Finance Research Letters 13 (2015) 243–257

Contents lists available at ScienceDirect

Finance Research Letters

journal homepage: www.elsevier.com/locate/frl

The instability of the Pearson correlation coefficient in the presence of coincidental outliers ☆

Yunmi Kim [a], Tae-Hwan Kim [b,*], Tolga Ergün [c]

[a] Department of Economics, University of Seoul, Republic of Korea
[b] School of Economics, Yonsei University, Republic of Korea
[c] State Street Corporation, USA

ARTICLE INFO

Article history:
Received 26 May 2014
Accepted 12 December 2014
Available online 23 December 2014

JEL classifications:
C13
C14
C18
C46

Keywords:
Correlation
Robust statistic
Outliers

ABSTRACT

It is well known that any statistic based on sample averages can be sensitive to outliers. Some examples are the conventional moments-based statistics such as the sample mean, the sample variance, or the sample covariance of a set of observations on two variables. Given that sample correlation is defined as sample covariance divided by the product of sample standard deviations, one might suspect that the impact of outliers on the correlation coefficient may be neither present nor noticeable because of a 'dampening effect' i.e., the effects of outliers on both the numerator and the denominator of the correlation coefficient can cancel each other. In this paper, we formally investigate this issue. Contrary to such an expectation, we show analytically and by simulations that the distortion caused by outliers in the behavior of the correlation coefficient can be fairly large in some cases, especially when outliers are present in both variables at the same time. These outliers are called 'coincidental outliers.' We consider some robust alternative measures and compare their performance in the presence of such coincidental outliers.

☆ We are grateful to an anonymous referee and the editor Ramazan Gencay for their invaluable comments which substantially improved the paper.

* Corresponding author at: Yonsei University, School of Economics, 134 Shinchon-dong, Seodaemun-gu, Seoul 120-749, Republic of Korea. Tel.: +82 2 2123 5461; fax: +82 2 2123 8638.
E-mail address: tae-hwan.kim@yonsei.ac.kr (T.-H. Kim).

http://dx.doi.org/10.1016/j.frl.2014.12.005
1544-6123/

1. Introduction

The sample correlation coefficient is probably the most frequently-used statistic for measuring the linear co-movement between two variables. It has been documented (e.g., see Stigler, 1989) that the essential idea of correlation or 'co-relation' was conceived by Francis Galton and was formally developed by Karl Pearson, which explains why it is sometimes called the 'Pearson correlation coefficient.' Although the correlation coefficient does not measure the causal relationship between two variables, it plays an important role in many scientific areas. For example, understanding how financial assets are moving together which is measured by the correlation coefficient is crucial in lowering portfolio risk through diversification.

Based on the main idea put forward by Kim and White (2004), Bonato (2011), Ergun (2011) and White et al. (2010), an intuitively appealing and easily computable robust measure of covariance has been proposed by Huo et al. (2012). They demonstrated that the conventional measure of covariance is heavily influenced by outliers. Given that sample correlation is defined as sample covariance divided by the product of two sample standard deviations, one might suspect that the impact of outliers on the correlation coefficient may not be either present or noticeable because of a 'dampening effect,' i.e., the effects of outliers on both the numerator and the denominator of the correlation coefficient can cancel each other.

In this paper, we formally investigate this issue. We first derive the analytical expression of the distortion caused by outliers. Then we attempt to gauge the size of such a distortion in many different situations using Monte Carlo simulations. As expected, there is a 'dampening effect' due to the standardization in many cases under consideration. However, there also exists a surprising twist in other cases, in particular when outliers are present in both variables at the same time. These outliers are called 'coincidental outliers.' In such cases, the discovered distortion is fairly large. We consider some robust alternative measures and compare their performance in the presence of such coincidental outliers.

2. The effect of outliers on the conventional measure of correlation

We consider two stochastic processes $\{x_t\}_{t=1,\ldots,T}$ and $\{y_t\}_{t=1,\ldots,T}$ where x_t are assumed to be independent and identically distributed (IID) with the cumulative distribution function (CDF) F_x and y_t are also assumed to be IID with the CDF F_y. The conventional measure of correlation (denoted by C), called the 'Pearson correlation coefficient,' is given by:

$$C = \frac{E[(x_t - \mu_x)(y_t - \mu_y)]}{\sqrt{\sigma_x^2}\sqrt{\sigma_y^2}}, \tag{1}$$

where $\mu_x = E(x_t)$, $\mu_y = E(y_t)$, $\sigma_x^2 = E[(x_t - \mu_x)^2]$, $\sigma_y^2 = E[(y_t - \mu_y)^2]$, and the expectation E is taken with respect to the joint CDF of x_t and y_t. The conventional measure C is, of course, a population parameter and thus must be estimated. Its usual estimation is achieved by replacing the population expectation E with its corresponding sample mean:

$$\widehat{C} = \frac{\frac{1}{T}\sum_{t=1}^{T}[(x_t - \widehat{\mu}_x)(y_t - \widehat{\mu}_y)]}{\sqrt{\widehat{\sigma}_x^2}\sqrt{\widehat{\sigma}_y^2}}, \tag{2}$$

where $\widehat{\mu}_x = \frac{1}{T}\sum_{t=1}^{T} x_t$, $\widehat{\mu}_y = \frac{1}{T}\sum_{t=1}^{T} y_t$, $\widehat{\sigma}_x^2 = \frac{1}{T}\sum_{t=1}^{T}(x_t - \widehat{\mu}_x)^2$, and $\widehat{\sigma}_y^2 = \frac{1}{T}\sum_{t=1}^{T}(y_t - \widehat{\mu}_y)^2$.

The above sample correlation $\widehat{C}$ is based on sample averages; thus, it may be influenced by any outliers in either x_t or y_t as explained in Kim and White (2004) and Huo et al. (2012). To determine the influence of outliers on the conventional measure, we assume without loss of generality that a single outlier (denoted as m_x) occurs at time $[\tau T]$ with $\tau \in (0, 1)$ in x_t, and y_t also has an outlier (m_y) at time