Stock Market Anomalies: 
An Extreme Bounds Analysis

Jae H. Kim*  
Abul Shamsuddin†  

October 17, 2019

Abstract
We conduct the extreme bounds analysis (EBA) to evaluate the robustness or fragility of a range of stock market anomalies, focusing on the U.S. market and using daily data from 1960. The EBA is a large-scale sensitivity analysis under model uncertainty, capable of isolating the effects of potential data-mining or p-hacking. The anomalies covered include the effects of Halloween, sports event (the FIFA World Cup), seasonal affective disorder (winter blues), weather (cloud cover), political cycle (U.S. presidential party), daylight saving, and lunar phase. We find that the empirical evidence for these anomalies is highly fragile, in terms of effect size estimates and their statistical significance.

Keywords: Data-mining, Market efficiency, Model uncertainty

JEL Classification: C52, G12

*Department of Economics and Finance, La Trobe University, Australia; Email: J.Kim@latrobe.edu.au
†Newcastle Business School, University of Newcastle, Australia; Email: A.Shamsuddin@newcastle.edu.au
1 Introduction

A large number of stock market anomalies have been identified in empirical finance. They include a range of calendar effects, such as the day-of-the-week effect (Gibbons and Hess, 1981), the Monday effect (Chang et al., 1993; and Abraham and Ikenberry, 1994), weekend effect (French, 1980; Keim and Stambaugh, 1984; Jaffe and Westerfield, 1985; Connolly, 1989), January effect (Keim 1983; Rozeff and Kinney, 1976), effect of tax days (Dyl and Maberly, 1992), and turn-of-the-month effect (Ariel, 1987) and holiday (non-weekend) effect (Lakonishok and Smidt 1988; Kim and Park, 1994). Other anomalies related to investors’ mood include the Halloween effect (Bouman and Jacobsen, 2002), seasonal affective disorder (or winter blues) (Kamstra et al., 2003), FIFA World Cup (Kaplaniski and Levy, 2010a), daylight saving (Kamstra et al., 2000), political cycles (Santa-Clara and Valkanov, 2003; Novy-Marx, 2014), lunar phases (Yuan et al., 2006), and weather (rain, cloudiness, and temperature; Saunders, 1993; Hirshliefer and Shumway, 2003; Cao and Wei, 2005). Other events or factors that have been found to affect stock markets include Ramadan (Bialkowski et al., 2010), aviation disasters (Kaplaniski and Levy, 2010b), sunspot numbers (Novy-Marx, 2014), terrorism activity (Drakos, 2010), among many others. They stand as the evidence directly against the market efficiency.

A number of authors, however, have raised concerns that these anomalies may be the outcomes of data-mining or p-hacking. For example, Black (1993; p.75) argues that “most of the so-called anomalies that have plagued the literature on investments seem likely to be the result of data mining”. Based on White’s (2000) reality check for data-snooping, Sullivan et al. (2001) provide evidence that the observed calendar effects are likely to be the outcome of data-mining. Harvey (2017) is concerned with widespread practice of p-hacking in finance research, with a claim of an embarrassing number of false positives in the empirical literature. Kim and Ji (2015) criticize the practice of mindlessly using the “p-value less than 0.05” criterion in finance studies, without considering the key factors of hypothesis testing such as sample size.
and statistical power (see also Kim et al., 2018). They provide evidence that statistical significance in a large number of published studies may not hold if alternative criteria for statistical significance are used. In particular, they demonstrate that the use of the \( p \)-value criterion under a large or massive sample size is meaningless since any statistical tests can reject the null hypothesis of no effect in this case, even though the effect is economically negligible. In the context of stock market anomalies, Kim (2017, 2019) criticize the exclusive use of the \( p \)-value criterion for statistical significance, with examples including the effects of the weather, FIFA world cup, and political cycles. Connolly (1991) report similar findings for the January effect, while Pinegar (2002) and Kelly and Meschke (2010) express their skepticism on the effects of daylight saving and winter blues.

In this paper, we argue that the practice of data-mining, \( p \)-hacking or multiple testing arises due mainly to the presence of model uncertainty (Avramov 2002; and Cremers, 2002). In the absence of compelling economic theories, it is possible that many reported results on stock market anomalies are accidentally obtained. That is, whether consciously or not, researchers try a range of alternative models and report only those found to be statistically significant; or only those that support “preconceived notions” (Christensen and Miguel, 2018; p.931). The process is closely related with the practice of data-mining (Black, 1993), data-snooping (Lo and MacKinlay, 1990), multiple testing (Harvey et al., 2016), or \( p \)-hacking (Harvey, 2017), generally referring to (often unintentional) cherry-picking of statistically significant results from a large number of competing empirical models. The literature of stock market anomalies is particularly prone to these practices because researchers have a large number of potential and empirically-driven explanatory variables, without concrete economic theories that can guide their model selection.

Model uncertainty refers to the situation where the variables to be included in the regression are not fully known to the researcher. As a result, the estimated models are different from studies to studies, and so are the empirical results. For example, in establishing the effect of seasonal affective disorder on stock return, the regression model of Kamstra et al. (2003) includes the key variable measuring seasonal depression, along with a range of weather variables such as cloud cover. They report the statistically significant effect of seasonal depression, but find their weather variables mostly insignificant. In contrast, Hirshleifer and Shumway (2003) report a highly significant weather effect from a regression model which does not include the
variables for seasonal affective disorder. In this case, it is not clear which model is statistically adequate; and whether either of these variables is a robust predictor for stock return.

The main point of this paper is that many of these anomalies are identified without a full consideration of model uncertainty. To this end, we assess the empirical validity of the anomalies by employing the extreme bounds analysis (EBA) of Leamer (1985) and Sala-i-Martin (1997). As an effective tool to address model uncertainty, it is designed to assess the robustness or fragility of the relationship between the dependent variable (stock return) and key explanatory variables, within a large number of alternative models that include all combinations of explanatory variables. From the EBA, we obtain distributions for the indicators of anomalies such as the parameter estimates (effect size) and measures for statistical significance (e.g., $t$-statistics). By examining the salient features of these distributions, we can determine the degree of robustness of the claimed anomalies, isolating the effects of data-mining, $p$-hacking, or multiple testing.

The main finding of the paper is that the anomalies examined are found to be highly fragile under the EBA. They include the effects of seasonal depression, weather, FIFA World cup, lunar phase, Halloween, political cycle, and daylight saving. This is evidence that the anomalies reported in the past studies are highly likely to be the products of data-mining, multiple testing, or $p$-hacking under model uncertainty. In the next section, we present a brief literature review of the anomalies, Section 3 presents the methodologies; Section 4 presents the empirical results; and Section 5 concludes the paper.

2 Literature review

This section provides a brief review of the literature on stock market anomalies. We first review the studies of widely known calendar anomalies; and then the recent anomalies that are largely related to investor mood. A summary of findings from typical studies on market anomalies is given in the table of Appendix 1. Most of these studies report statistically significant evidence of anomalies. However, Sullivan et al. (2001) argue that many of these anomalies are the outcomes of data-mining. We discuss the role of model uncertainty in motivating data-mining and generating spurious evidence of anomalies.
2.1 Calendar anomalies

2.1.1 The day-of-the-week

The weekend effect was first reported by Cross (1973) who finds that the average return on Mondays is +0.16% and the average return on Fridays is -0.12%. Fama (1980) tests two competing hypotheses regarding the day-of-the-week effect. The calendar time hypothesis postulates that the expected return on Mondays should be three times the expected return for other days of the week because the return generating process operates continuously. On the other hand, according to the trading time hypothesis, expected return should be the same for each day of the week because returns are generated only on trading days. Fama (1980) finds that the daily returns on the S&P 500 portfolio are inconsistent with both hypotheses. Specifically, the average return on Monday is negative while the average return on the other four days of the week is positive. Keim and Stambaugh (1984) also provide similar evidence. Lakonishok and Levi (1982) explain the Monday effect in terms of the delay between execution and settlement. For stocks traded on a Friday, it takes two additional days for settlement than those traded on another day. Hence, an amount equivalent to two days’ interest is embedded in price of a stock traded on a Friday. Furthermore, firms have a tendency to release bad news on a Friday following markets close, which dampens stock prices on Monday. Another source of the Monday effect is investors’ inclination to increase the number of sell transactions relative to buy transactions on Mondays (Lakonishok and Maberly, 1990; Abraham and Ikenberry, 1994). However, Connolly (1989) finds that both the day-of-the-week and weekend effects are not found in the US by 1975 if the significance level is adjusted downward to control for sample size and the estimation method.

2.1.2 The turn of the month effect

The turn of the month effect was first reported by Ariel (1987) in the US market. It refers to positive average returns during the first half of calendar months and close to zero average returns during the second half. Lakonishok and Smidt (1988) find that in the US the cumulative returns over the last day of the month and the subsequent three days of the following month surpass returns over the whole month. Kunkel et al. (2003), McConnell and Xu (2008) and Giovanis (2009) provide similar evidence for the turn of the month effect in an international context. This has remained a puzzle as there
is no risk-based, behavioural or institutional explanation for the turn of the month effect.

### 2.1.3 The month-of-the-year effect

The January effect was first reported by Rozeff and Kinney (1976) for NYSE stocks. They found higher returns in January than other months of the year. The January effect is typically explained in terms of tax-loss selling in the late December (Branch, 1977) and the window dressing practice of institutional investors who sell losing and risky stocks in the late December and reinvest the proceeds in January (Lakonishok & Smidt, 1988). Using 300 years of data for the UK, Zhang and Jacobsen (2012) find no January effect but they find evidence in support of July and October effects, the half yearly Halloween (or Sell-in-May) effect, and the winter effect.

### 2.1.4 Holiday Effect

The holiday effect refers to the evidence that on average pre-holiday returns are higher than non-pre-holiday returns. Fields (1934), Lakonishok & Smidt (1988) and Ariel (1990) provided the evidence for the holiday effect in the US. The holiday effect was also reported in Australia (Easton, 1990), the United Kingdom and Japan (Kim & Park, 1994), and advanced emerging markets (Seif et al., 2017). The holiday effect is attributed to behavioural and institutional factors as opposed to risk factors. Fabozzi et al. (1994) argue that the holiday effect may be due to improved investor mood around holidays, a delay in settlement procedures as a result of the holiday, and a greater perceived loss potential of a short position prior to a non-trading day.

### 2.2 Recent anomalies

As we have stated above in Section 2.1, Sullivan et al. (2001) report that all of the above calendar effect are the outcomes of data-mining. This is consistent with the finding of recent study by Plastun et al. (2019) who report that all of the calendar effects have disappeared from 1980’s.

There have been a new wave of the studies which report a different class of anomalies, especially in relation to the investors’ mood. The seminal studies in this literature are Saunders (1993), Hirshleifer and Shumway (2003) and Kamstra et al. (2003). The first two studies report statistically significant
weather effects on stock return, with a claim that cloudy weather adversely affect the investors’ mood, which in turn shows negatively affect stock return. Kamstra et al. (2003) study the effect of depression linked with seasonal affective disorder (SAD) on stock return, with a statistically significant impact. They claim that, through the link between SAD and depression, and the link between depression and risk aversion, seasonal variation in length of day can translate into seasonal variation in equity return.

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), Lucey and Dowling (2005), 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), sports events (Edmans et al., 2007; Kaplan-ski and Levy, 2010a; Chang et al., 2012), lunar phases (Yuan et al., 2006; Keef and Khaled, 2011), pollution (Lepori, 2016), Ramadan (Bialkowski et al., 2012), and political cycle (Santa-Clara and Valkanov, 2003; Novy-Marx, 2014). Most of these studies report statistically significant effects of investors’ mood on stock market, and their findings are presented as direct evidence for the anomalies against market efficiency.

There are studies that raise skepticisms 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) argue that the documented weather effects might be the consequence of “data-driven inference based on spurious correlation”. They provide evidence that the seasonal anomaly in stock return is unlikely to be caused by investors’ mood changes due to weather variations, stating
that it is premature to conclude that weather has an effect on stock return through mood changes of investors. In re-evaluating the effect of seasonal depression, Kelly and Meschke (2010) document that the observed effect is mechanically driven by an overlapping dummy-variable specification. In criticizing the effect of daylight saving on stock return, Pinegar (2002) uses the posterior odds ratio which strongly support the null hypothesis of no effect. Kim (2017) argues that the statistically significant weather effect is a spurious correlation caused by statistical power equal to one due to massive sample size.

Hence, while the literature is weighted heavily on the overwhelming evidence for the anomalies as a result of investors’ mood and calendar effects, a number of authors have been also skeptical about their existence, especially in relation to their statistical significance. This provides a good case for conducting a large scale robustness analysis, for which the EBA is designed.

2.3 Model uncertainty

Empirical studies in this area popularly adopt a linear regression for stock return as a function of relevant continuous and dummy variables. The specifications ranges from a simple regression with a single explanatory (dummy) variable; or a multiple regression which contain a combination of a large number of continuous and dummy variables. Some studies include dynamic terms of stock return to capture potential autocorrelation (e.g., Saunders, 1993), while there are studies which adjust the stock return with GARCH(1,1) conditional standard deviations to adjust for heteroskedasticity (see Edmans et al., 2007).

As an example, Kamstra et al. (2003) study the effect of depression linked with seasonal affective disorder (SAD) on stock return. They consider the regression model of the following form:

\[ R_t = \gamma_0 + \sum_{i=1}^{2} \gamma_i R_{t-i} + \gamma_3 SAD_t + \gamma_4 A_t + \gamma_5 M_t + \gamma_6 T_t + \gamma_7 C_t + \gamma_8 P_t + \gamma_9 G_t + \epsilon_t, \]

where \( R_t \) denotes the stock return in percentage on day \( t \); \( M \) a dummy variable for Monday; \( T \) a dummy for the last trading day or the first five trading days of the tax year; \( A \) a dummy for autumn months; \( C \) cloud cover, \( P \) a precipitation; and \( G \) temperature. \( SAD \) is a measure of seasonal depression, which takes the value of \( H - 12 \) where \( H \) represents the time
from sunset to sunrise if the day $t$ is in the fall or winter; 0 otherwise. The models adopted by Edmans et al. (2007) and Kaplanski and Levy (2010a) for the effect of sports sentiment are special cases of this model. Hirshleifer and Shumway (2003), on the other hand, adopt a simple linear model, expressing stock return as a cloud cover only:

$$R_t = \gamma_0 + \gamma_1 C_t + \epsilon_t.$$  \hspace{1cm} (2)

A number of other authors have adopted simple regression to provide evidence of anomaly. To name a few, Bouman and Jacobsen (2002) for the Halloween effect; Yuan et al. (2006) for the effect of lunar phases; Santa-Clara and Valkanov (2003) for the effect of political cycle; and Novy-Marx (2014) for a range of exotic factors including sunspot numbers. There are studies which take a simple model like (2), but with additional dummy variables. Hence, we observe that the model specifications of the studies in this area of literature range from a general one like (1) to a simple one like (2).

We should note that the choice of models in the past studies are made rather arbitrarily, without any concrete theoretical basis. It appears that no studies has chosen their model through extensive model specification testing. This indicates a strong degree of model uncertainty associated with the studies in this area. As a result, it is possible that the models often show conflicting results. From daily stock returns from U.S. and other international markets and using model (1), Kamstra et al. (2003; p.326) find the strong evidence of SAD effect with high statistical significance, while the the weather variables ($C, P$ and $G$) are found to be statistically insignificant. On the other hand, using a simple model like (3), Hirshleifer and Shumway (2003) have found highly statistically significant effect of cloud cover ($C$) on stock return. These two studies provide conflicting results, which are likely to be the outcomes of data-mining under model uncertainty. In this context, the following quote from Black (1993; p.75) provides a good description of the process of data-mining under model uncertainty:

When a researcher tries many ways to do a study, including various combinations of explanatory factors, various periods, and various models, we often say, he is “data mining.” If he reports only the more successful runs, we have a hard time interpreting any statistical analysis he does. We worry that he selected, from the many models tried, only the ones that seem to support his conclusions. With enough data mining, all the results that seem

9
significant could be just accidental. .... Data mining is not limited to single research studies. In a single study, a researcher can reduce its effects by reporting all the runs he does, though he still may be tempted to emphasize the result she likes. Data mining is most severe when many people are studying related problems.

The above statement also strongly suggests that a sensitivity analysis such as the EBA is called for to evaluate the robustness or fragility of the empirical results obtained under model uncertainty.

2.4 EBA for empirical finance

The EBA has been scarcely applied in empirical finance. Durham (2000, 2001) evaluate stock return predictability from a range of financial indicators, for the emerging stock markets as well as for the developed ones. Durham (2000) reports that, for the case of emerging markets, no financial indicators are found to be robust; while, for developed markets, Durham (2001) reports that a handful of financial indicators are found to be robust. We note that Durham (2000, 2001) adopts a conventional level of significance for statistical significance, which is not adjusted for the problem of multiple testing (see Section 3.2 for detailed discussions). Moosa and Cardak (2006) have identified a number of robust factors for the foreign direct investment. Kim et al. (2019) apply the EBA to evaluate whether energy prices are robust predictors for stock return for the U.S. market. They find that energy prices are fragile predictors for the stock returns over a range of holding periods, adopting the $p$-value criterion as a decision rule with its threshold adjusted for multiple testing. In criticising the practice of econometrics in empirical economics and finance, Moosa (2017; Chapters 6 and 7) proposes the EBA as a remedy for the abuse or misuse of econometric methods.

3 Methodology

The EBA is a large-scale sensitivity analysis, strongly recommended for observational studies (Leamer, 1985). As Christensen and Miguel (2018) point out, any individual study can be biased due to (conscious or unconscious) selective reporting of statistical results. Through the EBA, we can evaluate the extent of bias of a single study, and assess its results free from the
ill-effects of potential cheery-picking. We first present the basic structure of
the EBA in a general setting and then explain how it can be adapted to the
analysis of stock market anomalies.

3.1 Extreme bounds analysis

As a large-scale sensitivity analysis, the EBA examines whether the empir-
ical relationship between the dependent variable and the key explanatory
variable(s) of interest is robust or fragile, when the relationship is controlled
for various combinations of other possible explanatory variables. Through
the EBA, a researcher can identify a set of explanatory variables associated
with the dependent variable with a high degree of robustness. As discussed
in Section 2.3, the regression models used in the literature are exposed to
model uncertainty. In this case, researchers may judge predictive power
based largely on statistical significance, which may occur by chance when
they examine multiple models, a range of hypotheses, and different sample
periods. In addition, the researcher may selectively report her preferred re-
sult without providing the full details of the extensive analysis conducted.
The problems have identified in different names, such as data-mining (Black,
1993), data-snooping (Lo and MacKinlay, 1990), multiple testing (Harvey
et al., 2016), or p-hacking (Harvey, 2017). By employing the EBA, the re-
searchers can find the relationship when such model uncertainty is fully taken
into account.

The basic structure of the EBA can be described as follows: let $y$ be the
dependent variable and $X$ be the set of all possible explanatory variables.
The purpose of the EBA is to examine whether a particular variable, called
the focus variable, $Z \in X$ is robustly related to $y$. A subset $F$ (not including
$Z$) of $X$ is called the free variables, which are included in every regression
model along with $Z$. The rest of the variables in $X$ are called the doubtful
variables denoted $D$’s. In each regression model of the EBA, a set of $k$ vari-
ables are drawn from the entire set of doubtful variables. The $j$th regression
model in the EBA is written as

$$y = \alpha_j + \beta_j Z + \gamma_j F + \delta_j D_j + u_j,$$

where $\alpha_j$ is the intercept; $\beta_j$, $\gamma_j$, and $\delta_j$ are the (vector of) coefficients at-
tached to $Z$, $F$ and $D_j$ and $u_j$ is the error term. In each regression, different $k$
combinations from the entire set of doubtful variables are included (denoted
The value of $k$ is usually set at three, but other values may be chosen. Let $M$ be the total number of regressions: note that $M = \binom{N}{k}$ represents the number of $k$ combinations out of $N$, where $N$ is the number of all doubtful variables. With the EBA, we pay attention to the distribution $\{\hat{\beta}_j\}_{j=1}^M$, where $\hat{\beta}_j$ denotes the parameter estimate for $\beta_j$. For example, if $\{\hat{\beta}_j\}_{j=1}^M$ has a high proportion of number of positive and economically large values, it is an indication that the effect of the focus variable on $y$ is robust and meaningfully positive. We present a histogram of $\{\hat{\beta}_j\}_{j=1}^M$ for its visual summary. Researchers may pay attention to their statistical significance: for example, $\{t(\hat{\beta}_j)\}_{j=1}^M$, where $t(\hat{\beta}_j)$ denotes the $t$-statistic for $\hat{\beta}_j$, to assess statistical significance of $Z$. One may ask what proportion of $\{\hat{\beta}_j\}_{j=1}^M$ is statistically significant at $\alpha$ level of significance. The next subsection provides further discussion on the issue of evaluating statistical significance of the estimated coefficients in the EBA, especially for the choice of the value of $\alpha$.

Leamer (1985) proposes the use of extreme bounds of $\{\hat{\beta}_j\}_{j=1}^M$ to assess the robustness or fragility of the effect of a focus variable. For each member of $\{\hat{\beta}_j\}_{j=1}^M$, calculate the 100$(1-\alpha)$% confidence intervals. From this set of $M$ confidence intervals, we obtain lower and upper extreme bounds. Following Leamer (1985), a focus variable is judged to be fragile if the lower and upper extreme bounds cover the value of zero. The choice of value $\alpha$ is discussed in the next subsection. We note that, however, McAleer et al. (1985) criticize that this definition of fragility is too stringent. In response to this, Granger and Uhlig (1990) propose the use of these extreme bounds from a subset of $M$ regressions with superior specifications judged by the values of $R^2$. That is, the researchers consider only the regressions associated with the top 5% or 10% of the $R^2$ values for the calculation of these extreme bounds.

As Sala-i-Martin (1997) argues, it is important to consider the entire distribution of the estimated coefficients from the EBA. We present the histograms for an effective visual impression about the magnitude and dispersion of the regression coefficients. As a summary measure of distribution, Sala-i-Martin (1997) proposes the use of the cumulative distribution function (CDF) of $\{\hat{\beta}_j\}_{j=1}^M$. For examples, $\text{CDF}(0)$ is calculated as the proportion of all members of $\{\hat{\beta}_j\}_{j=1}^M$ less than 0. A low (high) value of $\text{CDF}(0)$ indicates that most of the values of regression coefficients are positive (negative). To calculate the $\text{CDF}(0)$ for $\hat{\beta}_j$, it is assumed that it follows a normal distribution with mean $\hat{\beta}_j$ and standard deviation $\hat{\sigma}_j$, where $\hat{\sigma}_j$ is the standard error of $\hat{\beta}_j$. Then, it is calculated as $\phi_j(0|\hat{\beta}_j, \hat{\sigma}_j^2)$, where $\phi_j$ is the cumulative
distribution of a normal random variable. Then, the aggregate CDF(0) for \( \hat{\beta} \) is calculated as the weighted average of all the individual CDF(0)’s. That is,

\[
\Phi(0) = \sum_{j=1}^{M} w_j \phi_j(0|\hat{\beta}_j, \hat{\sigma}^2_j),
\]

where \( w_j \) denotes the weight given to \( \hat{\beta}_j \). The weights are determined so that the regression with a better fit can have a higher weight. That is,

\[
w_j = \frac{R^2_j}{\sum_{j=1}^{M} R^2_j},
\]

where \( R^2_j \) represents the adjusted coefficient of determination from the \( j \)th regression. Note that, when \( R^2_j \) is negative, it is set to 0 making \( w_j = 0 \). We also pay attention to the weighted coefficient estimate of the focus variable and its standard error estimate. That is,

\[
\hat{\beta}_w = \sum_{j=1}^{M} w_j \hat{\beta}_j,
\]

and its standard error estimate denoted as \( se(\hat{\beta}_w) \).

### 3.2 Statistical significance

In the literature of financial anomalies, statistical significance is evaluated almost exclusively using the “\( p \)-value less than 0.05” criterion, as with other areas of empirical finance. There have been a number of criticisms and concerns expressed recently to this practice: see, for example, Kim and Ji (2015), Wasserstein and Lazar (2016), and Harvey (2017). It is argued that the \( p \)-value is a misleading measure of evidence, mainly because it completely ignores the probability of Type II error or statistical power (see Berger and Sellke, 1987). In addition, its popular threshold such as 0.05 is arbitrary and often too lenient, especially under a large sample (Kim and Ji, 2015) or when multiple testing is conducted under model uncertainty (Harvey et al., 2016).

In evaluating statistical significance of the estimated coefficients in the context of the EBA, there are two main issues to consider. First, it should be noted that, by construction, significance testing in the EBA subject to the problem of multiple testing. Statistical significance at a conventional
level of significance is misleading in this case, as Harvey et al. (2016) point out. To guard against the multiple testing problem, Harvey et al. (2016) propose the use of the critical value of 3 for the $t$-test, which is largely consistent with 0.001 level of significance. This level of significance has been also advocated by the scholars of other disciplines of science, who claim that the conventional significance levels are too lenient under a range of situations including multiple testing (Johnson, 2013; Benjamin et al., 2017). Hence, throughout the EBA in this paper, we use the 0.001 level of significance as a threshold for the $p$-value criterion; or equivalently adopt 99.9% confidence intervals.

Second, the studies which have provided evidence of anomalies usually adopt a large sample size. In this case, as Kim (2017) points out in the context of the weather effect on stock return, statistical significance based on the $p$-value less than 0.05 criterion is meaningless. This is mainly because the power of the $t$-test is effectively one (which makes the probability of Type II error zero), and the level of significance (or probability of Type I error) set at 0.05 in infinitely larger than the probability of Type II error. The consequence is that any negligible deviation from 0 (albeit economically meaningless) is found to be statistically significant. Kim and Ji (2015) and Kim (2017) propose the use of an alternative criterion such as the Bayes factor, which suggests a statistical decision by balancing the probabilities of Type I and II errors. In this paper, we adopt the Bayes factor proposed by Zellner and Siow (1980) (see also Liang et al., 2008). It is a simple function of $F$-test statistic (or the square of $t$-statistic) for a regression model, based on a Cauchy distribution for the prior. It is written as

$$BF = \frac{\Gamma[0.5(k_0 + 1)][1 + (k_0/v_1)F]^{0.5(v_1-1)}}{\pi^{0.5}(0.5v_1)^{0.5k_0}},$$

where $\Gamma()$ is the Gamma function and $v_1 = T - k_0 - k_1 - 1$, while $k_0$ is the number of parameters restricted under $H_0$ and $k_1$ is the number of unrestricted parameters (i.e., $K = k_0 + k_1$). According to Kass and Raftery (1995, p. 777), the evidence against $H_0$ of no effect is “not worth more than a bare mention” if $2\log(BF) < 2$; “positive” if $2 < 2\log(BF) < 6$; “strong” if $6 < 2\log(BF) < 10$; and “very strong” if $2\log(BF) > 10$, where $\log$ is the natural logarithm. Note that Harvey (2017) also proposes the use of a Bayesian alternative for more sensible hypothesis testing in empirical finance.
3.3 Evaluating anomalies

Our basic regression equation takes the following form:

$$R_t = \gamma_0 + \sum_{i=1}^{2} \gamma_i R_{t-i} + \beta F_t + \sum_{i=1}^{k} \delta_i D_{it} + \epsilon_t,$$

where $R_t$ is stock return at time $t$, $F_t$ is the focus variable related with the anomaly under investigation, and $D_{it}$ is a set of $k$ doubtful variables. Note that we use $R_t$ with two lag orders as a set of free variables for the EBA to ensure that the residuals of (6) show no sign of auto-correlation, following Kamstra et al. (2003). Since we use daily return data, the error term is likely to suffer from a high degree of conditional heteroskedasticity. Following Edmans et al. (2007), we fit a GARCH(1,1) model to the OLS residuals from equation (6) and calculate the normalized return by calculating $R_t/\hat{\sigma}_t$, where $\hat{\sigma}_t$ is the estimate of conditional standard deviation obtained from the GARCH(1,1) model estimation. We use the normalized return in estimating equation (6).

The common doubtful variables are derived from the model specifications from the past studies. The common doubtful variables are tabulated in Appendix A. We include the dummy variables for the January effect ($Jan$), Monday effect ($Mon$), autumn ($A$); and a dummy for the last trading days or the first five trading days in the tax year ($T$). In selecting these variables, we are guided by Kamstra et al. (2003), Kaplanis and Levy (2010a) and Edmans et al. (2007). We also include the weather variables such as cloud cover ($C$), temperature ($G$), and precipitation ($P$), guided by the studies such as Saunders (1993), Hirshleifer and Shumway (2003) and Cao and Wei (2005). Note that the cloud cover $C$ is available only to 1996, so it is included only in selected EBA evaluations. We also include dummy variables for the 10 days with the highest and lowest returns ($J_1$ and $J_2$), following Kaplanski and Levy (2010a). Other doubtful variables include seasonal affective disorder ($SAD$; Kamstra et al., 2003), dummy variable for the FIFA World cup event (Kaplanis and Levy, 2010a), dummy variable for the Halloween effect ($S$; Bouman and Jacobsen, 2002), lunar phase ($L$; Yuan et al., 2006), dummy variable for the political cycle or U.S. presidency ($PRE$; Santa-Clara and Valkanoc, 2003), and dummy variables for daylight saving changes ($DL_1$ for changes in Spring and $DL_2$ for the changes in Fall; Kamstra et al., 2000).

Note that, as defined in Section 3.2, the focus variable ($F$) is a variable
of interest among the set of all available variables \((X)\). For example, if we are interested in the robustness of the Halloween effect in the EBA, the Halloween dummy variable is focus variable and all the remaining variables are doubtful variables. If we pay attention to the effect of lunar phase, the lunar phase variable is the focus variable and all others (including Halloween dummy variable) are doubtful variables.

All computations, including generation of dummy variables, are conducted using programming language R (R Core Team, 2017) and its packages “ExtremeBounds” (Hlavac, 2016), “lunar” (Lazaridis, 2014), “timeDate” (Wuertz et al., 2017a), and “fGarch” (Wuertz et al., 2017b). The R codes are available from the authors on request.

4 Empirical results

For stock return, we use daily returns (value-weighted) from the NYSE composite index (CRSP) from 1950 to 2016. We pay attention to the anomalies reported that are not covered in the study by Sullivan et al. (2001) who examined the case of calendar anomalies. We analyze the winter blues (SAD) effect (Kamstra et al, 2003), weather effect (Saunders, 1993; Hirshleifer and Shumway, 2003), FIFA World Cup effect (Kaplanski and Levy, 2010a), Halloween effect (Bouman and Jacobsen, 2002), moonstruck (lunar phase) effect (Yuan et al, 2006), political cycle (Santa-Clara and Valkanov, 2003; Novy-Marx, 2015), and daylight saving effect (Kamstra et al. 2000).

We first replicate the results of the study that reports an anomaly, and then conduct the EBA. We present the EBA results with \(k = 5\), where up to 5 combinations of doubtful variables are included in each regression. We have tried other \(k\) values such as 3 and 7, but found the results qualitatively similar. The results with \(k = 7\) are available from the authors on request.

4.1 SAD Stock market cycle

Kamstra et al. (2003) study the effect of depression linked with the seasonal affective disorder (SAD) on stock return. They claim that, through the link between SAD and depression, and the link between depression and risk aversion, seasonal variation in length of day can translate into seasonal variation in equity return. They consider the regression model given in (1). Note that \(SAD_t\) is a measure of seasonal depression, which takes the value of \(H_t - 12\).
where $H_t$ represents the time from sunset to sunrise if the day $t$ is in the autumn or winter; 0 otherwise.

Kamstra et al. (2003; p.326) argue that lower returns should commence with autumn because depressed investors shunning risk and re-balance their portfolio in favor of safer assets. This is followed by abnormally higher returns when days begin to lengthen and SAD-affected investors begin resuming their risky holdings. They use the daily index return data from the markets around the world: U.S. (S&P 500, NYSE, NASDAQ, AMEX), Sweden, U.K., Germany Canada, New Zealand, Japan, Australia, and South Africa. They report, nearly for all markets, that the estimate of $\gamma_3$ in equation (1) is positive and statistically significant at a conventional level of significance; and that of $\gamma_4$ is negative and statistically significant.

We replicate their result using the US data daily from Jan 1965 to April 1996 (7886 observations). Our data range is limited by the cloud cover data which is available only from 1965 to 1996\(^1\). After adjusting the return with GARCH(1,1) conditional standard deviation, we have the following estimated values for the key coefficients: $\hat{\gamma}_3 = 0.033$ with $t$-statistic of 2.31; $\hat{\gamma}_4 = -0.053$ with $t$-statistic of $-2.10$; and $R^2 = 0.05$. These values are fairly close to those reported in Table 4A of Kamstra et al. (2003).

We have conducted the EBA with $SAD$ and $F$ as the focus variables and $R_{t-1}$ and $R_{t-2}$ as the free variables. The doubtful variables include all of the independent variables in equation (1), plus the lunar phase variable, daylight saving dummy variables, holiday dummy, Presidential dummy, FIFA world cup dummy, January dummy, dummy variables for the lowest and highest returns, and Halloween dummy variables. With this set up, the total number of regression models estimated in the EBA is 11829.

Figure 1 presents the histograms for the coefficient estimates of the focus variables. For the SAD variable, most of the estimates are positive with $\Phi(0) = 0.065$, but a large proportion of them are far away from the sample estimates of 0.033, which represents the 98.6th percentile of the distribution. This may suggest the the effect size may have been inflated due to data-mining. The weighted coefficient ($\hat{\beta}_w$) is 0.017 with standard error of 0.011, indicating much smaller value of FALL effect. At the significance level of 0.001, none of the estimated coefficients are statistically significant. Leamer’s (1985) lower and upper extreme bounds are $(-0.037, 0.087)$, meaning that the effect is statistically fragile. Those associated with the regressions with the

\(^1\)The data is downloaded from https://www.ncdc.noaa.gov/ (accessed March, 2019).
top 10% of adjusted $R^2$ values (Granger and Uhlig, 1990) are $(-0.032, 0.082)$, again indicating that the effect is statistically fragile.

For the FALL effect, it is noticeable that the distribution is bi-modal. This means that the coefficient estimates may be contingent on a set of doubtful variables which trigger this bi-modality. Most of the estimated coefficient are negative with $\Phi(0) = 0.9074$, but they are far away from the estimated value of $-0.053$, which is the 9.3th percentile of the estimated coefficients. The weighted coefficient ($\hat{\beta}_w$) is $-0.029$ with standard error of $0.023$, again indicating a much smaller value of FALL effect. This again suggests that the effect size may have been inflated due to data-mining. None of the estimated coefficients are statistically significant at the 0.001 level of significance. Leamer’s (1985) lower and upper extreme bounds are $(-0.156, 0.063)$, suggesting that the effect is statistically fragile, while those associated with the regressions with the top 10% of adjusted $R^2$ values are $(-0.145, 0.059)$.

When the Bayes factor is used as a measure of statistical evidence, the $2\log(BF)$ values from both variables are less than 2 for all cases, indicative of the effect that is “not worth more than a bare mention”. The adjusted $R^2$ values range from 0.045 to 0.063, indicating that all fitted models have a low predictive power for stock return. All of the results from the EBA
clearly indicate that both SAD and FALL effects are fragile under model uncertainty.

4.2 Weather effect on stock return

Saunders (1993) consider the following regression to evaluate the weather effect on stock return:

\[ R_t = \gamma_0 + \gamma_1 R_{t-1} + \beta C_t + \sum_{i=2}^{12} \tau_i M_{it} + \sum_{i=2}^{5} \gamma_i D_{it} + u_t, \]  

(7)

where \( M_{it} \) denotes the monthly dummy variable; \( D_{it} \) day-of-the-week dummy variable; and \( C_t \) is the cloud cover variable which takes 1 if cloud cover is between 0 to 20%, 0 if it is between 30% and 90%; -1 if it higher than 90%.

This study is followed by Hirshleifer and Shumway (2003) who estimate the model given in (2).

We replicate the regression (7) using the data in the previous subsection for the SAD effect. Using the returns adjusted with the GARCH(1,1) standard deviations, we find that \( \hat{\beta} = 0.0427 \) with \( t \)-statistic of 3.36 and \( R^2 \) of 0.06. These values are virtually identical to the values reported in Table 2 of Saunders (2003): for example, using NYSE/AMEX value weighted return, Saunders (1993) reports that \( \hat{\beta} = 0.052 \) with \( t \)-statistic of 3.27 and \( R^2 \) of 0.05. A simpler version of the above regression, only including the lagged return, January dummy and Monday dummy variables with cloud cover \( C \), provide the similar results: namely, \( \hat{\beta} = 0.0406 \) with \( t \)-statistic of 3.22 and \( R^2 \) of 0.05.

We conduct the EBA with \( C_t \) as the focus variable and \( R_{t-1} \) and \( R_{t-2} \) as the free variables, with doubtful variables including January dummy, Monday dummy, SAD variable, Fall dummy, precipitation, temperature, lunar phase, Halloween dummy, holiday dummy, Presidential dummy, FIFA world cup dummy, Tax day dummy variables, and daylight saving dummy variables, and the dummy variables for the lowest and highest returns. A total of 6885 regressions are conducted in the EBA.

The EBA results are reported in Figure 2 where the EBA distribution shows a clear bi-modality. While all estimated values are positive, most of the values are less than the estimated coefficient of 0.0427 from the replication results. This value represents the 85.8th percentile of the values from the EBA. The weighted coefficient \( (\hat{\beta}_w) \) is 0.039 with standard error of 0.013,
Figure 2: Histograms from EBA: Weather effect

![Histograms from EBA: Weather effect](image)

Note: The histogram presents the frequency distributions of the estimated coefficients from the EBA.

implying a much smaller value of the weather effect. These feature represent a sign that the effect size estimate may have been inflated due to data-mining. Around 38% of the estimated coefficients are statistically significant at the 0.001 level of significance. However, the values of $2 \log(BF)$ range from -4.18 to 3.11, indicating that most of the estimated coefficient show the effects which are “not worth of bare mention”. However, all of their $2 \log(BF)$ values are less than 6, which is a threshold for a strong effect. Leamer’s (1985) extreme bounds ($-0.013, 0.086$) cover the value of zero, indicating that the effect of cloud cover is fragile. Those associated with the regressions with the top 10% of the adjusted $R^2$ values ($-0.009, 0.084$), providing similar results.

Figure B.1 plots the Box-and-Whisker plots of stock return against the cloud cover. It appears that the median and other quartiles are virtually identical across different values of $C$, which further supports the EBA results that the cloud cover as a predictor for stock return is fragile. From the Box-whisker plot, we can observe the presence of outliers, especially when $C = -1$ and 0, which may suggest that outliers may have played a role in achieving statistical significance. The value of adjusted $R^2$ ranges from 0.045 to 0.062, meaning that the models can explain only around 5% of the total variation of the stock return. Again, the overall evidence strongly suggests that the
relationship between stock return and cloud cover is fragile.

Note that Hirshleifer and Shumway (2003) provides further evidence in favor of the weather effect, but Kim (2017) provides the discussion as to the problem in their research design and fragility of their results.

4.3 FIFA World cup effect

To test for the FIFA World Cup effect on stock return, Kaplanski and Levy (2010a) estimate the following equation:

\[ R_t = \gamma_0 + \sum_{i=1}^{2} \gamma_{1i} R_{t-i} + \sum_{i=1}^{4} \gamma_{2i} D_{it} + \gamma_3 H_t + \gamma_4 T_t + \gamma_5 B_t + \gamma_6 E_t + \sum_{i=1}^{2} \gamma_{7i} J_{it} + \epsilon_t, \]

where \( R_t \) is the daily stock market return, \( D_{it} \) are the dummy variables for the day of the week (Monday to Thursday), \( H_t \) is a dummy variable for days after a non-weekend holiday, \( T_t \) is a dummy variable for the first 5 days of the taxation year, \( B_t \) is a dummy variable for the annual event period (June and July), \( E_t \) is the dummy variable for the event days, and \( J_{it} \) are the dummy variables for the 10 days with the highest \((i = 1)\) and lowest \((i = 2)\) returns during the studied period. The above model follows those of Edmans et al. (2007) and Kamstra et al. (2003). For the dummy variable \( E_t \), Kaplanski and Levy (2010a) use EED (event effect days) and EPED (event period effect days): the former being the game days that are trading days and subsequent trading days, while the latter being all days of the FIFA World Cup (from the first game to the first day after the final game) plus two subsequent trading days. We follow their basic model with serial correlation reported in Panel E (2e) of their Table 2.

We replicate their results for the above equation using daily CRSP value-weighted return from 1950 to 2016 (16819 observations). We find \( \hat{\gamma}_6 = -0.130 \) with \( t \)-statistic of \(-2.50\) and \( R^2 = 0.040 \), which are close to those reported in Kaplanski and Levy (2010a; Table 2). Following Kamstra et al. (2003), the returns are scaled with GARCH(1,1) conditional standard deviations.

For the EBA, we set \( R_{t-1} \) and \( R_{t-2} \) as the free variables and \( E_t \) as the focus variable. We report the results with the entire sample from 1950 to 2016, but the result using the data to 2007 as in Kaplanski and Levy (2010a) provide the similar results. The doubtful variables include \( M, Jan, H, T, B, P, G, J_i, SAD, F, \) plus, day light saving dummy variables, lunar phase...
variable, Presidential dummy, and Halloween dummy variable. Note that the cloud cover $C$ cannot be included because it is available only to 1996. The total number of regressions in the EBA is 6885.

Figure 3 presents the histogram of the FIFA world cup event dummy coefficients from the EBA. The EBA distribution shows bi-modality, which indicates that the coefficients may be sensitive to a set of doubtful variables. It is found that the most of the value are negative with $\Phi(0) = 0.9927$. A large proportion (54%) of the estimated coefficients are greater than the estimated coefficient of $-0.130$, which is the value obtained from the model adopted by Kaplanski and Levi (2010a). The weighted coefficient ($\hat{\beta}_w$) is $-0.124$ with standard error of 0.051, indicating a much smaller value of the weather effect. These features again suggest that the process of data-mining has inflated the effect size estimate. Note that none of them are significant at the 0.001 level of significance. All of the estimated coefficients have $2\log(BF)$ values less than 2, indicating the evidence of “not worth more than a bare mention”. Leamer’s (1985) extreme bounds ($-0.305, 0.059$) for the event dummy variable cover the value of zero, while those associated with the regressions with the top 10% of adjusted $R^2$ are ($-0.292, 0.045$), indicating that the effect is fragile. The value of adjusted $R^2$ ranges from 0.018 to 0.039, which indicate little predictive power of the models in the EBA. All of the
statistical results from the EBA point to the observation the effect of FIFA World Cup on Stock market is fragile.

### 4.4 Halloween Effect

Bouman and Jacobsen (2002) provide evidence that stock returns are higher from the period of November to April, which is called the Halloween effect. They fit a simple regression model of the form

\[ R_t = \gamma_0 + \gamma_1 H a L_t + \epsilon_t, \]

where \( H a L_t \) is a dummy variable which takes 1 from November to April; and 0 otherwise. Using monthly data for a large number of stock markets around the world, they report positive and statistically significant (at a conventional level) values of \( \hat{\gamma}_1 \). For the U.S. market, Bouman and Jacobsen (2002; Table 1) report that \( \hat{\gamma}_1 = 0.96 \) with \( t \)-statistic of 1.95, using 344 monthly observations. Using the GARCH(1,1)-adjusted daily return (in percentage) from 1950 to 2016 (16819 observations), we have found that

\[ R_t = 0.004 + 0.135 R_{t-1} - 0.023 R_{t-2} + 0.051 H a L_t, \]
with all estimated slope coefficients statistically significant (t-statistic for 0.051 is 3.601). The estimated coefficient $\gamma_1$ is close to the value from monthly data of Bouman and Jacobsen (2002) when it is converted to the monthly value by multiplying 22.

The EBA is conducted using the Halloween dummy variable $HaL$ as the focus variable, and $R_{t-1}$ and $R_{t-2}$ as the free variables. The variables included for the doubtful variables are Monday dummy, January dummy, $H_t$, $T_t$, $A_t$, $P_t$, $G_t$, $J_t$, $SAD$, $F$, plus, day light saving dummy, lunar phase variable, and Presidential dummy variable, involving 6885 regression models. The histogram for the coefficients of $S$ is given in Figure 4. All values are positive with $\Phi(0) = 0.006$; and about 41% of the coefficients are statistically significant at the 0.001 level of significance. The weighted coefficient ($\hat{\beta}_w$) is 0.053 with standard error of 0.019. The values of $2 \log(BF)$ ranges from -9.04 to 3.72, indicating that, while a proportion of the results may represent a “positive” effect, but none of them appear to be “strong”. The value of adjusted $R^2$ ranges from 0.019 to 0.036, indicating little predictive of the models estimated. Leamer’s (1985) extreme bounds ($-0.043, 0.151$) cover the value of zero, indicating that the effect is fragile. The results are similar with those associated with the top 10% of the adjusted $R^2$ values, which are ($-0.034, 0.144$) Again, the results for the presence of the Halloween effect does not appear to be robust overall.

Figure B.2 presents the Box-and-Whisker plot for the Halloween dummy and GARCH(1,1)-adjusted stock return. It appears that the median and other quartiles are nearly identical, but there are outliers when the Halloween dummy takes the value of 0. This may have contributed to a relatively high proportion of statistical significance, since the effect size and the $t$-statistic may have been inflated by these outliers.

4.5 Lunar phase

Yuan et al. (2006) provide the empirical evidence that the stock returns are significantly affected by lunar phase with the regression of the form

$$R_t = \gamma_0 + \gamma_1 L_t + \epsilon_t,$$

where $L_t$ is a dummy variable which takes 1 during the 15-day full moon period; and 0 otherwise. From the pooled regression of the 48 global market returns daily from 1970 to 2001, they found that the estimated coefficients of
\( \gamma_1 \) are negative and statistically significant. We suspect that this statistical significance may be the outcome of conducting significance testing with a massive (pooled) sample size: see Kim (2017).

We replicate their results using the U.S. stock return (GARCH(1,1)-adjusted) from January 1973 to July 2001 (7222 observations), with the following results:

\[
R_t = 0.04 + 0.163R_{t-1} - 0.015R_{t-2} - 0.0015L_t.
\]

The coefficient \(-0.0015\) is statistical insignificant of with the \(p\)-value of 0.49. Although statistically insignificant (due possibly to much smaller sample size than that of Yuan et al., 2006), the estimated coefficient is negative in concert with the values reported in Yuan et al. (2006). Using the whole data set from 1950 to 2016 (16819 observations), we have the lunar coefficient of \(-0.0001\) with the \(p\)-value of 0.94. As it is possible that this statistical insignificance may be the outcome of model uncertainty, we conduct the EBA for the lunar phase effect.

The EBA is conducted using the whole sample with the lunar phase dummy variable \(S\) as the focus variable, and \(R_{t-1}\) and \(R_{t-2}\) as the free variables. The variables included for the doubtful variables are Monday dummy, January dummy, \(H_t\), \(T_t\), \(A_t\), \(P_t\), \(G_t\), \(J_{it}\), \(SAD\), \(F\), plus, day light saving
dummy, Halloween dummy, and Presidential dummy variable, involving 6885 regression models.

The histogram for the coefficients of $S$ is given in Figure 5. The distribution shows a clear bi-modality, which means that the results are sensitive to a set of doubtful variables. The weighted coefficient ($\hat{\beta}_w$) is $-0.001$ with standard error of 0.014, which represents little effect of lunar phase. Although all of the estimated coefficients from the EBA are negative, none of them are statistically significant at the 0.001 level. Statistical significance based on the Bayes factor also provide no evidence of any effect with the values of $2\log(BF)$ ranging from $-10.18$ to $-10.16$, well below 2 which is the threshold for the effect of “not worth more than a bare mention”. Leamer’s (1985) extreme bounds ($-0.049, 0.046$) cover the value of zero, indicating that the effect is fragile. Similar results are found based on extreme bounds ($-0.046, 0.043$), associated with the top 10% of best-fit regressions. The value of adjusted $R^2$ ranges from 0.018 to 0.035, indicating negligible predictive power of the model.

### 4.6 Political cycle

Santa-Clara and Valkanov (2003) and Novy-Marx (2014) consider the regression of the form

$$Y_t = \gamma_0 + \gamma_1 PRE_{t-1} + \epsilon_t,$$

where $Y$ is the excess return (in percentage) and $PRE$ is the dummy variable for the party of the U.S. president (1 for the Democratic president, 0 for the Republican). Using the monthly U.S. data from 1961 to 2012 ($T=642$), Novy-Marx (2014) reports that $\hat{\gamma}_1 = 0.75$ with the $t$-statistic of 2.08, concluding that the party of the presidency has an economically significant effect on stock return. Based on a more detailed and comprehensive analysis, Santa-Clara and Valkanov (2003) report the similar results, calling the evidence as “Presidential puzzle”.

Using the daily data from 1950 to 2016 (16819 observations), we have obtained the following results:

$$R_t = 0.021 + 0.135R_{t-1} - 0.022R_{t-2} + 0.018PRE_{t-1}.$$

While the coefficient is positive, it is statistically insignificant at a conventional level with $t$-statistic is 1.253 with $p$-value of 0.210 and $R^2 = 0.018$. 

26
The EBA is conducted using the Presidential dummy variable \( PRE \) as the focus variable, and \( R_{t-1} \) and \( R_{t-2} \) as the free variables. The variables included for the doubtful variables are Monday dummy, January dummy, \( Ht, \ Tt, \ At, \ Pt, \ Gt, \ Jit, \ SAD, \ F \), plus, day light saving dummy, Halloween dummy, and Presidential dummy variable, involving 6685 regression models. The histogram for the coefficients of \( PRE \) is given in Figure 5. Although nearly all values are positive with \( \Phi(0) = 0.1127 \), none of them are statistically significant at the 0.001 level of significance. Statistical significance based on the Bayes factor also provide no evidence of any effect, with all values of \( 2 \log(\text{BF}) \) less than 2. The estimated coefficient of 0.018 value from the basic model is at the far end tail of the distribution, representing the 86.7th percentile. The weighted coefficient \( (\hat{\beta}_w) \) is 0.017 with standard error of 0.014. The value of adjusted \( R^2 \) ranges from 0.018 to 0.036, indicating little explanatory power of the models. Leamer’s (1985) extreme bounds are \((-0.031, 0.065)\), which cover the value of zero; so do those associated with the top 10% of the best-fit models which are \((-0.028, 0.062)\). Again, all of the results from the EBA indicate that the effect is fragile.
4.7 Daylight Saving

Kamstra et al. (2000) report that stock returns are lower on the weekend after daylight saving changes. Using the daily stock returns from the U.S., U.K. and German markets, they found that the mean returns of the daylight saving weekends (both Spring and Fall) are lower than the mean return of other weekends. The differences are found to be statistically different from 0 at a conventional significance level for nearly all market induces they considered. In this paper, we replicate their results based on our basic regression. Using the U.S. stock return for the same data period of Kamstra et al. (2000) (from January 1967 to December 1997), we have

\[ R_t = 0.061 + 0.211R_{t-1} - 0.025R_{t-2} - 0.162MON_t - 0.314DL_{1t} - 0.401DL_{2t}, \]

where \( R_t \) is the value-weighted stock return adjusted with GARCH(1,1) conditional standard deviation, \( MON \) is the dummy variable for the weekend not including the daylight saving weekends, \( DL_1 \) is the dummy variable for the Spring daylight weekend, and \( DL_2 \) is the dummy variable for the Fall daylight saving weekend. All coefficients are statistically significant at a conventional level of significance with \( R^2 = 0.049 \). As Kamstra et al. (2000) reports, the estimated coefficient of \( MON \) is larger than those of \( DL_1 \) and \( DL_2 \), and the differences are statistically significant. Similar results are obtained using the whole sample, where the estimated coefficient of \( MON \) is found to be \(-0.198\); that of \( DL_1 \) is found to be \(-0.245 \) (\( t \)-statistic = \(-2.197\)); and that of \( DL_2 \) is \(-0.180 \) (\( t \)-statistic = \(-1.594\)).

We conduct the EBA with \( DL_1 \) and \( DL_2 \) as the focus variables, and \( R_{t-1} \) and \( R_{t-2} \) as the free variables. We use the whole sample from 1950 to 2016 (16819 observations) for the EBA, while the similar results are obtained if we use the same data range as in Kamstra et al. (2000). The usual list of doubtful variables apply, except that the dummy variable \( MON \) is used instead of usual Monday dummy \( Mon \). This involves 11829 regression models in the EBA.

Figure 7 presents the histograms of the estimates of the two focus variables. The estimated coefficients are consistently negative with \( \Phi(0) = 0.9755 \) for \( DL_1 \) and \( \Phi(0) = 0.8800 \) for \( DL_2 \). However, none of them are found to be statistically significant at the 0.001 level of significance, and the values of \( 2\log(BF) \) also indicate no evidence of daylight saving effects as they are less than 2 for all cases for both \( DL_1 \) and \( DL_2 \). The estimated coefficient of \( DL_2 \) and \( DL_2 \) (\(-0.245 \) and \(-0.180\)) represent 23.1th and 7.52nd
percentiles of the above histograms respectively, again implying the possibility that the effect size estimates are inflated as a result of data-mining. The weighted coefficients ($\hat{\beta}_w$) are $-0.229$ and $-0.135$ respectively for $DL_1$ and $DL_2$, with standard errors 0.112 and 0.113. Leamer's (1985) extreme bounds are ($-0.640, 0.181$) and ($-0.570, 0.297$) respectively for the coefficients of $DL_1$ and $DL_2$, indicating that the effect is statistically fragile. Similar results are also obtained from the extreme bounds associated with the top 10% best-fit models, although the details are reported. The adjusted $R^2$ values range from 0.019 to 0.036, indicating very low explanatory power of the models. Overall, the evidence of the EBA point to fragile evidence for the daylight saving effect.

5 Conclusion

The stock market anomalies have strong implications to many fundamental issues in finance, such as market efficiency, asset-pricing, and behavioral finance. A large number of anomalies have been reported in the literature, ranging from the calendar anomalies such as the Monday effect (e.g., Chang et al. 1993) and January effect (e.g., Keim 1983) to those based on investors’
mood such as seasonal depression (Kamstra et al. 2003), weather (Saunders, 1993; Hirshleifer and Shumway, 2003), and sports events (Edmans et al., 2007; Kaplanski and Levy, 2010a). There are anomalies which may have weak links to investors’ mood, such as lunar phase (Yuan et al., 2006), political cycles (Santa-Clara and Valkanov, 2003) and Halloween effect (Bouman and Jacobsen, 2002). There are authors who are skeptical about the presence of these anomalies, including Black (1993), Connolly (1991), Pinegar (2002), Kelly and Meschke (2010), and Kim (2017, 2019).

This paper makes a contribution to the literature by addressing the following two points in relation to the stock market anomalies. First, as Black (1993) points out, these anomalies are based on shaky theoretical grounds and there is a danger of data-mining where researchers selectively report only statistically significant results under model uncertainty. Harvey (2017) also raises similar concerns in the context of p-hacking in empirical finance. In other words, the evidence in favour of these anomalies are subject to a high degree of model uncertainty, which have not been fully addressed in each individual studies. Noting that none of the studies that reported these anomalies conduct extensive specification search, any individual study can be biased due to (intentional or unintentional) selective reporting of econometric results, as Christensen and Miguel (2018; p.931) point out. Second, statistical significance of these anomalies are established based solely on the “p-value less than 0.05” criterion, often under a large sample or massive sample size. There are concerns that the p-value is a misleading measure of statistical evidence (Berger and Sellke, 1987); and its conventional thresholds are arbitrary and inappropriate, especially when sample size is large (Kim and Ji, 2015; Kim, 2017; Kim et al., 2018; Kim and Choi, 2019). In addition, sole reliance on this criterion may provide incorrect statistical decisions (Harvey, 2017; Kim, 2019). That is, the anomalies reported are established based on statistical significance whose reliability is questionable. The recent statement made by the American Statistical Association further supports these concerns and criticisms (Wasserstein and Lazar, 2016).

In light of the above points, the aim of this paper is to re-evaluate the robustness of the statistical evidence reported in the recent literature of stock market anomalies. To this end, we employ the extreme bounds analysis pioneered by Leamer (1985) and further elaborated by Sala-i-Martin (1997), which provides an effective tool to evaluate the robustness of the statistical results obtained under model uncertainty. As an extensive sensitivity analysis strongly recommended for observational studies (Leamer, 1985), it is capable
of isolating the effects of data-mining, \( p \)-hacking, or multiple testing. As for the measure of statistical significance, in addition to the \( p \)-value criterion with its threshold adjusted with multiple testing following Harvey et al. (2016), we use the Bayes factor of Zellner and Siow (1980) as an alternative criterion for statistical significance. The latter is established based on a sensible balance between the probabilities of Type I and II errors, especially under a large or massive sample. We also employ graphical measures where necessary.

In this paper, we focus on the U.S. stock market using the daily data from 1950 to 2016. We examine a range of anomalies identified or reported after the study of Sullivan et al. (2001) who provide evidence that the calendar anomalies are the results of data-mining. We analyze the anomalies related with investors’ mood (winter blues, weather, sports events, daylight saving), sports event, calendar anomalies (Halloween effect), political cycle, and lunar phase. Our extreme bounds analysis reveals clear evidence that none of these anomalies represents a robust effect under model uncertainty. There are cases where the signs of the estimated coefficients are consistently negative or positive, but their statistical significance is questionable and explanatory power is often negligible. We have also found the evidence in some cases that the reported effect size estimates may have been inflated due to (unintentional) data-mining. In addition, we have identified the cases where the presence of outliers are not taken into account, which may have inflated the effect size estimates and their statistical significance. This is consistent with the point made by Sullivan et al. (2001; p.251) that “data with important outliers, such as those observed in stock market returns, are particularly prone to data-mining biases.”

Based on these findings, we conclude that the recent stock market anomalies reported in the literature are highly fragile when the model uncertainty is effectively taken into account. These results form the evidence that the U.S. stock market has been more efficient than previously considered in the literature. Hou et al. (2018) provide a similar finding from a large scale replication exercise of the other types of anomalies of the U.S. stock market.

**Acknowledgement**

We would like to thank Imad Moosa for his constructive comments.
### Appendix 1: A glance at stock market anomalies

<table>
<thead>
<tr>
<th>Studies</th>
<th>Types of Anomalies</th>
<th>Study context</th>
<th>Key findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Abraham &amp; Ikenberry (1994)</td>
<td>Day of the week effect</td>
<td>1963 to 1991; CRSP</td>
<td>The mean Monday return is -0.12%. When Friday's return is negative, Monday's return is also negative and vice versa.</td>
</tr>
<tr>
<td>3. French (1980)</td>
<td>Day of the week effect</td>
<td>1953 to 1977; S&amp;P 500</td>
<td>The mean Monday return is -0.17%. This outcome is inconsistent both in the trading time and calendar time hypothesis.</td>
</tr>
<tr>
<td>4. Keim &amp; Stambaugh (1984)</td>
<td>Day of the week effect</td>
<td>1928 to 1982; S&amp;P Composite</td>
<td>The mean Monday return over the whole sample period is -0.19%.</td>
</tr>
<tr>
<td>6. Birru (2018)</td>
<td>Day of the week effect</td>
<td>1963-2013; CRSP</td>
<td>Low (high) mood on Monday (Friday) leads to relatively low (high) returns for speculative stocks which is consistent with mispricing explanation that the pattern is driven by the speculative leg.</td>
</tr>
<tr>
<td>7. Jaffe &amp; Westerfield (1985)</td>
<td>Day of the week effect in international markets</td>
<td>USA: S&amp;P 500, 1962 to 1983; UK: The financial Times ordinary share index, 1950 to 1983; Japan: The Nikkei Dow index, 1970 to 1983; Canada: Toronto stock exchange index; 1976 to 1983; Australia: The Statex actuaries index; 1973 to 1982.</td>
<td>The mean Monday return for each of the countries: US: -0.126% UK: -0.142% Canada: -0.139% Japan: -0.02% Australia: -0.052% The time zone theory or the seasonality in foreign exchange rates cannot explain the weekend effect in international stock market.</td>
</tr>
<tr>
<td></td>
<td>Authors</td>
<td>Topic</td>
<td>Time Period</td>
</tr>
<tr>
<td>---</td>
<td>--------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td>8.</td>
<td>Zhang, Lai &amp; Lin (2017)</td>
<td>Day of the week effects in international markets</td>
<td>1990-2016; 15 emerging market indices and 13 developed market indices</td>
</tr>
<tr>
<td>9.</td>
<td>Rozeff &amp; Kinney Jr (1976)</td>
<td>January effect</td>
<td>1904-1974; NYSE listed stocks</td>
</tr>
<tr>
<td>10.</td>
<td>Keim (1983)</td>
<td>January effect</td>
<td>1963-1979; NYSE &amp; AMEX listed firms in CRSP</td>
</tr>
<tr>
<td>12.</td>
<td>Keloharju, Linnainmaa &amp; Nyberg (2016)</td>
<td>Seasonality in monthly returns</td>
<td>1963-2011; CRSP</td>
</tr>
<tr>
<td>13.</td>
<td>Ariel (1987)</td>
<td>Turn of the month effect</td>
<td>1963-1981; CRSP</td>
</tr>
<tr>
<td>14.</td>
<td>Urquhart &amp; McGroarty (2014)</td>
<td>Monday effect, January effect, Turn-of-the-month effect, and Halloween effect</td>
<td>1900-2013; Dow Jones Industrial Average</td>
</tr>
<tr>
<td>15.</td>
<td>Lynch, Puckett &amp; Yan (2014)</td>
<td>Turn of the year effect</td>
<td>1999-2005; CRSP; Abel Noser Solutions data for institutional trading</td>
</tr>
<tr>
<td>16.</td>
<td>Lakonishok &amp; Smidt (1988)</td>
<td>Turn of the week, turn of the month, turn of the year and holiday effect</td>
<td>1897-1986; Dow Jones Industrial Average</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>17. Kim &amp; Park (1994)</strong></td>
<td>Holiday effect</td>
<td>1963-1986; NYSE, AMEX &amp; NASDAQ</td>
<td>More than 20 times the normal rate of return is observed before holidays. The holiday effect exists in all the three major stock markets in the US. It is also present in the UK and Japan.</td>
</tr>
<tr>
<td><strong>18. Białkowski, Etebari &amp; Wisniewski (2012)</strong></td>
<td>Ramadan effect</td>
<td>1989-2007; 14 Muslim countries</td>
<td>The mean annualized return during the Ramadan was 38.09%, compared to an annualized return of 4.32% throughout the rest of the year.</td>
</tr>
<tr>
<td><strong>19. Bouman &amp; Jacobsen (2002)</strong></td>
<td>The Halloween effect</td>
<td>1970-1988; 37 countries</td>
<td>The sell-in-May effect or the Halloween effect is widespread. The Halloween strategy outperforms the Buy and Hold strategy in all countries except Hong Kong and South Africa. The Halloween strategy involves (a) buying stocks at the end of October and selling them at the beginning of May, and (b) investing in short-term Treasury bonds from the end of April to the end of October.</td>
</tr>
<tr>
<td><strong>20. Kamstra, Kramer &amp; Levi (2003)</strong></td>
<td>Seasonal affective disorder (SAD) effect</td>
<td>1928-2000; 12 stock indices in nine countries</td>
<td>The SAD effect exists in every northern country and it is greater for a country with a higher latitude.</td>
</tr>
<tr>
<td><strong>22. Santa-Clara &amp; Valkanov (2003)</strong></td>
<td>Political cycles effect</td>
<td>1927-1998; CRSP</td>
<td>The excess return in the equal (value) weighted portfolio is 16% (9%) higher under Democratic presidencies than that of Republican presidencies.</td>
</tr>
<tr>
<td><strong>23. Novy-Marx (2014)</strong></td>
<td>Party of the US president, the weather, global warming, sunspots, and the stars</td>
<td>1961-2012; US</td>
<td>The US stock market outperformed T-bills by 87 (12) basis points per month under Democratic (Republican) presidencies. The party of the US president, the weather, global warming, sunspots, and the conjunctions of the planets can predict the performance of popular anomalies.</td>
</tr>
<tr>
<td><strong>24. Yuan, Zheng &amp; Zhu (2006)</strong></td>
<td>Lunar effect</td>
<td>1973-2001; 48 countries of MSCI</td>
<td>Stock returns are lower on the days around a full moon than on the days around a new moon. The extent of the return gap is 3% to 5% per annum based on global portfolios.</td>
</tr>
<tr>
<td>No.</td>
<td>Author(s) (Year)</td>
<td>Effect</td>
<td>Period</td>
</tr>
<tr>
<td>-----</td>
<td>-------------------------</td>
<td>-------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>25.</td>
<td>Saunders (1993)</td>
<td>Weather effect</td>
<td>1927-1989; Dow Jones Industrial Average, NYSE &amp; AMEX</td>
</tr>
<tr>
<td>26.</td>
<td>Hirshleifer &amp; Shumway (2003)</td>
<td>Weather effect</td>
<td>1982-1997; 26 countries</td>
</tr>
<tr>
<td>27.</td>
<td>Cao &amp; Wei (2005)</td>
<td>Temperature effect</td>
<td>1962-2001; 8 countries</td>
</tr>
<tr>
<td>28.</td>
<td>Kaplanski &amp; Levy (2010a)</td>
<td>FIFA world cup effect</td>
<td>1950-2007; NYSE</td>
</tr>
<tr>
<td>29.</td>
<td>Kaplanski &amp; Levy (2010b)</td>
<td>Aviation disasters effect</td>
<td>1950-2007; NYSE Composite Index from CRSP</td>
</tr>
<tr>
<td>30.</td>
<td>Drakos (2010)</td>
<td>Terrorism effect</td>
<td>1994-2004; 22 country indices</td>
</tr>
</tbody>
</table>

*This table covers calendar anomalies and anomalies related to politics, weather, global warming, sunspots, stars, and religious and other occasions. For popular anomalies related to firm-specific characteristics, see Novy-Marx (2014, p.145).*
# Appendix

## A List of common doubtful variables

<table>
<thead>
<tr>
<th>Description</th>
<th>Notes</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Jan$</td>
<td>January dummy</td>
<td>1 for January days</td>
</tr>
<tr>
<td>$M$</td>
<td>Monday dummy</td>
<td>1 for Mondays</td>
</tr>
<tr>
<td>$A$</td>
<td>Dummy for autumn months</td>
<td>1 for autumn months</td>
</tr>
<tr>
<td>$SAD$</td>
<td>Seasonal affective disorder</td>
<td>time to set to sunrise</td>
</tr>
<tr>
<td>$T$</td>
<td>Dummy for tax-year days</td>
<td>1 for the last trading days or the first five trading days the tax year</td>
</tr>
<tr>
<td>$C$</td>
<td>Cloud cover</td>
<td>1 for low; 0 for medium; -1 for high cover</td>
</tr>
<tr>
<td>$G$</td>
<td>Temperature</td>
<td>Saunders (1993)</td>
</tr>
<tr>
<td>$P$</td>
<td>Precipitation</td>
<td>Saunders (1993)</td>
</tr>
<tr>
<td>$E$</td>
<td>FIFA World Cup event dummy</td>
<td>1 for the event days</td>
</tr>
<tr>
<td>$HaL$</td>
<td>Halloween dummy</td>
<td>1 from November to April</td>
</tr>
<tr>
<td>$DL$</td>
<td>Daylight Saving dummy</td>
<td>1 for Monday following daylight saving change</td>
</tr>
<tr>
<td>$PRE$</td>
<td>Political cycle</td>
<td>1 for Democratic Presidency</td>
</tr>
<tr>
<td>$H$</td>
<td>Holidays dummy</td>
<td>1 for non-weekend holidays</td>
</tr>
<tr>
<td>$L$</td>
<td>Lunar phase</td>
<td>1 for full moon days</td>
</tr>
<tr>
<td>$J_1, J_2$</td>
<td>Worst and best trading days</td>
<td>1 for 10 worst and best trading days</td>
</tr>
</tbody>
</table>
B   Box-and-Whisker Plots

Figure B.1: Stock return against cloud cover dummy

Figure B.2: Stock return against Halloween Dummy
References


