Journal of the Nigerian Association of Mathematical Physics
Volume **11 (November 2007), 627 - 634**

Contributions of influence function using the inverse autocorrelation function in the detection of outliers

N. P. Olewuezi
Department of Mathematics and Computer Science
Federal University of Technology, Owerri, Nigeria.
e-mail: ngolewe@yahoo.com. 08034933133

Abstract

Outliers in time series, depending on their nature may have a moderate to significant impact on the effectiveness of the standard methodology for time series analysis with respect to model identification, estimation and forecasting. The suggested procedure used for identifying the outliers graphically in time series data was investigated by considering the influence function for the inverse autocorrelation function (IACF). Form the findings, it was noticed that for large series the influence was almost positive in values while for relatively short series the large negative influence are noticeable. The model order determination technique was also proposed.

***Keywords*:** outliers; inverse autocorrelation function; influence function; order determination

1.0 Introduction

Time series data are often subject to uncontrolled or unexpected interventions from which various types of outlying observations are produced. An outlying observation is one that appears to deviate markedly from the other members of the sample in which it occurs. Outliers can take several forms in time series. Fox (1972) [10] proposed the formal definitions and a classification of outliers in time series context. He proposed a classification of time series outliers to type I and type II. These two types have later been renamed as additive and innovational outliers. For a properly deduced stationary process, let X_t be the observed series and Z_t be the outlier-free series. Consider a familiar time series model.

$$\Pi(B)z_t = a_1 \qquad (1.1)$$

where $\Pi(B)z = 1_1 B - \Pi_2 B^2 - \cdots$.

$\{\alpha_t\}$ is a sequence of independently distributed normal variables with zero mean and variance σ^2. The function Π(B) is often expressed as a ratio of

$$\frac{\phi(B)}{\vartheta(B)}$$

where

$\phi(B) = 1 - \phi_1 B \cdots \phi_\rho B^\rho$

and $\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$

are stationary and invertible operators sharing no common factors. The models commonly employed on the outlier-free time series Z_t are the additive outlier (AO) and the innovational outlier (IO) which are defined respectively of a single outlier for a simple cases as

$$x_t = z_t + D\xi_t^{(T)} \tag{1.2}$$

and

$$x_t = z_t + \left(\frac{\theta(B)}{\phi(B)}\right) D\xi_t^{(T)} \tag{1.3}$$

where x_t is the observed series, D is the magnitude of the outlier and $D\xi_t^{(T)} = 1$, if $t = T$ and 0 otherwise, which is the time indicator signifying the time occurrence of the outlier.

In general, the presence of more than one outlier of various types in a model is specified by

$$x_t = \sum_{k=1}^{n} V_k(B) D_k \xi_t^{(T)} + Z_t \tag{1.4}$$

where

$$z_t = \theta(B)\phi^{-1}(B)a_t$$
$$V_k(B) = 1 \text{ for AO model}$$
$$V_k(B) = \frac{\theta(B)}{\phi(B)} \text{ for IO model}$$

at time $t = T_k$ and n is the number of outliers.

Fox (1972) [10] made no methodological proposals to distinguish between the two basic types of outliers. This is remedied in the work of Muirhead (1986) [13] who presents a text of discordancy for a single outlier of unknown type and proceeds to examine the properties of an appealing likelihood–ratio based rule for distinguishing whether the outlier is of AO or IO type. He also compares this rule with corresponding Bayesian procedure. When the timing and type of an outlier are unknown, Abraham and Box (1979) [1] proposed a Bayesian approach, Martin (1980) 9[12], robust method and Chang and Tiao (1983) [4] an iterative procedure for resolving the estimation problem. An empirical study of the detection of outliers in time series of body temperature of cows is described by de Alba and Zartman (1980) [9]. They employ a model similar to that used by Stoodley and Mimia (1979) but incorporated the specific prospect of innovational outliers. Shangodoyin (1993) [16] found out that the estimates of the magnitude of outliers for additive model favoured the comments of Cook (1979) [8] on the sequential deletion of outliers and residual correlations.

2.0 Inverse autocorrelation function (IACF)

Cleveland (1972) [7] introduced the concept of the IACF. He defined the inverse autocovariances as the associated with the inverse of the spectral density of the series which Parzen (1974) [14] called the inverse spectral density. That is, let for the discrete stationary process $\{X_t\}$,

$$f_i(w) = \{f(w)\}^{-t}$$

be integrable on the interval (0, 1). The inverse autocovariances of $\{X_t\}$, are defined by

$$\gamma_i(k) = \int_{-\pi}^{\pi} e^{jkw} f_i(w)dw \tag{2.1}$$

and

$$\gamma_i(k) = \gamma_i(-k), \ k = 0, 1, \ldots$$

The inverse autocorrelations are defined by $\rho_i(k) = \dfrac{\gamma_i(k)}{\gamma_i(0)},\ \ k = 0, 1, \cdots.$

As suggested by Clevenland (1972) [7] the two methods of estimating the IACF stem from either the autoregressive method or the window method. Chatfield (1979) [5] gave the time domain definition of the IACF. Olewuezi and Shangodoyin (2005) [15] derived an orthogonal relationship between Autocorrelation function (ACF) and IACF and defined the inverse autocovariance function at lag *k* by

$$\gamma_i(k) = \begin{cases} \dfrac{2\pi^2}{\gamma_0},\ k = 0 \\ 4\pi \sum\limits_{j=1}^{k} \left(\dfrac{1}{j}\right)\left(\dfrac{1}{\gamma_j}\right) \sin\ 2\pi j,\ \ k \neq 0 \end{cases}$$

where γ_k is the autocovariance functions at lag k and γ_0 is the variance of the process.

The IACF of a time series are useful at the identification stage of model building. In practice this quantity must be estimated from the data. The IACF of an Autoregressive (AR) process cuts off at lag *k*. It turns out that IACF has similar properties to the Partial Autocorrelation function (PACF).

3.0 The influence function

This is a convenient tool for studying both outlier and robust estimation. An influence function for an estimate is the result of an infinitesimal change in the weight given to an observation in the theoretical distribution function. It depends on the parameters being estimated, the observation vector whose influence is measured and the distribution function of that observation vector. The parameter θ can be expressed as a function of the distribution function F and is written as T(F). The influence function as defined by Hampel (1974) [11] is given by

$$I(F, x, T(F)) = \lim_{\in \to 0} (R - T(F))/\in,$$

where

$$R = T(I - \in)F + \in_{\sigma_x}, x$$

is the point of interest in the observation space, $\in$ is a position number and σ_x is the distribution function that has all its probability mass concentrated at the point *x*.

Chernick et al (1982) [6] comment that outliers (of unspecified type) can seriously distort estimates of autocorrelations is stationary time series. They considered a discrete time series $x_1, x_2, \ldots x_n$ choosing a fixed number *m* of lags (with *m* much smaller than *n*), we are advised to consider an *nxm* matrix

$$\{I\lfloor H, \rho(k)(y_j, y_{j+k})\rfloor\}$$

where y_i is the standardized observation $(x_i - \mu)/\sigma$ (with μ and σ the mean and standard deviation of x_i independent of y_i and I(.) is the influence function, which the authors show to have the simple form.

$$y_i y_{j+k} - \rho(k)\left(y_j^2 + y_{j+k}^2\right)/2$$

Using the result employed by Shangodoyin (1993) in his study of the influence function of the autocorrelation function and with the transformation of the $y_i y_{j+k}$ to the Gaussian processes

$$U_{j,k,1} = \tfrac{1}{2}\left\lfloor \left(y_j + y_{j+k}\right)/\sqrt{1+\rho(k)} + \left(y_j - y_{j+k}\right)/\sqrt{1-\rho(k)} \right\rfloor$$

and
$$U_{j,k,2} = \tfrac{1}{2}\left\lfloor \left(y_j + y_{j+k}\right)/\sqrt{1+\rho(k)} - \left(y_j - y_{j+k}\right)/\sqrt{1-\rho(k)} \right\rfloor$$

we have
$$\left\lfloor 1 - \rho^2(k) \right\rfloor U_{j,k,1} U_{j,k,2} = y_j y_{j+k} - \rho(k)\left(y_j^2 + y_{j+k}^2\right)/2 \qquad (3.1)$$

and so $$I\lfloor H,\rho(k),\ (y_j, y_{j+k})\rfloor = \left(1-\rho^2(k)\right) U_{j,k,1} U_{j,k,2} \qquad (3.2)$$

are the influence function matrix of the autocorrelation function. Extending further, the similarity between ACF and IACF as model identification tolls, equation (3.2) becomes

$$I\lfloor H,\rho_i(k),\ (y_j, y_{j+k})\rfloor = \left(1-\rho_i^2(k)\right) U_{j,k,1} U_{j,k,2} \qquad (3.3)$$

where $$U_{j,k,1} = \frac{1}{2}\left\lfloor \left(y_j + y_{j+k}\right) / \sqrt{1+\rho(k)} - \left(y_j - y_{j+k}\right) / \sqrt{1-\rho(k)} \right\rfloor$$

and $$U_{j,k,2} = \frac{1}{2}\left\lfloor \left(y_j + y_{j+k}\right) / \sqrt{1+\rho(k)} - \left(y_j - y_{j+k}\right) / \sqrt{1-\rho(k)} \right\rfloor$$

Hence, for a stationary Gaussian process with μ, σ and $\rho_I(k)$ all known, $U_{j,k,1}$ and $U_{j,k,2}$ are observations from independent standard normal distributions. The quantity $I\lfloor H,\rho_i(k),(y_j, y_{j+k})\rfloor$ has the distribution of a constant times a product of standard normal random variables. This distribution can be used to determine what values for the influence function would be unusually large for a realization from a stationary Gaussian process. Since y_i's are Gaussian, then the distribution of the influence function depends on $\rho_i(k)$.

4.0 Detection procedure

To investigate the effect of outliers on times series data we investigate the performance of the estimate used in equation (7) based on the standard error,

$$\left[\frac{n-k}{n(n+2)}\right]^{\frac{1}{2}}$$

Let n be the number of observations and k be a fixed number of lags with k considerably less than n and we consider an n x n k matrix with the (j, k) entry given by $I\lfloor H,\rho_i(k),(y_j, y_{j+k})\rfloor$ where y_i is the *jth* standard observation.

The observation y_j which is an outlier have drastic effects on estimates of the correlation coefficients which influences several lagged autocorrelation estimates. Chernick et al (1982) [6] proposed that y_j appears in the computation of every element in the jth row and also in the diagonal elements of the proceeding rows beginning in Column I of row j - I and proceeding up to the right. An outlier will often have a very large positive or negative influence on each estimate of correlation. If all the elements in the jth row and the above diagonal are large in absolute value, this will indicate that the *jth* observation is probably an outlier. To clearly see the patterns in the influence function matrix, they proposed choosing a critical value. Influence function estimates exceeding this critical value in absolute value are designated plus (+) or minus (-) depending on the sign of the estimate. Other observations are left blank. The matrix will then appear with patterns of +s and –s and the cloths–pin effect should be evident to the eye. Thus lags (k) with the highest number of blanks and for which $\rho_i(k + 1)$ cuts off are the possible order of the model. If $\rho_i(k)$ cuts off after lag (k) it indicates that the process satisfies an Autoregressive (AR) model of order k.

5.0 Illustration

The influence function matrix (IFM) as described above was applied to three sets of data. Results of the analysis of the series are presented and influence function matrices of the series are given. Series A is the Stack – loss data [Brownlee (1965) [3]]. Series B is the Gas furnace data [Box and Jenkins (1976) [2]]. Series C is the Nigerian composite consumer prince index monthly series from January 1988 to December 19991.

5.1 Series A

The IFM for the IACF is shown in Table 1 which gives large values in rows 3 and 17 which indicates that observations at these points are possible outliers. The IFM also suggest that cut off takes place only after lag I. Hence from our procedure an AR of order I will be suitable for this series when outliers at the third and seventeenth positions have been screened out.

5.2 Series B

The IFM for this series is shown in Table 2 which indicates that there are large values in rows 9 and 11 and by our procedure it shows that these observations are possible outliers. The proposed procedure identified an Autoregressive moving average (ARMA) model of order (1, 1).

5.3 Series C

The IFM is shown in Table 3 which gives large positive values in rows 1 – 6, 20 – 22 and 39 – 46 which indicates 20 outlying observations. There are cut off after lag 1. Hence an AR(1) was identified.

Table 1: IFM for Series A (Inverse Autocorrelation)

Lag Time	1	2	3	4	5
1	+	-			
2	-				
3	-	-	+	-	
4					
5					
6					
7			+	-	
8	-				
9					
10	-				
11					
12	+				+
13		-		-	
14					
15	-				
16					
17	+	-	+	-	-
18					
19				+	
20					
21			-		-

Table 2: IFM for Series B (Inverse Autocorrelation)

Lag Time	1	2	3	4	5
1	+				
2	-				
3	-				
4		+			
5				-	

Lag Time	1	2	3	4	5
6					
7	-	+			
8	+	-			
9	-	-	-	-	-
10	-				
11	-	+	-	+	-
12		+			
13		+			
14					
15					
16					
17	+		+		
18					
19					
20	-				
21					

Table 3: IFM for Series A (Inverse Autocorrelation)

Lag Time	1	2	3	4	5
1	+	+	+	+	+
2	+	+	+	+	+
3		+	+	+	+
4		+	+	+	+
5		+	+	+	+
6		+	+	+	+
7					
8			+		
⋮					
19	+				
20	+	+	+	+	+
21	+	+	+	+	+
22		+	+	+	+
23				+	
⋮					
38		+	+	+	+
39		+	+	+	+
40		+	+	+	+
41		+	+	+	+
42		+	+	+	+
43		+	+	+	+
44		+	+	+	+
45	+	+	+	+	+
46		+			
47					
48					

6.0 Results and conclusion

From the results, the influence functions that are left blank indicate low influence functions of observations while the influence functions estimated and exceeding the chosen critical value (in magnitude) are designated plus (+) or minus (-) depending on the sign of the estimates. It was noticed that for large series ($n \geq 30$) the influences are almost positive in values while for relatively short series, large negative influences are noticeable. It was also noticed that since it is not unusual that the values of the IACF, particularly at lower lags, are small so the lag with smaller influence could be regarded as the possible order of the model if the value of the IACF is significant at this point which also depends on the nature of plot of the IACF.

The influence function approach should be extended to other methods like Extended Sample Autocorrelation function (ESACF) and the Inverse Partial Autocorrelation function (IPACF) because it is appropriate to say that model identification is both a science and an art. One should not use one method to the exclusion of others.

Acknowledgement

My sincere appreciation goes to Dr. D. K. Shangodoyin of the Department of Statistics, University of Ibadan for his immense contribution towards my educational attainment. I am also indebted to my beloved husband Ven. (Engr. E.C. Olewuezi and children for their endurance, support and prayers.

References

[1] B. Abraham and G.E. P. Box (1979). Bayesian analysis of some outlier problems in Time Series.Biometrika, 66, 229 – 236

[2] G. E. P. Box and G. M. Jenkins (1976). Times Series Analysis, forecasting and Control Holden Day, San Francisco.

[3] K. A. Brownlee (1965). Statistical Theory and Methodology in Science and Engineering, Wiley, New York.

[4] I. H. Chang and G.C. Tiao (1983). Estimation of time series parameters in the presence of outliersTechnical report, 8 Unviersity of Chicago, Statistics and Research centre.

[5] C. Chatfield (1979). Inverse autocorrelations, J.R. Statistics Soc. A. 142, Part 3, 363 – 377.M.R. Chrnick et al (1982). Detecting outliers in timer series data. J. Amer Stat. Assoc. 77, 743- 377.

[6] M. R. Chernick (1982) Detecting outliers in time series data. J. America Stat. Assoc. 77, 743 - 747

[7] W. S. Cleveland (1972). The inverse autocorrelations of a time series and their application Technometrics, 14, 277 – 298.

[8] R. D. Cook (1979) . Influence observations in linear regression. J. Amer. Stat. Asso. 74, 169-174.

[9] E. de Alba and D. L. Zartman, (1980). Testing outliers in time series, an application to remotely Sensed temperatures in cattle. Pp 127 – 138 of Anderson O. D. (Ed) (1980) Analyzing Time Series. North – Holland, Amsterdam.

[10] A. J. Fox (1972)/ Outliers in Time Series J. Roy Stat. Soc. B, 43, 350 – 363.

[11] F.R. Hampel (1974). The influence curve and its role in robust estimation. J. AMer. Stat. Asso. 69,383 – 393.

[12] R.D. Martin (1980). Robust estimation of Autoregressive models. In directions of time series, Inst.of Mathematical Statistics, 228 – 254.

[13] C. R. Muirhead (1986). Distinguishing outlier types in time series, J. Roy Stat. Soc. B 48, 39-47.

[13] E. Parzen (1974). Some recent advances in time series modeling I.E.E.E. J. Roy Stat. Trans. Aut.Control 723 – 741

[15] N.P. Olewuezi and D. K. Shangodoyin (2005). Contributions of Inverse autocorrelation function in Transfer function modeling, Far East J. Theo. Stat. 17(2), 227 – 238.

[16] D.K. Shangodoyin (1993). On the specification of time series models in the presence of aberrant observations. Ph.D. thesis, Dept. of Statistics, Univ. of Ibadan, Nigeria

[17] K.D.C. Stoodley and M. Mirnia (1979). The automatic detection of transients, step changes and slope changes in the monitoring of medical time series. The statistician, 28, 163 – 170..