Journal of Mathematics and Statistics 7 (4): 348-352, 2011
ISSN 1549-3644
© 2011 Science Publications

Specification of the Outlier Univariate Model

N.P. Olewuezi
Department of Statistics, Federal University of Technology Owerri, Nigeria

Abstract: Problem statement: An outlier is an observation that deviates so much from other observations so as to arouse suspicion that it was generated by a different mechanism. **Approach:** In this study, outlier univariate model was specified based on statistical measures, Outlier Free (OF) and Outlier Contaminated (OC) models were Stated and compared. **Results:** The number of expected outliers obtained and the effect of the presence of outliers were noted. **Conclusion:** The outlier univariate model incorporated the modeling change in an evolving extension.

Key words: Outlier free, outlier contaminated, univariate model, statistical measure

INTRODUCTION

Outliers are commonly encountered in time series analysis. Observations that deviate from the rest of the observations exist frequently in data. These observations are known by different names such as ‘outliers,’ ‘contaminants’, ‘discordant observations’ and ‘extreme values’. (Beckman and Cook, 1983) defined discordant observations as observations that appear discrepant to the investigator. Outlier contribution were presented using the inverse autocorrelation transfer function. On the other hand, an outlier is just a collective name referring to either contaminant or a discordant observation. It has been shown that statistics derived from data sets that include. Outliers may be misleading. They may be indicative of data points that belong to a different population than the rest of the sample set.

An extensive literature has been accumulated in recent years on statistical procedures for handling outliers. Most of these procedures are derived under the strong assumption that the observations under consideration are drawn independently from some populations. This independent assumption is often totally unrealistic in process data and in economic and business time series. A common approach to deal with outliers in a time series is to identify the locations and the types of outliers and then use intervention models discussed (Box and Tiao, 1975) to accommodate the outlier effects. This approach requires iterations between stages of outlier detection and estimation of an intervention models.

Zellner (1983) and Chang *et al.* (1988) considered procedures that are quite effective in detecting the locations outliers.

Outliers can take several forms in time series, (Fox, 1972) proposed the formal definitions and a classification of outliers in time series context. He proposed a classification of based on an autoregressive model. These two types were later renamed as additive and innovational outliers abbreviated as AO and IO respectively.

We can defines three other types of outliers namely, Level Shift (LS), Transient Change (TC) and Variance Change (VC). These level shifts and variance change are not strictly outliers, but rather structural changes.

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:

$$\overline{\Pi}(B)\ Z_t = a_t$$

Where:

$$\overline{\Pi}(B) = 1 - \overline{\Pi}_1 B - \overline{\Pi}_2 B^2 - ...$$

$\{a_t\}$ is a sequence of identically, independently distributed normal variates with zero mean and variance $\sigma^{2.}$. The function $\overline{\Pi}(B)$ is often expressed as a ratio of:

$$\varphi(B)/\theta(B)$$

Where:

$$\varphi(B) = 1 - \varphi_1 B - ..... - \varphi_p B^p$$

And:

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

Are stationary and invertible operators sharing no common factors.

The models commonly employed on the outlier free time series Z_t are the AO and the IO, which are defined respectively of a single outlier for a simple case.

An additive outlier affects a single observation which is either larger or smaller in value than expected. After this disturbance, the series returns to its normal path as if nothing has happened. An additive model is given by:

$$X_t = \begin{cases} Z_t \ , & t \neq T \\ Z_t + D, & t = T \end{cases}$$

Where:

X_t = The observed series
Z_t = The outlier-free series
T = The time at which the outlier occurs
D = The magnitude of the outlier

Alternatively, it can be written as Eq. 1:

$$X_t = Z_t = D I_t^{(T)} \ = \ \frac{\theta(B)}{\theta(B)} a_t + D I_t^{(T)} \qquad (1)$$

Where:

$$I_t^{(T)} = \begin{cases} 1 \ , & t = T \\ 0 \ , & t \neq T \end{cases}$$

Is an indicator variable which is zero at all lags except at time t = T.
Equivalently:

$$a_t = \overline{\Pi}(B)\left(X_t - D I_t^{(T)}\right)$$

An additive outlier can have serious effect on the properties of an observed time series. It will affect the estimated residuals and the estimates of the parameter values.

Assuming known parameters, the residuals, if the additive outlier does not occur could be obtained as:

$$a_t = \overline{\Pi}(B) X_t$$

whereas in the case of AO, they are obtained as:

$$e_t = \overline{\Pi}(B) X_t = \overline{\Pi}(B)\left(X_t + D I_t^{(T)}\right)$$

The relation between them is:

$$e_t = a_t + \overline{\Pi}(B)\, D I_t^{(T)}$$

So, the effect of the AO on the residuals depends on the $\overline{\Pi}$ weights. Innovational outlier corresponds to an internal error. It affects several observations. An IO model is given by Eq. 2:

$$X_t = Z_t + \frac{\theta(B)}{\varphi(B)} D I_t^{(T)} \ = \frac{\theta(B)}{\varphi(B)}\left(a_t + D I_t^{(T)}\right) \qquad (2)$$

Equivalently:

$$a_t = \overline{\Pi}(B) X_t - D I_t^{(T)}$$

A level shift (sometimes called Level Changes, LC) simply changes the level (or mean) of the series by a certain magnitude D, from a certain observation onwards. It can be seen as sequence of additive outlier of the same size. A transient change (or temporary change, TC) is a generalization of additive outlier and level shift in the sense that it causes an initial impact like an additive outlier but the effect is passed on to the following observations. The impact of a TC is not permanent, however, it decays exponentially.

A Variance Change (VC) is still further away from the additive outlier and innovational outlier types and is not usually considered in connection with outlier at all. It does not affect the level of the series directly like the other types considered. A VC simply changes the variance of the observed series by a new zero mean.

Univariate statistical based outlier: Most of the earliest univariate methods for outlier detection rely on the assumption of an underlying known distribution of the data which is assumed to be identically and independently distributed. Moreover, many discordance tests for detecting univariate outliers further assume that the distribution of the outliers are also known (Barnett and Lewis, 1994).

A central assumption in statistical based methods for outlier detection is a general model that allows a small number of observations to be randomly sampled from distribution $G_I,\ldots G_k$ differing from the target distribution F, which is often taken to be a normal distribution $N(\mu, \sigma^2)$. The outlier identification problem is then translated to the problem of identifying those observations that lie in an outlier region. It should be noted that the outlier definition does not identify which is the observation are contained that is, resulting from distribution $G_I,\ldots,$ $G_{k,}$ but rather it indicates observations that lie in the outlier region.

Given a data set of n observations of a variable, let $\overline{X}$ be the mean and let S be the standard deviation of the data distribution. One observation is declared as an outlier if it lies outside the interval Eq. 3:

$$\left(\overline{X} - ks, \overline{X} + Ks\right) \ldots \tag{3}$$

where, K is usually taken as 2 or 3. The justification of these values relies on the fact that assuming normal distribution; one expects to have 95% (99% respectively) percent of the data on the interval centered in the mean with a semi length-equal to two (three, respectively) standard deviation. Also one expects to have the whole data inside an interval centered at the mean and 3 standard deviations as semi-length, From Eq. 3 the observation X is considered an outliers if Eq. 4:

$$\frac{I x - \overline{x}\, I}{S} > k \tag{4}$$

The problem with the above criteria is that it assumes normal distribution of the data, something that frequently does not occur. Furthermore, the mean and standard deviation are highly sensitive to outliers.

MATERIALS AND METHODS

Let us assume that X and Y are jointly covariance stationary outlier series it is possible to model each series individually as Eq. 5 and 6:

$$X_t = V_x(B)\, U_{xt} \quad = \theta_x(B)\varphi_x(B)^{-1} U_{xt} \tag{5}$$

And:

$$Y_t = V_y(B)\, U_{xt} \quad = \theta_y(B)\varphi_y(B)^{-1} U_{yt} \tag{6}$$

where, U_{xt} and U_{yt} are white noise:

$$\varphi_x(B) = \varphi_0 - \varphi_{X_t} B - \ldots - \varphi_{X_p} B^p$$

$$\varphi_y(B) = \varphi_0 - \varphi_{Y_t} B - \ldots - \varphi_{Y_p} B^p$$

$$\theta_x(B) = \theta_0 - \theta_{X_t} B - \ldots - \theta_{Y_q} B^q$$

and

$$\theta_y(B) = \theta_0 - \theta_{Y_t} B - \ldots - \theta_{Y_q} B^q$$

Are the autoregressive and moving average matrix polynomials of order p and q respectively and ϕ_0 and θ_0 are non singular m x m matrices.

For any nondegenerate case where the covariance matrix is positive definite, we assume without no loss of generality that $\phi_0 = \theta_0 = I$, the m x m identity matrix Again, Eq. 5 and 6 may be written as Eq. 7:

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} V_x B & 0 \\ 0 & V_y B \end{bmatrix} \begin{bmatrix} U_{X_t} \\ U_{y_t} \end{bmatrix} \tag{7}$$

For a properly deduced outlier free time series Z_t which is assumed to follow an ARMA (p, q) model, the AO and IO models are defined as in Eq. 1 and 2 and we have:

$$\varphi(B) Z_t = \theta(B) a_t$$

Using Eq. 1 and 2 into Eq. 5 and 6 we have the identified outlier models as shown in 8 and 9.

For IO we have:

$$\left.\begin{aligned} Z_{X_t} + V_{Z_X}(B)\, D_X I_{X_t}^{(T)} &= V_X(B)\, U_{Xt} \\ Z_{Y_t} + V_{Z_Y}(B)\, D_Y I_{Y_t}^{(T)} &= V_Y(B)\, U_{Y_t} \end{aligned}\right\} \tag{8}$$

And for AO we have:

$$\left.\begin{aligned} Z_{X_t} + D_X I_{X_t}^{(T)} &= V_X(B)\, U_{X_t} \\ Z_{Y_t} + D_Y I_{Y_t}^{(T)} &= V_Y(B)\, U_{Y_t} \end{aligned}\right\} \tag{9}$$

where in both cases:

$$V(B) = \theta(B) / \varphi(B)$$

D represents the magnitude of the outlier and:

$$I_t^{(T)} = \begin{cases} I, & t = T \\ 0, & \text{otherwise} \end{cases}$$

Is the time indicator signifying the time occurrence of the outlier.

After the removal of outliers, Eq. 5 and 6 are respectively for AO and IO given by Eq. 10:

$$\left.\begin{aligned} X_t - D_{XA,T} &= V_X(B)\, U_{X_t} \\ Y_t - D_{YA,T} &= V_Y(B)\, U_{Yt} \end{aligned}\right\} \tag{10}$$

And:

$$\left.\begin{aligned} X_t - V_{Xt-T}\, D_{X1,T} &= V_X(B)\, U_{X_t} \\ Y_t - D_{Yt-T}\, D_{Y1,T} &= V_y(B)\, U_{Yt} \end{aligned}\right\} \quad (11)$$

Rewriting Eq. 10 and 11 in matrix form we have Eq. 12 and 13:

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} D_{XA,T} \\ D_{YA,T} \end{bmatrix} + \begin{bmatrix} V_X(B) & 0 \\ 0 & V_{Y(B)} \end{bmatrix} \begin{bmatrix} U_{X_t} \\ U_{Y_t} \end{bmatrix} \quad (12)$$

That is:

$$A_{(A)} = B_{(A)} \quad D_{(A)} + C_{(A)} \quad U_{(A)} \text{ where}$$

$$A_{(A)} = \begin{bmatrix} X_t \\ Y_t \end{bmatrix}, \ B_{(A)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad D_{(A)} = \begin{bmatrix} D_{XA,T} \\ D_{Ya,T} \end{bmatrix}$$

$$C_{(A)} = \begin{bmatrix} V_X(B) & 0 \\ 0 & V_Y(B) \end{bmatrix} \text{ and } \quad U_{(A)} = \begin{bmatrix} U_{Xt} \\ U_{Yt} \end{bmatrix} \quad (13)$$

$$\text{Also} \begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} V_{X_t-T} & 0 \\ 0 & V_{Y_t-T} \end{bmatrix} \begin{bmatrix} D_{X_{I,I}} \\ D_{Y_t-T} \end{bmatrix}$$

$$+ \begin{bmatrix} V_X(B) & 0 \\ 0 & V_Y(B \end{bmatrix} \begin{bmatrix} U_{Xt} \\ V_{Yt} \end{bmatrix}$$

That is:

$$A_{(I)} = B_{(I)} \; D_{(I)} + C_{(I)} \quad U_{(I)}$$

Where:

$$A_{(I)} = \begin{bmatrix} X_t \\ Y_t \end{bmatrix}, B_{(I)} = \begin{bmatrix} V_{X_t-T} & 0 \\ 0 & V_{Y_t-T} \end{bmatrix}$$

$$D_{(I)} = \begin{bmatrix} D_{X_I,T} \\ D_{Y_I,T} \end{bmatrix}, \quad C_{(I)} = \begin{bmatrix} V_X(B) & 0 \\ 0 & V_{Y(B)} \end{bmatrix}$$

and

$$U_{(I)} \begin{bmatrix} U_{X_t} \\ u_{Y_t} \end{bmatrix}$$

In the specification of the error model, the estimate of the outliers in the series and its effects on the Ui's cannot be ignored. We can clearly see the influence of the matrix U from Eq. 12 and 13 in the specification of the model linking X and Y. The joint process $\left(U_{X_t}, U_{Y_t}\right)$ is not bivariate white noise, since U_{X_t} and U_{Y_t} may be cross correlated at non zero lags and there is every need to check for the presence of outliers before we specify the model (Haugh and Box, 1977).

RESULTS

Let the resultant Outlier Free (OF) and Outlier Contaminated (OC) models be defined as:

OF model:

$$\begin{aligned} Y_t &= \sigma_1 Y_{t-1} + \ldots + \sigma_q X_{t-q} \\ &+ w_1 X_{t-1} + \ldots + w_r X_{t-r} + \ell_t \end{aligned} \quad (14)$$

And:

OC model:

$$\begin{aligned} Y_t &= \sigma_1 Y_{t-1} + \sigma_2 Y_{t-2} + \ldots + \sigma_q Y_{t-q} + w_1 Z_{t-1} \\ &+ \ldots + w_r Z_{t-r} + w_1 D_{t-1} + \ldots + W_r D_{t-r+} \ell_t \end{aligned} \quad (15)$$

Here, two series 14 and 15 are used and univariate time series models are fitted to each series. Series A is the well known First word-Gessel adaptive score Series B is the simulated data of size 100.

For series A, when the series is outlier contaminated, a Transfer model of order (2, 2) was fitted that is, TF (2, 2).

Similarly, for an outlier free series a TF (2, 2) was also fitted.

OF:

$$\begin{aligned} Y_t &= 0.331 Y_{t-1} + 0.480 Y_{t-2} \\ &+ 0.945 X_{t-1} + 0.312 X_{t-2} + e_t \end{aligned}$$

And:

OC:

$$\begin{aligned} Y_t &= 0.709 Y_{t-1} + 0.108 Y_{t-2} \\ &+ 0.866 X_{t-1} + 0.038 X_{t-2} + e_t \end{aligned}$$

From Table 1, the Standard error of the model reveals that the OC model is 1.624 multiple of that obtained for the OF model. The standard errors of the estimates for the OC model are greater than that obtained for the OF model.

For series B, a TF (2,2) was also fitted to both the OC and models given as:

OF:

$$\begin{aligned} Y_t &= -0.078 Y_{t-1} - 0.026 Y_{t-2} \\ &+ 0.044 X_{t-1} + 0.012 X_{t-2} + e_t \end{aligned}$$

Table 1: Estimates of the TF models Fitted with their standard errors in bracket

	A		B	
Type of series	Without outliers	With outliers	Without outliers	With outliers
No of outliers	-	5	-	10.00
σ estimate				
σ_1	0.331	0.7050	0.078	0.011
	-0.265	-0.3300	-0.106	-0.099
σ_2	0.480	0.1080	0.026	0.036
	-0.248	-0.3370	-0.109	-0.087
W estimate				
W_1	0.945	0.8666	0.440	0.005
	-0.552	-0.7740	-0.119	-0.018
W_2	0.312	0.0380	0.012	0.051
	-0.627	-0.7030	-0.106	-0.018
Prob (F)	0.000	0.0000	0.088	0.043
S.E	15.978	25.9470	1.019	1.790

And:

OC:

$$Y_t = -0.011\, Y_{t-1} + 0.036 + 0.005X_{t-1} + 0.051X_{t-2} + e_t$$

DISCUSSION

The univariate outlier model was specified both for the OF and OC models respectively, based on the results from the table. The model standard error shows that the OC model is about 1.76 multiple of that obtained for the OF model. The error variance was increase for the OC model which reduce the power of the test and might result in an inappropriate model.

CONCLUSION

This piece of work had specified a univariate outlier model. The number of outliers was noted for both series. For series A, the standard error of the model revealed that the OC model is about 1.624 multiple of that obtained for the OF model. Also, for series B, the OC model is about 1.76 multiple of that obtained for the OF model.

The joint univariate model can also be identified but I have not experimented with this option.

REFERENCES

Barnett, V. and T. Lewis, 1994. Outliers in Statistical Data. 3rd Edn., John Wiley and Sons, Chichester, ISBN: 0471930946, pp: 584.

Beckman, R.J. and R.D. Cook, 1983. Outlier…s. Technometrics, 25: 119-163.

Box, G.E.P. and G.C. Tiao, 1975. Intervention analysis with applications to economic and environmental problems. J. Am. Stat. Assoc., 70: 70-97.

Chang, I.H., G.C. Tiao and C. Chen, 1988. Estimation of time series parameters in the presence of outliers. Technometrics, 30: 193-204.

Fox, A.J., 1972. Outliers in time series. J. Royal Stat. Soc., 34: 350-363.

Haugh, L.D. and G.E.P. Box, 1977. Identification of Dynamic Regression (Distributed Lag) models connecting two time series. J. Am. Stat. Assoc., 72: 121-130.

Zellner, A., 1983. Applied Time Series Analysis of Economic Data. 1st Edn., U.S. Department of Commerce Bureau of the Census, Washington, pp: 399.