ASM Sc. J., 13, 2020
https://doi.org/10.32802/asmscj.2020.sm26(1.15)

Nonparametric Estimation for a Slope of a Replicated Linear Functional Relationship Model

Azuraini Mohd Arif[1], Yong Zulina Zubairi[2*] and Abdul Ghapor Hussin[3]

[1]*Institute for Advanced Studies, University of Malaya, 50603 Kuala Lumpur, Malaysia*
[2]*Centre for Foundation Studies in Science, University of Malaya, 50603 Kuala Lumpur, Malaysia*
[3]*Faculty of Defence Science and Technology, National Defence University of Malaysia, 57000 Kuala Lumpur, Malaysia*

In this paper, we propose a nonparametric method to estimate the slope of a replicated linear functional relationship model. The nonparametric method is a robust method in nature and does not affect when the observations have outliers. Additionally, the nonparametric method does not require the normality assumption. Using simulation studies, we compared the performance of the proposed nonparametric method with the traditional method using maximum likelihood estimation. It is found that without any outlier, the maximum likelihood estimation works well but when outliers exist in the data, our proposed nonparametric method gives a small mean square error, thus suggesting a better estimate.

Keywords: linear functional relationship model; nonparametric method; outliers

I. INTRODUCTION

Errors-in-variable model (EIVM) can be explained by the equation $Y = \alpha + \beta X$. In linear regression, the dependent variable X is assumed to be fixed and measured without error. However, in the errors-in-variable model, both X and Y variables are measured with error. In real situations, many experiments that involves relationships between two random variables that cannot be recorded correctly because there exists error (Gençay and Gradojevic, 2011; Patefield, 1985) thus, the errors-in-variable model is applicable rather than the linear regression model. The EIVM in this paper will emphasis on replicated Linear Functional Relationship Model (LFRM) where variable X is fixed and measured with error (Hassan *et. al.,* 2010; Ghapor *et. al.,* 2015).

Hussin (1997) termed the model as replicated LFRM for multiple x and y observations at each level of i. In replicated LFRM, it is often found that corresponding to a particular pair (X_i, Y_i) there may be replicated observations of X_i and Y_i occurring in p groups. As mentioned by Hussin (1997) and Barnett (1970), a linear relationship between X_i and Y_i are given by:

$$x_{ij} = X_i + \delta_{ij} \text{ and } y_{ik} = Y_i + \varepsilon_{ik}$$

$$\text{where} Y_i = \alpha + \beta X_i \quad (1)$$

$$\text{for} i = 1,2,\cdots,p\ , j = 1,2,\cdots,m_i \text{ and } k = 1,2,\cdots,n_i$$

We also assume that the observations on X_i and Y_i have been measured with errors δ_{ij} and ε_{ik} where $\delta_{ij} \sim N(0,\sigma^2)$ and $\varepsilon_{ik} \sim N(0,\tau^2)$. Numerous alternative methods of estimation have been suggested by many authors that required normality assumption which can lead to erroneous problems if outliers present in the data (Kendall and Stuart, 1979; Fuller, 1987). The robust method can be considered in

*Corresponding author's e-mail: yzulina@um.edu.my

estimating the parameter which does not require the normality assumption and also can diminish the effect of outliers (Ghapor *et. al.,* 2015). Although some authors like Ghapor *et. al.,* (2015) and Al-Nasser and Ebrahem (2005) using the nonparametric method, their research is limited to unreplicated LFRM. Thus, the proposed nonparametric method will be focused on replicated LFRM. In this paper, we proposed a new parametric estimation of the slope parameter based on the nonparametric method which was proposed by Ghapor *et. al.,* (2015).

The aim of this paper is to introduce the robust technique which is the trimmed mean to the replicated linear functional relationship model and also to compare this technique with the maximum likelihood estimation in estimating the slope parameter.

II. MATERIALS AND METHODS

A. Maximum Likelihood Estimation Method

Maximum Likelihood Estimation is the common method used in estimating the parameters. Replicated LFRM can be used when there is no information about the ratio of two variances in unreplicated LFRM or replication can be made on the observations (Hussin *et. al.,* 2005; Barnett, 1970).

In this case, the log-likelihood function can be expressed as:

$$\log L\left(\alpha,\beta,\sigma^2,\tau^2,X_1,\dots,X_p\right) =$$

$$Constant-\frac{1}{2}\left(\sum m_i\log\sigma^2+\sum n_i\log\tau^2\right)-\frac{1}{2}\left\{\sum\sum\frac{\left(x_{ij}-X_i\right)^2}{\sigma^2}+\sum\sum\frac{(y_{ik}-\alpha-\beta X_i)^2}{\tau^2}\right\}(2)$$

There are $(p+4)$ parameters to be estimated and may be obtained by differentiating the log likelihood function as given in equation (2) with respect to $\hat{\alpha},\hat{\beta},\hat{\sigma}^2,\hat{\tau}^2$ and $\hat{X}_i$respectively and equating to zero (Barnett, 1970). Thus, we can obtain the parameters in the order given by:

$$\hat{X}_i=\frac{1}{\hat{\Delta}_i}\left\{\frac{m_i\bar{x}_{i.}}{\hat{\sigma}^2}+\frac{n_i\hat{\beta}}{\hat{\tau}^2}(\bar{y}_{i.}-\hat{\alpha})\right\}\ \text{followed by}$$

$$\hat{\sigma}^2=\frac{\sum\sum(x_{ij}-\hat{X}_i)^2}{\sum m_i},\ \hat{\tau}^2=\frac{\sum\sum(y_{ik}-\hat{\alpha}-\hat{\beta}\hat{X}_i)^2}{\sum n_i},$$

$$\hat{\alpha}=\frac{\sum n_i(\bar{y}_{i.}-\hat{\beta}\hat{X}_i)}{\sum n_i}\text{and}\hat{\beta}_{MLE}=\frac{\sum n_i\hat{X}_i(\bar{y}_{i.}-\hat{\alpha})}{\sum n_i\hat{X}_i^2}.\qquad(3)$$

where$\bar{x}_{i.}=\frac{\sum x_{ij}}{m_i}$, $\bar{y}_{i.}=\frac{\sum y_{ik}}{n_i}$, and $\hat{\Delta}_i=\frac{m_i}{\hat{\sigma}^2}+\frac{n_i\hat{\beta}^2}{\hat{\tau}^2}$.

The estimates of $\hat{\alpha},\hat{\beta}_{MLE},\hat{\sigma}^2,\hat{\tau}^2$ and $\hat{X}_i$can be solved iteratively given initial values from the linear regression model. Our primary interest in this research is the estimated slope, $\hat{\beta}_{MLE}$.

B. The Nonparametric Method (The Proposed Method)

The method we propose for estimating the slope parameter is by considering the nonparametric method proposed by Ghapor *et. al.,* (2015). Normality assumption can be ignored in this method. In this paper, we will follow the steps as suggested by Ghapor (2015) but in Step 5, we changed the calculation of the slope by using trimmed mean. Hence, we used the trimmed mean to measure the slope of the model instead of the usual mean or median in order to reduce the effect of outlier present in the data. By trimming a certain amount of percentage around 10 to 20 percent, the influence of outlying observations is negated and still give a reasonable estimate of the slope (Wilcox, 2012; Welsh, 1987). As mentioned by Wilcox (2012), a good choice for trimmed mean is 20% to acquire a relatively small standard error among commonly occurring situations.

As stated by Ghapor (2015), the steps are listed down in detail:

Step 1: The observations are first arranged in ascending order, based on x value namely

$x_{(1)}\le x_{(2)}\le\cdots\le x_{(n)}$.

The associated values of $y$which may not be in ascending order are taken namely,

$y_{[1]}\le y_{[2]}\le\cdots\le y_{[n]}$.

The new pairs will be $\left(x_{(i)},y_{[j]}\right)$.

Step 2: All the data are divided into p –subsamples. These subsamples contain m elements such that $p\times m=N$ where m_i is the maximum divisor of N such that $p\le m$. The advantage we abstracting the number of paired slopes that need to be calculated (Al-Nasser and Ebrahem, 2005).

Step 3: Find all the possible slopes.

$$\left\{b_x(k)_{ij}=\frac{y_{[j]}-y_{[i]}}{x_{(j)}-x_{(i)}};\ i=1,2,\dots,j-1;j=2,3,\dots,m\right\};k=1,2,\dots,p$$

Step 4: Repeat Steps 1 to 3 by interchanging $y$and x to get possible paired of $b_y(k)_{ij}$

$$\left\{b_y(k)_{ij}=\frac{y_{(j)}-y_{(i)}}{x_{[j]}-x_{[i]}};\ i=1,2,\dots,j-1;j=2,3,\dots,m\right\};k=1,2,\dots,p$$

Step 5: Combine all the slopes from Step 4

Step 6: Find the trimmed mean of the slopes.

$$\hat{\beta}_{trim} = trimmed\ mean\{b_x(k)_{ij}, b_y(k)_{ij}\}$$

This gives us the new estimated slope ofthe parameter, $\hat{\beta}_{trim}$.

C. Simulation Study

The performance of the proposed method with the maximum likelihood estimation for the replicated linear functional relationship model was carried out using a simulation study in R software. The observations are then simulated using our model as described earlier. Without loss of generality, the true value is fixed at $\alpha = 1$ and $\beta = 1$. We simulate 10000 trials for $N = 20, 80$ and 100. Additionally, the performance of the slope parameter in replicated linear functional relationship model when the observation has no outlier, $5\%, 10\%$ and 15% outliers respectively are also considered. By using this relationship,

$y_C = 1 + X_C + \varepsilon_C$ with $\varepsilon_C \sim N(0, 25)$

We contaminate data points as suggested by Al-Nasser and Ebrahem (2005). For the proposed estimator the required values of p and m for each sample size are given in Table 1.

Table 1. Values of p and m

Sample Size, N	p	m
20	4	5
80	8	10
100	10	10

III. RESULTS AND DISCUSSIONS

Table 2 shows the estimated bias for the slope of the replicated linear functional relationship model. In general, as the sample size increase from 20 to 100, the estimated bias is decreasing for both maximum likelihood estimation method and the proposed nonparametric method. The same can be said as we introduce outliers in our data from 5% to 15% outliers, the estimated bias decreases as the number of observations increase. Nevertheless, the proposed method gives smaller estimated bias value than the maximum likelihood estimation method when the observations have outliers and none.

Table 3 shows the mean square error of the slope for both methods. When the data have no outlier, the mean square error (MSE) for both methods give almost the same value for each sample size. At each level of contamination from 5% to 15% outliers, the proposed nonparametric method shows consistently smaller values of mean square error than the maximum likelihood estimation.

To illustrate the practicality of the method, we use data from a study that measures the accuracy of some widely used body-composition techniques for children between the ages 4 and 10 years by two different techniques, namely skinfold thickness (ST) and bioelectrical resistance (BR)(Goran *et. al.,* 1996). As measurement error can occur in both variables for this experiment, we note that we can describe the relationship by replicated LFRM as given in equation (1). The data consists of 96 observations and we assume that the error terms follow a normal distribution. As mentioned by Kim (2000) and Imon & Hadi (2008), some original y values were substituted by outliers namely at 5%, 10%, and 15% level to form different conditions in investigating the slope effect by two different methods. The estimated slopes by two different methods were shown in Table 4. From Table 4, both methods showed a somewhat similar value of the slope estimates when there is no outlier. However, when outliers increased from 5% to 15 %, the estimates of the slope using the maximum likelihood method become huge compared to the proposed nonparametric method.

This shows that the proposed nonparametric method can be used in estimating the slope of the replicated linear functional relationship model in the presence of outliers. Additionally, we can possibly use replicated linear functional relationship model to estimate the parameter of interest when there is a lack of information on the ratio of two variances in the unreplicated linear functional relationship model. This problem does not arise in replicated LFRM as the number of parameters is fixed and only the degree of replication increases with an increasing number of observations.

Table 2. The Estimated Bias of the slope

Outlier	Method	N=20	N=80	N=100
No outlier	$\hat{\beta}_{MLE}$	9.434 E-04	9.521 E-04	9.158 E-04
	$\hat{\beta}_{trim}$	4.369	1.721	4.613

		E-03	E-03	E-04
5% outliers	$\hat{\beta}_{MLE}$	5.975 E-01	3.791 E-01	3.767 E-01
	$\hat{\beta}_{trim}$	2.359 E-02	9.458 E-03	3.377 E-03
10% outliers	$\hat{\beta}_{MLE}$	7.966 E-01	8.309 E-04	7.688 E-04
	$\hat{\beta}_{trim}$	1.431 E-01	1.137 E-03	5.012 E-03
15% outliers	$\hat{\beta}_{MLE}$	5.909 E-01	3.790 E-01	3.767 E-01
	$\hat{\beta}_{trim}$	2.687 E-01	9.484 E-02	5.511 E-02

Table 3. The Mean Square Error of the slope

Outlier	Method	N=20	N=80	N=100
No outlier	$\hat{\beta}_{MLE}$	1.307 E-04	3.121 E-05	2.499 E-05
	$\hat{\beta}_{trim}$	1.652 E-04	3.481 E-05	2.505 E-05
5% outliers	$\hat{\beta}_{MLE}$	3.573 E-01	1.437 E-01	1.419 E-01
	$\hat{\beta}_{trim}$	1.015 E-03	1.232 E-04	3.745 E-05
10% outliers	$\hat{\beta}_{MLE}$	6.348 E-01	8.170 E-05	6.627 E-05
	$\hat{\beta}_{trim}$	2.769 E-02	4.031 E-05	5.319 E-05
15% outliers	$\hat{\beta}_{MLE}$	3.495 E-01	6.209 E-01	1.420 E-01
	$\hat{\beta}_{trim}$	9.043 E-02	9.717 E-03	3.341 E-03

Table 4. Slopes Estimates Using Goran et al. Data (1996)

Contamination	Method	Slope
No outlier	$\hat{\beta}_{MLE}$	0.982
	$\hat{\beta}_{trim}$	0.974
5% outliers	$\hat{\beta}_{MLE}$	2.998
	$\hat{\beta}_{trim}$	1.014
10% outliers	$\hat{\beta}_{MLE}$	4.486
	$\hat{\beta}_{trim}$	1.023
15% outliers	$\hat{\beta}_{MLE}$	5.321
	$\hat{\beta}_{trim}$	1.002

IV. SUMMARY

In conclusion, by looking at the estimated bias and mean square error of the slope, we can conclude that the proposed nonparametric method is superior to the maximum likelihood estimation. This can be seen from the simulation studies when the percentage of outliers increase, the mean square error of the maximum likelihood estimation becomes huge and breaks down easily as compared to the proposed nonparametric method.

V. ACKNOWLEDGEMENT

We are most grateful to the University of Malaya (research grant PG128-2015B and GPF006H-2018) for financial support.

VI. REFERENCES

A. G. Hussin, A. G. 1997, Pseudo-replication in functional relationship with environmental application, PhD thesis, University of Sheffield, England.

Al-Nasser, Amjad D &Mohammed Al-Haj Ebrahem. 2005, "A New Nonparametric Method forEstimating the Slope of Simple Linear Measurement Model in the Presence ofOutliers." *Pak. J. Statist* 21 (3): 265–74.

Barnett, VD. 1970, Fitting Straight Lines-the Linear Functional Relationship with Replicated Observations. *AppliedStatistics*. 19 (2): 135-44.

Fuller, Wayne A. 1987, *Measurement Error Models*. John Wiley & Sons. Gençay, Ramazan & Nikola Gradojevic.2011. "Errors-in-Variables EstimationWavelets." *Journal of Statistical Computation and Simulation* 81 (11):1545–64.

Ghapor, A.A., Y.Z. Zubairi, A.S.M.A Mamun, & A.H.M.R. Imon. 2015, "A RobustNonparametric

Slope Estimation in Linear Functional Relationship Model."*Pak. J. Statist* 31 (3): 339–50.

Goran, MI, P Driscoll, R Johnson, TR Nagy, and G Hunter. 1996, "Cross-Calibration of Body-Composition Techniques against Dual-Energy X- Ray Absorptiometry in Young Children." *Am J Clin Nutr* 63 (3): 299–305.

Hussin, Abdul Ghapor, Nick Fieller & Eleanor Stillman. 2005, "Pseudo-Replicates in the Linear CircularFunctional Relationship Model. "*Journal of Applied Sciences* 5 (1): 138–43.

Imon, A H M R, and A S Hadi. 2008, "Identification of Multiple Outliers in Logistic Regression."*Communications in Statistics-Theory and Methods* 37 (11): 1697–1709.

Kendall, M.G. & Alan Stuart. 1979, *The Advanced Theory of Statistics*. *London:Griffin*. Vol. 2.

Kim, Myung Geun. 2000. Outliers and Influential Observations in the Structural Errors-in-Variables Model. Journal of Applied Statistics. 27 (4): 451–60

Patefield, W.M. 1985, "Information from the Maximized Likelihood Function." *Biometrika* 72 (3): 664–68.

S.F. Hassan, A.G. Hussin & Y.Z. Zubairi. 2010, "Estimation of functional relationship model for circular variables and its application in measurement problem". *Chiang Mai J. Sci.* 37: 195–205.

Welsh, A. H. 1987, "Trimmed Mean in the Linear Model". *The Annals of Statistics*, 15: 20–45.

Wilcox, R.R. 2012, *Introduction to Robust Estimation and Hypothesis Testing*. 3rd ed. Academic Press.