THALES Project No. xxxx

The Analysis of Outliers in Statistical Data

Research Team

Chrysseis Caroni, *Associate Professor (P.I.)*

Vasiliki Karioti, *Doctoral candidate*

Polychronis Economou, Christina Pierrakou, *Postgraduate students*

School of Mathematical & Physical Sciences, National Technical University of Athens, Greece

Introduction

Statistical *outliers* are unusual points in a set of data that differ substantially from the rest. An outlier could be different from other points with respect to the value of one variable (e.g. the breaking strain for a beam that broke at exceptionally low load) or, in multivariate data, it could be unusual in respect of the combination of values of several variables.

One particular reason for the importance of detecting the presence of outliers is that potentially they have strong influence on the estimates of the parameters of a model that is being fitted to the data. This could lead to mistaken conclusions and inaccurate predictions.

Figures 1 and 2 below give two examples of apparent outliers, one in a time series and the other in a set of bivariate data.

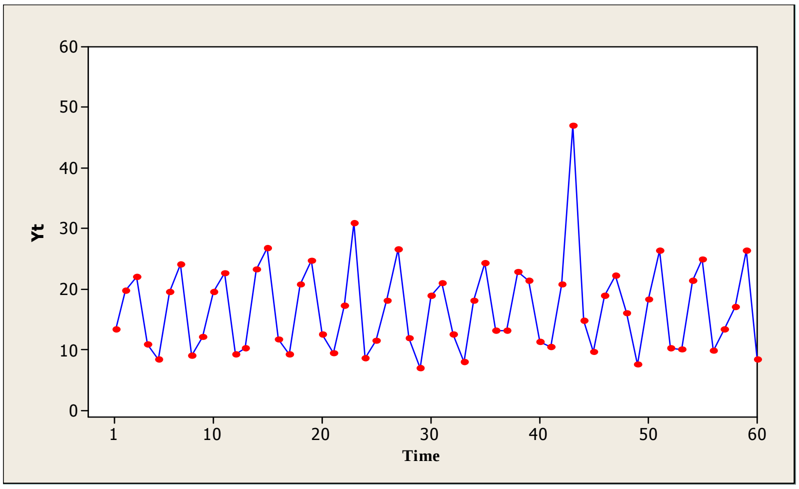

Figure 1. A possible outlier (at time 43) in a time series.

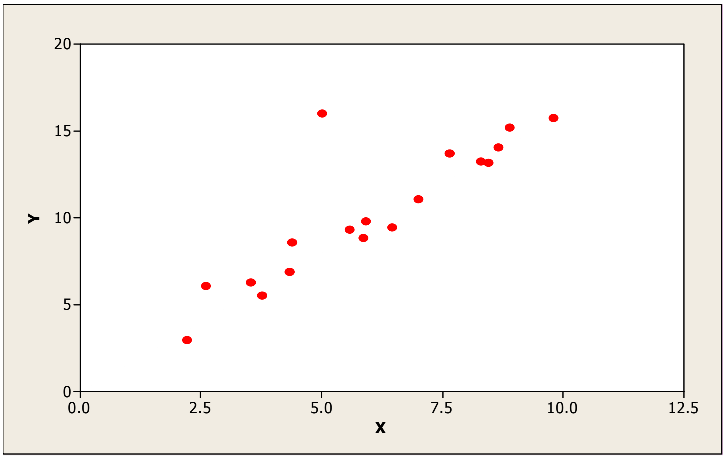

Figure 2. A possible outlier in a sample of bivariate data. The presence of this point has a strong influence on the value of the correlation between *X* and *Y*, reducing it from 0.97 to 0.86.

There is a very extensive bibliography on the topic of outliers. For example, Barnett & Lewis [1] give nearly 1000 references. However, relatively little work has been done on outliers in time series. Outliers are quite likely to arise in time series – for example in an economic time series affected at some point by an external event such as war or major strikes – and may have severe effects on model fitting and estimation.

Classical methods of time series analysis apply to a single series of long duration. However, in many situations, sets of relatively short time series arise. Our research focuses chiefly on the identification of outliers in data of this kind.

Detection of an outlying series

The first objective is to develop a method of detecting an outlying *series*, rather than outlying *points*, in a set of time series.

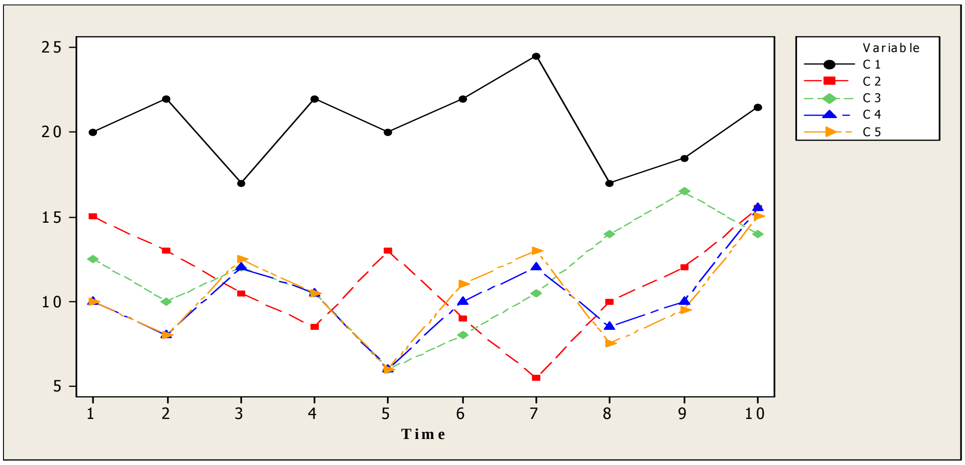

Figure 3. A possible outlying series (C1) among a set of 5 time series.

We assume the following model for a set of *m AR(1)* series:

$$y_{it}-\mu_i=\alpha(y_{i,t-1}-\mu_i)+u_{it,}\;\; i=1,...,m,\;\; t=1,...,n_i,\; u_{it}\text{ iid }N(0,\sigma^2)$$

Two different models for the series levels μ_i are investigated:

I) $\mu_i=\mu\;\forall i$ - all series have the same level

II) $\mu_i\sim N(\mu_0,\sigma_0^2)$ - a random effects model

In the presence of an outlier, the two cases are modified as follows:

I') $\mu_i=\mu,\, i\neq j;\; \mu_j=\mu+\delta$ for some j

II') $\mu_i\sim N(\mu_0,\sigma_0^2),\; i\neq j\;;\;\; \mu_j\sim N(\mu_0+\delta,\sigma_0^2),$ for some j

Within this framework:

a) we used the two-stage maximum likelihood method to construct test statistics for testing between the hypotheses I and I' and between II and II', and investigated the properties of the tests;
b) we examined the possibility of applying simple tests for an outlier in a single sample of univariate data, to the means of the series (or to a simple function of the means). Although unequal length of the series implies that their means have unequal variances, we found that this very simple approach works well.

Some of these results have been published in Karioti & Caroni [3].

Simultaneous outlier in every series

We suppose that an external factor affects every one of a set of time series, causing the appearance of an outlier at the same time in each series. We examine two cases, supposing the outlier to be an *innovative outlier (IO)* or an *additive outlier (AO)*. The theory is developed for a set of *AR(p)* series.

Random IO:

$$u_{it}\sim N(0,\sigma^2),\ \text{i=1,...,m;}\ \ \text{t=1,...,}n_i;\ t\neq q$$
$$u_{iq}\sim N(\Delta,\sigma^2+\sigma_\delta^2)$$

In this case, we used two-stage maximum likelihood to construct a test statistic for the presence of the outliers and obtained critical values by simulation.

Equal IO:

When the outlier has the same size in each series, the model can be written in the form of a time-series regression

$$\underset{\sim}{y}_i=X_i\underset{\sim}{\beta}+\underset{\sim}{\varepsilon}_i,\quad \text{i=1,...,m}$$

with $V(\underset{\sim}{\varepsilon}) = \Sigma \otimes I$. The presence of the outliers is equivalent to the addition of an extra column to the matrices X.

We examine three models, differing in the form assumed for the covariances:

$$\Sigma = \text{dia}\left(\sigma_1^2, \sigma_2^2, ..., \sigma_m^2\right) \quad \text{(heteroscedasticity between series)}$$
$$\Sigma \quad \text{unrestricted}$$
$$\Sigma = \sigma^2 \left\{(1-\rho)I + \rho J\right\} \quad \text{(equicorrelation)}$$

Applying the method of generalized least squares (GLS) gives the estimator

$$\underset{\sim}{\hat{\beta}} = \left(X'V^{-1}X\right)^{-1} X'V^{-1}\underset{\sim}{y}$$

for the regression coefficients, and a formula for the estimation of V which takes a different form for each of the three models. The two equations are solved iteratively. In this way, we are again able to obtain a two-stage maximum likelihood test statistic. Asymptotic critical values are obtained from the χ_1^2 distribution and their accuracy was verified by simulation. This material has been published in Caroni & Karioti [2].

Random AO: the case of a random AO was developed along the same lines as the analysis of the random IO.

Lifetime data

The methods of lifetime data analysis are used in studying survival and reliability (for example, the time until a patient dies, the time until a machine breaks down or the load under which a beam breaks). Outliers in lifetime data are unusually small or unusually large values. They may have a strong influence on the choice of model and on the estimates of the model's parameters. Some initial investigations of lifetime data models were undertaken in the course of this study (Economou & Caroni [4]).

Conclusions

Our two publications (Caroni & Karioti [2]; Karioti & Caroni [3]) are the first to present methods for detecting outliers in sets of time series. They represent a significant contribution to statistical methodology since data of this form are common in various areas of application of statistics. Further papers arising from this research project have appeared in the proceedings of various conferences (Economou & Caroni [4]; Karioti & Caroni [5], [6], [7]).

References

1. Barnett, V. and Lewis, T.: "*Outliers in Statistical Data*", 3rd ed., Wiley, 1994

2. Caroni, C. and Karioti, V.: "Detecting an innovative outlier in a set of time series", *Computational Statistics and Data Analysis* **46**, 561-570, 2004.

3. Karioti, V. and Caroni, C.: "Simple detection of outlying short time series", *Statistical Papers* **45**, 267-278, 2004.

Conference Papers

4. Economou, P. and Caroni, C.: "Investigation and development of models for survival analysis", *16th Panhellenic Statistics Conference, Hellenic Statistical Institute. Kavala,* 2003.

5. Karioti, V. and Caroni, C.: "Detecting an outlier in a set of time series", pp. 371-374. *Proceedings of the 17th International Workshop in Statistical Modelling. Chania, Greece*, 2002.

6. Karioti, V. and Caroni, C.: "Fixed and random innovative outliers in sets of time series". *International Workshop on Computational Management Science, Economics, Finance and Engineering. Limassol, Cyprus,* 2003.

7. Karioti, V. and Caroni, C.: " Detection of an additive or innovative outlier in a set of time series", *16th Panhellenic Statistics Conference, Hellenic Statistical Institute, Kavala*, 2003.