EfficientWald-Type Estimators for Simple Linear Measurement Error Model

Ahmed Al-Radaideh,
Department of Educational, psychological and Teaching Sciences, Doctoral Course in Sciences of the Mind and Human Relations, Salento University, Lecce, Italy
radaidehstat@yahoo.com

Amjad D. Al-Nasser,
Department of Statistics, Science Faculty, Yarmouk University, 21163 Irbid, Jordan
amjadn@yu.edu.jo

Enrico Ciavolino,
Dipartimento di Filosofia e Scienze Sciali, Università del Salento, Lecce, Italia
enrico.ciavolino@ateneo.unile.it

Abstract:
In this paper, Wald-type estimators for simple linear measurement error model presented based on L Ranked Set Sampling (LRSS) technique. The proposed estimators are compared to their counterparts based on Simple Random Sampling (SRS) and Ranked Set Sampling (RSS). It appears that the suggested estimators are more efficient. A real data set of student achievements is studied, a simulated data also is used to show how much efficient the use of the ranked data to estimate the EIV model parameter.

Keywords: Errors–in-Variables, Grouping Methods, Ranked Set Sampling, L Ranked Set Sampling, Simple Random Sampling.

1. Introduction

In the setting of the classical linear regression model the researcher assume that the explanatory variable X is measured exactly without errors, however in the most regression problems, measurement errors affect the explanatory and the response variable (Madansky A. , 1959). To illustrate the effect of measurement errors on explanatory variables; suppose that two mathematical variables ξ and η are assumed to have a relationship of the following form:

$$\boldsymbol{\eta} = \alpha + \beta \cdot \boldsymbol{\xi} \tag{1}$$

These two mathematical variables (ξ, η) are observed subject to random errors, where the observed values (x, y) are given by:

$$\mathbf{y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{2}$$

$$\mathbf{x} = \boldsymbol{\xi} + \boldsymbol{\delta} \tag{3}$$

Where ε_i and δ_i are assumed to be mutually independent random errors with known means and unknown variances $\left(\sigma^2, \tau^2\right)$, respectively. The model given in (1) - (3) is well known by Errors -In-Variables (EIV). There are many methodologies to estimate the unknown parameters α and β for model (1)-(3) suggested in the literatures, the easiest is the grouping methods which was proposed by Wald (1940) and Bartlett (1949), alternative methods of estimation of EIV model can be found in Chen and Van ness (1999), and Al-Nasser et al (2005). All the previous methods studied the EIV

model using a sample taken from simple random sampling (SRS). However, in order to reduce the time and the cost, we will use the ranked set sampling (RSS).

2. Grouping Methods

Various alternative methods for estimation EIV model parameters have been suggested by the researchers in this field. Grouping methods first proposed by Wald (1940) which is simpler than the other methods and hence/or otherwise may easily be applied in many practical cases. The grouping estimators are well known as wald-type estimators. Suppose that we have two variables *x*, *y*, and *n* pairs (x_i, y_i), then in general, the grouping methods can be described with the following steps:

1- Order the *n* pairs (x_i, y_i) by the magnitude of x_i; where $i=1,\ldots, n$.

2- Select proportions P_1 and P_2 such that $P_1 + P_2 \leq 1$, place the first nP_1 pairs in group one (G_1), and the last nP_2 pairs in another group (G_3), discarding G_2 the middle group of observations; that is to say:

$$\text{Place } (x_i, y_i)' \text{ in } \begin{cases} G_1 & if \quad x_i \leq x_{p_1} \\ G_2 & if \quad x_{p_1} < x_i \leq x_{1-p_2} \\ G_3 & if \quad x_i > x_{1-p_2} \end{cases} \tag{4}$$

where x_{p_i} is the p_i percentile. The slope can be estimated or formulated as follows:

$$\hat{\beta} = \frac{P_2^{-1}\sum_{G_3} Y_{np_2(i)} - P_1^{-1}\sum_{G_1} Y_{np_1(i)}}{P_2^{-1}\sum_{G_3} X_{np_2(i)} - P_1^{-1}\sum_{G_1} X_{np_1(i)}}, \; i=1,\ldots\ldots,n \tag{5}$$

Consequently, the intercept estimator will be:

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \tag{6}$$

Noting that when $P_1=P_2=1/2$ then the grouping method named by two groups (Wald ,1940), and when $P_1=P_2=1/3$ then the method is called three groups, (Nair and Shrivastava ,1942) and (Bartlett ,1949).

3. L Ranked Set Sampling for Bivariate Data

The LRSS proposed by Al-Nasser (2007) is a generalization procedure for many existing sampling techniques. Later, Al-Nasser and Al-Radaideh (2008) used LRSS to fit simple linear regression.
Assume that *(X, Y)* is a bivariate random vector such that variable *Y* is difficult to be measured, but the concomitant variable *X*, which is collected with *Y*, is easier to measure. Then one can follow the steps below in order to select LRSS:

Step1 Randomly draw *m* independent sets each containing *m* bivariate sampling units.

Step2 Rank the units within each sample with respect to the $X's$ by visual inspection or any other cheap method.

Step3 Select LRSS coefficient, $K = \lceil mp \rceil$ such that $0 \leq p < 0.5$, and $\lceil Z \rceil$ the largest integer value less than or equal to Z.

Step4 For each of the first $(k+1)$ ranked samples; select the unit with rankb $k+1$ and measure the Y value that corresponding to $x_{(k+1)i}$ and denote it by $y_{[k+1]i}$.

Step5 For $j = k+2,.....,m-k-1$, the unit with rank j in the j^{th} ranked sample is selected and measures the y value that corresponds.

Step6 The procedure continued until $(m-k)^{th}$ unit selected from the each of the last $(m-k)^{th}$ ranked samples, with respect to the first characteristic and measure the correspond y value.

4. Grouping Estimators Based on Ranked Data:

Because we believe that the use of the ranked data with the grouping methods will introduce more efficient estimates, in this section we introduce the two groups and the three groups estimates using the selected samples by the LRSS. In order to fit the EIV model, let:

$$L^{x}_{(i)j} = \begin{cases} X_{(k+1)j} & i \le k+1 \\ X_{(i)j} & k+2 \le i \le m-k-1 \; ; \; j = 1,2,..,r \\ X_{(m-k)j} & m-k \le i \le m \end{cases} \tag{7}$$

be LRSS of *X* variable, and

$$L^{y}_{[i]j} = \begin{cases} Y_{[k+1]j} & i \le k+1 \\ Y_{[i]j} & k+2 \le i \le m-k-1 \quad ; j = 1,2,...,r \\ Y_{[m-k]j} & m-k \le i \le m \end{cases} \tag{8}$$

be the correspondent variable *Y*. Then the two group's estimates will be

$$\begin{cases} \hat{\beta}_{LRSS_{.2g}} = \dfrac{\bar{L}^{y}_{2} - \bar{L}^{y}_{1}}{\bar{L}^{x}_{2} - \bar{L}^{x}_{1}} \\ \hat{\alpha}_{LRSS_{2g}} = \bar{L}^{y}_{LRSS} - \hat{\beta}_{LRSS_{2g}} \bar{L}^{x}_{LRSS} \end{cases} \tag{9}$$

Similarly; the three group's estimates can be formulated as

$$\begin{cases} \hat{\beta}_{LRSS_{.3g}} = \dfrac{\bar{L}^{y}_{3} - \bar{L}^{y}_{1}}{\bar{L}^{x}_{3} - \bar{L}^{x}_{1}} \\ \hat{\alpha}_{LRSS_{3g}} = \bar{L}^{y}_{LRSS} - \hat{\beta}_{LRSS_{3g}} \bar{L}^{x}_{LRSS} \end{cases} \tag{10}$$

5. Empirical Example:

We illustrate the proposed procedure using a real data set from Franklin et al. (1994), that contain of *2600* pairs of numbers (X, Y) where *Y* is the score (in percent) obtained by a student on a standardized calculus test administered at a certain university, and *X* is the number of hours (recorded to the nearest hour) that the students spent studying for this test. Here, the variable of interest is ranked based on exactly measured values. Hence we have actual quantifications. A simple random samples with sizes *n= 60, 80,100* and *120* is chosen from the population. Also, Using set size *m=3, 4, 5* and *6*, with repeated cycles *r= 20* for the chosen RSS and $LRSS_{k=1}$ samples. Two and three group methods are used to estimate the EIV model parameters based on grouping methods and the mean square of residuals for the model are also determined in (11), the results based on these samples are reported in Table 1.

$$MSE = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-2} \qquad (11)$$

			SRS			RSS			LRSS		
	m	**r**	$\hat{\alpha}$	$\hat{\beta}$	**MSE**	$\hat{\alpha}$	$\hat{\beta}$	**MSE**	$\hat{\alpha}$	$\hat{\beta}$	**MSE**
2G	3	20	74.59	-0.18	122.82	44.53	4.03	15.6	48.16	3.74	9.42
	4	20	67.85	-0.11	161.47	45.18	4.09	3.86	45.73	3.94	5.26
	5	20	67.61	0.4	194.12	42.09	4.34	7.93	43.55	3.91	9.72
	6	20	60.18	0.33	113.44	42.69	4.24	13.7	44.96	3.88	9.64
3G	3	20	76.03	-0.3	120.39	44.74	3.95	15.1	48.99	3.5	8.21
	4	20	68.2	0.23	150.23	45.03	4.11	4.34	47.07	3.74	4.75
	5	20	66.33	0.14	201.19	42.42	4.33	8.57	42.56	3.98	11.41
	6	20	59.85	0.28	65.01	42.29	4.29	16.5	44.51	3.89	12.94

Table 1 - The Estimated Slope & Intercept Using Grouping Methods

The results indicated that the estimators based on a ranked data have a smaller MSE comparing to the estimators based on the SRS. There is no effect of the sample size on reducing the value of MSE; Figures (1) – (2) gave a good visual comparison between the three sampling methods. The results indicate that the ranked data are more efficient and more curate estimators than SRS.

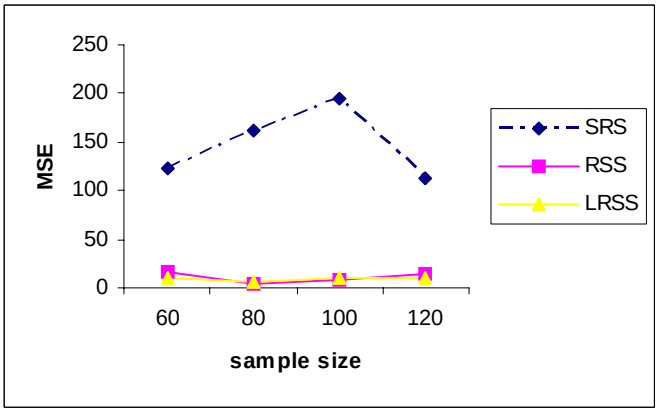

Figure 1 MSE using two-group method

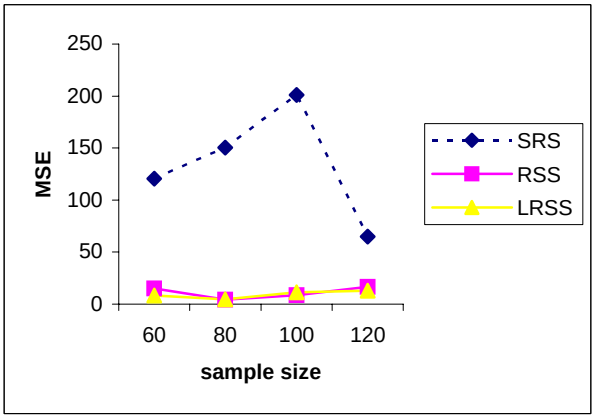

Figure 2 MSE using three-group method

References

Al-Nasser, Amjad D. and Al-Radaideh, Ahmed. 2008. Estimation of Simple Linear Regression Model Using L Ranked Set Sampling. International Journal of Open Problems in Computer Science and Mathematics (IJOPCM). 1(1):18-33.

Al-Nasser, Amjad D. (2007), **L** Ranked Set Sampling: A Generalization Procedure for Robust Visual Sampling. Communication in Statistics - Simulation and Computation. **36**:33-34.

Al-Nasser, Amjad, D., Mohammed E., and Al-Masri A. (2005), On the Use of L-Statistic in Wald-Type Estimators for Fitting a Straight Line When Both Variables are Subject to Error. Pakistan Journal of Statistics. 21(2): 179-194.

Bartlett, M. S. (1949), Fitting Straight Line when Both Variables are Subject to Error. Biometrics. **5:**207-212.

McIntyre, G. A. (1952). A Method for Unbiased Selective Sampling Using Ranked Set Sampling. Australian Journal Agriculture Research. **3:**385-390.

