A Study of Detecting Outliers in Time Series Using Simulation

Abdalla M. EL-HABIL, Ph.D. in Applied Statistics
Department of Applied Statistics, Faculty of Economics and Administrative Sciences; Al-Azhar University, Gaza - Palestine.

Abstract

In this paper, simulations for detecting outliers and studying their effects on the values of AR(1) time series, transfer function model with one-input variable, transfer function model with two-input variables processes, and simultaneous transfer function (STF) are conducted using the STFMODEL JOINTMDL paragraph in the Scientific Computing Associate Corporation (SCA) program. A simulation of a transfer function model is conducted to check its validity. By using the SCA program, Victor Gomez and Agustin Maravall's example for detecting outliers in time series by TRAMO program is pursued.

The conclusion, which we come up with, is that the presence of outliers, depending on their nature, may have a moderate to substantial impact on the effectiveness of the standard methodology for time series analysis with respect to model identification, estimation, and forecasting.

Keywords: Simulation - transfer function model – AR (1) time series - outliers.

 menggunakan

دراسة أسلوب المحاكاة لأكتشاف القيم الشاذة في السلسلة الزمنية

الدكتور عبد الله الهيل
استاذ الإحصاء التطبيقي
قسم الإحصاء التطبيقي
جامعة غزة - فلسطين

الملخص

تم في هذا البحث استخدام أسلوب المحاكاة لأكتشاف القيم الشاذة في السلسلة الزمنية ودراسة تأثيراتها في تقدير معالم كل من آليات الإحصائيات الذاتي من الدرجة الأولى وآليات التحويل في حالتين من تغير واحد ومتغيرين وكذلك انتروج دالة التحويل المتزامنة، وقد تم أيضاً باستخدام أسلوب المحاكاة التحقق من صحة آليات التحويل باستخدام البرنامج الإحصائي المتخصص لـ SCA، وكذلك في إطار اكتشاف القيم الشاذة في السلسلة الزمنية تم استخدام البرنامج نفسه تحقق نتائج أفضل من نتائج برنامج TRAMO باستخدام مثال من تابع مثال Victor Gomez.

تاريخ اسلام البحث 2009/6/25

تاريخ قبول النشر 2009/6/23
Introduction

Outliers have recently been studied more and more in the statistical time series literature and this interest is also growing in econometrics. Usually time series outliers are informally defined as somehow unexpected or surprising values in relation to the rest of the series.

Data of potential value in the formulation of public and private policy frequently occur in the form of time series. Most time series data are observational in nature. In addition to the possible gross errors, time series data are often subject to the influence of some uncontrolled or unexpected interventions; for example, implementations of a new regulation, major changes in political or economic policy, or occurrence of a disaster. Consequently, discordant observations and various types of structural changes occur frequently in time series data. Whereas, the usual time series model is designed to grasp the homogeneous memory pattern of a time series, the presence of outliers, depending on their nature, may have a moderate to substantial impact on the effectiveness of the standard methodology for time series analysis with respect to model identification, estimation, and forecasting. Therefore, there is a clear need to have available methods to detect, or accommodate, them.

Simulation data are derived from a sequence of pseudo random numbers. These pseudo random numbers are created by a random number generator. The generator requires an initial seed value from which to generate its first value. The random number generator creates both a random number and a new seed for the next value.

The SIMULATE paragraph in the Scientific Computing Associate Corporation (SCA) program may be used to estimate an ARIMA model or a transfer function model. The use of the SIMULATE paragraph for the estimation of a transfer function model is identical as its use for the estimation of an ARIMA model, except for the presence of input series. The SIMULATE paragraph will first generate a noise sequence using a pseudo random number generator. This sequence is then used according to a transfer function model specified lately using the TSMODEL paragraph.

The paper is organized as follows. Section 2 recalls the technical background of Transfer Function model. Section 3 detecting outliers of a simulated AR (1) time series. Section 4 detecting outliers by TRAMO and SCA programs. Section 5 simulation is a transfer function model. Section 6 simulation is a single-equation transfer function model (with two-input variables. Section 7 simulation is simultaneous transfer function (STF) model. Section 8 concludes.
The Transfer Function Model

In many cases, we may be able to relate the response (i.e., the observed value) of one series to its own past values, and also to the past and present values of other time series. So, we consider a time series $Y_t$ as an output time series whose values may be related to one or more input time series $X_{it}$ for example, sales may be related to advertising expenditures; daily electricity consumption may be related to certain weather variable series such as maximum daily temperature or relative humidity or both.

For a single explanatory variable, the transfer function model is

$$Y_t = C + B_1 X_t + N_t$$

where $Y_t$ represents a stationary ARMA process. If we assume that the input and output variables are both stationary time series, the general form of the single-input, single-output transfer function model can be expressed as

$$Y_t = C + \left[ \frac{\omega(B)}{\delta(B)} \right] X_t + N_t \quad (1)$$

where $N_t$ follows an ARMA model (i.e., $N_t = [\theta(B)/\phi(B)] at$), $at$ is a sequence of random errors that are independently and identically distributed with normal distribution $N(0, \sigma^2_a)$, and

$$\omega(B) = \omega_0 + \omega_1 B + \omega_2 B^2 + \ldots + \omega_{[s-1]} B^{[s-1]}$$

and

$$\delta(B) = 1 - \delta_1 B - \delta_2 B^2 - \ldots - \delta_r B^r.$$

In practice, the number of terms in $\omega(B)$ is small and the value for $r$ is usually 0 or 1. We can also represent the rational polynomial operator $\omega(B)/\delta(B)$ with a linear operator $\nu(B)$, where $\nu(B) = \nu_0 + \nu_1 B + \nu_2 B^2 + \ldots$.

The polynomial operators are related according to $\nu(B) = \omega(B)/\delta(B)$.

Since we assume the transfer function is stable, the coefficients $\nu_0, \nu_1, \nu_2, \ldots$ diminish to zero regardless the order of the $\delta(B)$ polynomial. If the linear operator $\nu(B)$ is used, the model in (1) can be written as:

$$Y_t = C + \nu(B) X_t + N_t \quad (2)$$

In the event that $\delta(B) = 1$ (i.e., $r = 0$), we have $\nu(B) = \omega(B)$ and $\nu(B)$ has a finite number of terms. In the case that $\delta(B) \neq 1$ (i.e., $r > 0$), then $\nu(B)$ has an infinite number of terms.

The representation in (1) can be extended directly to the case of multiple-input transfer function model as:

$$Y_t = C + \left[ \frac{\omega_1(B)}{\delta_1(B)} \right] X_{1t} + \ldots + \left[ \frac{\omega_m(B)}{\delta_m(B)} \right] X_{mt} + N_t \quad (3)$$

we can also use the linear form of the transfer function by writing (2) as:

$$Y_t = C + \nu_1(B) X_{1t} + \nu_2(B) X_{2t} + \ldots + \nu_m(B) X_{mt} + N_t \quad (4)$$

The values $\nu_0, \nu_1, \nu_2, \ldots$ are either referred to as the transfer function weights or the impulse response weights for the input series $X_t$ (see chapter 9 of Box and Jenkins, 1970). These weights provide a measure of how the input series affects the output series, and the weight given to each time lag.
A study of detecting outliers in time series … [12]

That is $v_0$, is a measure of how the current response is affected by the current value of the input series; $v_1$ is a measure of how the current response is affected by the value of the input series one period ago; $v_2$ is a measure of how the current response is affected by the value of the input series two periods ago; and so on. The sum of all weights, usually represented by $g$, is called the steady state gain and represents the total change in the mean level of the response variable if we maintain the input at a single unit increase above its mean level.

There are three assumptions of the model in (2) which describes the transfer function between $X_t$ and $Y_t$ (either in a linear form or as a rational polynomial) (Lon-Mu Liu, Gregory B. Hudak, 1992-2000):

1. The input series can affect the response variable, but not conversely (i.e., the relationship between $X_t$ and $Y_t$ is unidirectional).
2. The input series is assumed to be independent of the disturbance.
3. The model is stable; this is usually manifested as assuming the input and output series are stationary time series, and that the sum of the transfer function (TF) weights is finite.

The assumption that the output series does not affect the input series is often appropriate for physical or engineering processes. In these cases, the input may be viewed as a controller mechanism that is used to maintain a certain level in the response variable. If we model economic and business data, we may wish to use more dynamic models that allow for bi-directional (or feedback) relationships. Examples of such models include simultaneous transfer function (STF) models, vector ARMA models. However, although the assumption of a unidirectional relationship may not be strictly true, transfer function models can still be effectively in modeling business and economic data.

Note: There are some special cases of the transfer function model shown in (3).

1. If there are no explanatory variables, then the transfer function is the ARIMA model.
2. The intervention models can be obtained directly if all input series are binary series (that is, series consisting of only the values 0 and 1).

**Detecting Outliers of a Simulated AR (1) Time Series**

To facilitate our understanding of detecting outliers and their effects for example on the values of a simulated AR (1) process, we will assume that the constant of the proposed model is equal zero. For this purpose, 100 observations are simulated from the model

$$z_t = \left[1 / (1 - 0.6B) \right] a_t$$

with $\sigma a = 1.0$. The data are shown in figure 1.
To illustrate, for example, the effect of an AO on the base AR(1) model, we include an AO at time $t = 42$ with $\omega_A = 6$ (the value $\omega_A$ represents the amount of deviation from the "true" value of $Z_T$). The new shape of data is shown in figure 2.
By using the SCA program, only AO has been detected at \( t = 42 \), and we get estimation results for an AR (1) fit of the simulated AR(1) process as the following:

<table>
<thead>
<tr>
<th>Case</th>
<th>( \phi ) estimate</th>
<th>S.E. of ( \phi ) estimate</th>
<th>( \sigma_a ) estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without outlier</td>
<td>0.5921</td>
<td>0.0810</td>
<td>0.9783</td>
</tr>
<tr>
<td>AO at time ( t = 42 )</td>
<td>0.4683</td>
<td>0.0888</td>
<td>1.2177</td>
</tr>
</tbody>
</table>

From the table, with the additive outlier at time \( t = 42 \), we can see that the parameter estimate is decreased by approximately 0.13 and the estimated residual variance is inflated and in consequence the prediction intervals can be too wide. In turn, it will affects the model identification, estimation and forecasting.

**Detecting Outliers by TRAMO and SCA Programs**

Victor Gomez and Agustin Maravall consider an example for detecting outliers in time series in their (Beta version: November 1997, instructions for the user of program TRAMO (Time Series Regression with ARIMA Noise, Missing Observations, and Outliers).

TRAMO is a program for estimation and forecasting of regression models with possibly non-stationary (ARIMA) errors and any sequence of missing values. The program interpolates these values, identifies and corrects for several types of outliers, and estimates special effects such as Trading Day and Easter and, in general, intervention variable type of effects.

The program TRAMO has a facility for detecting outliers and for removing their effect; the outliers can be entered by the user or they can be automatically detected by the program, using an original approach based on those of Tsay (1986) and Chen and Liu (1993). The outliers are detected one by one, as proposed by Tsay (1986), and multiple regressions are used, as in Chen and Liu (1993), to detect spurious outliers. The procedure used to incorporate or reject outliers is similar to the stepwise regression procedure for selecting the "best" regression equation. This results in a more robust procedure than that of Chen and Liu (1993), which uses "backward elimination" and may therefore detect too many outliers in the first step of the procedure.

The four types of outliers considered are additive outlier (AO), innovational outlier (IO), level shift (LS), and transitory change (TC).

The example illustrates what could be a standard way of executing TRAMO for the monthly series of sales in retail stores in Chen Liu and Hudak (1990). The series consists of 153 observations. The data are shown in figure 3.
Victor Gomez and Agustin Maravall identify for the series the ARIMA (2,1,0) (0,1,1)12 model of the type
\[(1 + \varphi_1 B + \varphi_2 B^2)(1 - B)(1 - B^{12}) = (1 + \Theta 12 B^{12})a_t,\]
and only three outliers have been detected.

Their work is pursued, and retained the same model for the series according to the Auto-Correlation Function (ACF) of the residuals.

By using the SCA program, 15 outliers have been detected at \( t = 23, 26, 61 \) (AO-type), \( t = 38, 118 \) (IO-type), \( t = 142 \) (LS-type), and \( t = 40, 53, 66, 79, 88, 92, 101, 105, 114 \) (TC-type).

**Simulation of a Transfer Function Model**

A simulation for a transfer function model is conducted to see how close our estimates are to the true model (to check its validity).

To illustrate the simulation of a transfer function, we will simulate an input series and an output series. Specifically, we will simulate \( X_t \) and \( Y_t \) so that \( (1 - 0.65B)X_t = 13.0 + \epsilon_t \) and \( Y_t = 7.0 + \left[ 0.3B/(1 - 0.75B) \right]X_t + (1 - 0.7B)a_t \) with \( \sigma_\epsilon = 2.5 \) and \( \sigma_a = 1.5 \).

We will simulate 200 observations for \( X_t \) and \( Y_t \) and store the data in XDATA and YDATA, respectively. We intentionally simulate more than 200 observations and then select only the last 200 values of XDATA and YDATA to ensure that any potential irregularities in the beginning of the recursive computation of values are eliminated. The data are shown in figure 4.
We can check to see how consonant these series are to $X_t$ and $Y_t$ by computing the values of statistics based on our equations mentioned in the transfer function. In particular:

- $\mu_X = \frac{13.0}{1 - 0.65} = 37.143$; the ACF for $X_t$ is $(0.65)^l, l=1,2,3,\ldots$;
- the steady state gain of the transfer function is $g = \frac{3.0}{1 - 0.75} = 1.2$;
- $\mu_Y = 7.0 + g \mu_X = 7.0 + 1.2 (37.143) = 51.571$; $\nu_0 = 0$ and the values of the remaining transfer function weights are $(0.3) (0.75)^{l-1}$

This is not done here, instead, we estimate

$$Y_{DATA_t} = c + \left[\omega B/(1 - \delta B)\right]X_{DATA_t} + (1 - \Theta B)at,$$

to see how close our estimates are to the true model.

A summary from an exact estimation of this model is given below (t-values in parentheses):

- estimate of $c = 9.7141 (12.37)$, estimate of $\nu_1 = 0.3076 (41.06)$,
- estimate of $D_1 = 0.7268 (98.31)$, and estimate of $\Theta = 0.7754 (17.15)$.

The estimated values of $c, \omega, \delta, \text{ and } \Theta$ are in reasonable to good accord with the values used in the simulation. All diagnostic checks of this model support its validity. No outlier is detected.

When we include an IO at time $t = 50$ with $\omega I = 5$, only IO has been detected at time $t = 50$ by using the SCA program, and we get estimation results for a transfer function fit of the simulated transfer function process as the following:
Estimation results for a transfer function fit of the simulated transfer function, with the innovational outlier at $t = 50$, the parameters estimates are moderately changed, and the estimated residual variance is inflated and in consequence the existing of this outlier will affects the model identification, estimation and forecasting.

**Simulation of a Single-equation Transfer Function Model (with two-input variables)**

To detect outliers and study their effects on the values of a simulated single-equation transfer function model (with two-input variables), 300 observations are simulated from the model

$$z_{data} = 12.0 + (0.6)x_{data} + (0.7)y_{data} + at,$$

where the model of $x_{data}$ is

$$(1 - 0.66B)x_{data} = 12.0 + at,$$

and the model of $y_{data}$ is

$$(1 - 0.7B)y_{data} = 11.0 + (1 - 0.6B)at,$$ with $\sigma_a = 2.25$.

We select only the last 250 values of $x_{data}$, $y_{data}$ and $z_{data}$ to ensure that any potential irregularities in the beginning of the recursive computation of values are eliminated.

By using the SCA program, we estimated the model

$$z_{data} = 12.0 + (\omega_1)x_{data} + (\omega_2)y_{data} + at,$$

AO has been detected at $t=50, 82, 106$ and TC at $t = 160$. We get estimation results for a single-equation transfer function model (with two-input variables) fit of the simulated single-equation transfer function model (with two-input variables) process as the following:

<table>
<thead>
<tr>
<th>Case</th>
<th>estimate of $\omega_1$</th>
<th>S.E. of estimate of $\omega_1$</th>
<th>Estimate of $\omega_2$</th>
<th>S.E. of estimate of $\omega_2$</th>
<th>S.E. of estimate of $\sigma_a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without outlier</td>
<td>0.5741</td>
<td>0.0463</td>
<td>0.7416</td>
<td>0.0518</td>
<td>2.3085</td>
</tr>
<tr>
<td>AO at $t=50,82,106$ and TC at $t = 160$</td>
<td>0.5494</td>
<td>0.0424</td>
<td>0.7448</td>
<td>0.0476</td>
<td>2.3096</td>
</tr>
</tbody>
</table>

As we see from the table, the parameters estimates are moderately changed, and the estimated residual variance is inflated. Thus, the presence
of those extraordinary events could easily mislead the conventional time series analysis.

Simulation of Simultaneous Transfer Function (STF) Model

In order to detect outliers and study their effects on the values of a simulated simultaneous transfer function model, 150 observations are simulated from the models

\[ Z_{1\text{data}} = 17.0 + (1 - 0.5B) at, \]
\[ Z_{2\text{data}} = 25.0 + (1 - 0.6B) at, \]

with \( \sigma_a = 2.25 \). The data of \( z_{1\text{data}} \) and \( z_{2\text{data}} \) are shown in figure 5.

Figure 5

\[ z_{1\text{data}} \text{ and } z_{2\text{data}} \text{ for simultaneous transfer function model} \]

By using the SCA program, we estimated the two models simultaneously using the STFMODEL JOINTMDL paragraph
TC has been detected at $t = 112$ and IO at $t = 126$. We get estimation results for simultaneous transfer function model fit of the simulated simultaneous transfer function process as the following:

Estimation results for the simultaneous transfer function fit of the simulated transfer function process (first model)

<table>
<thead>
<tr>
<th>Case</th>
<th>estimate of $z_{1\text{data}}$</th>
<th>S.E. of estimate of $z_{1\text{data}}$</th>
<th>estimate of $z_{2\text{data}}$</th>
<th>S.E. of estimate of $z_{2\text{data}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without outlier</td>
<td>0.4579</td>
<td>0.0726</td>
<td>0.0702</td>
<td>0.0811</td>
</tr>
<tr>
<td>TC at $t = 112$</td>
<td>0.4786</td>
<td>0.0671</td>
<td>0.0786</td>
<td>0.0794</td>
</tr>
</tbody>
</table>

Estimation results for the simultaneous transfer function fit of the simulated transfer function process (second model)

<table>
<thead>
<tr>
<th>Case</th>
<th>estimate of $z_{1\text{data}}$</th>
<th>S.E. of estimate of $z_{1\text{data}}$</th>
<th>estimate of $z_{2\text{data}}$</th>
<th>S.E. of estimate of $z_{2\text{data}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without outlier</td>
<td>0.0061</td>
<td>0.0369</td>
<td>0.7262</td>
<td>0.0568</td>
</tr>
<tr>
<td>IO at $t = 126$</td>
<td>-0.0079</td>
<td>0.0706</td>
<td>0.7452</td>
<td>0.0553</td>
</tr>
</tbody>
</table>

As we see from the above two tables, the parameters estimates are changed, and the estimated residual variance is inflated. So, those outliers could easily mislead the conventional time series analysis.

Summary

In this paper, simulations for detecting outliers and studying their effects on the values of AR(1) time series, transfer function model with one-input variable, transfer function model with two-input variables processes, and simultaneous transfer function (STF) are conducted using the STFMODEL JOINTMDL paragraph in the Scientific Computing Associate Corporation (SCA) program. A simulation of a transfer function model is conducted to check its validity. Also, by using the SCA program, Victor Gomez and Agustin Maravall’s example for detecting outliers in time series by TRAMO program is pursued, in which, they detect only three outliers. However, by using the SCA program, 15 outliers have been detected.

The conclusion, which we come up with, is that the presence of outliers, depending on their nature, may have a moderate to substantial impact on the effectiveness of the standard methodology for time series analysis with respect to model identification, estimation, and forecasting.
A study of detecting outliers in time series …

References