JOURNAL OF THE NIGERIAN STATISTICAL ASSOCIATION VOL. 23, 2011 1-9

Assessment of Simultaneous Equation Techniques under the Influence of Outliers

[1]B. M. Oseni and [2]A. A. Adepoju

[1]Department of Mathematical Sciences, Federal University of Technology, Akure, Nigeria

[2]Department of Statistics, University of Ibadan, Ibadan, Nigeria

Abstract

Most simultaneous equations estimation techniques are based on the assumptions of normality which gives little consideration to some atypical data often called outliers which may be present in the observations. The outliers may have some obvious distorting influence on the estimates produce by these techniques. This study investigates the distorting effect of outliers on four simultaneous equation estimation techniques through Monte Carlo method. Outliers of various degrees were introduced into observations of different sizes. The estimators were ranked based on their ability to absorb the shock due to outliers in the observations. The Total Absolute Bias (TAB), Variance and Root Mean Square Error (RMSE) were used in ranking the performances of the estimators. Based on the criterion of tab, two stages least squares (2SLS) ranked the best, closely followed by three stage least squares (3SLS) and ordinary least squares (OLS) in that order, while limited information maximum likelihood (LIML) was the poorest when outliers of not more than 5% are present in the observation. It is however, not strange to observe that OLS outperformed the other estimators when variance was used. This could be misleading since variance may be measured around a wrong parameter. Based on the criterion of RMSE, ordinary least squares yields estimates with the least value of RMSE while LIML yields the greatest when outliers of not more than 10% are present in the observation. Also it was established that OLS has the greatest capacity to absorb the shock due to the presence of outliers in the observation.

Key words: Monte Carlo, Outliers, Estimators, Simultaneous Equation.

Introduction

Economic relationships are often of two-way dependency and may be conveniently represented using simultaneous equation models, thus simultaneous equations are important in econometrics. These equations in practice are mostly of mixed type i.e. comprising of mixture of just-identified and over-identified equations. A number of techniques have been devised to handle such equations among which the most common are the ordinary least square (OLS), Indirect Least squares (ILS), two stage least square (2SLS), three stage least square (3SLS), limited information maximum likelihood (LIML) and full information maximum likelihood (FIML). These techniques are based on the assumption that the random deviates are normally distributed. The assumption of normality gives little consideration to atypical data which are often found in most real life data [5, 11]. In reality, the behavior of many sets of data is only approximately normal, with the main discrepancy being that a small portion of the observations were quite atypical by virtue of being far from the bulk of the data. Such atypical data often called outliers, always come from the kind of approximately normal distribution with normal shape in the central region, but with heavier or fatter tails [5, 11]. Researches in recent times have shown that the presences of outliers in data set are not often tested by most statisticians [15, 18, 19], thus data set containing outliers are not always cleaned up before statistical techniques are applied. Outliers may have some distorting influence on the estimates produced by the estimators especially under the assumption of normality. The performances of these estimators in the presence of outliers should be of utmost importance. In this study, the performances of OLS, 2SLS, 3SLS and LIML were investigated. FIML is not considered because it generates outrageous estimates when outliers are presence in the data set and ILS; which do not have unique estimate for the over-identified equation of the model; is also not considered. The evaluation and comparisons of the techniques are carried out using total absolute bias (TAB), variance and root mean square error (RMSE).

Materials and Methods

Consider the simultaneous equation model consisting of over identified equation and just identified equation given below:

$$y_{1t} = \beta_{12}y_{2t} + \gamma_{11}x_{1t} + u_{1t}$$
$$y_{2t} = \beta_{21}y_{1t} + \gamma_{22}x_{2t} + \gamma_{23}x_{3t} + u_{2t}$$

(1)

This model can be represented in matrix form as

$$\mathrm{B}y_t = \Gamma x_t + U_t$$

(2)

Where: $y_t = \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix}$ is the vector of endogenous variables.

$x_t = \begin{pmatrix} x_{1t} \\ x_{2t} \end{pmatrix}$ is the vector of exogenous variables.

$U_t = \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}$ is independent and identically distributed (i.i.d.) random vector with mean zero and covariance matrix Σ.

$\mathrm{B} = \begin{pmatrix} 1 & -\beta_{12} \\ -\beta_{21} & 1 \end{pmatrix}$ is the matrix of coefficients of endogenous variables.

$\Gamma = \begin{pmatrix} \gamma_{11} & 0 & 0 \\ 0 & \gamma_{22} & \gamma_{23} \end{pmatrix}$ is the matrix of coefficient of exogenous variables.

Pre-multiplying both sides of (2) by B^{-1} yields

$$y_t = \mathrm{B}^{-1}\Gamma x_t + \mathrm{B}^{-1}U_t$$

(3)

Expanding and simplification of (3) gives

$$y_{1t} = \frac{\gamma_{11}}{1-\beta_{12}\beta_{21}}x_{1t} + \frac{\beta_{21}\gamma_{22}}{1-\beta_{12}\beta_{21}}x_{2t} + \frac{\beta_{21}\gamma_{23}}{1-\beta_{12}\beta_{21}}x_{3t} + \left(\frac{1}{1-\beta_{12}\beta_{21}}u_{1t} + \frac{\beta_{21}}{1-\beta_{12}\beta_{21}}u_{2t}\right)$$
$$y_{2t} = \frac{\beta_{12}\gamma_{11}}{1-\beta_{12}\beta_{21}}x_{1t} + \frac{\gamma_{22}}{1-\beta_{12}\beta_{21}}x_{2t} + \frac{\gamma_{23}}{1-\beta_{12}\beta_{21}}x_{3t} + \left(\frac{\beta_{12}}{1-\beta_{12}\beta_{21}}u_{1t} + \frac{1}{1-\beta_{12}\beta_{21}}u_{2t}\right)$$

(4)

The data used in the study are generated using Monte Carlo method. The exogenous variables are obtained from Uniform(0,1) distribution[8] using the standard random number generator. To generate the random disturbance $u_{it}(i=1,2;t=1,2,3,\ldots,n)$ with mean zero and covariance matrix Σ, independent series of uncorrelated standard normal random deviates $e_{it}(i=1,2;t=1,2,3,\ldots,n)$ of required sample size n are generated. These generated random deviates are transformed to be distributed as $N(0,\Sigma)$, using a predetermined covariance matrix $\Omega=\begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}$ such that $\Sigma=E(UU')=\Omega\otimes I_n$ where var(u_{1t}) = σ_{11}, var(u_{2t}) = σ_{22} and cov(u_1u_2)=σ_{12}[13]. The transformation requires that the matrix Ω be decomposed by a non-singular triangular matrix P such that $\Omega=PP'$[1, 2]. Assuming $P=\begin{pmatrix}a_{11} & a_{12}\\ 0 & a_{22}\end{pmatrix}$ i.e P is an upper triangular matrix, then transforming using $U_t=P\begin{pmatrix}e_{1t}\\ e_{2t}\end{pmatrix}$ yields

$$\left.\begin{aligned} u_{1t} &= a_{11}e_{1t}+a_{12}e_{2t}\\ u_{2t} &= a_{22}e_{2t}\end{aligned}\right\}\qquad t=1,2,\ldots,n$$

(5)

where $a_{22}=\sqrt{\sigma_{22}}$, $a_{12}={}^{\sigma_{12}}\!/_{a_{22}}$ and $a_{11}=\sqrt{\sigma_{11}-a_{12}^2}$.

If P is a lower triangular matrix, then the transformation is done using

$$\left.\begin{aligned} u_{1t} &= a_{11}e_{1t}\\ u_{2t} &= a_{12}e_{1t}+a_{22}e_{2t}\end{aligned}\right\}\qquad t=1,2,\ldots,n$$

(6)

where $a_{11}=\sqrt{\sigma_{11}}$, $a_{12}={}^{\sigma_{12}}\!/_{a_{11}}$ and $a_{22}=\sqrt{\sigma_{22}-a_{12}^2}$

The endogenous variables are generated using (4). For the purpose of the Monte Carlo data generation, the parameters of (2) are assigned as follows:

$\mathrm{B}=\begin{pmatrix}1 & -1.5\\ -1.8 & 1\end{pmatrix}, \Gamma=\begin{pmatrix}1.2 & 0 & 0\\ 0 & 0.5 & 2.0\end{pmatrix}$ while the predetermined covariance matrix $\Omega=\begin{pmatrix}5.0 & 2.5\\ 2.5 & 3.0\end{pmatrix}$.

To produce data set with such behaviour as described above, random deviates with approximately normal distribution are generated. These are obtained with the use of Tukey's contamination neighbourhood defined as follows:

$$\mathcal{F}_\delta(F_\theta)=\{F:F=(1-\delta)F_\theta+\delta F^*\}, \quad 0<\delta<1/2$$

(7)

where $\mathcal{F}_\delta$ is the approximately normal distribution with contamination level δ, F_θ is the central distribution and F^* is the contaminating distribution [6, 16]. Standardized random deviates (with the central distribution $N(0,1)$) of the sample sizes n = 20, 25 and 30 are generated and contaminated with samples from $N(10,1)$. The contamination is achieved with the aid of (7) by setting δ= 0.00, 0.05, 0.10. These produced random deviates with 0%, 5% and 10% outliers respectively. The contaminated random deviates then are transformed into stochastic disturbances with the aid of (5) and (6). These stochastic disturbances together with the exogenous variables and the assumed values of the parameters are used to generate the endogenous variables from (4). The Monte Carlo experiment is replicated 1000 times.

Analysis and Discussion of Results

The results obtained from the Monte Carlo experiment are compared using the criteria of total absolute bias (TAB), variance and root mean square error (RMSE). Tables 1-3 give the ranking of the results obtained. Table 1 presents the ranking of the techniques based on the results obtained using total absolute bias. On this criterion of tab, the best method of estimating the parameters of the over-identified equation, equation 1, is the 2sls. It ranks best irrespective of the percentage of outliers present in the sample (not more than 10% as considered in this work), closely followed by 3sls with identical estimates. LIML ranks the poorest at all contamination level using lower triangular matrix P_2 as in (6) but improves as the level of contamination increases if upper triangular matrix P_1 is used as in (5) making OLS the poorest at 10% contamination level. OLS ranks best in estimating the just-identified equation (equation 1) using P_1 while 2sls ranks best for outliers less than 5% in the sample using P_2. Though estimates of LIML, 2sls and 3sls are identical, at 0% contamination level, LIML ranks poorest closely followed by 2sls while under the influence of outliers 3sls ranks poorest closely followed by 2sls when P_1 is used. With P_2, LIML ranks poorest when no outlier is present while with outliers less than 5% OLS ranks poorest closely followed by LIML. At 10% contamination level, OLS ranks best followed by LIML though with identical estimates as 2sls and 3sls.

TABLE 1: Ranking of Estimators Using TAB

	Equation 1			Equation 2		
	0%	5%	10%	0%	5%	10%
P_1	2,3SLS	2,3SLS	2,3SLS	OLS	OLS	OLS
	OLS	OLS	LIML	2,3SLS	LIML	LIML
	LIML	LIML	OLS	LIML	2,3SLS	2,3SLS
P_2	2,3SLS	2,3SLS	2,3SLS	OLS	2,3SLS	OLS
	OLS	OLS	OLS	2,3SLS	LIML	LIML
	LIML	LIML	LIML	LIML	OLS	2,3SLS

Table 2 shows the ranking of techniques based on the results obtained using variance. In all cases, with both equations and triangular matrices, OLS ranks the best. 2SLS and 3SLS with identical estimates followed OLS while LIML ranks poorest with the over-identified equation (equation 1) irrespective of the triangular matrix used. When the lower triangular matrix P_2 is used in decomposing the variance-covariance matrix, with the just-identified equation, 3SLS ranks 2nd closely followed by 2SLS at 0% contamination level. But as contamination level increases to 5% and above LIML ranks 2nd while 3SLS is the poorest.

TABLE 2: Ranking of Estimators Using Variance

	Equation 1			Equation 2		
	0%	5%	10%	0%	5%	10%
P_1	OLS	OLS	OLS	OLS	OLS	OLS
	2,3SLS	2,3SLS	2,3SLS	2,3SLS	2,3SLS	2,3SLS
	LIML	LIML	LIML	LIML	LIML	LIML
P_2	OLS	OLS	OLS	OLS	OLS	OLS
	2,3SLS	2,3SLS	2,3SLS			

	LIML	LIML	LIML	2,3SLS	LIML	LIML
				LIML	2,3SLS	2,3SLS

Table 3 summarizes of the ranking of the techniques on the criterion of RMSE. In all cases (i.e. at all contamination level using either of the two triangular matrices) OLS is the best method. When estimating the parameters of the over-identified equation, 2SLS and 3SLS with identical estimates are 2nd and 3rd respectively while LIML is the poorest irrespective of the triangular matrix used. When outliers not more than 10% are present in the sample, with the parameters of the just-identified equation, 3SLS is the poorest when P_2 is used. When P_1 is used 2SLS becomes the poorest at 10% contamination level while 3SLS is the poorest at 5% contamination level.

TABLE 3: Ranking of Estimators Using RMSE

	Equation 1			Equation 2		
	0%	5%	10%	0%	5%	10%
P_1	OLS 2,3SLS LIML	OLS 2,3SLS LIML	OLS 2,3SLS LIML	OLS 2,3SLS LIML	OLS LIML 2,3SLS	OLS LIML 2,3SLS
P_2	OLS 2,3SLS LIML	OLS 2,3SLS LIML	OLS 2,3SLS LIML	OLS 2,3SLS LIML	OLS LIML 2,3SLS	OLS LIML 2,3SLS

Conclusion

The choice of estimation technique depends on many factors some of which are the purpose for which the estimation is embarked on, the identification condition of the equations of the model, the presence of other endogenous variable among the set of explanatory variables in any particular equation, the computational complexity of the problem, the importance which the researcher attributes to various statistical properties of the parameter estimates and even the presence of outliers in the observations. Under these factors, the researcher (or policy maker) is faced with the problem of choosing the estimator which will most likely yield the desired estimates. Obviously, the choice is conventionally based on the statistical properties possessed by the estimates of various methods, thus the methods are ranked based on the small sample properties. The RMSE, though may be influenced by either bias or variance but it is necessary as inference cannot be made on only variances or absolute bias. This is because a technique may have the smallest variance around the wrong estimate and conversely a small bias with the largest variance. Though the three criteria have been used in the investigation, but much depends upon the purpose for which the estimation is embarked upon.

Reference

1 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.

2 Adepoju A. A. and Olaomi J.O. (2009), Ranking of simultaneous equation Techniques to small sample properties and correlated random deviates, Journal of Mathematics and Statistics, Vol. 5 No. 4, pp 260-266.

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

4 Hendry D. F. (1971), The Structure of Simultaneous Equation Estimators. Journal of Econometrics, Vol 4 No 1, pp 51-88.

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

6 Jafar A. K. (2002), Globally Robust Inference for Simple Linear Regression Models with Repeated Median Slope Estimator. MSc thesis, University of British Columbia, Canada.

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

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

9 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.

10 Koutsoyiannis A. (2003), Theory of Econometrics (Second Edition), Palgrave New York.

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

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

13 Nagar A.L. (1960), a Monte Carlo Study of Alternative simultaneous equation estimators. Econometrica Vol. 28, 573-590.

14 Newey W.K. (1987), Efficient estimation of limited dependent variable models with endogenous explanatory variables, Journal of econometrics, vol. 36, issue3, pg 231-250, doi:10.1016/0304-4076(87)90001-7

15 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.

16 Tukey, J. (1960). A Survey of Sampling from Contaminated Distribution: *Contributions to Probability and Statistics*. I. Olkin, Ed., Stanford University Press, Stanford.

17 Wagner H.M. (1958), A Monte Carlo study of estimates of simultaneous linear structural equation. Econometrica, 26: 117-133. http:/www.jstor.org/stable/1907386

18 Zimmerman, D. W. (1994), A note on the influence of outliers on parametric and nonparametric tests. *Journal of General Psychology, 121*(4), 391-401.

19 Zimmerman, D. W. (1995), Increasing the power of nonparametric tests by detecting and downweighting outliers. *Journal of Experimental Education, 64*(1), 71-78.