MOMENT ESTIMATION OF MEASUREMENT ERRORS

Diarmuid O'Driscoll and Donald E. Ramirez

Department of Mathematics and Computer Studies, Mary Immaculate College, Limerick
E-mail address: diarmuid.odriscoll@mic.ul.ie

Department of Mathematics, University of Virginia, Charlottesville, VA 22904
E-mail address: der@virginia.edu

ABSTRACT

The slope of the best-fit line from minimizing a function of the squared vertical and horizontal errors is the root of a polynomial of degree four. We use second order and fourth order moment equations to estimate the ratio of the variances of errors in the measurement error model and this estimate is used to introduce two new estimators. A simulation study shows improvement in bias and mean squared error of each of these new estimators over the ordinary least squares estimator.

Keywords: Measurement errors; Maximum Likelihood estimation; Moment estimating equations; Oblique estimators

1 Introduction

With ordinary least squares $OLS(y|x)$ regression we have data $\{(x_1, Y_1|X = x_1),...,(x_n, Y_n)|X_n = x_n\}$ and we minimize the sum of the squared vertical errors to find the best-fit line $y = h(x) = \beta_0 + \beta_1 x$ where it is assumed that the independent or causal variable X is measured without error. The measurement error model does not assume that X is measured without error, has wide interest with many applications and has been studied in depth by Carroll et al. (2006) and Fuller (1987). As in the regression procedure of Deming (1943) to account for both sets of errors we determine a fit so that a function of both the squared vertical and the squared horizontal errors will be minimized.

2 Minimizing Squared Oblique Errors

From the data point (x_i, y_i) to the fitted line $y = h(x) = \beta_0 + \beta_1 x$ we define the vertical length $v_i = |y_i - \beta_0 - \beta_1 x_i|$ from which it follows that the sum of the squares of the oblique lengths from (x_i, y_i) to $(h^{-1}(y_i) + \lambda(x_i - h^{-1}(y_i)), y_i + \lambda(h(x_i) - y_i))$ is

$$SSE_o(\beta_0, \beta_1, \lambda) = (1-\lambda)^2 \sum v_i^2 / \beta_1^2 + \lambda^2 \sum v_i^2 \ . \qquad (1)$$

In a comprehensive paper by Riggs et al. (1978), the authors state that: "It is a poor method indeed whose results depend upon the particular units chosen for measuring the variables." So that our equation is dimensionally correct we consider the standardized weighted average

$$SSE_o(\beta_0, \beta_1, \lambda) = (1-\lambda)^2 S_{yy} \sum v_i^2 / \beta_1^2 + \lambda^2 S_{xx} \sum v_i^2$$

The solution of $\delta SSE_o / \delta \beta_0$ is given by $\beta_0 = \bar{y} - \beta_1 \bar{x}$ and the solutions of $\delta SSE_o / \delta \beta_1 = 0$ are the roots of the fourth degree polynomial, $P_4(\beta_1)$,

$$\lambda^2 (s_{xx} / s_{yy})^{1.5} \beta_1^4 - \lambda^2 \rho \, s_{xx} / s_{yy} \, \beta_1^3 + (1-\lambda)^2 \rho \beta_1 - (1-\lambda)^2 (s_{yy} / s_{xx})^{0.5} \quad (2)$$

From O'Driscoll et al. (2008), the Complete Discrimination System $\{D_1, ..., D_n\}$ of Yang is a set of explicit expressions that determine the number (and multiplicity) of roots of a polynomial. This system is used to show that the fourth order polynomial $P_4(\beta_1)$ has exactly two real roots, one positive and one negative with the global minimum being the positive (respectively negative) root corresponding to the sign of $s_{xy} = S_{xy} / n$.

With λ = 1 we recover the minimum squared vertical errors with estimated slope β_1^{ver} and with λ = 0 we recover the minimum squared horizontal errors with estimated slope β_1^{hor}. The geometric mean estimator $\beta_1^{gm} = \sqrt{s_{yy} / s_{xx}}$ has the oblique parameter λ =0.5 and for the measurement error model, when both the vertical and horizontal models are reasonable, a compromise estimator such as β_1^{gm} is widely used and is hoped to have improved efficiency. However, Lindley and El-Sayyad (1968) proved that the expected value of β_1^{gm} is biased unless $\kappa = \sigma_Y^2 / \sigma_X^2$ where $\kappa = \sigma_\tau^2 / \sigma_\delta^2$

3 Measurement Error Model; Second and Fourth Moment Estimation

We now consider the measurement error model as follows. In this paper it is assumed that X and Y are random variables with respective finite variances σ_X^2 and σ_Y^2, finite fourth moments and have the linear functional relationship $Y = \beta_0 + \beta_1 x$. The observed data $\{(x_i, y_i), 1 \le i \le n\}$ are subject to error by $x_i = X_i + \delta_i$ and $y_i = Y_i + \tau_i$ where it is also assumed that δ is $N(0, \sigma_\delta^2)$ and τ is $N(0, \sigma_\tau^2)$. It is well known, in a measurement error model, that the expected value for β_1^{ver} ($OLS(y|x)$) is attenuated to zero by the attenuating factor $\sigma_X^2 / (\sigma_X^2 + \sigma_\delta^2)$ called the reliability ratio by Fuller (1987) and similarly the expected value for β_1^{hor} ($OLS(x|y)$) is amplified to infinity by the amplifying factor $(\sigma_Y^2 + \sigma_\tau^2) / \sigma_Y^2$. From Gillard and Iles (2009), using the second moment estimating equation, we derive the Frisch hyperbola of Van Montfort (1989)

$$(s_{xx} - \tilde{\sigma}_\delta^2)(s_{yy} - \tilde{\sigma}_\tau^2) = s_{xy}^2 \quad (3)$$

and from the fourth order moments

$$(s_{xxxy} - 3 s_{xy} \tilde{\sigma}_\delta^2)(s_{xy}^2) = (s_{xx} - \tilde{\sigma}_\delta^2)^2 (s_{xyyy} - 3 s_{xy} \tilde{\sigma}_\tau^2) \quad (4).$$

We use these two equations to solve for $\tilde{\sigma}_\delta^2$ and $\tilde{\sigma}_\tau^2$ imposing suitable restrictions on the possible solutions; firstly the variances must be positive; secondly the kurtosis of the underlying distribution must be significantly different from the kurtosis of the normal distribution to assure the validity of Equation (4) and thirdly the sample sizes must be adequately large. We then use these solutions as estimates for the ratio κ in the maximum likelihood estimator as described in Section 4.

4 The Maximum Likelihood Estimator

If the ratio of the error variances $\kappa = \sigma_\tau^2 / \sigma_\delta^2$ is assumed finite, then Madansky (1959), among others, showed that the maximum likelihood estimator for the slope is

$$\beta_1^{mle} = \frac{(s_{yy} - \kappa s_{xx}) + \sqrt{(s_{yy} - \kappa s_{xx})^2 + 4\kappa\rho^2 s_{xx} s_{yy}}}{2\rho\sqrt{s_{xx} s_{yy}}} \tag{5}$$

It also follows that if κ = 1 in Equation (5) then the MLE (often called the Deming Regression estimator) is equivalent to the perpendicular estimator, β_1^{per} first introduced by Adcock (1878). In the particular case where $\kappa = s_{yy} / s_{xx}$ then β_1^{mle} has a λ value of 0.5. Using the solutions from equations (3) and (4) as estimates for κ in β_1^{mle}, we introduce a new estimator β_1^{kap} which performs very well in our Monte Carlo simulation.

5 Relation between kappa and lambda

The invertible function $\psi : [0, \infty] \rightarrow [0, 1]$ defined by $\lambda = \psi(\kappa) = c\kappa / (c\kappa + 1)$, $c = s_{xx} / s_{yy}$, creates a new estimator β_1^{lam} with κ estimated as in Section 4. This proposed oblique estimator also performs very well in our Monte Carlo simulation. Since the range of κ includes infinity, we do not compute its average value in our simulation. Instead, we compute the average λ value for β_1^{lam}, and use $\psi^{-1}(\bar{\lambda})$ as the effective average $\tilde{\kappa}$ for κ.

6 Monte Carlo Simulation

To determine the efficiency of the above six estimators we conducted a Monte Carlo simulation which assigns a Uniform Distribution over the interval (0,20) to X and sets $Y = X$. Both X and Y are subjected to errors $(\sigma_\delta^2, \sigma_\tau^2) \in \{1,4,9\} \times \{1,4,9\}$ and the sample size n is set to 100. Our simulations use R = 1000 and we report in Tables 1-4 the MSE and the Bias for the estimators $\{\beta_1^{ver}, \beta_1^{gm}, \beta_1^{hor}, \beta_1^{per}, \beta_1^{kap}, \beta_1^{lam}\}$. Table 5 reports the effective average for $\tilde{\kappa}$ for $(\sigma_\delta^2, \sigma_\tau^2) \in \{1,4,9\} \times \{1,4,9\}$

7 Summary

Our simulations support the claim that our estimators β_1^{kap} and β_1^{lam} are more efficient than the ordinary least squares estimator β_1^{ver}.

Table 1

X is UD(0,20), β_1=1, β_0=0, R=1000, n=100, σ_τ=1, σ_δ=3

	MSE 10^{-3}	%Bias	λ	θ_λ
β_1^{ver}	46.569	-21.189	1	51.76
β_1^{gm}	11.897	-9.947	0.500	95.99
β_1^{hor}	4.402	2.9572	0	134.17
β_1^{per}	15.130	-11.246	0.556	89.93
β_1^{kap}	4.625	-1.382	0.169	118.37
β_1^{lam}	4.442	-0.029	0.237	123.49

Table 2

X is UD(0,20), β_1=1.25., β_0=0, R =1000, n=100, σ_τ=1, σ_δ=3

	MSE 10^{-3}	%Bias	λ	θ_λ
β_1^{ver}	70.809	-20.929	1	45.33
β_1^{gm}	18.425	-10.036	0.500	83.29
β_1^{hor}	5.708	2.413	0	127.99
β_1^{per}	15.081	-8.546	0.434	89.90
β_1^{kap}	6.304	-1.180	0.171	114.70
β_1^{lam}	5.847	0.092	0.145	116.62

Table 3

X is UD(0,20), β_1=1, β_0=0, R =1000, n =100, σ_τ=2, σ_δ=2

	MSE 10^{-3}	%Bias	λ	θ_λ
β_1^{ver}	13.403	-10.688	1	48.23
β_1^{gm}	2.117	0.0989	0.500	89.94
β_1^{hor}	18.146	12.232	0	131.70
β_1^{per}	2.672	0.126	0.500	89.92
β_1^{kap}	4.432	0.295	0.495	90.38
β_1^{lam}	5.962	0.425	0.497	90.14

Table 4

X is UD(0,20), β_1=0.75, β_0=0, R =1000, n =100, σ_τ=2, σ_δ=2

	MSE 10^{-3}	%Bias	λ	θ_λ
β_1^{ver}	7.791	-10.518	1	56.13
β_1^{gm}	2.603	4.196	0.500	103.99
β_1^{hor}	28.487	21.417	0	137.68
β_1^{per}	2.041	0.169	0.640	89.96
β_1^{kap}	4.233	0.725	0.590	95.55
β_1^{lam}	5.402	-0.029	0.615	92.97

Table 5

Effective $\widetilde{\kappa}$ average, X is UD(0,20), $\beta_1=1$, $\beta_0=0$, R =1000, $n=100$

	$\sigma_\tau^2=1$	$\sigma_\tau^2=4$	$\sigma_\tau^2=9$
$\sigma_\delta^2=1$	1.1781	3.3975	6.1251
$\sigma_\delta^2=4$	0.3185	0.9169	1.9514
$\sigma_\delta^2=9$	0.1701	0.4090	1.1658

REFERENCES

Adcock, R. J. (1878). A problem in least-squares. *The Analys*t, 5, 53-54.

Carroll, R. J., Ruppert, D., Stefanski, L. A., Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models - A Modern Perspective, Second Edition*. Boca Raton: Chapman & Hall/CRC.

Deming, W. E. (1943). *Statistical Adjustment of Data.* New York: Wiley.

Fuller, W.A. (1987). *Measurement Error Models*. New York: Wiley.

O'Driscoll, D., Ramirez, D., Schmitz, R. (2008). Minimizing oblique errors for robust estimation. *Irish. Math. Soc. Bulletin*, 62,71-78.

Gillard, J., Iles T. (2009). Methods of fitting straight lines where both variables are subject to measurement error. *Current Clinical Pharmacology*, 4, 164-171.

Lindley, D., El-Sayyad, M. (1968). The Bayesian estimation of a linear functional relationship. *Journal of the Royal Statistical Society Series B (Methodological,),* 30, 190-202.

Madansky, A. (1959). The fitting of straight Lines when both variables are subject to error. *Journal of American Statistical Association,* 54, 173-205.

Riggs, D., Guarnieri, J., Addelman, S. (1978). Fitting straight lines when both variables are subject to error. *Life Sciences,* 22, 1305-1360.

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