.

The Winner:

"This was a clear winner distinguished by the clarity and quality of the presentation and the importance of the problem addressed. This teaching note provides the basis for presenting an introduction to time series outliers in a variety of ways in order to stimulate student interest and reinforce their understanding of the topic. This is clearly a very useful idea for anyone teaching a time series or forecasting course and thus is a very deserving winner of the first UTS Introductory Teaching Prize."

Understanding the effect of time series outliers on sample autocorrelations

By Wai-Sum Chan, The Chinese University of Hong Kong
Email: chanws@cuhk.edu.hk

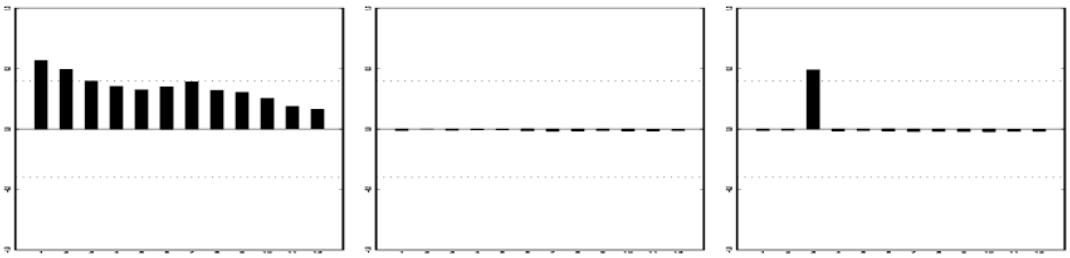
1. Introduction

Most introductory textbooks on linear time series analysis mention the fact that extreme data points could have a great influence on sample autocorrelations and hence the specification of time series models. However, not many textbooks provide a rigorous mathematical explanation of this phenomenon. In this note a way is suggested to fill this teaching gap.

2. Computer experiments

Computer experiments (e.g., via applet) are effective tools to demonstrate that extreme observations may have a large effect on the calculation of sample autocorrelations. An applet is able to show students interactively the effect of extreme data points on the sample autocorrelation function via the World Wide Web. Students often find this type of computer experiments amusing and at times entertaining.

As an illustration, we consider a well-known time series dataset (*Series A* of Box and Jenkins [1976]). Figure (a) plots the sample autocorrelation function (ACF) up to lag 12 for this observed series. In order to examine the effect of a large extreme value on sample autocorrelations, the data $y_{100} = 16.9$ is replaced by $y_{100} = 16.9 + \omega$. Students are asked to plot the sample ACF for various values of ω (say, $\omega = 0, 1, 10,$ and 100). The result of $\omega = 100$ is given in Figure (b). Observe that the contaminated observation pushes all the autocorrelations towards zero. For the next experiment, we study the effect of two outliers. The data $y_{100} = 16.9$ is replaced by $y_{100} = 16.9 + \omega_1$, as well as $y_{103} = 16.4$ is replaced by $y_{103} = 16.4 + \omega_2$. The result of $\omega_1 = \omega_2 = 100$ is given in Figure (c). We find that the contaminations create a large spurious autocorrelation at lag 3.



(a) No outlier (b) One outlier (c) Two outliers

3. A mathematical explanation

From the above computer experiments, students should be able to conclude that the sample ACF offers no resistance to extreme data, which can easily control the function and cause it to give a totally misleading correlogram. However, like other science experiments, the teacher is expected to provide students with a rigorous explanation for the observed phenomena from the experiment. We can show to students, algebraically, the results in the experiments are actually expected.

Given an observed time series $y_t, t = 1, \ldots, N$, the sample ACF is defined by $r_k = \frac{\sum_{t=1}^{N-k}(y_t - \overline{y})(y_{t+k} - \overline{y})}{\sum_{t=1}^{N}(y_t - \overline{y})^2}$ where $\overline{y} = \sum y_t / N$ for $k = 1, \ldots, N-1$. For the case of *one outlier*, let the value of y_T in the sample be replaced by $(y_T + \omega)$ for a fixed T $(k < T < N - k)$. Then, by breaking down the summations in the sample ACF, r_k, it is easy to show that

$$r_k \to -\frac{N + k}{(N)(N-1)}$$

as $\omega \to \infty$. This means that the information in the r_k could be completely destroyed by an extreme observation. When ω tends to ∞, the r_k becomes a small and negative constant and it does not contain any information about the underlying time series model. The situation will not be improved even though the sample size is increased. For the case of *two outliers*, the value of y_T in the sample be replaced by $(y_T + \omega_1)$ and the observation y_{T+m} be replaced by $(y_{T+m} + \omega_2)$. We can show that when N, ω_1 and ω_2 are all very large,

$$r_k \to \begin{cases} 0 & k \neq m, \\ 0.5 & k = m. \end{cases}$$

This result implies that two nearby outliers might create a significant spurious autocorrelation at lag $k = m$ which could lead to erroneous model specification.

4. Real applications

Economic time series observations are often influenced by interruptive events such as strikes, outbreaks of wars or diseases, and sudden political or economic crises. We can ask students to explore the effect of time series outliers using real datasets, for examples, among many others, stock return series during the global crashes in 1987; East Asian currency exchange rate data in mid-1997; and monthly number of foreign visitors to Hong Kong during the SARS (Severe Acute Respiratory Syndrome) period in early 2003.