Jour. Ind. Soc. Ag. Statistics
46(2), 1994 : 290–297

The Effect of Measurement Errors Under Survey Conditions on Least Squares Estimator of Regression Coefficients

R.C. Agrawal* and O.P. Kathuria
Indian Agricultural Statistics Research Institute, New Delhi
(Received : January, 1994)

Summary

For a set of survey conditions, the effect of correlated measurement errors on ordinary Least-Square (OLS) estimator of the regression coefficient for a finite bivariate population, when both variables are subject to measurement errors has been studied. In this work, the expressions for relative bias and relative mean-square error (m.s.e.) of regression estimates have been derived. A procedure to use the tables for relative absolute bias and relative m.s.e. given by Richardson and Wu [5] has been explained.

Key words : Measurement errors, Regression coefficients, Ordinary Least- Square (OLS) estimates, Confluent hypergeometric function.

Introduction

The data collected in a sample survey may sometimes be subject to measurement errors. Main sources of measurement errors are response errors but coding and other processing errors may also occur. Fuller [2] has given practical examples of measurement errors in many areas. Increased application of the theory of measurement errors has been made in recent years.

However most of the work has been oriented towards the development of models and their application to univariate cases, under some basic survey conditions. For a set of survey conditions, the effect of correlated measurement errors on the OLS estimator of the regression coefficient for a finite bivariate population, when both variables are subject to measurement errors is studied here.

2. *Survey Conditions, Assumptions and the Model*

We consider the following basic survey conditions–

(a) A large population of N elementary units is divided into L contiguous

* Present Address : National Bureau of Plant Genetic Resources, Pusa Campus, New Delhi-12.

groups or areas. Each group or area contains N_i elementary units and thus, $\sum_{i=1}^{L} N_i = N$

(b) A random sample of n elementary units is taken such that n_i units are drawn from the i^{th} area. Thus, $\sum_{i=1}^{L} n_i = n$

(c) Let there be L interviewers assigned at random to each of L areas.

(d) The survey can be repeated independently under the same survey conditions.

2.1 *Regression model*

Chai [1] has used a mathematical model under some survey conditions for deriving mathematical expressions for the bias factor of the ordinary least squares estimator of the regression coefficient for the two variable linear model when both variables are subject to correlated measurement errors. In the present study, we consider their mathematical model under some different assumptions and with changed notations.

Let x_{ijt}, y_{ijt} be the observed values of the variables x and y for the j^{th} sample unit of the i^{it} area at the t^{th} trial. The conditional expected values of x and y given the j^{th} unit of the i^{th} area are, say

$$E_t(x_{ijt} \mid i, j) = X_{ij}$$

$$E_t(y_{ijt} \mid i, j) = Y_{ij}$$

where the expectation is taken over all trials. Following Hansen et al [4], we define the response deviation for x and y variables given the j^{th} sample unit of the i^{th} area as follows:

$$\delta_{ijt} = x_{ijt} - X_{ij}$$

$$\eta_{ijt} = y_{ijt} - Y_{ij}$$

We assume that each of the error terms δ and η follows a normal probability distribution with mean zero i.e.

$$E_t(\delta_{ijt} \mid i, j) = 0$$

$$E_t(\eta_{ijt} \mid i, j) = 0$$

and with variance,

$$\sigma^2_{\delta(i,j)} = E_t(\delta^2_{ijt} \mid i, j)$$

$$\sigma^2_{\eta(i,j)} = E_t(\eta^2_{ijt} \mid i, j)$$

We consider the simple linear model given by,

$$Y_{ij} = \alpha + \beta x_{ij} \tag{2.1}$$

where α and β are the parameters of the model and X_{ij} and Y_{ij} are the conditional expected values of x_{ijt} and y_{ijt}.

We observe from a sample of n units a set of values x_{ijt} and y_{ijt}. Hence model (2.1) can be written as-

$$y_{ijt} = \alpha + \beta x_{ijt} + \varepsilon_{ijt} \tag{2.2}$$

where, $\varepsilon_{ijt} = \eta_{ijt} - \beta \delta_{ijt}$

In model (2.2), vector δ_{ijt} and η_{ijt} ($i = 1, 2. . . L, j = 1, 2. . . n_i$) are mutually independent and $(\varepsilon_{ijt}, \delta_{ijt})$ is bivariate normal with mean vector (0 0) and non-singular var-cov matrix-

$$\Sigma = \begin{pmatrix} \sigma^2_\eta + \beta^2 \sigma^2_\delta & -\beta \sigma^2_\delta \\ -\beta \sigma^2_\delta & \sigma^2_\delta \end{pmatrix}$$

where,

$$\sigma^2_\eta = \frac{1}{N} \sum_i^L \sum_j^{N_i} \sigma^2_{\eta(i,j)}$$

$$\sigma^2_\delta = \frac{1}{N} \sum_i^L \sum_j^{N_i} \sigma^2_{\delta(i,j)}$$

If ρ is the correlation coefficient between δ_{ij} and ε_{ij}, then,

$$\rho = \frac{-\beta \sigma_\delta}{\sqrt{(\sigma^2_\eta + \beta^2 \sigma^2_\delta)}}$$

We know that the ordinary least-square estimator of B is is given by,

$$b_{(t)} = \frac{s_{xy(t)}}{s^2_{x(t)}}$$

where $s_{xy(t)}$ and $s^2_{x(t)}$ may be defined as follows :

$$s^2_{x(t)} = \frac{1}{n} \sum_{i}^{L} \sum_{j}^{n_i} (x_{ijt} - \bar{x}_t)^2$$

$$s_{xy(t)} = \frac{1}{n} \sum_{i}^{L} \sum_{j}^{n_i} (x_{ijt} - \bar{x}_t)(y_{ijt} - \bar{y}_t)$$

where,

$$\bar{x}_t = \frac{1}{n} \sum_{i}^{L} \sum_{j}^{n_i} x_{ijt}$$

$$\bar{y}_t = \frac{1}{n} \sum_{i}^{L} \sum_{j}^{n_i} y_{ijt}$$

We shall be using the results derived by Halperin and Gurian [3] which are summarized in the following section.

3. *Halperin and Gurian's note*

Let $y_i = \alpha + \beta x_i + \varepsilon_i$, $x_i = \xi_i + \delta_i$, $i = 1, 2, \ldots N$, $N \geq 3$ where $(\varepsilon_i, \delta_i)$ are independent bivariate normal with zero means, variances σ_ε^2, σ_δ^2 and correlation coefficient ρ, the ξ_i's being unknown constants; at least two of which are distinct i.e. y_i's are independent drawings from distribution with expectation $\alpha + \beta x_i$, common variance σ^2_ε, α and β being unknown. They have considered the case where the vector $(\varepsilon_i\ \delta_i)$, $i = 1, 2, \ldots N$ is mutually independent and is bivariate normal with mean vector (0 0) and non-singular var-covariance matrix :

$$\Sigma = \begin{pmatrix} \sigma^2_\varepsilon & \rho\, \sigma_\varepsilon\, \sigma_\delta \\ \rho\, \sigma_\varepsilon\, \sigma_\delta & \sigma^2_\delta \end{pmatrix}$$

To simplify the results, they have used to following notations-

$$N-1 = m,\ z = \sum_{i=1}^{N} (\xi_i - \bar{\xi}_i)/2\sigma^2_\delta$$

and use the usual notation ${}_1F_1(\alpha, \beta, z)$ for the confluent hypergeometric function of arguments α, β and z.

Using these notations, their results may be summarized as follows.

$$E(b) = \left(\beta - \rho\frac{\sigma_\varepsilon}{\sigma_\delta}\right)\left(\frac{2z}{m}e^{-z}\right){}_1F_1\left(\frac{m}{2}, \frac{m}{2}+1, z\right) + \frac{\rho\,\sigma_\varepsilon}{\sigma_\delta} \tag{3.1}$$

$$E(b-\beta)^2 = \frac{e^{-z}}{m-2}\left\{\left[\left(\beta - \rho\frac{\sigma_\varepsilon}{\sigma_\delta}\right)^2 + \frac{\sigma_\varepsilon}{\sigma_\delta^2}(1-\rho^2)\right]{}_1F_1\left(\frac{m}{2}-1, \frac{m}{2}, z\right) + (m-3)\left(\beta - \rho\frac{\sigma_\varepsilon}{\sigma_\delta}\right)^2 {}_1F_1\left(\frac{m}{2}-2, \frac{m}{2}, z\right)\right\} \tag{3.2}$$

4. *The Relative Bias of the Regression Coefficient*

We can put the results defined in section 3 for model defined in eq. (2.2) after following substitutions

$$z = \sum_{i=1}^{L}\sum_{j=1}^{N_i}(X_{ij} - \bar{X})^2/2\sigma_\delta^2$$

$$\rho = -\frac{\beta\,\sigma_\delta}{\sqrt{(\sigma_\eta^2 + \beta^2\,\sigma_\delta^2)}}$$

$$\sigma_\varepsilon^2 = \sigma_\eta^2 + \beta^2\,\sigma_\delta^2$$

Using the recurrence relation (Slater) [6],

$$x\ {}_1F_1(a+1, b+1, x) = b\,[\,{}_1F_1(a+1, b, x) - {}_1F_1(a, b, x)]$$

and the relation ${}_1F_1(a, a, x) = e^x$, we can rewrite the expression defined in eq. (3.1) as

$$E(b) = \beta\left[1 - e^{-z}\cdot{}_1F_1\left(\frac{m}{2}-1, \frac{m}{2}, z\right)\right] - \frac{\rho\sigma_\varepsilon}{\sigma_\delta}\left[1 - e^{-z}\cdot{}_1F_1\left(\frac{m}{2}-1, \frac{m}{2}, z\right)\right] + \frac{\rho\sigma_\varepsilon}{\sigma_\delta}$$

The bias of b can be written as

$$E(b) - \beta = -e^{-z}\cdot{}_1F_1\left(\frac{m}{2}-1, \frac{m}{2}, z\right)\left[\beta - \frac{\rho\sigma_\varepsilon}{\sigma_\delta}\right]$$

Further relative bias can be written as

$$\frac{E(b)-\beta}{\beta} = -e^{-z}\cdot{}_1F_1\left(\frac{m}{2}-1, \frac{m}{2}, z\right)\left[1 - \frac{\rho\sigma_\varepsilon}{\beta\sigma_\delta}\right]$$

But, since $\frac{\rho\,\sigma_\varepsilon}{\beta\sigma_\delta} = -1$, the expression for relative bias will be

$$\frac{E(b)-\beta}{\beta} = -2e^{-z}\,{}_1F_1\left(\frac{m}{2}-1, \frac{m}{2}, z\right) \tag{4.1}$$

Since, $e^{-z}\,{}_1F_1\left(\frac{m}{2}-1, \frac{m}{2}, z\right) > 0$ for $z > 0$ and $m > 1$, the relative bias of b will be always negative.

Absolute relative bias of b can be derived from Table A–1 of Richardson and Wu [5] by multiplication factor 2.

In the tables of Richardson and Wu [5], notation n (defined as N–1) can be considered as m in our case.

For example, if $N = 21, \tau = 2\left[\text{where } z = \left(\frac{m}{2}\right)\tau\right]$ then $m = N-1 = 20$ and absolute value of relative bias = 0.6360

By employing the asymptotic expansion of ${}_1F_1(a, b, by)$ for large a and b, (b–a) and y bounded, we can obtain a large sample approximation to the relative bias of b.

It can be seen (Slater) [6] that

$${}_1F_1(a, b, by) = e^{by}(1+y)^{(a-b)}\left[1 - \frac{(b-a)(b-a+1)y^2}{2b(1+y)^2} + O(|b|^{-2}\right]$$

If we put $z = \left(\frac{m}{2}\right)\tau$, then,

$$\frac{E(b)-\beta}{\beta} = \frac{-2}{(1+\tau)}\left[1 - \frac{2\tau^2}{m(1+\tau)^2} + O(m^{-2})\right]$$

We can see from this expression that to the order of approximation, as the sample size decreases the absolute value of the relative bias increases. Further, in the limiting case of m (as $m \to \infty$), the absolute value of the relative bias tends to the factor $\left(\frac{2}{1+\tau}\right)$

For example, if $m \to \infty$

$\tau = 1$, absolute relative bias = 1.0000
$\tau = 2$, absolute relative bias = 0.6666

5. *The Mean-Square Error of the Regression Coefficient*

We can write the expression for relative m.s.e. using the eq (3.2) as

$$E(b-\beta)^2 = \frac{e^{-z}}{m-2}\left\{\left[\left(\beta-\frac{\rho\,\sigma_\varepsilon}{\sigma_\delta}\right)^2 + \frac{\sigma_\varepsilon}{\sigma_\delta^2}(1-\rho^2)\right]{}_1F_1\left(\frac{m}{2}-1,\frac{m}{2},z\right) + (m-3)\left(\beta-\frac{\rho\,\sigma_\varepsilon}{\sigma_\delta}\right)^2 {}_1F_1\left(\frac{m}{2}-1,\frac{m}{2},z\right)\right\}$$

After simplifications, we get

$$\frac{E(b-\beta)^2}{\beta^2} = \frac{4\,e^{-z}}{(m-2)}\left\{1+\left(\frac{\sigma_\varepsilon^2(1-\rho^2)}{(\beta\sigma_\delta-\rho\sigma_\varepsilon)^2}\right){}_1F_1\left(\frac{m}{2}-1,\frac{m}{2},z\right) + (m-3)\,{}_1F_1\left(\frac{m}{2}-1,\frac{m}{2},z\right)\right\}$$

We note the values of relative mean square error $\frac{E(b-\beta)^2}{\beta^2}$ can be obtained directly from tables A–2 (a–d) of Richardson and Wu[5] with the proviso that the tables which are computed for selected values of $\frac{\beta^2\sigma_\delta^2}{\sigma_\varepsilon^2}$ are taken to be for the same values of $\frac{(\beta\sigma_\delta-\rho\sigma_\varepsilon)^2}{\sigma_\varepsilon^2(1-\rho^2)}$. The values so obtained, when multiplied by 4 are values of relative mean square.

For example, if $\frac{\beta\sigma_\delta-\rho\sigma_\varepsilon}{\sigma_\varepsilon^2(1-\rho^2)} = 0.25, m = 10, \tau = 5$, relative m.s.e. = 0.4128

To obtain a large sample approximation to the relative m.s.e, we can apply the asymptotic expansion formula for large m to the Confluent hypergeometric function as has been done for relative bias.

We know that,

$$\frac{M.S.E.}{\beta^2} = \frac{var(b)}{\beta^2} + \frac{[E(b)-\beta]^2}{\beta^2}$$

where, for large sample approximation,

$$\text{var}(b) = \frac{4\beta^2}{(n-2)}\left\{\frac{\sigma_\varepsilon^2(1-\rho^2)}{(\beta\sigma_\delta - \rho\sigma_\varepsilon)^2(1+\tau)} + \frac{\tau(1+\tau^2)}{(1+\tau)^4}\right\} + O(n^{-2})$$

So, in limiting case, relative m.s.e. will tend to the square of the relative absolute bias and will not depend upon the factor $\frac{(\beta\sigma_\delta - \rho\sigma_\varepsilon)^2}{\sigma_\varepsilon^2(1-\rho^2)}$.

For example, if $\tau = 6$, relative m.s.e. = 0 0.0816

REFERENCES

[1] Chai., John J., 1971. Correlated measurement errors and the least square estimator of the regression coefficient. *J. Amer. Statist. Assoc.* **66**, 478–483.

[2] Fuller., W.A., 1988. Measurement errors models. *Wiley Series in Probability and Mathematical Statistics*. 440 p.

[3] Halperin., M., and Gurian., J., 1971. A note on estimation in straight line regression when both variables are subject to errors. *J. Amer. Statist. Ass.* **66**, 587–589.

[4] Hansen., M.H., Hurvitz., W.N. and Bershad., M. A., 1961. Measurement errors in censuses and surveys. Bull. *Int. Statist Inst.*, **38**, Part II, 359–374.

[5] Richardson., D.H. and Wu., De-min, 1970. Alternative estimators in the errors in variables model. *J. Amer. Statist. Assoc.* **65**, 724–748.

[6] Slater., L.J., 1960. Confluent Hypergeometric Functions. Cambridge, Cambridge University Press.