Testing for Signal-to-Noise Ratio in Linear Regression: A Test under Large or Massive Sample

Jae H. Kim[*]

Philip I. Ji[†]

April 27, 2023

Abstract

This paper proposes a test for the signal-to-noise ratio applicable to a range of significance tests and model diagnostics in a linear regression model. It is particularly useful when sample size is large or massive, where, as a consequence, conventional tests frequently lead to inappropriate rejection of the null hypothesis. The test is conducted in the context of the traditional F-test, with its critical values increasing with sample size. It maintains desirable size properties under a large or massive sample size, when the null hypothesis is violated by a practically negligible margin. The test is widely applicable to many empirical studies in business and management.

Keywords: Effect size, Large sample size bias, Statistical inference, False positive

JEL Classification: C1, G1

[*] Independent researcher; Email: jaekim8080@gmail.com

[†] Corresponding author: Department of Economics, Dongkuk University Seoul; Email: philipji0422@dongguk.edu

> All models are wrong, but some are useful
>
> George Box, 1976

1 Introduction

In the era of big data, it is now commonplace to conduct statistical research with large or massive samples. It is particularly so in the business and management disciplines, in which researchers have access to databases that provide extensive and comprehensive samples: see, for example, Lin et al. (2013) and Michaelides (2020). In practical applications of statistical inference, the p-value required for attaining statistical significance decreases as a function of sample size. As Wasserstein and Lazar (2016) point out, "any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough...". This is essentially because a null hypothesis is always violated and a practically negligible deviation from it almost always occur (De Long and Lang, 1992; Startz, 2014; Rao and Lovric, 2016). The conventional distributional theory is based on the assumption that the null hypothesis holds exactly and literally. However, this assumption is likely to be violated even when the null hypothesis is practically true. The consequence is that, if a conventional level of significance is maintained, any null hypothesis can be rejected as long as the sample size is sufficiently large (Spanos, 2018). This leads to a situation in which any variable in a regression model can be statistically significant under a large sample size. Similarly, any model may be judged to be mis-specified when a model specification test is employed. If the effect size or deviation from the true model is practically negligible, rejection of the null hypothesis constitutes a Type I error or a false positive.

This problem has contributed to what is called the significance test crisis where most of research findings based on statistical significance are arguably false or cannot be replicated: see, for example, Ioannidis (2005) and Johnstone (2022). A number of authors in a variety of fields of enquiry have raised alarms. These authors include Lin et al. (2013), Harford (2014), Andraszewicz et al. (2015), Kim and Ji (2015), Harvey (2017), Kim et al. (2018), and Wasserstein et al. (2019). In the era of big data, the conventional statistical test is facing a large challenge, casting serious doubt on its relevance and applicability as a basis for decisions (Gandomi and Haider, 2015; Algaba et al., 2020; Kim, 2022). This suggests that the way we conduct statistical inference be modified, especially in the era of big data, as Rao and Lovric (2016) argue. This is particularly so in business disciplines where critical decisions are routinely made based on results from large samples.

The purpose of this paper is to propose an alternative method of statistical inference, one that does not suffer from the effect of large sample size. Many popular statistical tests in linear regression can be expressed (roughly) as the sample size multiplied by signal-to-noise ratio. By formulating a test on the latter, it is possible to separate the effect of large sample size from hypothesis testing. The proposed test has a desirable property that its critical value increases with sample size. As a result, the test is not likely to reject a (practically) true null hypothesis when sample size is large enough. As an application, the FIFA (Fédération Internationale de Football Association) World Cup effect on stock return is examined using the test for signal-to-noise ratio. The proposed test is also applied to the results gathered from the past meta-studies in accounting and finance. The next section provides a simple example to motivate the paper, followed by Section 3 for methodological details. Sections 4 and 5 present applications, and Section 6 concludes the paper.

2 Motivation

To further motivate the paper, we use the regression results reported in Hong et al. (2012) as examples. To test how sensitive stock prices are to earnings news in relation to whether stock is actively shorted or not, Hong et al. (2012) employ the event study methodology and conduct the regression of CAR (cumulative abnormal return) against two indicator (dummy) variables: UEHIGH for high unexpected earnings; HISR for high short ratio; and their interaction (UEHIGH×HISR). The sample consists of 119,785 quarterly observations of the stocks listed on the U.S. exchanges from 1994 to 2007. Two of their regression results are reproduced in Table 1, Regression I is the linear regression of CAR against UEHIGH and HISR, and Regression II includes their interaction term as an explanatory variable, additional to UEHIGH and HISR.

Table 1: An example of regression results under a massive sample

	Regression I		Regression II	
	Coefficient	Standard Error	Coefficient	Standard Error
UEHIGH	3.46	0.06	3.27	0.07
HISR	-0.09	0.06	-0.27	0.07
UEHIGH × HISR			0.55	0.12
R^2	0.042		0.043	

The regression results are reproduced from Table 2 of Hong et al. (2012). Other coefficients such as the intercept terms are not included for simplicity.

The interaction term in Regression II has the coefficient estimate of 0.55 with the t-statistic 4.58 (p-value = 0), which indicate its statistical significance at any level. However, it has increased the model's explanatory power only by 0.001, in terms of R^2. That is, the incremental contribution of this additional term to the model's fit is negligible at best. Can we say that this additional term makes a significant contribution to the model's explanatory power? If not, how can we reconcile this negligible contribution with the t-statistic of such a large magnitude and the p-value of 0? This paper is aimed to explain why these conflicting results occur and how we can make a more sensible statistical decision in such a situation.

3 Methodology

Consider a linear regression model

$$Y = \beta_1 + \beta_2 X_2 + ... + \beta_K X_K + u, \tag{1}$$

where Y is the dependent variable, Xs are independent (non-stochastic) variables, and β's are unknown coefficients, and u is an error term that follows an independent normal distribution with zero mean and a fixed variance σ^2. Researchers are often interested in testing for a linear restriction $R\beta = r$, where R is a $J \times (K-1)$ matrix with a full rank, $\beta = (\beta_2, ..., \beta_K)'$ and r is $(J \times 1)$ vector of scalars.

The proposed test is applicable to the situation where a restriction is violated by an practically negligible margin, which is highly likely in practice as Rao and Lovric (2016) have demonstrated. In this case, the researcher should not reject such a restriction based on practical insignificance or negligible effect size. However, with the conventional test, a test statistic is an increasing function of sample size, and it will almost always reject the restriction with a sufficiently large sample size, leading to what Spanos (2018) calls the fallacy of rejection. This is largely because the critical value of the test does not change with sample size. The proposed test has a desirable property of its critical value increasing with sample size, at the same rate as the test statistic.

3.1 Effect size and signal-to-noise ratio

Effect size, defined as the "the value measuring the strength of the relationship" (Cohen, 2013), should be the focus of statistical research, as many authors have argued: see, among others, McCloskey and Ziliak (1996), Bakker et al. (2019), and Kim (2022). The main question is whether it is practically large enough to be useful. Consider a simple case where the researcher compares the mean effects from two different groups in response to a treatment. A natural measure of effect size is Cohen's (2013) d, which can be written in a general form as

$$d \equiv \frac{\bar{x_1} - \bar{x_2}}{s},$$

where $\bar{x_i}$ is mean effect for group i and s is the pooled standard deviation of effects. It measures the difference in the effects per unit of their variability. As benchmark values, Cohen (2013) suggests the effect to be small if d less than 0.2; medium if it is 0.5, large if it is greater than 0.8.

Returning to our motivating example in Section 2, the t-statistic can roughy be written as $\frac{\sqrt{119785}\times 0.55}{s} = 4.58$, where s is a measure of variability. This means that $0.55/s = 0.013$ is an estimate of Cohen's d, indicating a negligible effect size. This shows that how severely the t-statistic can be inflated by the sample size and reject a H_0 that is practically true. At a conventional level of significance such as 0.05, the power of the test will be virtually equal to 1 when the sample size is as large as 119,785. This means that the probability of Type II error (failure to reject false H0) of the test is 0 (1-power), implying that the probability of Type I error (level of significance) is infinitely larger than that. Hence, when H_0 is violated even with a negligible margin, Type I error (rejecting a true H_0) occurs almost surely.

In the context of linear regression in (1), a general measure of effect size is what is called the (observed) signal-to-noise ratio, which is defined as

$$\eta \equiv \frac{R_1^2 - R_0^2}{1 - R_1^2}, \tag{2}$$

where R_0^2 is the coefficient of determination from (1) with the restriction $R\beta = r$ imposed and R_1^2 the coefficient of determination without imposing it. Taking a simple restriction $\beta_2 = 0$ as an example, the signal-to-noise ratio η given in (2) measures the contribution of the variable X_2 to the model's fit or in-sample predictability, relative to its noise-component. It is also known as Cohen's f^2 in behavioral science as a measure of effect size. According to Cohen (2013, Chapter 9), the η values of 0.02, 0.15, and 0.35 respectively serve as thresholds for a small, medium, and large effect. An advantage of signal-to-noise ratio (Cohen's f^2) is that the effect size is expressed using a unit-free measure of R^2.

Now consider what may be called the population signal-to-noise ratio

$$\eta_p \equiv \frac{R_{p1}^2 - R_{p0}^2}{1 - R_{p1}^2}, \tag{3}$$

where R^2_{p0} denotes the population coefficient of determination under $R\beta = r$, R^2_{p1} denotes the population coefficient of determination without imposing it. While $\eta_p \geq 0$, $\eta_p = 0$ represents the case where the restriction $R\beta = r$ holds and it adds exactly zero contribution to the model's fit or in-sample predictability. When it adds a non-zero contribution, $\eta_p > 0$. The question is whether this contribution is substantively important or significant.

By focusing on the signal-to-noise ratio, researchers are able to conduct an incremental goodness-of-fit analysis in relation to the restriction being tested, which can provide dramatically different research outcomes from those of the conventional statistical test. Ohlson (2015) has identified the lack of such incremental goodness-of-fit analyses as a major shortcoming of contemporary empirical research in business areas.

3.2 A test for signal-to-noise ratio

The proposed test is one-tailed, with the following null and alternative hypotheses:

- $H_0 : \eta_p \leq \eta_0$
- $H_1 : \eta_p > \eta_0$

where η_0 denotes the value of η_p under H_0. The test is conducted with the conventional F-test statistic, which can be expressed as

$$F = \frac{(R^2_1 - R^2_0)/J}{(1 - R^2_1)/(T - K)} = \frac{(T - K)\eta}{J}, \tag{4}$$

where T is the sample size. The F-statistic given in (4) follows a non-central F-distribution $F(J, T - K; \lambda)$ with the non-centrality parameter given by (Peracchi, 2001, Theorem 9.2)

$$\lambda = T\frac{R^2_{p1} - R^2_{p0}}{1 - R^2_{p1}} \equiv T\eta_p. \tag{5}$$

Under H_0, the F-statistic follows $F(J, T - K; \lambda = T\eta_0)$: see Appendix for the statistical details. The decision rule is to reject H_0 at the α significance level if the F-statistic in (4) is greater than the critical value $F_{1-\alpha}(J, T-K; \lambda = T\eta_0)$, which is the $(1-\alpha)$ percentile from $F(J, T-K; \lambda = T\eta_0)$. The critical value, as with the F-test statistic, is an increasing function of sample size because the distribution $F(J, T-K; \lambda = T\eta_0)$ moves away from 0 as the value of non-centrality λ increases with T. The proposed test is applicable to a range of significance test in linear regression, as well as to many popular model diagnostics and specification tests.

3.3 Choice of η_0

The proposed test can be conducted by controlling the value of η_0, as the value of the parameter of interest under H_0. This should be set at the minimum value of the signal-to-noise ratio (η_p), where the restriction $R\beta = r$ is practically violated with substantive importance. This value is also associated with what McCloskey and Ziliak (1996) called the "minimum oomph", which represents the smallest value of effect size that matters practically. Researchers may also be guided by the thresholds given by Cohen (2013) given above. Alternatively, the value can be determined by researchers' judgement based on their experience or prior knowledge of the subject matter. For example, suppose Y is a stock return and X_2 is its predictor. In testing for restriction that $\beta_2 = 0$, the researcher may set $\eta_0 = 0.05$, requiring that the predictor X_2 contribute to the explanatory power of the model with a signal-to-noise ratio of at least 5%.

The conventional test (t or F) for the significance of β's is a special case of the proposed framework where $\eta_0 = 0$. One may question whether testing for this boundary point of the entire parameter space is of substantive importance or practical interest. This is because $\eta_0 = 0$ is associated with a point hypothesis such as $\beta_i = 0$ where $R^2_{p1} = R^2_{p0}$: i.e., the variable X_i makes an exactly nil contribution to the in-sample predictability or fit of the model. Such a sharp hypothesis is unlikely to hold exactly and literally in practice, as De Long and Lang (1992) pointed out. In particular, Leamer (1988, p.331) stated

> Genuinely interesting hypotheses are neighbourhoods, not points. No parameter

> is exactly equal to zero; many may be so close that we can act as if they were zero.

Furthermore, Startz (2014, p.123) stated that

> Economic hypotheses are usually best distinguished by some parameter being small or large, rather than some parameter being exactly zero versus non-zero.

On this point, the proposed test is related to the pioneering work of Hodges Jr and Lehmann (1954) on interval-based hypothesis testing and the subsequent literature of testing for equivalence and non-inferiority: see Wellek (2010).

3.4 Monte Carlo results

A Monte Carlo experiment is conducted to evaluate the properties of the proposed test. A linear regression models are considered, whose explanatory power are ranging from being negligible to strong in terms of its signal-to-noise ratio. The probability of rejecting the null hypothesis of the conventional t-test and the test for η_p is evaluated and compared, as the sample size increases.

Consider a simple linear regression model

$$Y_t = \beta_1 + \beta_2 X_{2t} + u_t, \tag{6}$$

where $\beta_1 = 0$ with $Var(X_{2t}) = Var(u_t) = 1$. We consider a range of β_2 to examine the negligible to strong effects. That is, we simulate $\beta_2 \in \{0.05, 0.1, 0.5, 0.75, 1\}$, corresponding to

- $R^2_{p1} \in \{0.0025, 0.01, 0.2, 0.36, 0.5\}$; and
- $\eta_p \in \{0.0025, 0.01, 0.25, 0.56, 1\}$,

while $R^2_{p0} = 0$. These values cover a range of negligible to large effects, as classified by Cohen (2013). All experiments are conducted using 1000 Monte Carlo trials, over a grid of sample sizes from 100 to 10,000, at $\alpha = 0.05$.

We first compare the rejection probabilities of the conventional t-test for $H_0 : \beta_2 = 0$ and the proposed test for $H_0 : \eta_p \leq \eta_0$, when the effect is negligible with $\beta_2 = 0.05$. Figure 1 plots the probability of rejecting $H_0 : \beta_2 = 0$ for the conventional t-test (black line) and the probability of rejecting $H_0 : \eta_p \leq 0.0025$ (blue line). As expected, the conventional t-test rejects the null hypothesis with increasing probability, but the test for signal-to-noise ratio shows rejection probability almost equal to 0.05 even when the sample size is massive. The results demonstrate that, when the signal-to-noise ratio is negligible, the conventional test rejects the null hypothesis of no effect with the probability increasing sharply with sample size, but the proposed test maintains the rejection probability close to the Type I error rate.

Table 2: Power properties of of the test

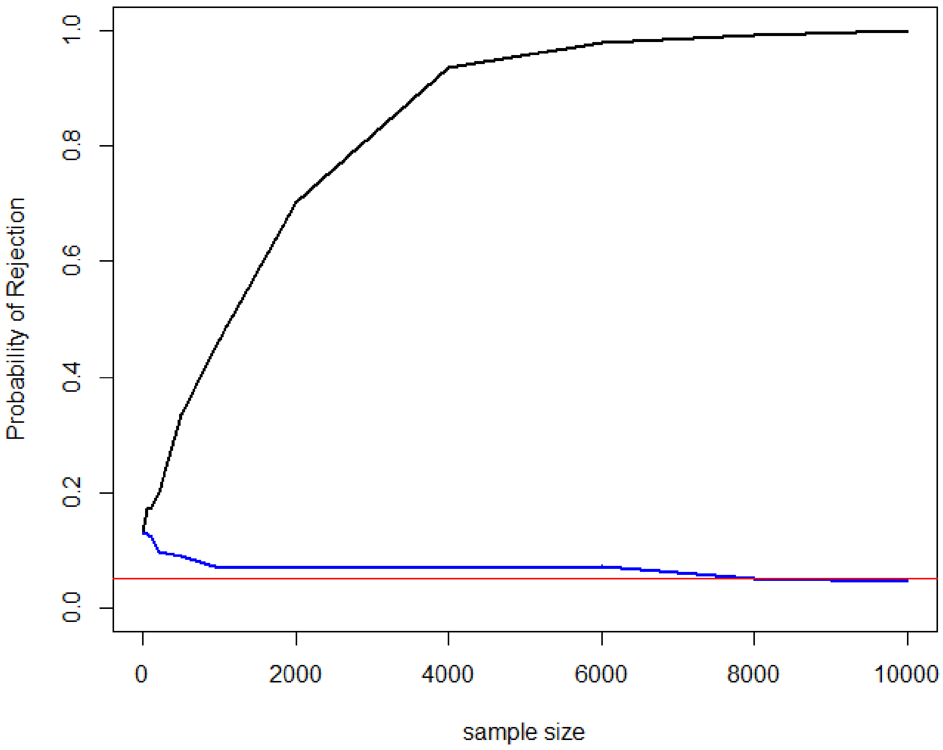
β_2	η_p	100	200	500	1000	5000	10000
0.1	0.01	0	0	0	0	0	0
0.5	0.25	0.065	0.272	0.645	0.813	1	1
0.75	0.56	0.694	0.995	1	1	1	1
1	1	0.995	1	1	1	1	1

The entries are the probabilities of rejecting $H_0 : \eta_p \leq 0.15$ at $\alpha = 0.05$.

Table 2 reports the rejection probabilities (size and power) when the $\beta_2 \in (0.1, 0.5, 0.75, 1)$. We test for $H_0 : \eta_p \leq 0.15$, consistent with value of medium effect as suggested by Cohen (2013). That is, $\beta_2 = 0.1$ represents the case where the effect is negligible, while $\beta_2 \in (0.5, 0.75, 1)$ the cases where the effects are larger than the medium. It appears that, when $\beta_2 = 0.1$, the proposed test does not reject $H_0 : \eta_p \leq 0.15$ with zero frequencies for all sample sizes. In contrast, it rejects $H_0 : \eta_p \leq 0.15$ when the effects are large ($\beta_2 \geq 0.5$) with high frequencies. These rejection probabilities approach 1 quickly as the sample size gets larger.

Note that the proposed test is valid under the assumption that the error term follows an independent normal distribution. Further Monte Carlo experiments are conducted to

Figure 1: Size Properties of Alternative Tests



Black line: probability of rejecting $H_0 : \beta_2 = 0$ using the t-test; Blue line: probability of rejecting $H_0 : \eta_p \leq 0.025$ at $\alpha = 0.05$ ($\beta_2 = 0.05$); Red line is at 0.05.

evaluate the size properties under a range of non-normal distributions including chi-squared, Student's t, and conditionally heteroskedastic error terms. That is, we consider chi-squared distribution with 5 degree of freedom (χ^2_5) and Student-t distribution with 5 degree of freedom (t_5) to represent the error term u in equation (6) which is asymmetric and fat-tailed. We also considered the GARCH(1,1) model of Bollerslev (1986), which is widely observed in financial data, which can be written as

$$\sigma_t^2 = 0.001 + 0.1u_{t-1}^2 + 0.8\sigma_{t-1}^2,$$

where $u_t \sim N(0, \sigma_t^2)$. All non-normal error terms are standardizes so that they have zero mean and unit variance. As reported in Table 3, the proposed test shows the correct size properties in large samples as the probability of rejecting the true H_0 converges to 0.05 with increasing sample size, as also observed in Figure 1. This strongly suggests that the proposed test is robust to a wide range of non-normal and heteroskedastic error terms widely encountered in practice.

Table 3: Size properties of of the test under non-normal errors

T	χ^2_5	t_5	GARCH
10	0.152	0.112	0.132
20	0.143	0.138	0.118
50	0.137	0.139	0.141
100	0.125	0.136	0.126
200	0.107	0.111	0.102
500	0.090	0.083	0.093
1000	0.079	0.097	0.073
2000	0.076	0.06	0.077
4000	0.056	0.058	0.050
6000	0.075	0.063	0.068
8000	0.053	0.063	0.072
10000	0.056	0.063	0.060

The entries are the probabilities of rejecting $H_0 : \eta_p \leq 0.025$ at $\alpha = 0.05$ when $\beta_2 = 0.05$.

4 Application: FIFA World Cup Effect

In empirical finance, it has been widely reported that investors' moods from major sports events have a systematic effect on stock (index) return with exploitable abnormal profits: see, among others Edmans et al. (2007). This accumulated evidence is at odds with the concept of stock market efficiency where such anomalies have little impact on the value of stocks: see Fama (1970).

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

$$R_t = \beta_1 + \sum_{i=2}^{5} \beta_i D_{it} + \beta_6 H_t + \beta_7 T_t + \beta_8 P_t + \beta_9 E_t + u_t, \tag{7}$$

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, P_t is a dummy variable for the annual event period (June and July), and E_t is the dummy variable for the event (FIFA World Cup) days. Kaplanski and Levy (2010) claimed that β_9 is expected to be negative, due to global negative mood effects of the FIFA World Cup, induced by all losing countries' fans at an international level.

4.1 Results based on the conventional test

Using the NYSE composite index return (CRSP) from 1950 to 2007 (14,679 observations), Kaplanski and Levy (2010, Table 2) reported a statistically significant (negative) effect of the FIFA World Cup games on stock returns. In this paper, the model (7) is estimated using the same return data (value-weighted) updated to June 2016 ($T = 16,819$).

It is found the FIFA World Cup event coefficient (β_9) estimate is -0.0015 with the t-statistic of -2.86 (p-value $= 0.0042$) and $R^2 = 0.005392$, similar to those reported in Kaplanski and Levy (2010), as reported in Table 4. The F-statistic for the joint significance of all the slope coefficients ($\beta_2 = ... = \beta_9 = 0$) is 11.39 with the p-value of 0. The event coefficient indicates that the stock return is expected to be lower by 0.15% during the FIFA World Cup days. As for model diagnostics, the RESET of Ramsey (1969) with three augmentation terms for model (7) gives F-statistic of 3.69 with p-value of 0.01. The ARCH(5) test of Engle (1982) for the error term of (7) gives the F-statistic of 615.80 with p-value of 0. From these significance tests and model diagnostics, it is possible to conclude that the event coefficient is negative and statistically significant, with the model showing the error term with strong conditional heteroskedasticity.

Table 4: Test for $H_0 : \eta_p \leq 0.01$ for FIFA World Cup Effect

Test	F-statistic	η	$F_{0.95}(J, T-K; \lambda = T\eta_0)$
$\beta_9 = 0$	8.19	0.00048	213.72
$\beta_2 = ... = \beta_9 = 0$	11.39	0.00542	27.64
$RESET(3)$	3.69	0.00066	71.94
$ARCH(5)$	615.80	0.18318	43.59

Note: RESET(3) refers to the RESET with three augmentation terms, and ARCH(5) the test for ARCH effect with 5 lags, both for model (7). For $\beta_9 = 0$, $F = -2.86^2$. The critical values are calculated using qf function in R (R Core Team (2020)).

4.2 Results based on the proposed test

The problem with the above regression results is that their economic significance is questionable with negligible values of signal-to-noise ratio. That is, in testing for $\beta_9 = 0$, we have $R_0^2 = 0.004907$, $R_1^2 = 0.005392$, which gives $\eta = 0.00048$. In addition, in testing for $\beta_2 = ... = \beta_9 = 0$, we have $\eta = 0.00542$, calculated from $R_1^2 = 0.005392$ and $R_0^2 = 0$. For the RESET, the value of $\eta = 0.00066$, calculated from $R_1^2 = 0.006046$, $R_0^2 = 0.005392$. The ARCH(5) test has $\eta = 0.183$ with $R_1^2 = 0.1548$, $R_0^2 = 0$.

Table 4 reports the results of the test for $H_0 : \eta_p \leq 0.01$ against $H_1 : \eta_p > 0.01$. It should be pointed out that $\eta_0 = 0.01$ is a conservative value, lower than the threshold for a small effect according to Cohen (2013), meaning that a 1% of signal-to-noise ratio is required for the effect size or model to be economically important. From Table 1, for significance of the FIFA event ($H_0 : \beta_9 = 0$) and significance of all slope coefficients ($H_0 : \beta_2 = ... = \beta_9 = 0$), the F-statistics are well below the 5% critical values (213.72 and 27.64 respectively), which indicates that the presence of negligible effects cannot be rejected at the 5% significance level for both cases. The RESET's F-statistic is also far less than the critical value, indicating that the model shows evidence of specification error of negligible magnitude. The model only shows evidence of the strong ARCH(5) effect in the error term with a substantially large value of η of 0.18. That is, $H_0 : \eta_p \leq 0.01$ is clearly rejected with its F-statistic well above the 5% critical value of 43.59.

The above results indicate that the FIFA World Cup effect is associated with practically negligible signal-to-noise ratios and effect sizes. Although the model shows strong ARCH effect in the error term, it shows no evidence of specification error of practical importance. These findings are consistent with the implications of stock market efficiency, where anomalies such as the FIFA World Cup effect are economically negligible. The effect reported by Kaplanski and Levy (2010) is likely to be an outcome of large sample size bias (Michaelides, 2020) or fallacy of rejection (Spanos, 2018).

5 Evaluation from Past Meta-Studies

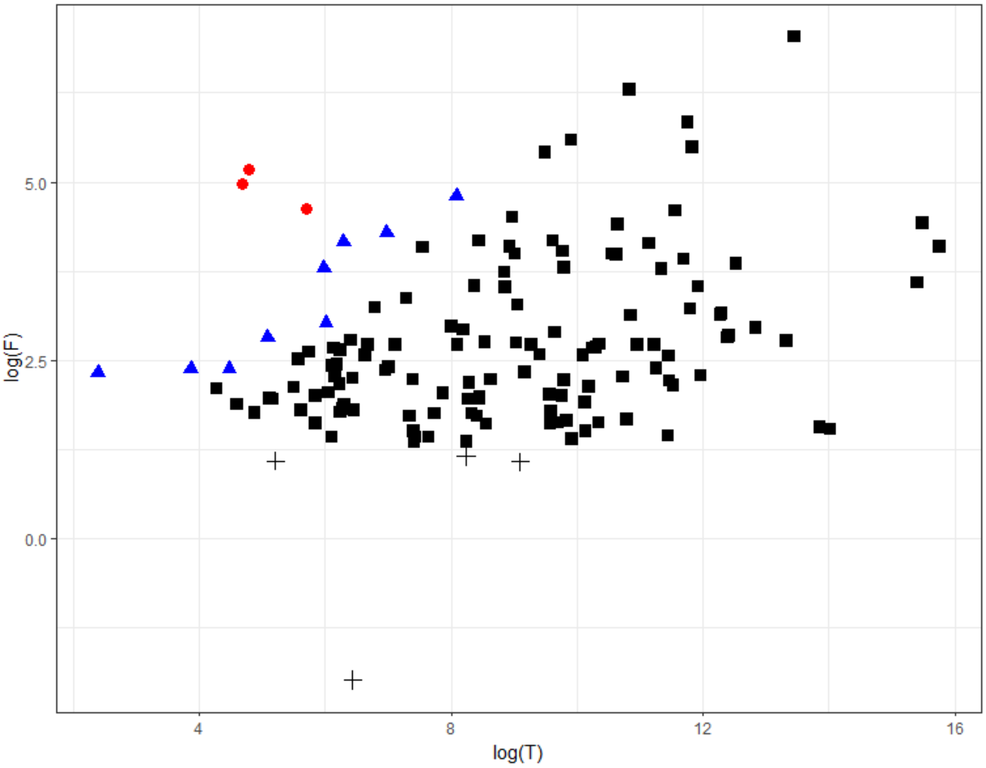
In this section, the test for signal-to-noise ratio is applied to the t-test and F-test results reported in the meta-studies Kim and Ji (2015) and Kim et al. (2018) conducted. From Kim and Ji (2015), we obtain the t-statistics for individual significance of the key variable of the regressions, reported in 161 published articles in four major journals in finance in 2012. Using these t-statistic values, we have conducted the tests for the signal-to-noise ratio $H_0 : \eta_p \leq \eta_0$ with $\eta_0 = \{0, 0.02, 0.15\}$, at the 5% level of significance. Note that $\eta_0 = 0$ corresponds to the conventional t-test, while the values 0.02 and 0.15 are chosen following Cohen (2013), as thresholds for small and medium effects. Figure 2 reports the natural logarithm of the squared t-statistic[1]. against the natural logarithm of sample size ($\log(T)$). The conventional test with $\eta_0 = 0$ is rejected for all cases except four (four cross dots), consistent with the findings of Kim and Ji (2015). For test with $\eta_0 = 0.02$, 129 cases (91%) are not rejected at the 5% level of significance (black square dots and cross), meaning that the signal-to-noise ratio is less than 0.02 for these studies statistically, indicative of small effects. For the test with $\eta_0 = 0.15$, only three cases are rejected, indicating a medium or large effect.

Kim et al. (2018) report the R^2 values from the regressions, reported in 191 article published in major accounting journals in 2014. From these values, the F-statistics for the joint significance of all the explanatory variables are calculated[2], and they are subject

[1] Note that the t-test and F-test are equivalent with $F = t^2$ in testing for individual significance of regression coefficient.

[2] The calculation is based on (4) where the reported R^2 values correspond to R_1^2 ($R_0^2 = 0$).

Figure 2: Test for signal-to-noise ratio for the studies from Kim and Ji (2015)

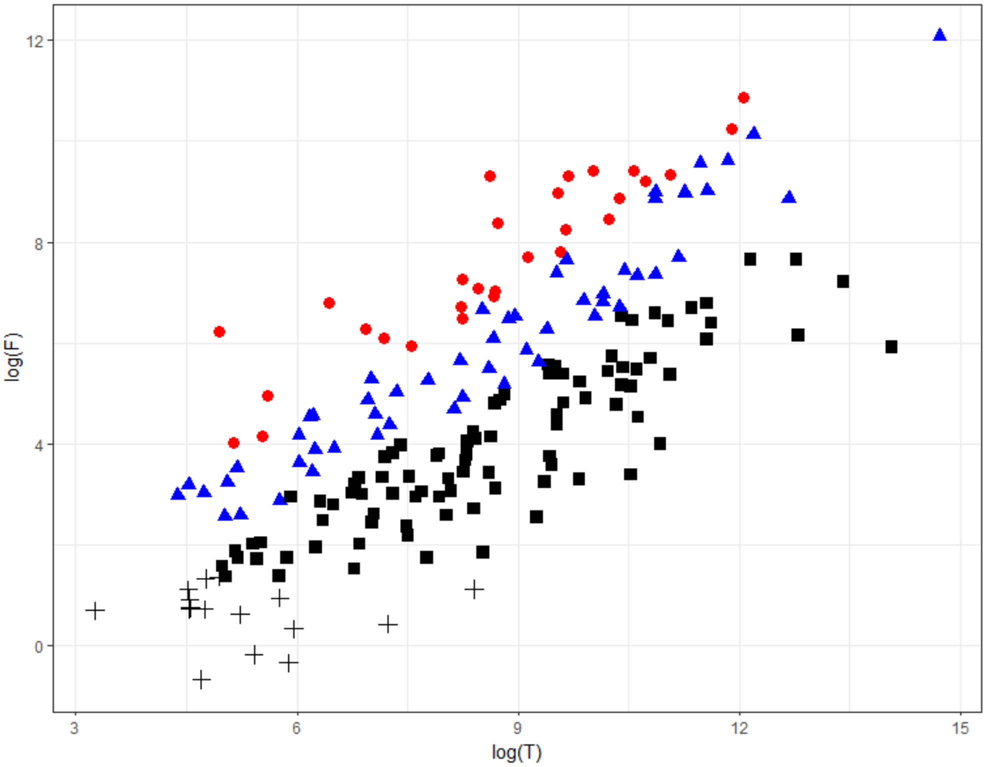


The dots indicate the combinations of $\log(F)$ (where $F = t^2$) and $\log(T)$. The cross represents the cases where the test for $\eta_0 = 0$ is not rejected. The cross and black square dots represent those where the test for $\eta_0 = 0.02$ is not rejected. The cross, black square, and blue triangle dots represent those where the test for $\eta_0 = 0.15$ is not rejected. The red dots indicate those rejected by the test with $\eta_0 = 0.15$. All tests are done at the 5% level of significance.

to the tests for the signal-to-noise ratio with $\eta_0 = \{0, 0.02, 0.15\}$. Figure 2 reports the natural logarithm of the F-statistic ($\log(F)$) against the natural logarithm of the sample size ($\log(T)$). For the conventional F-test with $\eta_0 = 0$, only 8.4% of cases are statistically insignificant at the 5% level (cross dots), which means that more than 90% of the studies have found their model to be statistically significant using the conventional F-test. For the test for $\eta_0 = 0.02$, 57% of the cases found to have signal-to-noise ratio less than 0.02 (square dots and cross), at the 5% level of significance. For the test for $\eta_0 = 0.15$ (medium effect), 85% of the cases found to have signal-to-noise ratio less than 0.15 (square, cross, and blue triangle), at the 5% level of significance. Only 15% of the cases show the joint significance of all the explanatory variables with $\eta_0 = 0.15$, indicated by the red dots.

From the evaluation of the results from the past meta-studies, it is found that only a small

Figure 3: Test for signal-to-noise ratio for the studies from Kim et al. (2018)



The dots indicate the combinations of $\log(F)$ and $\log(T)$. The cross represents the cases where the test for $\eta_0 = 0$ is not rejected. The cross and black square dots represent those where the test for $\eta_0 = 0.02$ is not rejected. The cross, black square, and blue triangle dots represent those where the test for $\eta_0 = 0.15$ is not rejected. The red dots indicate those rejected by the test with $\eta_0 = 0.15$. All tests are done at the 5% level of significance.

proportion of the studies in empirical accounting and finance show the signal-to-noise ratio statistically greater than 0.15, for both t-test for individual significance of an explanatory variable and the F-test for the joint significance. A large proportion of the studies are found to be associated with negligible or small effect sizes. Concerns on the tiny effect sizes in empirical research in accounting and finance have been raised recently by Ohlson (2015), Harvey (2017), and Mitton (2022).

6 Conclusion

With the availability of big data, the conventional test for significance is losing its ground for applicability and relevance in statistical research (Algaba et al. (2020); Gandomi and

Haider (2015)). The conventional distribution theory under a point null hypothesis should be modified, because it is applicable only when the null hypothesis holds exactly and literally. The latter situation is unrealistic because a point null hypothesis is almost always violated: see De Long and Lang (1992) and Rao and Lovric (2016). Under a large or massive sample size, conventional tests are problematic because they can be subject to a very high probability of Type I error when the null hypothesis is violated by a practically negligible margin. This problem has contributed to accumulation of false positive research findings: see Ioannidis (2005), Harvey (2017), and Johnstone (2022). Many authors including Rao and Lovric (2016), Wasserstein and Lazar (2016), and Wasserstein et al. (2019) have raised alarms, with a proposal that the current paradigm of statistical testing should be modified. In business disciplines, a number of authors have called for changes recently: they include Andraszewicz et al. (2015), Ohlson (2015), Harvey (2017), Johnstone (2022), Li et al. (2022), Mitton (2022), and Kim (2022).

In this paper, a test for the signal-to-noise ratio is proposed, making use of the property that a conventional test statistic can be expressed (roughly) as sample size multiplied by the signal-to-noise ratio. By focusing on the signal-to-noise ratio, it is possible to identify the effect size from the test statistic by removing the effect of sample size. This is also closely related with the incremental goodness-of-fit analysis, which Ohlson (2015) proposes as a method that can overcome the problems of the statistical research that uses large samples. The proposed test is implemented in the context of the conventional F-test, with its critical values obtained from a non-central F-distribution. The test has a desirable property that its critical value is an increasing function of sample size, at the same rate as the test statistic. As a result, the test is able to effectively control the Type I error rate, when the null hypothesis is practically true under a large or massive sample size, as demonstrated in a Monte Carlo experiment. As an application, the FIFA (Fédération Internationale de Football Association) World Cup effect on stock return is tested and re-evaluated using the test for signal-to-noise ratio. It is found that the effect is economically and statistically negligible in terms of signal-to-noise ration, in contrast with the past studies that reported statistically significant results using the conventional tests. The test is also applied to the past empirical results in accounting and finance reported by Kim and Ji (2015) and Kim et al. (2018), and it is found that a large proportion of the studies are associated with the signal-to-noise ratio indicative of small or negligible effect sizes.

The proposed test has a wide applicability under a large or massive sample size. It can be applied to a range of significance tests, tests for linear restrictions, and model diagnostics. The test can also be conducted in general linear models suitable for time series data as well as for panel data, under a wide range of non-normal error terms.

Data Availability Statement

The data used in this paper and R code are available from the authors on request.

Appendix

We provide an outline of the proof of the results given in (4) and (5). A formal and rigorous proof can be found in Peracchi (2001, Theorem 9.2), while simpler proofs are available in Kelley and Maxwell (2008, p.176) and Kim and Choi (2021, p.16-18).

The F-test statistic for $H_0 : R\beta = r$ against $H_1 : R\beta \neq r$ is defined as

$$F \equiv \frac{(SSR_0 - SSR_1)/J}{SSR_1/(T-K)},$$

where SSR_i is the residual sum of squares from the model under H_i $(i = 0, 1)$. The result in (4) follows, by dividing the numerator and denominator of the above F-statistic by SST (total sum of squares) and noting that $R_i^2 = 1 - \frac{SSR_i}{SST}$. Under $H_0 : R\beta = r$, since

$$\frac{SSR_0 - SSR_1}{\sigma^2} \sim \chi^2_J; \frac{SSR_1}{\sigma^2} \sim \chi^2_{T-K},$$

the F-statistic is distributed as $F(J, T-K; \lambda = 0)$, as the ratio of the independent χ^2_J and χ^2_{T-K} distributions scaled by their respective degrees of freedom. Under $H_1 : R\beta \neq r$ where $\eta_0 > 0$, it can be shown that

$$\frac{SSR_0 - SSR_1}{\sigma^2} \sim \chi^2_{J,\lambda},$$

where $\chi^2_{J,\lambda}$ denotes the non-central χ^2 distribution with J degrees of freedom with the non-centrality parameter λ given in (5). By definition, the F-statistic in (4) follows the non-central F-distribution $F(J, T-K; \lambda > 0)$ with the non-centrality parameter λ: see Greene (2012, p.1053) .

R code

This is a shortened version of the R code. The data and the full code are available from the authors on request. This R code replicates the results of the tests for $\beta_9 = 0$ and $\beta_2 = ... = \beta_9 = 0$ reported in Table 4.

```
# Regression in Equation (7) Unrestricted
Reg1 = lm(ret~Holiday+Mon+Tue+Wed+Thu+Tax+JunJul+Event,data=data)


# Regression in Equation (7) under restriction beta9 = 0
Reg0 = lm(ret~Holiday+Mon+Tue+Wed+Thu+Tax+JunJul,data=data)
# Regression results
summary(Reg1); summary(Reg0)
# Calculation of signal-to-noise ration and critical values
T=16819;          # Sample size
K= 9;             # Number of X variables in (7)
eta0=0.01         # The value under H0
R12=0.005392      # R-square under H1
R02 = 0.004907    # R-square under H0
```

```
eta = (R12-R02)/(1-R12)  # Signal-to-noise ratio
J=1             # number of restriction under H0: beta9=0
F=eta*(T-K)/J   # F-statistic (F=t^2)
Fc=qf(0.95,df1=J,df2=T-K,ncp=T*eta0)  # Critical value
J=8                # number of restriction under H0: beta2= ... = beta9 = 0
eta = R12/(1-R12)   # Signal-to-noise ratio
F=eta*(T-K)/J       # F-statistic
Fc=qf(0.95,df1=J,df2=T-K,ncp=T*eta0) # Critical value
```

References

Algaba, A., D. Ardia, K. Bluteau, S. Borms, and K. Boudt (2020). Econometrics meets sentiment: An overview of methodology and applications. *Journal of Economic Surveys 34*(3), 512–547.

Andraszewicz, S., B. Scheibehenne, J. Rieskamp, R. Grasman, J. Verhagen, and E.-J. Wagenmakers (2015). An introduction to bayesian hypothesis testing for management research. *Journal of Management 41*(2), 521–543.

Bakker, A., J. Cai, L. English, G. Kaiser, V. Mesa, and W. Van Dooren (2019). Beyond small, medium, or large: Points of consideration when interpreting effect sizes. *Educational Studies in Mathematics 102*, 1–8.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics 31*(3), 307–327.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. Ebook.

De Long, J. B. and K. Lang (1992). Are all economic hypotheses false? *Journal of Political Economy 100*(6), 1257–1272.

Edmans, A., D. Garcia, and Ø. Norli (2007). Sports sentiment and stock returns. *The Journal of finance 62*(4), 1967–1998.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the econometric society 50*(4), 987–1007.

Fama, E. F. (1970). Efficient capital markets: a review of theory and empirical work. *The Journal of Finance 25*(2), 383–417.

Gandomi, A. and M. Haider (2015). Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management 35*(2), 137–144.

Greene, W. H. (2012). *Econometric analysis* (7th ed.). Pearson Education.

Harford, T. (2014). Big data: A big mistake? *Significance 11*(5), 14–19.

Harvey, C. R. (2017). Presidential address: The scientific outlook in financial economics. *The Journal of Finance 72*(4), 1399–1440.

Hodges Jr, J. and E. Lehmann (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society: Series B (Methodological) 16*(2), 261–268.

Hong, H., J. D. Kubik, and T. Fishman (2012). Do arbitrageurs amplify economic shocks? *Journal of Financial Economics 103*(3), 454–470.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine 2*(8), e124.

Johnstone, D. (2022). Accounting research and the significance test crisis. *Critical Perspectives on Accounting 89*, 102296.

Kaplanski, G. and H. Levy (2010). Exploitable predictable irrationality: The fifa world cup effect on the us stock market. *Journal of Financial and Quantitative Analysis 45*(2), 535–553.

Kelley, K. and S. E. Maxwell (2008). Sample size planning with applications to multiple regression: Power and accuracy for omnibus and targeted effects. *The SAGE Handbook of Social Research Methods*, 166–192.

Kim, J. H. (2022). Moving to a world beyond p-value¡ 0.05: a guide for business researchers. *Review of Managerial Science 16*(8), 2467–2493.

Kim, J. H., K. Ahmed, and P. I. Ji (2018). Significance testing in accounting research: a critical evaluation based on evidence. *Abacus 54*(4), 524–546.

Kim, J. H. and I. Choi (2021). Choosing the level of significance: A decision-theoretic approach. *Abacus 57*(1), 27–71.

Kim, J. H. and P. I. Ji (2015). Significance testing in empirical finance: A critical review and assessment. *Journal of Empirical Finance 34*, 1–14.

Leamer, E. E. (1988). Things that bother me. *Economic Record 64*(4), 331–335.

Li, G., M. K. So, and K. Y. Tam (2022). Identifying the big shots—a quantile-matching way in the big data context. *ACM Transactions on Management Information Systems (TMIS) 13*(2), 1–30.

Lin, M., H. C. Lucas Jr, and G. Shmueli (2013). Research commentary—too big to fail: large samples and the p-value problem. *Information Systems Research 24*(4), 906–917.

McCloskey, D. N. and S. T. Ziliak (1996). The standard error of regressions. *Journal of economic literature 34*(1), 97–114.

Michaelides, M. (2020). Large sample size bias in empirical finance. *Finance Research Letters 41*, 101835.

Mitton, T. (2022). Economic significance in corporate finance. *The Review of Corporate Finance Studies in press.* cfac008.

Ohlson, J. A. (2015). Accounting research and common sense. *Abacus 51*(4), 525–535.

Peracchi, F. (2001). *Econometrics.* Wiley. https://lccn.loc.gov/00043915.

R Core Team (2020). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological) 31*(2), 350–371.

Rao, C. R. and M. M. Lovric (2016). Testing point null hypothesis of a normal mean and the truth: 21st century perspective. *Journal of Modern Applied Statistical Methods 15*(2), 2–21.

Spanos, A. (2018). Mis-specification testing in retrospect. *Journal of Economic Surveys 32*(2), 541–577.

Startz, R. (2014). Choosing the more likely hypothesis. *Foundations and Trends in Econometrics 7*, 119–189.

Wasserstein, R. L. and N. A. Lazar (2016). The asa statement on p-values: context, process, and purpose. *The American Statistician 70*(2), 129–133.

Wasserstein, R. L., A. L. Schirm, and N. A. Lazar (2019). Moving to a world beyond "p < 0.05". *The American Statistician 73*(sub1), 1–19.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority* (2nd ed.). Chapman and Hall/CRC.