ZULFAQAR International Journal of Defence Science, Engineering & Technology

Journal homepage: www.zulfaqar.upnm.edu.my

Nonparametric Robust Estimator for Slope Parameter in Linear Structural Relationship Model

Amel Saad Alshargawi[a,*], Abdul Ghapor Hussin[b], Ummul Fahri binti Abd Rauf[b]

[a] *Department of Statistics, Faculty of Science, Tripoli University*
[b] *Centre for Defence Foundation Studies, National Defence University of Malaysia, Kuala Lumpur*

*Corresponding author: *amelsh@hotmail.com*

ARTICLE INFO

Article history:
Received
21-11-2018
Received in revised
15-12-2018
Accepted
30-12-2018
Available online
31-12-2018

Keywords:
Linear structural relationship model,
Maximum likelihood method,
Outliers,
Trimmed mean

e-ISSN:
Type: Article

ABSTRACT

In this study, the slope parameter of linear structural relationship model is determined by using the proposed robust nonparametric method based on trimmed mean. This method is an upgrade to the nonparametric method that was introduced by Al-Nasser et al. (2005) by employing trimmed mean for all likely paired slopes rather than median slopes. Simulation study and real data were used to compare the proposed method's performance versus the traditional maximum likelihood method. In the simulation study, based on both methods' mean square error, it was inferred that the MLE method break down due to the presence of outliers even though its elaborate was not affected when there was no outlier in the data set. Based on the real life examples, it can be concluded that the performance of our proposed method was better in determining the slope parameter and thus provides a good alternative to MLE method when outliers are present.

Introduction

A family in the errors-in-variables model (EIVM) is the linear structural relationship model (LSRM). There are numerous studies that focused on LSRM; however, for just a few special cases, the formulas are available regarding these estimators' accuracy for the model (Bolfarine & Cordani; 1993, Patefield;1985, Wong; 1989). An article by Hood et al. (1999) provides a brief description regarding the estimates applicable for the LSRM parameters. Recently, some studies have applied on LSRM as Mamun et al. (2013) . In this study, the use of trimmed mean in a nonparametric method is proposed rather than the median when there are outliers in the data set.

We get the following linear relationship form by assuming two random variables X and Y:

$$X = \alpha + \beta Y, \tag{1}$$

when measuring X, Y without the error which cannot be observed directly, i.e. their measurements are subject to error. Suppose that for each i, x_i and y_i are considered instead of X_i and Y_i respectively and (i=1,2,.....,n), the two respective errors represented by δ_i and ε_i to measure X_i and Y_i as:

$$\left.\begin{aligned} x_i &= X_i + \delta_i \\ y_i &= Y_i + \varepsilon_i \end{aligned}\right\}, \tag{2}$$

where normally distributed parameters δ_i and ε_i with zero mean and variance σ_δ^2 and σ_ε^2, respectively, and when covariances are zero for variables and errors, then the model in equation (1) is written as:

$$y_i = \alpha + \beta x_i + (\varepsilon_i - \beta\delta_i), \tag{3}$$

which suggests that both recognisable x_i and y_i are linked to the error term $(\varepsilon_i - \beta\delta_i)$, and this error term is dependent on the slope parameter β. Thus, we get six unknown parameters $\beta, \alpha, \mu, \sigma_x^2, \sigma_\delta^2$ and σ_ε^2, which require estimation. To achieve estimation, additional assumption is required to get unique and consistent solutions regarding the parameters associated with the model in equation (1). Some of the authors have employed the maximum likelihood method to determine those parameters that definitely need normality assumption as Lakshminarayanan and Gunst (1984). Hood et al. (1999) put forward this concept to address the normal equations related to the model as well as to determine the model's parameters. Thus, it is imperative as well as sufficient to consider one amongst the variances or to know the ratio between the two variances. In this study, a new robust estimator has been put forward for the slope by employing the nonparametric method, by refining the idea of Al-Nasser et al. (2005), by making use of the trimmed mean for all likely combination for the paired slopes rather than the median. It should be noted that nonparametric methods are developed without relying on the assumption of normality of error distributions. Furthermore, the trimmed mean was employed as it clearly did not discard numerous observations such as in the case of the median that would discard nearly everything. The paper is categorised as follows: In Section 2, the slope parameter estimation is reviewed for LSRM employing the MLE, and we put forward a new method to determine the slope. Next, in Section 3, the performance is compared via a simulation study performed for both methods. In section 4, a real data set is employed to examine the performance of both methods estimating the slope.

Estimation of Slope Parameter for LSRM

i. Maximum likelihood estimation method

The most commonly employed method in LSRM is the maximum likelihood estimation (MLE) method. Hood et al. (1999) employed various assumptions to provide a brief description regarding the estimation process for the parameters of model in equation (1). One amongst these assumptions was regarding ratio of error variance $\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\delta^2}$ that is considered to be known. Under the normality assumption, the MLE for the slope parameter can be defined as follows:

$$\hat{\beta} = \frac{\left(s_y^2 - \lambda s_x^2\right) + \sqrt{(s_y^2 - \lambda s_x^2)^2 + 4\lambda s_{xy}^2}}{2s_{xy}} \tag{4}$$

where $s_x^2 = {\sum_i^n (x_i - \bar{x})^2}/{n}$, $s_y^2 = {\sum_i^n (y_i - \bar{y})^2}/{n}$ and $s_{xy} = {\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}/{n}$.

ii. Proposed method

In this study, an estimator was put forward for the linear structural relationship model's slope according to a nonparametric method, which does not need normality assumption. Ghapor et al.

(2015) showed that the maximum likelihood method appears to be quite unreliable when there are outliers and the break-down of mean square error is effortless in the study employing the linear functional model. The method we put forward is to develop the methods by Al-Nasser et al. (2005) by employing trimmed mean applicable to all possible paired slopes instead of median slopes.

The observations (x_i, y_i) were coordinated in an ascending manner in accordance to the x values, i.e. $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$. The related values associated with y that are not in ascending order can be taken as $y_{[1]}, y_{[2]}, \ldots y_{[n]}$, to consequently get the pairs $(x_{(i)}, y_{[i]})$. The i values that are in a bracket, ( ) indicates that they are arranged in ascending order, while i values that are in a square bracket, [ ] indicates that they are not arranged in ascending order. The following steps are involved in the process:

1. The classification of the total observations is done into m subsamples, where each has the r elements in a manner that $n = m \times r$ when $m \leq r$. Thus, the coordination of the samples is done in the following manner:

$$\begin{matrix} (x_{(1)}, y_{[1]}) & (x_{(2)}, y_{[2]}) & \ldots\ldots & (x_{(r)}, y_{[r]}) \\ (x_{(r+1)}, y_{[r+1]}) & (x_{(r+2)}, y_{[r+2]}) & \ldots\ldots & (x_{(2r)}, y_{[2r]}) \\ \vdots & \vdots & \vdots & \vdots \\ (x_{((m-1)\times(r+1))}, y_{[(m-1)\times(r+1)]}) & \ldots\ldots & \ldots\ldots & (x_{(mr)}, y_{[mr]}) \end{matrix}$$

2. Calculate the number of all possible paired slopes using the form

$$b(K)_{ij} = \frac{y_{[j]} - y_{[i]}}{x_{(j)} - x_{(i)}}, i = 1,2 \ldots j-1; j = 2,3 \ldots . r, and\ K = 1,2,..m$$

3. Find the median of these slopes as follows

$$\beta = medain(b(K)_{ij}) i = 1,2 \ldots j-1; j = 2,3 \ldots . r, and\ K = 1,2,..m$$

4. Calculate trimmed mean of all possible paired slopes as

Trimmed mean=mean $(b(K)_{ij}$, trim=20%)

The steps defined in steps 1 to 3 for determining the slope parameter are according to the nonparametric estimation method as put forward by Al-Nasser et al. (2005). The trimmed mean in step 4 replaces by step 3 to get the proposed estimator.

The benefits of employing trimmed mean are: first, it is known to be a popular robust centre indicator as median. Secondly, unlike median, its calculation involves discarding the highest and lowest p% of the values, and then finally computing the remaining data's mean. Thus, when p=50%, it can be considered equivalent to median, and when p=0%, it is equivalent to mean Kitchenham et al.(2016). This indicates that the trimmed mean is still the best even when the median's means squares are the least. In this paper, we selected p=20%, which seems to be a reasonable proportion for striking a balance between controlling a type 1 error's probability and achieving a small standard error Wilcox & Keselman (2003).

Simulation Study

In this section, a simulation study was performed to compare the performance of the proposed method with the MLE method in the presence of outliers. We simulated the initial observations from the LSRM

$$Y_i = 1 + X_i\ , x_i = X_i + \delta_i\ and\ \ y_i = Y_i + \varepsilon_i, \qquad (5)$$

where $X_i \sim N(15,4)$ and both errors by $\delta_i, \varepsilon_i \sim N(0, 0.1)$. Then, exchanging of the original observations with the contaminated observations, We generated the contaminated data points for example at point *d* for the variable y, through replace y with the equation $y_d = 1 + X_d + \varepsilon_d$, where $\varepsilon_d \sim N(0,25)$. We generate samples of size 50, 100 and 150 from the sampling distribution with 10000 trials. The non-normal error terms were considered for studying the proposed method's robustness. Moreover, three different Beta distributions were considered for generating error terms: right skewed Beta distribution (3, 7), left skewed Beta distribution (7,3) and symmetric Beta distribution (2,2) with the relationships that were employed above. Both methods' performance is evaluated by mean square error (MSE) of the slope parameter. The following form defines MSE:

$$MSE = \sum(\widehat{w_i} - w)^2 /_S,$$

where $\bar{\widehat{w}} = \frac{\sum \widehat{w}}{10\,000}$, w is a generic term for the parameters and s is the number of trials.

For the nonparametric method, Table 1 shows the values of *r* and *m* for each sample size, in which *m* represents the number of groups and *r* signifies the number of elements for each group.

Table 1 Shows Values of *m* and *r*

Sample Size n	m	r
50	5	10
100	10	10
150	10	15

The Tables (2, 3, 4, and 5) show simulation results for MSE with regards to the slope estimator by employing both the proposed and MLE methods.

Table 2 Shows MSE Values of the Slope: Normal-case

Contamination	Method	n=50	n=100	n=150
No outlier	MLE	2.96E-04	5.30E-04	3.45E-04
	Proposed method	1.88E-03	1.20E-03	9.82E-04
Single Outlier	MLE	2.07E-02	4.87E-03	2.31E-03
	Proposed method	2.05E-03	1.22E-03	1.00E-03
10%	MLE	2.71E-01	1.84E-01	1.67E-01
	Proposed method	3.19E-03	1.75E-03	1.33E-03
20%	MLE	1.08E+00	8.23E-01	7.57E-01
	Proposed method	7.03E-03	3.40E-03	2.44E-03

Table 3 Shows MSE Values of the Slope: Beta (7,3)

Contamination	Method	n=50	n=100	n=150
No outlier	MLE	1.01E-02	5.05E-03	3.36E-03
	Proposed method	3.01E-02	2.91E-02	2.95E-02
Single Outlier	MLE	1.74E+05	2.12E+04	3.62E+01
	Proposed method	3.23E-02	2.96E-02	2.97E-02
10%	MLE	3.09E+07	2.14E+07	4.80E+05
	Proposed method	4.89E-02	3.69E-02	3.39E-02
20%	MLE	1.27E+07	1.67E+07	7. 63E+10
	Proposed method	1.53E-01	7.06E-02	5.07E-02

Table 4 Shows MSE Values of the Slope: Beta (3,7)

Contamination	Method	n=50	n=100	n=150
No outlier	MLE	9.91E-03	4.99E-03	3.35E-03
	Proposed method	3.00E-02	2.94E-02	2.94E-02
Single Outlier	MLE	2.77E+06	3.58E+03	2.60E+02
	Proposed method	3.33E-02	3.06E-02	2.98E-02
10%	MLE	7.99E+05	1.95E+07	3.06E+07
	Proposed method	5.51E-02	4.29E-02	4.05E-02
20%	MLE	1.07E+07	1.15E+07	3.40E+07
	Proposed method	1.64E-01	8.56E-02	6.78E-02

Table 5 Shows MSE Values of the Slope: Beta (2,2)

Contamination	Method	n=50	n=100	n=150
No outlier	MLE	3.05E-02	1.47E-02	9.69E-03
	Proposed method	1.26E-01	1.29E-01	1.30E-01
Single Outlier	MLE	3.55E+04	2.94E+03	4.66E+02
	Proposed method	1.33E-01	1.31E-01	1.31E-01
10%	MLE	9.33E+05	1.35E+07	6.26E+06
	Proposed method	1.65E-01	1.52E-01	1.48E-01
20%	MLE	2.97E+06	4.75E+06	6.11E+06
	Proposed method	2.55E-01	1.82E-01	1.65E-01

The above results clearly show the benefits of both methods. Table 2 shows MLE demonstrated a marked difference in MSE values along with outliers in the data set versus the proposed method. According to MSE for the slope estimator, MLE was seen to break down when there is a small sample size as well as with increased contamination level. Meanwhile, smaller MSE values are generated with the proposed method when outliers are present. In general, no significant changes were found between both methods when outlier was not present in data. The curve in Figure 1 gives a clear view of MSE of the slope by using both MLE and the proposed method with different levels of contaminations in normal distribution.

The results in Tables 3 to 5 show the simulation results of the MSE values of slope estimator in non-normal distribution using MLE breaks down quickly and become huge when the data are contaminated. MSE values using the proposed method are not affected by outliers regardless of the percentage of contamination or sample size, thus the proposed method as robust estimation gives a consistently smaller values of the MSE than MLE.

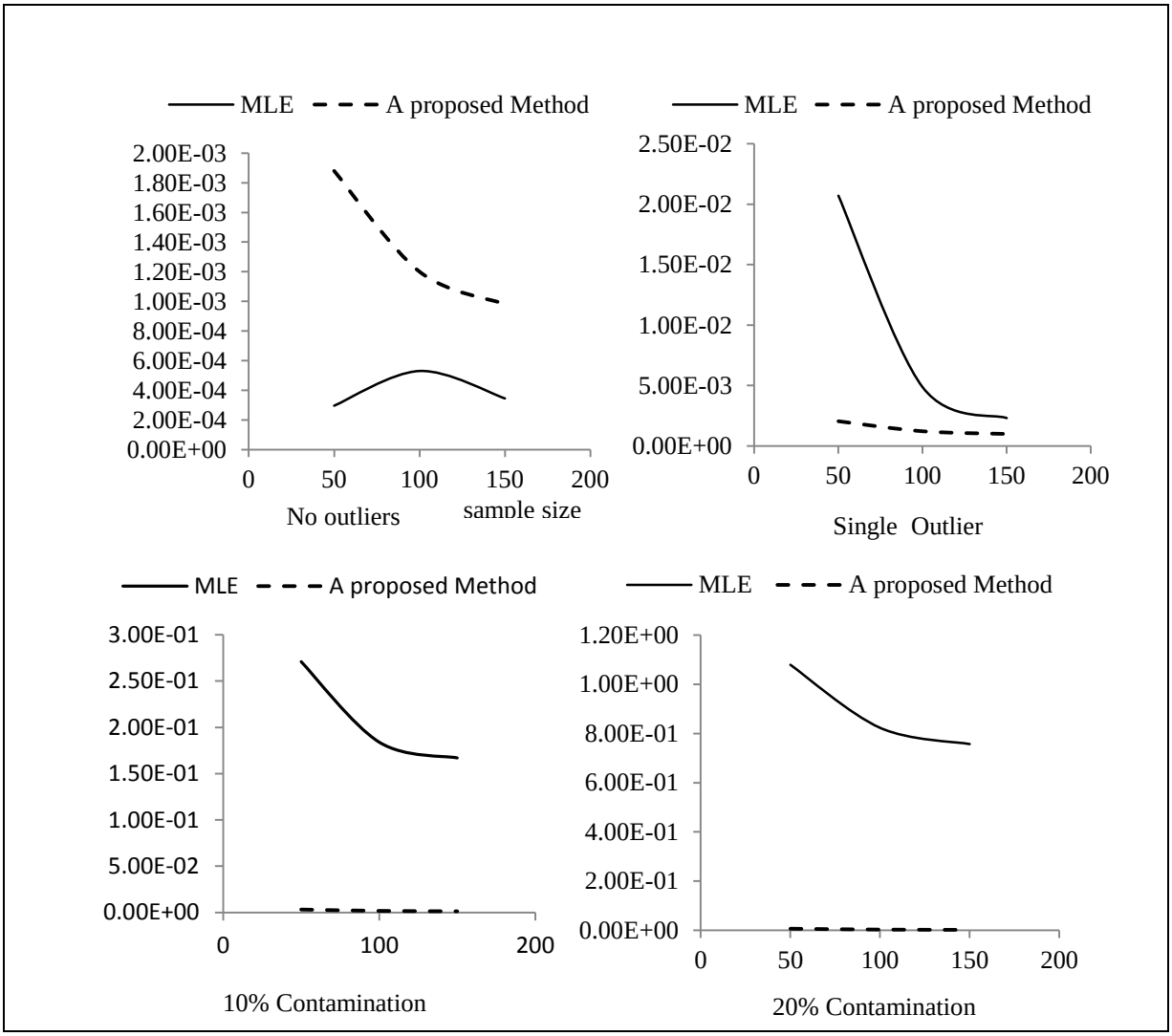

Figure 1 Shows MSE of the slope: normal-case with different level of contamination

Example of Real Data

In this section, we apply the MLE method and proposed method for real data application. Real data was obtained from Hand et al. (1994). This data have 50 results regarding crushed blast-furnace slag as determined by two different ways. The data were used to evaluate both methods' performance. Data sets were segmented into 5 groups for applying the proposed method of trimmed mean in determining the slope parameter, in which every group has 10 elements. To create different situations, Kim (2000) and Imon & Hadi (2008) inserted outliers by getting rid of few observations and substituting them each with the outliers' observation, in which the selected cases were single outlier, 10% and 20%, (See Table 6).

Table 6 Shows Estimated Slope Using Different Methods from Iron in Slag Data

Contamination	Method	Slope Estimator	Percentage change of slope estimator (%)
No outlier	MLE	0.9339	19.56
	Proposed method	0.7512	
Single outlier	MLE	1.1234	30.04
	Proposed method	0.7859	
10%	MLE	4.3371	77.96
	Proposed method	0.9560	
20%	MLE	12.1066	89.58
	Proposed method	1.5618	

As presented in Table 6, a small difference was observed between both methods when no outliers were present. In contrast, a huge difference is observed when outliers are present in the data set. That is to say, the MLE of slope parameter breaks down quickly with the increase in the percentage of outliers as compared to the proposed method, whereas proposed estimator is not much affected by the existence of outliers even though trim was only 20% from data.

conclusion

The results are concluded in this section based on the simulation study as well as real data example. For the slope parameter in LSRM, the MLE estimated value is provided for the case in which variance ratio is one. The MLE method gives a slightly better result when compared with our proposed method in the absence of outliers in the data. Otherwise, in the case of normal distribution and other distributions for the error terms, as Beta distributions, there is a complete break-down of the MLE method in the presence of outliers. The performance of the proposed method was the best for all situations when there were outliers with various distributions. While alternatively, to get the desired results, the discarded observations' percentage needs to be controlled.

Acknowledgement

The authors would like to express our appreciation and thanks to the editors and reviewers for their valuable comments and feedback on this paper.

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.

Bolfarine, H. & Cordani, L.K. (1993). Estimation of a structural linear regression model with known reliability ratio. *Ann. Ins. of Statist. Math.,* 45, 531-540.

Ghapor, A. A., Zubairi, Y. Z., Mamun, A.S.M.A. &. Imon A.H.M.R (2015). A robust nonparametric slope estimation in the linear functional relationship model. *Pakistan Journal of Statistics*, Vol. 31(3), 339-350.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J., & Ostrowski, E. (1994*). A handbook of small data sets*. Chapman and Hall, London.

Hood, K., Barry A.J. and Terence C. (1999). Asymptotic information and variance-covariance matrices for the linear structural model. *J. Roy. Statist. Soc., D,* 48, 477-493.

Imon, A.H.M.R. & Hadi, A.S. (2008). Identification of multiple outliers in logistic regression. *Commun. Stat. Theor. Methods*, *37*(11), 1697-1709.

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

Kitchenham, B., Madeyski, L., Budgen, D., Keung, J., Brereton, P., Charters, S., Gibbs, S., & Pohthong, A. (2016). Robust Statistical Methods for Empirical Software Engineering. Empir Software Eng, 22(2), 579–630.

Lakshminarayanan, M. Y., & Gunst, R. F. (1984). Estimation of parameters in linear structural relationships: Sensitivity to the choice of the ratio of error variances. Biometrika, 71(3), 569–573.

Mamun, A. S. M. A. (2014). Estimation of parameters and analysis of missing values in linear structural relationship model. PhD thesis, University of malaya, kualua lumpur.

Mamun, A. S. M. A., Hussin, G. A., Zubairi, Y. Z., & Imon, A. H. M. R. (2013). Maximum likelihood estimation of linear structural relationship model parameters assuming the slope is known. Scienceasia, 39(5), 561–565.

Patefield, W.M. (1985). Information from the maximized likelihood function. *Biometrika,* 72, 664-668.

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

Wong, M.Y. (1989). Likelihood estimation of a simple linear regression model when both variables have error. *Biometrika,* 76, 141-148.