent variable on the performance of the OLS slope estimator is also discussed.

1. Introduction

The problem of fitting a straight line when both variables are subject to error has received considerable attention in the literature. Among others, Mandel [4] considered the ordinary least squares (OLS) procedure in this situation, and suggested that the "sensitivity of y with respect to x" be small for applicability of the OLS method. This paper examines the large-sample properties of the OLS slope estimator, and demonstrates that the "relative dispersion of the independent variable to its error variance" must be also considered in evaluating the performance and applicability of the OLS estimator when the independent variable is also subject to error. Further, the effect of refining or using more precise measuring instrument on the bias and variance of the OLS slope estimator is also discussed.

* Department of Industrial Engineering, KAIST, 373-1 Gusung-Dong, Yusong-Gu, Taejon 305-701.
** Department of Industrial Engineering, Gyeongsang National University, Chinju 660-701.

2. Large-Sample Properties and Applicability of OLS Procedure

The present investigation considers the following model.

$$\left.\begin{aligned} x_i &= \xi_i + \delta_i \\ y_i &= \eta_i + \varepsilon_i \\ \eta_i &= \alpha + \beta\xi_i \end{aligned}\right\} \quad i=1, 2,\cdots,N \tag{1}$$

where ξ_i and η_i are respectively the true, unobservable values of the independent and dependent variables, α and β are unknown constants, and δ_i and ε_i represent random measurement errors distributed as

$$(\delta_i, \varepsilon_i) \sim \text{BVN}\{(0, 0), \text{diag}(\sigma_\delta^2, \sigma_\varepsilon^2)\}$$

where 'BVN' stands for 'bivariate normal'. We further assume that error vectors are independent across i.

The model described in Eq. (1) is frequently called an errors-in-variables model (EVM) in the literature. The EVM is further classified into functional and structural one according as ξ, and hence η, is fixed and random, respectively (Kendall and Stuart [2]). In this paper, ξ is assumed to be fixed. A general survey of various estimation methods for the EVM is beyond the scope of this paper, and the reader is referred to Madansky [3], Moran [5], Kendall and Stuart [2], and Fuller [1] among others.

The OLS estimator of β is given by

$$b=\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})/\sum_{i=1}^{N}(x_i-\bar{x})^2.$$

Define the relative bias and variance of b as

$$RB(b) = \{E(b)-E(b)\}/\beta$$
$$RV(b) = V(b)/\beta^2.$$

Although the exact expressions for $RB(b)$ and $RV(b)$ are available in Richardson and Wu [6], they are so complicated that the relationship between parameters may not be easily assessed. In this paper we instead adopt the asymptotic expressions, the accuracy of which can be found in Yum [7]. Then the probability limit (plim), the asymptotic relative bias (ARB), and the asymptotic relative variance (ARV) of b are respectively given by (e.g., see Yum [7])

$$b_{\text{lim}} = \text{plim } b = \beta\psi(1+\psi)$$
$$ARB(b) = 1/(1+\psi) \tag{2}$$
$$ARV(b) = N^{-1}\{\theta/(1+\psi)+\psi(1+\psi^2)/(1+\psi)^4\} \tag{3}$$

where

$$\psi=\sum_{i=1}^{N}(\xi_i-\bar{\xi})^2/(N\sigma)=\sigma_\xi^2/\sigma_\delta^2$$
$$\theta=\sigma_\xi^2/(\beta^2\ \sigma_\delta^2).$$

If there had not been any error in ξ, then the OLS estimator would have been obtained by

$$b_0=\sum_{i=1}^{N}(\xi_i-\bar{\xi})(y_i-\bar{y})/\sum_{i=1}^{N}(\xi_i-\bar{\xi})^2.$$

with

$$RB(b_0)=\{\beta-E(b_0)\}/\beta=0 \tag{4}$$
$$\begin{aligned}RV(b_0)&=V(b_0)/\beta^2\\&=\sigma_\xi^2/\{\beta^2\sum_{i=1}^{N}(\xi_i-\bar{\xi})^2\}\\&=N^{-1}\theta/\psi.\end{aligned} \tag{5}$$

In the above, we evaluate the performance of b or b_0 in terms of relative biases and variances instead of absolute ones. This allows us to represent the quantities of interest in simpler forms as a function of θ and ψ only. Once $ARB(b)$ and $ARV(b)$ are estimated, we may conveniently assess the (asymptotic) bias and variance of b as percentages of β and $RV(b_0)$, respectively.

For applicability of the OLS procedure, Mandel [4] suggested that ϕ, the "sensitivity of y with respect to x", be small. That is,

$$\phi=|\beta|\sigma_\delta/\sigma_\xi \ll 1.$$

Since $\phi=1/\sqrt{\theta}$, the above inequality is equivalent to

$$\theta \gg 1.$$

However, it is clear from Eqs. (2)-(5) that θ is not the only parameter to be considered in evaluating the performance of b. Note that $ARB(b)$ cannot be controlled by θ. Further, as θ increases (or equivalently, as ϕ decreases) for given N and ψ, both $ARV(b)$ and $RV(b_0)$ increases, implying that we are comparing b and b_0 in a highly undesirable circumstances. All of these suggest that considering θ (or ϕ) alone as a criterion may not be appropriate and could be misleading.

Table 1 was constructed based upon Eqs. (2), (3), and (5). It is clear from the table that, for given N,

(1) as ψ increases $ARB(b)$ and $ARV(b)$ respectively approach to $RB(b_0)$ and $RV(b_0)$ for any θ, and

(2) for given ψ, the relative magnitudes of $ARV(b)$ and $RV(b_0)$ are determined by θ.

Table 1. Asymptotic Relative Bias and Variance of b and the Relative Variance of b_0.

ψ	$ARB(b)$	$N \cdot ARV(b)$	$N \cdot RV(b_0)$
10	0.09091	$0.09091\theta+0.06898$	0.01θ
20	0.04762	$0.04762\theta+0.04124$	0.05θ
30	0.03226	$0.03226\theta+0.02927$	0.03333θ
40	0.02439	$0.02439\theta+0.02266$	0.025θ
50	0.01961	$0.01961\theta+0.01848$	0.02θ
100	0.00990	$0.00990\theta+0.00961$	0.01θ
200	0.00498	$0.00498\theta+0.00490$	0.005θ
500	0.00200	$0.00200\theta+0.00198$	0.002θ
1000	0.00100	$0.00100\theta+0.00100$	0.001θ

Figure 1 illustrates contours of $N \cdot ARV(b)$ with respect to θ and ψ, and Figure 2 shows $ARB(b)$ as a function of ψ. These figures can be used to evaluate the performance of b as shown in the following sections. In summary, to evaluate the relative as well as absolute performance of b, θ and ψ must be jointly considered.

3. Sensitivity Analysis

Reduction of σ_δ^2 by scaling the variables does not affect θ and ψ. On the other hand, refining the measuring instrument may physically reduce σ_δ^2, and subsequently changes the magnitude of θ and ψ. Suppose that σ_δ^2 is reduced to σ_δ^2/c $(c>1)$. Then, the corresponding θ and ψ are increased to $c\theta$ and $c\psi$, respectively. The effect of such changes on the bias and variance of b can be assessed from Figures 1 and 2. For instance, suppose that currently $\theta=1.83$ and $\psi=54.17$ (point A in Figures 1 and 2). If σ_δ^2 is reduced to $\sigma_\delta^2/2$, then θ and ψ are increased to 3.66 and 108.34, respectively (point B in Figures 1 and 2). Varing c continuously in this way we obtain the bias and variance traces as shown in Figures 1 and 2. Note that as c increases (or equivalently, as σ_δ^2 goes to 0), the trace of the bias approaches to 0. To which value the trace of the variance approaches is not clear from Figure 1. However, theoretically it should approach to the minimum attainable value,

$$N \cdot RV(b_0) = \theta/\psi = 0.034.$$

The analyses in Sections 2 and 3 can be used to assess the applicability of the OLS estimator when the independent variable is also subject to error. Let h_B and h_V be the tolerable

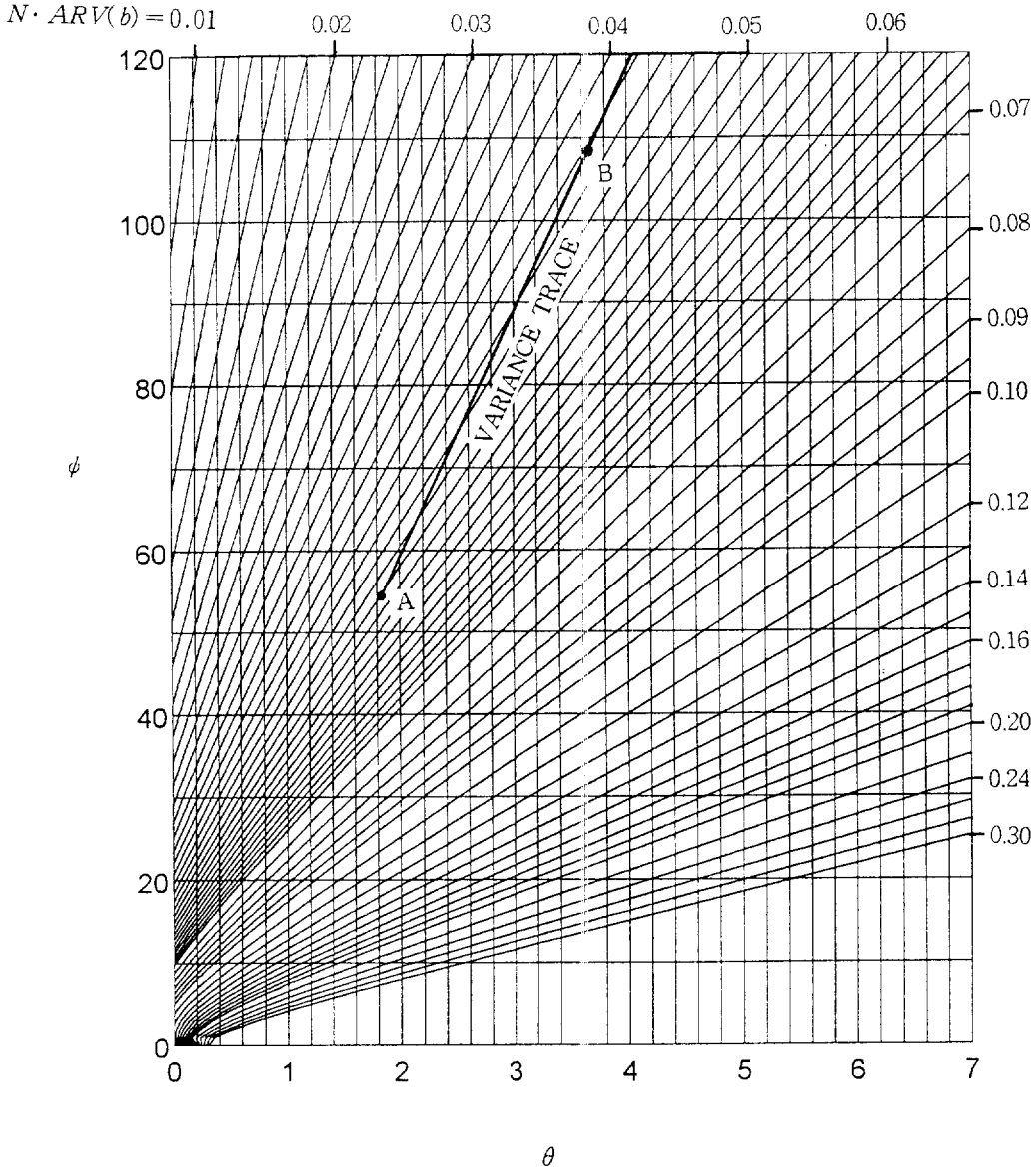

Figure 1. Contours of N times the Asymptotic Relative Variance of b with respect to θ and ψ.

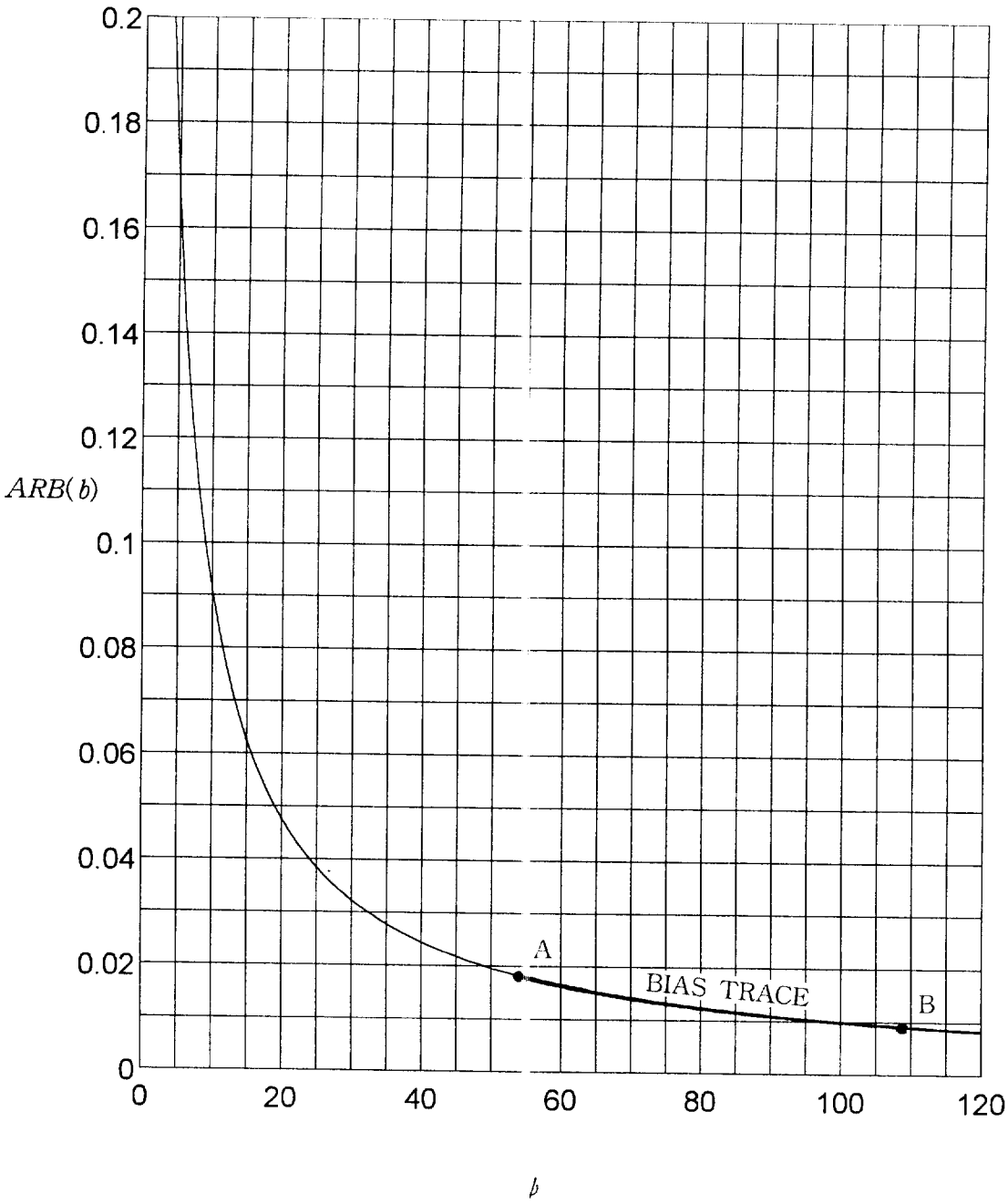

Figure 2. Asymptotic Relative Bias of b as a Function of ψ.

values (set by the user) for the relative bias of b and the ratio of $ARV(b)$ to $RV(b_0)$, respectively. If the estimated $ARB(b)$ and $ARV(b)/RV(b_0)$ are less than h_B and h_V, respectively, then we may use the OLS estimator even though the independent variable is subject to error. Otherwise, we need to consider physically reducing σ_δ^2 and its effects on $ARB(b)$ and $ARB(b)/RV(b_0)$ to see if the requirements can be satisfied within an affordable amount of effort.

4. Example

Suppose that in an estimation experiment with $N=25$, the following data were obtained.

$(x_i, y_i)=$(10.0, 14.7), (9.5, 13.4), (10.0, 15.2), (10.0, 15.7), (9.7, 15.5),
(11.9, 16.8), (11.1, 17.3), (12.1, 18.3), (12.5, 16.2), (11.8, 18.7),
(14.4, 19.8), (13.5, 21.6), (14.0, 22.8), (14.1, 21.3), (14.7, 20.1),
(16.3, 26.4), (16.0, 23.1), (16.5, 23.2), (16.1, 25.2), (16.1, 24.8),
(18.1, 24.8), (18.5, 28.4), (17.8, 25.3), (17.8, 27.5), (17.9, 27.1)

Further, assume that estimates of error variances were available from the historical data as $\hat{\sigma}_\delta^2=0.16$ and $\hat{\sigma}_\varepsilon^2=0.62$. These variance estimates could be also obtained by measuring the same quantity repeatedly to obtain replicated observations of (x, y) and calculating sample variances of x and y. From the above data the OLS estimate of β is obtained as $b=1.4569$. Then, "rough" estimates of θ and ψ are respectively given by

$$\hat{\theta}=\hat{\sigma}_\varepsilon^2/(b^2\hat{\sigma}_\delta^2)=1.83$$
$$\hat{\psi}=\sum_{i=1}^{N}(x_i-\bar{x})^2/(N\hat{\sigma}_\delta^2)=54.17$$

Finally, from point A in Figures 1 and 2, we obtain $N\cdot ARV(b)\approx 0.05$ and $ARB(b)\approx 0.018$, respectively, and $N\cdot RV(b_0)\approx 0.0338$ from Eq. (5). Suppose h_B and h_V are 0.05 and 1.3, respectively. The relative bias of b is about 0.02 and the variance ratio of b to b_0 is about 1.48 (0.05/0.0338). Note that the variance ratio exceeds h_V. To satisfy the requirement on the variance ratio, $N\cdot ARV(b)$ must be less than 0.044 ($=h_V\cdot N\cdot RV(b_0)$). This implies that ψ must be at least 90 from the variance trace in Figure 1, which can be achieved by reducing σ_δ^2 to σ_δ^2/c where $c\geq 1.66(=90/54.17)$. If the experimenter could reduce σ_δ^2 by half, say, then the relative bias could be decreased to 0.009 (point B in Figure 2) and $N\cdot ARV(b)$ becomes approximately 0.042 (point B in Figure 1). That is, the variance ratio is decreased to 1.24

(=0.042/0.0338), which meets the requirements for the applicability of the OLS estimator.

5. Concluding Remarks

When both variables are subject to error it is shown that θ and ψ must be jointly considered to evaluate the relative as well as the absolute performance of the OLS slope estimator. The figures and table may be utilized to assess to what extent errors in the variables affect the performance of the OLS estimator, and to evaluate the effectiveness of physically reducing measurement errors in the variables.

Although the present study assumes that $Cov(\delta, \varepsilon)=0$, it is straightforward to conduct a similar analysis for the case of nonzero covariance.

References

[1] Fuller, W.A., *Measurement Error Models*, Wiley, New York, 1987.

[2] Kendall, M.G. and A. Stvart, *The Advanced Theory of Statistics*, Vol.2, Macmillan, New York, 1979.

[3] Madansky, A., "The Fitting of Straight Lines When Both Variables Are Subject to Error", *J. Amer. Statist.*, Vol.54(1959), pp.173-205.

[4] Mandel, J., "Fitting Straight Lines When Both Variables Are Subject to Error", *J. Quality Technology*, Vol.16(1984), pp.1-14.

[5] Moran, P.A.P, "Estimating Structural and Functional Relationships", *J. Multivariate Analysis*, Vol.1(1971), pp.232-255.

[6] Richardson, D.H. and D. Wu, "Least Squares and Grouping Method Estimators in the Errors in the Variables Model", *J. Amer. Statist. Ass.*, Vol.65(1970), pp.724-748.

[7] Yum, B.J., "Asymptotic Properties of the OLS and GRLS Estimator for the Replicated Functional Relationship Model", *Comm. Statist.-Theor. Meth.*, Vol.14(1985), pp.1981-1996.