Estimate the Slope Parameter in Replicated Linear Structural Relationship Model

Amel Saad Alshargawi[1,*], Abdul Ghapor Hussin[2], Ummul Fahri binti Abd Rauf[3]

[1] Department of Statistics, Faculty of Science, Tripoli University
Libya

[2] Centre for Defence Foundation Studies, National   Defence University of Malaysia,
Kuala Lumpur, Malaysia

[3] National   Defence University of Malaysia
Kuala Lumpur, Malaysia

[*]*Corresponding author's email: amelsh [AT] hotmail.com*

ABSTRACT---- *Replication of observation allows consistent estimation of slope parameter of a linear structural model when the ratio of variances is unknown or when some external information about parameters is not available. In this paper, we look at the way a linear structural relationship model work by replicating observations with two different estimation methods of slope parameter and different cases of existence of outliers. The maximum likelihood estimate (MLE) and a new nonparametric robust estimation method are used to estimate the slope parameter in replicated linear structural relationship model (RLSRM). The simulation studies and the application of real data are used to investigate the performance of the estimated parameters.*

Keywords— Maximum likelihood method, A nonparametric method, Trimmed mean, Outlier, Linear structural relationship model with replicated.

1. INTRODUCTION

Some studies of literatures on replicated measurement error model for different models can be found over the years. Maybe the most detailed study being that of Dorff and Gurland (1961) who compared various consistent estimators in terms of their asymptotic standard errors for both in the replicated case and the unreplicated case. Then, Dolby (1976) who derived maximum likelihood estimators of slope parameter for an ultrastructural model, also Chan and Mak (1979) showed that the maximum likelihood estimate of the slope parameter is a root of a fourth degree polynomial. Isogawa (1984) gave study on multivariate linear structural model in replicated case under the assumption of normally distributed measurement error. In 1985, he proposed on the same model the asymptotic covariance matrix of the maximum likelihood estimator when the sample size goes to infinity, and he got maximum likelihood estimator of the coefficient vector. Recently, Shalabh et al. (2009) proposed consistent estimation parameters under replicated ultrastructural model. Singh et al. (2012) proposed consistent estimators satisfying the exact linear restrictions of a replicated ultrastructural measurement error regression model.

Consider two random variables Xi and Yi satisfying the following linear relationship with unknown parameter α and β to be estimated

$$Y_i = \alpha + \beta X_i \tag{1}$$

For each (X_i, Y_i), r repeated observations x_{ij} and y_{ij}, $j=1,2,...r$. The replicated linear structural relationship model is given with errors as

$$x_{ij} = X_i + \delta_{ij} \qquad y_{ij} = Y_i + \varepsilon_{ij}, \tag{2}$$

when $X_i \sim N(0, \sigma_X^2)$, and with errors $\delta_{ij} \sim N(0, \sigma_\delta^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, as well as $\forall_{i \neq j}$, $cov(\delta_{il}, \delta_{jm}) = cov(\varepsilon_{il}, \varepsilon_{jm}) = cov(\delta_{il}, \varepsilon_{jm}) = 0$.

The model in equation (1) known as a replicated linear structural relationship model with "balanced replicated", which means the number of elements in each group is similar. In this paper we display the estimation of the slope parameter for RLSRM using the MLE and propose a new method. The organization of this article is as follows; Section 1 begins with the introduction. In Section 2 elaborates on the two estimation methods for slope parameter of RLSRM. A simulation study is carry out in Section 3 to compare the performance of both methods in the presence of outliers. Next, we applied the two methods to real life data to estimated the slope parameter. Lastly, conclusion is presented in Section 5.

2. ESTIMATION METHOD FOR SLOPE PARAMETER IN RLSRM

2.1. Maximum Likelihood Estimation Method

Maximum likelihood estimation (MLE) method is commonly used in linear structural relationship model. Based on Chan and Mak (1979), when the ratio of error variance $\left(\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\delta^2}\right)$ is unknown, the MLE of the slope by assuming normality for replicated model is

$$\beta^* = \frac{\left(s_y^2 - \frac{w_y^2}{w_x^2}s_x^2\right) + \sqrt{\left(s_y^2 - \frac{w_y^2}{w_x^2}s_x^2\right)^2 + 4\frac{w_y^2}{w_x^2}s_{xy}^2}}{2s_{xy}}$$

When

$$s_x^2 = \frac{\sum_{i=1}^n (\bar{x}_i - \bar{x}_{..})^2}{n} \qquad s_y^2 = \frac{\sum_{i=1}^n (\bar{y}_i - \bar{y}_{..})^2}{n} \qquad s_{xy} = \frac{\sum_{i=1}^n (\bar{x}_i - \bar{x}_{..})(\bar{y}_i - \bar{y}_{..})}{n}$$

$$w_x^2 = \frac{\sum\sum(x_{ij} - \bar{x}_{i.})^2}{nr} \qquad , \qquad w_y^2 = \frac{\sum\sum(y_{ij} - \bar{y}_{i.})^2}{nr}$$

2.2. The Proposed Nonparametric Robust Estimation Method

We propose a new nonparametric robust estimation method whereby normality assumption is not required to estimate slope parameter. In this paper, we have used trimmed mean of all possible paired slopes instead of the average slopes that used by Al-Nasser and Ebrahem (2005), then comparison to the well-known MLE method. The method consider the observed pairs (x_{ij}, y_{ij})'s, $i = 1,2\dots n; j = 1,2\dots r,$ are put in one group for each variable then ordered according to the magnitude of the x value, then we distribute the observations into several groups to calculate all possible paired slopes. The following are the steps involved in our proposed method

1. The total observations are distributed into m subsamples each having elements r whereas $n = m \times r$ provided that $m \leqq r$.

2. The number of all possible paired slopes is calculate by using the form
$$b(K)_{ij} = \frac{y_{[j]} - y_{[i]}}{x_{(j)} - x_{(i)}}, i = 1,2\dots j - 1; j = 2,3\dots r, \text{and } K = 1,2,\dots m$$

3. Find the median of these slopes, as follows
$$\beta = medain(b(K)_{ij})i = 1,2\dots j - 1; j = 2,3\dots r, and\ K = 1,2,\dots m$$

4. Calculate trimmed mean of all possible paired slopes as follows
$$\beta = trimmed\ mean(b(K)_{ij})$$

The steps described in step 1 to step 3 for estimating the slope parameter is based on the nonparametric estimation method as introduced by Al-Nasser and Ebrahem (2005). To improve the estimation, step 3 is replaced with trimmed mean in step 4. We used trimmed mean because it is based on removing outliers with fixed proportion of smallest and

largest values in the data set compare with median. The 20% trimmed mean was chosen in this study based on Wilcox (2003), when p=20% it can make a reasonable default to make balance between achieving a small standard error and controlling the probability of a type 1 error (Wilcox, 2003).

3. SIMULATION STUDY

A simulation study is carried out to test the performance of estimator slope for different levels of contamination by exchange the original observation by contaminated observations. In the beginning we simulating observations from the following model, where $X_i \sim N(15,4)$ and both errors by $\delta_{ij}, \varepsilon_{ij} \sim N(0,0.1)$ then generate the observation from

$$Y_i = 1 + X_i \ , x_{ij} = X_i + \delta_{ij} \ and \ y_{ij} = Y_i + \varepsilon_{ij}$$

Without loss of generality the replicated numbers of the observations are chosen with 6, 10, 10, 12 and 15 which are paired with different sample sizes of 30, 50, 100, 132 and 150 respectively from the sampling distribution with 1000 iteration. Table 1 shows the general sample and values of size sample and number of replicates. To determine the performance of the estimates, it was measured using mean square error (MSE) and the standard deviation (SD) in normal distribution along with the non-normal error terms as Beta distribution in order to investigate the robustness of our proposed method. The MSE are defined by

$$MSE = \Sigma(\widehat{w_i} - w)^2 / 1000$$

Where $\bar{\bar{w}} = \frac{\Sigma \hat{w}}{1000}$

The simulation results are highlighted in Table 2 and Table 3 for normal case, tables 4 to 9 for Beta distribution. From Table 2 and Table 3, we observe a marked difference when the data set gets contaminated, by looking at Table 2, the MSE of MLE gives marginally better result than our proposed method when no outliers present in the data, and can say the results in general of the MSE of the proposed method is somewhat similar to that of the MLE when outlier under 10% of contamination. MLE breaks down with an increase level of contamination to 20%. The same results can see it on SD of the slope estimator.

Table 4 to Table 9 display the results for MSE and SD of the slope estimator for the error terms with that of non-normal symmetric case with Beta distribution (2,2), Beta distribution (3,7) and Beta distribution (7,3) on respectively with different levels of contamination.

As we see from Table 4 the MSE of MLE method works best in case where the data are free from outliers or if there is one single outlier, but it breaks down easily when there are about 10% and above contamination in the data. Table 5 gives more satisfactory results of the SD for our proposed method than MLE method. Tables 6 to 9 display the results of MSE and SD of the slope estimator where the error terms skewed to the right with Beta distribution (3, 7) and skewed to the left with Beta distribution (7, 3). Clearly, the results show that MSE values of two estimation methods somewhat similar when no outlier exist in the data. But when outliers are present in the data, a marked difference of the MSE and SD of the slope is observed and the breakdown completely of MLE method is noticed.

4. REAL DATA EXAMPLES

In this part, we applied our proposed nonparametric method to real life data and compare it with the MLE method to estimated the slope parameter. Example 1, we considered a real data set from a study taken by Bland and Altman (1999). The data are taken from the measurements of plasma volume expressed as a percentage of 60 and 90 subjects using two alternative sets of normal values with Nadler (x) as dependent variable and Hurley (y) as independent variable. In this example, unreplicated data are transferred into replicated data by setting 60 subjects to 6 groups and 90 subjects to 9 groups where each group contains 10 elements.

Example 2, the data used are a subset of systolic blood pressure data from a study in which simultaneous measurements were made by two experienced observers (denoted J and R) on an individual patient. We consider J as x_{ij} and R as y_{ij}. For every variables, three observations in quick succession are made on twenty subjects. The data is available from Dr E O'Brien, and a detailed description of the data is given by also Bland and Altman (1999). We extract the estimation from the original data insert a few outliers to create different situations namely single outlier, 10%, 20%. The real data results are highlighted in Table 10 and table 11.

Based on Table 10 and Table 11, it clearly show that both methods show somewhat similar results in terms of the slope estimator when there are no outliers in the data set. However, as the percentage of outlier increases, the slope estimator for MLE method also increases when compared to the proposed method. That means our proposed method was not affected by the outliers as the MLE.

5. CONCLUSION

In this article, a new nonparametric method is proposed to estimate slope parameter of a replicated linear structural relationship model RLSRM and compare it with MLE method. From the simulation study show that the proposed method has better accuracy in comparison with the MLE in existence of outliers.

From the real data set, we can conclude that when there is no outlier in the data set, the MLE showed somewhat consistent results in terms of slope parameter with our proposed method. However, as the percentage of contamination increases, the MLE method is shown to break down compare by our proposed method which performs very well in this case. Thus the proposed estimator is considered as a good alternative to the MLE method in RLSRM.

6. ACKNOWLEDGEMENT

7. REFERENCES

- Al-Nasser, A.D. & Ebrahem, M.A.H. (2005). A new nonparametric method for estimating the slope of simple linear measurement error model in the presence of outliers. Pakistan Journal of Statistics, 21(3), 265-274.

- Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. Statistical Methods in Medical Research, 8(2), 135–160.

- Chan, L. K., & Mak, T. K. (1979). On the maximum likelihood estimation of a linear structural relationship when the intercept is known. Journal of Multivariate Analysis, 9, 304–313.

- Dolby, G. R. (1976). The ultrastructural relation: a synthesis of the functional and structural relations. Biometrika, 63(1), 39–50.

- Dorff, M., & Gurland, J. (1961). Estimation of the parameters of a linear functional relation. Journal of the Royal Statistical Societ,. 23(1), 160–170.

- Isogawa, Y. (1984). Large sample theory for a multivariate structural relationship with replication. Communications in Statistics - Theory and Methods, 13(17), 2111–2118.

- Isogawa, Y. (1985). A note on a linear with structural replication. Journal Japan Statistical Society, 15(1), 71–74.

- Kim, M. G. (2000). Outliers and influential observations in the structural errors-in-variables model. Journal of Applied Statistics, 27(4), 451–460.

- Shalabh, Paudel, C. M., & Kumar, N. (2009). Consistent estimation of regression parameters under replicated ultrastructural model with non-normal errors. Journal of Statistical Computation and Simulation, 79(3), 251–274.

- Singh, S., Jain, K., & Sharma, S. (2012). Replicated measurement error model under exact linear restrictions. Stat Papers. Springer 55(2), 253–274.

- Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: measures of central tendency. Psychological Methods, 8(3), 254–274.

Table 1: Values of Sample Size and the Number of Replicates

N	Sample size	Replicates No.
30	5	6
50	5	10
100	10	10
132	11	12
150	10	15

Table 2: MSE of the Slope Parameter: Normal-Case

Contamination	Method	Simple Size				
		30	50	100	132	150
No outlier	MLE	4.91E-03	2.37E-03	7.51E-04	4.80E-04	4.75E-04
	Proposed method	1.42E-02	1.40E-02	3.85E-03	3.21E-03	3.99E-03
Single Outlier	MLE	1.90E-02	7.62E-03	1.60E-03	1.01E-03	9.70E-04
	Proposed method	1.80E-02	1.48E-02	3.43E-03	3.12E-03	4.47E-03
10%	MLE	5.41E-02	3.02E-02	8.83E-03	6.65E-03	6.21E-03
	Proposed method	2.07E-02	1.84E-02	4.28E-03	3.81E-03	5.07E-03
20%	MLE	1.11E-01	7.66E-02	1.98E-02	1.32E-02	1.15E-02
	Proposed method	3.92E-02	3.10E-02	8.28E-03	6.60E-03	6.93E-03

Table 3: SD of the Slope Parameter: Normal-Case

Contamination	Method	Simple Size				
		30	50	100	132	150
No outlier	MLE	7.00E-02	4.86E-02	2.74E-02	2.19E-02	2.18E-02
	Proposed method	9.01E-02	8.12E-02	4.22E-02	3.40E-02	3.62E-02
Single Outlier	MLE	1.38E-01	8.73E-02	4.00E-02	3.17E-02	3.12E-02
	Proposed method	1.03E-01	8.71E-02	3.93E-02	3.32E-02	3.94E-02
10%	MLE	2.32E-01	1.74E-01	9.37E-02	8.16E-02	7.89E-02
	Proposed method	1.17E-01	9.79E-02	4.70E-02	4.01E-02	4.48E-02
20%	MLE	3.33E-01	2.77E-01	1.41E-01	1.15E-01	1.07E-01
	Proposed method	1.75E-01	1.41E-01	7.34E-02	6.27E-02	5.88E-02

Table 4: MSE of the slope: Symmetric-Case, Beta(2, 2)

Contamination	Method	Simple Size				
		30	**50**	**100**	**132**	**150**
No outlier	MLE	3.00E-02	1.76E-02	9.58E-03	7.63E-03	6.55E-03
	Proposed method	7.73E-02	7.85E-02	1.00E-01	1.03E-01	9.99E-02
Single Outlier	MLE	4.76E-01	1.83E-01	7.28E-02	4.33E-02	3.40E-02
	Proposed method	8.82E-02	7.95E-02	1.02E-01	1.03E-01	1.00E-01
10%	MLE	6.41E-01	4.30E-01	2.55E-01	2.34E-01	1.72E-01
	Proposed method	9.11E-02	8.53E-02	1.01E-01	1.04E-01	1.16E-01
20%	MLE	0.16E+01	8.23E-01	4.34E-01	4.34E-01	3.40E-01
	Proposed method	2.15E-01	1.34E-01	1.20E-01	1.16E-01	1.02E-01

Table 5: SD of the slope: Symmetric-Case, Beta(2, 2)

Contamination	Method	Simple Size				
		30	**50**	**100**	**132**	**150**
No outlier	MLE	1.73E-01	1.33E-01	9.77E-02	8.72E-01	8.10E-02
	Proposed method	1.34E-01	1.01E-01	7.14E-02	6.28E-02	5.91E-02
Single Outlier	MLE	6.90E-01	4.28E-01	2.70E-01	2.08E-01	1.85E-01
	Proposed method	1.63E-01	1.13E-01	7.67E-02	6.71E-02	5.99E-02
10%	MLE	8.01E-01	6.56E-01	4.96E-01	4.83E-01	4.15E-01
	Proposed method	1.99E-01	1.52E-01	1.08E-01	9.94E-02	8.81E-02
20%	MLE	0.13E+01	9.06E-01	6.58E-01	6.58E-01	5.83E-01
	Proposed method	3.98E-01	2.85E-01	2.01E-01	1.74E-01	9.75E-02

Table 6: MSE of the Slope: Right Skewed-Case, Beta(3, 7)

Contamination	Method	Simple Size				
		30	50	100	132	150
No outlier	MLE	1.01E-02	5.66E-03	3.55E-03	2.88E-03	2.48E-03
	Proposed method	1.89E-02	1.67E-02	2.15E-02	2.21E-02	2.24E-02
Single Outlier	MLE	4.67E-01	1.64E-01	6.76E-02	3.97E-02	3.01E-02
	Proposed method	2.36E-02	1.90E-02	2.29E-02	2.26E-02	2.16E-02
10%	MLE	5.80E-01	4.05E-01	2.70E-01	2.21E-01	1.78E-01
	Proposed method	2.78E-02	2.31E-02	2.49E-02	2.41E-02	2.88E-02
20%	MLE	0.18E+01	7.83E-01	4.90E-01	4.19E-01	3.19E-01
	Proposed method	1.30E-01	6.47E-02	4.16E-02	3.86E-02	2.76E-02

Table 7: SD of the Slope: Right Skewed-Case, Beta(3, 7)

Contamination	Method	Simple Size				
		30	50	100	132	150
No outlier	MLE	1.00E-01	7.52E-02	5.96E-02	5.36E-02	4.98E-02
	Proposed method	9.59E-02	6.68E-02	5.28E-02	4.74E-02	4.56E-02
Single Outlier	MLE	6.84E-01	4.06E-01	2.60E-01	1.99E-01	1.74E-01
	Proposed method	1.17E-01	8.32E-02	6.03E-02	4.86E-02	4.55E-02
10%	MLE	7.62E-01	6.37E-01	5.20E-01	4.70E-01	4.22E-01
	Proposed method	1.40E-01	1.10E-01	8.40E-02	7.31E-02	7.05E-02
20%	MLE	0.13E+01	8.85E-01	7.00E-01	6.47E-01	5.65E-01
	Proposed method	3.49E-01	2.33E-01	1.64E-01	1.41E-01	7.27E-02

Table 8: MSE of the Slope: Right Skewed-Case, Beta(7, 3)

Contamination	Method	Simple Size				
		30	50	100	132	150
No outlier	MLE	9.90E-03	5.65E-03	3.56E-03	2.91E-03	2.53E-03
	Proposed method	1.82E-02	1.67E-02	2.19E-02	2.29E-02	2.18E-02
Single Outlier	MLE	4.47E-01	1.64E-01	6.72E-02	3.97E-02	2.99E-02
	Proposed method	2.40E-02	1.84E-02	2.24E-02	2.30E-02	2.22E-02
10%	MLE	5.64E-01	4.08E-01	2.74E-01	2.25E-01	1.74E-01
	Proposed method	2.91E-02	2.40E-02	2.44E-02	2.62E-02	2.64E-02
20%	MLE	0.18E+01	7.90E-01	4.86E-01	4.23E-01	3.19E-01
	Proposed method	1.40E-01	6.18E-02	4.17E-02	3.90E-02	2.32E-02

Table 9: SD of the slope: right skewed-case, Beta(7, 3)

Contamination	Method	Simple Size				
		30	50	100	132	150
No outlier	MLE	9.94E-02	7.52E-02	5.97E-02	5.39E-02	5.02E-02
	Proposed method	9.48E-02	7.02E-02	5.47E-02	4.74E-02	4.26E-02
Single Outlier	MLE	6.69E-01	4.05E-01	2.59E-01	1.99E-01	1.73E-01
	Proposed method	1.16E-01	7.94E-02	5.90E-02	5.17E-02	4.57E-02
10%	MLE	7.51E-01	6.39E-01	5.24E-01	4.75E-01	4.18E-01
	Proposed method	1.39E-01	1.11E-01	8.36E-02	7.58E-02	6.86E-02
20%	MLE	0.13E+01	8.89E-01	6.97E-01	6.50E-01	5.65E-01
	Proposed method	3.62E-01	2.29E-01	1.61E-01	1.46E-01	7.31E-02

Table 10: Estimated the Slope Parameter using two Different Methods
for Plasma Volume Data(Replicated Data)

Contamination	Method	Slope (n=6, r=10) N=60	Slope (n=9, r=10) N=90
No outlier	MLE	0.9106	0.9177
	Proposed method	0.9113	0.9190
Single Outlier	MLE	1.0415	0.9914
	Proposed method	0.9288	0.9277
10%	MLE	0.8149	1.0557
	Proposed method	0.9173	0.9442
20%	MLE	0.7455	1.1765
	Proposed method	0.9814	0.9951

Table 11: Estimated the Slope Parameter using two Different Methods
from Systolic Blood Pressure Data

Contamination	Method	Slope
No outlier	MLE	0.9694
	Proposed method	0.9646
Single Outlier	MLE	1.0228
	Proposed method	0.9686
10%	MLE	1.0199
	Proposed method	0.9591
20%	MLE	1.0542
	Proposed method	0.9625