Stock Returns and Investors' Mood: Good Day Sunshine or Spurious Correlation?

Kim, Jae

12 April 2016

Stock Returns and Investors' Mood: Good Day Sunshine or Spurious Correlation?

Jae H. Kim*
Department of Economics and Finance
La Trobe University, VIC 3086
Australia

Abstract

This paper examines the validity of statistical significance reported in the seminal studies of the weather effect on stock return. It is found that their research design is statistically flawed and seriously biased against the null hypothesis of no effect. This, coupled with the test statistics inflated by massive sample sizes, strongly suggests that the statistical significance is spurious as an outcome of Type I error. The alternatives to the p-value criterion for statistical significance soundly support the null hypothesis of no weather effect. As an application, the effect of daily sunspot numbers on stock return is examined. Under the same research design as that of a seminal study, the number of sunspots is found to be highly statistically significant although its economic impact on stock return is negligible.

Keywords: Anomaly, Behavioural finance, Data mining, Market efficiency, Sunspot numbers, Type I error, Weather.

JEL Classification: G12, G14.

April 2016

* Tel: +613 9479 6616; Email address: J.Kim@latrobe.edu.au

I would like to thank Xiangkang Yin for helpful comments on an earlier version of the paper.

1. Introduction

The question as to whether investors' mood affects the stock market (i.e. their emotional states or feelings unrelated to market fundamentals or rational pricing of financial assets) has been an issue of considerable interest in economics and finance (see, for a survey, Lucey and Dowling, 2005). The seminal studies in this literature are Saunders (1993) and Hirshleifer and Shumway (2003) where statistically significant weather effect on stock return is reported. Subsequent studies overall support the existence of the weather effect (cloudiness, sunshine, temperature, or wind) on stock return or other trading activities: see Cao and Wei (2005), Dowling and Lucey (2005, 2008), Goetzmann and Zhu (2005), Chang et al. (2008), Chang et al. (2006), Keef and Roush (2002, 2005, 2007), Yoon and Kang (2009), Kang et al. (2010), Lee and Wang (2011), Lu and Chou (2012), and Novy-Marx (2014). The literature has proliferated over the years in the publication of studies examining the effects of investors' moods derived from disparate sources such as: daylight saving (Kamstra et al. 2000), seasonal depression (Kamstra et al., 2003), sport events (Edmans et al., 2007; Kaplanski and Levi, 2010; Chang et al., 2012), lunar phases (Yuan et al., 2006; Keef and Khaled, 2012), pollution (Lepori, 2016), and the Ramadan (Bialkowski et al., 2012). Most of these studies report a statistically significant effect of investors' mood on the stock market, and their findings are presented as direct evidence for the anomalies against market efficiency.

On the other hand, there are studies that raise suspicions that a statistically significant weather effect may be the result of data mining or spurious correlation. In replicating Saunders' (1993) results using a German data set, Krämer and Runde (1997) report that statistical significance of the weather effect depends largely on how the null hypothesis is phrased. Trombley (1997) provides evidence that Saunders' (1993) results depend on the type of the return used and sample period employed. Loughran and Schultz (2004), in the context of localized trading of NASDAQ stocks, examine the weather effect in the city where the company is based and find

that the weather effect is too slight to establish a profitable weather-based trading strategy. They (p.363) state that "we would not dismiss the possibility that the relationship between cloud cover in New York and stock returns is spurious". Jacobsen and Marquering (2008) state that the documented weather effects might be the consequence of "data-driven inference based on spurious correlation", providing evidence that seasonal anomaly in stock return is unlikely to be caused by investors' mood changes due to weather variations.

The purpose of this study is to examine the validity of statistical significance reported in the two seminal studies of the weather effect on stock return, i.e. Saunders (1993) and Hirshleifer and Shumway (2003), in order to shed light on the possibility of a spurious relationship between investor's mood and stock return. First, paying attention to Hirshleifer and Shumway (2003), I evaluate whether their research design that "maximizes the power of the test by pooling the all available data jointly" is statistically sensible. This is important since many subsequent studies in this area (and also elsewhere in finance) adopt large or massive sample sizes in the same spirit. However, there is a danger that the use of massive sample size produces spurious statistical significance (see, for example, McCloskey and Ziliak, 1996). Second, statistical significance reported in these seminal studies is re-evaluated using the Bayesian method (Zellner and Siow, 1978) and the adaptive level of significance (Perez and Pericchi, 2014). These are the alternatives to the p-value criterion for statistical significance exclusively adopted in prior studies. Note that the American Statistical Association (Wasserstein and Lazar, 2016) recently issued a statement expressing grave concerns that improper use of the p-value criterion is distorting the scientific process and invalidating many scientific conclusions. Third, as an application, the effect of sunspot numbers on stock return is examined, using the same research design as that of Hirshleifer and Shumway (2003). This is to demonstrate that a variable with little economic relevance on stock return can be shown to be statistically significant through a simple data mining process.

The main finding of the paper is that statistically significant weather effect reported in the past studies is highly likely to be spurious and an outcome of Type I error. In particular, the research design employed by Hirshleifer and Shumway (2003) is problematic, with the probability of Type I error disproportionately higher than that of Type II. It is severely biased against the null hypothesis of no effect in its implied specification of loss function and prior probabilities. The alternatives to the p-value criterion show an overwhelming support for the null hypothesis of no weather effect. It is also demonstrated that a balanced research design in the context of the data employed by Hirshleifer and Shumway (2003) requires the sample size only under 2000. The results from the empirical application further confirm these findings. While suspicions concerning spurious statistical significance of the weather effect on stock return have been raised previously, this study is the first to assess the underlying statistical issues in the research design of the seminal studies and re-evaluates statistical significance of their results. In the next section, the research design of Hirshleifer and Shumway (2003) is examined. Section 3 presents further analyses based on the alternative criteria for statistical significance; and discussion on the effect size estimates reported in the past seminal studies. Section 4 presents the empirical application, and Section 5 concludes the paper.

2. Issues related with the Research Design

In this section, the research design of Hirshleifer and Shumway (2003) is discussed with reference to the possibility of spurious statistical significance. The weather effect is typically tested in the regression model of the form:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \ldots + \beta_K X_{Kt} + u_t, \quad (1)$$

where Y is the stock return, X_1 is a weather variable (e.g. cloud cover), and other X's represent the possible control variables. Hirshleifer and Shumway (2003) also consider the logit model where Y is an indicator variable. Under H_0: $\beta_1 = 0$, the weather has no effect on the stock

return. The t-test statistic can be written as $t = \frac{b_1}{s/\sqrt{n}}$, where b_1 is the least-square estimator for β_1, n is the sample size, and $s/\sqrt{n}$ denotes the standard error of b_1. Throughout the paper, following convention, the regression parameters are denoted as β_i's; and the probability of Type I and II errors as α and β, respectively (the power ≡ 1- β).

2.1 Background

One common feature of the studies of the weather effect is the use of large or massive sample sizes: a survey of twenty papers in the literature finds that the average sample size used is around 6000 with the maximum being 92808. In addition, they conduct their statistical tests almost exclusively at the conventional level of significance such as 0.05. A number of authors warn that spurious statistical significance may occur in this scenario (Neal, 1987, p. 524; Connolly, 1989, p. 139; McCloskey and Ziliak 1996; p.102). However, many researchers seem to believe that a large or massive sample size is necessarily a desirable feature of a research design, delivering strong power to their statistical tests. Hirshleifer and Shumway (2003, p.1014) justify their use of a massive panel data set, asserting that "the panel increases our power to detect an effect. … Given high variability of returns, it is useful to maximize power by using a large number of markets". However, as we shall see, this can cause statistical inference severely biased towards Type I error. The extreme power leads to an acute imbalance between the α and β, if a conventional level of significance is maintained. For example, suppose an extreme power (1-β) of 0.99999 is achieved by pooling a massive panel data set. If the researcher conducts a test at the 5% level (α= 0.05), the Type I error is 5000 times more likely to occur than the Type II error. As a result, if an error occurs, it is highly likely to be that of Type I, rejecting the true null hypothesis of no effect. This is particularly so when the effect size (e.g. the magnitude of the regression coefficient β_1 in (1)) is small.

The null hypothesis is often violated by an economically trivial deviation (see De Long and Lang, 1992). It is unrealistic that the null hypothesis of no effect holds exactly in practice: McCloskey and Ziliak (1996, p.98) provides an example in the context of the purchasing power parity. In reality, a null hypothesis is often violated by a negligible margin even when the true effect is economically unimportant. That is, $\beta_1 = 0+\Delta$, where Δ represents a deviation from the null hypothesis. As De Long and Lang (1992, p. 1269) find, all economic hypotheses are false with $\Delta \neq 0$, and the key question in empirical research is whether the value of Δ is large enough to be economically meaningful (see McCloskey and Ziliak, 1996). An important point is, even when the value of Δ is economically unimportant, the t-statistic (in absolute value) approaches infinity as the sample size increases. In this case, if a fixed level of significance is maintained, the probability of rejecting the null hypothesis approaches one as the sample size increases (see, for details, Kim and Ji, 2015, Section 5.1).

Moreover, as discussed in Section 3.3, the effect sizes reported in published seminal studies on the weather effect are indeed fairly small, which strongly suggests that the values of Δ is economically negligible. Hence, one may validly suspect that the statistically significant relationship between investors' mood and stock return reported in many studies are spurious, on the basis that their significance testing is conducted at the fixed conventional level (such as 0.05) under a large or massive sample size. A sensible strategy in this case is to adjust the level of significance as a decreasing function of sample size so that a reasonable balance between α and β is maintained (see, for example, Leamer, 1978, Chapter 4; and DeGroot and Schervish, 2012; Section 9.9). Alternatively, the researcher can choose a sample size in such a way that a reasonable balance between the two error probabilities is reached with a high statistical power.

In a recent statement, the American Statistical Association (ASA) contends that "Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough …" (Wasserstein and Lazar, 2016). The ASA further warns that "Widespread use of 'statistical significance' (generally interpreted as 'p < 0.05') as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process". Note that the empirical studies in the literature of weather effect on stock return exclusively use the "p-value less than 0.05" as the criterion for statistical significance under large or massive sample sizes, which is also the general practice in finance research (Kim and Ji, 2015). The ASA warns that this practice can lead to erroneous beliefs and poor decision-making (Wasserstein and Lazar, 2016). Gigerenzer (2004, p. 601) also argues that a low p-value (statistical significance) attained by a large sample size has little scientific value.

2.2 Problems with Power Maximization

Consider Hirshleifer and Shumway (2003), where the weather effect is tested using the simple regression between Y and X_1, where Y is stock return and $X_1 = SKC^*$ is a measure of average cloudiness at the city where the stock market is located. They collect the data from 26 stock markets around the world, each on average having the sample size of 3570. As mentioned earlier, Hirshleifer and Shumway (2003; p.1024) "design more powerful tests of the adverse weather explanation by considering all cities' returns jointly", which gives them the total number of pooled observations of 92808. However, this research design is problematic since it leads to an extreme imbalance between α and β values. As McCloskey and Ziliak (1996, p. 102) point out, "At such large sample sizes, the authors need to pay attention to the tradeoff between power and the size of the test, and to the economic significance of the power against alternatives". In addition, increasing the sample size to the point that statistical significance is achieved at a conventional level of significance constitutes an act of data mining.

I set the variance of the error term u in (1) to 1 (equal to the variance of stock return in percentage) and the standard deviation of X_1 to 2.19: these values are as reported in Hirshleifer and Shumway (2003). Under the assumption of normality, I calculate the power functions for H_0: $\beta_1 = 0$, plotted against the sample size (ranging from 10 to 100000) in Figure 1. That is, the probability of rejecting H_0: $\beta_1 = 0$ is plotted against increasing sample size, when the value of Δ is set at -0.007, -0.010, and -0.013, at the 5% level of significance. These three values are chosen on the basis that the estimated value of X_1 coefficient is -0.01 from the pooled regression of Hirshleifer and Shumway (2003, Table III), which means that daily stock return changes only by -0.01%, on average, in response to a one-unit increase of X_1. The plot shows that when the sample size is in the 2000 to 4000 range, the power is under 0.4. This is consistent with the individual stock market results reported in Hirshleifer and Shumway (2003, Tables III and IV), which show no strong evidence of statistical significance at the 5% level. However, the power reaches 1 when the sample size gets larger than 40000 when $\Delta = -0.01$. This indicates the pooled regression of Hirshleifer and Shumway (2003) conducted with the sample size of 92808 are clearly flawed at the 5% level of significance. The ratio of α to β (1-power) is infinite for a range of possible Δ values that are economically negligible. That is, Type I error occurs with the probability of one when the sample size as large as 92808 is employed, if the 5% level of significance is maintained.

2.3 Is a Larger Sample Size Informative?

Another common feature of the studies of the weather effect is that their R^2 values are seriously low. A low R^2 value indicates not only a poor in-sample fit, but also poor predictive ability of the model (both in-sample and out-of-sample). To this end, I conduct a survey of the past studies to collect the reported R^2 values and sample sizes. Figure 2 plots 94 pairs of R^2 and the sample size from these studies (listed at the bottom of the figure). As is clearly shown in the plot, the R^2 values are tiny with only handful of them higher than 0.10. Nearly 81% of

the R^2 values are less than 0.05, and around 48% of them are less than 0.01. This means that, for most cases, the regression models that include a weather variable can explain less than 5% of the total stock return variation. Since nearly all regressions use a number of control variables, the contribution of the weather variable can be a lot smaller. As indicated in Figure 2, there are studies which use the weather variable as the only explanatory variable, and their reported R^2 values are virtually zero. For example, Akhtari (2011) reports seven simple regressions with X_1 being the cloud cover (no other control variable), and the median of the reported R^2 values are 0.003. Goetzmann and Zhu (2005, Table 14) report one simple regression with X_1 being total sky cover, and the reported R^2 value is 0.01. This means the weather variable's contribution to the total variation of index return is negligible. Although not included in Figure 2, Loughran and Schultz (2004, Table 7) report the R^2 values from the regression of (localized) daily city portfolio returns on local and New York cloudiness, with the sample size varying between 3448 and 3529. From their 25 regression results reported, the largest adjusted R^2 value is 0.003. They (2004, p.359) note that these low values imply that little variation in return is explained by the weather variables.

In Figure 2, a negative relationship between the sample size and R^2 values is evident, which suggest either a negatively linear or a reciprocal functional form between the two. For the latter case, the red line in Figure 2 plots the line implied by the (R^2, n) relationship from the estimated regression:

$$R^2 = -0.002 + 32.52\, n^{-1} + 0.05\, E + 0.002 \log(K),$$

where K is the number of independent variables and E is the dummy variables for the study that use the equal-weighted returns. The latter is included as there is a strong tendency that the R^2 values from the regression with equal-weighted returns are higher than those associated with value-weighted returns, as indicated in Figure 2. The results indicate that the goodness-of-fit and predictive ability of the model deteriorate dramatically, as an additional sample is

incorporated into the models for the weather effect on stock return. From a linear model with n as an explanatory variable (instead of n^{-1}), the estimate of the elasticity of R^2 with respect to n is -0.45 (evaluated at means). This means that, if the researcher doubles the sample size, the value of R^2 is expected to decline by 45%. In other words, a larger sample does not contribute to the explanatory power of the model, but only inflates the value of the test statistic. This is further evidence that the research design of pooling as much as data as possible, employed by Hirshleifer and Shumway (2003), is not statistically sound and constitutes a data mining process.

2.4 Level of Significance and the Required Sample Size

The selection of the level of significance (α) is a critical element of statistical research. While a conventional level (0.05, 0.01, and 0.10) serves as a benchmark for Type I error, the level of Type II error (β) or power is often completely ignored (see MacKinnon, 2002, p. 633; Ziliak and McCloskey, 2008; Kim and Ji, 2005). If the researcher maintains a certain level of α, then the value of β can be controlled by selecting an appropriate sample size. For example, if the researcher wishes to maintain (α,β) = (0.05, 0.10), the sample size required can be obtained. The first question is how the relative error probabilities (α/β) should be determined. If the researcher wants to be more conservative regarding Type I error, the value of β should be set at a higher level than α. That is, (α,β) = (0.05, 0.10) means that the researcher wants to control the Type I error probability at 5%, allowing for Type II error twice more likely to occur than Type I error. This also means that the researcher wishes to conduct the test with a sufficiently high power of 0.90.

The choice of (α,β) combination may depend on the researcher's subjective assessment (priors) on the likelihood of H_0 and H_1 and the relative losses from Type I and II errors,

among others (Myers and Melcher, 1969; Leamer, 1978). The relative error probabilities may be written as:

$$\frac{\alpha}{\beta} = \frac{L_2}{L_1}\frac{P(H_1)}{P(H_0)},$$

where L_1 and L_2 are the losses from Type I and II errors; and $P(H_0)$ is the prior probability that H_0 is true and $P(H_1)=1-P(H_0)$. For example, if L_1 is larger than L_2; or the researcher strongly believes that the null hypothesis true with a high $P(H_0)$, then a low value of α should be chosen relatively to that of β. In the neutral case where $L_2 = L_1$ and $P(H_0) = P(H_1) = 0.5$, the value of β is set equal to that of α. Hirshleifer and Shumway (2003) aims to maximize the power (or minimize the value of β) by pooling as many as data points as possible. This leads to the value of α/β being extremely large or even infinite, if α is fixed at a conventional value. This case is equivalent to the situation where either $P(H_0)$ is extremely small; or the value of L_2 is extremely large relative to that of L_1. This means that the research design employed by Hirshleifer and Shumway (2003) is arbitrary in its specification of the loss function and prior probabilities. More importantly, their research design is extremely biased towards H_1 in the implied values of the losses and prior probabilities.

Once the choice is made for a desired value of (α,β), the required sample size can easily be calculated. In the regression context such as in equation (1), consider H_0: $\beta_1 = 0$ against H_1: $\beta_1 < 0$. Let the standard error estimator for the estimation of β_1 de denoted as $s/\sqrt{n}$, where s is the function of data (Y and X). Then, under the assumption of normality, the required sample size is determined as $n^* = (CR_{1-\beta} - CR_{\alpha})^2 \Delta^{-2} s^2$, where CR_{τ} represents the τth percentile from the standard normal distribution and Δ represents a deviation of β_1 from the value under H_0. Note that, in the event of H_1: $\beta_1 > 0$, the formula is $n^* = (CR_{\beta} - CR_{1-\alpha})^2 \Delta^{-2} s^2$. Table 1 presents the values of required sample size for a selection of reasonable (α, β) combinations

under a range of Δ values, in the context of Hirshleifer and Shumway's (2003) simple regression model. Again, the variance of the stock return is assumed to be 1 and Var(X_1) = 2.19^2, as reported in their paper. As might be expected, a larger sample size is required if the researcher wishes to have a smaller Type II error probability or a higher power, given the fixed values of α and Δ. When Δ = -0.01, the required sample sizes are a lot smaller than 92808 which Hirshleifer and Shumway (2003) use for their pooled regression, under a range of (α, β) combinations considered. When (α, β) = (0.05, 0.05), the sample size 92808 is justifiable only when the value of Δ is less than -0.005. This means that, in a well-designed research, the effect size of the weather on stock return implied by the sample size of 92808 is nearly zero. When the effect size as large as -0.10 is assumed, the required sample size ranges from 200 to 500 for all (α, β) combinations considered. These results indicate that the researchers in the studies of the weather effect on stock return use sample sizes that are too large to justify the conventional level of significance (such as 0.05) they almost exclusively employed.

2.5 Research Design in Other Related Studies

Note that the research designs implemented by Hirshleifer and Shumway (2003) are widely adopted in the literature on investors' mood effect on stock market. In the analysis of seasonal depression on stock return, Kamstra et al. (2003) use 12 daily stock indices, each over long time periods, with the sample sizes ranging from 3000 to 19000. They establish statistical significance at the conventional levels of significance, with tiny R^2 values. For the effect of sport matches, Edmans et al. (2007, Table II) use the data from 39 countries over an average of 4690 trading days, with the total of 182919 observations[1]; Kaplanski and Levy (2010) use the data of 14679 trading days covering the period from 1950 to 2007; while Chang et al. (2012) use the daily data from 1972 to 2004 for all firms in the CRSP and Compustat

[1] Häring and Storbeck (2009; p.222) raise the possibility of data mining and spurious correlation for the results of Edmans et al. (2007)

databases, with the total number of observations of more than 4 million. In analysing the effect of lunar phase, Yuan et al. (2006) use daily data from 48 stock markets from 1973 to 2001. Again, statistical significance in all of these studies is judged using the p-value criterion, at the level of significance not lower than 1%, in spite of massive sample sizes.

As Dyckman and Zeff (2014, p. 697) point out, selection of sample period or range is a key element of research design and a clear justification should be given for research findings to be convincing. Ioannidis and Doucouliagos (2013) also argue that the sample size is a key parameter of research design that affect the research credibility. Although the above studies do not provide such clear justifications, it seems that their intention is to maximize the power of their tests for the null hypothesis of no effect by pooling all available data points as possible, in the same spirit as Hirshleifer and Shumway (2003). For example, Yuan et al. (2006; p.5) state "a broad sample of 48 countries is examined, which constitutes a more comprehensive and powerful test". In response to Kelly and Meschke's (2010) criticisms on the existence of the effect of seasonal depression, Kamstra et al. (2012; p.935) defend their results with an argument that their joint tests using panel data are more powerful than single equation tests that Kelly and Meschke (2010) use. As we have seen earlier in this section, it is well expected that the power of the test is extreme under the sample sizes adopted by these studies. Furthermore, their tests are likely to be extremely biased towards Type I error if a conventional level of significance is maintained, rejecting the true null hypothesis too often.

3. Further Analyses and Discussions

All of the papers in the literature on weather effect use the "p-value less than 0.05" (or 0.01 and 0.10) as a sole criterion for statistical significance, as with many other areas of finance research (see, for example, Kim and Ji, 2015). However, the ASA (Wasserstein and Lazar, 2016) is gravely concerned with this research practice, stating that "Scientific conclusions and

business or policy decisions should not be based only on whether a p-value passes a specific threshold". They further state that the p-value should not be used as a measure of the importance or the evidence regarding a model or a hypothesis. They propose the Bayesian method of statistical inference and estimation-based methods such as the confidence interval as possible alternatives, which are not widely adopted in empirical research in economics and finance (see, for example, Kim and Ji, 2015).

The other problem associated with the p-value criterion is that its widely used thresholds (0.05, 0.01, 0.10) are arbitrary and lack scientific justifications (see, for example, Lehmann and Romano, 2005, p.57). Several authors have proposed methods for choosing the optimal threshold (or the level of significance) given the sample size, prior probabilities, and relative losses from Type I and II errors (Manderscheid, 1965; Leamer, 1978; DeGroot and Schervish, 2012). Kim and Ji (2015) provide an example of choosing the optimal level of significance for an asset pricing model, while Kim and Choi (2016) apply the method to unit root testing. An alternative is to use the Bayesian method of hypothesis testing, which implies the critical value as an increasing function of sample size (see, for example, Leamer, 1978). Additionally, Perez and Perichhi (2014) propose a simple adaptive rule that adjusts the level of significance as a decreasing function of sample size. In this section, I present these alternatives and re-evaluate the results reported in Saunders (1993) and Hirshleifer and Shumway (2003). I also discuss the effect size of the weather effects on stock return reported in these two seminal papers.

3.1 Alternatives to "p-value < 0.05" criterion

The Bayesian method of significance testing is based on the posterior odds ratio in favour of the alternative hypothesis (H_1) over the null (H_0), which is defined as

$$P_{10} \equiv \frac{P(H_1 \mid D)}{P(H_0 \mid D)} = \frac{P(D \mid H_1)}{P(D \mid H_0)} \frac{P(H_1)}{P(H_0)}, \tag{2}$$

where $P(H_i)$ is the prior probability for H_i ; D indicates the data; $P(D|H_i)$ is the marginal distribution of data under H_i; and $P(H_i|D)$ is posterior probability for H_i. Note that $B_{10} \equiv P(D|H_1)/P(D|H_0)$ is referred to as the Bayes factor. The evidence favours H_1 over H_0 if $P_{10} > 1$. In this paper and as documented in Kim and Ji (2015), I use the version of P_{10} proposed by Zellner and Siow (1979) which is given by:

$$P_{10} = \left[\frac{\pi^{0.5}}{\Gamma[0.5(k_0+1)]} \frac{(0.5v_1)^{0.5k_0}}{[1+(k_0/v_1)F]^{0.5(v_1-1)}} \right]^{-1}, \tag{3}$$

where F is the F-statistic for H_0, $\Gamma()$ is the gamma function and $v_1 = T - k_0 - k_1 - 1$, while k_0 is the number of X variables restricted under H_0 and k_1 is the number of those unrestricted. Note that the same expression in Kim and Ji (2015) does have a typo. For simplicity, it is assumed that $P(H_0) = P(H_1)$, which means that the researcher is neutral in the likelihood of H_0 and H_1, where $P_{10} = B_{10}$. According to Kass and Raftery (1995, p. 777), the evidence against H_0 is "not worth more than a bare mention" if $2\log(B_{10}) < 2$; "positive" if $2 < 2\log(B_{10}) < 6$; "strong" if $6 < 2\log(B_{10}) < 10$; and "very strong" if $2\log(B_{10}) > 10$, where log is the natural logarithm. Perez and Perichhi (2014) propose a simple adaptive rule for the level of significance derived by reconciling the Bayes factor and likelihood ratio principle. In the context of the regression given in (1) and H_0: $\beta_1 = \beta_2 = \ldots = \beta_q = 0$, their adaptive rule is written as:

$$\alpha^* = \frac{\left[\chi^2_\alpha(q) + q\log(n)\right]^{0.5q-1}}{2^{0.5q-1} n^{0.5q} \Gamma(0.5q)} \exp(-0.5\chi^2_\alpha(q)) \tag{4}$$

where q is the number of parameters under H_0 and the value of α is set at the most popular value of 0.05.

Saunders (1993, Table 2) reports the results of eight regressions for the weather effect on a range of U.S. stock returns, controlling for the lagged return, Monday effect, and January

effect. Note that the weather variable is formulated in such a way that the value of β_1 is positive under H_1. The daily data is for the period from 1927 to 1989, with the number of observations ranging from 6298 to 9990. The variation of the sample size depends on: firstly, the type of index used; and secondly, whether large index changes are excluded. With an observation that the weather effect is positive on 6 out of 8 regressions and statistically significant at the 5% level (or at a lower level such as 0.01%), Saunders (1993, p.1342) concludes that "New York city weather is significantly correlated with index return". In Table 2, the above two methods are applied to the regression results of Saunders (1993). According to the Bayes factor ($P_{10} = B_{10}$) given in (3), strong evidence against H_0 of no weather effect is found on only one occasion, which is the case for an equal-weighted index where large index changes have been excluded. Under the adaptive rule of Perez and Perichhi (2014) given in (4), the weather variable is statistically significant only in the two cases where equal-weighted indices are used. Hence, under these alternative criteria, statistical significance of the weather effect does not hold in general.

Table 3 presents the case of regression results reported in Hirshleifer and Shumway (2003, Tables III and IV), where the weather variable is formulated in such a way that the value of β_1 is negative under H_1. They consider daily data from 26 markets and estimate the weather effect using the data from individual markets as well as the pooled data. For the regression model with only the weather variable as an explanatory variable, seven regressions show statistically significant weather effects at the 5% level (one-tailed), including the pooled regression. On this basis, Hirshleifer and Shumway (2003) claim that the weather has an effect on stock return. However, according to the Bayes factor given in (3), none of the regression provides strong evidence against H_0. If the adaptive rule of Perez and Perichhi (2014) is used, only the pooled regression shows a statistically significant weather effect. For the regression model that includes other control variables, six regressions show a statistically

significant weather effect at the 5% level. However, both Bayesian method and adaptive rule do not support the weather effect on stock return for all regressions, including the pooled regression. Again, the alternatives to the p-value criterion are in strong support for the null hypothesis of no weather effect.

3.3. **Discussion on effect size**

In empirical research, the most important target of analysis is the effect size (see, for example, Ziliak and McCloskey, 2008). If it is large and economically meaningful, then it is most likely that the rejection of the null hypothesis does not represent the occurrence of Type I error. However, when it is small or economically unimportant, the question regarding the spurious statistical significance may arise, as discussed earlier. In Saunders (1993), the median value of the estimated coefficients of the cloud cover variable reported in his Table 2 is 0.000415. This value is obtained from the regression of daily stock return in percentage against the cloud cover variable, which take 1 if cloud cover is: 0%-20%; 0 if 30%-90%; and -1 if 100%. This means that one-unit change of cloudiness is expected to increase the daily stock return by 0.0004 in percentage terms. Subsequent studies which conduct the analysis in a similar setting as in Saunders (1993), for example Gotesman and Zhu (2005) and Akhtari (2011), report the coefficients of similar magnitude.

In Hirshleifer and Shumway (2003), the estimated coefficients of the weather variable is around -0.01 (when they are negative), from the regression of daily stock return in percentage and the sky cover variable whose standard deviation is 2.19. This means that a one-standard deviation increase of sky cover decreases daily stock return around -0.02%, on average. Hirshleifer and Shumway (2003) often use the signs of estimated coefficients as the evidence of weather effects. For example, regarding their regression results in their Table III, they (p.1019) comment that “7 out of 26 coefficients are statistically significantly negative”. In the

same table, they note 25 out of 26 coefficients in their logit regressions are negative, with a comment that "this is quite strong evidence that cloudiness is correlated with returns". While 25 negative values may show a strong degree of consistency, this cannot be the evidence for a strong and meaningful effect size (see MaCloskey and Ziliak, 1996). In addition, these negative signs may be the reflection of some other problems, such as estimation bias from model mis-specification (functional form or missing variables). The effect size of their logit regressions are also negligible. Their logit coefficients are around -0.020, which means that 1 unit increase in their cloud cover variable (*SKC**) is expected to increase the probability of the stock return being positive by 0.02. While Hirshleifer and Shumway (2003, p. 1011) claim that the magnitude of sunshine effect is substantial, they note that the profit from a weather-based investment strategy can be limited by trading cost and diversification. Note that small effect size estimates are not confined in these two seminal studies. For example, Loughran and Schultz (2004) also conclude that the US weather effect is too slight to develop profitable trading strategies (on NASDAQ stocks).

4. Application: Sunspot Numbers and Stock Return

As an application, I examine whether daily sunspot numbers can explain stock return. The data is abundantly available and widely analysed in statistics and physics. While its influence on the climate and weather changes has been hypothesized, its effect on weather in the long run or short run is not clear and yet to be fully established (see, Haigh, 2007; Hipel and McLeod, 1994, p. 191). Hence, it is difficult to establish economically that the number of daily sunspots can explain daily variations of stock index return[2]. In this section, under the research design of Hirshleifer and Shumway (2003), I demonstrate that the sunspot numbers have statistically significant effect on stock return using the p-value criterion. Then I show

[2] See also: https://www.cxoadvisory.com/3942/calendar-effects/sunspot-cycle-and-stock-returns/

that this statistical significance cannot stand under the alterative criteria for statistical significance.

I collect daily data of sunspot numbers and stock return from January 1988 to February 2016 (7345 observations)[3]. The log returns in percentage are calculated from the MSCI price index obtained from DataStream. I cover 24 markets around the world, which are the same 26 markets as in Hirshleifer and Shumway (2003) except for the Milan and Johannesburg markets whose MSCI indices are not available from DataStream. The markets included are Amsterdam, Athens, Bangkok, Brussels, Buenos Aires, Copenhagen, Dublin, Helsinki, Istanbul, Kuala Lumpur (KL), London, Madrid, Manila, New York, Oslo, Paris, Rio de Janeiro (Rio), Santiago, Singapore, Stockholm, Sydney, Taipei, Vienna, and Zurich. The total number of pooled observations are 176280. Following Hirshleifer and Shumway (2003), I consider simple regression of stock return against the sunspot numbers; and pool the data from these cross-sectional units in order to maximize the power of the test for the null hypothesis of no sunspot effect. To highlight the effect of increasing sample size on statistical significance, I conduct pooled regressions by accumulatively pooling the data from the Amsterdam to Zurich markets (increasing the sample size from 7345 to 176280).

Figure 3 plots the estimated slope coefficients of the regression (black line) and R^2 values (red line) of stock return against sunspot numbers, as the additional sample is pooled into the regression. For all regressions, the estimated coefficient is tiny, no more than 0.0005 for most cases. The standard deviation of sunspot numbers is around 75, and this estimated coefficient means that one-standard deviation increase of sunspot numbers is expected to increase the daily stock return by no more than 0.0375%. Hence, the impact of sunspot numbers on stock

[3] The sunspot numbers are obtained from https://www.quandl.com/data/SIDC/SUNSPOTS_D-Total-Sunspot-Numbers-Daily

return is fairly small although it is not clear if the positive sign is economically sensible, given the unknown effect of sunspot number on stock return. As might be expected, the R^2 values are also tiny, with the maximum value is close to 0.0003. The increasing sample size does not improve the goodness-of-fit either. Hence, the estimated coefficient (effect size) and R^2 values indicate that the sunspot numbers show little effect on stock return.

Figure 4 plots the corresponding t-statistics. As might be expected, the t-statistic increases with sample size; and it becomes greater than the 5% one-tailed critical value of 1.645 when the sample size reaches 44070 or more. The t-statistics are often larger than 3, with which one may claim the existence of strong statistical significance. Hence, based on the "p-value less than 0.05" criterion, a statistically significant relationship between sunspot numbers and stock return is established, despite negligible effect size and goodness-of-fit of the model. The blue line in Figure 4 indicates the critical values associated with the adaptive level of significance of Perez and Pericchi (2014) given in (4). For all cases but one, the t-statistics are less than the critical values associated with the adaptive levels of significance, indicating no statistical significance. Figure 5 plots the $2\log(B_{10})$ values obtained from (3) from the same regressions, again as the sample size increases with data pooling. On no occasions is this value greater than 6, which is the critical value for strong evidence against H_0, while it is greater than 2 (critical value for positive evidence against H_0) for only one case. Hence, if alternative criteria for statistical significance are considered, statistical significance based on the p-value criterion cannot be defended.

Given that the explanatory variable is identical for all cross-sectional units and its effect sizes on stock returns are negligible, it is apparent that the sample size is the major contributing factor to an increasing t-statistic. This indicates that a statistically significant result at a conventional level is spurious. Note that for the sunspot data, a research design based on (α,β)

= (0.05, 0.05) requires a sample size of around 2000, when the value of economically meaningful deviation from the null hypothesis is set at 0.001. Again, the massive sample sizes adopted in the pooled regressions are too large to justify the conventional level of significance.

5. Conclusion

The question as to whether investors' mood influences stock return has strong implications for many areas of finance, such as asset pricing, behavioural finance, and the study of market efficiency. Many studies report statistically significant relationship, which has been presented as the direct evidence for the anomalies of market efficiency. This paper questions the validity of statistical significance reported in the two seminal studies in this literature, which study the weather effect on stock return: Saunders (1993) and Hirshleifer and Shumway (2003). The studies in this literature (as with generally the case in empirical finance research as reported in Kim and Ji, 2015) typically adopt the research design favouring a large or massive sample size, conducting significance tests using the p-value criterion with a fixed level of significance. This has a strong potential for spurious statistical significance, as pointed out in Neal (1987, p. 524), Connolly (1989, p. 139), McCloskey and Ziliak (1996) and Kim and Ji (2015). The possibility of data mining or spurious correlation also has been raised by Krämer and Runde (1997), Trombley (1997), Loughran and Schultz (2004), and Jacobsen and Marquering (2008).

It is found that the statistical significance claimed in Saunders (1993) and Hirshleifer and Shumway (2003) are indeed questionable. Their research designs that favour a large or massive sample size can lead to a serious imbalance between Type I and II error probabilities. The sample sizes they adopt are too large to justify the level of significance they employed. In addition, their test is severely biased against the null hypothesis of no weather effect, in its implicit specification of loss function and prior probabilities. These points, combined with fairly small effect size estimates, strongly suggest that the reported statistical significance represent the occurrence of Type I error (rejection of the true null hypothesis of no weather

effect). As alternatives to the p-value criterion for statistical significance, the Bayes factor and adaptive level of significance are applied to the results of these seminal studies. These alternatives strongly support the null hypothesis of no weather effect. As an application, the effect of daily sunspot numbers on stock return is examined, following the research design of Hirshleifer and Shumway (2003). By pooling the data from a large number of stock markets jointly, the number of daily sunspot numbers is found to be statistically significant for stock return using the p-value criterion. This is in spite of the fact that the estimated effect sizes are negligible and the R^2 values are tiny. If the alternative criteria for statistical significance are adopted, the statistical significance of sunspot numbers cannot be supported.

The recent statement made by the American Statistical Association (Wasserstein and Lazar, 2016) sends a clear warning that improper use of statistical methods are making the validity of scientific conclusions questionable. As Wasserstein and Lazar (2016) point out, the p-value criterion is problematic when it is used with a large or massive sample size and its mindless use as a licence for statistical significance is distorting the scientific process. De Prado (2015) also warns that "empirical finance is in crisis", with many statistical and mathematical tools used by empirical researchers being flawed. The studies of the weather effect on stock return represent a strand of literature in empirical finance where such flawed statistical tools may have distorted scientific findings. More broadly, empirical finance is an area of research where the data is abundant, and a large proportion of published studies favour large or massive sample sizes (see Kim and Ji, 2015). It is most likely that the researchers adopt this strategy in order to maximize the power in their statistical inference, in the same spirit as in Hirshleifer and Shumway (2003). However, as shown in this paper, this practice represents a poor research design which can deliver spurious statistical significance, especially when it is combined with the exclusive use of the "p-value less than 0.05" criterion. More sensible alternatives have been discussed in this paper.

References

Akhtari, M. 2011, Reassessment of the Weather Effect: Stock Prices and Wall Street Weather, *Undergraduate Economic Review*, 7(1), http://digitalcommons.iwu.edu/uer/vol7/iss1/19.

Bialkowski, J., Etebari, A., Wisniewski, T.P. 2012, Fast profits: Investor sentiment and stock returns during Ramadan, *Journal of Banking & Finance*, 36, 835-845.

Cao, M. Wei, J. 2005, Stock market returns: A note on temperature anomaly, *Journal of Banking & Finance*, 29, 1559-1573.

Chang, S-C., Chen S-S, Chou, R. K., Lin Y-H. 2012, Local sports sentiment and returns of locally headquartered stocks: A firm-level analysis, *Journal of Empirical Finance*, 19, 309-318.

Chang, T. C.-C. Nieh, M.J. Yang, T.-Y. Yang 2006, Are stock market returns related to the weather effects? Empirical evidence from Taiwan, *Physica A*, 364, 343-354.

Connolly, R. A. 1989, An Examination of the Robustness of the Weekend Effect, *The Journal of Financial and Quantitative Analysis*, 24(2), 133-169.

Connolly, R. A. 1991, A posterior odds analysis of the weekend effect, *Journal of Econometrics*, 49, 51-104.

DeGroot, M. H. and Schervish, M. J. 2012, Probability and Statistics, 4th edition, Addison-Wesley, Boston.

De Long, J.B. and Lang, K. 1992, Are all Economic Hypotheses False? *Journal of Political Economy*, 100(6), 1257-1271.

De Prado, M. L. 2015, The Future of Empirical Finance, *Journal of Portfolio Management*, Summer, 140-144.

Dyckman, T. R. and Zeff, S. A. 2014, Some Methodological Deficiencies in Empirical Research Articles in Accounting, *Accounting Horizons*, 28(3), 695-712.

Edmans, Alex, García, Diego, Norli, Øyvind 2007, Sports sentiment and stock returns. *Journal of Finance*, 62, 1967–1998.

Gigerenzer, G. 2004, Mindless statistics: Comment on "Size Matters", *Journal of Socio-Economics*, 33, 587-606.

Goetzmann, W.N., Zhu, N. 2005, Rain or shine: where is the weather effect? *European Financial Management,* 11, 559–578.

Haigh, J. D. 2007, The Sun and the Earth's Climate, Living Reviews in Solar Physics, http://www.livingreviews.org/lrsp-2007-2.

Häring, N. and Storbeck, O. 2009, Economics 2.0: What the Best Minds in Economics Can Teach You about Business and Life, Palgrave.

Hipel, K. W., McLeod. A. I. 1994, *Time Series Modelling of Water Resources and Environmental Systems*, Elsevier, Amsterdam.

Hirshleifer, D., & Shumway, T. 2003, Good day sunshine: Stock returns and the weather, *Journal of Finance*, 58(3), 1009– 1032.

Ioannidis, J.P.A. and Doucouliagos, C. 2013, What's to know about credibility of empirical economics? *Journal of Economic Surveys*, 27, 5, 997–1004.

Jacobsen, B., Marquering W. 2008, Is it the weather? *Journal of Banking & Finance*, 32, 526-540.

Kamstra, Mark J., Kramer, Lisa A., Levi, Maurice D. 2000, Losing Sleep at the Market: The Daylight Saving Anomaly, *American Economic Review*, 90(4), 1005–11.

Kamstra, M. J., Kramer, L. A., & Levi, M. D. 2003, Winter blues: A sad stock market cycle, American Economic Review, 93(1), 324–343.

Kamstra, M. J., Kramer, L. A., & Levi, M. D. 2012, A careful re-examination of seasonality in international stock markets: Comment on sentiment and stock returns, Journal of Banking and Finance, 36, 934-956.

Kang, S., Jiang, Z., Lee, Y., Yoon, S., 2010, Weather effects on the returns and volatility of the Shanghai stock market, *Phys. A Stat. Mech. Appl*., 389, 91–99.

Kaplanski, G., Levy. H. 2010, Exploitable Predictable Irrationality: The FIFA World Cup Effect on the U.S. Stock Market, *Journal of Financial and Quantitative Analysis*, 45(2), 535-553.

Keef, S. P. and Khaled, M. S. 2011, Are investors moonstruck? Further international evidence on lunar phases and stock returns, *Journal of Empirical Finance*, 18, 56-63.

Keef, S. P. and Roush, M. L. 2002, The weather and stock returns in New Zealand, *Quarterly Journal of Business and Economics*, 41, 61–79.

Keef, S. P. and Roush, M. L. 2005, The influence of weather on New Zealand financial securities, *Accounting and Finance*, 45, 415–37.

Keef, S.P., Roush, M.L. 2007, Daily weather effects on the returns of Australian stock indices, *Applied Financial Economics,* 17, 173-184.

Kelly, P.J., Meschke, F., 2010. Sentiment and stock returns: the SAD anomaly revisited. Journal of Banking and Finance 34, 1308 – 1326.

Kim, J. H., Choi, I. 2016, Unit Roots in Economic and Financial Time Series: A Re-Evaluation at the Optimal Level of Significance.
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2700659

Kim, J. H., Ji, P. 2015, Significance Testing in Empirical Finance: A Critical Review and Assessment, *Journal of Empirical Finance*, 34, 1-14.

Krämer, W. and Runde, R. 1997, Stocks and the weather: an exercise in data mining or yet another capital market anomaly? *Empirical Economics*, 11, 637–41.

Leamer, E. 1978, *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, Wiley, New York.

Lehmann E.L. and Romano, J.S. 2005, *Testing Statistical Hypothesis*, 3rd edition, Springer, New York.

Lepori, G.M. 2015, Air pollution and stock returns: Evidence from a natural experiment, *Journal of Empirical Finance*, 35, 25-42.
Loughran, T. and Schultz, P. 2004, Weather, stock returns and the impact of localized trading behaviour, *Journal of Financial and Quantitative Analysis*, 39, 343–64.

Lucey B., Dowling, M. 2005, The role of feelings in investor decision making, *Journal of Economic Surveys*, 19, 2, 211-237.

MacKinnon, J. G. 2002, Bootstrap inference in Econometrics, *Canadian Journal of Economics*, 35(4), 615-644.

Manderscheid, L.V. 1965, Significance Levels-0.05, 0.01, or ?, *Journal of Farm Economics*, 47(5), 1381-1385.

McCloskey, D. and Ziliak, S. 1996, The standard error of regressions, *Journal of Economic Literature*, 34, 97–114.

Myers, L.B., and Melcher J.A. 1969, On the Choice of Risk Levels in Managerial Decision-Making, *Management Science*, 16(2), B31-B39.

Neal, R. 1987, Potential Competition and Actual Competition in Equity Options, *The Journal of Finance*, 42(3), 511-53.

Noby-Marx, R. 2014, Predicting anomaly performance with politics, the weather, global warming, sunspots, and the stars, *Journal of Financial Economics*, 112, 137-146.

Perez, M-E, Pericchi, L.R. 2014, Changing statistical significance with the amount of information: The adaptive α significance level, *Statistics and Probability Letters*, 85, 20-24.

Saunders, E. M. 1993, Stock prices and wall street weather, *American Economic Review*, 83(5), 1337-1345.

Trombley, M. A. 1997, Stock prices and Wall Street weather: additional evidence, *Quarterly Journal of Business and Economics*, 36, 11–21.

Wasserstein R. L. Lazar, N. A. 2016, The ASA's statement on p-values: context, process, and purpose, *The American Statistician*, DOI: 10.1080/00031305.2016.1154108

Yoon, S.M., Kang, S.H. 2009, Weather effects on return: evidence from the Korean stock Market, Physica A, 388, 682–690.

Yuan, K., Zheng, L., Zhu, Q. 2006, Are investors moonstruck? Lunar phases and stock returns. *Journal of Empirical Finance*, 13, 1–23.

Zellner, A. and Siow, A. 1979, Posterior odds ratio of selected regression hypotheses, http://dmle.cindoc.csic.es/pdf/TESTOP_1980_31_00_38.pdf.

Ziliak, S. T., McCloskey, D.N. 2008, *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Ann Arbor, The University of Michigan Press.

Table 1. Required Sample Size: Sunshine Regression (Hirshleifer and Shumway, 2003)

	α	β	α	β	α	β	α	β
	0.05	0.05	0.05	0.10	0.05	0.01	0.01	0.01
Δ								
-0.005	90258		71423		131527		180543	
-0.010	22564		17855		32881		45135	
-0.050	902		714		1315		1805	
-0.100	225		178		328		451	
-0.150	100		79		146		200	
-0.200	56		44		82		113	

The results are obtained under the assumption that the variance of stock return (or error term) is one percent and the standard deviation of the cloud cover variable (SKC*) is 2.19, which are the values reported in Hirshleifer and Shumway (2003).

Table 2. Re-evaluation of the results obtained by Saunders (1993, Table 2)

Index	Type	*n*	t-statistic	$2\log(B_{10})$	CR*
	All Data				
DJIA	Value-Weighted	9990	1.33	-7.89	3.41
DJIA	Value-Weighted	6911	2.72*	-1.90	3.36
NYSE/AMEX	Value-Weighted	6911	3.27*	1.39	3.36
NYSE/AMEX	Equal-Weighted	6911	3.65*	4.02	3.36*
	Large Index Changes Excluded				
DJIA	Value-Weighted	8694	1.18	-8.13	3.39
DJIA	Value-Weighted	6298	2.72*	-1.80	3.34
NYSE/AMEX	Value-Weighted	6298	3.29*	1.62	3.34
NYSE/AMEX	Equal-Weighted	6298	3.91*	6.07*	3.34*

t-statistic: t-test statistic for the coefficient of cloud cover in the regression equation (1) of Saunders (1993). The starred t-statistics are those significant at the 5% level of significance (one-tailed).
2log(B10): Posterior odds ratio (Bayes factor) in log-likelihood scale given in (3).
Evidence against H_0 is strong if $2\log(B_{10}) > 6$, "not worth more than a bare mention" if $2\log(B_{10}) < 2$; and "positive" if $2 < 2\log(B_{10}) < 6$. The starred value in $2\log(B_{10})$ column indicates strong evidence.
CR*: Critical value based on the adaptive level of significance of Perez and Pericchi (2015) given in (4). The starred value indicates the case where t-statistic is greater than these critical values.

Table 3. Re-evaluation of the results obtained by Hirshleifer and Shumway (2003, Tables III and IV)

		Sunshine Regression (k=1)			Sunshine Regression Controlling for other Weather Conditions (k=3)		
Location	*n*	t-stat	$2\log(B_{10})$	CR*	t-stat	$2\log(B_{10})$	CR*
Amsterdam	3984	-1.07	-7.60	-3.27	-0.69	-8.26	-3.27
Athens	2436	0.71	-7.75	-3.20	0.77	-7.66	-3.20
Bangkok	3617	0.45	-8.44	-3.26	0.63	-8.25	-3.26
Brussels	3997	-3.25*	1.80	-3.27	-1.76*	-5.65	-3.27
Bueno Aires	2565	-0.98	-7.34	-3.21	-0.64	-7.89	-3.21
Copenhagen	4042	-0.30	-8.67	-3.28	-0.10	-8.75	-3.28
Dublin	3963	-0.002	-8.74	-3.27	0.08	-8.73	-3.27
Helsinki	2725	-1.67*	-5.57	-3.21	-1.51	-6.08	-3.21
Istanbul	2500	0.32	-8.17	-3.20	0.33	-8.17	-3.20
Johannesburg	3999	0.47	-8.52	-3.27	0.28	-8.67	-3.27
KL	3863	0.26	-8.64	-3.27	0.38	-8.57	-3.27
London	4003	-1.52	-6.44	-3.27	-1.14	-7.45	-3.27
Madrid	3760	-1.60	-6.12	-3.26	-1.42	-6.67	-3.26
Manila	2878	0.83	-7.73	-3.22	0.63	-8.02	-3.22
Milan	3961	-2.03	-4.62	-3.27	-1.99*	-4.78	-3.27
New York	4013	-1.28	-7.11	-3.27	-0.31	-8.65	-3.27
Oslo	3877	-1.92*	-5.03	-3.27	-1.76*	-5.62	-3.27
Paris	3879	-1.27	-7.10	-3.27	-1.53	-6.37	-3.27
Rio	2988	-1.93*	-4.73	-3.23	-2.16*	-3.79	-3.23
Santiago	2636	0.05	-8.33	-3.21	0.17	-8.30	-3.21
Singapore	3890	0.37	-8.58	-3.27	0.32	-8.61	-3.27
Stockholm	3653	-1.54	-6.28	-3.26	-1.22	-7.17	-3.26
Sydney	4037	-1.96*	-4.92	-3.28	-1.51	-6.47	-3.28
Taipei	3784	-0.97	-7.75	-3.27	-1.13	-7.41	-3.27
Vienna	3907	-2.14*	-4.15	-3.27	-2.00*	-4.72	-3.27
Zurich	3851	-1.28	-7.07	-3.27	-0.31	-8.61	-3.27
All cities	92808	-3.96*	3.79	-3.72*	-3.47*	0.15	-3.72

t-statistic: t-test statistic for the coefficient of cloudiness in the regression equation (1) of Hirshleifer and Shumway.

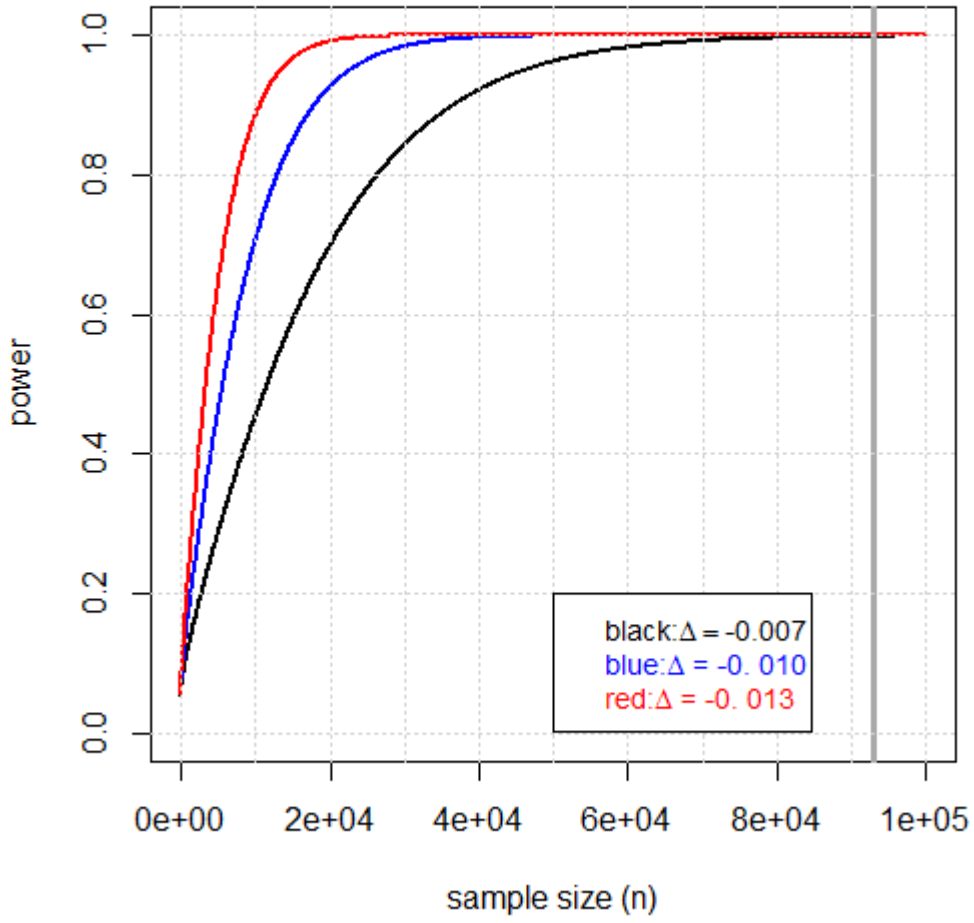
The starred t-statistics are those significant at the 5% level of significance (one-tailed).

$2\log(B_{10})$: Posterior odds ratio (Bayes factor) in log-likelihood scale given in (3).

Evidence against H_0 is strong if $2\log(B_{10}) > 6$, "not worth more than a bare mention" if $2\log(B_{10}) < 2$; and "positive" if $2 < 2\log(B_{10}) < 6$.The starred value in $2\log(B_{10})$ column indicates strong evidence.

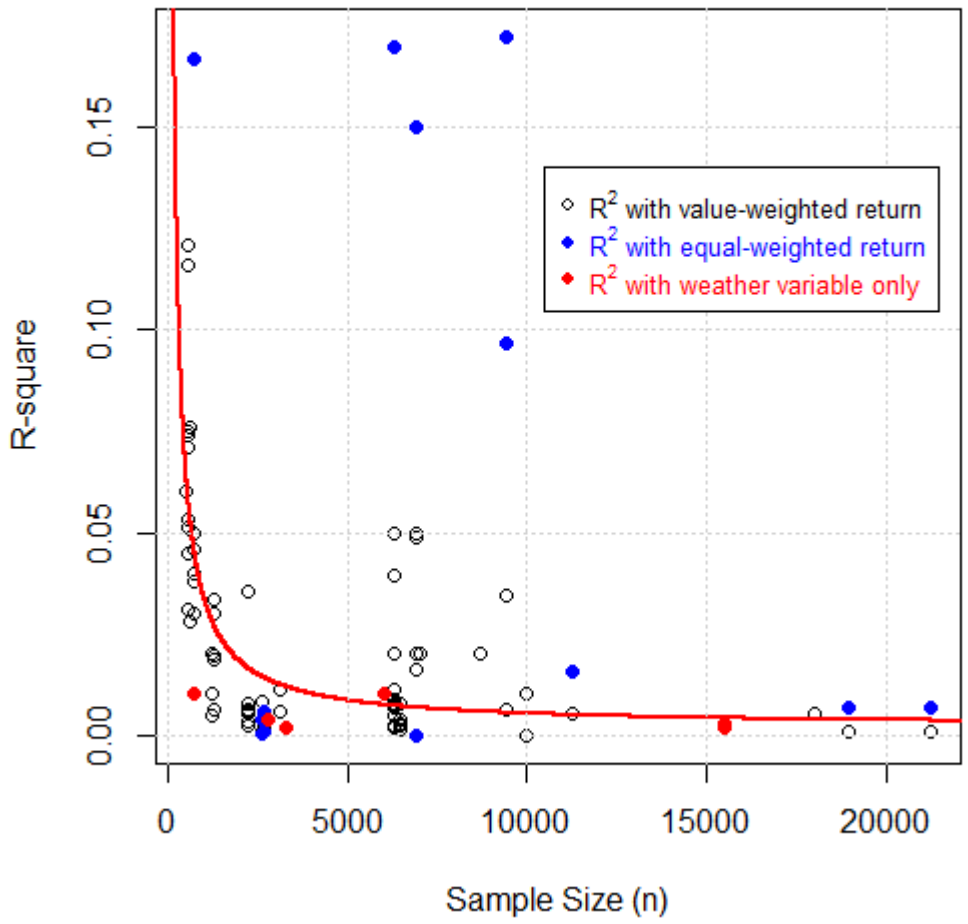
CR*: One-tailed critical value based on the adaptive level of significance of Perez and Pericchi (2015) given in (4). The starred value indicates the case where t-statistic is less than these critical values.

Figure 1. Power functions for the test regarding weather effect on stock return



The power functions (in blue) are associated with H0: $\beta_1 = 0$; H_1: $\beta_1 = \Delta$, where $\Delta \in$ (-0.007, -0.010, -0.013) in the context of the regression equation in Hirshleifer and Shumway (2003). The black line is associated with Δ = -0.007, while the blue one is linked with -0.010 and the red one with -0.013. The grey vertical line indicates the sample size of 92808, which is the sample size of the pooled regression of Hirshleifer and Shumway (2003).

Figure 2. R^2 and sample size derived from the regressions results of stock return on weather

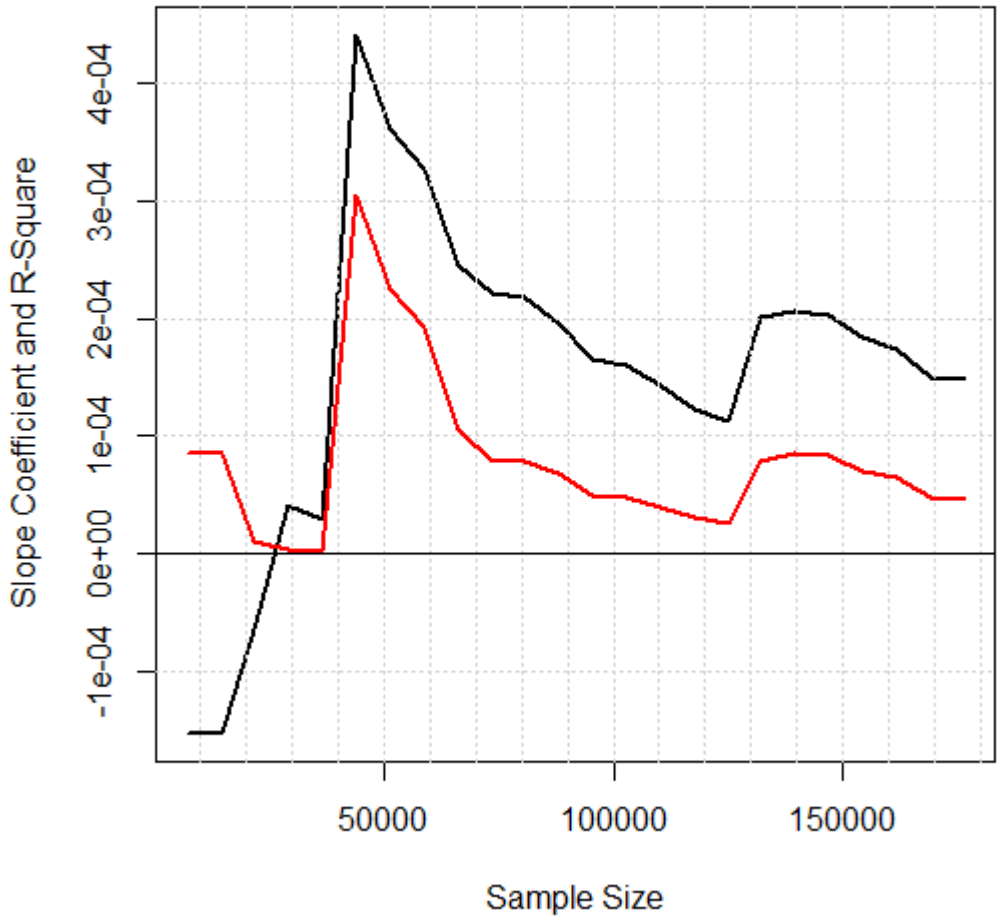


The (adjusted) R^2 values and sample sizes are obtained from Saunders (1993), Trombley (1997), Cao and Wei (2005), Goetzsmann and Zhu (2005), Chang et al. (2008), Akhtari (2011), and Lu and Chou (2012). In Cao and Wei (2005), the sample sizes of several regressions are approximated based on the information provided in their paper. Hirshleifer and Shumway (2003) do not report the R^2 values.

The blue dots are associated with the regressions using equally-weighted return, while black open dots originate from the regression with value-weighted return. The red dots are associated with the regression with the weather variable as the only explanatory variable.

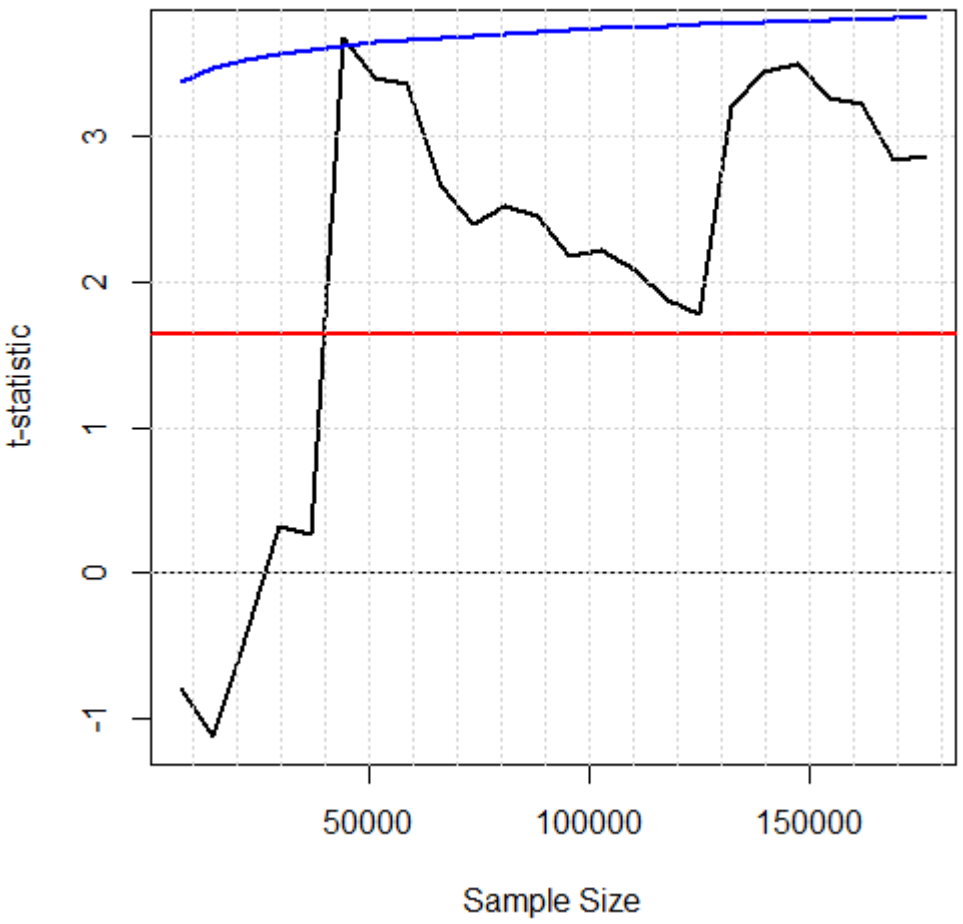
The red line plots (R^2, n) pairs implied by the regression $R^2 = -0.002 + 32.52\ n^{-1} + 0.05\ \mathrm{E} + 0.002\ \log(\mathrm{K})$.

Figure 3. Sunspot Regression: Effect Size Estimates and R^2 values



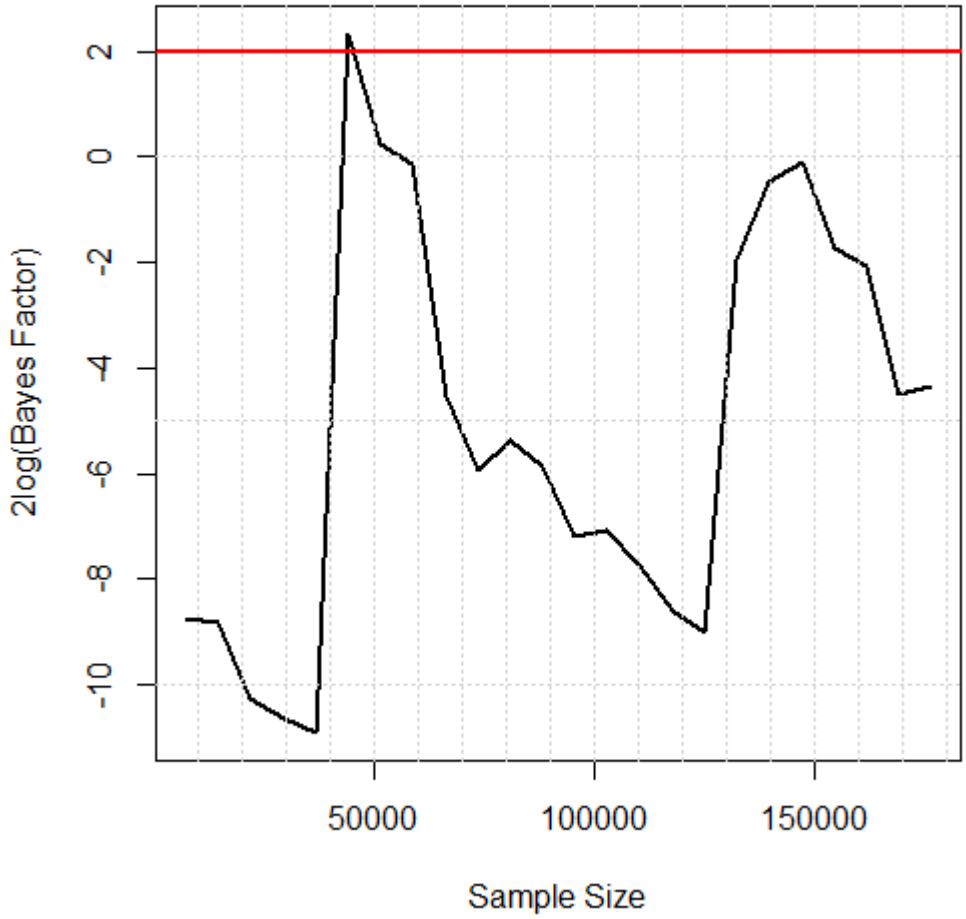
The estimated slope coefficients from the regression of stock returns against sunspot numbers are plotted in black, as additional cross-sectional units are progressively included in the model (from Amsterdam to Zurich markets). The red line indicates the value of R^2 from each regression. The sample size ranges from 7345 to 176280.

Figure 4. Sunspot Regression: t-statistics for H_0 of no effect on stock return



The t-statistics for the null hypothesis that the slope coefficient is zero from the regression of stock returns against sunspot numbers are plotted, as additional cross-sectional units are progressively included in the model (from Amsterdam to Zurich markets). The sample size ranges from 7345 to 176280.
The red horizontal line corresponds to 1.645, which is the 5% one-tailed critical value.
The blue line indicates the one-tailed critical values associated with the adaptive level of significance given in (4).

Figure 5. Sunspot Regression: $2\log(B_{10})$ for H_0 of no effect on stock return



The Bayes factor in (3) (in log-likelihood scale) for the slope coefficient from the regression of stock returns against sunspot numbers is plotted, as additional cross-sectional units are progressively included in the model (from Amsterdam to Zurich markets). The sample size ranges from 7345 to 176280.
Evidence against H_0 is strong if $2\log(B_{10}) > 6$, "not worth more than a bare mention" if $2\log(B_{10}) < 2$; and "positive" if $2 < 2\log(B_{10}) < 6$. The red horizontal line corresponds to 2.