Research Division
Federal Reserve Bank of St. Louis
Working Paper Series

Asymptotic Inference for Performance Fees and the Predictability of Asset Returns

Michael W. McCracken
and
Giorgio Valente

Working Paper 2012-049C
http://research.stlouisfed.org/wp/2012/2012-049.pdf

October 2012
Revised July 2016

FEDERAL RESERVE BANK OF ST. LOUIS
Research Division
P.O. Box 442
St. Louis, MO 63166

The views expressed are those of the individual authors and do not necessarily reflect official positions of the Federal Reserve Bank of St. Louis, the Federal Reserve System, or the Board of Governors.

Federal Reserve Bank of St. Louis Working Papers are preliminary materials circulated to stimulate discussion and critical comment. References in publications to Federal Reserve Bank of St. Louis Working Papers (other than an acknowledgment that the writer has had access to unpublished material) should be cleared with the author or authors.

Asymptotic Inference for Performance Fees and the Predictability of Asset Returns*†

Michael W. McCracken[a] and Giorgio Valente[b]

a: Federal Reserve Bank of St. Louis

b: City University of Hong Kong

July 2016

Abstract

In this paper, we provide analytical, simulation, and empirical evidence on a test of equal economic value from competing predictive models of asset returns. We define economic value using the concept of a performance fee – the amount an investor would be willing to pay to have access to an alternative predictive model used to make investment decisions. We establish that this fee can be asymptotically normal under modest assumptions. Monte Carlo evidence shows that our test can be accurately sized in reasonably large samples. We apply the proposed test to predictions of the US equity premium.

JEL classification: C53, C12, C52

Keywords: Utility-based comparisons, economic value, out-of-sample forecasting, predictability.

*This work was partly written while Giorgio Valente was visiting the Federal Reserve Bank of St Louis whose hospitality is gratefully acknowledged. We gratefully acknowledge helpful comments from Andrew Patton, Enrique Sentana, Nicholas Taylor and conference participants at the Society for Financial Econometrics Annual Meeting in Oxford and the seminar participants at Baptist University of Hong Kong, Chinese University of Hong Kong, and National Taiwan University. The views expressed herein are solely those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of St. Louis, Federal Reserve System, or any of its staff.

†*McCracken (corresponding author)*: Research Division; Federal Reserve Bank of St. Louis; P.O. Box 442; St. Louis, MO 63166; `michael.w.mccracken@stls.frb.org`. *Valente*: Department of Economics and Finance; College of Business; City University of Hong Kong; Kowloon, Hong Kong; `g.valente@cityu.edu.hk`.

1 INTRODUCTION

A wide variety of studies have debated whether asset returns are, or should be, predictable using information available to investors. Many, including Barberis (2000), Lettau and Ludvigson (2001), Campbell and Thompson (2008), and Cochrane (2008, 2011), conclude that asset returns are predictable. Others, including Goyal and Welch (2008) and Boudoukh, Richardson, and Whitelaw (2008), remain skeptical.

In many instances, the evidence of predictability (or the lack thereof) is based on out-of-sample conditional mean predictions of asset returns. These predictions are, in turn, evaluated using *statistical* measures of out-of-sample predictive accuracy. In most cases, mean squared error (MSE) is used (e.g., Campbell and Thompson, 2008; Goyal and Welch, 2008; Ferreira and Santa Clara, 2011; and the references therein). In many, though not all, cases the question of interest is whether a newly developed predictive model is better at guiding investment decisions than a pre-existing baseline model. As such, when statistical measures of predictive accuracy (such as MSE) are used to evaluate the performance of a new model, not only are MSEs reported but inference is conducted to determine whether any differences in the MSEs are statistically significant. Inference is often conducted using common approaches to out-of-sample inference (West, 1996; Clark and McCracken, 2001; McCracken, 2007).

While statistical metrics of evaluation are informative, there is increasing interest in evaluating the predictability of asset returns using *economic value* measures. Examples of economic measures of predictability include Sharpe ratios (Fleming et al., 2001), performance fees (Patton, 2004), certainty-equivalent returns (Ingersoll et al., 2007), and profits (Leitch and Tanner, 1991). As for the case of statistical measures, the most common question of interest is whether a newly developed predictive model is "better" at generating higher economic value than an established baseline model. However, despite their increasing use, these economic measures are typically, though not always, reported with no indication of whether any empirical differences are statistically significant.[1]

[1] Wachter and Warusawitharana (2015) conduct a test of the null that certainty equivalent relations (CERs) are

In this paper, we develop an asymptotically valid approach to inference when performance fees are used as the economic measure of predictive accuracy. Since the performance fee (Φ) is the amount an investor would be willing to pay to have access to an alternative predictive model, and can take both negative and positive values, the null and alternative hypotheses are both composite and take the form $H_0 : \Phi \leq 0$ and $H_A : \Phi > 0$, respectively. As is usual when the null hypothesis is composite (e.g., Hansen, 2005), we implement our test by using critical values associated with the asymptotic distribution of the estimated performance fee ($\hat{\Phi}$) under the least favorable alternative within which $\Phi = 0$.

Specifically, we first show that when Φ is zero, $\hat{\Phi}$ is asymptotically normal with zero mean and, following West (1996) and West and McCracken (1998), has an asymptotic variance that is affected by estimation error – additional variation induced by the fact that the models must be estimated prior to their use. In some instances, consistent estimation of this asymptotic variance is straightforward and thus standard Normal critical values can be used to conduct inference. In others, estimating the asymptotic variance can be quite complicated. We therefore also investigate the efficacy of a bootstrap approach to inference based on the methods developed in Calhoun (2015). Our Monte Carlo evidence suggests that both asymptotic and bootstrap-based critical values can provide reasonably well-sized tests but, in particular, obtaining substantial power sometimes requires large samples. In addition, the simulation evidence shows that there can be large differences in statistical versus economic measures of predictive ability. In fact, under some circumstances, the performance fee can be zero or even negative despite having strong conditional mean predictive ability in asset returns.

We are, of course, not the first to emphasize that statistical evidence of predictability need not imply anything about economic predictability. In particular, Cochrane (1999), Sentana (1999, 2005), and Taylor (2013) have developed formulas linking R^2s of linear predictive equations to both Sharpe ratios and performance fees. Our main contribution is not in extending these formulas, but

zero versus an alternative that they are positive. Their critical values are generated using Monte Carlo methods assuming no return predictability.

rather in providing a method for conducting asymptotically valid inference on the null hypothesis that a performance fee is at most zero versus an alternative in which it is positive. Our results reinforce many of their findings by showing that the non-monotonic link between statistical and economic predictability leads to difficulties in conducting inference.

The paper is organized as follows. Section 2 provides a simple example of the type of application we have in mind. Section 3 develops the theoretical results and Section 4 discusses practical methods related to inference. Section 5 provides simulation results designed to investigate the size and power properties of the test statistic. In Section 6 we apply our analytical results to predictions of the equity premium by means of various predictors (as in Goyal and Welch, 2008). A final section concludes.

2 A SIMPLE ILLUSTRATIVE EXAMPLE

In this section, we delineate a simple example of performance fees in the context of a portfolio of two assets: a single risk-free and a single risky asset (e.g., a stock index) that is predictable by means of a single variable. While simplistic, we use this example not only so that we can convey certain analytical details in closed form, but also because it is consistent with the existing literature that uses utility-based comparisons to assess asset returns predictability. It is worth emphasizing that our analytical results, discussed in the following section, are not restricted to this particular environment.

Let r_t denote the return on the risky asset and r_t^f the rate of return on the risk-free asset. Define $ep_{t+\tau} = r_{t+\tau} - r_{t+\tau}^f$ as the excess stock index return, or equity premium, in period $t+\tau$ and let z_t denote a variable observed at time t that is believed to predict ep at a future time $t+\tau$. The investor uses the predictive regression,

$$ep_{t+\tau} = \alpha_{1,0} + \alpha_{1,1} z_t + e_{1,t+\tau}, \tag{1}$$

to make conditional mean forecasts of future stock index excess returns. The variable z_t has predictive content for $ep_{t+\tau}$ if $\alpha_{1,1} \neq 0$. If z_t has no predictive content, then $\alpha_{1,1} = 0$ and Equation

(1) collapses to

$$ep_{t+\tau} = \alpha_{0,0} + e_{0,t+\tau},$$

where stock index excess returns are equal to their historical mean plus an unpredictable error term. Throughout the paper we denote the competing model as 1 and the baseline model as 0. In addition, the investor uses a parametric model to estimate the conditional variance of the excess returns. Define $\sigma^2_{i,t+\tau}(\vartheta_i)$ as the time t conditional variance of $ep_{t+\tau}$ implied by model $i = 0, 1$ as a function of the finite-dimensioned parameter estimates $\hat{\vartheta}_{i,t}$.

At each forecast origin $t = T, ..., T + P - \tau$, the investor uses the conditional mean $(ep_{t+\tau}(\hat{\alpha}_{i,t}))$ and conditional variance $(\sigma^2_{i,t+\tau}(\hat{\vartheta}_{i,t}))$ predictions of excess returns to decide how much of her wealth to invest in the risky and the risk-free assets. If the investor is endowed with mean-variance preferences, the optimal allocation to the risky asset $w_{i,t}$ at any time t from model $i = 0, 1$ is given by the conventional formula

$$w_{i,t}(\hat{\beta}_{i,t}) = \hat{w}_{i,t} = \frac{ep_{t+\tau}(\hat{\alpha}_{i,t})}{\gamma\sigma^2_{i,t+\tau}\left(\hat{\vartheta}_{i,t}\right)}, \tag{2}$$

where $\hat{\beta}_{i,t} = (\hat{\alpha}'_{i,t}, \hat{\vartheta}'_{i,t})'$ and γ is the investor's known coefficient of relative risk aversion (RRA). If the investor is endowed with initial wealth $W = 1$ at each forecast origin, the time $t + \tau$ realized gross return implied by model $i = 0, 1$ equals

$$\hat{R}_{i,t+\tau} = 1 + r^f_{t+\tau} + \hat{w}_{i,t}(\hat{\beta}_{i,t})ep_{t+\tau}.$$

Any improvements in predictive ability between models 0 and 1 are assessed using utility-based comparisons. That is, for a prespecified utility function $U(R)$, economic value is evaluated by comparing the average utility implied by models 0 and 1 across all forecast origins. If we let $\tilde{P} = P - \tau$ denote the number of τ-step-ahead forecasts, this takes the form

$$\overline{U}(\hat{R}_i) = \tilde{P}^{-1} \textstyle\sum_{t=T}^{T+P-\tau} U(\hat{R}_{i,t+\tau}).$$

As in Fleming et al. (2001), the difference in utilities is characterized using the concept of a performance fee. This fee is the value of $\hat{\Phi}$ that satisfies

$$\overline{U}(\hat{R}_1 - \hat{\Phi}) - \overline{U}(\hat{R}_0) = 0.$$

We interpret Φ as the maximum fraction of wealth the investor would be willing to pay per period to switch from model 0 to model 1. If the two conditional variance models are identical, this criterion measures how much a risk-averse investor is willing to pay for conditioning on the information in the predictive variable z_t. It follows that, if there is no predictive power embedded in the variable z_t, then $\Phi = 0$; whereas, if z_t helps to predict stock excess returns, one expects $\Phi > 0$.

We can better understand the behavior of Φ when $\alpha_{1,1} \neq 0$ if a few more assumptions are made. In particular, let the forecast horizon be $\tau = 1$ and assume that z_t follows a stationary AR(1) process of the form

$$z_t = \mu_z(1-\rho) + \rho z_{t-1} + v_t,$$

where (e_t, v_t) are i.i.d. normally distributed with zero means and variances σ_e^2 and σ_v^2. Let μ_z and σ_z^2 denote the unconditional mean and variance of z_t. Finally, assume that the conditional mean models are estimated by OLS and the conditional variance models are identical so that $\sigma_{0,t+1}^2(\hat{\vartheta}_{0,t}) = \sigma_{1,t+1}^2(\hat{\vartheta}_{1,t})$. More specifically, for ease of presentation, assume that the predictions of these conditional variances are obtained simply by using a consistent estimate of the unconditional variance.[2]

Straightforward algebra shows that

$$\Phi = \left(\frac{\alpha_{1,1}^2\sigma_z^2}{\gamma(\alpha_{1,1}^2\sigma_z^2 + \sigma_e^2)}\right)\left(\frac{\sigma_e^2 - 3(\alpha_{1,0} + \alpha_{1,1}\mu_z)^2}{2(\alpha_{1,1}^2\sigma_z^2 + \sigma_e^2)}\right). \tag{3}$$

Equation (3) is the product of two terms. If the first term is zero, as it is when $\alpha_{1,1} = 0$, then $\Phi = 0$. This first term can be interpreted as the (population) R^2 from the predictive model 1, scaled by γ. Using this interpretation, the smaller the R^2 from model 1, the closer Φ is to zero. The second term is less easily interpretable but arises due to the marginal differences in the variance components of the mean-variance utility functions.

Since the first term in parentheses increases monotonically with $|\alpha_{1,1}|$, it seems likely that larger (absolute) values of $\alpha_{1,1}$ imply larger values of Φ. In this case, statistical measures of predictive

[2] While this may seem odd given our framework, rolling window estimates of the unconditional variance of ep_{t+1} are often used as estimates of the conditional variance of excess returns in the empirical literature (Goyal and Welch, 2008; Campbell and Thompson, 2008; and Ferreira and Santa Clara, 2011).

accuracy coincide with utility-based economic measures of predictability. In fact, stronger evidence of statistical predictability, represented by large t-statistics on $\alpha_{1,1}$ or large R^2 recorded for the unrestricted regression, implies larger utility gains to investors, and subsequently larger values of $\Phi > 0$.

However, it is worthwhile noting that this *intuitive* case need not be the only case in which Φ equals zero, nor is it trivially true that larger (absolute) values of $\alpha_{1,1}$ imply larger values of Φ. Note that the second term in parentheses also depends on $\alpha_{1,1}$. Taken as a whole, this implies that as $\alpha_{1,1}$ deviates from zero, Φ can be positive, negative, or zero depending on the specifics of the data-generating process. In Figure 1, we plot Φ as a function of $\alpha_{1,1}$ on the basis of parameter values loosely calibrated using the return on the NYSE value-weighted index as our risky asset, its dividend yield as our predictor, and the yield on the 3-month T-bill as our return on the risk-free asset.[3] Given these parameter values, Figure 1 shows the values of Φ obtained as a function of $\alpha_{1,1} \in [0, 2]$. As expected, Φ is zero when $\alpha_{1,1}$ is zero and initially increases as $\alpha_{1,1}$ increases. And yet for large enough values of $\alpha_{1,1}$, Φ begins to decline and eventually becomes negative. Clearly, large and statistically significant statistical measures of accuracy such as MSEs and R^2s need not provide any indication of economic performance when measured using performance fees.

As shown in the following section, the fact that Φ can have multiple roots (as a function of $\alpha_{1,1}$) makes inference more complicated than we might want. In particular, while we are able to establish that $\tilde{P}^{1/2}\hat{\Phi}$ is asymptotically normal with zero mean at each root, the asymptotic variance differs whether $\alpha_{1,1}$ is zero or nonzero. We consider two approaches to manage this problem. First, we show that there is a straightforward estimator of the asymptotic variance that is robust to both instances. The simulation evidence suggests this estimator can work reasonably well but sometimes requires large samples to provide accurately sized tests. In addition, we consider a percentile-bootstrap approach to inference based on the work of Calhoun (2015) that is designed

[3] Specifically, we use these data to estimate empirically relevant values of μ_z, σ_z^2, σ_e^2, and $E(ep_{t+1})$. The latter two terms are used to parameterize $\alpha_{1,0} = E(ep_{t+1})$ and $\sigma_e^2 = Var(ep_{t+1})$. The coefficient of RRA is set to $\gamma = 3$. Some ad hoc adjustments are made to induce the hump-shaped pattern for Φ.

explicitly for out-of-sample inference when the test statistic is asymptotically Normal. While the two approaches yield different results, the bootstrap approach to inference is much easier to implement and yields comparable size and power in our simulations.

3 THEORETICAL RESULTS

This section provides the asymptotic distribution of a per-period performance fee measure $\hat{\Phi}$ on the boundary of the null hypothesis and hence $\Phi = 0$. The performance fee Φ is estimated as a function of two sequences of pseudo-out-of-sample forecasts: one each for models 0 and 1. In the context of the example from Section 2, these forecasts consist of both conditional mean and conditional variance forecasts. To calculate the performance fee, we assume that the investor has access to the necessary observables over the time frame $s = 1, ..., T + P$. This sample is split into an in-sample period $s = 1, ..., T$ and an out-of-sample period $t = T + 1, ..., T + P$. At each forecast origin $t = T, ..., T + P - \tau$, both of the parametric τ-period-ahead investing models are estimated and used to construct a forecast that is then used to construct portfolio weights. The assumptions used to derive the asymptotic results are presented below and closely follow those in West (1996) with some modest deviations.

Assumption 1. Let $\beta = (\beta_0', \beta_1')'$. (a) There exists a function $f(X_{t+\tau}, \beta) = f_{t+\tau}(\beta)$, with $f_{t+\tau}(\beta^*) = f_{t+\tau}$, that is twice continuously differentiable in β and satisfies $\tilde{P}^{1/2}\hat{\Phi} = \tilde{P}^{-1/2} \sum_{t=T}^{T+P-\tau} f_{t+\tau}(\hat{\beta}_t) + o_p(1)$. (b) For the max norm $|.|$ and some open neighborhood N of β^*, $E(\sup_{\beta \in N} \partial^2 f_{t+\tau}(\beta)/\partial\beta\partial\beta') < D$ some finite scalar D.

Assumption 2. The parameters are estimated using one of two sampling schemes: the recursive or the rolling. The recursive parameter estimates satisfy $\hat{\beta}_{i,t} - \beta_i^* = B_i(t)H_i(t)$, where $B_i(t) \rightarrow^{a.s.} B_i$ is a non-stochastic matrix, $H_i(t) = t^{-1} \sum_{s=1}^{t-\tau} h_{i,s+\tau}$ with $Eh_{i,s+\tau} = 0$, and β_i^* denotes the population counterparts of the parameter estimates $\hat{\beta}_{i,t}$. The rolling parameter estimates $\hat{\beta}_{i,t}$ are defined similarly but are constructed using data over the ranges $s = t - T, ..., t$. Define

$\hat{\beta}_t = (\hat{\beta}'_{0,t}, \hat{\beta}'_{1,t})'$, $h_{t+\tau} = (h'_{0,t+\tau}, h'_{1,t+\tau})'$, and $B = diag(B_0, B_1)$. All parameter estimates are constructed using the same sampling scheme.

Assumption 3. Define $f_{t+\tau,\beta} = \partial f_{t+\tau}(\beta^*)/\partial\beta$ and let $w_t = (f'_{t+\tau,\beta}, f_{t+\tau}, h'_{t+\tau})'$. (a) For some $d > 1$, $\sup_t E||w_t||^{4d} < \infty$. (b) w_t is strong mixing of size $-3d/(d-1)$. (c) w_t is covariance stationary.

Assumption 4. The number of in-sample observations associated with the initial forecast origin T, and the number of predictions $\tilde{P} = P - \tau + 1$, are arbitrarily large; in particular, they satisfy the restriction that $\lim_{P,T\to\infty} \frac{P}{T} = \pi \in (0, \infty)$,

Assumption 5. Define $F = Ef_{t+\tau,\beta}$. If the models are nested, $FB \neq 0$.

Before providing the main result, it is important to explain some key assumptions and their implications for the validity of our testing procedure. Assumption 1 maps the problem of inference on $\hat{\Phi}$ into a framework in which the theoretical results in West (1996) can be applied directly. While the assumption is stated at a very high level, it is actually very simple to verify. For example, in the context of the mean-variance example from Section 2, Assumption 1 is satisfied for the function

$$f_{t+\tau}(\hat{\beta}_t) = (\hat{R}_{1,t+\tau} - \frac{\gamma}{2}(\hat{R}_{1,t+\tau} - ER_{1,t+\tau})^2) - (\hat{R}_{0,t+\tau} - \frac{\gamma}{2}(\hat{R}_{0,t+\tau} - ER_{0,t+\tau})^2) \tag{4}$$

if $\bar{R}_i \to^p ER_{i,t+\tau}$. As another example suppose that power utility is used and hence $U(\hat{R}_{i,t+\tau}) = \hat{R}^{1-\rho}_{i,t+\tau}/(1-\rho)$. Since $\hat{\Phi}$ is defined as a root and $U(.)$ is continuously differentiable in its argument, we obtain $\hat{\Phi} = (\bar{U}(\hat{R}_1) - \bar{U}(\hat{R}_0))/\tilde{P}^{-1}\sum_{t=T}^{T+P-\tau} \partial U(\hat{R}_{1,t+\tau} - \tilde{\Phi})/\partial\Phi$ some $\tilde{\Phi}$ on the line between $\hat{\Phi}$ and 0.[4] If $\tilde{P}^{-1}\sum_{t=T}^{T+P-\tau} \partial U(\hat{R}_{1,t+\tau} - \tilde{\Phi})/\partial\Phi \to^p E(\partial U(R_{1,t+\tau})/\partial\Phi) \neq 0$, Assumption 1 is satisfied with

$$f_{t+\tau}(\hat{\beta}_t) = (\hat{R}^{1-\rho}_{1,t+\tau} - \hat{R}^{1-\rho}_{0,t+\tau})/(E(\partial U(R_{i,t+\tau})/\partial\Phi)(1-\rho)).$$

The requirement that $f_{t+\tau}(\beta)$ is twice continuously differentiable in β is nontrivial for our results. In the vast majority of studies on economic value calculations, the utility function $U(\cdot)$ itself is continuously differentiable in the gross return. Even so, there are cases where the assumption might

[4] When $\Phi = 0$.

fail because $\hat{R}_{i,t+\tau}$ is not twice continuously differentiable in β. For example, in some applications the estimated portfolio weights can be bounded (or winsorized) to limit the maximal amount of leverage in the constructed portfolio (Ferreira and Santa Clara, 2011). We pursue this issue further in simulations reported in section 5.

Given Assumption 1, Assumptions 2, 3, and 4 are nearly identical to those in West (1996). The predictive models must be parametric and estimated using a framework that can be mapped into GMM. The observables must have sufficient moments and mixing conditions to satisfy a central limit theorem,[5] and the number of in-sample and out-of-sample observations must be of the same order. Assumption 5 states that if the two models are nested under the null hypothesis, and hence $R_{1,t+\tau} = R_{0,t+\tau}$, it must be the case that a certain product of two moments is nonzero. We do so since, as shown below, there is the potential for the asymptotic variance of $\tilde{P}^{1/2}\hat{\Phi}$ to be zero. To avoid this problem, we state a high-level assumption that ensures that the asymptotic variance is nonzero. As a practical matter, the condition is likely to hold as long as the model parameters are not estimated using the utility function $U(\cdot)$ as the objective function.[6] Given the assumptions, our main result follows immediately from Theorem 4.1 of West (1996).

Theorem 1 *Maintain Assumptions 1-5 and let* $\Phi = 0$. $\tilde{P}^{1/2}\hat{\Phi} \to^d N(0,\Omega)$ *with*

$$\Omega = S_{ff} + 2\Lambda_0(\pi)FBS'_{fh} + \Lambda_1(\pi)FBS_{hh}B'F'$$

where $S_{ff} = \lim_{T\to\infty} Var\left(T^{-1/2}\sum_{s=1}^{T} f_{s+\tau}\right)$, $S_{hh} = \lim_{T\to\infty} Var\left(T^{-1/2}\sum_{s=1}^{T} h_{s+\tau}\right)$, $S_{fh} = \lim_{T\to\infty} Cov\left(T^{-1/2}\sum_{s=1}^{T} f_{s+\tau}, T^{-1/2}\sum_{s=1}^{T} h_{s+\tau}\right)$, *and*

scheme	$\Lambda_0(\pi)$	$\Lambda_1(\pi)$
recursive	$(1-\pi^{-1}\ln(1+\pi))$	$2(1-\pi^{-1}\ln(1+\pi))$
rolling $0<\pi\leq 1$	$\frac{\pi}{2}$	$\pi - \frac{\pi^2}{3}$
rolling $1\leq\pi<\infty$	$1-\frac{1}{2\pi}$	$1-\frac{1}{3\pi}$.

.

Theorem 1 shows that on the boundary of the null hypothesis, the estimated performance fee is asymptotically normal with zero mean and an asymptotic variance that reflects not only variation

[5] West uses the central limit theorem of Wooldridge and White (1998).

[6] This assumption precludes a few isolated applications including Skouras (2007) and Cenesizoglu and Timmermann (2011).

in the difference in utilities, via S_{ff}, but also the influence of the estimation error via the remaining components of the variance, FBS'_{fh} and $FBS_{hh}B'F'$, respectively.

When the models are nested there are two distinct cases in which Theorem 1 applies. The leading case is when $\Phi = 0$ because the models are identical and hence the "competing" and "baseline" models are better thought of as "unrestricted" and "restricted." In this situation, there exists a selection matrix J such that $J\beta_1^* = \beta_0^*$, and the asymptotic variance simplifies to

$$\Omega = \Lambda_1(\pi)(E\frac{\partial U_{1,t+\tau}}{\partial \Phi})^{-2}E(\frac{\partial U_{1,t+\tau}}{\partial \beta_1})(-JB_0J + B_1)S_{h_1h_1}(-JB_0J + B_1)E(\frac{\partial U'_{1,t+\tau}}{\partial \beta_1}). \tag{5}$$

The asymptotic variance simplifies because in this specific case, $R_{1,t+\tau} = R_{0,t+\tau}$ for all t and hence S_{ff} and S_{fh} are both trivially zero. In addition, the fact that the models are nested implies $J'h_{1,t+\tau} = h_{0,t+\tau}$ and $J'\frac{\partial U_{1,t+\tau}}{\partial \beta_1} = \frac{\partial U_{0,t+\tau}}{\partial \beta_0}$ and thus the asymptotic variance can be simplified even further. While not immediately obvious in Theorem 1, it is for this case that we have Assumption 5. To achieve asymptotic normality in any useful sense we need Ω to be positive. When predictability exists and hence $\beta_{1,1}^*$ is not zero, S_{ff} is nonzero and hence Ω is nonzero as long as S_{ff} does not cancel with the remaining terms. By imposing Assumption 5, we ensure that Ω is positive for the case in which the unrestricted model perfectly nests the restricted model – a possibility that we must allow for under the null hypothesis.

A less intuitive case arises when the models are ostensibly nested and yet $\Phi = 0$ despite the fact that $\beta_{1,1}^*$ is nonzero. In this case, as stated in Theorem 1, $\hat{\Phi}$ is asymptotically normal with mean zero but with an asymptotic variance that does not simplify as it does when $\beta_{1,1}^* = 0$. Note also that our results are applicable to the comparison of two models that are non-nested under the null. While theoretically plausible, such a comparison does not appear to exist in the literature and hence we do not pursue it any further herein.

4 INFERENCE

The theorem from the previous section provides a means of assessing the statistical significance of performance fees for a given utility function. Specifically, if $\hat{\Omega}$ is a consistent estimate of Ω,

it immediately follows that $\tilde{P}^{1/2}\hat{\Phi}/\hat{\Omega}^{1/2} \rightarrow^{d} N(0,1)$, and therefore one can use standard normal critical values to conduct a conservative test of the null hypothesis that $\Phi \leq 0$ against an alternative in which $\Phi > 0$. In our simple illustrative example discussed in Section 2, forming a consistent estimate of Ω is not too difficult. However, the standard errors become increasingly complicated to estimate as the number of risky assets increases and we move from mean-variance utility to other utility functions.

In this section, we therefore discuss two approaches to inference: one in which we estimate Ω and use standard normal critical values for inference and one in which critical values are obtained by bootstrapping the distribution of $\hat{\Phi}$ without any normalization. To facilitate application of our results and to emphasize some peculiar features of our results, we delineate both approaches in the context of the simple environment discussed in Section 2.

Recall that since utility is mean/variance, the percentage of wealth invested in the risky asset takes the form

$$w_{i,t}(\hat{\beta}_{i,t}) = \hat{w}_{i,t} = \frac{ep_{t+1}\left(\hat{\alpha}_{i,t}\right)}{\gamma\sigma^2_{i,t+1}\left(\hat{\vartheta}_{i,t}\right)}$$

for a known RRA parameter γ, conditional mean prediction $ep_{t+1}\left(\hat{\alpha}_{i,t}\right)$, and conditional variance prediction $\sigma^2_{i,t+1}\left(\hat{\vartheta}_{i,t}\right)$. The baseline model 0 uses the historical mean of ep_{t+1} $(\alpha_{0,0})$ as the conditional mean prediction and the historical unconditional variance of ep_{t+1} as the conditional variance prediction. The competing model 1 forms the conditional mean prediction using the OLS-estimated regression $ep_{t+1} = \alpha_{1,0} + \alpha_{1,1}z_t + e_{1,t+1} = \alpha_1' x_t + e_{1,t+1}$ and also uses the historical unconditional variance of ep_{t+1} as the conditional variance prediction.

We allow for three distinct methods of estimating each of these predictions. Under the recursive scheme, both α_i and ϑ_i $i = 0, 1$ are estimated using all available observations from $i = 1, ..., t$ for each forecast origin $t = T, ..., T + P - 1$. Under the rolling scheme, α_i and ϑ_i $i = 0, 1$ are estimated using only the most recent T observations from $i = t - T + 1, ..., t$ for each forecast origin $t = T, ..., T + P - 1$. Under a "mixed" scheme, used by Goyal and Welch (2008), α_i is estimated using the recursive scheme, while ϑ_i is estimated using a small rolling window of size $M << T$. The

first two schemes align with the maintained assumptions of Section 3 and, in particular, define $\hat{\beta}_{i,t}$ as $(\hat{\alpha}'_{i,t}, \hat{\vartheta}'_{i,t})'$. The latter *may* align with the maintained Assumptions but only after reinterpreting the variance parameter $\hat{\vartheta}_{i,t} = M^{-1}\sum_{s=t-M+1}^{t}(ep_s - \bar{ep}_{t,M})^2$, where $\bar{ep}_{t,m}$ denotes the sample mean of ep_s over the sample $s = t-M+1, ..., t$. Since $M << T$, it is not unrealistic to interpret the parameter estimates in the manner put forth by Giacomini and White (2006). There, asymptotics are developed whereby *all* parameters are assumed to be estimated using a rolling window of fixed and finite length M. By taking this approach, they effectively treat the parameter estimates as just another time series of observables in much the same way as we treat ep_{t+1} and z_t as observables. If we take this view of the rolling window estimator of the unconditional variance, *but* still estimate the conditional mean parameters using the recursive scheme, our theoretical results continue to be applicable with two distinctions: (i) Contributions to Ω due to estimation error arise only via $\hat{\alpha}_{i,t}$ since we have de facto redefined $\hat{\beta}_{i,t}$ as just $\hat{\alpha}_{i,t}$ and (ii) the proofs treat $\hat{\vartheta}'_{i,t}$ as just a component of the function $f_{t+\tau}$ that continues to satisfy, in particular, Assumption 3.

4.1 Asymptotic Critical Values

Since Φ can be zero regardless of whether $\alpha_{1,1}$ is zero, for robust inference we need a consistent estimate of Ω that does not require $\alpha_{1,1}$ to take any particular value. This can be achieved by estimating every component of Ω using the formula in the theorem. Under the recursive or rolling schemes, for which $h_{s+1} = (e_{0,s+1}, (ep_{s+1} - Eep_{s+1})^2 - E(ep_{s+1} - Eep_{s+1})^2, e_{1,s+1}, z_s e_{1,s+1}, (ep_{s+1} - Eep_{s+1})^2 - E(ep_{s+1} - Eep_{s+1})^2)'$, these elements take the form $B = diag(B_0, B_1)$, $B_0 = I_2$, $B_1 = diag((Ex_t x'_t)^{-1}, 1)$, and $F = (-E\frac{\partial U'_{0,t+1}}{\partial \beta_0}, E\frac{\partial U'_{1,t+1}}{\partial \beta_1})'$ where $E\frac{\partial U_{i,t+1}}{\partial \beta_i} = E\frac{\partial w_{i,t+1}}{\partial \beta_i} ep_{t+1}(1 - \gamma(R_{i,t+1} - ER_{i,t+1}))$ $i = 0, 1$. Under the mixed scheme, for which $h_{s+1} = (e_{0,s+1}, e_{1,s+1}, z_s e_{1,s+1})'$, these elements take the form $B = diag(B_0, B_1)$, $B_0 = 1$, $B_1 = (Ex_t x'_t)^{-1}$, and $F = (-E\frac{\partial U'_{0,t+1}}{\partial \beta_0}, E\frac{\partial U'_{1,t+1}}{\partial \beta_1})$, where $E\frac{\partial U_{i,t+1}}{\partial \beta_i} = E\frac{\partial w_{i,t+1}}{\partial \beta_i} ep_{t+1}(1 - \gamma(R_{i,t+1} - ER_{i,t+1}))$ $i = 0, 1$ but with β_i appropriately redefined as just α_i.

In either case, as discussed in West (1996), sample analogs can be used to estimate each component of Ω. Under the recursive or rolling schemes, consistent estimates of B, F, and both

$\Lambda_0(\pi)$ and $\Lambda_1(\pi)$ can be formed using $\hat{B} = diag(I_2, diag(((T+P)^{-1}\sum_{t=1}^{T+P} x_t x_t')^{-1}, 1))$, $\hat{F} = (-P^{-1}\sum_{t=T}^{T+P-1} \frac{\partial \hat{w}_{0,t+1}}{\partial \beta_0} e\hat{p}_{t+1}(1-\gamma(\hat{R}_{0,t+1}-\bar{R}_0)), P^{-1}\sum_{t=T}^{T+P-1} \frac{\partial \hat{w}_{1,t+1}}{\partial \beta_1} e\hat{p}_{t+1}(1-\gamma(\hat{R}_{1,t+1}-\bar{R}_1)))'$, $\Lambda_0(P/T)$, and $\Lambda_1(P/T)$ respectively. Estimates of S_{ff}, S_{fh}, and S_{hh} can be constructed using conventional nonparametric kernel estimators of the long-run variances and covariances of h_{t+1} and f_{t+1} if we define $\hat{h}_{t+1} = (\hat{e}_{0,t+1}, (e\hat{p}_{t+1} - e\bar{p})^2 - \hat{\sigma}^2_{ep}, \hat{e}_{1,t+1}, z_t\hat{e}_{1,t+1}, (e\hat{p}_{t+1} - e\bar{p})^2 - \hat{\sigma}^2_{ep})'$ and $\hat{f}_{t+1} = (\hat{R}_{1,t+1} - \frac{\gamma}{2}(\hat{R}_{1,t+1} - \bar{R}_1)^2) - (\hat{R}_{0,t+1} - \frac{\gamma}{2}(\hat{R}_{0,t+1} - \bar{R}_0)^2)$. Under the mixed scheme, a consistent estimate of B takes the form $\hat{B} = diag(1, diag(((T+P)^{-1}\sum_{t=1}^{T+P} x_t x_t')^{-1}))$. The estimate of F changes only insofar as $\frac{\partial \hat{w}_{i,t+1}}{\partial \beta_i}$ $i = 0, 1$ is adjusted to account for the fact that $\hat{\vartheta}_{i,t}$ is treated as an observable rather than a component of $\hat{\beta}_{i,t}$. Finally, consistent estimates of S_{ff}, S_{fh}, and S_{hh} can be constructed similarly if we redefine $\hat{h}_{t+1}$ as $(\hat{e}_{0,t+1}, \hat{e}_{1,t+1}, z_t\hat{e}_{1,t+1})'$. It is important to note that in general, both h_{s+1} and f_{s+1} will be serially correlated even if $e_{1,s+1}$ is a martingale difference sequence. This occurs when $\alpha_{1,1}$ is not zero and hence the error term in model 0 ($e_{0,s+1}$) exhibits serial correlation induced by the omitted regressor z_s.

4.2 Bootstrap Critical Values

The bootstrap used in this paper is consistent with one developed by Calhoun (2015). We have chosen this specific bootstrap since it is explicitly designed to be applicable in cases when out-of-sample methods are used to conduct inference and the relevant test statistic is asymptotically normal. Furthermore, the bootstrap is designed to allow for estimation error in the standard errors. This final feature is nontrivial for our results. To see this, recall that the theorem implies that $\hat{\Phi}$ is asymptotically normal with an asymptotic variance that is a linear combination of three elements: S_{ff}, FBS_{fh}, and $FBS_{hh}B'F'$. The latter two of these terms exist because the portfolio weights are functions of estimated parameters and are not known a priori. As a consequence, accounting for estimation error in the standard errors is crucial in this context.

Let X_t denote the vector of time t observables. In the simple example discussed in Section 2, this consists of $X_t = (ep_t, z_t)'$. The first stage of the bootstrap consists of using the moving blocks, circular blocks, or stationary bootstrap with block lengths drawn from the geometric distribution

to generate the time series of bootstrapped observations $X_1^*...X_{T+P}^*$. In each method the block length l satisfies $l/P \to 0$ as $l, P \to \infty$. In the second stage, the bootstrapped data are used to construct the bootstrapped performance fee measure $\hat{\Phi}^*$. This process is repeated many times so that we have a collection of $\hat{\Phi}_j^*$ $j = 1, ..., N$ of bootstrapped performance fees that can be used to estimate asymptotically valid critical values.

As in White (2000), we recenter each of the bootstrapped $\hat{\Phi}_j^*$ statistics to ensure that the empirical critical values remain bounded under the alternative hypothesis, $\Phi > 0$. Unfortunately, as shown in Corradi and Swanson (2007), White's approach to recentering does not allow for estimation error in the asymptotic distribution. The issue is that White recommends recentering each $\hat{\Phi}_j^*$ by the empirical value of the performance fee $\hat{\Phi}$, which is a function of an entire sequence of parameter estimates $\hat{\beta}_t$, $t = T..., T + P - \tau$. In contrast, Calhoun recommends recentering in one of two ways: (i) by the constant $\bar{\Phi} = \hat{\Phi}(\hat{\beta}_{T+P-\tau})$, which is constructed exactly as was $\hat{\Phi}$ but where $\hat{\beta}_t = \hat{\beta}_{T+P-\tau}$ for all $t = T..., T + P - \tau$ or (ii) by the constant $\bar{\Phi} = N^{-1} \sum_{i=j}^{N} \hat{\Phi}_j^*$. Given the recentering constant $\bar{\Phi}$ and the bootstrapped performance fees $\hat{\Phi}_j^*$, critical values are then estimated based on the empirical distribution of $\hat{\Phi}_j^* - \bar{\Phi}$. After some experimentation in our simulations, we found that the second of the two centering methods performed best and is therefore uses in both our Monte Carlo and empirical evidence.

5 MONTE CARLO EVIDENCE

In this section, we provide Monte Carlo evidence on the finite sample properties of the asymptotic results. Specifically, we provide simulation evidence on the efficacy of both asymptotic and bootstrap approaches to testing the null hypothesis $H_0 : \Phi \leq 0$ against the alternative $H_A : \Phi > 0$. In our experiments, we consider the problem of a US investor who faces the problem of choosing how to optimally allocate her wealth between the value-weighted index of stocks traded on the NYSE and the 3-month T-bill. We assume that the investor is endowed with mean-variance preferences and the performance fees are computed using a mean-variance utility function, as outlined in Sec-

tion 2. To be clear, our experiments are not designed to perfectly match the data. Rather, they are intended to delineate the difficulties associated with inference when multiple roots of Φ exist in an environment that is not too dissimilar from the one we discuss in the empirical section of this paper.

5.1 Experiment Design

The experiments are conducted as follows. For all cases, we use a data-generating process loosely calibrated on the empirical properties of excess returns to the NYSE value-weighted index and its dividend yield z_t both at a monthly frequency:

$$\begin{aligned} ep_{t+1} &= 0.05 + b(c_b^{1/2} z_t) + (c_b^{1/2} u_{t+1}) h_t^{1/2} \\ z_t &= 0.05 + 0.9 z_{t-1} + v_t \\ h_t &= 0.1 + 0.4 h_{t-1} + 0.05 u_{t-1}^2 + z_{t-1}^{-2} \\ r_{t+1}^f &= 0.036 \end{aligned} \tag{6}$$

with

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} \sim i.i.d.N\left(0, \begin{bmatrix} 0.1716 & -0.0038 \\ -0.0038 & 0.0012 \end{bmatrix}\right).$$

With an eye toward our empirical section, in which we consider annual, quarterly, and monthly frequency data, we primarily consider three overall sample sizes $T + P = 88$, 352, and 1056. A few additional results are given for a larger sample size $T + P = 2112$.[7] For each we consider three sample splits: $P/T = 1/3$, 1, and 3. Conditional heteroskedasticity is modeled using a GARCH-X form in the equity premium equation.[8] The risk aversion parameter γ is set to 3. For brevity, we focus exclusively on the recursive and mixed schemes. For the mixed scheme we set the rolling window size, used to estimate the conditional variance, at values indicative of a 5-year window and

[7] We use a burnout period of 500 observations to remove the effects of initial conditions.

[8] In the current simulations, Assumption 5 requires that $FB = E(\frac{\partial U_{1,t+\tau}}{\partial \beta_1})(-JB_0J + B_1)$ is nonzero. Straightforward algebra reveals that Assumption 5 fails if ep_{t+1} and ep_{t+1}^2 are *both* mean independent of $(1, z_t)$. It is for this reason that we introduce conditionally heteroskedastic errors. The issue is relevant only when $\alpha_{1,1} = 0$. In unreported results for which $\alpha_{1,1} = 0$ and the errors are conditionally homoskedastic, the rejection frequencies are less than 1 percent for nominally 5 percent tests.

hence take the values 5, 20, 60, and 120 depending on the sample size. To facilitate interpretation of our results, the data-generating process for the equity premium ep is designed so that for all values of b, the unconditional variance of ep takes the constant value $c_0 = Var(u_{t+1})E(h_t)$. We achieve this through the scaling factor $c_b = c_0/(b^2 Var(z_t) + c_0)$.

For all size and power experiments we consider both asymptotic and bootstrap-based critical values. For asymptotic critical values, we estimate the standard errors using the formulas described in Section 4. Newey-West (1987) HAC estimators were used to estimate S_{ff}, S_{hh}, and S_{fh} with a lag length fixed at 4. For bootstrap critical values, we use the stationary block-bootstrap approach described in Section 4 with the block length set at $(T+P)^{0.6}$ and the number of bootstrap replications set at 299. For insight on the impact winsorizing may have on our asymptotics, we conduct all experiments twice: once with and once without winsorized portfolio weights. In the experiments with winsorized weights, we restrict them to lie between -0.5 and 1.5. All results are based on 2,000 Monte Carlo replications.

Our experiment design is intended to not only emulate data from our empirical work but also to exemplify the the difficulty of conducting inference when Φ has multiple roots. For nonnegative values of b, unreported simulation evidence indicates that under the recursive scheme Φ is hump-shaped in b and takes the value of zero when $b = 0$ and $b = 2.35$ in much the same way seen in Figure 1.[9] When the mixed scheme is used, the presence of a hump shape now depends on the value of M. As M increases from 20, to 60, to 120, the nonzero roots of Φ increase from 0.8, to 1.7, to 2.1. When $M = 5$, Φ has a unique root when $b = 0$ and is negative for all nonzero values of b.

5.2 Size Results

Tables 1 and 2 report the actual rejection frequencies of nominally 5% tests in experiments in which $\Phi = 0$. Table 1 corresponds to rejection frequencies when asymptotic critical values are used, while Table 2 corresponds to rejection frequencies when bootstrap critical values are used.

[9] The formula for Φ in section 2 is not applicable due to the introduction of conditionally heteroskedastic errors.

Within each table the columns are associated with various sample splits, roots of Φ, and presence of winsorization. The rows correspond to various estimation schemes and sample sizes.

Consider the left-hand side of Table 1, within which $b = 0$ and the portfolio weights are not winsorized. In most cases the test is undersized. At the smaller sample sizes, rejection frequencies are near 1%. As the sample size increases, the test is typically more accurately sized approaching 3% when $T + P = 1056$ and 4% for the largest sample size. The asymptotics appear to work better under the recursive scheme than under the mixed scheme for which the undersizing is more acute. There is some variation across sample splits: For the smaller samples, size is a little better for smaller out-of-sample periods, while the opposite is true for the larger sample sizes. There are very few changes in rejection frequencies when we winsorize the portfolio weights and in all instances the differences occur under the mixed scheme.

On the right side of Table 1, within which $b > 0$, the rejection frequencies are sometimes less and sometimes greater than 5%. Under the recursive scheme, the rejection frequencies range from 0.028 to 0.086. Under the mixed scheme the range is wider, ranging from 0.003 to 0.157. As before, winsorization has little effect on the rejection frequencies.

In Table 2 we report rejection frequencies when bootstrap-based critical values are used and restrict attention to the case in which $P/T = 1$. In all cases for which $b = 0$, the bootstrap rejection frequencies are more accurate than those associated with asymptotic critical values from Table 1. When $b > 0$, the bootstrap generally helps under the mixed scheme, bringing the rejection frequencies closer to 5%. Under the recursive scheme, when $b > 0$ the rejection frequencies are a bit higher relative to those found in Table 1 but still remain below 10%.

5.3 Power Results

Figures 2 and 3 plot rejection frequencies of nominal 5% tests over a range of values for the tuning parameter b. In particular, we allow b to increase from zero up to, and a bit beyond, the value of b for which Φ has its second root as delineated in the size experiments. For brevity, we restrict attention to sample splits for which $P/T = 1$ and cases for which the portfolio weights are not

winsorized. As a benchmark, Figure 2 provides rejection frequencies for a test of equal MSE when we allow b to increase from zero to 2.585 in increments of 0.235 – a range of values for which Φ has two roots under the recursive scheme.[10] As expected, rejection frequencies increase monotonically with b and the sample size.

Figure 3 reports rejection frequencies associated with the performance fee measure. There are four distinct plots, each associated with some permutation of the recursive or mixed scheme and asymptotic or bootstrap critical values. Note that for the mixed scheme there are two scales for the horizontal axis: one for the largest sample size and one for the two smaller sample sizes. There are a number of items of note.

- In each plot, the path of rejection frequencies is very different than that found in Figure 2. In many cases, there is a clear hump-shaped pattern that is particularly clear under the recursive scheme. As one might expect given our example in Figure 1, the hump arises from the fact that there are two nonnegative roots for Φ and, moreover, between these two roots Φ is positive but beyond the largest root, Φ is negative.

- For the largest sample sizes, power can be substantial. But for the smallest sample sizes, especially when $T + P = 88$, power is essentially nonexistent and in some instances the actual power of the test is lower than the nominal size of the test.

- While a direct comparison is invalid because the mixed and recursive schemes have different roots for Φ, power appears to be a bit stronger under the recursive scheme than under the mixed scheme.

- For a large enough sample size, bootstrap-based inference provides reasonable power. Even so, the bootstrap approach to inference seems to come at a small loss in power relative to the asymptotic approach that estimates the asymptotic variance directly and uses standard normal critical values.

[10] We use the MSE-t test as delineated in McCracken (2007) and its associated critical values.

6 THE ECONOMIC VALUE OF PREDICTIONS OF THE US EQUITY PREMIUM

In this section, we use the testing procedure discussed in the previous sections to revisit the findings of the recent literature on the predictability of the US equity premium. Methodologically our framework is identical to the one highlighted in Section 2. We use monthly value-weighted returns from the S&P 500 index from January 1927 to December 2011 from the Centre for Research in Security Prices and Robert Shiller's website. Stock returns are continuously compounded including dividends, and the predictive variables z_t are a selection of 14 variables from the ones used by Goyal and Welch (2008, and additional appendix).[11] We begin forecasting in January 1965 and continue through December 2011, giving us $P = 564$ out-of-sample observations out of a total of $T + P = 1692$.

Following a conventional practice used in several studies (see, among others, Ferreira and Santa-Clara, 2011; and the references therein), we compute the weights $\hat{w}_{i,t}$ using the mixed scheme: A small rolling window of 5 years' worth of past observations is used to construct an estimate of the conditional variance, while the recursive scheme is used to estimate the conditional means. The coefficient of relative risk aversion is set to 3 (Goyal and Welch, 2008; Campbell and Thompson, 2008; and Ferreira and Santa Clara, 2011). We compute the portfolio weights both unconstrained and winsorized by imposing a maximum value of the investment in the risky asset to 150% (i.e., $-0.5 \leq \hat{w}_{i,t} \leq 1.5$).

The results of our empirical exercise are reported in Table 3 for three different data frequencies: monthly, quarterly, and annual in Panels A, B and C, respectively. For all data frequencies the p-values of the null hypothesis are reported for both unconstrained and winsorized portfolio weights and they are computed using the asymptotic and bootstrap critical values as discussed in Sections 4.1 and 4.2 of the main text. The bootstrapped critical values are computed using 999 replications

[11] For further details on data construction, refer to Goyal and Welch (2008, and additional appendix). The full dataset used in this empirical exercise can be downloaded from Amit Goyal's website (http://www.bus.emory.edu/agoyal/Research.html).

and a sample-size-dependent block length like that used in the Monte Carlo section.

The results in all panels of Table 3 suggest unambiguously that none of the predictive variables in our sample is able to generate economic values from forecasting that are statistically different from the one provided by the no-predictability benchmark. Put differently, the positive performance fees recognizable for some predictors across data frequencies are found to be statistically indistinguishable from zero when our testing procedure is applied. The results are confirmed using both asymptotic and bootstrap critical values. The only exception is represented by the term spread that is able to generate positive performance fees at the monthly and quarterly frequency. However, the statistical significance is only at the 10% level at the quarterly frequency and is not confirmed when bootstrap critical values are computed.

Overall, the results reported in this simple exercise reinforce the primary point of this paper. The mere evidence of a positive estimated performance fee does not provide conclusive evidence of superiority of a given predictive model against a given (in this case, no-predictability) benchmark. All panels of Table 3 provide examples of estimated performance fees that are positive and yet are not significantly different from zero based on a standard z-score-based approach to inference. Since this level of rigor is commonly used when reporting evidence of statistical (MSE-based) predictability, it seems only fitting to do the same when reporting evidence of economic (performance-fee-based) predictability. This is particularly true given that the data used to construct the MSEs are the same as those used to construct the performance fees!

7 CONCLUSION

Out-of-sample methods are a common approach to evaluating the predictive content of a model. As such, a healthy literature has developed that provides methods for conducting inference on measures of forecast accuracy. This literature is almost completely focused on statistical measures. Economic measures of predictive accuracy are becoming increasingly common and are used to complement the evidence provided by statistical measures. In this paper, we derive asymptotics

that can be used to conduct inference on one economic measure of forecast quality – performance fees. In particular, building on the theoretical results in West (1996), we are able to establish that these performance fee measures are asymptotically normal with an asymptotic variance that is affected by parametric estimation error. Monte Carlo evidence suggests that the theoretical results can be useful but also suggest that large samples are sometimes required.

8 APPENDIX

To understand the results reported in Section 3, it is instructive to consider the forms of $\hat{\Phi}$ and $f_{t+\tau}(\hat{\beta}_t)$ for three commonly used functional forms for utility: mean-variance, quadratic, and power. In addition, we also characterize the moment F when the two models are nested under the null hypothesis and hence $\beta^*_{1,1} = 0$.

(i) When utility is mean-variance, the average utility obtained using model $i = 0, 1$ is

$$\bar{U}(\hat{R}_i) = \bar{R}_i - \frac{\gamma}{2}\tilde{P}^{-1}\sum\nolimits_{t=T}^{T+P-\tau}(\hat{R}_{i,t+\tau} - \bar{R}_i)^2,$$

where $\bar{R}_i = \tilde{P}^{-1}\sum_{t=T}^{T+P-\tau}\hat{R}_{i,t+\tau}$ and γ is a known preference parameter. For this functional form, we trivially obtain

$$\hat{\Phi} = \bar{U}(\hat{R}_1) - \bar{U}(\hat{R}_0).$$

As stated in the text, for this utility function Assumption 1 is satisfied for the function

$$f_{t+\tau}(\hat{\beta}_t) = (\hat{R}_{1,t+\tau} - \frac{\gamma}{2}(\hat{R}_{1,t+\tau} - ER_{1,t+\tau})^2) - (\hat{R}_{0,t+\tau} - \frac{\gamma}{2}(\hat{R}_{0,t+\tau} - ER_{0,t+\tau})^2)$$

if $\bar{R}_i \to^p ER_{i,t+\tau}$. When the models are nested, straightforward algebra implies

$$F = (-E\frac{\partial U'_{0,t+\tau}}{\partial \beta_0}, E\frac{\partial U'_{1,t+\tau}}{\partial \beta_1})',$$

where $E\frac{\partial U_{i,t+\tau}}{\partial \beta_i} = E\frac{\partial w_{i,t+\tau}}{\partial \beta_i} ep_{t+\tau}(1 - \gamma(R_{i,t+\tau} - ER_{i,t+\tau}))$ $i = 0, 1$.

(ii) When utility is quadratic, the average utility obtained using models $i = 0, 1$ takes the similar but distinct form

$$\bar{U}(\hat{R}_i) = \bar{R}_i - \gamma\tilde{P}^{-1}\sum\nolimits_{t=T}^{T+P-\tau}\hat{R}^2_{i,t+\tau}.$$

For this utility function, there are actually two roots that satisfy the definition of Φ. If we use the larger of the two as our estimate of Φ, we obtain the following closed form for the performance fee:

$$\hat{\Phi} = (\bar{R}_1 - (2\gamma)^{-1}) + [(\bar{R}_1 - (2\gamma)^{-1})^2 + \gamma^{-1}(\bar{U}(\hat{R}_1) - \bar{U}(\hat{R}_0))]^{1/2}.$$

For this utility function Assumption 1 is satisfied for the function

$$f_{t+\tau}(\hat{\beta}_t) = ((\hat{R}_{1,t+\tau} - \gamma\hat{R}^2_{1,t+\tau}) - (\hat{R}_{0,t+\tau} - \gamma\hat{R}^2_{0,t+\tau}))/(-1 + 2\gamma E(R_{1,t+\tau}))$$

if $-1+2\gamma\bar{R}_1 \to^p -1+2\gamma E(R_{1,t+\tau}) \neq 0$. When the models are nested, straightforward algebra implies

$$F = (-E\frac{\partial w'_{0,t+\tau}}{\partial \beta_0} ep_{t+\tau}(1 - 2\gamma R_{0,t+\tau}), E\frac{\partial w'_{1,t+\tau}}{\partial \beta_1} ep_{t+\tau}(1 - 2\gamma(R_{1,t+\tau})))'/(-1 + 2\gamma E(R_{1,t+\tau})).$$

(iii) When utility is power, the average utility obtained using model $i = 0, 1$ is

$$\bar{U}(\hat{R}_i) = \frac{\tilde{P}^{-1}\sum_{t=T}^{T+P-\tau} \hat{R}^{1-\gamma}_{i,t+\tau}}{1-\gamma}.$$

For this functional form of utility, we do not obtain a closed form for the performance fee $\hat{\Phi}$. Estimating $\hat{\Phi}$ is done numerically using the definition and hence we have

$$\hat{\Phi} = \arg_\Phi root(\frac{\tilde{P}^{-1}\sum_{t=T}^{T+P-\tau}(\hat{R}_{1,t+\tau} - \Phi)^{1-\gamma}}{1-\gamma} - \bar{U}(\hat{R}_0)).$$

As stated in the text, for this utility function, Assumption 1 is satisfied for the function

$$f_{t+\tau}(\hat{\beta}_t) = (\hat{R}^{1-\rho}_{1,t+\tau} - \hat{R}^{1-\rho}_{0,t+\tau})/(E(\partial U(R_{1,t+\tau})/\partial\Phi)(1-\rho))$$

if $\tilde{P}^{-1}\sum_{t=T}^{T+P-\tau} \partial U(\hat{R}_{1,t+\tau} - \tilde{\Phi})/\partial\Phi \to^p E(\partial U(R_{1,t+\tau})/\partial\Phi) \neq 0$. When the models are nested, straightforward algebra implies

$$F = (-E\frac{\partial w_{0,t+\tau}}{\partial \beta_0} ep_{t+\tau} R^{-\gamma}_{0,t+\tau}, E\frac{\partial w_{1,t+\tau}}{\partial \beta_1} ep_{t+\tau}(R_{1,t+\tau})^{-\gamma})'/E(R_{1,t+\tau})^{-\gamma}.$$

9 REFERENCES

Barberis, N. (2000), "Investing for the Long Run When Returns Are Predictable," *Journal of Finance*, 55, 225-264.

Boudoukh, J., Richardson, M., and Whitelaw, R. (2008), "The Myth of Long-Horizon Predictability," *Review of Financial Studies*, 24, 1577-1605.

Calhoun, G. (2015), "A Simple Block Bootstrap for Asymptotically Normal Out-of-Sample Test Statistics," Iowa State University working paper.

Campbell, J.Y., and Thompson, S.B. (2008), "Predicting Excess Returns Out of Sample: Can Anything Beat the Historical Average?" *Review of Financial Studies*, 21, 1509-1531.

Cenesizoglu, T., and Timmermann, A. (2011), "Do Return Prediction Models Add Economic Value?" UCSD working paper.

Clark, T.E., and McCracken, M.W. (2001), "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics*, 105, 85-110.

Cochrane, J.H. (1999), "New Facts in Finance," *Economic Perspectives*, 23, 36-58.

Cochrane, J.H. (2008), "The Dog That Did Not Bark: A Defense of Return Predictability," *Review of Financial Studies*, 21, 1533-1575.

Cochrane, J.H. (2011), "Presidential Address: Discount Rates," *Journal of Finance*, 66, 1047-1108.

Corradi, V., and Swanson, N.R. (2007), "Nonparametric Bootstrap Procedures for Predictive Inference Based on Recursive Estimation Schemes," *International Economic Review*, 48, 67-109.

Ferreira, M.A., and Santa Clara, P. (2011), "Forecasting Stock Market Returns: The Sum of the Parts Is More than the Whole," *Journal of Financial Economics*, 100, 514-537.

Fleming, J., Kirby, C., and Ostdiek, B. (2001), "The Economic Value of Volatility Timing," *Journal of Finance*, 56, 329-352.

Giacomini, R., and White, H. (2006), "Tests of Conditional Predictive Ability," *Econometrica*, 74, 1545-1578.

Goyal, A., and Welch, I. (2008), "A Comprehensive Look at the Empirical Performance of Equity Premium Predictions," *Review of Financial Studies*, 21, 1455-1508; see additional appendix at http://www.hec.unil.ch/agoyal/docs/AllTables2009.pdf.

Hansen, P.R. (2005), "A Test for Superior Predictive Ability," *Journal of Business and Economic Statistics,* 23, 365-380.

Ingersoll, J., Speigel, M., Goetzmann, W., and Welch, I. (2007), "Portfolio Performance Manipulation and Manipulation-proof Performance Measures," *Review of Financial Studies*, 20, 1503-1546.

Leitch, G., and Tanner, J.E. (1991), "Economic Forecast Evaluation: Profits Versus the Conventional Measures," *American Economic Review*, 81, 580-590.

Lettau, M., and Ludvigson, S. (2001),"Consumption, Aggregate Wealth and Expected Stock Returns," *Journal of Finance*, 56, 815-849.

McCracken, M.W. (2007), "Asymptotics for Out-of-Sample Tests of Granger Causality," *Journal of Econometrics*, 140, 719-752.

Newey, W.K., and West, K.D. (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703-708.

Patton, A.J. (2004), "On the Out-of-Sample Importance of Skewness and Asymmetric Dependence for Asset Allocation," *Journal of Financial Econometrics*, 2, 130-168.

Sentana, E. (1999), "Least Square Predictions and Mean-Variance Analysis," CEPR Discussion Paper No. 2088.

Sentana, E. (2005), "Least Square Predictions and Mean-Variance Analysis," *Journal of Financial Econometrics*, 1, 56-78.

Skouras, S. (2007), "Decisionmetrics: A Decision-based Approach to Econometric Modeling," *Journal of Econometrics*, 137, 414-440.

Taylor, N. (2013), "A Formula for the Economic Value of Return Predictability," *European Journal of Finance*, 19, 37-53.

Wachter, J.A., and Warusawitharana, M. (2015), "What Is the Chance That the Equity Premium Varies Over Time? Evidence from Regressions on the Dividend-Price Ratio," *Journal of Econometrics*, 186, 74-93.

West, K.D. (1996), "Asymptotic Inference About Predictive Ability," *Econometrica,* 64, 1067-1084.

West, K.D. and McCracken, M.W. (1998), "Regression-based Tests of Predictive Ability," *International Economic Review*, 39, 817-840.

White, H. (2000), "A Reality Check For Data Snooping," *Econometrica*, 68, 1097-1127.

Wooldridge, J.M., and White, H. (1998), "Central Limit Theorems for Dependent, Heterogeneous Processes with Trending Moments," in *Topics in Econometric Theory: The Selected Works of Halbert White*, ed. H. White, Cheltenham: Edward Elgar.

Table 1. Actual Size of Nominal 5% Tests Using Asymptotic Critical Values

			$b = 0$						$b > 0$					
			Not Winsorized			Winsorized			Not Winsorized			Winsorized		
	$T + P$	P/T	1/3	1	3	1/3	1	3	1/3	1	3	1/3	1	3
Recursive	88		0.018	0.010	0.006	0.018	0.010	0.006	0.065	0.036	0.028	0.065	0.036	0.028
	352		0.012	0.013	0.014	0.012	0.013	0.014	0.056	0.054	0.063	0.056	0.054	0.063
	1056		0.018	0.030	0.033	0.018	0.030	0.033	0.064	0.082	0.086	0.064	0.082	0.086
	2112		0.025	0.040	0.054	0.025	0.040	0.054	0.064	0.063	0.075	0.064	0.063	0.075
Mixed	88		0.012	0.004	0.002	0.011	0.003	0.002	0.026	0.011	0.006	0.018	0.010	0.003
	352		0.003	0.002	0.002	0.003	0.002	0.002	0.022	0.021	0.023	0.022	0.021	0.023
	1056		0.009	0.021	0.026	0.009	0.021	0.026	0.136	0.146	0.157	0.136	0.146	0.157
	2112		0.022	0.036	0.049	0.018	0.035	0.048	0.111	0.094	0.095	0.111	0.094	0.095

NOTES: The data-generating process is a bivariate VAR(1) with coefficients and error variance given in Section 5. The left-hand set of the two subpanels corresponds to the case in which the null holds because $b = 0$, while the set on the right corresponds to the case in which the null holds despite $b > 0$. In each set, the two subpanels report rejection frequencies under different sampling schemes. Columns denote different sample splits P/T and whether or not the weights are truncated between -0.5 and 1.5. Rows denote different values under different sample sizes $T + P = 88$, 352, 1056, and 2112. Each value is the percentage of rejections out of 2,000 replications.

Table 2. Actual Size of Nominal 5% Tests Using Bootstrapped Critical Values

		$b = 0$		$b > 0$	
	$T + P$	Not Winsorized	Winsorized	Not Winsorized	Winsorized
Recursive	88	0.011	0.011	0.065	0.065
	352	0.018	0.018	0.091	0.091
	1056	0.035	0.035	0.084	0.084
Mixed	88	0.057	0.023	0.064	0.034
	352	0.007	0.007	0.020	0.020
	1056	0.030	0.030	0.095	0.095

NOTES: The data-generating process is a bivariate VAR(1) with coefficients and error variance given in Section 5. The left-hand set of the two subpanels corresponds to the case in which the null holds because $b = 0$, while the set on the right corresponds to the case in which the null holds despite $b > 0$. In each set, the two subpanels report rejection frequencies under different sampling schemes. Columns denote whether or not the weights are truncated between -0.5 and 1.5. Rows denote different values under different sample sizes $T + P = 88$, 352, and 1056. Each value is the percentage of rejections out of 2,000 replications.

Table 3. US Equity Premium Predictability

Panel A) Monthly data

		Without winsorization			With winsorization		
	Variables	Φ	Asymptotic p-value	Boot p-value	Φ	Asymptotic p-value	Boot p-value
svar	Stock variance	−0.195	[0.59]	(0.67)	−0.010	[0.50]	(0.55)
d/e	Dividend payout ratio	−0.192	[0.99]	(0.82)	−0.038	[0.81]	(0.75)
lty	Long-term yield	−0.055	[0.56]	(0.67)	0.107	[0.32]	(0.12)
tms	Term spread	−0.005	[0.50]	(0.52)	0.177	[0.05]	(0.13)
ltr	Long-term return	−0.097	[0.72]	(0.75)	0.104	[0.20]	(0.20)
infl	Inflation	0.015	[0.46]	(0.47)	0.131	[0.03]	(0.17)
tbl	T-bill rate	0.030	[0.46]	(0.42)	0.171	[0.19]	(0.12)
dfr	Default return spread	0.051	[0.39]	(0.36)	0.069	[0.34]	(0.30)
dfy	Default yield spread	−0.089	[0.65]	(0.59)	0.074	[0.20]	(0.22)
ntis	Net equity expansion	−1.476	[0.90]	(0.87)	0.051	[0.41]	(0.33)
d/p	Dividend-price ratio	0.056	[0.43]	(0.41)	−0.054	[0.60]	(0.66)
d/y	Dividend yield	−0.069	[0.56]	(0.59)	0.018	[0.47]	(0.44)
e/p	Earnings-price ratio	−0.126	[0.59]	(0.69)	0.152	[0.26]	(0.17)
b/m	Book-market ratio	−0.693	[0.90]	(0.82)	−0.130	[0.69]	(0.77)

Panel B) Quarterly data

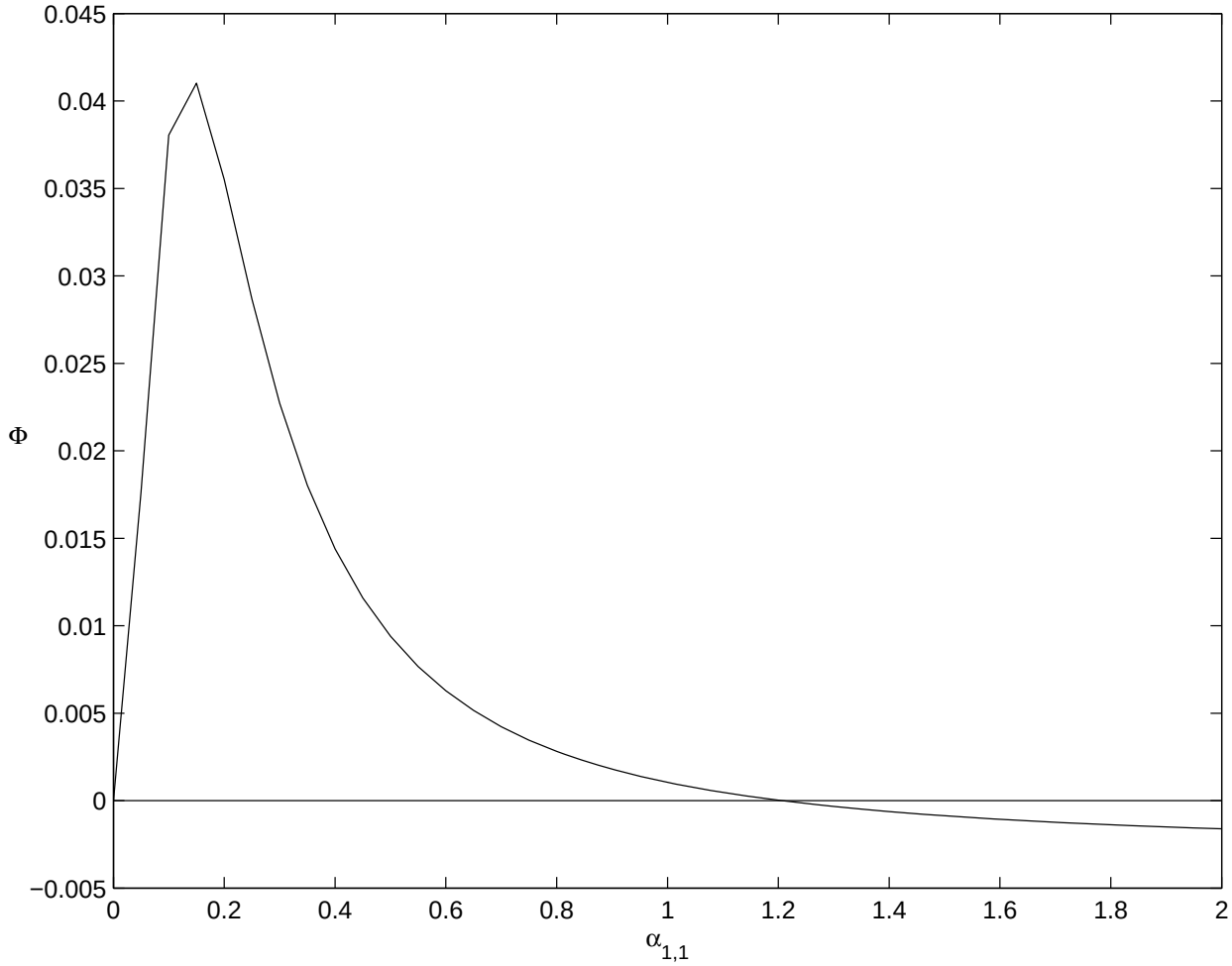
		Without winsorization			With winsorization		
	Variables	Φ	Asymptotic p-value	Boot p-value	Φ	Asymptotic p-value	Boot p-value
svar	Stock variance	−1.084	[0.73]	(0.74)	0.078	[0.39]	(0.39)
d/e	Dividend payout ratio	0.116	[0.36]	(0.42)	−0.044	[0.56]	(0.61)
lty	Long-term yield	0.424	[0.39]	(0.28)	0.362	[0.34]	(0.14)
tms	Term spread	−0.165	[0.54]	(0.58)	0.620	[0.08]	(0.15)
ltr	Long-term return	−1.928	[0.94]	(0.84)	0.017	[0.47]	(0.47)
infl	Inflation	0.332	[0.34]	(0.34)	0.355	[0.21]	(0.22)
tbl	T-bill rate	0.176	[0.46]	(0.41)	0.618	[0.22]	(0.11)
dfr	Default return spread	−8.296	[0.82]	(0.85)	−0.182	[0.52]	(0.73)
dfy	Default yield spread	−0.219	[0.55]	(0.53)	0.342	[0.24]	(0.25)
ntis	Net equity expansion	−16.589	[0.90]	(0.84)	0.074	[0.48]	(0.42)
d/p	Dividend-price ratio	−2.448	[0.73]	(0.74)	0.037	[0.48]	(0.47)
d/y	Dividend yield	0.065	[0.49]	(0.48)	0.160	[0.43]	(0.37)
e/p	Earnings-price ratio	0.022	[0.49]	(0.50)	0.546	[0.23]	(0.21)
b/m	Book-market ratio	−10.976	[0.96]	(0.82)	−0.321	[0.59]	(0.63)

Panel C) Annual data

		Without winsorization			With winsorization		
	Variables	Φ	Asymptotic p-value	Boot p-value	Φ	Asymptotic p-value	Boot p-value
svar	Stock variance	0.023	[0.46]	(0.49)	−0.009	[0.51]	(0.50)
d/e	Dividend payout ratio	−4.205	[0.91]	(0.87)	−0.567	[0.72]	(0.76)
lty	Long-term yield	2.146	[0.31]	(0.19)	0.666	[0.42]	(0.22)
tms	Term spread	1.892	[0.21]	(0.22)	1.580	[0.16]	(0.13)
ltr	Long-term return	0.032	[0.48]	(0.49)	1.758	[0.16]	(0.11)
infl	Inflation	−0.859	[0.91]	(0.66)	−0.417	[0.87]	(0.72)
tbl	T-bill rate	3.174	[0.21]	(0.16)	1.338	[0.29]	(0.14)
dfr	Default return spread	−8.092	[0.82]	(0.83)	0.418	[0.39]	(0.31)
dfy	Default yield spread	0.158	[0.47]	(0.46)	0.054	[0.48]	(0.46)
ntis	Net equity expansion	−24.312	[0.87]	(0.85)	−0.433	[0.52]	(0.65)
d/p	Dividend-price ratio	11.934	[0.04]	(0.12)	1.617	[0.33]	(0.22)
d/y	Dividend yield	11.961	[0.07]	(0.13)	2.310	[0.30]	(0.18)
e/p	Earnings-price ratio	9.319	[0.06]	(0.16)	0.591	[0.39]	(0.35)
b/m	Book-market ratio	7.226	[0.15]	(0.18)	−0.264	[0.54]	(0.56)

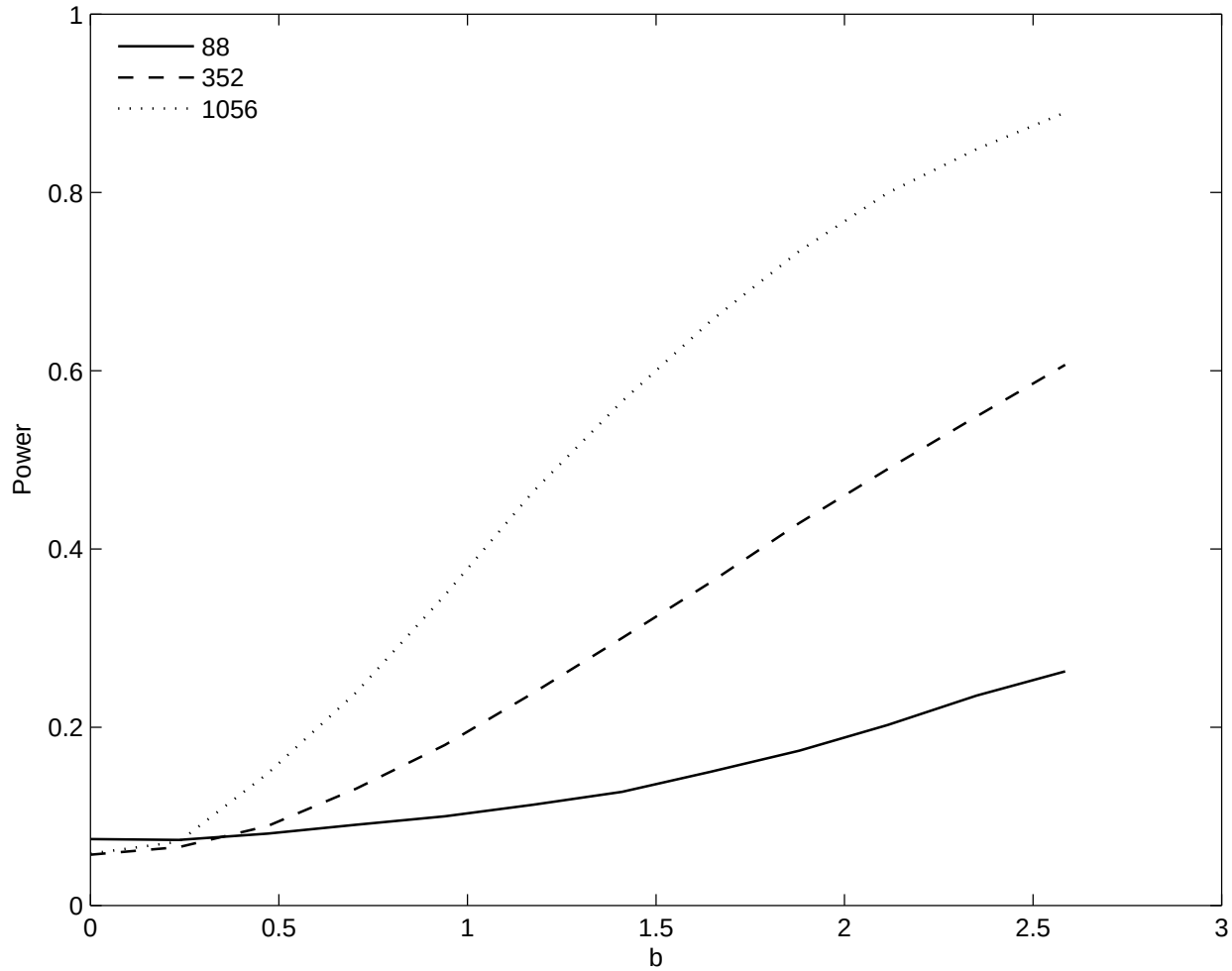
NOTES: The table reports performance fees, Φ based on out-of-sample forecasts of the conditional mean of stock index excess returns from predictive models with alternative predictive variables (Variables) against the benchmark represented by the historical mean of excess returns. Performance fees denote the amount investors with a mean-variance utility function and a coefficient of relative risk aversion (RRA) $\gamma = 3$ would be willing to pay to switch from each one of the predictive models to the historical average benchmark. The allocation to the risky asset is computed using a mixed scheme, where the conditional mean of excess returns is forecast using a recursive scheme and the variance of excess returns is estimated using the equivalent of the past 5 years' worth of past observations. The predictive variables are a sample of the ones used by Goyal and Welch (2008). The sample period is from January 1871 until December 2011 and the forecasts begin in January 1965. Φ are expressed as percentage per period (i.e., month, quarter, and year, respectively). Values in brackets are the p-values for the null hypothesis that $H_0 : \Phi \leq 0$. The p-values are computed using both the asymptotic scheme and bootstrap as discussed in Section 4. Bootstrapped p-values are computed with 999 replications and a block length selected according to the size of the sample. The recentering constant of the bootstrapped distribution is equal to the median of the empirical distribution of the test statistic computed across the 999 replications. When portfolio weights are winsorized, $-0.5 \leq w_{i,t}\left(\widehat{\beta}_{i,t}\right) \leq 1.5$ (i.e. a cap of 150% applies).

Figure 1: Φ as a Function of $\alpha_{1,1}$



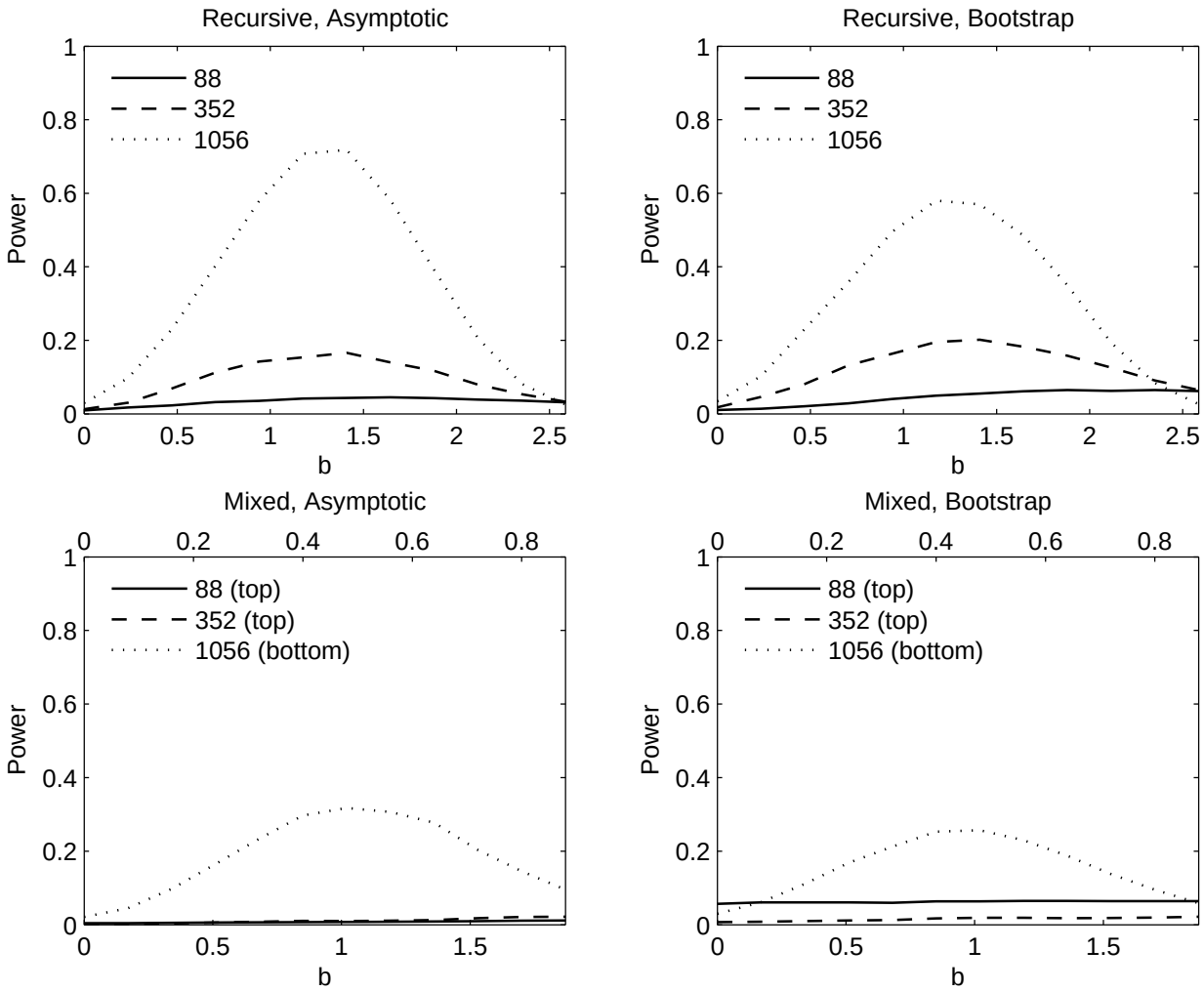
NOTES: The figure shows Φ as a function of $\alpha_{1,1}$ based on equation (3) in the text. Moments in the formula are loosely calibrated to the return on the NYSE value-weighted index as our risky asset, its dividend yield as our predictor, and the yield on the 3-month T-bill as our return on the risk-free asset. See Section 2 for more details.

Figure 2: Power of *MSE-t* Statistic



NOTES: The figure shows rejection frequencies of nominally 5% tests of the null hypothesis $E(e_{0,t+1}^2 - e_{1,t+1}^2) = 0$ using the *MSE-t* statistic in McCracken (2007) and its associated critical values. The simulation design is described in Section 5 and is the same as that used in our other Monte Carlo experiments that conduct inference on $\hat{\Phi}$.

Figure 3: Power Properties of Test Statistic



NOTES: The figure shows rejection frequencies of nominally 5% tests. Details on the simulation design are provided in section 5. Recursive results arise when all model parameters are estimated using the recursive scheme. Mixed results arise when the conditional mean and conditional variance are estimated using the recursive and rolling schemes, respectively. Asymptotic critical values indicate that the asymptotic standard errors were estimated and standard normal critical values were used for inference on the Studentized value of $\hat{\Phi}$. Bootstrap critical values are based on the percentile-bootstrap approach of Calhoun (2015) as described in the text. Note that under the mixed scheme, the horizontal axis varies by sample size.