Journal of the Nigerian Association of Mathematical Physics
Volume **22 (November, 2012), pp 265 – 272**

Ranking of Simultaneous Equation Estimators to Outliers From Heavy-Tailed Quasi-Uniform Distribution

[1]Oseni B.M., [2]Adepoju A. A. and [2]Olubusoye O. E.

[1]Federal University of Technology, Akure, Nigeria
[2]University of Ibadan, Nigeria

Abstract

In this work, the ranking of the performances of two-equation simultaneous models when outliers are presumed present in a convoluted exogenous variable is carried out. The exogenous variable is a convolution of normal and uniform distribution. Monte Carlo experiment was carried out to investigate the performances of four estimators namely: Ordinary Least Squares (OLS), Two Stage Least Squares (2SLS), Limited Information Maximum Likelihood (LIML) and Three Stage Least Squares (3SLS). Five sample sizes were used to allow for measure of asymptotic properties of these estimators. The experiment was replicated 1000 times and the results were evaluated using Total Absolute Bias (TAB), Variance and Root Mean Squared Error (RMSE). It is observed that the performances of the estimators when lower triangular matrix is used are better than that of upper triangular matrix. OLS using TAB as evaluation criterion is better than the other estimators when an exogenous variable is convoluted for the just-identified equation. The performance of 2SLS is best for the over-identified equation. OLS possesses the least variance for both equations and both matrices while LIML has the worst variance in most cases. OLS possesses the smallest RMSE for both matrices and equations except with the over-identified equation using lower triangular matrix when an exogenous variable is convoluted.

Keywords: Outliers, Convolution, Normal Distribution, Uniform Distribution, Monte Carlo, Estimators, Simultaneous Equation.

1.0 Introduction

The assumption of normality is central to most statistical procedures. With real life data, in some cases this assumption holds approximately while it may not hold in other cases. This violation may be attributed to a number of factors among which are outliers. Outliers are either legitimate or illegitimate observation far away from the rest of the observation. Outliers often come from a family of approximately normal distribution with heavy tail or long tail [1, 2]. These distributions are mostly convolutions of two or more distributions [1]. Outliers have distorting influence on the statistical procedures [1, 2, 3], especially when the underlying distribution is normal. Outliers could also contain useful information on abnormal behavior of the system being described by the data [4]. This behavior is often explore in a number of applications such as credit card fraud, financial applications, marketing and host of other areas.

This has lead to some researchers suggesting robust inference as alternative [5] and some are even of the opinion that the normality assumption be substituted with that which accommodate heavy tail distribution (such as the multivariate-t distribution) [6] or scale mixtures of normal distribution [7].

A good number of estimators that are robust to outliers have been developed. These estimators such as the Generalized Median of Slopes [8], M-estimators [9], MM-estimator [10], $\tau -$ estimates [11] and host of other estimators were developed for linear regressions. Those that were developed for simultaneous equation are often a modified version or weighted versions of the established classical estimators e.g. the weighted 2SLS considered by [12]. Despite these development, most statistician still shows preference for the classical estimators and yet, they neither test for the assumptions nor check for outliers [3].

In this work attempt is made to compare and rank some simultaneous equation estimators when one of the variables is assumed to come from quasi-uniform distribution instead of the uniform distribution [13] as found in most literature. The underlying distribution of the error term is still assumed to be normal. This is done in other to simulate a real life scenario where outliers are present in a variable. Ordinary Least Squares (OLS), Two Stage Least Squares (2SLS), Limited

[1]Corresponding author: ***Oseni B. M.,*** E-mail: osenibm@yahoo.com, Tel. +234 806 624 5700

Information Maximum Likelihood (LIML) and Three Stage Least Squares (3SLS).

2.0 Model Specification

Consider the two-equation simultaneous model

$$y_{1i} = \beta_{12} y_{2i} + \gamma_{11} x_{1i} + \mu_{1i}$$
$$y_{2i} = \beta_{21} y_{1i} + \gamma_{22} x_{2i} + \gamma_{23} x_{3i} + \mu_{2i} \quad (1)$$

where β_{12} and β_{21} are the parameter of the endogenous variables y_{2i} and y_{1i} respectively, $\gamma_{rc}\,(r=1,2; c=1,2,3)$ are the parameters of the exogenous variables x_{ci} and μ_{ri} are the disturbance terms.

This model can represented more compactly by

$$\mathrm{B}y + \Gamma x = U \quad (2)$$

where $y = \begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix}, x = \begin{pmatrix} x_{1i} \\ x_{2i} \\ x_{3i} \end{pmatrix}, U = \begin{pmatrix} \mu_{1t} \\ \mu_{2t} \end{pmatrix}, \mathrm{B} = \begin{pmatrix} 1 & -\beta_{12} \\ -\beta_{21} & 1 \end{pmatrix}$ and $\Gamma = \begin{pmatrix} -\gamma_{11} & 0 & 0 \\ 0 & -\gamma_{22} & -\gamma_{23} \end{pmatrix}$

The first equation of the model is said to be over-identified, since the total number of variables excluded from it but included in other equation of the model is greater than the number of equation less one, while the second equation is just-identified.

Outliers were introduced into the exogenous variable x_2, by assuming that x_2 is a convolution of the variables z_1 and z_2 (i.e. $x_2 = z_1 + z_2$) where $Z_1 \sim U(0,1)$and $Z_2 \sim N(0,1)$ .Thus Z_1 follows uniform distribution while Z_2 follows normal distribution. The other exogenous variables are assumed to come from $U(0,1)$ [13]. Thus;

$$f_{X_2}(x_2) = \int h(x_2 - z_1) g(z_1) dz_1 \quad (3)$$

where $f_{Z_2}(z_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z_2^2}{2}}$, and $f_{Z_1}(z_1) = I_{(0,1)}$, $0 < z_1 < 1$

which gives

$$f_{X_2}(x_2) = \frac{1}{\sqrt{2\pi}} \int_0^1 e^{-(x_2 - z_1)^2/2} dz_1$$

Hence

$$f_{X_2}(x_2) = \left(\vartheta(x_2) - \vartheta(x_2 - 1)\right) \text{ where } \vartheta(z_1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_1} e^{\frac{-t}{2}} dt \quad (4)$$

The function $f_{X_2}(x_2)$ is a quasi-uniform distribution with heavy tail. Hence x_{2i} was simulated from a distribution with heavy tail. As an illustration of the possibility of the sample from the above distribution containing outliers, consider the Tables 1a and 1b.

Table 1: Sample data simulated from $U(0,1)$ and $N(0,1)$

i	***z1i***	***z2i***	***x2i***	***I***	***z1i***	***z2i***	***x2i***
1	0.198675	-0.84636	-0.64769	16	0.498459	-0.00386	0.494596
2	0.915006	1.372241	2.287247	17	0.754509	0.688748	1.443257
3	0.786218	0.793366	1.579584	18	0.868648	1.120025	1.988673
4	0.589343	0.225855	0.815198	19	0.459517	-0.10165	0.357868
5	0.015595	-2.15464	-2.13904	20	0.649861	0.384946	1.034807
6	0.768059	0.732471	1.50053				
7	0.784173	0.786365	1.570538				
8	0.583666	0.211282	0.794949				
9	0.741081	0.646681	1.387762				
10	0.483291	-0.04189	0.441396				
11	0.412458	-0.22123	0.19123				
12	0.353404	-0.37615	-0.02274				
13	0.326456	-0.44972	-0.12326				
14	0.645009	0.37188	1.016889				
15	0.138127	-1.08877	-0.95065				

(a)

i	z_{1i}	z_{2i}	x_{2i}	*I*	z_{1i}	z_{2i}	x_{2i}
1	0.36198	-0.35317	0.008808	16	0.51561	0.039139	0.554749
2	0.697012	0.515827	1.212839	17	0.608203	0.27464	0.882843
3	0.258827	-0.64696	-0.38814	18	0.532945	0.082674	0.615619
4	0.689047	0.493151	1.182198	19	0.281442	-0.57856	-0.29712
5	0.011689	-2.26721	-2.25552	20	0.458266	-0.1048	0.353463
6	0.340403	-0.41136	-0.07096				
7	0.21015	-0.8059	-0.59575				
8	0.534562	0.086743	0.621305				
9	0.18778	-0.8861	-0.69832				
10	0.211585	-0.80094	-0.58935				
11	0.64742	0.378363	1.025783				
12	0.993286	2.472207	3.465493				
13	0.285928	-0.56532	-0.27939				
14	0.807947	0.870356	1.678303				
15	0.419477	-0.20323	0.216246				

(b)

The convoluted variable x_{2i} from the samples are plotted in Figure 1 ((a) and (b)) below.

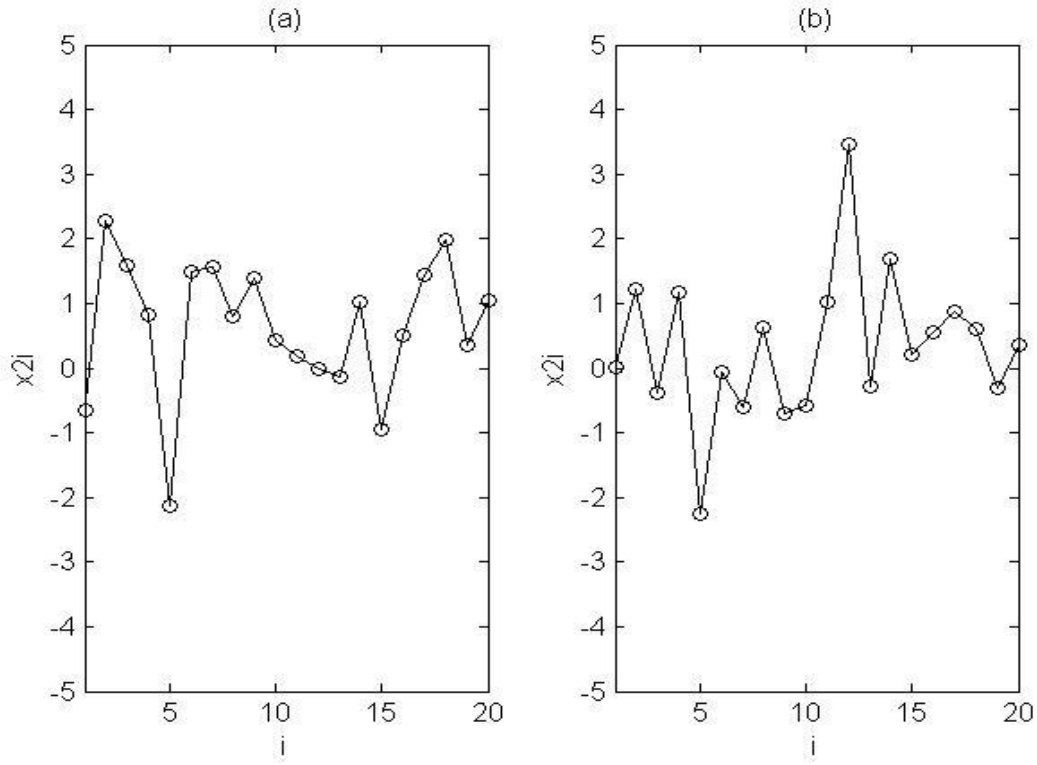

Figure 1: Graphs of sample data in Table 1 (a) and (b)

From Figure 1(a), the lowest observation i.e. value at $i = 5$ (-2.13904) appears to be an outlying observation. To confirm this, the value is subjected to Grubbs test for outliers [14]. Using the Grubbs T_1 test statistic

$$T_1 = \frac{(\bar{x} - x_1)}{s} \tag{5}$$

where $\bar{x}$ is the sample mean, x_1 is the lowest observation and s is the sample standard deviation.

The mean is 0.65106 and the sample standard deviation is 1.0724, thus substituting into (5), the statistics

$$T_1 = \frac{0.65106 - (-2.13904)}{1.0724} = 2.60173 \quad (6)$$

For n = 20, the Grubbs critical value for 5% level of significant is 2.56 which is less than 2.60173. Hence, it can be concluded that the value -2.13904 is an outlier.

In Figure 1b, two values (-2.2555 and 3.46549 at $i = 5$ and 12 respectively) appear to be outlying observations. Since these values are the lowest and greatest observations and no independent estimate of the variance is available, the ratio of range to sample standard deviation test of David et al [15] is used. The test statistic is

$$\frac{w}{s} = \frac{x_n - x_1}{s} \quad (7)$$

where w, x_n, x_1 and s are sample range, greatest observation, lowest observation and sample standard deviation respectively.

The sample standard deviation is 1.15083, thus substituting into (7), the statistic

$$\frac{w}{s} = \frac{3.46549 - (-2.2555)}{1.15083} = 5.4254 \quad (8)$$

For n = 20, the critical value for 5% level of significant as tabulated by David et al is 4.49 which is less than 5.4254. Thus, both extremes are either outliers or one of the extreme is an outlier. Since the sample mean of the observations is 0.33215, these extremes are 3.13334 units above and 2.58767 units below it. The extremes are not symmetric about the mean and the greatest observation (3.46549) is farther therefore it's either both are outliers or only 3.46549 is an outlier. That 3.46549 is an outlier can be verified by the use of Grubbs T_n statistic below.

$$T_n = \frac{(x_n - \bar{x})}{s} \quad (9)$$

The statistic $T_n = 2.72267$ and it is less than the critical value for 5% level (2.56). Since it has been confirmed that 3.46549 is an outlier, the remaining 19 observations are considered as new sample. Using (5) as illustrated above, the lowest observation (-2.2555) is also an outlier.

3.0 Methodology

Simultaneous equation estimators are based on assumptions, such as normality assumption which gives no room to most problems (such as multi-collinearity, autocorrelation, errors of measurements, problems of outlier and so on) that are associated with most real life econometric data [16]. To examine the effect of any of these problems on the estimators, there is a need to isolate the problem from the rest. This is impossible without Monte Carlo experiment where data synonymous to real life data but devoid of other problems could be generated. Thus in this work, Monte Carlo method was employed in investigating the performances of the estimators.

The parameters B and Γ of model (1) were fixed as

$$\mathrm{B} = \begin{pmatrix} 1 & -1.5 \\ -1.8 & 1 \end{pmatrix}, \ \Gamma = \begin{pmatrix} 1.2 & 0 & 0 \\ 0 & 0.5 & 2.0 \end{pmatrix} \quad (10)$$

The unequal variability of the endogenous variable (as it's often the case with most real life data) was introduced by assuming that the variance covariance matrix is given by

$$\Omega = \begin{pmatrix} 5.0 & 2.5 \\ 2.5 & 3.0 \end{pmatrix} \quad (11)$$

To generate the endogenous variable with the above variance-covariance structure, there is a need to decompose (11) into upper triangular matrix P_1 or lower triangular matrix P_2. Such that $\Omega = P_k P_k'$, $k = 1, 2$. Both cases were examined.

I Suppose

$\Omega = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$ is decomposed using an upper triangular matrix $P_1 = \begin{pmatrix} \omega_{11} & \omega_{12} \\ 0 & \omega_{22} \end{pmatrix}$, i.e. when $k = 1$. Then,

$$\begin{pmatrix} \omega_{11} & \omega_{12} \\ 0 & \omega_{22} \end{pmatrix} \begin{pmatrix} \omega_{11} & 0 \\ \omega_{12} & \omega_{22} \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}. \quad (12)$$

Multiplying (the matrices on the LHS), equating the elements and simplifying yields

$$\omega_{11} = \sqrt{\sigma_{11} - \omega_{12}^{\,2}}$$

$$\omega_{12} = \frac{\sigma_{12}}{\omega_{22}} \quad (13)$$

$$\omega_{22} = \sqrt{\sigma_{22}}$$

then if Ω is as given by (11), performing backward substitution on (13) gives

$$P_1 = \begin{pmatrix} 1.707825128 & 1.443375673 \\ 0 & 1.73205808 \end{pmatrix} \quad (14)$$

II Similarly, for $k = 2$, Ω is decomposed using a lower triangular matrix $P_2 = \begin{pmatrix} \omega_{11} & 0 \\ \omega_{21} & \omega_{22} \end{pmatrix}$, such that $\Omega = P_2 P_2'$, then P_2 is obtained as

$$P_2 = \begin{pmatrix} 2.236067977 & 0 \\ 1.118033989 & 1.322875656 \end{pmatrix}$$

The generation of the endogenous variable is in 2 stages. Firstly the disturbance terms are generated using the triangular matrix above. This is achieved using equation (15) below.

$$U = P_k E \quad (15)$$

where $\mathrm{E} = \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}$ is Gaussian distributed random error.

Finally, the exogenous variables, the disturbance terms and the fixed parameters were used to generate the endogenous variables using the reduced form equation below.

$$y = \mathrm{B}^{-1}\Gamma x + \mathrm{B}^{-1}U \quad (16)$$

The experiment is performed using sample sizes *n* = 20, 40, 60, 80, 100 and replicated 1000 times each.

4.0 Results

The results of the experiment are evaluated using three criteria, namely: Total Absolute Bias (TAB), Variance and Root Mean Square Error (RMSE). The results are examined across the sample sizes when x_2 is convoluted and when it is not convoluted and for lower and upper triangular matrices. Using TAB, the parameter estimates of OLS are inconsistent for the case of no convolution but decrease consistently with increase in sample size for the case when the variable x_2 is convoluted for the over-identified equation. For the parameters of the just-identified equation, OLS is inconsistent for both cases of when x_2 is convoluted and when it is not convoluted. The parameter estimates of OLS are smallest for the just identified equation for convoluted x_2 using TAB. 2SLS and 3SLS consistently decrease for the over-identified equation but are inconsistent for the just identified equation. LIML is inconsistent for both equations but possesses the lowest TAB for the just identified equation when no variable is convoluted.

Surprisingly, the results obtained with the convolution tend to have lower TAB than those without convolution. Using lower triangular matrix, though the performances of the estimators are similar to that of the upper triangular matrix, but the TAB values are higher. Table 2 shows the ranking of the estimator.

Table 2: Results Using Total Absolute Bias

Equation	Upper triangular Matrix		Lower Triangular Matrix	
	No Convolution	Convoluted	No Convolution	Convoluted
Just Identified	3SLS	OLS	LIML	OLS
	LIML	3SLS	3SLS	3SLS
	2SLS	2SLS	2SLS	LIML
	OLS	LIML	OLS	2SLS
Over-Identified	2SLS	2SLS	2SLS	2SLS
	3SLS	3SLS	3SLS	3SLS
	LIML	OLS	LIML	LIML
	OLS	LIML	OLS	OLS

Using Variance in evaluating the estimators, OLS possesses the smallest variance of all estimators though inconsistent for the parameters of the endogenous variables. For all the parameters, excluding the parameters of the exogenous variables in the just-identified equation, variance of OLS when no variable is convoluted is smaller than when an exogenous variable is convoluted. When lower triangular matrix is used, the variance OLS still remains the smallest and it consistently decreases as the sample size increases. Also, with lower triangular matrix, the variance of the estimators when no variable is convoluted is smaller than the variance when an exogenous variable is convoluted for both equations. The variance obtained with using the lower triangular matrix is considerably smaller than that obtained using the upper triangular matrix.

Table 3: Results Using Variance

Equation	Upper triangular Matrix		Lower Triangular Matrix	
	No Convolution	Convoluted	No Convolution	Convoluted
Just Identified	OLS	OLS	OLS	OLS
	2SLS	2SLS	2SLS	2SLS
	LIML	3SLS	LIML	LIML
	3SLS	LIML	3SLS	3SLS
Over-Identified	OLS	OLS	OLS	OLS
	2SLS	2SLS	2SLS	2SLS
	3SLS	3SLS	3SLS	3SLS
	LIML	LIML	LIML	LIML

Using the RMSE, OLS possesses the smallest RMSE for all the estimators except for the over-identified equation when an exogenous variable is convoluted. The RMSE obtained with an exogenous variable convoluted is quite in the same range as with a variable convoluted but the performances using the lower triangular matrix is better than that of the upper triangular matrix for both equations.

Table 4: Results Using RMSE

Equation	Upper triangular Matrix		Lower Triangular Matrix	
	No Convolution	Convoluted	No Convolution	Convoluted
Just Identified	OLS	OLS	OLS	OLS
	2SLS	2SLS	2SLS	2SLS
	LIML	LIML	LIML	3SLS
	3SLS	3SLS	3SLS	LIML
Over-Identified	OLS	OLS	OLS	2SLS
	2SLS	2SLS	2SLS	3SLS
	3SLS	3SLS	3SLS	OLS
	LIML	LIML	LIML	LIML

5.0 Conclusion

Since most real life data comes from a distribution with heavy tail or longer tail, we assumed a convolution of the normal and uniform with the hope to obtaining a flat-tailed distribution which may possibly contain outlier. A mixed model of one over-identified and just-identified equation was assumed as most practical models are of mixed status.

A Monte Carlo experiment was performed to examine performances of four estimators of parameters of simultaneous equation econometric models namely; OLS, 2SLS, LIML and 3SLS under the condition described above. The estimates were subjected to 3 evaluation criteria. The performances of the estimators when lower triangular matrix is used are better than that of upper triangular matrix. The performance of OLS using TAB as evaluation criterion is better than those of the other

estimators when an exogenous variable is convoluted for the just-identified equation while the performance of 2SLS is best for the over-identified equation. Using variance as evaluation criterion, OLS possesses the least variance for both equations and both matrices while LIML has the worst variance in most cases. Using RMSE, OLS possesses the smallest RMSE for both matrices and equations except with the over-identified equation using lower triangular matrix when an exogenous variable is convoluted.

It must be stressed that the performances varies depending on the criteria employed. A bias estimator with lower variance may eventually be the best estimator using RMSE since variance could compensate for the larger bias.

References

[1] Huber, P. J. (1981), Robust Statistics. John Wiley & Sons Ltd, New York.

[2] Maronna, R. A., Douglas, R. M. and Yohai, V. J. (2006), Robust Statistics: *Theory and Methods*. John Wiley & Sons Ltd, West Sussex.

[3] Osborne, J. W., Christiansen, W. R. I. and Gunter, J. S. (2001). Educational psychology from a statistician's perspective: *A review of the quantitative quality of our field.* Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

[4] Aggarwal, C. C. and Yu, P. S. (2001), Outlier Detection for High Dimensional Data, In: Proceeding of the ACM SIGMOD Conference 2001

[5] Rey, W. J. J. (1978), Lecture Notes in Mathematics (Robust Statistical Methods), Springer-Verlag, Belin Heidelberg, New York.

[6] Maruyama, Y. And Strawderman, W. E. (2010), Improved robust Bayes estimators of the error variance in linear models, Arxiv [math.ST], 1004.0234v1.

[7] West, M. (1984), Outlier Models and Prior distributions in Bayesian Linear Regression. Journal of Royal Statistical Society, Series B (Methodological), Vol 46, No 3, pp.431-439

[8] Brown, G. W. and Mood, A. M. (1951), On median tests for linear hypothesis, In: Proceedings of the second Berkeley Symposium on Mathematical Statistics and Probability, Univ. of California Press, Berkeley. Pages 159-166.

[9] Huber, P. J. (1964), Robust Estimator of a Location Parameter, The Annals of Mathematical Statistics, Vol.35, No. 1, pp73-101.

[10] Yohai, V. J. (1987), High Breakdown Point and High Efficiency Robust Estimates for Regression, The Annals of Statistics, Vol. 15, No. 2, pp642-656.

[11] Yohai, V. J. and Zamar, R. H. (1988), High Breakdown Point Estimates of Regression by Means of the Minimization of an Efficient Scale. Journal of American Statistical Association. Vol. 83, No. 42, pp 456-413.

[12] Mishra, S. K. (2008), Robust Two-Stage Least Square: Some Monte Carlo Experiments. Journal of Applied Economic Sciences, Issue 4(6), Vol. 3 (4(6)): 434-443.

[13] Kmenta, J. K. (1971), Elements of Econometrics, MacMillian Press Ltd, New York.

[14] Grubbs, F. E. (1969), Procedures for detecting outlying observations in samples. Technometrics, Vol. 11, No. 1, pp 1-21.

[15] David, H. A., Hartley, H. O. and Pearson, E. S. (1954), The distribution of the ratio in a single sample of range to standard deviation. Biometrics, Vol. 41, pp482-493.

[16] Adepoju, A. A. (2009), Comparative Assessment of Simultaneous Equation Techniques to Correlated Random Deviates. European Journal of Scientific Research, Vol. 28 No .2, pp. 253-265.

[17] Gujarati, D. N. (2004), Basic Econometric Methods (Fourth Edition), McGraw-Hill, New York.

[18] Johnston, J. and Dinardo, J. (1998), Econometric Methods (Fourth Edition), McGraw-Hill, New York.

[19] Kmenta, J. and Gilbert, R. F. (1967), Small Sample properties of Alternative Estimators of Seemingly Unrelated regressions. Journal of the American Statistical Association, Vol. 63, 1180-1200.

[20] Koutsoyiannis, A. (2003), Theory of Econometrics (Second Edition), Palgrave New York.

[21] Maronna, R. A. and Yohai, V. J. (1994), Robust Estimation in Simultaneous Equation Models. Statistics and Econometrics, Series 14, Working Paper 94-37.

[22] Oseni, B. M. and Adepoju, A. A. (2011), Assessing the Performances of Simultaneous Equation Estimators under the Influence of Outliers. Journal of Nigerian Statistical Association, Vol. 23, pp. 1-8.