MJFAS MALAYSIAN JOURNAL OF FUNDAMENTAL AND APPLIED SCIENCES
PRINT ISSN: 2289-5981 | ONLINE ISSN: 2289-599X

RESEARCH ARTICLE

Parameter estimation in replicated linear functional relationship model in the presence of outliers

Azuraini Mohd Arif [a], Yong Zulina Zubairi [b,*], Abdul Ghapor Hussin [c]

[a] Institute of Graduate Studies, Universiti Malaya, 50603 Kuala Lumpur, Malaysia.
[b] Centre for Foundation Studies in Science, Universiti Malaya, 50603 Kuala Lumpur, Malaysia.
[c] Faculty of Defence Science and Technology, Universiti Pertahanan Nasional Malaysia, 57000 Kuala Lumpur, Malaysia.

* Corresponding author: yzulina@um.edu.my

Article history
Received 13 May 2019
Revised 20 July 2019
Accepted 29 January 2020
Published Online 15 April 2020

Abstract

The relationship between two linear variables where both variables are observed with errors can be modeled using a linear functional relationship model. However, when there is no knowledge about the ratio of error variance, we proposed that one can use the replicated linear functional relationship model. The aim of this study is to compare the parameter estimates between unreplicated and replicated linear functional relationship model. The study also extends to examine the behavior of the estimators of the replicated linear functional relationship model in the presence of outliers. A simulation study is performed to investigate the performance of the model. In the absence of outlier, it is found that the value of the parameter estimates is almost similar for both models. Whereas in the presence of outliers, the parameter estimates of the replicated linear functional relationship model have a smaller mean square error as the number of observations increased. This suggests the superiority of the replicated model.

Keywords**:** Linear functional relationship model, outliers, replicated

INTRODUCTION

Suppose the variables X and Y are related by the equation $Y=\alpha+\beta X$. There is no statistical problem if variable X and Y can be observed exactly. If Y is observed with error, then we use ordinary linear regression. If both variable X and Y are observed with errors, errors-in-variable model (EIVM) is used. EIVM is an extension of a linear regression model that looks at the relationship between two variables where both variables are subjected to measurement error. The study of the EIVM dates back to the late 18th century when Adcock (1878) investigated the problem of fitting a linear relationship when both the dependent variable and independent variable are subject to error. The EIVM occurs in many fields such as in the industrial experiment, quality control, epidemiological studies, economics, and environmental science (Buonaccorsi, 2010; Gencay and Gradojevic, 2011). EIVM can be divided into three categories which are functional relationship model, structural relationship model, and ultrastructural relationship model (Fuller, 1987). In this study, we consider the Linear Functional relationship Model (LFRM) where the variable X is fixed.

In LFRM, there are $(n+4)$ parameters that need to be estimated, namely the intercept α, the slope β, the two error variances σ^2, τ^2, and the incidental parameters X_i. The log likelihood function is given by

$$\log L\,(\alpha,\beta,\sigma^2,\tau^2,X_i\,;x_1,\dots,x_n;\;y_1,\dots,y_n)=-n\log(2\pi)$$

$$-\frac{n}{2}(\log\sigma^2+\log\tau^2)-\frac{\sum(x_i-X_i)^2}{2\sigma^2}-\frac{\sum(y_i-\alpha-\beta X_i)^2}{2\tau^2}.$$

However, when the number of observations increases, the number of parameters will also increase. This can lead to the parameter estimation problem as it has been reported in another study as inconsistencies with the existence of the incidental parameter (Neyman and Scott, 1951; Lindley, 1953; Ghapor *et al.*, 2015). In order to solve the problem, either the knowledge of ratio of two variances is known or replication can be made (Barnett 1970; Kendall and Stuart, 1979; Hussin *et al.*, 2005). The goal of this research is to propose a solution in estimating parameters of interest when we only have unreplicated data and the ratio of the error variances is unknown. Thus, in this study, we will consider the parameter slope β and error variance σ^2 using the maximum likelihood method for both unreplicated and replicated linear functional relationship model and investigate the performance of both estimators for the replicated model in the presence of outliers.

MAXIMUM LIKELIHOOD ESTIMATION METHOD

Unreplicated linear functional relationship model

Maximum likelihood estimation (MLE) method is a common method used in estimating the parameters of LFRM. Consider X and Y are linearly related but observed with error, unreplicated LFRM can be expressed by the equation

$$Y_i=\alpha+\beta X_i \quad \text{for } i=1,2,3,\dots,n \tag{1}$$

For any fixed X_i, we observe x_i and y_i from continuous linear variable subject to errors δ_i and ε_i respectively, i.e.

$$x_i = X_i + \delta_i \text{ and } y_i = Y_i + \varepsilon_i \tag{2}$$

where the error terms δ_i and ε_i are assumed to be mutually independent and normally distributed random variables, i.e.

$$\delta_i \sim N(0, \sigma^2) \text{ and } \varepsilon_i \sim N(0, \tau^2) \tag{3}$$

An assumption must be made in order to avoid the problem in unreplicated LFRM (Solari, 1969; Moran, 1971). With the assumption that the ratio of error variances is known, $\tau^2 = \lambda\sigma^2$ and there are $(n+3)$ parameters to be estimated namely α, β, σ^2, and the incidental parameters X_i (Fuller, 1987; Kendall and Stuart, 1979), the log likelihood function can be expressed as

$$\log L\,(\alpha, \beta, \sigma^2, X_i\,; x_1, \dots, x_n;\ y_1, \dots, y_n) = -n\log(2\pi)$$

$$-\frac{n}{2}(\log\lambda) - n\log\sigma^2 - \frac{\sum(x_i - X_i)^2}{2\sigma^2} - \frac{\sum(y_i - \alpha - \beta X_i)^2}{2\lambda} \tag{4}$$

The parameters may be obtained by differentiating the log likelihood function as given in equation (4) with respect to namely $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}^2$, and $\widehat{X_\iota}$ and equating to zero. Thus, we can obtain the parameters given by

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}\,,\ \hat{\beta} = \frac{S_{yy} - \lambda S_{xx}\sqrt{\left(S_{yy} - \lambda S_{xx}\right)^2 - 4\lambda S_{xy}^2}}{2S_{xy}}, \tag{5}$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\left\{\sum\left(x_i - \hat{X}_i\right)^2 + \sum\left(y_i - \hat{\alpha} - \hat{\beta}\hat{X}_i\right)^2\right\}$$

and $\hat{X}_i = \frac{\lambda x_i + \widehat{\beta}(y_i - \widehat{\alpha})}{\lambda + \widehat{\beta}^2}$

where $\bar{y} = \frac{1}{n}\sum y_i, \bar{x} = \frac{1}{n}\sum x_i$, $S_{xx} = \frac{1}{n}\sum\left(x_i - \hat{X}_i\right)^2$,

$S_{yy} = \frac{1}{n}\sum(y_i - \bar{y})^2$, $S_{xy} = \frac{1}{n}\sum\left(x_i - \hat{X}_i\right)(y_i - \bar{y})$

Replicated linear functional relationship model

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. 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. A linear relationship between X_i and Y_i is given by

$$x_i = X_i + \delta_{ij} \text{ and } y_i = Y_i + \varepsilon_{ik} \text{ where} \tag{6}$$

$$Y_i = \alpha + \beta X_i \quad \text{for } i = 1,2,\dots,p\,, j = 1,2,\dots,m_i\,, k = 1,2,\dots,n_i \tag{7}$$

It is assumed that $\delta_{ij} \sim N(0, \sigma^2)$ and $\varepsilon_{ik} \sim N(0, \tau^2)$.

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

$$\log L\left(\alpha, \beta, \sigma^2, \tau^2, X_1, \dots, X_p\,; x_{11}, \dots, x_{pm_p};\ y_{11}, \dots, y_{pn_p}\right) =$$

$$-constant - \frac{1}{2}\left(\sum m_i \log\sigma^2 + \sum n_i \log\tau^2\right)$$

$$-\sum\sum\frac{\left(x_{ij} - X_i\right)^2}{2\sigma^2} - \sum\sum\frac{(y_{ik} - \alpha - \beta X_i)^2}{2\tau^2} \tag{8}$$

There are $(p+4)$ parameters to be estimated and may be obtained by differentiating the log likelihood function as given in equation (8) with respect to $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}^2$, $\hat{\tau}^2$, and $\hat{X}_i$ and equating to zero (Barnett, 1970). Thus, we can obtain the parameters in the order given by

$$\hat{\sigma}^2 = \frac{\sum\sum(x_i - \hat{X}_i)^2}{\sum m_i}, \quad \hat{\tau}^2 = \frac{\sum\sum\left(y_i - \hat{\alpha} - \widehat{\beta}\hat{X}_i\right)^2}{\sum n_i},$$

$$\hat{\alpha} = \frac{\sum n_i\left(\bar{y}_{i.} - \widehat{\beta}\hat{X}_i\right)}{\sum n_i}, \quad \hat{\beta} = \frac{\sum n_i \hat{X}_i(\bar{y}_{i.} - \hat{\alpha})}{\sum n_i \hat{X}_i^2}, \quad \text{and}$$

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

$$\bar{x}_{i.} = \frac{1}{m_i}\sum x_{ij}\ ,\ \bar{y}_{i.} = \frac{1}{n_i}\sum y_{ik} \text{ and } \hat{\Delta}_i = \frac{m_i}{\hat{\sigma}^2} + \frac{n_i\widehat{\beta}^2}{\hat{\tau}^2}.$$

The estimates of $\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2$, $\hat{\tau}^2$, and $\hat{X}_i$ can be solved iteratively given some suitable initial values at the estimate. An initial estimate can be obtained from the unreplicated linear functional relationship model.

Simulation study

A simulation study was carried out using R software in order to evaluate the performance of the unreplicated and replicated linear functional relationship model. The parameters of interest in this study are the slope $\hat{\beta}$ and the error variance $\hat{\sigma}^2$. The observations are then simulated using our model as described earlier. Without loss of generality, the true value is fixed at $\hat{\alpha} = 1$ and $\hat{\beta} = 1$ for both models. Additionally, in unreplicated linear functional relationship model, we assume that the ratio of two variance or λ is equal to one. We simulate 1000 trials for $n = 20, 50, 80,$ and 100. Additionally, the performance of the parameter of interest in replicated linear functional relationship model when the observation has no outlier, single outlier, 10%, and 20% outliers are also considered. This is done by contaminating data points as suggested by Al-Nasser and Ebrahem (2005) using relationship $y_C = 1 + X_C + \varepsilon_C$ with $\varepsilon_C \sim N(0{,}25)$. The mean square error is used to assess the performance of the slope and the error variance.

RESULTS AND DISCUSSION

Tables 1 and 2 show the performance of both models using measures of mean and estimated bias. It can be seen that from Table 1, the value for the slope is almost the same between unreplicated and replicated linear functional relationship model. The estimated bias for both models is almost the same as the number of observations increase. However, in Table 2, the value of the estimated bias for the unreplicated model is less than the replicated model as the number of observations increase. Next, from Table 3, where the errors δ_i and ε_i are normally distributed, when the data have no outlier, the mean square error (MSE) decreases as the number of observations increases. As we introduce outliers in our data from single outlier to 20% outliers, the MSE decreases as the number of observations increase. Nevertheless, the replicated model has a smaller MSE value than the unreplicated model, suggesting the replicated model is much a better model than the unreplicated model when the data has outliers. Similar trend also can be observed for Table 4 which the MSE values decrease with the increase of sample sizes. Comparing the replicated linear functional relationship model with the unreplicated model, at each level of contamination, the replicated model shows consistently smaller values of MSE than the unreplicated model. Through simulation study, the slope estimates for replicated model remain resistant when outliers exist in the data. This shows that we can possibly use replicated linear functional relationship model to estimate the parameter of interest when we are lacking information on the λ or the ratio of two variances in unreplicated linear functional relationship model is not available. Furthermore, when the number of observations increases, the estimation of unreplicated LFRM becomes complicated as the number of parameters will also increase. 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 1 Mean and estimated bias for slope $\hat{\beta}$ estimates.

Sample Size	Model	Mean	Estimated Bias
N=20	Unreplicated	1.0053	0.0053
	Replicated	1.0042	0.0042
N=50	Unreplicated	0.9989	0.0011
	Replicated	0.9988	0.0012
N=80	Unreplicated	1.0044	0.0044
	Replicated	1.0041	0.0041
N=100	Unreplicated	1.0013	0.0013
	Replicated	1.0012	0.0012

Table 2 Mean and estimated bias for error variance $\hat{\sigma}^2$ estimates.

Sample Size	Model	Mean	Estimated Bias
N=20	Unreplicated	0.9996	0.0004
	Replicated	0.8446	0.1554
N=50	Unreplicated	0.9944	0.0056
	Replicated	0.9305	0.0695
N=80	Unreplicated	0.9997	0.0003
	Replicated	0.9362	0.0638
N=100	Unreplicated	1.0039	0.0039
	Replicated	0.9474	0.0526

Table 3 Mean square error (MSE) of the slope.

Contamination	Model	N=20	N=50	N=80	N=100
No outlier	Unreplicated	1.410E-02	5.406E-03	3.459E-03	2.586E-03
	Replicated	1.390E-02	5.285E-03	3.380E-03	2.540E-03
Single outlier	Unreplicated	5.764E+02	9.035E-01	2.009E-01	1.082E-01
	Replicated	5.576E+02	5.723E-02	2.306E-02	1.518E-02
10% outliers	Unreplicated	9.054E+04	4.255E+01	3.774E+01	3.707E+01
	Replicated	9.053E+04	5.352E-03	3.502E-03	2.639E-03
20% outliers	Unreplicated	3.076E+02	1.577E+02	1.359E+02	1.348E+02
	Replicated	1.648E+02	4.689E+00	3.484E-03	2.626E-03

Table 4 Mean square error (MSE) of the error variance.

Contamination	Model	N=20	N=50	N=80	N=100
No outlier	Unreplicated	1.241E-01	4.184E-02	2.486E-02	2.068E-02
	Replicated	1.158E-01	4.883E-02	3.036E-02	2.280E-02
Single outlier	Unreplicated	6.749E+01	2.609E+01	1.239E+01	8.436E+00
	Replicated	1.327E+01	5.052E-02	3.323E-02	2.599E-02
10% outliers	Unreplicated	7.574E+01	5.250E+01	5.253E+01	5.219E+01
	Replicated	5.022E+01	5.056E-02	3.368E-02	2.668E-02
20% outliers	Unreplicated	6.599E+01	6.002E+01	6.031E+01	6.001E+01
	Replicated	4.853E+00	4.002E-01	3.368E-02	2.669E-02

Real example

A study that measures the accuracy of some widely used body-composition techniques for children between the ages 4 and 10 years old, two different techniques, namely skinfold thickness (ST) and bioelectrical resistance (BR), are used to illustrate the use of replicated linear functional relationship model (Ghapor *et al.*, 2015; Goran *et al*., 1996). In this data, we assumed that the measurement error can occur on both variables and that the error term follows a normal distribution. Some original *y* values were replaced by the values of the outliers namely, a single outlier, 10%, and 20% outliers to examining the slope effect by following Kim (2000). The estimated slope for both unreplicated and replicated linear functional relationship model are presented in Table 5.

From Table 5, we can see the slope parameter of the linear functional relationship model is almost the same between the unreplicated and replicated model. This shows that when the information about lambda (the ratio of errors variance) is not available, we can use replicated linear functional relationship model as an alternative to estimate the slope parameter.

Table 5 Slope estimates from Goran *et al.* (1996) data.

Contamination	Unreplicated	Replicated
No outlier	1.09969	1.09838
Single outlier	1.48694	1.25986
10% outliers	5.76879	1.01953
20% outliers	13.54840	0.99039

CONCLUSION

In conclusion, the proposed MLE can be used for unreplicated and replicated LFRM for estimating the slope and the error variance parameter. However, when the ratio of two variances is unknown, the replicated LFRM is the better model in estimating the parameter of interest. More importantly, the replicated LFRM can give better estimates in the presence of outliers.

ACKNOWLEDGMENT

This work was financially supported by University of Malaya (research grant PG128-2015B and BKS010-2016), GE STEM grant with vot no. 07397, and Ministry of Higher Education Malaysia.

REFERENCES

Adcock, R. J. 1878. A problem in least squares. *Annals of Mathematics*. 5 (2): 53–54.

Al-Nasser, Amjad D., Ebrahem, M. A. 2005. A New nonparametric method for estimating the slope of simple linear measurement model in the presence of outliers. *Pakistan Journal Statistics*. 21 (3): 265–74.

Barnett, V. D. 1970. Fitting straight lines-the linear functional relationship with replicated observations. *Applied Statistics*. 19 (2): 135–44.

Buonaccorsi, J. P. 2010. *Measurement Error Models, Methods and Applications*. New York: Chapman and Hall.

Fuller, Wayne A. 1987. *Measurement Error Models*. John Wiley & Sons.

Gencay, R., Gradojevic, N. 2011. Errors-in-variables estimation with wavelets. *Journal of Statistical Computation and Simulation*. (81): 1545-1564.

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

Goran, M. I., Driscoll, P., Johnson, R., Nagy, T. R., Hunter, G.. 1996. Cross-Calibration of Body-Composition Techniques against Dual-Energy X-Ray absorptiometry in young children. The American Journal of Clinical Nutrition. 63 (3): 299–305.

Hussin, A. G., Fieller, N., Stillman, E. 2005. Pseudo-replicates in the linear circular functional relationship model. *Journal of Applied Sciences*. 5: 138-143.

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

Kim, Geun, M. 2000. Outliers and Influential observations in the structural errors-in-variables model. *Journal of Applied Statistics*. 27 (4): 451–60.

Lindley, D.V. 1953. Estimation of a functional relationship. *Biometrika*. 40 (1/2): 47–49.

Moran, P. A. P. 1971. Estimating structural and functional relationships. *Journal of Multivariate Analysis*. 1 (2): 232–55.

Neyman, J., Scott, E. L. 1951. On certain methods of estimating the linear structural relation. *The Annals of Mathematical Statistics*. 22 (3): 352–61.

Solari, Mary E. 1969. The 'maximum Likelihood Solution' of the Problem of Estimating a Linear Functional Relationship." J*ournal of the Royal Statistical Society. Series B (Methodological).* 31 (2): 372–75.