Quest for Research Excellence On Computing, Mathematics and Statistics

Editors
Kor Liew Kee
Kamarul Ariffin Mansor
Asmahani Nayan
Shahida Farhan Zakaria
Zanariah Idrus

Faculty of Computer and Mathematical Sciences
iCMS 2015

Quest for Research Excellence on Computing, Mathematics and Statistics

Chapters in Book

The 2nd International Conference on Computing, Mathematics and Statistics (iCMS2015)

Editors:

Kor Liew Lee
Kamarul Ariffin Mansor
Asmahani Nayan
Shahida Farhan Zakaria
Zanariah Idrus

UNIVERSITI TEKNOLOGI MARA
CAWANGAN KEDAH

iCMS 2015

Quest for Research Excellence on Computing, Mathematics and Statistics

Chapters in Book

The 2nd International Conference on Computing, Mathematics and Statistics

(iCMS2015)

4-5 November 2015
Langkawi Lagoon Resort
Langkawi Island, Kedah
Malaysia

Publication by
Faculty of Computer & Mathematical Sciences
UiTM Kedah

ISBN 978-967-0314-26-6

Content

CHAPTER 28
Outlier Detection in Time Series Model

Nurul Sima Mohamad Shariff, Nor Aishah Hamzah, and
Karmila Hanim Kamil

Abstract. Difficulty occurs in time series when the series are contaminated with outliers typically (i) Innovational Outlier (IO) and (ii) Additive Outlier (AO). As such, before estimating the parameters, one needs to overcome the effect of outliers. There are two approaches employed in this study to identify outliers: (i) iterative outlier detection and joint parameter estimates proposed by Chen and Liu [2] and (ii) application of regression diagnostic tools. A simulation study is performed in an effort to assess the performance of both methods. The identification based on the regression diagnostic tools is seems superior compared to those proposed by Chen & Liu. The results also indicate that the proposed technique based on the regression diagnostic tools can be used to determine the outlier effects and the identification on the type of outlier. Moreover, it can also be applied to more complicated time series models that are widely use in practice particularly in the area of statistics research.

Keywords: outliers; autoregressive model; regression diagnostic tool; robust method

Nurul Sima Mohamad Shariff (✉)• Karmila Hanim Kamil
Faculty of Science of Technology, Universiti Sains Islam Malaysia (USIM).
e-mail: nurulsima@usim.edu.my, karmila@usim.edu.my

Nor Aishah Hamzah
Institute of Mathematical Sciences, University of Malaya, Kuala Lumpur.
e-mail: naishah.hamzah@gmail.com

1 Introduction

The main objective of time series analysis is to have a clear understanding about the phenomenon represented by the sequence of observation given. Difficulty occurs in time series when the series are contaminated with outliers. Thus, there is a clear need to understand the nature of outliers and to have available methods to detect, or accommodate them. Examination of outliers allows more appropriate model to be formulated or enables us to assess any liability that may arise from inferences based on the normality assumption.

The existence of multiple outliers (MO) in the dataset adds to the complexity of identifying and detecting outliers. The time series analysis with MO is hindered by two problems: masking and swamping. While swamping occurs when the observation is not an outlier but is misjudged as an outlier, masking on the other hand occurs when an outlier is being masked by other observations. These two problems are typically caused by the other adjacent outliers. Thus, the regression line will rotate or shift, causing some of the original observations to appear as outliers (swamping). If the additional outliers are added to the original outlier closely, these outliers may not be picked up as outliers because of the more pronounced rotation or shift of the regression line towards these outliers (masking).

In order to overcome masking, several methods are proposed and this includes those of Atkinson and Marco [1]. These methods however, typically involve removing the 'masking' outliers from the data set before the 'masked' outliers can be identified. The misidentification of outliers may result in biasness to parameter estimates and thus provide an inappropriate model in the panel analysis. Many other useful references for the detection of outliers in the time series model can be found in the literatures [2-5].

Due to such interest, the main purpose of this paper is to illustrate the problem involved in time series data in the presence of outliers, particularly: AO and IO. The identification, detection and removal of outliers are discussed, in particular those related to procedures proposed by Chen and Liu [2]. Subsequently, this study focuses on another method in detecting and identifying outliers through the application of regression diagnostic tools. A simulation study is carried out to understand the behavior of two competing methods, namely Chen and Liu [2] method of iterative procedure and those modified procedure as in Pena [6] based on the regression diagnostic tools. A comparative study on the performance of both procedures based on the power function is also conducted to describe the behavior of the procedures.

2 Model and Tests

Let the univariate time series, y_t follows a general AR(p) process of the form

$$\phi(B)y_t = \varepsilon_t, \qquad t = 1, 2, 3, \ldots, n \tag{1}$$

where n is number of observation, B is the backshift operator, with roots outside the unit circle, and ε_t is random error with $E(\varepsilon_t) = 0$ and $Var(\varepsilon_t) = 1$. In general, (1) can be written as

$$y_t = \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} + \varepsilon_t, \qquad t = 1, 2, \ldots, n \tag{2}$$

In order to allow for data contamination, this study assumes an outlier-contaminated series, z_t is observed, and given by

$$z_t = y_t + \omega V(B) I_t(\tau) \tag{3}$$

where ω is the magnitude of outlier and $V(B)$ being the outlier type of $V(B) = 1$ and $V(B) = \dfrac{1}{\phi(B)}$ for AO and IO respectively. Here, I_t is the indicator a variable takes the value of $I_t(\tau) = 1$ when contamination occurs at $t = \tau$ and $I_t(\tau) = 0$, elsewhere.

2.1 Iterative Outlier Detection and Joint Estimation by Chen and Liu

In dealing with multiple types of outliers, the procedure proposed by Chen and Liu [2] is applied iteratively. The procedure can be used to capture the outlier's effect in such a way that the identification can be done in stages. The procedure is applied to non-seasonal ARMA process, where no constant term was used. The procedure comprised three stages: i. Initial outlier detection, ii. Joint estimation of outlier effects and, iii. Detection of outliers based on final parameter estimates.

Stage 1: Initial outlier detection

1.1 Find the maximum likelihood estimates (MLE) of the model parameters to obtain the residuals. Then, the test statistics, $\hat{\tau}_{IO}(\tau), \hat{\tau}_{AO}(\tau)$ are computed as follows:

$$\hat{\tau}_{IO} = \hat{\omega}_{IO}(\tau)/\hat{\sigma}_a \text{ and } \hat{\tau}_{AO} = \{\hat{\omega}_{AO}(\tau)/\hat{\sigma}_a\}\left(\sum_{t=\tau}^{n} x_t^2\right)^{1/2} \tag{4}$$

where $\hat{\omega}_{IO}(\tau)$ and $\hat{\omega}_{AO}(\tau)$ are magnitude of ω at time τ :

$$\hat{\omega}_{IO}(\tau) = \hat{e}(\tau) \text{ and } \hat{\omega}_{AO}(\tau) = \frac{\sum_{t=\tau}^{n} \hat{e}_t x_t}{\sum_{t=\tau}^{n} x_t^2} \tag{5}$$

and σ_a is the standard deviation of the residual process and are obtained the 'omit one' method. Here, x_t is defined as:

$$x_t = \begin{cases} 0 & ; t < \tau \\ 1 & ; t = \tau \\ -\pi_k & ; t > \tau, k = 1,2 \end{cases} \tag{6}$$

1.2 Find the maximum test statistics, $(\hat{\tau}_{IO}(\tau)/\hat{\tau}_{AO}(\tau))$, denoted as η_t. When max η_t > C, where C is predetermined critical value, the type of outlier is detected at t_1.

1.3 The effect of the outlier from the residuals is then discarded by adjusting observation z_t in **(3)** at τ, denoted by $\tilde{z}_\tau$: (i) IO: $\tilde{z}_\tau = E(z_\tau | z_{\tau-1,\tau-2,\ldots,1})$, and (ii) AO: $\tilde{z}_\tau = \sum_{i=\tau-1}^{\tau+5} z_i / 10$, $i \neq \tau$.

1.4 Then, repeat the procedure using the remaining residual series to detect another outlier. When the number of outliers is greater than 1, move to Stage 2.

Detections of outliers are done in sequence where, in this stage, one by one outlier is detected in descending order of $\hat{\tau}$ statistics. The purpose of this procedure is to simplify the computation involved in joint estimation of multiple outliers [3]. The advantage of applying this procedure is that one can reduce the effects of masking.

Stage 2: Joint Estimation of Outlier Effects

2.1 Let y_t be the series subject to *m* intervention at time $t_1, t_2, \ldots, t_m$. This series are contaminated with various types of outliers resulting the new model of z_t, $z_t = \sum_{j=1}^{m} \omega_j V_j(B) I_t(t_j) + \frac{1}{\phi(B)} a_t$

(7)

By fitting an AR model to z_t, the residuals $\hat{e}_t$ may be viewed as:

$$\hat{e}_t = \sum_{j=1}^{m} \omega_j \frac{1}{\phi(B)} V_j(B) I_t(t_j) + a_t$$

(8)

Assume that there are *m* points are detected as possible outliers, either IO or AO. The outlier effects $\omega_j{}'s$ can be estimated jointly using the multiple regression models in equation **(8)** with $\{\hat{e}_t\}$ and $\{V_j(B) I_t(t_j)\}$ are regarded as the output and input variables, respectively.

2.2 The $\hat{\tau}$ statistics of $\omega_j{}'s$ is $\hat{\tau}_j = \hat{\omega}_j / \sigma(\hat{\omega}_j), j = 1, \ldots, m$. An outlier at time point $t = \tau$ from the set can be removed if $\min_j |\hat{\tau}_j| \leq C$, for same critical value, C as used in step 1.2. Repeat step 2.1 with the $m-1$ remaining outliers, one at a time. Otherwise, move the next step.

2.3 The adjusted series is obtained by removing the outlier effects using the most recent estimates of $\omega_j{}'s$ at step 2.1. The MLE of the model parameters are calculated based on the adjusted series and the step 2.1 until 2.3 are repeated until convergence.

Stage 3: Detection of Outliers Based on Final Parameter Estimates

3.1 The residuals are calculated by filtering the original series based on the parameter estimates obtained at step 2.3. Repeat Stage 1 and 2 using the MLE estimates obtained in 2.3.

In this final stage, only the most recent outliers will be detected. The residuals and the observation of the series are then adjusted according to equation in 1.3. The later series are now free from spurious outliers' effects and thus ready for modeling.

2.2 Application of Regression Diagnostics Tools

Recall in the previous section, model **(2)** can be written in matrix form as $Y = X\beta + \varepsilon$ where Y is an $n \times 1$ vector of responses, X is the $n \times p$ matrix of explanatory variables and β is a p-vector unknown parameters and ε is an n-vector random errors usually assume an $\varepsilon \sim NID(0, \sigma^2)$. Here,

$$Y = \begin{bmatrix} y_{P+1} \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} y_p & y_{p-1} & \cdots & y_1 \\ \vdots & \vdots & & \vdots \\ y_{n-1} & y_{n-2} & \cdots & y_{n-p} \end{bmatrix} \tag{9}$$

and X is assumed full rank of size p.
The least squares estimate $\hat{\beta}$ is that value of β which minimize the sum of squares;

$$S(\beta) = (Y - X\beta)^T (Y - X\beta) \tag{10}$$

This can be obtained by differencing **(10)** with respect to β and equating to zero, resulting in $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$ and the corresponding fitted responses

$$\hat{Y} = X\hat{\beta} \tag{11}$$

The *i*-th elements of H is,

$$h_i = x_i (X^T X)^{-1} x_i^T \tag{12}$$

where h_{ii} is the diagonal elements of H which takes values $0 \le h_{ii} \le 1$. According to Huber [7], the leverage point at *i*-th observation when $2 \le \frac{n}{p} h_{ii} \le 3$. Note that, h_{ii} is related to the variance of the residuals, $Var(e_i)$ When h_{ii} tends to zero, $Var(e_i)$ tends to σ^2. Otherwise, $Var(e_i)$ tends to zero if h_{ii} tends to 1. A large h_{ii} may influence the parameter, then pulling the fitted model towards it, resulting bad fit. Some diagnostic tools in regression analysis are used to overcome this situation. Amongst them are:

Studentized residuals: The studentized residuals are given by:

$$r_i^* = \frac{e_i}{s(-i)\sqrt{1-h_{ii}}} \tag{13}$$

where $s(-i)$ is the standard deviation when the *i*-th observation is omitted. The significance of discarding the *i*-th observation is to remove the effects of outlier if it occurs at *i*-th position. The standard variance of residuals may vary with leverage in **(12).** The value of r_i^* is then compared with the critical value, C and observation with r_i^* value is greater than 3.0 will be flagged as outliers.

DFFits$_i$: Another common tool used with the aims at the flagging outliers or influential observations is $DFFits_i$. $DFFits_i$ is a scale measure of the change in the fitted value of $\hat{y}$ when the *i*-th observation is deleted.

$$DFFits_i = \frac{e_i\sqrt{h_{ii}}}{s(-i)(1-h_{ii})} \tag{14}$$

The *i*-th observation is declared as influential point if the $DFFits_i$ value is greater than $2\sqrt{\frac{(p+1)}{n}}$ as suggested by Belsley et al.[8].

2.3 Examining the Residuals

The initial stage of this procedure is the examination of the residuals of the fitted model as in equation **(11)**. Normal probability plot of the residual can be used to identify the obvious outlying points easily. These observations are flagged as outliers if they deviate from the straight line. To reduce the influence of outliers on the estimated model, this point is isolated and the standard deviation is computed as:

$$s^* = \sqrt{\frac{\sum_{i=1,\neq j}^{n} e_i}{n - n_1 - p}}, i \neq j \tag{15}$$

where e_j is an outlier, and n_1 is the number of outlier.

The purpose of isolating these obvious outliers is to allow other statistics measures such as studentized residuals and $DFFits_i$ to detect other suspicious points that may not be apparent from residuals plots. These points are being masked by the present of the obvious outliers.

3 A Study on the Power of the Outlier Detection-Joint Estimation of Chen and Li and the Techniques Based on the Regression Diagnostic Tools

In this section, the power of the outlier-detection and joint estimation procedure by Chen and Liu and the proposed techniques using diagnostic tools in the regression analysis are investigated. The performance of the procedure is said to be powerful and effective if the probability of outlier detection is high and correct identification of outliers is made.

To set the idea, we consider an AR (3) model with the parameters $\phi_1 = 0.87, \phi_2 = 0.02, \phi_3 = 0.01$; and the standard deviation σ = 1. The sample size used is n = 100 and the magnitude of outliers is set to be $\omega = 5\sigma$. To assess the power of the procedure, the following case will be considered: (i) Single outlier of AO / IO, (ii) Multiple outliers AOs / IOs, and (iii) Mixed multiple outliers AOs and IOs. Five situations will be examined in the analysis. Two situations for single outlier AO and IO and the remaining are for the two-outlier case. The location of single outlier is set to be at $\tau = 54$ which in the middle of observational period and for the two outlier cases at $t_1 = 17$ and $t_2 = 64$.

3.1 Results and Discussion

Table 1 provides the results on the simulations of 1000 replications in correct identification outliers for one-outlier and the two-outlier cases. The percentage of correct identification of type and location of outliers are reported in the table. For the two-outlier cases (2AOs, 2IOs and both AO and IO), row labeled by 1st and 2nd show the result of detecting and identifying the first and second outlier, respectively. The correct identification of outliers' type and location are determined by comparing the test statistics as in equation **(4)** for the joint estimation procedure with critical value, C = 3.0 and studentized residual, r_i^* in equation **(13)** with critical value, $\left|r_i^*\right|$ = 3.0 for the proposed technique.

Table 1. Comparison of the power two competing methods**.**

Case / Procedure	Outlier-detection and Joint Estimation	Proposed Technique using Robust Measure, r^*_i
Single AO	0.890	0.993
Single IO	0.966	0.979
Two AOs		
1st outlier	0.887	0.908
2nd outlier	0.901	0.932
Two IOs		
1st outlier	0.925	0.956
2nd outlier	0.914	0.942
Both outliers		
1st outlier (IO)	0.912	0.937
2nd outlier (AO)	0.863	0.931

*Critical value, C and $\left|r_i^*\right|$ = 3.0.

The power of outliers' identification using proposed techniques based on the regression diagnostic tools gives higher percentages than outlier-detection and joint estimation procedure with 90.8% to 99.3% and 86.3% to 96.6%, respectively. For one-outlier cases, more than 97% of correct detection is reported in the proposed techniques which based on the regression diagnostic tools, while Chen and Liu's method with percentage of 89.0% for single AO and 96.6% for an IO. For the two-outlier cases, proposed techniques based on the regression diagnostic tools dominate the power of detecting outliers with 90.8%-93.2% for two AOs, 95.6%-94.2% for two IOs and 93.7%-93.1% for the mixed outliers, AO and IO. The technique based on regression diagnostic tools seems performing better in detecting AO. For example, in the mixed outliers, the power of the outlier-detection and joint estimation in detecting AO is 86.3% while in proposed techniques is 93.1%.

The identification based on the regression diagnostic tools seems to outperform those techniques proposed by Chen and Liu. The accuracy in identifying the type and location of outlier based on the regression diagnostic tools are slightly higher than the outlier-detection and joint estimation of Chen and Liu [2]. Hence, this concludes that the proposed techniques based on the regression diagnostic tools can be used to detect outlying observations and the identification of the type of outliers can be done through the actual observation in the series.

4 Conclusion

This paper proposes an alternative method to those proposed by in Chen and Liu's in capturing spurious observation in time series model. One of the hitches involved in Chen and Liu's procedure is due to the complication in doing an adjustment on the contaminants because they are affected by the neighboring points. In addition, it involves iterative steps in detecting initial outliers to ensure that the subsequent series is free from outliers' effect. The results from the simulation study provides avenue to the proposed technique based on the regression diagnostic tools to be one of the superior techniques in determining the outlier effects and the identification of type of outlier. The proposed procedure uses the simple procedure that is consisting only two straightforward techniques in outliers' identification; (i) detecting obvious outlier using normal probability plot and (ii) capturing other outliers determined by the robust measures. This technique can also be applied to the more advanced time series models that are commonly used in practice, especially in the area of statistics research. With the advancement and flexibility of the software packages nowadays, future researchers could further extend this study to the higher order of time series model.

Acknowledgements.

The authors gratefully acknowledge the grant support **USIM/RAGS/FST/36/50214** from the Universiti Sains Islam Malaysia (USIM), Malaysia.

References

1. Atkinson, A. and Riani, M. *Robust Diagnostic Regression Analysis*. Springer. (2000).
2. 2. Chen, C, and Liu, L. M. (1993). Joint Estimation of Model Parameter and Outlier Effects in Time Series. *Journal of the American Statistical Association* 88(421): 284-297.
3. Chang, I., Tioa, G. C. and Chen, C. (1989). Estimation of Time Series Parameters in the Presence of Outliers. *Technometrics* 30(2):193-202.
4. Saches, M. J., Estadisca, L. and Pena, D. (2003). The Identification of Multiple Outliers in ARIMA Models. *Communication in Statistics – Theory & Methods* 31(6):1265-1287.
5. Kaya, A. (2010). Statistical Modelling for Outlier Factors. *Ozean Journal of Applied Sciences* 3(1):185-194.
6. Pena, D. (1990). Influential Observation in Time Series. *Journal of Business and Economic Statistics* 8(2): 235-241.
7. Huber, P. J. (1981). *Robust Statistics.* John Wiley & Sons, New York.
8. Belsley, D. A., Kuh, . and Welsch, R.E. (1980). *Regression diagnostics: identifying influential data and sources of collinearity*. Wiley series in probability and mathematical statistics.

iCMS 2015

ISBN 978-967-0314-26-6