ABACUS A Journal of Accounting, Finance and Business Studies

THE UNIVERSITY OF SYDNEY

ABACUS, Vol. 57, No. 1, 2021 doi: 10.1111/abac.12172

JAE H. KIM AND IN CHOI

Choosing the Level of Significance: A Decision-theoretic Approach

In many areas of science, including business disciplines, statistical decisions are often made almost exclusively at a conventional level of significance. Serious concerns have been raised that this contributes to a range of poor practices such as *p*-hacking and data-mining that undermine research credibility. In this paper, we present a decision-theoretic approach to choosing the optimal level of significance, with a consideration of the key factors of hypothesis testing, including sample size, prior belief, and losses from Type I and II errors. We present the method in the context of testing for linear restrictions in the linear regression model. From the empirical applications in accounting, economics, and finance, we find that the decisions made at the optimal significance levels are more sensible and unambiguous than those at a conventional level, providing inferential outcomes consistent with estimation results, descriptive analysis, and economic reasoning. Computational resources are provided with two *R* packages.

Key words: Bootstrapping; Expected loss; Optimal significance level; Power analysis; Statistical significance.

If economists have natural constants,
then the most well-known is 0.05.
Keuzenkamp and Magnus (1995)

INTRODUCTION

Hypothesis testing is an integral part of statistical research in many fields of science, including business disciplines such as accounting, economics, and finance. The level of significance is a key input into hypothesis testing. It controls the critical value and power of the test, thus having a consequential impact on the test outcome and decision making. It is the probability of rejecting the true null hypothesis, representing the degree of risk that the researcher is willing to take for

JAE H. KIM (J.Kim@latrobe.edu.au) is with the Department of Economics and Finance, La Trobe University. IN CHOI is with the Department of Economics, Sogang University. Dan Nordman, Benjamin Scheibehenne, Abul Shamsuddin, and Xiangkang Yin provided their comments on an earlier version of the paper, which was circulated with the title ‘How to Choose the Level of Significance: A Pedagogical Note’. We also thank Denzil Fiebig, Imad Moosa, Hal Stern, Dick Startz, and Tom Stanley for their comments on this version of the paper. An earlier version of the paper was presented to the 2018 Asia Meeting of the Econometrics Society, Seoul, Korea.

making a wrong decision. It is a convention to set the level at 0.05, while 0.01 and 0.10 levels are also widely used. However, researchers should be reminded that this choice is only a convention, based on R. A. Fisher's argument that a one in twenty chance represents an unusual sampling occurrence (Moore and McCabe, 1993, p. 473); and that there is no scientific basis for it (Arrow, 1960, p. 70; Lehmann and Romano, 2005, p. 57). This arbitrary threshold was established in the 1920s when the sample size of more than 100 was rarely used (Lindley, 2014, p. 358).

Concerns have been raised that empirical researchers often use the significance level in a ritualistic and mindless way (see, for example, Keuzenkamp and Magnus, 1995; Gigerenzer, 2004; and Kim *et al.*, 2018). This poor practice has been recognized as a main contributor to widespread distortion of scientific process and conclusions, as expressed in the recent statement made by the American Statistical Association (Wasserstein and Lazar, 2016). The problems include data-mining (Black, 1993), multiple testing (Harvey *et al.*, 2016), *p*-hacking (Harvey, 2017), replication crisis (Peng, 2015), publication bias (Basu and Park, 2014), and a large number of possible false positive research outcomes (Kim and Ji, 2015; Harvey, 2017; Kim *et al.*, 2018). Accounting scholars who have raised similar alarms include Johnstone (1990, 1994), Basu (2015), Dyckman and Zeff (2014, 2015), Dyckman (2016), Ohlson (2015, 2018), and Stone (2018).

In fact, the choice of significance level should be made with careful consideration of the key factors of hypothesis testing. For example, the level of significance should be set as a decreasing function of sample size in consideration of statistical power (Arrow, 1960; Leamer, 1978; Degroot and Schervish, 2012; Spanos, 2017, p. 21), and with a full consideration of the implications of Type I and Type II errors (see, for example, Arrow, 1960; Skipper *et al.*, 1967; Das, 1994; Poirier, 1995). Recently, Kim and Choi (2017) apply a decision-theoretic approach to choosing the optimal level of significance for popular unit root tests, following Leamer (1978). They find that the test conducted at the optimal level often provides inferential outcomes consistent with economic reasoning, overturning a range of empirically motivated puzzles produced at the conventional level.

Although a good deal of academic research has been conducted on the issue of choosing the level of significance for many years, these studies are not readily accessible to students or researchers. Modern textbooks in statistics and econometrics provide surprisingly little guidance on this fundamental question, as Goldberger (1991, p. 238) points out. In addition, the proposals made by previous authors need to be updated with modern econometric methods for further improvements. The purpose of this paper is to introduce a decision-theoretic approach to choosing the optimal significance level for classical hypothesis testing in the context of the linear regression model. Calculating the optimal significance level requires an extensive power analysis. In this paper, we use two alternative methods of power estimation, one valid under the assumption of normality and the other using bootstrapping. The latter often provides a superior alternative to the former in small samples, taking full account of sampling variability, also robust to non-normality or heteroscedasticity. The optimal level is chosen so that the expected loss from hypothesis testing is minimized. This choice is made from all possible combinations of the level of significance and probability of Type II error (1-power), which Leamer (1978) refers to as the line of enlightened judgement. We also propose methods of calculating the weighted optimal significance level, with the power calculated from a range of possible points under the alternative hypothesis. We discuss the roles that prior belief and relative loss play in determining the optimal level of significance.

We present empirical applications in accounting, economics, and finance, along with a numerical example used by Gelman and Stern (2006), where it is demonstrated that more economically sensible and unambiguous inferential outcomes are obtained at the optimal level of significance. The decisions at the conventional level can be ambiguous, often in conflict with economic reasoning or estimation results. In the next section, we discuss the key factors affecting hypothesis testing with simple illustrative and numerical examples. This is followed by the sections that present a brief review of the past studies; and the methods of estimating the power of the test (or probability of Type II error); and calculating the optimal level of significance. Empirical applications are also presented, using the accompanying packages written in R (R Core Team, 2017). The details of R packages and programs used for empirical applications are provided in the Appendix.

KEY FACTORS OF HYPOTHESIS TESTING

In this section, we discuss the key factors affecting hypothesis testing using simple illustrative and numerical examples. They include:

- losses from incorrect decisions;
- the researcher's prior beliefs for the null (H_0) and alternative hypotheses (H_1);
- the power of the test (or sample size); and
- substantive importance (and the effect size) of the relationship being tested.

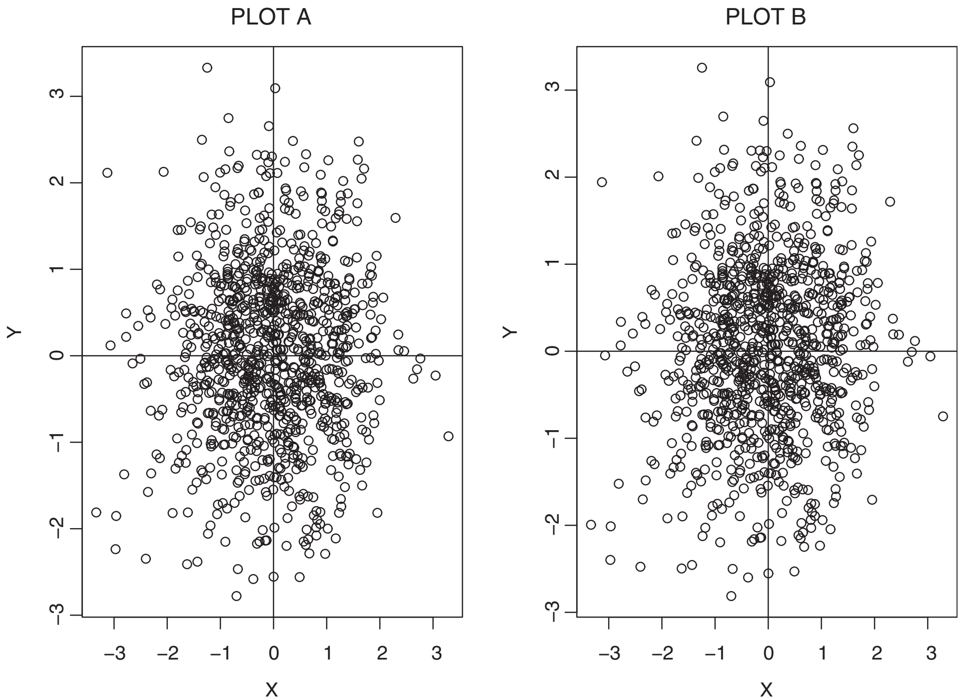
Let α represent the level of significance, which is the probability of rejecting the true null hypothesis (Type I error); and β the probability of accepting the false null hypothesis (Type II error), while $1 - \beta$ is the power of the test. Let $P \equiv Prob(H_0)$ denote the researcher's prior belief for H_0 with $Prob(H_1) = 1 - P$, while L_1 and L_2 denote the losses from Type I and II error respectively, with $k \equiv L_2/L_1$ being the relative loss. Note that $P = 0.5$ means that the researcher *a priori* believes that H_0 and H_1 are equally likely; and $k = 1$ means that Type I and II errors have equal consequences.

Testing For No Pregnancy

Consider a doctor testing if a patient is pregnant or not. The doctor tests for H_0 that the patient is not pregnant and against H_1 that she is. Type I error is diagnosing a patient as pregnant when in fact she is not; Type II error is that the

FIGURE 1

STATISTICAL SIGNIFICANCE AND SAMPLE SIZE: A LARGE SAMPLE



Note: X and Y are generated from $N(0, 1)$ with a sample size of 1000. In Plot A, Y and X are independent; and the regression slope coefficient is statistically insignificant ($t = 1.23$, p-value = 0.22). In Plot B, Y and X are related with negligible correlation of 0.05, but the regression slope coefficient is statistically significant at the 1% level ($t = 2.82$, p-value = 0.004). The same random numbers are used for both cases.

patient is told that she is not pregnant when in fact she is. Suppose only two clinical tests for pregnancy are available: Tests A and B. Test A has a 5% chance of showing evidence for pregnancy when the patient is in fact not pregnant (Type I error); but it has a 20% chance of indicating evidence for no pregnancy when in fact she is pregnant (Type II error). Test B has a 20% chance of Type I error and a 5% chance of Type II error. Test A has a four times smaller chance of committing Type I error, yet it has a four times higher chance of Type II error.

The doctor believes that Type II error has substantially more serious consequences than Type I error ($L_2 > L_1$) since the former risks the lives of the patient and baby. On this basis, the doctor prefers Test B ($\alpha = 0.20$, $\beta = 0.05$) as a safer option, which also has a higher power of 0.95. Note that the value of $k = L_2/L_1$ can be substantially higher than 1, due to the seriousness of misdiagnosing a pregnant patient. This means that the doctor is highly cautious about committing Type II error, favouring H_1. In this case, the doctor may also have a strong prior

TABLE 1

BURDEN OF PROOF IN LEGAL TRIALS

Burden of Proof	Description	Trials
Preponderance of evidence	Greater than 50% chance	Civil, Family: Child support, Unemployment benefit
Clear and convincing evidence	Highly and substantially probable	Civil, Criminal: Paternity, juvenile delinquency, probate, decision to remove life support
Beyond reasonable doubt	No plausible reason to believe otherwise	Criminal: Imprisonment, death penalty

belief that the patient is pregnant (a low value of $P < 0.5$). These values of P and k are reflected in the doctor's choice of Test B that has a lower value of β.

Legal Trial as Hypothesis Testing

Hypothesis testing is often likened to a legal trial where a defendant is assumed to be innocent (H_0) until the evidence showing otherwise is presented. The jury returns a guilty verdict when they are convinced by the evidence presented. If the evidence is not sufficiently compelling, then they deliver a 'not guilty' verdict. In making their judgement, the jury tries to avoid the incorrect decisions; namely returning a guilty verdict to an innocent defendant (Type I error) and a not guilty verdict to a guilty defendant (Type II error). In a court of law, there are different standards of evidence that should be presented, as Table 1 shows. For a civil trial, a low burden of proof (preponderance of evidence) is required since the consequences of wrong decisions are not severe. However, for a criminal trial where the final outcome may be imprisonment or even a death penalty, a high bar (beyond reasonable doubt) is required to reject the null hypothesis of innocence.

This means that the legal system is using different levels of significance (or critical values) depending on the consequences of wrong decisions. That is, the level of significance for 'preponderance of evidence' may be as high as 0.40; and that for 'clear and convincing evidence' can be as low as 0.01. To meet the level of 'beyond reasonable doubt', the level of significance should be much lower (say 0.005), which puts in place a high bar for a guilty verdict. That is, as the gravity (or loss) of committing Type I error increases, the value of α should be adjusted downward.

A jury who strongly believes *a priori* that the defendant is innocent until proven guilty (the high value of $P > 0.5$) will assign a higher value to L_1 than L_2 (a low of value of $k < 1$). This is because the jury is highly cautious about committing Type I error, favouring H_0. For example, when the burden of proof is 'beyond reasonable doubt', a high value of P and/or a low value of k will lead to choosing a very low value of α. Using a conventional value such as 0.05 in this case is too lenient and may result in a wrong decision with a serious consequence.

Power and Sample Size

Suppose $(X_1, \ldots, X_T)$ is a random sample from a normal distribution with the population mean μ and known standard deviation of 2. We test for H_0: $\mu = 0$ against H_1: $\mu > 0$. The test statistic is $Z = 0.5T^{0.5}\bar{X}$, where $\bar{X}$ is the sample mean. At the 5% level of significance, H_0 is rejected if Z is greater than the critical value of 1.645. Note that the Z-statistic tends to increase as sample size T grows; or, equivalently, the p-value tends to decrease as T increases. This means that when the level of significance (or the critical value) is fixed, the null hypothesis is more and more likely to be rejected as the sample size increases. Assuming $k = 1$ and $P = 0.5$ for simplicity and without loss of generality, we demonstrate that it is reasonable to set the level of significance as a decreasing function of sample size, as the following example shows.

Consider H_1: $\mu = 0.5$. Table 2 presents $\beta = P(Z < 1.645|\mu = 0.5, \sigma = 2)$, along with the power and critical values for a range of sample sizes. The upper panel presents the case where α is fixed at 0.05 for all sample sizes, while the lower panel presents the case where it is set as a decreasing function of sample size and in balance with the value of β. The upper panel shows that, when the sample size is small, the value of β is unreasonably high compared to $\alpha = 0.05$, with a woefully low power of the test. When the sample size is large, the power of the test is high, but it appears that the value of α is unreasonably high compared to that of β. For example, when the sample size is 300, $\alpha = 0.05$ is 12.5 times higher than the value of β. In this case, a negligible deviation from the null hypothesis may appear to be statistically significant (see Figure 1 and the next subsection and the related discussion).

From the lower panel, we can see that, by achieving a perfect balance between the probabilities of committing Type I and II errors, the test enjoys a

TABLE 2

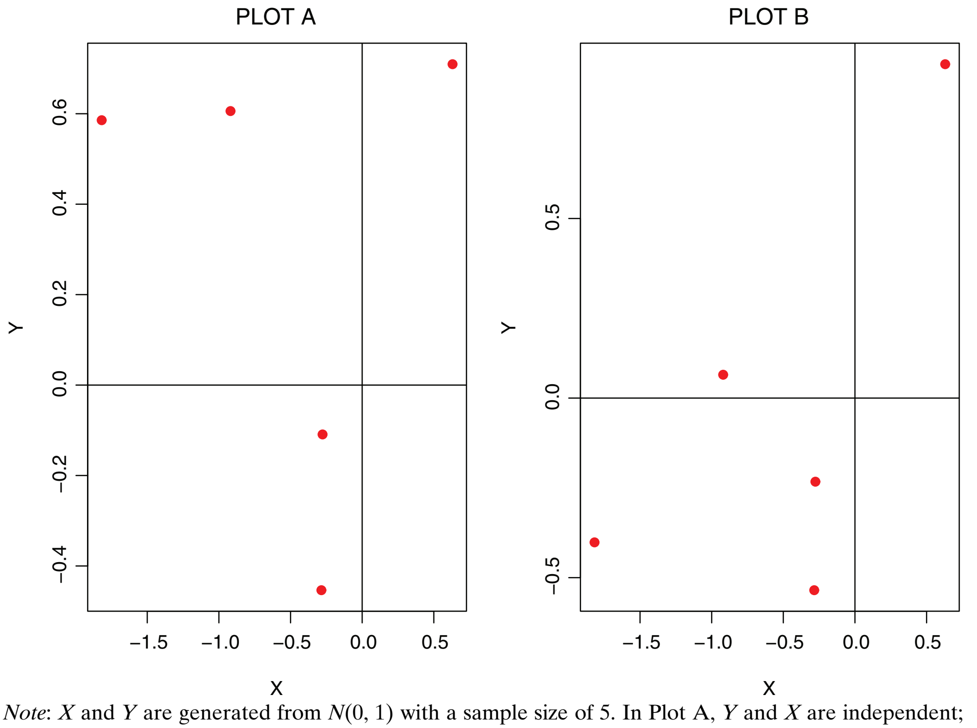
α, β, SAMPLE SIZE, POWER, AND CRITICAL VALUES

T	α	β	$1-\beta$	CR
α: fixed at 0.05				
10	0.05	0.80	0.2	1.645
50	0.05	0.45	0.55	1.645
100	0.05	0.20	0.80	1.645
200	0.05	0.03	0.97	1.645
300	0.05	0.004	0.996	1.645
α: decreases with sample size				
T	α	β	$1-\beta$	CR
10	0.35	0.35	0.65	0.40
50	0.19	0.19	0.81	0.89
100	0.11	0.11	0.89	1.25
200	0.04	0.04	0.96	1.76
300	0.015	0.015	0.985	2.17

T: sample size; α: level of significance; β: Probability of Type II error; CR: critical value.

Figure 2

STATISTICAL SIGNIFICANCE AND SAMPLE SIZE: A SMALL SAMPLE



Note: X and Y are generated from $N(0, 1)$ with a sample size of 5. In Plot A, Y and X are independent; and the regression slope coefficient is statistically insignificant (t = 0.20, p-value = 0.85). In Plot B, Y and X are related with a moderate correlation of 0.50, but the regression slope coefficient is statistically insignificant at the 10% level (t = 1.49, p-value = 0.23). The same random numbers are used for both cases.

substantially higher power for nearly all cases. For example, when the sample size is 10 with $\alpha = 0.05$, the power of the test is only 0.20. However, if it is set at 0.35, the power of the test is 0.65. When $n = 300$, setting $\alpha = 0.015$ provides a balance with the value of β. More importantly, the sum of the probabilities of Type I and II errors $\alpha + \beta$ is always smaller than when α is fixed at 0.05. As we shall see later, this value represents the expected loss from hypothesis testing, and the optimal level of significance is chosen so that the expected loss is minimized. In general, a higher power of the test can be achieved when α is set as a decreasing function of sample size and in balance with the value of β (see also Figure 3 and the related discussion). In fact, Arrow (1960) proposes a test in which the value of α is set equal to that of β, which he called the equal-probability test.

Another factor that determines the power of the test is the point at which the power is calculated under H_1; for example, H_1: $\mu = 0.5$ in the above example. This point should be the minimum value of substantive importance under H_1. That is, this value should represent an economically meaningful deviation from H_0, which is what Ziliak and McCloskey (2004) refer to as the 'minimum oomph'. Obviously, the power will be higher (lower) when this value is further from (closer to) the value under H_0. Most ideally, the value can be determined by in-depth economic analysis, which can often be subjective. It can also be difficult to determine when the test is conducted jointly on a number of parameters of the model. One way of overcoming these difficulties is to consider the power of the test over a grid of possible values. We will elaborate on this proposal in the section on optimal level of significance.

Substantive Importance (Effect Size)

Figure 1 presents two scatter plots (labeled A and B) between random variables Y and X, both with a sample size of 1,000. The two plots are almost identical, showing no linear associations. In fact, Y and X are independent in Plot A; but in Plot B, they are related with the correlation of 0.05. Regressing Y on X in Plot A, the slope coefficient is 0.04 with t-statistic 1.23 and p-value 0.22, indicating there is no statistical significance at any reasonable level. In Plot B, the regression slope coefficient is 0.09 with t-statistic 2.82 and p-value 0.004. In this case, although X and Y are related with a negligible correlation, the regression slope coefficient is statistically significant at the 1% level.

Figure 2 plots two scatter plots (labelled A and B) when the sample size is 5. In Plot A, Y and X are independent; but in Plot B, they are related with a moderately strong correlation of 0.50 with a clear positive relationship. In Plot A, the estimated slope coefficient is small (− 0.09) and statistically insignificant, as might be expected, yet in Plot B, the estimated slope coefficient (0.42) is large but statistically insignificant (t-statistic = 1.49 and p-value = 0.23) at the 5% level. In this case, although X and Y are related with a moderately high linear association, the slope coefficient is statistically insignificant at any conventional level of significance.

The data presented in Figures 1 and 2 illustrate that the substantive importance or effect size of the relationship should be taken into account in hypothesis testing. A regression with negligible effect size should be dismissed even if it is statistically significant at a conventional level, while a regression with a large effect size should be taken seriously even if it is statistically insignificant. On this point, we note that presenting t-statistic and p-value alone, without discussing the effect size or performing basic data analyses, can give a wrong impression or illusion about the true nature of the relationship, especially when the decision is made at a conventional significance level. From a survey of academic economists, Soyer and Hogarth (2012) find that regression statistics can create an illusion of strong association. They find that the surveyed economists provide better predictions when they are presented with a simple visual representation of the data than when they are confronted only with regression statistics such as t-statistic and p-value. Gigerenzer (2004, p. 599) emphasizes the importance of conducting descriptive and exploratory analyses, rather than mechanical hypothesis testing with yes/no answers. Leamer (1988) is also concerned that empirical researchers do not widely use graphs, pointing out that they can help identify the critical features of a data set.

In this paper, we propose that the level of significance be adjusted based on the key factors of hypothesis testing, such as the sample size. For the data set presented in Figure 1, considering the large sample size (and high power), a much lower level of significance than 0.05 (such as 0.001) should be used, which will deliver the decision of no statistical significance. For the data set presented in Figure 2, considering the small sample size (and low power), the level of significance should be set at a much higher level than 0.05 (such as 0.30), leading to the rejection of no linear association in Plot B. These proposals for the adjustment of significance level will be justified in the section on optimal level of significance, below.

It is clear from the illustrative and numerical examples discussed in this section that it is not sensible to use the same level of significance (such as 0.05) every time for all applications. It is unscientific and can lead to fallacious claims (Spanos, 2017, p. 20), potentially incurring huge social costs (see Ziliak and McCloskey, 2008). The choice should be made in a scientific way, in consideration of the key factors of hypothesis testing discussed in this section and also of the context of the application at hand.

LITERATURE REVIEW

Grave concerns have been raised that statistical significance based on the '*p*-value less than 0.05' criterion is being widely abused and misused in statistical research in many fields of science. Recently, the American Statistical Association issued a statement that improper use of the *p*-value criterion is distorting the scientific process and invalidating many scientific conclusions (Wasserstein and Lazar, 2016, p. 131). For example, Black (1993) argues that most investment anomalies identified in finance are likely to be the result of data-mining; while Kandel and Stambaugh (1995) find that the *p*-value as a measure of evidence is often in conflict with economic significance in the context of asset-allocation decisions. Abuse and misuse of the *p*-value criterion are also closely related to the publication bias and replication crisis (see, for example, Basu and Park, 2014; Peng, 2015), where an unreasonably high proportion of published results is statistically significant but often not reproducible by replication exercises. In this section, we provide a brief review of the literature in this area: the first subsection covers the general econometric and statistical issues; while the second reviews the studies from accounting and finance.

Statistical and Econometric Issues

While the *p*-value itself is often misunderstood or misinterpreted (see, for example, Wasserstein and Lazar, 2016), it has been pointed out that the arbitrary threshold of statistical significance has been contributing to a range of problems. Theoretically, it is well-known to be a root cause of a conflict between the classical and Bayesian methods of hypothesis testing (see Lindley, 1954). In practical applications, it has contributed largely to the misuse or abuse of the *p*-value criterion. For example, Keuzenkamp and Magnus (1995, p. 20) report, from a survey of the papers published in economics journals, that 'the choice of significance levels seems arbitrary and depends more on convention and, occasionally, on the desire of an investigator to reject or accept a hypothesis'. In his comprehensive critique of econometrics, Moosa (2017, p. 98) considers the arbitrary choice of significance level as one of the shortcomings of econometric analysis. On this point, we believe that, if this threshold is chosen scientifically and objectively, the problems associated with abuse and misuse of statistical significance, and those associated with the shortcomings of econometric analysis, can largely be avoided. In fact, a number of authors have been making the point that the level of significance should not be fixed, but chosen carefully in consideration of the key factors of hypothesis testing. In this section, we provide a brief review of the related past studies.

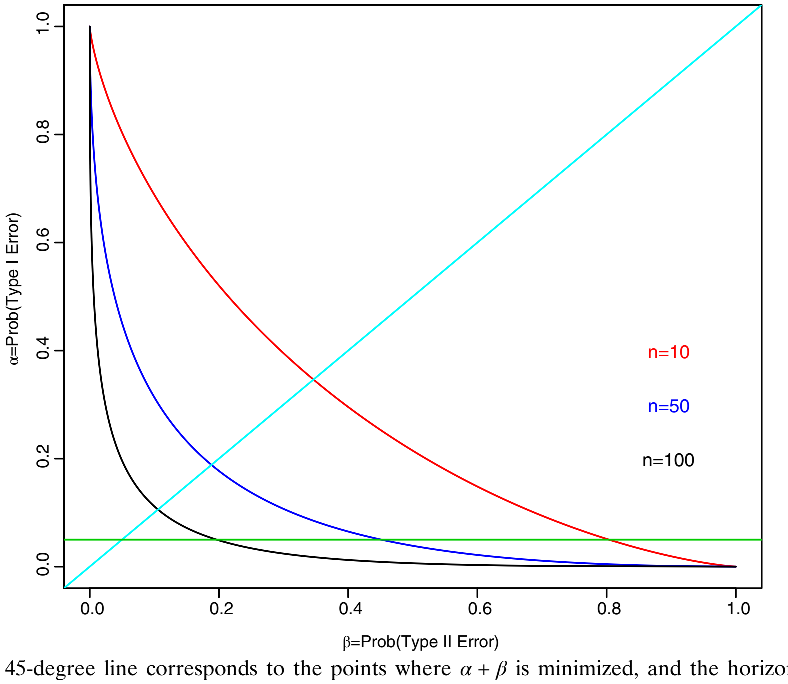
In econometrics, Arrow (1960) and Leamer (1978, Chapter 4) make the most notable early contributions to this issue by presenting a detailed analysis as to how the level of significance should be chosen in consideration of sample size, statistical power, and expected loss. Arrow (1960) proposes the equal-probability test where α is set equal to β; while Leamer (1978) introduces the line of enlightened judgement where the choice of an optimal value of α can be made. The latter is obtained by plotting all possible combinations of (α, β) given the sample size. Leamer (1978) demonstrates how the optimal level of significance can be chosen by minimizing the expected losses from Type I and II errors, as a function of sample size, prior belief, and expected loss. Other authors who made the similar proposals include Manderscheid (1965), Das (1994), and Perrichi and Pereira (2016), while Poirier (1995) and DeGroot and Schervish (2012) cover the issue at an advanced textbook level.

Figure 3 presents three lines of judgement corresponding to the (α, β) values given in Table 2 when the sample size is 10, 50, and 100. The line shows a trade-off between α and β, where a higher (lower) value of α is associated with a lower (higher) value of β given the sample size. As the sample size increases, the line shifts towards the origin as the power (the value of β) increases (decreases). The horizontal line at $\alpha = 0.05$ represents the case where the level of significance is fixed at 0.05. The (α, β) values in the upper panel of Table 2 correspond to the points where this line and the lines of enlightened judgement intersect. The 45-degree line connects the points where the expected loss from hypothesis testing ($\alpha + \beta$ in a simple case) is minimized for each line of enlightened judgement, corresponding to the (α, β) values in the lower panel of Table 2. These points represent the optimal levels of significance based on a decision-theoretic approach: more technical details will be given in the next section.

From Figure 3, it is clear that, when the sample size is small and the power is low, the optimal levels appear to be much higher than 0.05. On this point, Winer (1962) states that 'when the power of the tests is likely to be low, and when Type I and Type II errors are of approximately equal importance, the 0.3 and 0.2 levels of significance may be more appropriate than the .05 and .01 levels' (cited in Skipper *et al.*, 1967). In proposing the equal-probability test, Arrow (1960, p. 73)

FIGURE 3

EXAMPLES OF THE LINE OF ENLIGHTENED JUDGEMENT



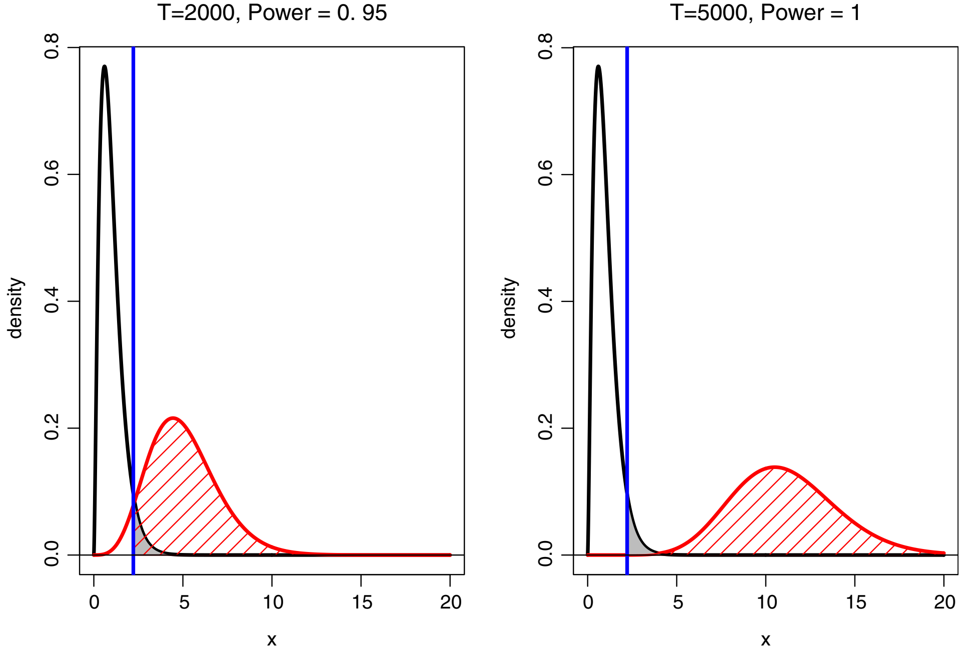
Note: The 45-degree line corresponds to the points where $\alpha + \beta$ is minimized, and the horizontal one to $\alpha = 0.05$.

states that, when the power of a test is low, the level of significance as high as 0.4 should be chosen, although it may be perceived to be 'outrageously higher' than 0.05. In fact, these claims have been further supported by two subsequent studies for statistical tests with low power. Fomby and Guilkey (1978) show, through extensive Monte Carlo simulations, that the optimal level of significance for the Durbin-Watson test should be around 0.5, which is much higher than the conventional levels. Based on the line of enlightened judgement, Kim and Choi (2017) report that the optimal levels of unit root testing are in the 0.20 to 0.40 range, with the findings that many economic puzzles driven by unit root testing are the products of conducting the test at a conventional significance level. These results are alarming in light of Ioannidis *et al.* (2017), who report that most economic research studies are under-powered based on their meta-analytic survey.

It is also clear from Figure 3 that the level of significance should be adjusted downward as the sample size increases and the power becomes higher. That is, when the sample size is large, the level of significance should be set at a much lower value than 0.05 or 0.01. For example, Arrow (1960, p. 74) states that, as the sample size increases, 'the inadequacy of the constant-level-of-significance testing becomes more glaring … all the gain due to large sample size is taken out in the

Figure 4

POWER OF THE F-TEST: $R^2_{p1} = 0.01$, $R^2_{p0} = 0$



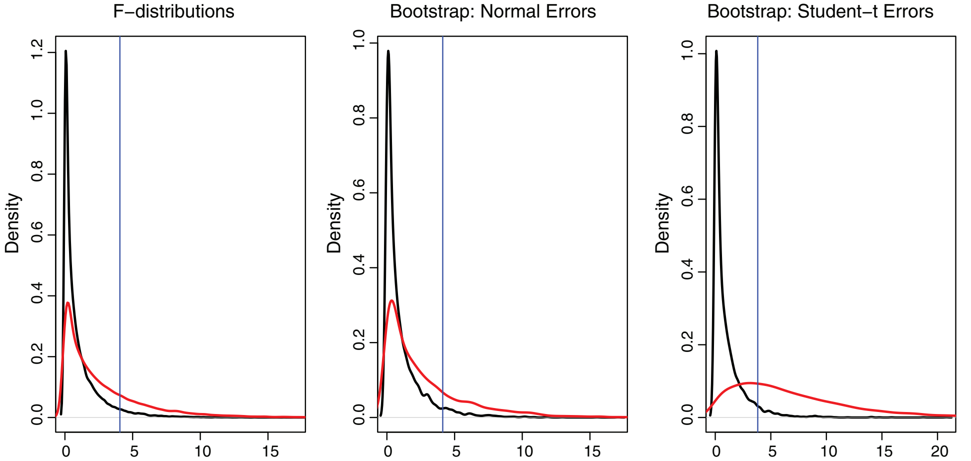
Note: H_0: $\gamma_1 = \ldots = \gamma_5 = 0$ where $\gamma' s$ are regression slope coefficients. The vertical line indicates the critical value of the F-test at $\alpha = 0.05$ (approximately 2.22). In each panel, the tall curve on the left plots the density under H_0 and the one on the right plots the density under H_1. The dark shaded area under the density of H_0 corresponds to the level of significance (Type I error probability) while the shaded area under the density of H_1 indicates the power of the test. T: sample size, $J = 5$, $K = 6$, $\alpha = 0.05$.

form of increased power'. The American Statistical Association also raises an alarm, stating: 'Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough …' (Wasserstein and Lazar, 2016, p. 132). McCloskey and Ziliak (1996, p. 102) also argue that when the sample size is massive, a researcher should pay attention to the trade-off between the power and size (α) of the test. Gigerenzer (2004, p. 601) argues that 'the combination of large sample size and low p-value is of little value in itself'.

Keuzenkamp and Magnus (1995, p. 17) point out that Fisher's theory of significance testing is intended for small samples, stating that 'Fisher does not discuss what the appropriate significance levels are for large samples', a point also made by Lindley (2014). Labovitz (1968) argues that sample size is one of the key factors for selecting the level of significance, along with the power or probability of Type II error of the test. Kish (1959) asserts that (at the conventional level of significance) 'in small samples, significant, that is, meaningful, results may fail to appear statistically significant. But if the sample size is large enough, the most insignificant relationships will appear statistically significant'. This phenomenon is

FIGURE 5

DENSITY FUNCTIONS OF F-STATISTIC



Note: In each panel, the tall curve on the left plots the density under H_0, and the shorter curve on the right the density under H_1. The vertical line indicates the critical value at $\alpha = 0.05$.

closely related with what Spanos (2017, p. 20) calls ‘the fallacy of acceptance’ and ‘the fallacy of rejection’, where the fallacious claims stem from ignoring the power of the test. For example, as Spanos (2017) points out, in the context of a model-specification test, all models are found to be mis-specified when the sample size is large enough; while all models are judged to be statistically adequate when the sample size is small. As such, as Engsted (2009, p. 401) argues, using the conventional level ‘mechanically and thoughtlessly in each and every application’ is meaningless.

In classical hypothesis testing, the roles that prior belief and loss function play are not widely appreciated and understood. For example, Koop and Steel (1994, p. 99) consider the lack of formal development of loss function as a serious weakness of hypothesis testing. They argue that the classical analysis has an implicitly defined loss function in setting the level of significance, in which losses are asymmetric. However, as the examples in the above section outlining key factors of hypothesis testing show, these key factors influence statistical decisions. Ziliak and McCloskey (2008, p. 8) state that ‘without a loss function, a test of statistical significance is meaningless’, arguing that hypothesis testing without considering the potential losses is not ethically and economically defensible. Gigerenzer (2004, p. 591) provides an example as to how the values of α and β are set in consideration of the losses from Type I and II errors, in the context of Neyman-Pearson decision theory. Kim and Choi (2017) examine the roles of relative loss and prior belief in unit root testing, and demonstrate that unit root testing at the conventional significance level is often implicitly associated with relative loss and prior belief, which is inconsistent with economic reasoning, resulting in misleading inferential outcomes. Startz (2014) contends that the classical method is based on implicit and unstated priors for the null and alternative hypotheses, which are often unlikely to be scientifically neutral.

By reconciling the classical and Bayesian methods of significance testing for a large number of the papers published in psychology journals, Johnson (2013) finds that p-values of 0.005 and 0.001 correspond to strong and very strong evidence against H_0, while the p-values in the neighborhood of 0.05 and 0.01 reflect only modest evidence. On this basis, Johnson (2013) recommends adoption of the 'revised standards for statistical evidence' by setting the level of significance at 0.005 or 0.001, instead of 0.05 and 0.01 (as in Figure 1 of this paper). Recently, this proposal has been further supported by a group of 72 statisticians, psychologists, economists, sociologists, and political scientists who argue that statistical significance should be redefined with a tighter threshold of 0.005 (Benjamin *et al.*, 2017). However, as Benjamin *et al.* (2017) acknowledge, the revised threshold can still be arbitrary, although it can improve research integrity and reproducibility in many research fields. They (p. 8) further note that, the appropriate threshold for statistical significance should be different for different research communities', emphasizing that, in general, other factors such as statistical power, prior beliefs, and losses from Type I and II errors should also be carefully considered.

Studies from Accounting and Finance

In the accounting literature, a number of authors have raised concerns about and criticisms of the practice of significance testing. Recent studies include Basu (2015), Dyckman and Zeff (2014, 2015), Dyckman (2016), Ohlson (2015, 2018), and Stone (2018). They cover a wide range of issues, including arbitrary selection of sample size, confusion between economic and statistical significance, lack of power analysis, lack of replication studies, the practice of 'p-hacking', the practice of what McCloskey and Ziliak (1996) call the 'sign econometrics' or 'asterisk econometrics', and near-exclusive use of 'null hypothesis statistical tests'. These authors have proposed a number of alternatives to the p-value criterion, such as confidence intervals, Bayesian inference, mandatory reporting of standardized effect size, goodness-of-fit testing, and meta-analytic approach.

In particular, several accounting researchers have proposed that the level of significance be adjusted as a function of sample size or statistical power. For example, in the context of auditing, Johnstone (1990, p. 27) argues that the level of significance adjusted as a decreasing function of sample size would lead to auditing decisions largely consistent with those of a Bayesian. Johnstone (1994) further shows that many auditing decisions based on a conventional significance level can be 'incoherent' in Bayesian terms, often leading to 'over-servicing'. As such, Johnstone (1997) argues that auditing practitioners following the classical methods would benefit from embracing Bayesian principles, which suggests adjustment of significance level. Johnstone (1994) also proposes that the decision rule (or level of significance) be set so as to minimize the expected loss, which is fundamental to a decision-theoretic approach to hypothesis testing. Basu (2015, p. 240) contends that, if accounting researchers wish to maintain the null hypothesis testing paradigm, the best practices should include the adjustment of the significance level and power analysis, which again are the elements of the decision-theoretic approach. Ohlson (2015) suggests that a *t*-statistic should exceed $\sqrt{0.004T}$ (in absolute value) to be taken seriously, that is, the level of significance should be adjusted as a decreasing function of sample size (see also Ohlson, 2018). Burgstahler (1987) has made a point that a statistically significant test result has little value if it is obtained with a low power. This is because it has a considerable effect on a researcher's posterior belief against the null hypothesis, only when the test is conducted with a high power. This point lends its support to the proposal for adjusting the level of significance, since a higher power can be achieved by adjusting the significance level higher when the power is low at a conventional level.

In finance, early studies that have suggested a lower significance level for a larger sample size include Klein and Brown (1984), Neal (1987), Shanken (1987), and Connolly (1989). Kim and Ji (2015) and Kim (2017c) report that many finance studies may have reported false positives or spurious correlations by conducting significance tests at a fixed conventional level of significance under large or massive sample sizes. They propose adjustment of significance level as a possible solution to the problem. Harvey *et al.* (2016) propose that the critical value of the *t*-test be raised to 3 for more credible statistical significance, in the context of making adjustment for multiple testing. In his presidential address to the American Finance Association, Harvey (2017) provides a comprehensive critique on the practice of *p*-hacking in finance, with a proposal of a higher threshold for statistical significance (by adjusting significance level) as a possible solution.

Despite these calls for change and criticisms, it appears that little has changed in the practice of statistical testing in accounting and finance. For example, from a survey of published papers in major accounting journals, Kim *et al.* (2018) report that the conventional levels of significance are adopted almost exclusively, without considering the key factors of hypothesis testing such as sample size. They further report that most statistically significant results obtained using the '*p*-value less than 0.05' are found not to be statistically significant under alternative criteria such as the Bayes factor and adaptive level of significance. More broadly, as the above-mentioned accounting critiques point out, the practice of 'mindless statistics' based on what Gigerenzer (2004) calls the 'null ritual' is still widespread in the current research in accounting, as well as that in finance (Kim and Ji, 2015).

OPTIMAL LEVEL OF SIGNIFICANCE

In this section, we present methods of choosing the optimal significance level in the context of linear regression based on a decision-theoretic approach. The methods are presented in the context of a linear regression model for time series or cross-section data, but it is applicable to the models with a more complicated structure, such as seemingly unrelated regression models or fixed-effects models.

To obtain the optimal significance level, we need to estimate the value of β or the power of the test, given the value of α. In this paper, we consider two alternative methods of estimating the power: one based on the assumption of normality and the other based on the bootstrap method (Efron, 1979). While the former is based on the finite-sample distribution theory, it may be too restrictive in practice where the assumption of normality fails or is unjustifiable. The latter may find more wide applications in practice where the data often show features like non-normality and heteroscedasticity. It also takes full account of sampling variability.

Power Calculation Under Normality
Consider the linear regression model

$$y = X\gamma + u, \tag{1}$$

where X is a $(T \times K)$ matrix of fixed regressors and $\gamma = (\gamma_1, \ldots, \gamma_K)'$ is the corresponding $(K \times 1)$ vector of unknown parameters. Assume that $(T \times 1)$ vector $u \overset{d}{=} N(0, \sigma^2 I)$ and that u and X are independent.

We test for the null hypothesis of the form

$$H_0 : R\gamma = r, \tag{2}$$

against

$$H_1 : R\gamma \neq r, \tag{3}$$

where R ia a $(J \times K)$ matrix of full row rank and r is a $(J \times 1)$ vector of known elements. Denoting the OLS estimator of γ and the residual vector as $\hat{\gamma}$ and $\hat{u}$, respectively, the F-test statistic for H_0 is written as

$$F = (R\hat{\gamma} - r)' \left[s^2 R (X'X)^{-1} R' \right]^{-1} (R\hat{\gamma} - r)/J, \tag{4}$$

where $s^2 = \hat{u}'\hat{u}/(T-K)$. The F-statistic (4) can also be written as

$$F = \frac{\left(R_1^2 - R_0^2\right)/J}{\left(1 - R_1^2\right)/(T-K)},$$

where R_j^2 represents the coefficient of determination under H_j (j = 0, 1). The F-statistic follows the F-distribution with J and $T - K$ degrees of freedom, denoted as $F(J, T - K)$. While this setting is for a test concerning a linear restriction in a linear regression model, it can also be applied to a range of model diagnostics whose test statistics are obtained from a test of a linear restriction from auxiliary regressions. Examples include the Lagrange multiplier tests for autocorrelation or heteroscedasticity such as the Breusch-Pagan test and Breusch-Godfrey test.

Let H_A denote a hypothesis at a particular point of interest under H_1. That is,

$$H_A : R\gamma = s \ (r \neq s),$$

we have

$$\begin{aligned} R\hat{\gamma} - r &= R\hat{\gamma} - s + (s - r) \\ &\stackrel{d}{=} N\left(\mu, \sigma^2 R(X'X)^{-1}R'\right), \ (\mu = s - r), \end{aligned}$$

which gives

$$(R\gamma - r)'\left[\sigma^2 R(X'X)^{-1}R'\right]^{-1}(R\gamma - r) \stackrel{d}{=} \chi^2_J(\lambda),$$

where $\chi^2_J(\lambda)$ denotes the non-central chi-square distribution with degrees of freedom J and the non-centrality parameter λ. Note that

$$\lambda = \mu'\left[\sigma^2 R(X'X)^{-1}R'\right]^{-1}\mu. \tag{5}$$

Since $\frac{(T-K)s^2}{\sigma^2} \stackrel{d}{=} \chi^2_{n-K}$ under the alternative hypothesis, we have

$$F \stackrel{d}{=} F(J, T-K; \lambda),$$

where $F(J,\ T-K;\lambda)$ denotes the non-central F-distribution with degrees of freedom $(J,\ T-K)$ and the non-centrality parameter λ as given in (5). Note that the non-centrality parameter can also be written as

$$\lambda = T\frac{R^2_{p1} - R^2_{p0}}{1 - R^2_{p1}}, \tag{6}$$

where R^2_{pj} denotes the population coefficient of determination under H_j, following from Peracchi (2001, Theorem 9.2).

The pdf of F is given as

$$\begin{aligned} pdf_F(f) = e^{-\lambda/2}{}_1F_1\left(\frac{1}{2}(J+n-K); \frac{1}{2}J; \frac{\frac{1}{2}\frac{J}{n-K}\lambda f}{1 + \frac{J}{n-K}f}\right) \\ \times \frac{\Gamma\left[\frac{1}{2}(J+n-K)\right]}{\Gamma\left(\frac{1}{2}J\right)\Gamma\left(\frac{1}{2}n-K\right)} \cdot \frac{f^{J/2-1}\left(\frac{J}{n-K}\right)^{J/2}}{\left(1 + \frac{J}{n-K}f\right)^{(J+n-K)/2}}, \ (f > 0), \end{aligned} \tag{7}$$

where

$$ {}_1F_1(p;q;r) = \sum_{k=0}^{\infty} \frac{(p)_k}{(q)_k} \frac{r^k}{k!} $$

and

$$ (x)_k = x(x+1)\cdots(x+k-1). $$

Using the pdf of F, the probability of Type II error of the test is calculated as

$$ \beta(\alpha,\lambda) = \int_0^{C_\alpha} pdf_F(f)df, $$

where C_α is the critical value of the test at the α significance level.

In practical applications, the value of λ is unknown and should be estimated. One may use the sample estimator, which can be written as

$$ \hat{\lambda} = \hat{\mu}' \left[s^2 R(X'X)^{-1} R' \right]^{-1} \hat{\mu} = T \frac{R_1^2 - R_0^2}{1 - R_1^2}, \tag{8} $$

where $\hat{\mu} = R\hat{\gamma} - r$. Using (8) and the pdf of F-distribution given in (7), an estimate of the Type II error probability or power can be obtained. Since $\hat{\mu}$ is consistent for μ as T goes to infinity, this approach has an asymptotic justification. In the section below discussing the decision-theoretic approach, alternative ways of estimating and specifying the λ values are explored.

Power and Statistical Significance

In this section, we present an example of power calculation to discuss its sensitivity to sample size and implications for statistical significance. Figure 4 presents the density functions of the F-test statistic for H_0 that all slope coefficients are jointly equal to 0, given the values of α = 0.05, J = 5, and K = 6, when $R_{p1}^2 = 0.01$ and $R_{p0}^2 = 0$. The tall curve on the left is the density under H_0 and the one on the right is that under H_1. The power is the shaded area under the density under H_1, while α = 0.05 is indicated by the dark area under the density under H_0. When T = 2000, the power is 0.95 at α = 0.05, representing an ideal case where the test is conducted with sufficiently high power with a balance between α and β. As the sample size increases to 5000, the power attains the value of 1. The plot illustrates that, although the model explains only 1% of the total variation of the dependent variable, an extreme power can be obtained when the sample size is large enough. This is because the sample size T is a dominant factor for the position of the distribution of the F-test statistic under H_1, as clear from the non-centrality parameter given in (6). One key point to note is that the F-statistic is generated from the density under H_0 only when H_0 is exactly true with $R^2_{p1} = 0$. As long as $R^2_{p1} \geq 0.01$, H_0 is always rejected at the 5% level of significance.

It is clear from Figure 4 that H_0 is almost always rejected in repeated sampling when T = 5000. Note that this rejection may indicate the model's statistical significance, but not necessarily its economic significance (or substantive importance). This is because the X variables jointly explain only 1% of the total variation of Y. As also illustrated in Figure 1, this represents the case where an economically negligible deviation from H_0 appears to be statistically significant. This occurs since H_0 is always violated in practice, even when H_0 is practically true. That is, it is unrealistic in practice that all of the slope coefficients literally take the exact numerical values of zero ($R^2_{p1} = 0$), although they may be small and economically negligible (see Hodges and Lehmann, 1954; De Long and Lang, 1992, p. 1269). For example, Grossman and Stiglitz (1980) show that a perfectly efficient market (if Y is a stock return and X's are predictors) is impossible because if prices fully reflect all available information, traders would not have any incentive to acquire costly information. As such, the main research question in empirical testing of market efficiency should be whether the observed deviation from market efficiency is economically important. This means, in practice, that the value of λ is always positive with $R^2_{p1} > 0$. As a result, rejection of H_0 will surely occur with increasing sample size, even when the value of R^2_{p1} is negligible. The key question in empirical research is whether this non-zero deviation from H_0 is large enough to be economically meaningful. McCloskey and Ziliak (1996) provide a similar example in the context of the purchasing power parity.

Power Calculation Based on Bootstrapping

The method of estimating the power presented in the previous sub-section is valid under the assumption of normality. When the error term does not follow a normal distribution, the method of power estimation given above may be invalid. In this case, the bootstrap (Efron, 1979) can provide a useful alternative as a means of approximating the unknown sampling distribution in small samples, which is conducted by repeated re-sampling of observed data. In the present context, the bootstrap is employed to approximate the distributions of a F-statistic under H_0 and H_1. We can obtain the critical values of the test as the quantiles from the bootstrap distribution under H_0; and an estimate of the power from the bootstrap distribution under H_1.

Let $\hat{\gamma}_0$ denote the restricted estimator for γ under H_0: $R\gamma = r$ and $\hat{u}_0$ the associated residual vector. Similarly, let $\hat{\gamma}_1$ denote the unrestricted estimator for γ under H_1: $R\gamma \neq r$ and $\hat{u}_1$ the associated with residual vector. We generate a set of artificial data set as

$$y^*_0 = X\hat{\gamma}_0 + \hat{u}^*_0 \tag{9}$$

where $\hat{u}_0^*$ is a $(T\times 1)$ vector whose elements are random re-samples with replacement from the elements of $\hat{u}_0$. Using $\left(y_0^*, X\right)$, estimate the model under H_0 and calculate the F -statistic denoted as F_0^*. Repeat this process sufficiently many times, say B, to obtain the bootstrap distribution of the F-statistic under H_0, denoted as $\left\{F_0^*(j)\right\}_{j=1}^{B}$.

In a similar way, generate $\left(y_1^*, X\right)$ as

$$y_1^* = X\hat{\gamma}_1 + \hat{u}_1^* \tag{10}$$

where $\hat{u}_1^*$ is a $(T\times 1)$ vector whose elements are random re-samples with replacement from the elements of $\hat{u}_1$. Using $\left(y_1^*, X\right)$, estimate the model under H_1; and calculate the F -statistic denoted as F_1^*. Repeat this B times, to obtain the bootstrap distribution of the F-statistic under H_1, denoted as $\left\{F_1^*(j)\right\}_{j=1}^{B}$.

We use these bootstrap distributions $\left\{F_i^*(j)\right\}_{j=1}^{B}$ to approximate the sampling distributions of the F-statistic under H_i. Namely, the α-level bootstrap critical value C_α^* for the test is obtained as the $(1-\alpha)$ percentile from $\left\{F_0^*(j)\right\}_{j=1}^{B}$; while the value of β can be estimated as the proportion of $\left\{F_1^*(j)\right\}_{j=1}^{B}$ less than C_α^*.

As an illustration, we consider a multiple regression of the form

$$y = \gamma_0 + \gamma_1 X_1 + \ldots + \gamma_5 X_5 + u.$$

Setting $(\gamma_0, \gamma_1, \ldots, \gamma_5) = (0, 0.1, 1, 1, 1, 1)$ and $u \sim NID(0, 0.25)$, we generate y with $T = 50$ with all X variables generated from $NID(0, 1)$. We test for H_0: $\gamma_1 = 0$ against H_1: $\gamma_1 \neq 0$. Following (9), an artificial data set under H_0 is generated as

$$y_0^* = \hat{\gamma}_{00} + \hat{\gamma}_{02} X_2 + \ldots + \hat{\gamma}_{05} X_5 + \hat{u}_0^*,$$

where $(\hat{\gamma}_{00}, 0, \hat{\gamma}_{02}, \ldots, \hat{\gamma}_{05})$ denotes the parameter estimators under H_0 and $\hat{u}_0^*$ random re-sample with replacement from the residual vector $\hat{u}_0$. Calculate the F-statistic from repeated generations of $\left(y_0^*, X\right)$ to obtain $\left\{F_0^*(j)\right\}_{j=1}^{B}$. Following (10), an artificial data set under H_1 is generated as

$$y_1^* = \hat{\gamma}_{10} + \hat{\gamma}_{11} X_1 + \ldots + \hat{\gamma}_{15} X_5 + \hat{u}_1^*,$$

where $(\hat{\gamma}_{10}, \hat{\gamma}_{11}, \ldots, \hat{\gamma}_{15})$ denotes the parameter estimators under H_1 and $\hat{u}_1^*$ random re-sample with replacement from the residual vector $\hat{u}_1$. Calculate the F-statistic from repeated generations of $\left(y_1^*, X\right)$ to obtain $\left\{F_1^*(j)\right\}_{j=1}^{B}$.

Figure 5 plots the densities from a set of realizations of Y and X's with $R_0^2 = 0.9537$ and $R_1^2 = 0.9486$. The first figure plots $F(1,\ T-K)$ and $F\left(1, T-K; \hat{\lambda}\right)$; and the second and third show their bootstrap counterparts $\left\{F_0^*(j)\right\}_{j=1}^{B}$ and $\left\{F_1^*(j)\right\}_{j=1}^{B}$, under normal and Student-t with 5 degrees of freedom, respectively. The densities look similar in the first and second cases, which means that the bootstrap provides a good approximation of the exact densities under normality. The vertical line represents the 5% critical value, which is 4.06 for the exact distributions and 4.12 for the bootstrap distributions under normality. The power (1-β)associated with the exact distributions are 0.2286, while that of bootstrap is 0.2232. In the third figure, the distributions are largely different from those under normal errors. The 5% critical value is 3.82 and the power takes a much higher value 0.6528. Note that, under non-normal errors frequently encountered in practice, the bootstrap may provide more accurate power estimation in small samples. In addition, with the bootstrap method, the test does not suffer from size-distortion, because the estimated level of significance is always equal to the chosen level of significance.

When the error term is heteroscedastic, the above bootstrap procedure based on residual resampling is mis-specified. In this case, one can use the wild bootstrap (Liu, 1988; Mammen, 1993; Davidson and Flachaire, 2008), where resampling is conducted by scaling the residuals with a random variable with the mean 0 and variance equal to 1. That is, $\hat{u}_{it}^* = \eta_t \hat{u}_{it}$ where i = 0, 1. In this paper, we choose Mammen's (1993) two-point distribution, where $\eta_t = -0.5\left(\sqrt{5}-1\right)$ with the probability $\left(\sqrt{5}+1\right)/2\sqrt{5}$; $\eta_t = 0.5\left(\sqrt{5}-1\right)$ with the probability $\left(\sqrt{5}-1\right)/2\sqrt{5}$. This distribution is well-known to have a higher order refinement.

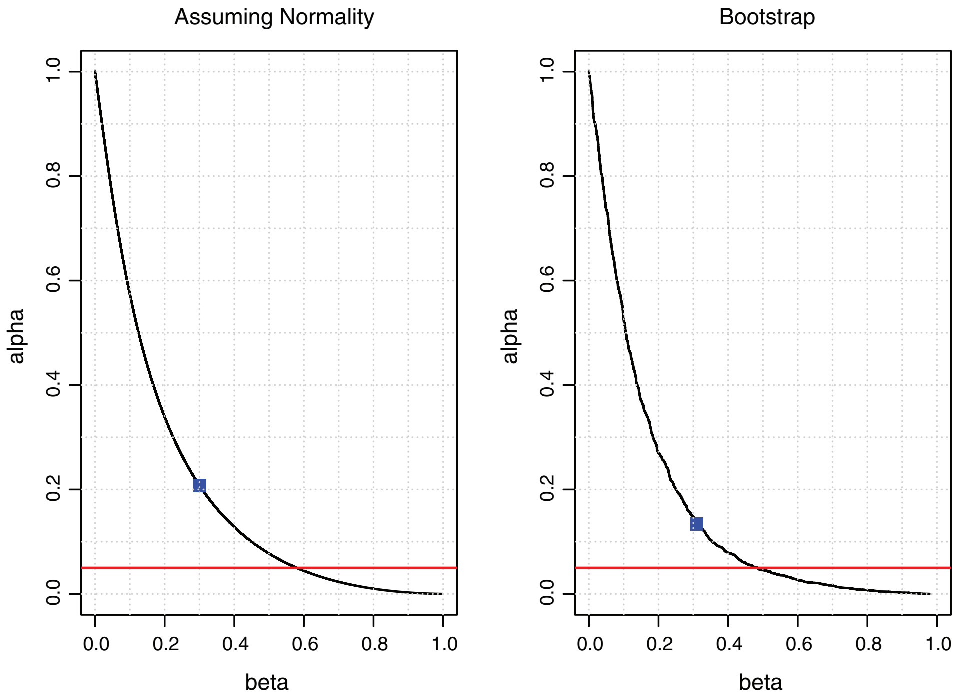
Note that, in estimating $\left\{F_1^*(j)\right\}_{j=1}^{B}$, we use unrestricted estimator $\hat{\gamma}_1$, which is equivalent to using the sample estimator $\hat{\lambda}$ for λ. We note that other choices can be made in the context of bootstrapping, which will be discussed in the next section.

A Decision-theoretic Approach

It is well-known that there is a trade-off between the two error probabilities α and β. A higher value of α is associated with a lower value of β, or *vice versa*. As such, one can consider a range of values of $\alpha \in (0, 1)$ and the corresponding β values, given the values of T, K, J, and λ. By plotting all possible combinations of (α, β), we obtain what is called the line of enlightened judgement (Leamer, 1978). Figure 3 presents three lines of judgement corresponding to the (α, β) values in Table 2 when the sample size is 10, 50, and 100. The horizontal line represents the case where the level of significance is fixed at 0.05. Given these infinitely many combinations of (α, β), one may naturally ask how we should choose the most desirable combination. Obviously, there is no reason to believe that the points associated with α = 0.05 represent such a choice.

FIGURE 6

OPTIMAL LEVEL OF SIGNIFICANCE: TEST FOR CONSTANT RETURNS TO SCALE



Note: The horizontal line in each panel indicates H_0 and the square dot indicates the point of the optimal level at $\alpha^* = 0.207$ and $\alpha^* = 0.134$.

With the decision-theoretic approach, the optimal choice is made by minimizing the expected loss from hypothesis testing. Namely, choose (α, β) such that the expected loss is minimized, which can be written as $PL_1\alpha + (1 - P)L_2\beta$. Without loss of generality, let $L_1 = 1$, then $k = L_2$. Noting that the value of β is driven by α and the non-centrality of the distribution under H_1, we can express the expected loss as a function of α, given the values of P, k and λ. That is,

$$EL(\alpha;P,k,\lambda) = P\alpha + (1-P)k\beta(\alpha;\lambda). \tag{11}$$

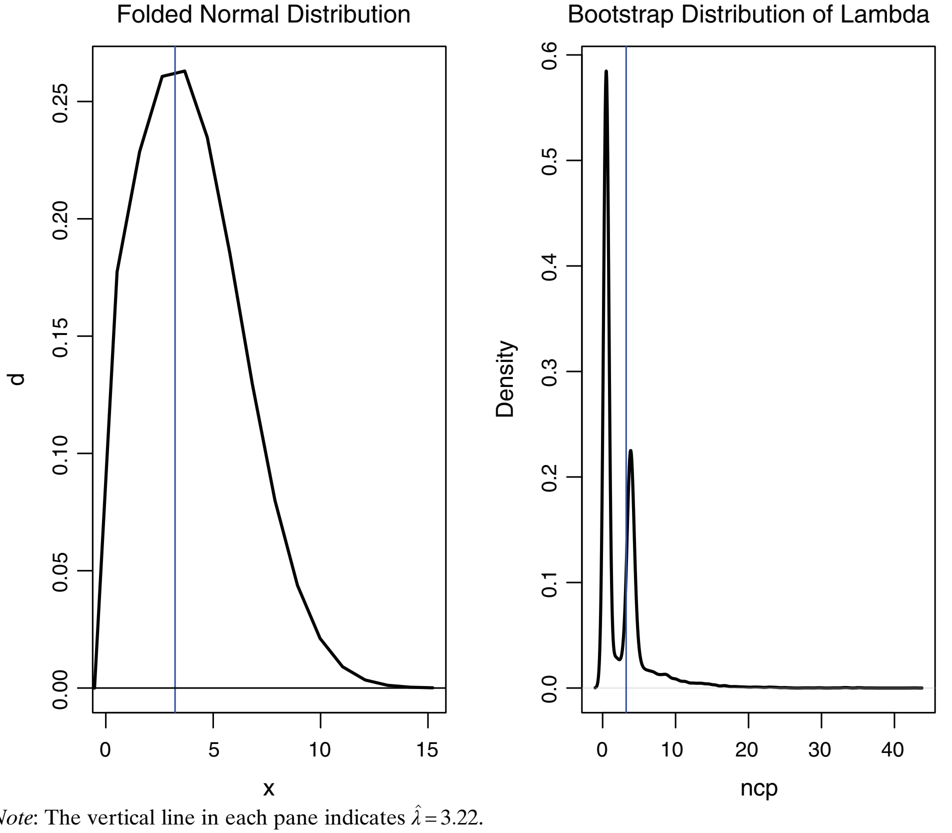
From the first-order condition of minimizing (11) with respect to α, we have

$$\frac{d\alpha}{d\beta} = -\frac{(1-P)k}{P}, \tag{12}$$

which represents the slope of the line of enlightened judgement at the point of minimization. The optimal value of α that satisfies this condition is denoted as α^*;

FIGURE 7

DISTRIBUTIONS FOR THE NON-CENTRALITY PARAMETER



Note: The vertical line in each pane indicates $\hat{\lambda} = 3.22$.

and $\beta^* \equiv \beta(\alpha^*;\lambda)$. In practice, the optimal level α^* can be calculated in the following steps:

Step 1 Choose a grid of $\alpha \in (0, 1)$ values
Step 2 For each value of α, calculate $\beta(\alpha;\lambda)$ and expected loss (11)
Step 3 Find the value of α^* that minimizes the expected loss

If the researcher believes that H_0 and H_1 are equally likely ($P = 0.5$) and the losses from Type I and II errors are equal ($k = 1$), then the expected loss (11) is minimized when the slope of the line (12) is −1. In Figure 3, this gives $\alpha^* = \beta^*$, indicated by the 45 degree line. This is what Arrow (1960) called the equal-probability test where the two error probabilities are set equal. These values correspond to the (α, β) combinations in the second panel of Table 2 for T = 10, 50, 100. For example, when $T = 10$, $\alpha^* = 0.35$; and when $T = 50$, $\alpha^* = 0.19$. Note that, in many applications, it is often the case that the researcher is impartial between H_0 and H_1, in terms of prior belief for H_0 and losses resulting from incorrect decisions. In this case, it is reasonable to set P = 0.5 and k = 1 when the expected loss (11) is minimized.

If the researcher strongly believes that H_0 is true ($P > 0.5$) or that his or her loss from Type I error is substantially higher than that from Type II error ($k < 1$), then the slope given in (12) becomes much smaller than 1 (in absolute value) at the point of minimization, and we have $\alpha^* < \beta^*$, reflecting the researcher's attitude favouring H_0. This researcher is similar to a jury who strongly believes in the defendant's innocence, in the example given in the previous section. Conversely, if the value of P is low or the value of k is high (a higher prior belief for H_1 or a higher loss from Type II error), then the slope will be much higher than 1 (in absolute value), which gives $\alpha^* > \beta^*$, reflecting the researcher's attitude favouring H_1. This researcher is similar to the doctor who is deeply cautious about misdiagnosing a pregnancy in the example given in the previous section.

It is also clear from Figure 3 that a conventional value of α such as 0.05 cannot be optimal in general. It can only be optimal only under special values of P and k, which are unspecified and implicit in hypothesis testing. In other words, by setting α = 0.05, the researcher implicitly favours either H_0 or H_1, unwittingly giving a much heavier weight to it. As Startz (2014) and Kim and Choi (2017) point out, this arbitrariness can often be the cause of a wrong decision, especially when the researcher's implicit attitude is not consistent with the context of hypothesis testing (e.g., the jury adopts the 0.05 significance level to meet the burden of proof 'beyond reasonable doubt').

Weighted Optimal Significance Level

In calculating the optimal level of significance as detailed in the previous subsection, a choice should be made for the value under H_1 under which the power is calculated. If this value is too close to the value under H_0, then the power will be too low; while this value is too far away from it, the power can be too high or even equal to 1. Where possible, the choice should be made based on a careful economic analysis or reasoning as an economically meaningful deviation from H_0 (see Ziliak and McClosekey, 2004). However, sometimes, the choice can be subjective or may be difficult to make. In the previous subsection, we have used the sample estimators $\hat{\lambda}$ given in (8) under H_1. While this sample estimator may provide a representative value for λ, other values may also be considered. Since it is a point estimator for λ, a more sensible approach is to calculate the optimal levels of significance over a grid of possible λ values. These optimal levels can be weighted to produce a weighted optimal level of significance.

In specifying the distribution for λ, we propose the use of the folded-normal distribution whose density function is given as

$$w(b) = \frac{1}{\delta\sqrt{2\pi}} e^{-\frac{(b-\hat{\lambda})^2}{2\delta^2}} + \frac{1}{\delta\sqrt{2\pi}} e^{-\frac{(b+\hat{\lambda})^2}{2\delta^2}},\ s \geq 0$$

In using this distribution, the choice for λ and δ values should be made, which represent the centre and spread of the distribution. For λ, one may use the sample estimator $\hat{\lambda}$; choose a value of λ based on economic reasoning; or try a number of alternative values around $\hat{\lambda}$. Similarly, the value of δ may be chosen based on economic reasoning or the researcher may try a number of alternative values. The weighted optimal level is calculated as

$$\alpha_w^* = \int_0^\infty \alpha^*(b)w(b)db.$$

Alternatively, the sampling distribution of $\hat{\lambda}$ can be approximated by the bootstrap. That is, based on the estimated model $y = X\hat{\gamma}_1 + \hat{u}_1$ under H_1, generate $\left(y_1^*, X\right)$ by $y_1^* = X\hat{\gamma}_1 + \hat{u}_1^*$ where $\hat{u}_1^*$ is a vector whose elements are random re-samples with replacement from $\hat{u}_1$. Estimate the model under H_1 using $\left(y_1^*, X\right)$; and estimate the unrestricted regression parameters and non-centrality parameter, denoted as $\left(\hat{\lambda}^*; \hat{\gamma}_1^*\right)$. Repeat this B times ro obtain the bootstrap distribution of $\left\{\hat{\lambda}^*(j); \hat{\gamma}_1^*(j)\right\}_{j=1}^B$. A grid of λ values are drawn from this distribution, from which we calculate the optimal levels of significance. Let $\left(\hat{\lambda}^*(\tau); \hat{\gamma}_1^*(\tau)\right)$ represents such a value of λ and parameter estimates from this distribution. Then the bootstrap distribution $\left\{F_1^*(j)\right\}_{j=1}^B$ under H_1 is generated using the corresponding $\hat{\gamma}_1^*(\tau)$ values and the associated residuals. The optimal significance levels obtained in this way can be weighted using the density of the bootstrap distribution $\left\{\hat{\lambda}^*(j); \hat{\gamma}_1^*(j)\right\}_{j=1}^B$. These methods will be demonstrated in the next section where empirical applications are presented.

Extreme Power (Under Massive Sample)

As discussed in the section on power and statistical significance above, the power of the test can sometimes be equal to one or extremely close to one, especially under a massive sample size. In this case, the decision-theoretic approach presented above may not be useful, since the optimal level of significance in this case is also extremely close to zero or equal to 0. As an alternative, a researcher can use a simple formula proposed by Perez and Perrichi (2014). However, when the power is extreme, this means that the value of α should be set at an infinitesimal value. This may not represent a sensible research design since an extreme power in this case means that any negligible deviation from the null hypothesis is strongly rejected.

We note that the power can be extreme under the following circumstances:

1. the value being tested under H_0 is unreasonably far away from the true population value;

2. the sampling error is extremely small; or
3. the sample size is extremely large or massive.

In the first case, an economically more sensible value closer to the population value may be tested under H_0; in the second case, the researcher will have a clear idea on the value of population parameter, as long as sampling is conducted in a random and unbiased way.

The third case is frequently encountered in modern applications in the business disciplines with the availability of large or massive data sets. The use of a conventional level of significance is particularly problematic in this case, since the two error probabilities are severely unbalanced, as discussed previously. The consequence is that a negligible deviation from H_0 is almost always rejected at a conventional level of significance, even if the effect size estimate is economically negligible or the model's explanatory power (e.g., R^2) is close to 0 (see Figures 1 and 4 and the accompanying discussions). This is a case of what Spanos (2017, p. 20) calls the 'fallacy of rejection' where the evidence against H_0 is misinterpreted as evidence for a particular H_1 when the test in question has high power to detect substantively minor discrepancies. As a result, an economically negligible effect may spuriously be judged to be statistically significant: see, for example, Kim (2017c) for critical evaluations of the weather effect on stock return. This also represents a clear conflict between model estimation and statistical inference. From their survey of the papers published in top accounting and finance journals, Kim and Ji (2015) and Kim *et al.* (2018) report that a high proportion of the studies in their survey employ large or massive sample sizes. However, they almost exclusively adopt a conventional level of significance, which may lead to a large number of false discoveries.

When the researcher is confronted with the power extremely close to or equal to one, due to the use of large or massive sample size, we have the following recommendations.

First, the focus of the research in this case should be the economic plausibility of the hypothesis (Harvey, 2017); effect size estimate and its economic implications (Ziliak and MacCloskey, 2008); and the model's explanatory power typically measured by R^2 (Kim, 2017c). This is because statistical inference at a conventional level of significance is meaningless in this case, since the test rejects H_0 by construction as illustrated in Figure 4. If the economic impact or explanatory power is found to be negligible, the test outcome indicating statistical significance should not be taken seriously because a large value of the test statistic (in absolute value) or a small p-value is obtained only due to the effect of a large sample size. As Black (1993, p.75) cautions, 'Lack of theory is a tipoff: watch out for data mining!'

Second, a Bayesian method should be considered as a credible alternative. For example, Kim and Ji (2015) propose the use of the Bayes factor proposed by Zellner and Siow (1980); while Harvey (2017) suggests using the Bayesianized p-value. These alternatives can be calculated as a simple transformation of the F-test statistic or t-statistic. Note that, when the sample size is large or massive, a p-value less than 0.05 does not necessarily mean that there is strong evidence against H_0, as demonstrated by Johnstone (1990) and Johnstone and Lindley (1995). As an example, suppose the observed F-statistic is 3 in Figure 4 when $T = 5000$. The p-value $Prob(F > 3|H_0) = 0.01$ (from the density under H_0), leading to rejection of the null at 0.05 level. However, $Prob(F < 3|H_1) = 0.0001$ (from the density under H_1), much smaller than the p-value. This means that the F-test outcome is incompatible with H_0 with an odd of 0.01, while it is incompatible with H_1 with a substantially lower odd of 0.0001. That is, the p-value of 0.01 in fact indicates strong evidence for H_0, not against it, in comparison with the evidence for H_1. The ratio of the two probabilities is called the Bayes factor, a measure of evidence widely used in the Bayesian method. Keuzenkamp and Magnus (1995, p. 20) also point out that, with such large samples, ‘a case can be made for using a Bayesian procedure’. On this point, Gigerenzer (2004) argues that students and researchers should be taught to use a statistical toolbox (a range of alternatives including the Bayesian) for sound statistical research, rather than being instructed to use a single hammer (null hypothesis testing).

Third, instead of using all available data points, a range of sub-sample analyses over time or over cross-section can be conducted, also employing a battery of descriptive and exploratory methods. Ohlson (2015, 2018) also argues that accounting researchers can benefit from ‘sample splitting’ or sub-sample analyses in their research. Note that a large data set may suffer from selection bias, structural breaks, cross-sectional variations, and noises, which cannot be captured fully by the analysis using the whole sample (Kaplan *et al.*, 2014). For example, Hand (2016, p. 631) cautions that big (data) does not necessarily mean accurate or comprehensive, emphasizing the importance of the underlying theory and the analysis of small data sets. Harford (2014) also discusses the problems associated with the use of big data, with an example of misleading prediction based on a large number of observations. Fiebig (2017) makes a point that, while big data opens up opportunities to improve empirical research, it is not synonymous with better or more informative data.

EMPIRICAL APPLICATIONS

In this section, we present three empirical applications of the optimal significance level: they include a textbook level example of testing for the returns to scale of a production function; testing for empirical validity of asset-pricing models following Fama and French (1993, 2015); and an application to auditing research taken from Johnstone (1990) and Dodd *et al.* (1984). We initially assume an impartial researcher with $P = 0.5$ and $k = 1$, but we also examine the cases where this assumption is relaxed. As a further example, we re-examine a numerical example taken from Gelman and Stern (2006) at the optimal level of significance. The R codes used in this section are presented in the Appendix.

Testing for Constant Returns to Scale

We consider the log-linear Cobb-Douglas production function estimation discussed in Gujarati (2015). The data contains 51 cross-sectional observations of output, capital, and labour in natural logarithm, covering 50 states and Washington DC of the US for the year 2005. The regression equation is written as

$$\log(Q) = \gamma_0 + \gamma_1 \log(L) + \gamma_2 \log(K) + u, \tag{13}$$

with OLS estimates $\hat{\gamma}_1 = 0.468$ and $\hat{\gamma}_2 = 0.521$. The hypothesis of the constant returns to scale can be tested using H_0: $\gamma_1 + \gamma_2 = 1$. The F-test statistic of 0.141 with the p-value of 0.709, which indicates that H_0 cannot be rejected at any reasonable level of significance. This is not surprising given that $\hat{\gamma}_1 + \hat{\gamma}_2 = 0.989$ is practically equal to 1.

Suppose one tests for $H_0^*: \gamma_1 + \gamma_2 = 0.94$, arguing that the returns to scale is less than 1. The F-test statistic is 3.225 with the p-value of 0.079, which indicates that, at the 5% significance level, H_0^* cannot be rejected. This means that $\hat{\gamma}_1 + \hat{\gamma}_2 = 0.989$ is statistically indistinguishable from 0.94 at a conventional level of significance, casting doubt on the result in favour of the constant returns to scale. We note that the power of the test for H_0^* at the 5% level of significance is 0.421 (evaluated at $\hat{\lambda} = 3.22$), which indicates a reasonably high chance of Type II error. We calculate the optimal level of significance, which is found to be 0.207 under the assumption of normality and 0.134 based on the bootstrap (3000 bootstrap replications), as Figure 6 shows. This means that, we reject H_0^* at the optimal level of significance, further strengthening the earlier finding that the production function satisfies the constant returns to scale.

To calculate the weighted optimal level of significance, we have considered a number of λ values following the folded-normal distribution. First, we try the distribution with $\hat{\lambda} = 0.14$ and $\delta = 3$; and the one with $\hat{\lambda} = 3.22$ and $\delta = 3$. These $\hat{\lambda}$ values are sample estimates of the non-centrality parameters for H_0: $\gamma_1 + \gamma_2 = 1$ and $H_0^*: \gamma_1 + \gamma_2 = 0.94$, respectively; with the value of δ set at 3 to ensure sufficient spread in the distribution of λ. The folded normal distribution for the latter case is plotted in Figure 7. The weighted optimal level is 0.238 when $\hat{\lambda} = 0.14$ and 0.195 for $\hat{\lambda} = 3.22$. Since the p-value of the test for H_0 is 0.709, the hypothesis again cannot be rejected. Similarly, since the p-value of the test for H_0^* is 0.079, the hypothesis is rejected, which is consistent with the earlier findings based on the un-weighted optimal level of significance.

Figure 7 also presents the bootstrap distribution of $\hat{\lambda}$ for $H_0^*: \gamma_1 + \gamma_2 = 0.94$. From this distribution, we have selected nine percentiles (from the 10th to the 90th) and use them to calculate the values of α* based on bootstrap power calculation. With the weights obtained from the density estimates of the bootstrap distributions, we find that the weighted α* value is 0.248, leading to acceptance of H_0 and rejection of H_0^*. By employing the optimal level of significance, we have found clear evidence for H_0: $\gamma_1 + \gamma_2 = 1$, while that against $H_0^*: \gamma_1 + \gamma_2 = 0.94$. If a conventional level such as 0.05 was adopted, the evidence could have been ambiguous since H_0^* cannot be rejected.

To test for the presence of heteroscedasticity in the error term of (13), we employ the Breusch-Pagan test, which can be conducted using an auxiliary regression of the form

$$e^2 = \delta_0 + \delta_1 \log(L) + \delta_2 \log(K) + v,$$

as the F-test for H_0: $\delta_1 = \delta_2 = 0$, where e represents the residual vector from the regression (13). The F-statistic is 2.53 with the p-value of 0.09. At a conventional level of significance, the decision is marginal and indecisive, depending on whether the 0.05 or 0.10 level is chosen. In this case, the researcher may choose a level arbitrarily; or the choice may depend on the researcher's desire to accept or reject H_0, as Keuzenkamp and Magnus (1995, p.20) observe from their survey.

The optimal level of significance is found to be around 0.20 (evaluated at $\hat{\lambda} = 5.06, P = 0.5, k = 1$), based on a normality assumption as well as using the bootstrap. The weighted optimal levels are found to take similar values. Hence, at the optimal level of significance, a decisive inferential outcome is delivered as H_0 is clearly rejected. This is also consistent with our descriptive analyses (plots that are not reported), which show mild linear relationships between e^2 and the explanatory variables. As a further check, we also conduct the Glejser test by replacing e^2 with $|e|$ in the auxiliary regression. The F-statistic for H_0 is 5.38 with the p-value of 0.008. The optimal levels calculated are around 0.10, rejecting H_0 and further strengthening the evidence for the presence of heteroscedasticity. Note that, at the optimal levels, the two alternative tests are in clear agreement.

Empirical Validity of an Asset-pricing Model

In this section, we apply the method of choosing the optimal level of significance to testing for the empirical validity of an asset-pricing model. An asset-pricing model explains the variation of asset return as a function of a range of risk factors. The most fundamental is the capital asset pricing model (CAPM), which stipulates that an asset (excess) return is a linear function of market (excess) return. The slope coefficient (often called beta) measures the sensitivity of an asset return to the market risk. While the CAPM is theoretically motivated, it is often the case that the market risk alone cannot fully capture the variation of asset return. In response to this observation, several multi-factor models have been proposed, which augment the CAPM with a number of empirically motivated risk factors such as the size premium or value premium (see, for example, Fama and French, 1993). The most recently proposed multi-factor model is the five-factor model of Fama and French (2015), which can be written as

$$R_{it} - R_{ft} = a_i + b_i\left(R_{Mt} - R_{ft}\right) + s_i SMB_t + h_i HML_t + r_i RMW_t + c_i CMA_t + e_{it},$$

where R_{it} is the return on an asset or portfolio i at time t ($i = 1, \ldots, N; t = 1, \ldots, T$), R_{ft} is the risk-free rate, R_{Mt} is the return on a (value-weighted) market portfolio at time t, SMB_t is the return on a diversified portfolio of small stocks minus the return on a diversified portfolio of big stocks, the HML_t is the spread in returns between diversified portfolios of high book-to-market stocks and low book-to-market stocks, RMW_t is the spread in returns between diversified portfolios of stocks with robust and weak profitability, and the CMA_t is the spread in returns between diversified portfolios of low and high investment firms. Fama and French (1993, 2015) provide empirical justifications of these factors.

If these factors fully or adequately capture the variation of asset return, then the intercept terms a_i should be zero or sufficiently close to it. This intercept term a_i may be interpreted as the risk-adjusted return for asset i. On this basis, the model's empirical validity is evaluated by testing for H_0: $a_1 = \ldots = a_N = 0$. The F-test for H_0 is widely called the GRS test, proposed by Gibbons *et al.* (1989). Let $a = (a_1, \ldots, a_N)'$ be the vector of N intercept terms, and Σ be the $N \times N$ covariance matrix of error terms. The model is estimated using the ordinary least-squares: $\hat{a}$ denotes the estimator for a and $\hat{\Sigma}$ the estimator for Σ.

The F-test statistic is written as

$$F = \frac{T(T-N-K)}{N(T-K-1)} \frac{\hat{a}'\hat{\Sigma}^{-1}\hat{a}}{1+\hat{\mu}'\hat{\Omega}^{-1}\hat{\mu}},$$

where T is the sample size, $K = 5$ is the number of risk factors, $\hat{\Omega}$ is the $K \times K$ covariance matrix of risk factors, and $\hat{\mu}$ is the $K \times 1$ mean vector. Under the assumption that the error terms e's follow a multivariate normal distribution, the statistic follows the $F(N, T-N-K; \lambda)$ distribution, with the non-centrality parameter

$$\lambda = \left(\frac{T}{1+\hat{\theta}^2}\right) a'\Sigma^{-1}a = \left(\frac{T}{1+\hat{\theta}^2}\right)\left(\theta^{*2} - \theta^2\right),$$

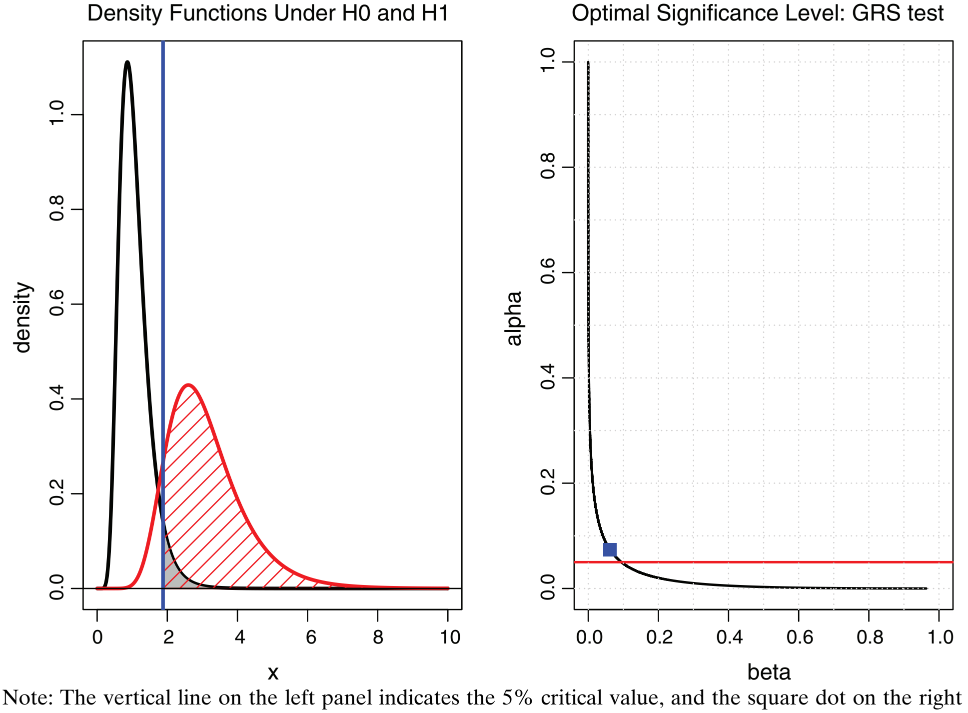
where $\hat{\theta}$ is the *ex post* maximum Sharpe ratio of K -factor portfolio, θ is the *ex ante* maximum Sharpe ratio of K -factor portfolio, and θ^* is the slope of the *ex ante* efficient frontier based on all assets. Gibbons *et al.* (1989) call θ/θ^* the proportion of the potential efficiency.

The data is available from French's data library monthly from 1963 to 2015.[a] We use 25 portfolio returns ($N = 25$) sorted by size and book-to-market ratio extensively analyzed by Fama and French (1993, 2015). We estimate the five-factor model above and examine the model's ability to capture the systematic variation of the returns by testing for H_0. In so doing, we demonstrate how the optimal level of significance can be used to make a more informed decision. In

[1] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.html

Figure 8

POWER AND THE OPTIMAL LEVEL OF SIGNIFICANCE: GRS TEST (PERIOD 1, $T = 60$, $N = 25$, $P = 0.5$, $K = 1$)



Note: The vertical line on the left panel indicates the 5% critical value, and the square dot on the right panel indicates the point of $\alpha^* = 0.074$.

calculating the optimal level, we use $\hat{\lambda}$ as the estimator for λ, which is obtained by replacing the unknowns with their estimators.

We take four sub-samples, consisting of the last 60, 90, 120, and 150 observations of the data set. Table 3 reports the GRS test results, the power values and the optimal levels of significance, with graphical illustrations of the power and the optimal significance level given in Figures 8 and 9 for Period 1 and 4. For the first two periods with sample sizes 60 and 90, the p-values of the GRS test are 0.546 and 0.199 respectively, accepting H_0 at a conventional level of significance. The mean of the absolute value of the 25 intercept estimates ($|a|$) is 0.105 for Period 1. This means that the average risk-adjusted monthly return is around 0.1%, which is not by any means economically large considering the transaction costs and market volatility. The mean of the R^2 values from 25 regressions is 0.939, indicating that the model can explain nearly 95% of the total variation of the portfolio returns, on average. Similar estimation results are evident for Period 2. The optimal levels of significance calculated are 0.074 and 0.039 respectively for Periods 1 and 2, which indicates that the decision to accept

TABLE 3

GRS TEST RESULTS: FAMA-FRENCH FIVE-FACTOR MODEL ($N = 25$, $K = 5$, $P = 0.5$, $K = 1$)

Period	T	GRS	p-value	$\lvert a\rvert$	R^2	Power	α^*
1	60	0.951	0.546	0.105	0.939	0.905	0.074
2	90	1.304	0.199	0.108	0.952	0.972	0.039
3	120	1.727	0.033	0.136	0.944	0.995	0.017
4	150	1.829	0.017	0.130	0.939	0.997	0.014

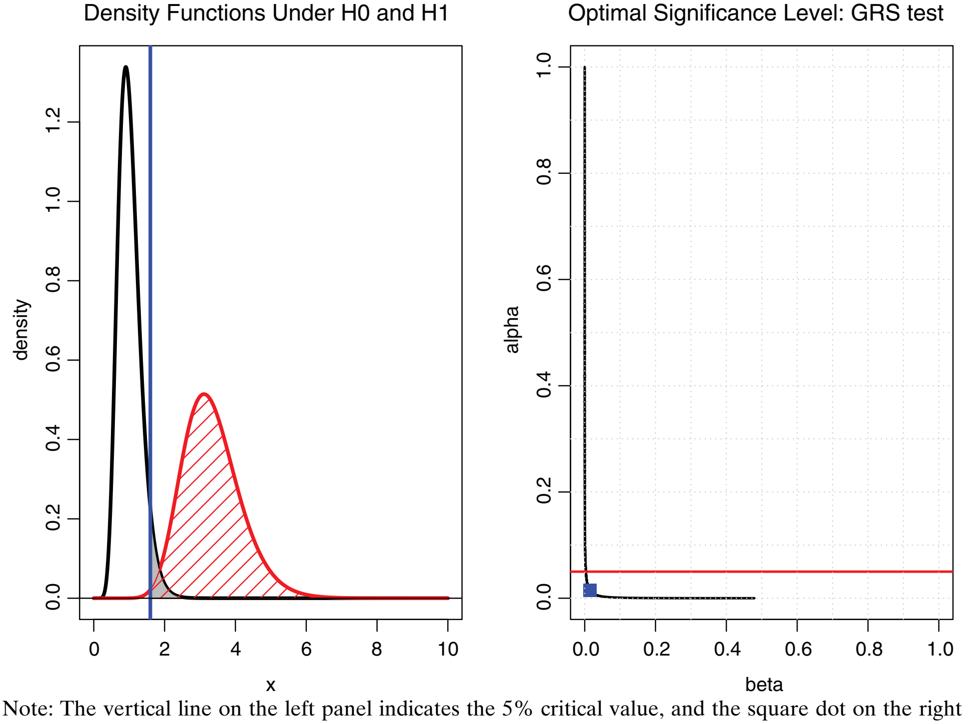
Period 1: Jan 2011 to Dec 2015; Period 2: July 2008 to Dec 2015; Period 3: Jan 2006 to Dec 2015; Period 4: July 2003 to Dec 2015; $\lvert a\rvert$: the mean of intercept estimates in absolute value; R^2: the mean of the coefficient of determination; α^*: the optimal level of significance.

H_0 remains unchanged at the optimal levels. We note that, under these moderate sample sizes, the conventional levels such as 0.05 or 0.10 appear to be close to the optimal values.

For Periods 3 and 4 where larger samples are employed, the p-values of the GRS test are less than 0.05, which leads to the rejection of H_0 at the 5% level.

FIGURE 9

POWER AND THE OPTIMAL LEVEL OF SIGNIFICANCE: GRS TEST (PERIOD 4, $T = 150$, $N = 25$, $P = 0.5$, $K = 1$)



Note: The vertical line on the left panel indicates the 5% critical value, and the square dot on the right one indicates the point of $\alpha^* = 0.014$.

TABLE 4

WEIGHTED OPTIMAL SIGNIFICANCE LEVELS: GRS TEST

Period	T	Folded Normal	Bootstrap
1	60	0.0801	0.0104
2	90	0.0453	0.0058
3	120	0.0210	0.0198
4	150	0.0022	0.0022

Period 1: Jan 2011 to Dec 2015; Period 2: July 2008 to Dec 2015; Period 3: Jan 2006 to Dec 2015; Period 4: July 2003 to Dec 2015; and $N = 25$, $P = 0.5$, $k = 1$.

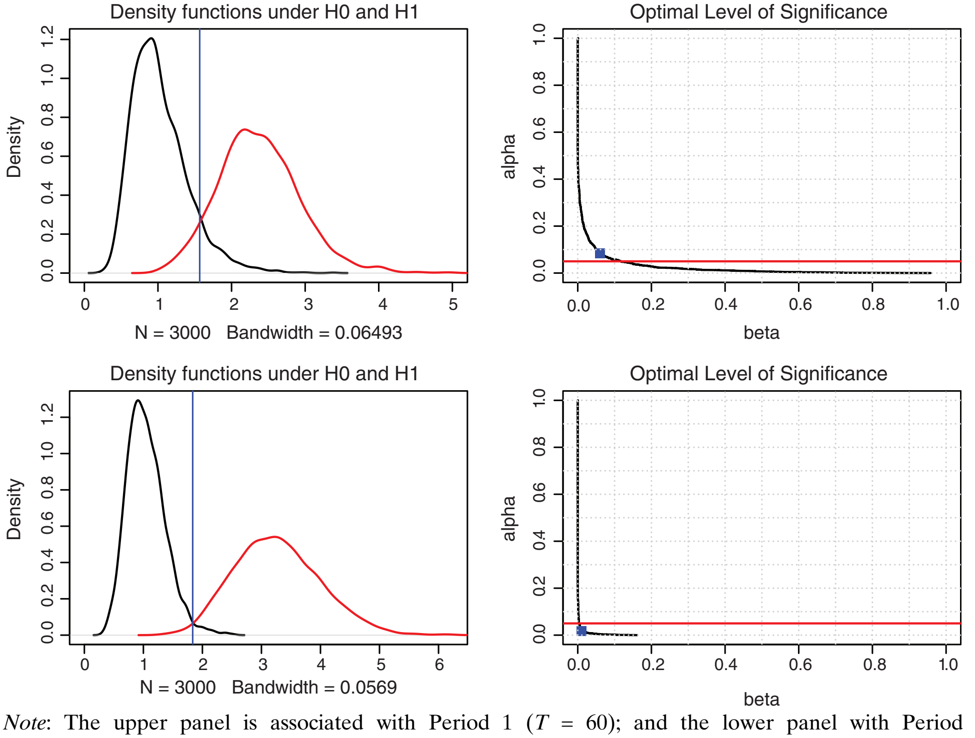
This is in conflict with the model estimation results, where the values of $|a|$ remain economically negligible (around 0.13%) with sufficiently high values of R^2 (nearly 0.95 on average). In contrast, the optimal levels in this case are around 0.01, which are smaller than the respective p-values, leading to acceptance of H_0. Hence, at the optimal level of significance, we have inferential outcomes consistent with model estimation results.

The optimal levels for the GRS test are also calculated based on the bootstrap method. As we observe strong signs of heteroscedasticity in the monthly returns, the wild bootstrap is conducted with 3,000 iterations. The optimal levels calculated are 0.084, 0.039, 0.020, 0.018, respectively for Periods 1 to 4, which are overall similar to those under the assumption of normality. Figure 10 shows the examples of the bootstrap density functions of the GRS test under H_0 and H_1 and the optimal levels of significance for Periods 1 and 4. Table 4 reports the weighted optimal levels for the GRS test when the distribution for λ is assumed to be a folded-normal (the mode at $\hat{\lambda}$ and $\delta = 10$) and wild bootstrap distributions (nine percentiles as done in the previous sub-section). The calculated levels are much lower than the p-values for all periods, leading to acceptance of H_0 for all cases.

An interesting point from the results reported in Table 3 is that the model estimation results ($|a|$ and R^2) do not change sensitivity to increasing sample size. In contrast, the p-value decreases sharply with increasing sample size, switching the inferential outcomes from acceptance to rejection of H_0 at the 5% level, despite qualitatively similar estimation results. That is, as the sample size increases, the empirical validity of the model becomes questionable, if the researcher maintains a conventional level. As mentioned in the literature review, this is widely observed in the context of model mis-specification test as Spanos (2017) points out. This phenomenon occurs because the conventional level totally ignores the increasing power of the test, providing the test outcomes that are in conflict with model estimation results. In contrast, the optimal level is adjusted downward in explicit consideration of the increasing power, and provides inferential outcomes consistent with model estimation results. Fama and French (1993, 1995) are in fact puzzled by the result that their multi-factor models do not pass the GRS test, although the values of $|a|$ are economically negligible and R^2 values are sufficiently high (see, for example, Fama and French, 1993, p.41). This

FIGURE 10

OPTIMAL LEVEL OF SIGNIFICANCE: GRS TEST BASED ON THE WILD BOOTSTRAP



Note: The upper panel is associated with Period 1 (T = 60); and the lower panel with Period 4 (T = 150). For both cases, N = 25, P = 0.5, k = 1. The density curve on the left is for H_0 and the one on the right is for H_1. The vertical line indicates the critical value at the optimal significance level. The square dots in the line of enlightened judgement indicate the points of the optimal levels of significance.

is because they conduct their test at a conventional level of significance, despite an extreme power of the test derived from a massive sample size obtained by including all available data points from 1963 in their hypothesis testing. For example, Fama and French (2015) use the total number of observations of $T \times N = 606 \times 25 = 15150$.

When P *and* k *Take General Values*

We have so far assumed that $P = 0.5$ and $k = 1$. As mentioned before, these values mean that the researcher is impartial between H_0 and H_1, in terms of prior belief and relative loss. However, in practice, the researchers may favour either of H_0 and H_1, similarly to the jury or doctor as in our illustrative examples given earlier. When such prior belief or relative loss is economically sensible, the test can provide inferential outcomes consistent with economic reasoning.

TABLE 5

OPTIMAL SIGNIFICANCE LEVEL AND RELATIVE LOSS (N = 25, P = 0.5)

Period	T	GRS	p-value	α*		
				k = 1	k = 10	k = 0.1
1	60	0.951	0.546	0.074	0.225	0.015
2	90	1.304	0.199	0.039	0.128	0.008
3	120	1.727	0.033	0.017	0.058	0.004
4	150	1.829	0.017	0.014	0.049	0.003

Period 1: Jan 2011 to Dec 2015; Period 2: July 2008 to Dec 2015; Period 3: Jan 2006 to Dec 2015; Period 4: July 2003 to Dec 2015; α*: the optimal level of significance; k: relative loss.

Consider the case of US production function estimation. In testing for $H_0^*: \gamma_1 + \gamma_2 = 0.94$, suppose the researcher strongly believes that this hypothesis is unlikely. This belief may have been formed based on a careful economic analysis; and/or the observation that $\hat{\gamma}_1 + \hat{\gamma}_2 = 0.989$ is close to one and far from the claimed value under H_0^*. We reflect this belief by setting P = 0.2. The optimal level evaluated at $\hat{\lambda} = 3.22$ is 0.702 under the normality; and 0.489 based on the bootstrap: both are much higher than those under P = 0.5. The weighted optimal level from the folded-normal distribution with $\hat{\lambda} = 3.22$ and δ = 3 is 0.658 again much higher than the value when P = 0.5. Hence, H_0^* is again clearly rejected. This example shows that, when the prior belief against the null hypothesis is reflected, the optimal significance level increases, leading to a more decisive rejection of the null hypothesis. Consider another researcher who believes that this hypothesis is still likely with P = 0.8. The optimal level in this case evaluated at $\hat{\lambda} = 3.22$ is 0.041, and the weighted optimal level evaluated at $\hat{\lambda} = 3.22$ and δ = 3 is 0.033, both leading to the acceptance of H_0^*. Given the value of $\hat{\gamma}_1 + \hat{\gamma}_2 = 0.989$, the researcher with P = 0.2 is likely to be more economically sensible. Note that the optimal levels when P = 0.8 are close to 0.05, which means that the use of a conventional level in this case may represent a biased and economically unreasonable prior belief for H_0.

Turning to the GRS test for the asset-pricing model, Table 5 reports α* values when $k \neq 1$ under P = 0.5. The case where k = 0.1 means that the loss from Type I error is ten times bigger than that of Type II error, which indicates that the researcher is highly cautious about Type I error, favouring H_0. For all periods, α* values are much smaller than when k = 1, leading to acceptance of H_0, reflecting the researcher's attitude. We can see that, by setting k = 0.1, the decision to accept H_0 has become much more decisive and unambiguous. The case where k = 10 means that the loss from Type II error is ten times bigger than that of Type I error, which indicates that the researcher is highly cautious about Type II error, favouring H_1. For Periods 3 and 4, α* values are higher than the p-values, leading to rejection of H_0. Given the estimated values of $|a|$ and R^2, we argue that the researcher who favours H_0 with k = 0.1 is more economically sensible. In addition, note that, when $k = 10$ for Periods 3 and 4, α^* values are close to the conventional value of 0.05, which means that the conventional level in this case is associated with a researcher whose attitudes towards H_0 and H_1 are economically unreasonable. Although not discussed in detail, the weighted optimal levels provide similar results.

Gelman-Stern Example

In expressing their caution on the interpretation of statistical significance in practice, Gelman and Stern (2006) provide the following example of three independent studies. Under the same sample size, the first obtains the effect estimate of 25 with the standard error estimate of 10; while the second the effect estimate of 10 with the standard error estimate of 10. The third is conducted with a much higher precision (a larger sample size), with the effect estimate of 2.5 and standard error estimate of 1.

Suppose we test for the null hypothesis that the effect is 0. In the first study, $t = 2.5$ with p-value of 0.006 (one-tailed, assuming normal); in the second study, $t = 1$ with p-value of 0.160. As might be expected, the effect is statistically significant in the first study, but not in the second, at the 5% level of significance. However, the difference between the two effects is 15 with the standard error estimate of $\sqrt{10^2 + 10^2} = 14$, which means that the difference is statistically insignificant at the 5% level (p-value = 0.14). This is contradictory because the effect is found to be statistically significant in the first study but otherwise in the second, but their difference is found to be statistically no different from 0. Gelman and Sterm (2006) use this example to demonstrate a case that the difference between 'significant' and 'not significant' is not itself statistically significant. The third study finds the same t-statistic as the first one, but the effect size estimate is much smaller and closer to 0. While this study finds a positive effect with statistical significance at the 5% level as in the first one, it is not clear if it replicates the first study.

We find that the contradictory and unclear outcomes in the above comparison occur because the power is totally ignored and statistical significance is evaluated at the 5% level of significance for all cases. Suppose, for economic significance, the effect should at least be 10, which is the point at which the power should be calculated. Assuming the sample size of 10 with $P = 0.5$ and $k = 1$, the optimal level of significance in this case is found to be around 0.32 (the effect size $d = 10/\left(10\sqrt{10}\right)$), where d denotes Cohen's d). This will give the same inferential outcomes as those at the 5% level. In comparing the two effect sizes, suppose that the difference between the two effects should be at least be 10 in order to be economically significant. Calculating the power at this point, the optimal level of significance is found to be 0.31, which indicates that the effect size difference is statistically different from 0. Hence, the difference between 'significant' and 'not significant' is indeed statistically significant, at the optimal level of significance.

As for the third study, the effect size estimate of 2.5 is well below 10, so it is highly likely to be economically insignificant. Suppose the standard error estimate of 1 is obtained using a larger sample size such as 30. We find that the optimal level of significance in this case is practically 0 since the power of 1 is achieved. Hence, we find that the effect size of 2.5 is statistically and economically no different from 0, with a clear conclusion that the third study does not replicate the first.

We note that our power analysis above may be sensitive to the value of effect size at which the power is evaluated (e.g., 10). However, as long as this value is economically justifiable, our analysis provides economically and statistically sensible inferential outcomes, overturning contradictory and unclear results obtained at a conventional significance level. The computation in this subsection is conducted using the one-sample and two-sample *t*-tests (one-sided): the detailed *R* codes are given in the Appendix.

Qualified Audit Opinions and Stock Prices

In this subsection, we present an accounting application, revisiting the example given in Johnstone (1990). It is based on the work of Dodd *et al.* (1984) who examine the security returns of firms with qualified financial reports. Let X denote the number of firms observed with negative abnormal returns, which is assumed to follow a binomial distribution $B(T, \theta)$ where T is the sample size (the total number of firms observed) and θ is the probability with which firms with qualified financial reports yield negative abnormal returns. It is hypothesized that 50% of firms with qualified reports earn negative abnormal returns around the time of qualification: H_0: $\theta = 0.5$; against H_0: $\theta > 0.5$. Two subcategories of firms are considered, those with qualifications attributed to litigation and asset realizing (see, for more details, Dodd et al., 1984). Dodd *et al.* (1984) report the following inferential results over the three-day observational interval: for litigation firms, $T = 130$, $x = 76$, $\hat{\theta} \equiv x/T = 0.58$, p-value = 0.0325; for asset realizing firms, $T = 93$, $x = 56$, $\hat{\theta} = 0.60$, p-value = 0.0307; and for all firms, $T = 283$, $x = 165$, $\hat{\theta} = 0.58$, p-value = 0.0031. For the five-day observational interval, they also report that result for all firms with $T = 283$, $x = 153$, $\hat{\theta} = 0.54$, p-value = 0.0954.

It appears that the effect size estimates ($\hat{\theta}$) are well above 0.5 in most cases, indicating that H_1 is more likely than H_0. However, at the conventional level of significance, the decision can be marginal and ambiguous. That is, at the 1% level of significance, H_0 cannot be rejected for all cases except for one; but it is rejected for all cases at the 10% level. One may choose the 5% level of significance, but this choice may be challenged, given that many p-values are close to it.

We calculate the optimal level of significance (assuming $P = 0.5$; $k = 1$), based on a normal approximation to binomial distribution: the computation details with *R* codes are given in the appendix. Note that the power is calculated at the observed value of θ. It is found that $\alpha^* = 0.167$ for litigation firms with $T = 130$, $x = 76$; $\alpha^* = 0.162$ for asset realizing firms with $T = 93$, $x = 56$; $\alpha^* = 0.081$ for all firms with $T = 283$, $x = 165$; and $\alpha^* = 0.247$ for all firms with $T = 283$, $x = 153$. The results show that the p-value is less than α^* for all cases, clearly rejecting H_0 in favour of H_1 at the optimal significance levels. This example again demonstrates that the optimal level of significance provides decisive inferential outcomes consistent with the estimation results.

CONCLUDING REMARKS

The researchers in the business disciplines of science (and also elsewhere) universally adopt a conventional level of significance such as 0.05 in their statistical research and decision making. While a popular benchmark, it is well-known that a conventional level is arbitrary and has no scientific justification. It is a key input to hypothesis testing with a consequential impact on statistical decisions. Many authors have raised serious concerns that mindless and mechanical use of a conventional level has often misled statistical decisions with serious social costs. The gravity of the problem is well summarized by the recent statement made by the American Statistical Association (Wasserstein and Lazar, 2016), warning against the widespread abuse and misuse of 'p -value less than 0.05' criterion. Other authors also have expressed growing concerns about the problems such as data-mining, p -hacking, replication crisis, and publication bias: see, for example, Black (1993), Keuzenkamp and Magnus (1995), Ziliak and McCloskey (2008), Basu and Park (2014), Kim and Ji (2015), Peng (2015), and Harvey (2017), to name a few. Accounting authors who raise similar concerns include Johnstone (1990, 1994), Basu (2015), Dyckman and Zeff (2014, 2015), Dyckman (2016), Ohlson (2015, 2018), Stone (2018), and Kim *et al*. (2018).

We note that central to these problems are the use of a fixed level of significance combined with testing for a sharp null hypothesis, as Leamer (1988) emphatically points out. We argue that these problems can largely be avoided if empirical researchers adjust the level of significance in a scientific way, keeping in mind that a point null hypothesis cannot exactly and literally hold in practice. For more informed and sensible decision making, the level of significance should be chosen with care, as a function of the key factors of hypothesis testing such as the sample size, statistical power, prior belief, and losses from incorrect decisions. Depending on the context and nature of the research, the researcher may also indicate his or her level of prior belief and relative loss, also providing economic justifications where possible and appropriate, which can be reflected in choosing the optimal significance level. In this way, we can promote statistical thinking and judgement, instead of mindless and mechanical practice of what Gigerenzer (2004) calls the 'null ritual'.

In this paper, we propose a decision-theoretic approach to choosing the level of significance, in the context of the linear regression model. The crux of the method we present is not new, since it has been proposed by several authors since the 1960s, such as Arrow (1960) and Leamer (1978), among others. In order to implement the method, a detailed and accurate power analysis should be conducted. In addition to the power calculation based on the assumption of normality, which may have asymptotic justification in small samples, we propose the bootstrap method as an alternative. The latter may be a superior alternative in small samples, taking full account of sampling variability, especially in the presence of non-normality or heteroscedasticity. In calculating the power, the point at which the power is calculated under H_1 is also important. To overcome the subjectivity of this choice, we propose methods of weighting the optimal levels of significance obtained from a grid of possible points under H_1.

We present three empirical applications (from accounting, economics, and finance) and a numerical example, using the accompanying *R* packages. We demonstrate that the tests at the optimal significance level proposed in this paper provide more economically sensible and unambiguous inferential outcomes, than those at a conventional level. The roles that prior belief and relative loss play in statistical decision making, in the context of classical hypothesis testing, are also highlighted. We note that a conventional level often implies a researcher's prior belief and expected loss, which are not consistent with economic reasoning or model estimation results. While the method of choosing the optimal level of significance is presented in the context of a linear regression model in this paper, we note that the same principle can be applied to other models and tests where a meaningful power analysis can be conducted.

It is unfortunate that the pioneering works of early contributors such as Arrow (1960) and Leamer (1978) have been largely ignored and now almost forgotten. The sad consequence is that we are now faced with credibility and reproducibility crises in many areas of science (see, for example, Ziliak and McCloskey, 2008; Basu and Park, 2014; Basu, 2015; Peng, 2015; Kim and Ji, 2015; Wasserstein and Lazar, 2016; Harvey, 2017). We hope that the present paper rekindles the interest in this area of research and contributes to re-building of credibility and integrity in our statistical research.

REFERECNES

Arrow, K. (1960), 'Decision Theory and the Choice of a Level of Significance for the t-test', in I. Olkin, S. G. Ghurye, W. Hoeding, W. G. Madow, and H. B. Mann (eds), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, Stanford University Press, pp. 70–8.

Basu, S. and H.-U. Park (2014), *Publication Bias in Recent Empirical Accounting Research*, Working Paper, http://ssrn.com/abstract=2379889.

Basu, S. (2015), 'Is There a Scientific Basis for Accounting? Implications for Practice, Research, and Education', *Journal of International Accounting Research*, Vol. 14, No. 2, pp. 235–65.

Benjamin, D. J., J. O Berger, M. Johannesson, et al. (2017), 'Redefine Statistical Significance', *Nature Human Behaviour*, https://www.nature.com/articles/s41562-017-0189-z

Black, F. (1993), 'Beta and Return', *The Journal of Portfolio Management*, Vol. 20, No. 1, pp. 8–18.

Burgstahler, D. (1987), 'Inference from Empirical Research', *The Accounting Review*, Vol. LXII, No. 1, pp. 203–14.

Champely, S. (2017) 'pwr: Basic Functions for Power Analysis. R package version 1.2-1', retrieved from https://CRAN.R-project.org/package=pwr.

Das, C. (1994), 'Decision Making by Classical Test Procedures Using an Optimal Level of Significance', *European Journal of Operational Research*, Vol. 73, pp. 76–84.

Davidson, R, and E. Flachaire (2008), 'The Wild Bootstrap, Tamed at Last', *Journal of Econometrics*, Vol. 146, No. 1, pp. 162–9.

DeGroot, M. H. and M. J. Schervish, (2012), *Probability and Statistics*, Addison-Wesley, Boston

Dodd, P., N. Dopuch, R. Holthausen and R. Leftwich (1984), 'Qualified Audit Opinions and Stock Prices: Information Content, Announcement Dates, and Concurrent Disclosures', *Journal of Accounting and Economics*, Vol. 6, pp. 3–38.

Dyckman, T. R. (2016), 'Significance Testing: We Can Do Better', *Abacus*, Vol. 52, No. 2, pp. 319–42.

Dyckman, T. R. and S. A. Zeff (2015), 'Accounting Research: Past, Present, and Future', *Abacus*, Vol. 51, No. 4, pp. 511–24.

—— (2014), 'Some Methodological Deficiencies in Empirical Research Articles in Accounting', *Accounting Horizons.* Vol. 28, No. 3, pp. 695–712.

Efron, B. (1979), 'Bootstrap Methods: Another Look at the Jackknife', *The Annals of Statistics*, Vol. 7, No. 1, pp. 1–26.

Engsted, T. (2009), 'Statistical vs. Economic Significance in Economics and Econometrics: Further Comments on McCloskey and Ziliak', *Journal of Economic Methodology*, Vol. 16, No. 4, pp. 393–408.

Fama, E. F. and K. R. French (1993), 'Common Risk Factors in the Returns on Stocks and Bonds', *Journal of Financial Economics*, Vol. 33, pp. 3–56.

—— (2015), 'A Five-factor Asset-pricing Model', *Journal of Financial Economics*, Vol. 116, pp. 1–22.

Fiebig, D. G. (2017), 'Big Data: Will It Improve Patient-Centered Care?', *Patient*, Vol. 10, pp. 133-139.

Fomby, T. B. and D. K. Guilkey (1978), 'On Choosing the Optimal Level of Significance for the Durbin-Watson test and the Bayesian alternative', *Journal of Econometrics*, Vol. 8, pp. 203–13.

Gelman, A. and H. Stern (2005), 'The Difference between "Significant" and "Not Significant" is Not Itself Statistically Significant', *The American Statistician*, Vol. 60, No. 4, pp. 328–31.

Gibbons, M. R., S. A. Ross, and J. Shanken (1989), 'A Test of the Efficiency of a Given Portfolio', *Econometrica*, Vol. 57, pp. 1121–52.

Gigerenzer, G. (2004), 'Mindless Statistics: Comment on "Size Matters"', *Journal of Socio-Economics*, Vol. 33, pp. 587–606.

Goldberger, A. S. (1991), *A Course in Econometrics*, Harvard University Press, Cambridge, MA.

Grossman, S. J. and J. E. Stiglitz (1980), 'On the Impossibility of Informationally Efficient Markets', *The American Economic Review*, Vol. 70, No. 3, pp. 393–408.

Gujarati, D. (2015), *Econometrics by Example*, Palgrave, London.

Hand, D. J. (2016), 'Editorial: Big Data and Data Sharing', *Journal of the Royal Statistical Society A*, Vol. 179, No. 3, pp. 629–31.

Harford, T. (2014), 'Big Data: Are We Making a Big Mistake?', *Significance*, December, pp. 14–19.

Harvey, C. R., Y. Lin, and H. Zhu (2016), '… and the Cross-Section of Expected Returns', *The Review of Financial Studies*, Vol. 29, No. 1, pp. 5–68.

Harvey, C. R. (2017), 'Presidential Address: The Scientific Outlook in Financial Economics', *Journal of Finance*, Vol. 72, No. 4, pp. 1399–440.

Hodges, J. L. Jr. and E. L. Lehmann (1954), 'Testing the Approximate Validity of Statistical Hypotheses', *Journal of the Royal Statistical Society, Series B (Methodological)*, Vol. 16, No. 2, pp. 261–68.

Ioannidis, J. P. A., T. D. Stanley, and H. Doucouliagos, (2017), 'The Power of Bias in Economics Research', *The Economic Journal*, Vol. 125, F236-F265.

Johnson, V. E. (2013), 'Revised Standards for Statistical Evidence', *Proceedings of the National Academy of Sciences*, www.pnas.org/cgi/doi/10.1073/pnas.1313476110

Johnstone, D. and D. Lindley (1995), 'Bayesian Inference Given Data Significant at Level α: Tests of Point Hypotheses', *Theory and Decision*, Vol. 38, pp. 51–60.

Johnstone, D. J. (1990), 'Sample Size and the Strength of Evidence: A Bayesian Interpretation of Binomial Tests of the Information Content of Qualified Audit Report', *Abacus*, Vol. 26, pp. 17–35.

—— (1994), 'A Statistical Paradox in Auditing', *Abacus*, Vol. 30, No. 1, pp. 44–9.

—— (1997), 'Comparative Classical and Bayesian Interpretations of Statistical Compliance Tests in Auditing', *Accounting and Business Research*, Vol. 28, pp. 53–82

Kandel, S. and R. F. Stambaugh (1996), 'On the Predictability of Stock Returns: An Asset-Allocation Perspective', *The Journal of Finance*, Vol. 51, No. 2, pp. 385–424.

Kaplan, R. D., D. A. Chambers, and R. E. Glasgow (2014), 'Big Data and Large Sample Size: A Cautionary Note on the Potential for Bias', *Clinical and Translational Science*, Vol. 7, No. 5, pp. 342–46.

Keuzenkamp, H. A. and J. Magnus (1995), 'On Tests and Significance in Econometrics', *Journal of Econometrics*, Vol. 67, No. 1, pp. 103–28.

Kim J. H. (2017a), OptSig: Optimal Level of Significance for Regression and Other Statistical Tests, R package version 1.0. https://cran.r-project.org/web/packages/OptSig/index.html

—— (2017b), GRS.test: GRS Test for Portfolio Efficiency, Its Statistical Power Analysis, and Optimal Significance Level Calculation. R package version 1.1, https://cran.r-project.org/web/packages/GRS.test/index.html

Kim, J. H. (2017c), 'Stock Returns and Investors' Mood: Good Day Sunshine or Spurious Correlation?', *International Review of Financial Analysis*, Vol. 52, pp. 94–103.

Kim J. H. and I. Choi (2017), 'Unit Roots in Economic and Financial Time Series: A Re-evaluation at the Decision-based Significance Levels', *Econometrics*, Vol. 5, No. 3, pp. 41.

Kim, J. H. and P. Ji (2015), 'Significance Testing in Empirical Finance: A Critical Review and Assessment', *Journal of Empirical Finance*, Vol. 34, pp. 1–14.

Kim J. H., P. Ji, and K. Ahmed (2018), 'Significance Testing in Accounting Research: A Critical Evaluation based on Evidence', *Abacus*, Vol. 54, No. 4, pp. 524–46.

Kish, L. (1959), 'Some Statistical Problems in Research Design', *American Sociological Review*, Vol. 24, pp. 328–38.

Klein, R. W. and S. J. Brown (1984), 'Model Selection When There is Minimal Prior Information', *Econometrica*, Vol. 52, pp. 1291–312.

Koop, G. and M. F. J. Steel (1994), 'A Decision-Theoretic Analysis of the Unit-Root Hypothesis Using Mixtures of Elliptical Models', *Journal of Business and Economic Statistics*, Vol. 12, pp. 95–107.

Labovitz, S. (1968), 'Criteria for Selecting a Significance Level: A Note on the Sacredness of 0*.05', The American Sociologist*, Vol. 3, pp. 200–22.

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

—— (1988), 'Things that Bother Me', *Economic Record*, Vol. 64, No. 4, pp. 331–335.

Lehmann E. L. and J. S. Romano (2005), *Testing Statistical Hypothesis*, Springer, New York.

Lindley, D. V. (1957), 'A Statistical Paradox', *Biometrika*, Vol. 44, pp. 187–192.

Lindley, D.V. (2014), *Understanding Uncertainty*, Wiley, New York.

Liu, R. Y. (1988). 'Bootstrap Procedures under some Non-I.I.D. Models', *The Annals of Statistics*, Vol. 16, No. 4, pp. 1696–708.

Mammen, E. (1993), 'Bootstrap and Wild Bootstrap for High Dimensional Linear Models', *The Annals of Statistics*, Vol. 21, No. 1, pp. 255–85.

Manderscheid, L. V. (1965), 'Significance Levels-0.05, 0.01, or?', *Journal of Farm Economics*, Vol. 47, No. 5, pp. 1381–385.

Moore, D. S. and G. P. McCabe (1993), *Introduction to the Practice of Statistics*, W.H. Freeman and Company, New York.

Moosa, I. A. (2017), *Econometrics as a Con Art: Exposing the Limitations and Abuses of Econometrics*, Edward Elgar, Cheltenham.

Morrison, D. E. and R. E. Henkel (1970), *The Significance Test Controversy: A Reader*, Aldine Transactions, New Brunswick, NJ.

Ohlson, J. A. (2015), 'Accounting Research and Common Sense', *Abacus*, Vol. 51, No. 4, pp. 525–35.

—— (2018), 'Researchers' Data Analysis Choices: An Excess of False Positives?', retrieved from SSRN: https://ssrn.com/abstract=3089571.

Peng, R. (2015), 'The Reproducibility Crisis in Science: A Statistical Counterattack', *Significance*, Vol. 12, No. 3, pp. 30–32.

Peracchi, F. (2001), *Econometrics*, Wiley, New York.

Perez, M-E. and L. R. Pericchi (2014), 'Changing Statistical Significance with the Amount of Information: The Adaptive Significance Level', *Statistics and Probability Letters*, Vol. 85, pp. 20–24.

Poirier, D. J. (1995), *Intermediate Statistics and Econometrics: A Comparative Approach*, The MIT Press, Cambridge, Massachusetts.

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, retrieved from https://www.R-project.org/.

Shanken, J. (1987), 'A Bayesian Approach to Testing Portfolio Efficiency', *Journal of Financial Economics*, Vol. 19, No. 2, pp. 195–215.

Skipper, J. K. Jr, A. L. Guenther, and G. Nass (1967), 'The Sacredness of .05: A Note on Concerning the Use of Statistical Levels of Significance in Social Science', *The American Sociologist*, Vol. 2, pp. 16–18.

Soyer, E. and R. M. Hogarth (2012), 'The Illusion of Predictability: How Regression Statistics Mislead Experts', *International Journal of Forecasting*, Vol. 28, pp. 695–711.

Spanos, A. (2017), Mis-specification Testing in Retrospect', *Journal of Economic Surveys*, doi: https://doi.org/10.1111/joes.12200

Startz, R. (2014). 'Choosing the More Likely Hypothesis', *Foundations and Trends in Econometrics*, Vol. 7, pp. 119–89.

Stone, D. N. (2018), 'The 'New Statistics' and Nullifying the Null: Twelve Actions for Improving Quantitative Accounting Research Quality and Integrity', *Accounting Horizon*, Vol. 32, No. 1, pp. 105–20.

Wasserstein R. L. and N. A. Lazar (2016), 'The ASA's Statement on p-values: Context, Process, and Purpose', *The American Statistician*, Vol. 70, pp. 129–33.

Winer, B. J. (1962), *Statistical Principles in Experimental Design*, McGraw-Hill, New York.

Zellner, A. and A. Siow (1980), 'Posterior Odds Ratios for Selected Regression Hypotheses', in J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (eds), *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia*, University of Valencia Press, Valencia, Spain, pp. 585-603.

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

APPENDIX

In this section, we provide the details concerning the accompanying *R* packages and examples of *R* commands used in the empirical applications.

APPENDIX A
R PACKAGES

The computational resources are available from two *R* packages called OptSig (Kim, 2017a) and GRS.test (Kim, 2017b). The former provides the functions for a linear restriction test in the regression model, and the latter those for the GRS test. The former also includes the functions for a range of other statistical tests including those for the population mean and population proportion, using the *R* package pwr (Champely, 2017).

The *R* functions for the calibration rules for the optimal level of significance for a range of unit root tests, proposed by Kim and Choi (2017), are available from http://www.mdpi.com/2225-1146/5/3/41 as a supplementary material.

APPENDIX B
R EXAMPLES

The following *R* codes replicate the part of the results for the test for the constant returns to scale in the testing for constant returns to scale:

```
rm(list=ls(all=TRUE))
library(OptSig)
data(data1) #Table 2.1 of Gujarati (2015)
# Extract Y and X
y=data1$lnoutput; x=cbind(data1$lncapital,data1$lnlabor)
# Restriction matrices for the slope coefficients sum to 0.94
Rmat=matrix(c(0,1,1),nrow=1); rvec=matrix(0.94,nrow=1)
# Model Estimation
M=R.OLS(y,x,Rmat,rvec); print(M$coef)
# Degrees of Freedom and estimate of non-centrality parameter
K=ncol(x)+1; T=length(y); df1=nrow(Rmat);df2=T-K; NCP=M$ncp
# Optimal level of Significance: Under Normality
Result=Opt.Sig(df1,df2,ncp=NCP,p=0.5,k=1, Figure=TRUE)
print(Result);
Power.F(df1,df2,ncp=NCP,alpha=Result$alpha.opt,Figure=TRUE)
# Weighted optimal levels
Opt.SigWeight(df1,df2,m=NCP,delta=3,p=0.5,k=1,Figure=TRUE)
# Optimal level of Significance: Bootstrapping
Opt.SigBoot(y,x,Rmat,rvec,p=0.5,k=1,Figure=TRUE)
Opt.SigBootWeight(y,x,Rmat,rvec,p=0.5,k=1,nboot=1000)
# Breusch-Pagan test for heteroscedasticity
e = M$resid[,1]
# Restriction matrices for the slope coefficients being 0
Rmat=matrix(c(0,0,1,0,0,1),nrow=2); rvec=matrix(0,nrow=2)
# Model Estimation for the auxiliary regression
M1=R.OLS(e^2,x,Rmat,rvec);
# Degrees of Freedom and estimate of non-centrality parameter
K=ncol(x)+1; T=length(y); df1=nrow(Rmat);df2=T-K; NCP=M1$ncp
# Optimal level of Significance: Under Normality
Opt.Sig(df1,df2,ncp=NCP,p=0.5,k=1, Figure=TRUE)
Opt.SigWeight(df1,df2,m=NCP,delta=3,p=0.5,k=1,Figure=TRUE)
#Optimal level of Significance: Bootstrapping
Opt.SigBoot(e^2,x,Rmat,rvec,p=0.5,k=1,Figure=TRUE)
Opt.SigBootWeight(e^2,x,Rmat,rvec,p=0.5,k=1,nboot=1000)
```

The following *R* codes replicate the part of the results for the GRS test in the section about empirical validity of an asset-pricing model:

```
rm(list=ls(all=TRUE))
library(GRS.test)
data(data) # Fama-French data
# Choose the last n observations from the data set
n=60; m1=nrow(data)-n+1; m2=nrow(data)
factor.mat = data[m1:m2,2:6] #Fama-French 5-factors
ret.mat = data[m1:m2,8:ncol(data)] #25 size-BM portfolio
returns
# Model Estimation
M1=GRS.test(ret.mat,factor.mat)
M2=GRS.MLtest(ret.mat, factor.mat)
# Print the statistics
print(M1$GRS.stat) # GRS test statistic
print(M1$GRS.pval); # its p-value
mean(abs(M1$coef[,1])) # |a| value
mean(M1$R2) # Mean of R2 values
# Optimal level of Significance under Normality
GRS.optimal(T=n, N=25, K=5, theta=M2$theta,
 ratio=M2$ratio, Graph = TRUE)
GRS.Power(T=n, N=25, K=5, theta=M2$theta,
 ratio=M2$ratio,alpha=0.01,xmax=10,Graph=TRUE)
GRS.optimalweight(T=n,N=25,K=5,theta=M2$theta,
 ratio=M2$ratio,delta=10,p=0.5,k=1)
# Optimal level of Significance based on wild bootstrapping
GRS.optimalboot(ret.mat,factor.mat,p=0.5,k=1,
 nboot=3000,wild=TRUE,Graph=TRUE)
GRS.optimalbootweight(ret.mat,factor.mat,p=0.5,k=1,
 nboot=3000,wild=TRUE,Graph=TRUE)
```

The *R* codes for the optimal level of significance in the Gelman-Stern examples are

```
library(OptSig)
Opt.sig.t.test(d=10/(10*sqrt(10)),n=10,type="one.sample",
 alternative="greater")
Opt.sig.t.test(d=10/(14/sqrt(0.2)),n=20,type="two.sample",
 alternative="greater")
Opt.sig.t.test(d=10/sqrt(30),n=30,type="one.sample",
 alternative="greater")
```

Note that d refers to the (standardized) effect size (Cohen's d) at the point of power calculation. The *R* codes for the optimal level of significance in relation to qualified audit opinions and stock prices are

```
library(OptSig)
n=283; x=153
theta=0.5
mean=n*theta # mean of the normal distribution
var=n*theta*(1-theta) # variance
sigma=sqrt(n)*sqrt(var)
d=(x-mean)/sigma # calculating Cohen’s d (standardized
effect size)
Opt.sig.norm.test(d=d,n=n,alternative="greater")
```