

How to Dominate the Historical Average

Kai Li

Peking University HSBC Business School and the PHBS Sargent Institute of Quantitative Economics and Finance, P.R. China

Yingying Li

Hong Kong University of Science and Technology, Hong Kong SAR

Changlei Lyu

Shanghai University of Finance and Economics, P.R. China

Jialin Yu

Hong Kong University of Science and Technology, Hong Kong SAR

We present a novel methodology for the out-of-sample forecast of the equity premium. Our predictive slope coefficient is a conservative constant that has a lower bias than the zero slope employed by the historical average, but has the same variance. We demonstrate that, theoretically and empirically, our method dominates the historical average in forecast performance. Our methodology establishes a simple yet powerful paradigm for exploiting the real-time equity premium predictability derived from a predictor. Applications of our method reveal that many predictors can forecast the equity premium, and that parameter estimates in previous studies add value to out-of-sample forecasts. (*JEL* C53, G11, G12)

Received: April 20, 2023; Editorial decision: July 21, 2024

Editor: Lauren Cohen

Authors have furnished an [Internet Appendix](#), which is available on the Oxford University Press Web site next to the link to the final published paper online.

The current gold standard for forecasting the equity premium is the historical average. Even though previous studies propose many predictors of the equity premium, [Goyal and Welch \(2008\)](#) show that these predictors perform poorly in

We are deeply grateful to Lauren Cohen (the Editor) and two anonymous referees. We thank Binying Liu, Zhentao Shi, Bo Sun, Yizhou Xiao, and Xinghua Zheng; seminar participants at HKUST; and conference participants at the 16th Annual SoFiE Meeting and the 36th Australasian Finance and Banking Conference for their helpful comments. We also thank Weiyang Wen and Bo Zhou for their excellent research assistance. We gratefully acknowledge financial support from the Research Grants Council of Hong Kong [project numbers: 26503316, SRF52324-6H02]. [Supplementary data](#) can be found on *The Review of Financial Studies* web site. Send correspondence to Jialin Yu, jialin@ust.hk.

The Review of Financial Studies 38 (2025) 3086–3116

© The Authors 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons

Attribution-NonCommercial-NoDerivs licence (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.
<https://doi.org/10.1093/rfs/hhaf010>

Advance Access publication February 12, 2025

out-of-sample (OOS) forecasts when using regressions. More recently, [Goyal, Welch, and Zafirov \(2024\)](#) confirm the findings of [Goyal and Welch \(2008\)](#) using a larger set of return predictors over a longer sample period. In response to [Goyal and Welch \(2008\)](#), [Campbell and Thompson \(2008\)](#) ask “Predicting excess stock returns out of sample: Can anything beat the historical average?” in the title of their paper.

This paper provides a novel methodology for OOS forecasts of the equity premium. We reframe the question by [Campbell and Thompson \(2008\)](#) to “Can anything beat the historical average by providing a lower forecast bias?” Our aim is more modest than previous studies. We target a lower bias than the historical average instead of attempting to remove as much bias as possible or achieve unbiasedness.

To help motivate our method, we consider a guess-the-number game similar to “The Price Is Right” game, where the competitors guess the actual slope coefficient of a predictor like the dividend-to-price ratio. The closest guess to the correct number wins. The competitor, whose name is Mr. Historical Average, guesses zero. Suppose the actual slope is 1. Guessing according to an unbiased regression estimate is a good idea if the statistical estimation noise is low. However, the stock market data are noisy. For instance, if the estimation variance is 1, the regression approach loses to Mr. Historical Average if the estimate is more than one-standard-deviation away from the mean (below 0 or above 2). In our method, if one believes that the actual slope is positive so that the guess of 0 is downward biased, guessing a conservative positive constant, such as 0.2, is a bias reduction relative to Mr. Historical Average’s guess. This beats Mr. Historical Average unless the actual slope is near 0 (below 0.1). This heuristic game illustrates that a conservative and deterministic slope can detect the potential of bias reduction while insulating against estimation variance.

Consistent with the intuition in the game, we use a conservative predictive slope, which refers to the slope coefficient of a predictor. We obtain the sign of the slope from restrictions from economic theory as in [Campbell and Thompson \(2008\)](#), previous predictability studies, or statistical inference. Any of these sign decisions allows our forecast to outperform the historical average in the empirical analysis. Other than the sign decision, our slope coefficient is a constant. Guessing a larger slope in the game, such as 0.5, potentially exploits the predictor more but runs the risk of overshooting the actual slope. If a larger slope can reduce bias, the associated predictor is a stronger predictor of the equity premium.

Our approach neutralizes the main advantage of the historical average, which sets the slope coefficient to a deterministic zero for any predictor and consequently incurs no estimation variance. Our forecast uses a constant as the predictive slope, matching the historical average. Meanwhile, the bias reduction implies a better forecast performance than the historical average. Theories of time-varying expected returns do not often incorporate estimation

noise. The zero estimation variance in our approach mirrors such theories better than standard estimators.

We present theorems showing that, *ex ante*, our forecast stochastically dominates the historical average in the first order and has a lower mean squared forecast error than the historical average when the population predictive slope differs from zero. First-order stochastic dominance is a strong indication that our forecast is superior to the historical average. According to [Huang and Litzenberger \(1988, p. 40\)](#), anyone with a strictly increasing utility function over gains will prefer A to B if A first-order stochastically dominates B.

When the sign of the predictive slope is inferred statistically, we theoretically characterize the accuracy of the sign inference required for our forecast to stochastically dominate the historical average. The conventional statistical significance levels, such as the 90% to 99% levels, suffice to ensure that our forecast using the inferred sign outperforms the historical average. The result holds if the unexpected return follows the Student's *t* distributions with arbitrary degrees of freedom, which are often used in statistical analysis to accommodate differences in tail thickness. We also characterize the required accuracy if the unexpected return follows other probability distributions.

Using simulations, we confirm that our method delivers a better OOS forecast R^2 than the historical average. The simulations illustrate our method's advantage of a lower bias and comparable variance relative to the historical average. The simulations also confirm the first-order stochastic dominance of our forecast over the historical average.

We empirically apply our method to the predictors in [Goyal and Welch \(2008\)](#) and the predictors available on Robert Shiller's website. Our method outperforms the historical average in OOS forecasts. [Goyal and Welch \(2008\)](#) use time-series plots to show that the performance gains of many models concentrate especially on the years of the Oil Shock of 1973–1975. However, our forecast exhibits a steady accumulation of performance gains, as illustrated in [Figure 1](#) using the dividend-to-price ratio as a predictor. A forecast outperforms (underperforms) the historical average if the corresponding line on the plot is above (below) zero. Our method outperforms the historical average for various choices of parameter *A*, where a larger *A* corresponds to a more conservative predictive slope that is closer to zero. Our method consistently outperforms the historical average over time, a radical improvement over the forecast using ordinary least squares (OLS) regressions in the plot. This is because we design our forecast to minimize estimation variance. Such a figure is the key evidence in [Goyal and Welch \(2008\)](#) for the superiority of the historical average (see figure 1 in their paper). [Figure 1](#) shows that our forecast is a clear improvement over the historical average. This improvement is statistically significant, with a 99.2% confidence level.

The magnitude of our cumulative performance gain often exceeds that from the OLS approach, as illustrated by [Figure 1](#), despite our modest aim for a lower bias than the historical average instead of unbiasedness. In addition,

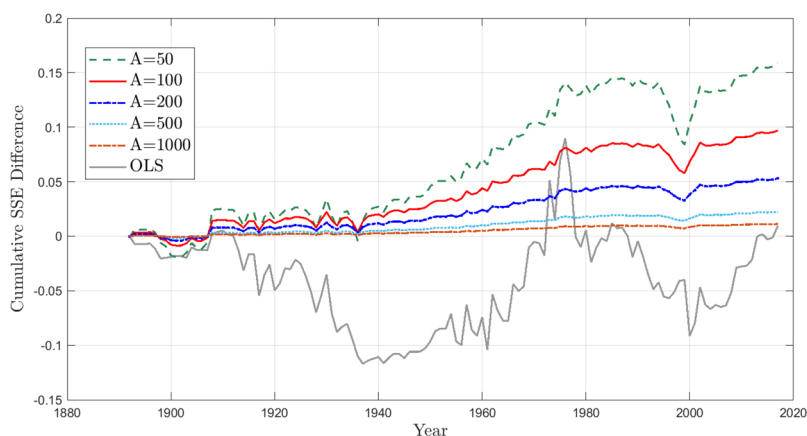


Figure 1
Cumulative annual performance using predictor dp

This figure shows the out-of-sample performance of annual return prediction using the predictor dp . Specifically, it shows the cumulative SSE (sum of squared prediction errors) using the historical mean minus the cumulative SSE using dp . We illustrate the performance of our method for A between 50 and 1,000 and for OLS.

we document evidence that our forecast first-order stochastically dominates the historical average. Consistent with the implication of first-order stochastic dominance, we also find a steady improvement in utility using our method.

Our theory also implies that the confidence bounds in previous predictability studies will be useful for OOS forecasts. For a predictor with a significantly positive slope, the lower confidence bound is above zero (by definition of positive significance) and below the population slope with a high chance (by construction of the confidence bounds). Therefore, there is a high probability that the lower confidence bound will have a lower bias than the historical average. Similarly, for a predictor with a significantly negative slope, there is a high probability that the upper confidence bound has a lower bias than the historical average. We confirm this in the empirical analysis. For simplicity of terminology, we refer to the lower (upper) confidence bound of a positively (negatively) significant estimate as the “confidence bound near zero” in this paper. For example, simply setting the predictive slope to the confidence bound near zero reported by [Campbell and Shiller \(1988a\)](#) produces a positive OOS R^2 in the subsequent sample period using the dividend-to-price ratio. Similarly, the confidence bound near zero in [Cochrane \(1991\)](#) generates a positive OOS R^2 using the investment-to-capital ratio. The fact that the numbers in previous papers add value to subsequent forecasts is direct evidence of OOS predictability.

Our paper makes the following contributions to the literature. First, given a predictor, we propose a novel methodology to deliver better OOS forecasts of the equity premium than the historical average. [Stambaugh \(1999\)](#) demonstrates that there is a bias in the regression coefficient of the

predictive slope. Lewellen (2004), Cochrane (2008), Campbell and Yogo (2006), and Campbell and Thompson (2008) use additional information derived from finance theory to address the bias Stambaugh (1999) identifies, as discussed in Campbell (2018). Eliazs (2005) and Jansson and Moreira (2006) study median unbiased estimators. Our approach focuses on reducing estimation variance. When the population predictive slope is not zero, the historical average is biased. However, the historical average is special because it uses a deterministic predictive slope of zero. The complete lack of variance can offset its bias and lower the mean squared error. We use a nonzero deterministic predictive slope. By not attempting to uncover the full extent of predictability, we avoid the substantial estimation variance associated with stock return data. Johnson (2019) studies a weighted least squares estimator (WLS). Despite being more efficient than OLS, WLS is fundamentally in the same class of regression estimators as OLS. WLS estimates often do not differ substantially from OLS estimates. For example, figure 2 in Johnson (2019) shows that the WLS estimate is similar to the OLS estimate over most of the sample period. We also replicate Figure 1 using WLS and find that the fluctuation of its performance is similar to the OLS.

Our focus on delivering a positive OOS R^2 differs from Clark and West (2006). They show that finite-sample estimation noises imply that predictive regressions generate a negative OOS R^2 under the null hypothesis of no predictability. Therefore, a zero OOS R^2 can be interpreted as weak evidence for predictability, and they propose a statistical test that accounts for this effect. Clark and West (2007) generalizes the test to comparing a parsimonious null model to a larger model that nests the null model. Clark and West (2006) is not about how to construct superior OOS forecasts, as discussed in Campbell and Thompson (2008, footnote 5). The same applies to Clark and West (2007).

Second, we contribute a theory on the outperformance over the historical average if the population predictive slope is not zero. Therefore, we answer the question “Can anything beat the historical average?” posed by Campbell and Thompson (2008) both theoretically and empirically. As long as we target a modest bias reduction and choose the sign of the predictive slope with sufficient accuracy, ours is the first forecast method for the expected market return that is proven ex ante to outperform the historical average. The ex ante validation is crucial for a viable alternative to the historical average, which is the current gold standard and widely taught in classrooms.

Third, we contribute to the literature by providing recommendations on how to target the modest bias reduction and estimate the sign of the predictive slope. These facilitate the implementation of our theory. We demonstrate that the confidence bound near zero satisfies the condition of our theorem with near certainty because there is a high probability that it has the right sign and does not overshoot the population slope. Therefore, our methodology forms a benchmark sufficient and necessary with near certainty to detect the OOS predictability of a given predictor, provided that there are sufficient

observations to evaluate the OOS performance. We are unaware of any other method that forms such a (nearly) necessary condition. When using the OLS, the estimation noise can be so large that the OOS R^2 is negative despite a nonzero population slope. Consistent with our theory, we find evidence that the confidence bounds in published studies add value to subsequent OOS forecasts, which contributes fresh evidence for OOS predictability. Our finding that the confidence bounds near zero, instead of the point estimates, have a high power to outperform the historical average also contributes to the OOS forecast methodology.

Fourth, we contribute strong evidence that supports the real-time predictability of time-varying equity premiums. Bossaerts and Hillion (1999), Goyal and Welch (2008), and Goyal, Welch, and Zafirov (2024) argue that prediction models have no OOS forecasting power. By contrast, using our method, 15 of the 23 predictors we study have significantly positive OOS R^2 at the 90% level, and eight are 95% significant. The eight predictors with 95% significance include the cyclically adjusted price-to-earnings ratio (Campbell and Shiller 1988b); cyclically adjusted total return price-to-earnings ratio (Bunn and Shiller 2014); dividend-to-price ratio and dividend yield (Ball 1978; Campbell and Shiller 1988a; Hodrick 1992; Lewellen 2004; Menzly, Santos, and Veronesi 2004; Cochrane 2008; van Binsbergen and Koijen 2010); consumption-wealth ratio (Lettau and Ludvigson 2001); investment-to-capital ratio (Cochrane 1991); long-term rate of returns (Campbell 1987; Fama and French 1989); and corporate bond returns (Fama and French 1989; Keim and Stambaugh 1986). The seven predictors with 90% significant OOS R^2 include the book-to-market ratio (Kothari and Shanken 1997; Pontiff and Schall 1998); long-term government bond yield using either the Goyal and Welch (2008) data or the Robert Shiller data, Treasury-bill rate, and term spread (Campbell 1987; Fama and French 1989; Hodrick 1992); inflation rate from the Robert Shiller data (Campbell and Vuolteenaho 2004b; Fama 1981; Fama and Schwert 1977); and earnings-to-price ratio (Campbell and Shiller 1988b, 1998; Lamont 1998). Superior OOS predictability helps exploit the predictability in real time and validates theories of time-varying expected return.

Some studies use multiple variables to forecast the equity premium. Rapach, Strauss, and Zhou (2010) and Elliott, Gargano, and Timmermann (2013) combine forecasts from different predictors. Pástor and Stambaugh (2009) exploit information in the lagged returns and predictors. Campbell and Vuolteenaho (2004a), Eleswarapu and Reinganum (2004), Polk, Thompson, and Vuolteenaho (2006), Kelly and Pruitt (2013), and Dong et al. (2022) use cross-sectional data to forecast the expected market return. Kelly, Malamud, and Zhou (2024) use machine learning and examine the Sharpe ratio, which suggests the potential informativeness of predictors of expected market return for volatility forecasts. Our findings show that even an individual predictor can outperform the historical average, and our method can potentially extend to multivariate forecasts.

1. Dominate the Historical Average

In this section, we present our forecast methodology. Consider a predictor x of the future return r that satisfies:

$$r = \mu + bx + e, \quad (1)$$

where b is the predictive slope coefficient of the predictor. e is a residual with zero mean and is uncorrelated with the predictor. x has zero mean. μ is the unconditional expected return. $\mathbf{x}$ denotes a vector containing n observations of the predictors $x_1, x_2, \dots, x_n$. Similarly, $\mathbf{r}$ and $\mathbf{e}$ denote the vectors of r and e , respectively. We use mean squared error (MSE) as the loss function to compare different forecasts.

Although the population coefficients are constants in Equation (1), it is without loss of generality, given our focus on OOS forecasts. If these coefficients vary over time, μ and b in Equation (1) correspond to the population intercept and slope at the time of the forecast.

We demonstrate the superiority of our forecast using the criterion of first-order stochastic dominance, which is defined as follows: random variable A has first-order stochastic dominance over random variable B if, for any outcome c , A gives at least as high a probability of receiving at least c as does B , and for some c , A gives a higher probability of receiving at least c , such that $P(A \geq c) \geq P(B \geq c)$ for all c and $P(A \geq c) > P(B \geq c)$ for some c .

We apply the definition of first-order stochastic dominance, which is over gains, to the negative loss function $-\text{MSE}$ in our theorems. Let MSE_A and MSE_B denote the MSE of forecasts A and B . Forecast A first-order stochastically dominates forecast B if and only if $P(\text{MSE}_A \leq c) \geq P(\text{MSE}_B \leq c)$ for all c and $P(\text{MSE}_A \leq c) > P(\text{MSE}_B \leq c)$ for some c .

We start by comparing the return forecast μ with $\mu + \delta x$ in Theorem 1. The first forecast does not use the predictor x . We assume that x is not always zero, otherwise the two forecasts will be the same.

Theorem 1. The forecast $\mu + \delta x$ first-order stochastically dominates the forecast μ for predicting return r under the following conditions:

- (1) δ is a constant between 0 and b ;
- (2) the probability density function of the vector $\mathbf{e}$ is strictly decreasing in $\|\mathbf{e}\|$. $\|\cdot\|$ indicates the l^2 norm.

The unconditional mean μ is often estimated by the historical average, denoted $\hat{\mu}$, which satisfies:

$$\mu = \hat{\mu} + \zeta, \quad (2)$$

where ζ has zero mean. Substituting Equation (2) into Equation (1) gives:

$$r = \hat{\mu} + bx + \epsilon, \quad (3)$$

where $\epsilon \equiv e + \zeta$. We compare the return forecast $\hat{\mu}$ with $\hat{\mu} + \delta x$ in the following theorem.

Theorem 2. The forecast $\hat{\mu} + \delta x$ first-order stochastically dominates the forecast $\hat{\mu}$ for predicting return r under the following conditions:

- (1) δ is a constant between 0 and b ;
- (2) the probability density function of the vector ϵ is strictly decreasing in $\|\epsilon\|$.

The monotonicity restriction that the probability density function decreases in the distance to the origin holds for many commonly used distributions, such as the t distribution and the normal distribution with zero mean.¹ First-order stochastic dominance implies that our method tends to have a smaller MSE than the historical average, as shown by [Corollary 1](#).

Corollary 1. Under the assumptions of [Theorem 2](#), the MSE using $\hat{\mu} + \delta x$ satisfies:

$$\text{MSE}_{\hat{\mu} + \delta x} \stackrel{d}{=} \text{MSE}_{\hat{\mu}} - \eta, \quad \eta \geq 0, \quad (4)$$

where $\stackrel{d}{=}$ indicates that both sides have the same probability distribution. $\text{MSE}_{\hat{\mu}}$ is the MSE using the historical average. η is a nonnegative random variable.

1.1. Implementation

To implement the forecast, we set the predictive slope to:

$$\delta = \begin{cases} 1/A & \text{if } b > 0 \\ -1/A & \text{if } b < 0, \end{cases} \quad (5)$$

where $A > 0$ is a constant. A larger A helps ensure δ is between 0 and b , as required by [Theorem 2](#), though at the expense of magnitude. We provide an interpretation of A in [Section 3.3](#). The confidence interval of the estimate of b can help select A and determine the sign of b . Consider a 95% confidence interval $[\hat{b}_L, \hat{b}_U]$ for the population slope b , where $\hat{b}_L$ and $\hat{b}_U$ are the lower and upper bounds, respectively. This implies that:

$$\hat{b}_L \leq b \leq \hat{b}_U \quad (6)$$

with 95% probability. For simplicity, let us assume for now that the slope is statistically significantly positive (i.e., $0 < \hat{b}_L$). The following equation then holds (with a probability close to one):

$$0 < \hat{b}_L \leq b. \quad (7)$$

In this case, if we set $\delta = 1/A$ in [Equation \(5\)](#) and choose $1/A$ according to the lower confidence bound $\hat{b}_L$, the requirement that δ is between 0

¹ [Theorem 1](#) can be proved without the monotonicity restriction using, instead, the orthogonality between ϵ and x . However, for [Theorem 2](#), the estimation error ξ of the unconditional mean can correlate with x , especially for a price-based predictor x influenced by past returns ([Stambaugh 1999](#)). Therefore, we impose the monotonicity restriction so that [Theorem 2](#) does not require orthogonality between ϵ and x .

and b in [Theorem 1](#) and [Theorem 2](#) is likely satisfied. A wider confidence interval implies a lower bound that is closer to zero. This translates into a more conservative slope (larger A) when the estimation variance is high. The confidence interval $[\hat{b}_L, \hat{b}_U]$ is typically derived from an asymptotic distribution, which is suitable for our method as long as the asymptotic distribution provides an accurate approximation in the finite sample.

In the empirical analysis, we standardize the predictor x to have a variance of one by dividing x by its standard deviation $\hat{\sigma}_x$ estimated from past data. This facilitates the interpretation and comparison of slopes across different predictors. Using a slope of $\hat{b}$ with a predictor x gives the same forecast as using a slope of $\hat{b}\hat{\sigma}_x$ with a standardized predictor $x/\hat{\sigma}_x$.

As an example, [Campbell and Shiller \(1988a, table 4\)](#) estimate that the dividend-price ratio has a predictive slope of 0.129 with a standard error of 0.057. Their table 2 shows that the standard deviation of the predictor is 0.277. Therefore, the 95% lower confidence bound for the slope of the standardized predictor is $(0.129 - 1.96 \times 0.057) \times 0.277 = 0.0048$, where 1.96 is the cutoff value for 95% significance. Setting $1/A$ to this number implies $A = 209$, where a one-standard-deviation change in the predictor affects the forecast of subsequent annual market return by $1/209$, or approximately 0.5%. In [Internet Appendix C.1](#), we use similar calculations based on estimates in published studies to guide A selections for other predictors.

[Equation \(7\)](#) indicates that our slope is closer to the population slope than the historical average with a 95% chance. In [Internet Appendix B.4](#), we use simulations to demonstrate that our forecast outperforms the historical average despite the 5% error rate. We also use a larger A to examine a more conservative slope than the lower confidence bound.

The choice of δ is similar if the estimate of b is significantly negative. In this case, the upper confidence bound satisfies $b \leq \hat{b}_U < 0$ with a probability close to one, which corresponds to setting $\delta = -1/A$ in [Equation \(5\)](#) and choosing $1/A$ according to $|\hat{b}_U|$.

We use statistical estimation to infer the sign of b above. Alternatively, theory can help restrict the sign ([Campbell and Thompson 2008](#)). In either case, an investor may not be sure about the sign, which we study in the next section.

2. Uncertainty about the Sign of the Predictive Slope

In this section, we analyze the performance of our forecast when the sign of the predictive slope is inferred with error. In this case, our forecast becomes $\mu + dx$, where

$$d = \begin{cases} \delta & \text{with probability } p \\ -\delta & \text{with probability } q \\ 0 & \text{with probability } 1 - p - q. \end{cases}$$

With probability p , the sign is correctly inferred. In this case, our forecast is the same as the forecast $\mu + \delta x$ in [Theorem 1](#). When a wrong sign is used, $d = -\delta$. When the sign cannot be determined, because of statistical insignificance, for example, we use the historical average, which corresponds to $d = 0$. Therefore, $d = s\delta$, where $s = 1, -1$, or 0 .

We apply the forecast over n periods when the predictor takes the value $\mathbf{x} = (x_1, x_2, \dots, x_n)$. Let $f(\mathbf{e})$ denote the probability density function (PDF) of $\mathbf{e}$, which is the vector containing the residual in [Equation \(1\)](#). Let $F(c, s)$ denote the probability of achieving a root mean squared error (RMSE) below c when using the slope $d = s\delta$ for $\mathbf{x}$. Let $F_c(c, s) \equiv \frac{\partial}{\partial c} F(c, s)$ denote the derivative of F with respect to c . The following condition, if it holds for all c , ensures that our forecast first-order stochastically dominates the historical average:

$$\frac{p}{q} > \frac{F(c, 0) - F(c, -1)}{F(c, 1) - F(c, 0)}. \quad (8)$$

The following theorem provides conditions under which [Equation \(8\)](#) holds for all c . The proof of [Theorem 3](#), in [Internet Appendix A.1](#), also derives [Equation \(8\)](#).

Theorem 3. For a given $\mathbf{x}$, if

- (1) there exists a finite constant B_1 , such that $B_1 > \lim_{c \rightarrow +\infty} \frac{F_c(c, 0) - F_c(c, -1)}{F_c(c, 1) - F_c(c, 0)}$,
- (2) the probability density function $f(\mathbf{e})$ is continuous;
- (3) all the conditions in [Theorem 1](#) hold,

then

- (i) there exists a finite constant C , such that $B_1 > \frac{F(c, 0) - F(c, -1)}{F(c, 1) - F(c, 0)}$ for all $c > C$;
- (ii) $B_2 = \sup_{0 < c \leq C} \frac{F(c, 0) - F(c, -1)}{F(c, 1) - F(c, 0)}$ is a finite constant;
- (iii) the forecast $\mu + d\mathbf{x}$ first-order stochastically dominates the forecast μ for predicting return r when $p/q > \max(B_1, B_2)$.

We do not make the dependence of the constants C , B_1 , and B_2 on $\mathbf{x}$ explicit in [Theorem 3](#) to simplify the notations.

To illustrate the magnitude required for p/q , in the next Lemma, we characterize the constant B_1 for the Student's t distribution for the one-dimensional case when $\mathbf{x} = (x)$, which corresponds to the evaluation of the next forecast when the predictor takes the value x . In this case, q is the probability of using a wrong sign for the next forecast of the expected return. The t distribution with degree of freedom ν captures many tail behaviors. It exhibits fat tails when ν is small. When $1 < \nu \leq 2$, the t distribution has a variance of infinity. When $\nu = 1$, it becomes the Cauchy distribution, for which the variance is even undefined. When $\nu \rightarrow \infty$, it becomes the standard normal distribution in the limit, which has thin tails.

Lemma 1. If the residual e in Equation (1) satisfies $e = \sigma \epsilon$, where ϵ is a random variable following the Student's t distribution with degree of freedom ν , then any $B_1 > \frac{2b+\delta}{2b-\delta}$ satisfies condition (1) of Theorem 3.

In Lemma 1, σ is a scale transformation of the t distribution. When the t distribution has a finite variance, σ changes the volatility of the residual. The cutoff for B_1 in Lemma 1 is the same regardless of σ and the degree of freedom ν .

We numerically investigate the magnitude of B_2 , which is defined over a finite region between 0 and C according to Theorem 3, in Internet Appendix A.2. We find that B_2 is typically higher than the B_1 cutoff in Lemma 1. However, they are close to each other, especially if δ is near the population slope b . In Internet Appendix A.2, we show that the required sign accuracy, $\max(B_1, B_2)$, is higher when δ is closer to b . Therefore, when investors are unsure about b , they should assume it is near δ to allow a sufficient buffer for the required sign accuracy. In this case, the B_1 cutoff in Lemma 1 provides a good approximation to B_2 . For example, when choosing δ according to the confidence bound near zero, we find a proportional difference of no more than 2% between the B_1 cutoff in Lemma 1 and the required sign accuracy if δ is near b .

Therefore, $\frac{2b+\delta}{2b-\delta}$ from Lemma 1 provides a rule of thumb on how large p/q should be. When δ/b is near zero, $\frac{2b+\delta}{2b-\delta} \approx 1$, our forecast stochastically dominates the historical average as long as the chance of a correct sign relative to a wrong sign is somewhat better than 50-50. Since $\frac{2b+\delta}{2b-\delta}$ is increasing in δ/b , a more aggressive choice of δ relative to the population slope b raises the required p/q . At the maximum $\delta/b = 1$ allowed in Theorem 1, $\frac{2b+\delta}{2b-\delta} = 3$.² When the sign is statistically significant at the 95% confidence interval, p/q is at least $0.95/0.05 = 19$, which substantially exceeds the requirement of 3. A statistical significance level of 90% still suffices since $0.9/0.10 = 9$.³

For the normal distribution, no finite bound B_1 can satisfy condition (1) of Theorem 3. This is because the tail of the normal distribution is so thin that the proportional difference between the probability densities of two observations a fixed distance apart converges to infinity in the extreme tails. Extreme tails refer to realizations far away from zero in terms of standard deviations in this paper. In Internet Appendix A.3, we elaborate why such convergence to infinity makes a finite bound for B_1 impossible.

² In Theorem 1, we exclude the scenario in which $1 < \delta/b < 2$. In this case, δ overshoots but is still closer to b than the historical average. However, $\frac{2b+\delta}{2b-\delta} \rightarrow \infty$ when $\delta/b \rightarrow 2$. Therefore, when there is uncertainty about the sign, the requirement that $p/q > B_1$ can be too high to meet if δ overshoots b .

³ In Internet Appendix A.2, we show that, holding δ/b fixed, the rule of thumb approximates B_2 less precisely if δ has a bigger absolute value. When using a δ far away from zero, investors should examine B_2 , in addition to the rule of thumb $\frac{2b+\delta}{2b-\delta}$, for the required sign accuracy. Conventional statistical significance levels, such as the 95% or 90% level, satisfy the required sign accuracy for all δ in this paper.

However, other than the extreme tails, the normal distribution can be closely approximated by a t distribution using a high degree of freedom. In [Internet Appendix A.3](#), we investigate whether the rule of thumb $\frac{2b+\delta}{2b-\delta}$, which is derived from the t distribution, works well for the normal distribution if we exclude the extreme tails. The answer is yes. For example, using parameters calibrated to the dp process, we show that $\frac{2b+\delta}{2b-\delta}$ is a good rule of thumb within a seven-standard-deviation range of the market return. Seven standard deviations imply that, if the required sign accuracy is met, for a RMSE value c that may appear in almost all scenarios in practice, our forecast has a higher probability of achieving a RMSE below c than the historical average. Therefore, our method will perform well in practice for normal distributions, which we confirm in [Internet Appendix B.2](#) using simulations.⁴

We also use the empirical distribution derived from the residual stock market return data in the simulation in [Internet Appendix B.2](#). The simulation results confirm that our method performs well using the empirical distribution.

First-order stochastic dominance is quite a high bar to clear, and naturally there are some return distributions under which we cannot make such strong statements. But our simulations (and empirical analyses in [Internet Appendix C](#)) suggest that investors on average will benefit from our method when signs are estimated from the data and predictors are used after passing a conventional significance threshold.

Next, we theoretically characterize the performance of the forecast $\hat{\mu} + dx$ when we estimate the unconditional mean μ using the historical average $\hat{\mu}$. In this case, the residual ϵ is given by [Equation \(3\)](#). Let ϵ denote the vector containing ϵ over n periods, and $F^\epsilon(c, s)$ denote the probability of achieving a RMSE below c when using the slope $d = s\delta$ for x . Let $F_c^\epsilon(c, s) \equiv \frac{\partial}{\partial c} F^\epsilon(c, s)$ denote the derivative of F^ϵ with respect to c .

Theorem 4. For a given x , if

- (1) there exists a finite constant B_1 , such that $B_1 > \lim_{c \rightarrow +\infty} \frac{F_c^\epsilon(c, 0) - F_c^\epsilon(c, -1)}{F_c^\epsilon(c, 1) - F_c^\epsilon(c, 0)}$,
- (2) the probability density function $f^\epsilon(\epsilon)$ is continuous;
- (3) all the conditions in [Theorem 2](#) hold,

then

- (i) there exists a finite constant C , such that $B_1 > \frac{F_c^\epsilon(c, 0) - F_c^\epsilon(c, -1)}{F_c^\epsilon(c, 1) - F_c^\epsilon(c, 0)}$ for all $c > C$;
- (ii) $B_2 = \sup_{0 < c \leq C} \frac{F_c^\epsilon(c, 0) - F_c^\epsilon(c, -1)}{F_c^\epsilon(c, 1) - F_c^\epsilon(c, 0)}$ is a finite constant;
- (iii) the forecast $\hat{\mu} + dx$ first-order stochastically dominates the forecast $\hat{\mu}$ for predicting return r when $p/q > \max(B_1, B_2)$.

⁴ In addition, the nonnormality of returns is a stylized fact demonstrated by past studies, such as [Fama \(1965\)](#) and [Mandelbrot \(1967\)](#).

Corollary 2. Under the assumptions of [Theorem 4](#), the MSE using $\hat{\mu} + dx$ satisfies:

$$\text{MSE}_{\hat{\mu}+dx} \stackrel{d}{=} \text{MSE}_{\hat{\mu}} - \eta, \quad \eta \geq 0, \quad (9)$$

where $\stackrel{d}{=}$ indicates that both sides have the same probability distribution. $\text{MSE}_{\hat{\mu}}$ is the MSE using the historical average. η is a nonnegative random variable.

3. Additional Interpretations of Our Forecast Method

Our expected return forecast, $\hat{\mu} + \delta x$, where δ is from [Equation \(5\)](#), consists of two components: the historical average $\hat{\mu}$ and a bias reduction governed by the predictor x and a deterministic slope δ . In this section, we provide some additional interpretations of our predictor.

3.1. Understand the opponent: why is the historical average difficult to beat in OOS forecasts?

To see the intuition behind why the historical average is difficult to beat in OOS forecasts, we can compute the OOS MSE difference between the historical average and a forecast using an estimate $\hat{b}$ of the predictive slope. Let $\text{MSE}_{\hat{\mu}}$ and $\text{MSE}_{\hat{\mu}+\hat{b}x}$ denote the MSE using the historical average $\hat{\mu}$ and $\hat{\mu} + \hat{b}x$, respectively, to forecast the next return r .

$$\begin{aligned} E \left[\text{MSE}_{\hat{\mu}} - \text{MSE}_{\hat{\mu}+\hat{b}x} \right] &= E \left[(r - \hat{\mu})^2 - (r - \hat{\mu} - \hat{b}x)^2 \right] \\ &= E \left[-\hat{b}^2 x^2 + 2\hat{b}bx^2 + 2\hat{b}x\epsilon \right] \\ &= E \left[b^2 x^2 - (\hat{b} - b)^2 x^2 + 2\hat{b}x(e + \mu - \hat{\mu}) \right] \\ &= (b^2 - \text{Bias}(\hat{b})^2 - \text{Variance}(\hat{b}))x^2 - 2E \left[(\hat{\mu} - \mu)\hat{b} \right]x, \end{aligned} \quad (10)$$

which is derived by plugging in [Equation \(3\)](#) for r and then substituting $\epsilon = e + \mu - \hat{\mu}$ following [Equation \(2\)](#).

The predictive slope is zero for the historical average. This results in a bias if the predictor is informative. The bias can be reduced using an estimate $\hat{b}$ of the predictive slope. However, if the variance of $\hat{b}$ is sufficiently large, it can more than offset the reduction in bias. This is consistent with [Clark and West \(2006\)](#), who show that estimation noises imply that the predictive regressions produce a negative OOS R^2 when there is no predictability. When there is no predictability, $b=0$, and the coefficient of x^2 is negative in [Equation \(10\)](#).

In addition, the noisiness of $\hat{b}$ may adversely affect the forecast performance through the last term in Equation (10) if $E[(\hat{\mu} - \mu)\hat{b}]x > 0$. For example, if $\hat{\mu}$ is already higher than μ , a big positive $\hat{b}x$ makes the forecast overshoot the population mean even further.

By contrast, our forecast uses a constant slope to match the constant slope of the historical average. When $\hat{b}$ is a constant, the last term in Equation (10) becomes zero if $\hat{\mu}$ is an unbiased estimate of μ . So Equation (10) becomes $(b^2 - \text{Bias}(\hat{b})^2)x^2$, which is positive if $\hat{b}$ is between 0 and b .

3.2. Gradient descent

Our forecast method is a one-step correction of the historical average. To see this, let us consider the gradient descent method in optimization (Lemaréchal 2012), which finds the minimum of a function $f(a)$. Gradient descent starts from an initial guess of a_0 and then finds the next guess according to:

$$a_1 = a_0 - \gamma f'(a_0), \quad (11)$$

where $\gamma > 0$ is a small step size. Gradient descent takes a step toward the direction indicated by $-f'$ to form the next guess of a_1 . A negative f' suggests that a bigger a leads to a further decrease in f and, consistently, $a_1 - a_0 > 0$ in this case.

In our context of return prediction, the function f and parameter a above map to the forecast loss function and the predictive slope, respectively. We consider the historical average as an initial guess. Since the historical average ignores any predictability, this implies $a_0 = 0$. Our method uses a predictive slope of $a_1 = \delta$, where δ is given by Equation (5). We can map δ in our method to $-\gamma f'(a_0)$ in Equation (11). In other words, our predictive slope can be viewed as taking a step of size $1/A$ from the initial $a_0 = 0$ toward the direction guided by the sign of the population predictive slope b .

Therefore, gradient descent links the historical average and our method. From the historical average as a starting point, our method can be viewed as the next step toward predictability. Similar to gradient descent, taking this step requires additional information. In gradient descent, the information comes from f' . In our method, the information derives from either theoretical restrictions or statistical inferences of the predictive slope's sign.

To match the zero variance of the historical average, we use a constant step size, which is equivalent to using only the sign, but not the magnitude of f' in gradient descent.

3.3. Amortization of a predictor's deviation from the mean

Next, we provide an interpretation of the step size in gradient descent, which is $1/A$ in Equation (5). For concreteness, we consider the log dividend-to-price ratio as a predictor, denoted by dp . Recall that we standardize the predictor to have a zero mean and a variance of one.

When dp is at its mean of 0, it does not affect the return forecast. If $dp = -1$, the log dividend-to-price ratio is one-standard-deviation below its mean. The price is excessively high relative to the price implied by the long-run mean dp . A predictive slope of $1/A$ resembles the amortization of the excess valuation as if a $1/A$ fraction of the excess valuation is mean reverted each period. For example, if $A = 100$, a predictor that is one-standard-deviation away from the mean affects the return forecast by 1% each period.

A larger A implies a more conservative amortization; hence, the expected return forecast responds less to a fluctuation in valuation. A conservative amortization also helps our forecast method guard against time variation in the population predictive slope (Lettau and Nieuwerburgh 2008). As long as $1/A$ is between the population predictive slope and zero, Theorem 2 continues to hold. If the population predictive slope differs substantially from zero, the associated predictor is strong. A strong predictor will likely outperform the historical average with more statistical significance for a given A or achieve a required statistical significance using a small A (equivalently, a large predictive slope).

Such linear amortization approximates the mean reversion dynamics of a predictor when the actual dynamics are nonlinear. For example, assuming the data-generating process of dp is an $AR(1)$ process with ρ being the autoregressive coefficient.⁵ The expected dp ratio at time t relates to its time 0 value by $dp_t = \rho^t dp_0$. The implied magnitude of mean reversion is:

$$dp_0 - dp_t = dp_0(1 - \rho^t) \approx dp_0(1 - \rho)t,$$

where the last step follows from a Taylor approximation around $\rho = 1$. The Taylor approximation implies a linear mean reversion at the rate of $1 - \rho$ each period, which our method can match by setting $1/A = 1 - \rho$. Many predictors are persistent with ρ near 1 (Stambaugh 1999), which implies that $1/A$ is near 0 and A is large.

3.4. A better limit for shrinkage

Shrinkage aims to improve a standard estimator, such as a regression estimator, by shrinking it toward a constant limit. The idea is that the reduction in estimation variance can outweigh the worsening of the estimate introduced by biasing the estimate toward the arbitrary constant value, which resembles the bias-variance trade-off in Equation (10).

However, a shrinkage estimator is still random since it is a weighted average of the regression estimator and the constant limit. By contrast, our slope δ in Equation (5) is a constant. Only by matching the zero variance while making an incremental reduction in bias are we able to characterize ex ante our method's superior performance relative to the historical average in Sections 1 and 2.

⁵ See Section 4 for a more detailed description of the $AR(1)$ process. Stambaugh (1999), among others, also uses $AR(1)$ to characterize the evolution of the predictor.

The historical average need not be the only choice for the shrinkage limit. When there is predictability, the historical average is biased. The bias can be of the same order of magnitude as the standard deviation of the estimate, as shown by our simulation study in [Section 4](#), and negate the benefit of shrinkage. Additionally, shrinkage toward zero is logically inconsistent, at least in the limit, if the hypothesis test reveals with 95% confidence that the slope is not zero. Therefore, replacing the historical average with our forecast, which offers a lower forecast bias than the historical average, can potentially improve shrinkage. The simulations in [Internet Appendix B.5](#) confirm that shrinkage using our forecast as the limit can outperform shrinkage using the historical average as the limit. We also compare the empirical performance of our method to methods based on shrinkage, such as the Ridge and Lasso regressions, in [Internet Appendix C.6](#) and find that our method has a better OOS performance.

4. Simulations

In this section, we use simulations to study the performance of our forecast method. We also compare our forecast's bias and variance with the historical average.

4.1. Simulation setup

In the simulations, we consider the following first-order vector autoregression (VAR) representation of log returns r , log dividend-to-price ratio dp , and log dividend growth Δd following [Cochrane \(2008\)](#):

$$\begin{aligned} r_{t+1} &= a_r + b_r dp_t + \varepsilon_{t+1}^r \\ dp_{t+1} - \overline{dp} &= \rho(dp_t - \overline{dp}) + \varepsilon_{t+1}^{dp} \\ \Delta d_{t+1} &= a_d + b_d dp_t + \varepsilon_{t+1}^d, \end{aligned} \quad (12)$$

where $\overline{dp}$ is the mean of dp . The [Campbell and Shiller \(1988b\)](#) decomposition implies:

$$r_{t+1} = \kappa_0 - \kappa_1 dp_{t+1} + dp_t + \Delta d_{t+1}, \quad (13)$$

in which $\kappa_0 = \log(1 + e^{-\overline{dp}}) + \kappa_1 \overline{dp}$, $\kappa_1 = e^{-\overline{dp}} / (1 + e^{-\overline{dp}})$. [Equation \(12\)](#) and [\(13\)](#) imply that the predictive slope b_r satisfies:

$$b_r = 1 - \kappa_1 \rho + b_d.$$

The data-generating process captures the realistic features of the data. [Equation \(13\)](#) also implies that the shocks of the VAR system satisfy $\varepsilon_{t+1}^r = \varepsilon_{t+1}^d - \kappa_1 \varepsilon_{t+1}^{dp}$, which reflects the correlation between the return disturbance and the dp innovation studied by [Stambaugh \(1999\)](#).

Table 1
Simulation parameters

Parameters	Value
$\overline{dp}$	-3.2239
ρ	0.8911
$\kappa_1 = \frac{\exp(-\overline{dp})}{1+\exp(-\overline{dp})}$	0.9617
b_d	0
$b_r = 1 - \kappa_1 \rho + b_d$	0.1430

Table 1 shows the simulation parameters. We calibrate $\overline{dp}$ using the sample mean of dp . ρ is from an annual $AR(1)$ regression of dp , which corresponds to the forecasts of annual returns. We set $b_d=0$ following [Cochrane \(2008\)](#). These parameters imply a population predictive slope of $b_r=0.143$ in the simulations. The residuals in [Equation \(12\)](#) follow a multivariate normal distribution, whose covariance matrix is calibrated to its sample counterpart. The initial dp is drawn from its stationary distribution in each simulation sample.

We calculate the historical average of r using a 20-year rolling window, which is consistent with our empirical analysis in [Section 5](#). We impose the theoretical restriction that dp positively forecasts the expected return so that $\delta=1/A$ in [Equation \(5\)](#). When forecasting the return using our method, we standardize dp , such that a one-standard-deviation change in dp affects the return forecast by $1/A$ (e.g., the return forecast changes 1% if $A=100$). We use rolling 20-year windows to estimate the mean and standard deviation of dp for the standardization. We also obtain the OLS estimates from the 20-year rolling windows.

4.2. Simulation results

4.2.1. Bias-variance trade-off. We first analyze the bias-variance trade-off in [Equation \(10\)](#). The historical mean (HM) sets the slope to zero and ignores the predictor. We illustrate our forecast using a predictive slope $\delta=1/A$ with $A=50$. Therefore, a one-standard-deviation change in the predictor x affects the return forecast by two percentage points.

In [Figure 2](#), panel A shows that the predictive slopes used in both the HM and our forecast are biased. Since the population predictive slope is $b_r=0.1430$, the predictive slopes used in HM and our forecast both undershoot the population slope. Our forecast is closer to the true b_r and has a smaller bias than HM. The predictive slopes are deterministic in both HM and our forecast, as illustrated by their zero variance in panel B. The same zero variance, combined with a smaller bias, implies a smaller RMSE in panel C for the predictive slope from our method than the HM. We also run simulations using a population predictive slope different from the estimated value of $b_r=0.1430$. [Figure 2](#) similarly shows that our method has a smaller RMSE than the HM due to a smaller bias while maintaining the same variance.

We also examine the OLS regression often used in return predictions in [Figure 2](#). Panel B shows that the standard deviation of the OLS slope estimate

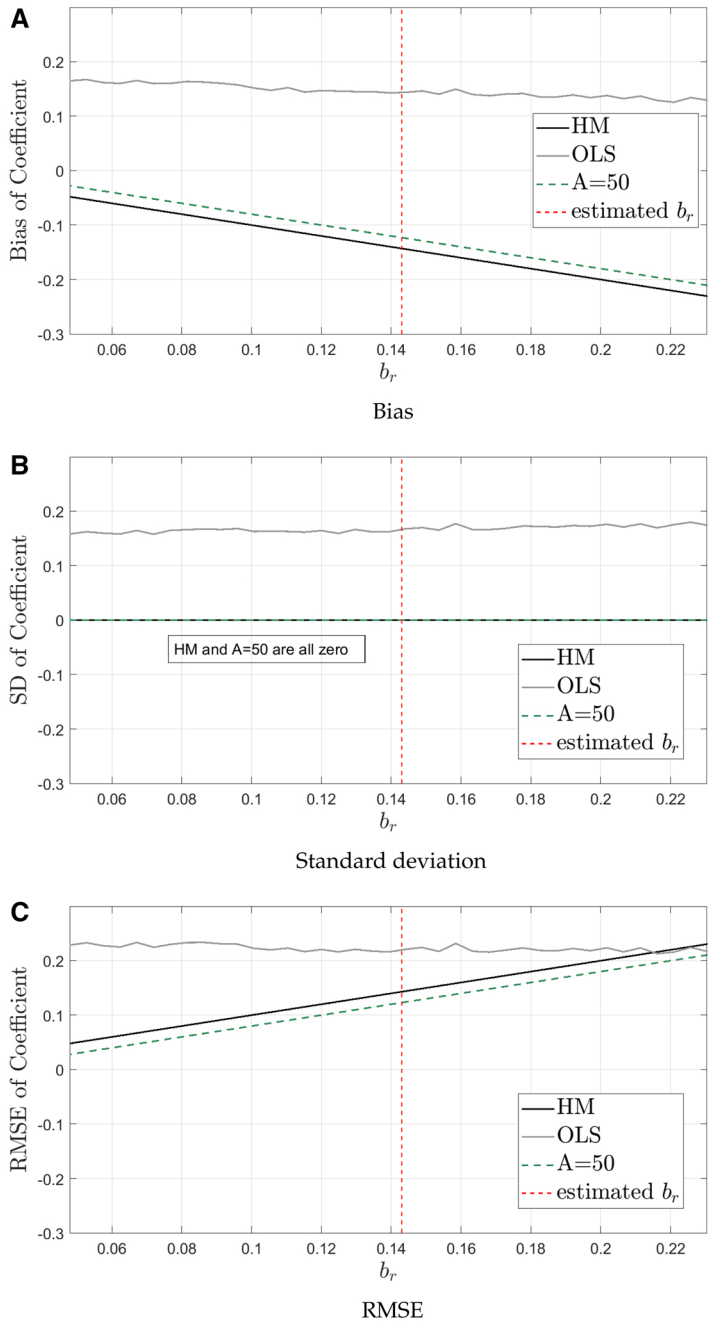


Figure 2
Simulation: Bias-variance trade-off
This figure shows the estimation bias, standard deviation, and root mean squared error (RMSE) of the predictive slope b_r using simulations. We compare our method using $A=50$ with the historical mean (HM) and OLS. In the data, $b_r=0.1430$. We conduct simulations for b_r in the range $[0.0479, 0.2306]$, which corresponds to the predictor's autocorrelation ρ within $[0.8, 0.99]$.

is almost as large as the bias of the HM in panel A. Therefore, OLS forecasts need not outperform the HM, consistent with the empirical results in [Goyal and Welch \(2008\)](#).

Figure 2, panel A, shows that the OLS slope estimate exhibits the bias documented by [Stambaugh \(1999\)](#). In [Internet Appendix B.1](#), we use simulations to examine forecasts using the unbiased OLS. The results show that the unbiased OLS produces better OOS MSE than the OLS, consistent with [Equation \(10\)](#). However, the unbiased OLS forecast inherits the high variance of the OLS forecast and, therefore, the instability of forecast performance, a key concern in [Goyal and Welch \(2008\)](#).

4.2.2. Out-of-sample forecast R^2 . After confirming the bias-variance advantage of our method relative to the HM, we turn to the OOS forecast. [Goyal and Welch \(2008\)](#) show that it is difficult to outperform the HM in OOS forecasts. We compute the OOS R^2 following [Goyal and Welch \(2008\)](#) and [Campbell and Thompson \(2008\)](#) to compare the performance of a forecast with the forecast based on the historical average:

$$R^2 = \frac{\text{MSE}_{\hat{\mu}} - \text{MSE}_{\hat{\mu} + \delta x}}{\text{MSE}_{\hat{\mu}}}, \quad (14)$$

where $\text{MSE}_{\hat{\mu}}$ is the MSE under the null hypothesis (using the HM $\hat{\mu}$), while $\text{MSE}_{\hat{\mu} + \delta x}$ is the MSE under the alternative hypothesis (using a forecast with the slope δ for the predictor x).

The R^2 statistic measures the reduction in the mean squared prediction error for a forecasting model under the alternative hypothesis relative to the forecast using the HM. When $R^2 > 0$, the forecast model under the alternative hypothesis outperforms the HM forecast.

Figure 3 shows the distribution of OOS R^2 from the simulations. It illustrates the results using the predictive slope $\delta = 1/A$ for A ranging from 50 to 500. All the distributions from our method are to the right of the origin, indicating a lower MSE than the HM, which is consistent with [Corollary 1](#). When A increases, the predictive slope δ converges toward zero. This has two effects. First, the outperformance relative to the HM becomes smaller since the predictive slope converges to zero, the slope of the HM.⁶ Second, the outperformance becomes less noisy since a smaller slope reduces the exposure to the variation of the predictor.

For comparison, Figure 3 also shows the distribution of OLS OOS R^2 . The OLS R^2 , even though centered on a positive R^2 , has a large fluctuation, consistent with the instability documented by [Goyal and Welch \(2008\)](#). The variation translates to a substantial probability of underperforming the HM

⁶ In Figure 3, the OOS R^2 tends to increase when A decreases. However, eventually if A is so small that $1/A$ substantially exceeds the true predictive slope, the performance deteriorates with further decreases in A .

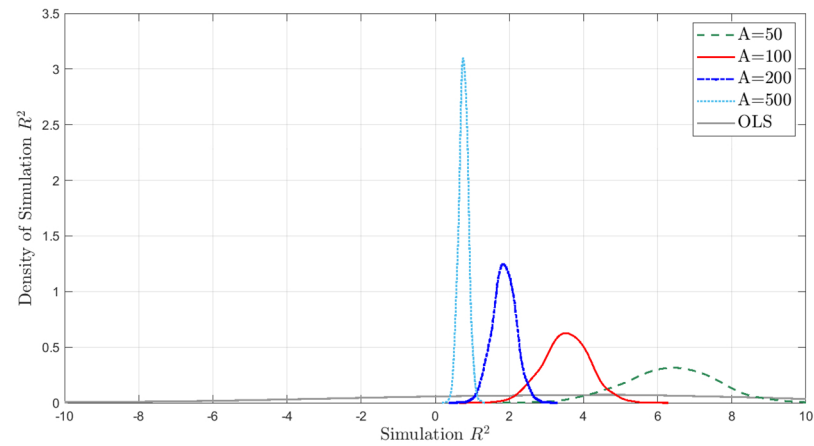


Figure 3
Simulation: OOS R^2
This figure shows the probability density function of out-of-sample R^2 using simulations. In each simulation, we compute the R^2 from a window consisting of 146 annual returns, which corresponds to the number of yearly dp observations in Table 2.

Table 2
Sample period of predictors

Predictor	Sample start year	Sample end year
bm	1921	2017
cape Shiller	1881	2017
cay	1945	2017
corpr	1926	2017
csp	1937	2002
de	1872	2017
dfr	1926	2017
dfy	1919	2017
dp	1872	2017
dy	1872	2017
ep	1872	2017
eqis	1927	2017
ik	1947	2017
infl	1914	2017
infl Shiller	1872	2017
ltr	1926	2017
lty	1919	2017
lty Shiller	1872	2017
ntis	1926	2017
svar	1885	2017
tbl	1920	2017
tms	1920	2017
Trcape Shiller	1881	2017

(i.e., $R^2 < 0$). Each simulated sample consists of 146 annual returns, matching the available yearly dp observations in Table 2. Even with such a long sample period, the simulation shows that OLS can underperform the HM with nontrivial probability.

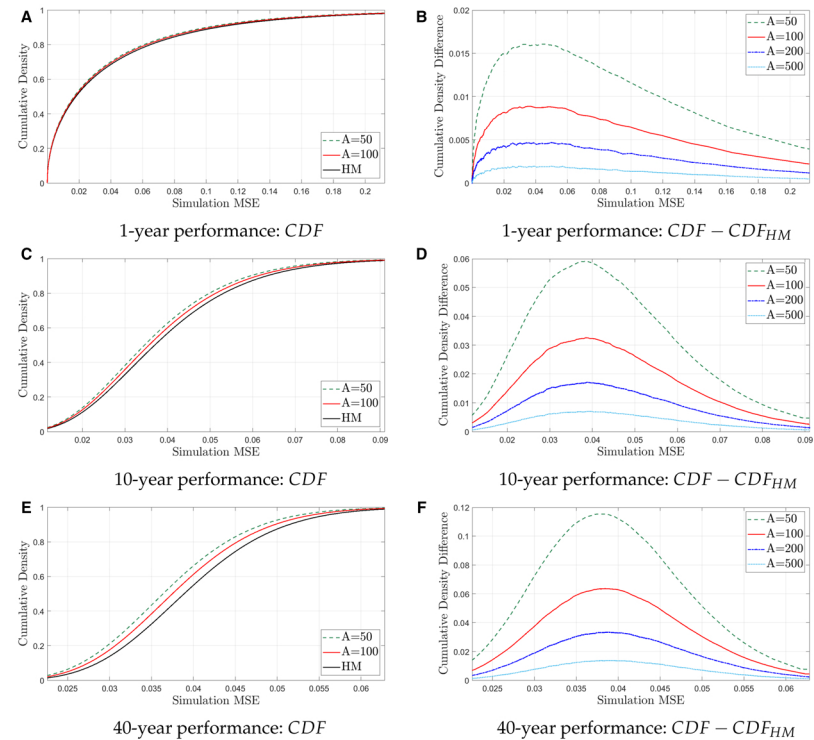


Figure 4
Simulation: first-order stochastic dominance

This figure shows the cumulative distribution function (CDF) of out-of-sample prediction MSE and its difference with the CDF of out-of-sample prediction MSE of the historical mean (HM) using simulations. We illustrate the forecast performance for 1-, 10-, and 40-year windows, during which we make successive 1-year return forecasts.

4.2.3. First-order stochastic dominance. Having confirmed our method’s superior OOS R^2 , we next study the first-order stochastic dominance shown in [Theorem 2](#). First-order stochastic dominance over the HM implies that the cumulative distribution function (CDF) of MSE using our method is everywhere above the CDF of MSE using the HM.

[Figure 4](#) shows the CDFs of MSE from the HM and our method using A from 50 to 500. We illustrate the forecast MSE for 1-, 10-, and 40-year windows, during which we make successive one-year return forecasts. We show the 1-, 10-, and 40-year CDFs in panels A, C, and E, respectively. We show the CDF difference between our method and the HM in panels B, D, and F.

The CDF using our method is above the CDF using the HM everywhere, confirming the first-order stochastic dominance in [Theorem 2](#). The first-order stochastic dominance is easier to discern using a longer horizon, where the MSE is less noisy (e.g., in [Figure 4](#), the horizontal axis range for the MSE from the 40-year window in panel F is substantially narrower than that from the

1-year window in panel B). At a shorter horizon, the range of MSE realization is more spread out, and the CDFs from our method and the HM are much closer (see [Figure 4](#), panel A). The first-order stochastic dominance still holds at the short horizon. However, we can pinpoint the CDF by simulating many sample paths. In the empirical analysis, where we estimate the CDF, a longer horizon can facilitate the inference of stochastic dominance.

4.3. Additional simulations

We conduct additional simulations to study the performance of our forecast method. We analyze uncertainties about the sign of the predictive slope in [Internet Appendix B.2](#), false positives associated with the conventional statistical significance levels in [Internet Appendix B.3](#), and variations of the predictive slope over time in [Internet Appendix B.4](#). Our forecast outperforms the historical average in the simulations.

5. Empirical Performance

In this section, we empirically evaluate the OOS R^2 and stochastic dominance of our method using the dividend-to-price ratio as the predictor. [Internet Appendix C](#) provides the results for other predictors and utility gains.

To determine the sign of the predictive slope in [Equation \(5\)](#), we first use the theoretically expected sign according to [Campbell and Thompson \(2008\)](#). This demonstrates our contribution to the literature where theory provides guidance on the sign. We then examine the empirical performance when setting the sign of the predictive slope according to estimates in previous empirical studies. We also conduct our own statistical analysis to determine the sign, and these results are in [Internet Appendix C.3](#).

In our empirical analysis, we normalize the predictor to have a variance of one and use a range of A values from 50 to 1,000 for [Equation \(5\)](#). These choices of A are around the 95% lower confidence bound from the estimates in [Campbell and Shiller \(1988a\)](#), which corresponds to $A=209$ as discussed in [Section 1.1](#). We do not focus on A values much larger than 1,000, as a huge A implies negligible predictability despite helping prevent our predictive slope from overshooting the population counterpart as required by [Theorem 2](#). For example, if $A=10,000$, a one-standard-deviation change in a predictor affects the annual expected return by only 0.01%. We require the predictability to have an empirically meaningful magnitude by capping the maximum A value.

5.1. Data

We analyze nonoverlapping annual market returns. We download the value-weighted index returns including dividends from the Center for Research in Security Prices (CRSP). We download data on the predictors, the risk-free rate, and the S&P 500 Index dividends and returns from Amit Goyal's website, which are extensions of the data used in [Goyal and Welch \(2008\)](#). We use the

CRSP value-weighted index return as the market return after 1926 and the S&P Index return as the market return before 1926.

Table 2 shows the predictors we examine and their sample periods. We include all the predictors in Goyal and Welch (2008): the aggregate book-to-market ratio (bm), consumption-wealth ratio (cay), corporate bond return (corpr), cross-sectional premium (csp), dividend-payout ratio (de), default return spread (dfr), default yield spread (dfy), dividend-to-price ratio (dp), dividend yield (dy), earnings-to-price ratio (ep), percent equity issuing (eqis), investment-to-capital ratio (ik), inflation (infl), long-term government bond rate of returns (ltr), long-term government bond yield (lty), net equity expansion (ntis), stock variance (svar), Treasury-bill rate (tbl), and term spread (tms).

In addition, we obtain four predictors from Robert Shiller's website: the cyclically adjusted price-to-earnings ratio ("cape Shiller" in Table 2), inflation ("infl Shiller"), long-term interest rate ("lty Shiller"), and cyclically adjusted total return price-to-earnings ratio ("Trcape Shiller").⁷ Campbell and Shiller (1988b) and Bunn and Shiller (2014) study the cyclically adjusted price-to-earnings ratio and the cyclically adjusted total return price-to-earnings ratio. Shiller's data on inflation and long-term government bond interest rates cover longer sample periods than the corresponding series from Goyal and Welch (2008), allowing us to study the forecast performance in earlier sample periods.

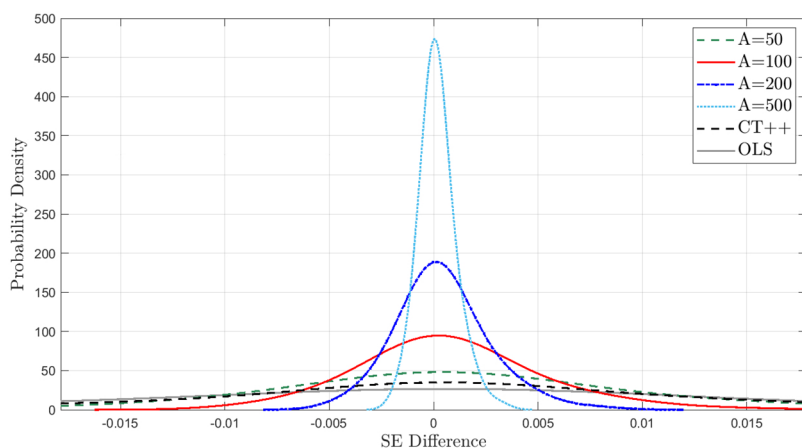
We standardize each predictor to mean zero and a standard deviation of one. The mean and standard deviation for the standardization in a given period are computed from the lagged 20-year window. We also use the lagged 20-year window to calculate the HM of the market return. Therefore, the first year of our forecast starts 20 years after the sample start year in Table 2. Campbell and Thompson (2008) and Goyal and Welch (2008) also use 20-year windows to compute the required parameters. For robustness, we also use a recursive expanding estimation window that includes all historical data up to the forecasting period, and the results are similar.

5.2. Out-of-sample forecast R^2

To begin, Figure 1 shows the cumulative performance of the annual return prediction. Specifically, it shows the cumulative OOS squared prediction errors of the HM minus the cumulative OOS squared prediction errors of another forecast. This is equivalent to figure 1 in Goyal and Welch (2008). Consistent with Goyal and Welch (2008), the OLS performance is volatile and substantially driven by its superior performance in the 1970s.

By contrast, the performance of our forecast method is much steadier. The cumulative squared error difference between the HM and our method steadily increases over time, revealing forecast performance gains across large parts of

⁷ Robert Shiller's data can be found online (<http://www.econ.yale.edu/~shiller/data.htm>).

**Figure 5****Distribution of annual performance using predictor dp**

This figure shows the probability density function (PDF) of out-of-sample performance of annual return prediction using the predictor dp . Specifically, it shows the PDF of annual squared prediction errors using historical mean minus the annual squared prediction errors using dp (denoted “SE Difference”). We estimate the PDF using kernel density. We illustrate the performance of our method for A between 50 and 1,000, OLS, and the CT++ method in Campbell and Thompson (2008).

the sample period. The smoothness of the performance over time is consistent with the bias-variance trade-off in Equation (10). The biggest forecast error using our method occurs during the dot-com boom in the late 1990s when the market return was high despite a low dp ratio. However, even for the dot-com episode, the performance loss is substantially mitigated using a smaller predictive slope of $\delta = 1/A$, or equivalently a large A , such as $A=200$ or higher.

When $A=200$, a two-standard-deviation change in dp affects the annual return forecast by 1%.⁸ This utilizes the predictor more conservatively than the OLS forecast, which can often deviate more aggressively from the HM. Such a conservative departure mitigates the performance loss relative to the HM during periods when the predictor does not work, such as the dot-com period. The forecast loss can be further mitigated using an even larger A , such as $A=500$ or 1,000. When using $A=1,000$, even large shocks in the dot-com episode or the 1970s oil shock do not have a pronounced impact on the forecast performance. Even for a large $A=1,000$, which affects the annual return forecast by a mere 0.1% for each standard deviation change in dp , the final cumulative performance at the end of the sample is still better than that of the OLS forecast.

Figure 5 shows the distribution of the yearly difference in squared prediction errors between the HM and various forecasts. We estimate the distribution

⁸ Since we standardize the predictors, a two-standard-deviation change in dp implies that the standardized predictor changes by 2, which affects the return forecast by $1/200 \times 2 = 1\%$ when $A=200$.

using nonparametric kernel density. The figure shows that the OLS forecast performance is volatile. The fat tail indicates that the OLS forecast has a large probability of doing either much better or much worse than the HM. By contrast, the yearly performance of our method is much more stable.⁹ On average, our method outperforms the HM. The performance gain from using our method relative to the HM is centered on the positive region in Figure 5, which can be seen more clearly from the positive cumulative gain in forecast performance in Figure 1.

The result in Figure 5 resembles the simulation R^2 distribution in Figure 3, both of which show that the forecast performance of our method is substantially less noisy than the regression approach. Unlike Figure 3, where the distribution of the simulation R^2 can be computed for a much longer window, we show in Figure 5 the yearly prediction error, where the performance gain of our method can be negative in some years by chance. Nonetheless, the result in Figure 5 confirms that a higher A is associated with a less noisy performance. Figure 1 also confirms the simulation result in Figure 3 that a higher A is associated with a smaller average performance gain relative to the HM.

Campbell and Thompson (2008) impose restrictions from economic theory to discipline the OLS forecasts. They use two methods: CT+ and CT++. CT+ requires the OLS slope to have the right sign according to economic theory. Otherwise, the predictive slope is replaced by zero. CT++ uses the CT+ predictive slope but further requires the expected return forecast to be positive. Otherwise, the expected return forecast is replaced by zero. Figure 5 confirms that the CT++ forecast is less noisy than the OLS forecast. Our method goes even further and has a zero variance for the predictive slope, the same as the historical average. As a result, Figure 5 shows that our forecast has even lower noise than with the CT++ forecast.

We next turn to the statistical significance of the performance gain of our method relative to the HM in Table 3. We calculate the p -value for the R^2 according to Diebold (2015) and adjust the standard errors for return heteroscedasticity. We also examine the p -value according to Harvey, Leybourne, and Newbold (1997), which provides a modified version of the Diebold and Mariano (1995) test statistic, and find similar results.

Table 3 contains the OOS R^2 using dp as the predictor. When $A=50$, the R^2 is 3.4% (p -value=.044). Using a larger A provides stronger statistical evidence for the OOS predictability. When A increases, the R^2 decreases but becomes statistically more significant, consistent with the simulation evidence in Figure 3. When $A=100$, each standard deviation change in dp affects the annual return forecast by 1%, and the resultant R^2 is 2.1% (p -value=.019). When $A=500$ and 1,000, the p -value is statistically significant at the 99%

⁹ We omit the distribution for $A=1,000$ in Figure 5 to make the distribution for smaller A and OLS easier to distinguish. The distribution for $A=1,000$ is qualitatively similar to that of other A but has a much higher peak probability density due to its smaller variance.

Table 3
OOS R^2 using dividend-to-price ratio

	A=1,000	A=500	A=200	A=100	A=50	OLS	CT+	CT++
R^2 (%)	0.2	0.5	1.1	2.1	3.4	0.2	1.3	1.8
	(.008)	(.009)	(.012)	(.019)	(.044)	(.477)	(.357)	(.286)

This table shows the out-of-sample R^2 for the dividend-to-price ratio (dp). This table also shows the out-of-sample R^2 using the OLS, as well as the [Campbell and Thompson \(2008\)](#) CT+ and CT++ methods. The p -value s are in parentheses. We calculate the p -value for the R^2 according to [Diebold \(2015\)](#) and adjust the standard errors for return heteroscedasticity. Following [Goyal and Welch \(2008\)](#), this table shows one-sided p -values.

level, with R^2 being 0.5% and 0.2%, respectively. In [Internet Appendix C.2](#), we illustrate the confidence bands for the cumulative performance in [Figure 1](#) and find that our method is statistically significantly better than the historical average from as early as the 1960s. In comparison, OLS generates a statistically insignificant R^2 . The CT+ and CT++ forecasts from [Campbell and Thompson \(2008\)](#) tend to have a higher R^2 and a better p -value than the OLS forecast. Our method even generates a more significant p -value than the CT+ or CT++ forecasts.

Although our method does not focus on exploiting the full magnitude of predictability, it generates a nontrivial gain in forecast performance. For example, in [Table 3](#), the R^2 is 2.1% when $A=100$. According to [Campbell and Thompson \(2008, Section 2\)](#), the correct way to judge the magnitude of R^2 is to compare it with the squared Sharpe ratio, which gives approximately the proportional increase in the expected return for an investor who recognizes the predictability. In the sample where dp is available, the annual market Sharpe ratio is 0.278, and the squared Sharpe ratio is 0.078. An R^2 of 2.1% implies that an investor can increase the expected return proportionally by $0.021/0.078=27\%$.

5.3. First-order stochastic dominance

Next, we empirically examine whether our forecast first-order stochastically dominates the HM according to [Theorem 2](#). We use the empirical distribution function to estimate the CDF of the OOS forecast MSE. We compute the MSE using 40-year windows. As simulations in [Section 4.2.3](#) show, first-order stochastic dominance is easier to discern over a longer horizon where the MSE is less noisy. [Figure 6](#) shows that the CDF from our method is everywhere higher than the CDF from the HM. Therefore, given any MSE threshold, a long-term investor using our method has a higher probability of achieving an MSE below the threshold than the HM.

The empirical distribution function does not smooth across observations. This results in a jagged CDF estimate but allows us to confirm stochastic dominance using the actual occurrences of loss events below a given threshold. In [Internet Appendix C.5](#), we use kernel smoothing, which also confirms that our forecast first-order stochastically dominates the HM.

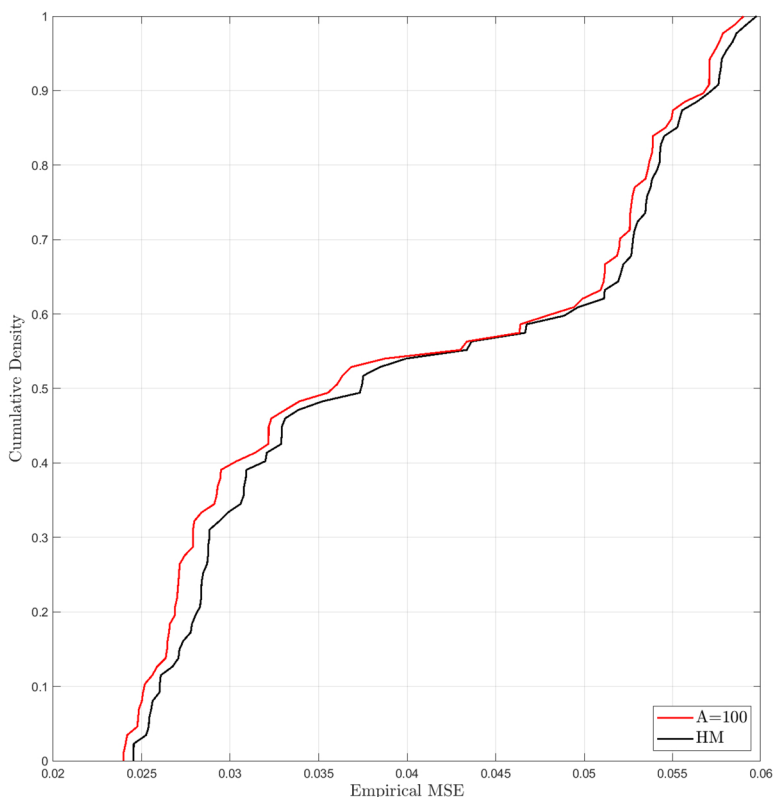


Figure 6
First-order stochastic dominance using predictor dp

This figure shows the cumulative distribution function (CDF) of out-of-sample prediction MSE using the predictor dp and the historical mean (HM), respectively. The CDF is estimated by the empirical distribution function of the out-of-sample prediction MSEs in 40-year rolling windows, during which we make successive 1-year return forecasts.

5.4. Informative estimates from previous studies

In this section, we investigate the performance of our method when relying on previously published empirical studies to determine the sign of the predictive slope, rather than relying solely on theoretical restrictions. Previous studies sometimes include sophisticated theoretical restrictions in their statistical analysis, which can make their estimates more informative than purely statistical methods. For example, [Cochrane \(2008\)](#) uses the absence of dividend growth predictability to help infer a positive predictive slope for the dividend price ratio, while [Lewellen \(2004\)](#) uses stationarity to restrict the slope estimate.¹⁰ Using existing estimates also helps us to avoid data mining and is easy to implement.

¹⁰ [Ang and Piazzesi \(2003\)](#) examine the fixed-income market and impose no-arbitrage conditions in their estimation. [Aït-Sahalia and Duarte \(2003\)](#) contain a review of the literature on estimation under shape restrictions.

Campbell and Shiller (1988a) document a significantly positive predictive slope for the dividend-to-price ratio using the sample period 1871–1986. When using a positive slope with a range of different A values, including $A=209$ as implied by the 95% lower bound from Campbell and Shiller (1988a), we observe a positive OOS R^2 in the sample period starting from 1987. We also use the 95% lower bound from Campbell and Shiller (1988a) without standardizing the predictor to a variance of one, and the OOS performance gain relative to the historical average remains similar. This shows that the sign and the lower confidence bound in Campbell and Shiller (1988a) add value to the OOS forecasts using our method.

The estimates in Campbell and Shiller (1988a) are statistically significant at the 95% level. An investor may find that a 90% confidence level, another common significance threshold, will provide acceptable confidence. This has the advantage of earlier detection of the sign. A back-of-the-envelope calculation implies that 90% significance requires a sample period that is only 70% as long as the sample period needed for 95% significance.¹¹ We confirm positive OOS R^2 values starting from earlier years using a variety of A . The detailed results using the 95% lower bound from Campbell and Shiller (1988a) with various start years are suppressed for brevity, but the good OOS performance can be seen from the stable OOS performance in Figure 1, where $A=200$ approximates the 95% lower bound in Campbell and Shiller (1988a).

6. Conclusion

The debate on equity premium predictability is shifting from in-sample statistical significance to real-time out-of-sample (OOS) predictability. Real-time prediction uses fewer observations than full-sample analysis, exacerbating the noise of regression estimates relative to the historical average.

We provide a novel methodology to forecast the equity premium. Our method fuses the zero variance feature of the historical average with the bias reduction from a predictor. Our approach offers a new paradigm to affirm the OOS predictability of a predictor with high power. We demonstrate theoretically that our forecast first-order stochastically dominates the historical average and has a better mean squared forecast error than the historical average when the population predictive slope is not zero. Our paradigm forms a sufficient and necessary (with near certainty) condition for the OOS predictability of a given predictor, as long as there are sufficient data to evaluate the OOS performance.

Empirically, our method reveals strong evidence of real-time predictability. Fifteen of the 23 predictors we study can individually generate positive OOS R^2 at the 90% significance level and eight are 95% significant. Our approach

¹¹ This is because $(1.645/1.96)^2 = 70\%$, where 1.645 is the cutoff value at the 90% significance. Our back-of-the-envelope calculation assumes that the population slope is stable over time.

drastically improves the stability of forecast performance relative to existing forecasts, as shown in Figure 1. This solves the key concern of instability in Goyal and Welch (2008) (e.g., see the abstract and figure 1 in their paper).

Our study shows that the lower confidence bound of a statistically significantly positive predictive slope probably has a smaller bias than the zero slope employed by the historical average. Similarly, the upper confidence bound of a negatively significant slope can likely add value to OOS forecasts. We confirm that such confidence bounds in previous studies help subsequent forecasts, which contributes direct evidence of OOS predictability. Our method can improve further with future advances in time-series econometrics. A better estimator or a more efficient test will improve the implementation of our method due to better sign estimation and larger bias reduction from the tighter confidence bounds.

The historical average is the *de facto* benchmark for the expected market return. The simplicity of our forecast, its *ex ante* (near) guarantee of outperforming the historical average, and its superior performance in empirical analyses contribute toward a benchmark for the time-varying expected market return.

Code Availability: The replication code and data are available in the Harvard Dataverse at <https://doi.org/10.7910/DVN/9VNJUN>.

References

- Ait-Sahalia, Y., and J. Duarte. 2003. Nonparametric option pricing under shape restrictions. *Journal of Econometrics* 116:9–47.
- Ang, A., and M. Piazzesi. 2003. A no-arbitrage vector autoregression of term structure dynamics with macroeconomic and latent variables. *Journal of Monetary Economics* 50:745–87.
- Ball, R. 1978. Anomalies in relationships between securities' yields and yield-surrogates. *Journal of Financial Economics* 6:103–26.
- Bossaerts, P., and P. Hillion. 1999. Implementing statistical criteria to select return forecasting models: What do we learn? *Review of Financial Studies* 12:405–28.
- Bunn, O. D., and R. J. Shiller. 2014. Changing times, changing values: A historical analysis of sectors within the us stock market 1872-2013. Working Paper, Barclays Bank PLC.
- Campbell, J. Y. 1987. Stock returns and the term structure. *Journal of Financial Economics* 18:373–99.
- . 2018. *Financial decisions and markets: A course in asset pricing*. Princeton, NJ: Princeton University Press.
- Campbell, J. Y., and R. Shiller. 1988a. The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* 1:195–228.
- . 1988b. Stock prices, earnings, and expected dividends. *Journal of Finance* 43:661–76.
- . 1998. Valuation ratios and the long-run stock market outlook. *Journal of Portfolio Management* 24:11–26.
- Campbell, J. Y., and S. Thompson. 2008. Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies* 21:1509–31.

- Campbell, J. Y., and T. Vuolteenaho. 2004a. Bad beta, good beta. *American Economic Review* 94:1249–75.
- . 2004b. Inflation illusion and stock prices. *American Economic Review* 94:19–23.
- Campbell, J. Y., and M. Yogo. 2006. Efficient tests of stock return predictability. *Journal of Financial Economics* 81:27–60.
- Clark, T. E., and K. D. West. 2006. Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of Econometrics* 135:155–86.
- . 2007. Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics* 138:291–311.
- Cochrane, J. H. 1991. Production-based asset pricing and the link between stock returns and economic fluctuations. *Journal of Finance* 46:209–37.
- . 2008. The dog that did not bark: A defense of return predictability. *Review of Financial Studies* 21:1533–75.
- Diebold, F. X. 2015. Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold-mariano tests. *Journal of Business & Economic Statistics* 33.
- Diebold, F. X., and R. S. Mariano. 1995. Comparing predictive accuracy. *Journal of Business & Economic Statistics* 13:253–63.
- Dong, X., Y. Li, D. E. Rapach, and G. Zhou. 2022. Anomalies and the expected market return. *Journal of Finance* 77:639–81.
- Eleswarapu, V., and M. Reinganum. 2004. The predictability of aggregate stock market returns: Evidence based on glamour stocks. *Journal of Business* 77:275–94.
- Eliasz, P. 2005. Optimal median unbiased estimation of coefficients on highly persistent regressors. Working Paper, Princeton University.
- Elliott, G., A. Gargano, and A. Timmermann. 2013. Complete subset regressions. *Journal of Econometrics* 177:357–73.
- Fama, E. F. 1965. The behavior of stock-market prices. *Journal of Business* 38:34–105.
- . 1981. Stock returns, real activity, inflation, and money. *American Economic Review* 71:545–65.
- Fama, E. F., and K. French. 1989. Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics* 25:23–49.
- Fama, E. F., and G. Schwert. 1977. Asset returns and inflation. *Journal of Financial Economics* 5:115–46.
- Goyal, A., and I. Welch. 2008. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies* 21:1455–508.
- Goyal, A., I. Welch, and A. Zafirov. 2024. A comprehensive 2022 look at the empirical performance of equity premium prediction. *Review of Financial Studies* 37:3490–557.
- Harvey, D., S. Leybourne, and P. Newbold. 1997. Testing the equality of prediction mean squared errors. *International Journal of Forecasting* 13:281–91.
- Hodrick, R. J. 1992. Dividend yields and expected stock returns: alternative procedures for inference and measurement. *Review of Financial Studies* 5:257–86.
- Huang, C.-F., and R. H. Litzenberger. 1988. *Foundations for financial economics*. Amsterdam, the Netherlands: North-Holland.
- Jansson, M., and M. J. Moreira. 2006. Optimal inference in regression models with nearly integrated regressors. *Econometrica* 74:681–714.
- Johnson, T. L. 2019. A fresh look at return predictability using a more efficient estimator. *Review of Asset Pricing Studies* 9:1–46.

- Keim, D. B., and R. F. Stambaugh. 1986. Predicting returns in the stock and bond markets. *Journal of Financial Economics* 17:357–90.
- Kelly, B., S. Malamud, and K. Zhou. 2024. The virtue of complexity in return prediction. *Journal of Finance* 79:459–503.
- Kelly, B., and S. Pruitt. 2013. Market expectations in the cross-section of present values. *Journal of Finance* 68:1721–56.
- Kothari, S., and J. Shanken. 1997. Book-to-market, dividend yield, and expected market returns: A time-series analysis. *Journal of Financial Economics* 44:169–203.
- Lamont, O. 1998. Earnings and expected returns. *Journal of Finance* 53:1563–87.
- Lemaréchal, C. 2012. Cauchy and the gradient method. *Doc. Math. Extra*.
- Lettau, M., and S. Ludvigson. 2001. Consumption, aggregate wealth, and expected stock returns. *Journal of Finance* 56:815–49.
- Lettau, M., and S. V. Nieuwerburgh. 2008. Reconciling the return predictability evidence. *Review of Financial Studies* 21:1607–52.
- Lewellen, J. 2004. Predicting returns with financial ratios. *Journal of Financial Economics* 74:209–35.
- Mandelbrot, B. 1967. The variation of some other speculative prices. *Journal of Business* 40:393–413.
- Menzly, L., T. Santos, and P. Veronesi. 2004. Understanding predictability. *Journal of Political Economy* 112:1–47.
- Pástor, L., and R. F. Stambaugh. 2009. Predictive systems: Living with imperfect predictors. *Journal of Finance* 64:1583–628.
- Polk, C., S. Thompson, and T. Vuolteenaho. 2006. Cross-sectional forecasts of the equity premium. *Journal of Financial Economics* 81:101–41.
- Pontiff, J., and L. D. Schall. 1998. Book-to-market ratios as predictors of market returns. *Journal of Financial Economics* 49:141–60.
- Rapach, D. E., J. K. Strauss, and G. Zhou. 2010. Out-of-sample premium prediction: Combination forecasts and links to the real economy. *Review of Financial Studies* 23:821–61.
- Stambaugh, R. F. 1999. Predictive regressions. *Journal of Finance* 54:375–421.
- van Binsbergen, J. H., and R. S. J. Koijen. 2010. Predictive regressions: A present-value approach. *Journal of Finance* 65:1439–71.