Forecast-Agnostic Portfolios*

November 20, 2025

Abstract

We introduce forecast-agnostic (FA) portfolios that exhibit out-of-sample market-timing ability without relying on estimated predictive coefficients. These portfolios go long or short the market based on the level of a predictor variable, thereby avoiding the instability and estimation error that undermine traditional market-timing strategies. Despite using predictor variables that typically deliver negative out-of-sample (OOS) R^2 values (Goyal et al., 2024), FA portfolios deliver significantly positive alphas on average. We explain these seemingly contradictory phenomena by interpreting regression coefficients as portfolio returns: genuine predictability is necessary for high portfolio returns, whereas achieving a positive OOS R^2 additionally requires the ability to forecast the returns on the FA portfolios themselves. Simulations show that FA alphas have power to detect predictability that extends beyond in-sample diagnostics and the OOS R^2.

Keywords: Market timing, Efficient Market Hypothesis, Return Predictability

JEL codes: G12, G14, G40

*The authors do not have conflicts to disclose.

1 Introduction

The search for successful market-timing strategies is likely as old as organized equity markets themselves. As early as 1900, Bachelier posited that stock prices followed an arithmetic random walk, implying that price changes are unpredictable. The notion of unpredictable returns became standard market lore, with Fama's 1970 Efficient Markets Hypothesis and Malkiel's 2003 *Random Walk Down Wall Street*. There is good reason to believe that markets are unpredictable: should investors perceive that stocks have an unusual advantage over bonds, they will quickly shift their portfolios toward stocks, thereby causing stock prices to rise and eliminating such predictability. Nonetheless, Campbell and Shiller (1988) found an economically and statistically significant relation between scaled-price measures and future excess returns. A vast empirical and theoretical literature followed.

Subsequently, Goyal and Welch (2008) evaluated the many variables that the literature had identified as having market-timing ability. They showed that in-sample significance was fleeting, and out-of-sample (OOS) performance was almost uniformly poor. Campbell and Thompson (2008) showed that performance improved when one imposes the economic requirement that the forecast of the excess return – which equals the conditional equity premium under rational expectations – must be nonnegative. The literature incorporated both the Goyal and Welch OOS metric and the Campbell and Thompson adjustment, making it all the more surprising that newly proposed predictor variables continue to perform poorly (Goyal et al., 2024).

Poor performance in an out-of-sample forecast exercise need not indicate that a predictor variable is useless. We propose an investment-performance-based metric, which we refer to as the *forecast-agnostic* alpha. According to this metric, the majority of the 46 predictor variables studied in Goyal et al. (2024) would have been useful to investors. In fact, the forecast-agnostic alphas are significantly positive overall. Simulation results further show that, under the null of no predictability, alphas are on average zero. Under the null of predictability, alphas are on average positive—unlike the OOS R^2, which remains negative. The forecast-agnostic alphas are hence more powerful than OOS R^2 even if used on a standalone basis, responding to Campbell and Thompson (2005)'s critique on out-of-sample statistics.

The construction of the forecast-agnostic portfolios proceeds as follows. In the first stage, we take a predictor variable, de-meaned using data available up to that point, and form the portfolio with weight on the market equal to the de-meaned predictor variable. We then scale the portfolio weight by the volatility of this portfolio's returns up to that point and multiply by a predetermined target volatility. Lastly, we change the sign of the portfolio if the expanding-window predictive regression coefficient of stock market returns on the predictor values is negative, and Winsorize at 300% and -300%, as leverage close to these positions is likely to be unachievable. The result is a portfolio that assigns a time-varying weight to the market based on a predictor's value, while simultaneously aiming for market neutrality and a target volatility. The portfolio is forecast-agnostic, as it only utilizes the sign of the predictive regression. Unlike the OOS R^2, it is not *forecast-driven*. It does not use expanding window beta coefficients or the forecasts that they lead to.

Although not the focus of their paper, Goyal et al. (2024) also examine the investment performance of predictor variables, specifically by looking at what they term "bull" and "bear" portfolios that reduce the dependence on the forecasting coefficient beta. They find that the portfolios' performance only weakens the use case of the predictor variables. The reason is that Goyal et al. examine the average return and ask whether it is greater or less than what one would have achieved investing in equities. This is too stringent a test for a portfolio that takes on varying degrees of risk, given that the excess return on the market, is positive most of the time. Rather, the test of whether a long-short portfolio adds value is whether it has an alpha. If the investor is holding the market, then the relevant question is whether the portfolio has an alpha relative to the CAPM.[1]

To better understand the properties of the forecast-agnostic alphas, as well as the OOS R^2 statistics, we conduct Monte Carlo simulations in which we seek to match the level of

[1]Ultimately, under both the Goyal et al. (2024) interpretation and ours, the investor holds a portfolio that consists of the market, together with the riskfree asset with time-varying weights in the market. However, under our interpretation, the investor tilts around a market portfolio of 100% depending on the predictor variable, whereas under their interpretation, the tilt is around a weight closer to zero.

predictability seen in the data. Given this level of predictability, a forecast-agnostic approach to out-of-sample performance is superior to a forecast-driven approach. Our simulations do not require any degree of parameter instability; they reflect sampling noise. The in-sample t-statistics indicate a noticeable, but not high, level of statistical significance. By using expanding-window beta coefficients, forecast-driven portfolios, and R^2 coefficients tend to load on noise, and thus generate poor performance. One can think of the forecast-driven/agnostic distinction as another example of the variance-bias tradeoff. By restricting how portfolios can use the information in the predictive coefficient, we introduce bias, but we greatly reduce noise, in a way that makes our metric more reliable in the type of sample we have. Simulations show, however, that if predictability were far stronger and we had ten times the amount of data, a forecast-driven approach would add value to a forecast-agnostic one.

A deeper question is why we do not see sufficient predictability to allow forecast-driven portfolios to outperform. We show that the beta coefficient in a predictive regression equals the excess return on a portfolio – not exactly our forecast-agnostic portfolio – but one that is highly similar, though uses the full sample. If the out-of-sample R^2 were close to the in-sample R^2, then this beta coefficient would also be close, namely, the mean excess return on what is approximately the forecast-agnostic portfolio would be near a constant throughout the sample. If it were a constant, it would be an arbitrage opportunity, as it implies an infinite information ratio. Thus, the greater the precision in the beta estimate, the closer the out-of-sample results come to the in-sample results, and the closer the forecast-agnostic portfolio approaches an arbitrage opportunity. The absence of arbitrage is consistent with predictable returns on the market. However, these returns cannot be predicted in a sufficiently predictable way.

In line with the reasoning that arbitrage is at work to limit predictability, we show that the forecast-agnostic alpha associated with a strategy becomes statistically insignificant post-publication of that strategy, echoing Mclean and Pontiff (2016). In constrast, point estimates are similar pre-sample, suggesting that it is indeed arbitrage, and not instability in alpha estimates, that drives variation across samples.

The above reasoning presupposes that return predictability stems from a failure of rational expectations, and is not, say, a result of time-varying risk or risk aversion. Most likely, observed return predictability has more than one underlying cause. However, the degree to which it is a failure of rational expectations, as opposed to time-varying risk, is reflected in the results of several additional tests that we run. We show that while imposing the economic constraint that the risk premium cannot go below zero, as recommended by Campbell and Thompson (2008), does improve performance, nearly the same degree of improvement arises from imposing the symmetric positive constraint. Thus, the improvement appears to arise not so much from imposing an economic constraint, but from reducing the variance in the estimation of the forecasting coefficient. Similarly, for variables where there is a strong economic prior for what the beta should be, imposing this prior also improves out-of-sample performance. The superior performance could arise because the model in question is correct, but it also could be that the model gives a "reasonable" value of the predictive coefficient, and imposing this coefficient leads to reduced noise, even as it introduces bias, similarly to our forecast-agnostic alphas. Finally, post-publication fall-off in results also occurs for predictor variables with risk-based explanations, as does the lack of pre-sample fall-off. Taken together, these patterns cast further doubt on the arguments that these variables capture risk-based fluctuations in the equity premium and that they solely reflect statistical bias, but are consistent with the hypothesis that they got arbitraged away post-publication.

Besides the papers mentioned above, our paper also relates to early work by Kandel and Stambaugh (1996a), who emphasize that it is returns, not R^2s, that ultimately matter for investors. Other papers that make use of this bias-variance trade-off include Rapach et al. (2010), who find market-timing capabilities for a portfolio that equally weights component portfolios. Like our forecast-agnostic alphas, equal-weighting throws away what seems to be information but is in fact just noise. Wachter and Warusawitharana (2009) show that imposing an economic prior on the R^2 leads to shrinkage of predictive coefficients and superior out-of-sample performance of a forecast-driven portfolio. More recently, Kelly et al. (2024)

show that, with a properly chosen shrinkage parameter, the portfolios formed based on their high-complexity forecasts generate positive information ratios even though the high-complexity forecasts also produce negative OOS R^2. Our approach shows that this dichotomy also reflects what might be called a virtue of simplicity – the use of a portfolio that estimates fewer regression coefficients than even a linear projection.

2 Stock return predictability: Summary of existing evidence

The standard predictive regression assumes

$$r_{t+1} = \alpha + \beta x_t + v_{t+1} \tag{1}$$

where r_{t+1} is the excess return over the Treasury-bill at the monthly or quarterly horizon, where x_t is a variable known at time t, and v_{t+1} is mean zero and is uncorrelated with any time-t variables.[2]

For an investor choosing between an aggregate stock portfolio and the Treasury bill, and assuming (unrealistically of course) that parameters are known, $\alpha + \beta x_t$ is an input into portfolio choice. If it is not matched one-for-one with the conditional variance of r_{t+1}, or with changes in risk aversion, then x_t will lead the investor to "time the market," potentially earning far more than holding a static position. It is this possibility that accounts for the enduring interest in this simple regression.

In practice, variables are not known, and so an investor estimates (1) using ordinary least squares (OLS), and then use the OLS estimates
alphat and $\hat{\beta}$ to form a forecast for stock returns and as an input into portfolio choice.[3]

[2] Throughout, we use excess returns in levels, so that r_{t+1} is the level return on the market, minus the level return on the riskfree asset.

[3] OLS is consistent for α and β but not unbiased (Stambaugh, 1999), with the degree of bias depending on the persistence of x_t and the correlation between shocks to x_t and v_{t+1}. One can incorporate the statistical estimation directly into the decision problem using Bayesian methods, given prior beliefs on α and β; see Kandel and Stambaugh (1996b).

Over time, the literature has uncovered many such variables. Goyal et al. (2024) organize 46 of these conveniently, expanding the data both backwards and forwards in time (the first two columns of Table 1 given the start and date). The next two columns replicate their in-sample results. Barring researcher error, all variables at one time were statistically significant in at least some specification at standard levels (otherwise they would not have been published). However, only 5 of the 46 achieve t-statistics of above 2.[4] This tells us that stock return predictability studies usually do not hold up over time.

The next 4 columns display a second problem. Goyal et al. (2024) ask whether the predictors are better forecasts than the expanding window mean. Specifically, they form the out-of-sample R^2:

$$\text{OOS}R^2 = 1 - \frac{\sum_{t=0}^{T} r_t - (\hat{\alpha}_{t-1} + \hat{\beta}_{t-1} x_{t-1})}{\sum_{t=0}^{t} r_t - \bar{r}_{t-1}}$$

Here r_t is the excess return realized at time t and $\hat{\alpha}_t$ and $\hat{\beta}_t$ are estimated using an expanding window of data , as is the sample mean $\bar{r}_t$.[5] A reduced mean-squared-error (MSE) from using the predictor variables will imply a numerator that is small relative to the denominator, and a positive OOS R^2. The smaller the numerator, the closer the OOS R^2 comes to 1. In contrast, if the numerator is larger than the denominator, the predictor variable adds noise, and the OOS R^2 is negative.

Column 6 Table 1 reports the OOS R^2. Echoing the main conclusion of Goyal et al. (2024), this table shows that the OOS performance tends to be poor. Slightly more than half of the variables have negative OOS R^2s, implying that the predictor variable adds noise in a forecasting exercise.[6] Following Goyal et al. (2024), we also consider OOS R^2 statistics under the adjustment proposed by Campbell and Thompson (2008) (the CTOOS R^2). In equilibrium, assuming investors are risk averse, the equity premium cannot be negative.

[4] Assuming these variables are chosen at random and errors are independent, one might expect to see 2 or 3 above 2 by chance. However, neither of these assumptions are satisfied, and this number alone is not decisive regarding rejecting the null of no predictability.

[5] The computation uses all the data available, but never less than twenty years.

[6] We also use data going back to 1871 to form the expanding window mean, whereas Goyal et al. (2024) use the data over the time period for which the predictor is available. Our method leads to slightly lower OOS R^2 statistics.

Thus any negative values of $\hat{\alpha} + \hat{\beta} x_t$ can only represent noise. Table 1 shows that imposing this constraint leads OOS R^2 statistics that are less negative on average. The table considers two modification of the CTOOS R^2: the first uses a minimum of ten rather than 20 years of data, the second starts in 1946. While individual values of the OOS R^2 change, neither make much of a difference, with the median, in both cases, remaining negative.

3 Forecast-agnostic alphas

In the cross sectional return forecasting literature, it is common practice to form long-short portfolios according to a predictor's values across stocks. For example, Fama and French (1993) create the HML factor, which takes a long position in the 30% of stocks with the highest book-to-market ratios and shorts the 30% of stocks with the lowest book-to-market ratios. Fama and MacBeth (1973) create unit-beta-exposure, market-neutral portfolios that long high-beta stocks and short low-beta stocks with weights linear in betas. Unlike the calculation of an R^2, these portfolios require no explicit forecasts for stock returns. In fact, they only require the predictors underlying those forecasts, and their returns are used to evaluate a predictor's success.

We construct analogous portfolios for time-series predictors. Our construction is not new to the literature—its logic goes back to Fama and MacBeth (1973), but it involves modifications to i) circumvent look-ahead bias in the portfolio construction process and ii) target a fixed risk budget. We then ask whether such portfolios created for the same set of predictors in Table 1 can *add value* in a portfolio.[7] despite generating overwhelmingly negative out-of-sample R^2. Specifically, we examine whether these portfolios have positive CAPM alphas.[8]

For a generic predictor x, we construct a zero-cost market-timing strategy that rebalances positions in i) aggregate stock market and ii) the risk-free bond. The strategy implements

[7]This is a less stringent criterion than that of Goyal et al. (2024), which is whether these predictors generate portfolios that individually beat the market.

[8]See Baks et al. (2001) for why α is the relevant metric rather than the raw return.

two rules with data available in real time. First, it places proportional weights on the market relative to the predictor value, with a target mean of zero. Second, the portfolio weight is scaled so that the strategy has a target volatility of, say, 5% per month. This volatility target should come from investors' preference, and we use 5% for the purpose of illustration. As shown in Table B1, our main results are not sensitive to this choice.

Let x_t be the predictor value at the end of period t. Then, the unscaled, unsigned portfolio weight on the market is simply the predictor value demeaned by its expanding-window mean, $x_{t-1} - \overline{x}_{t-1}$. The expanding-window mean $\overline{x}_{t-1}$ is the mean of x computed with data up to $t-1$. We then regress the market excess return on this demeaned predictor and multiply the *sign* of the resulting expanding-window coefficient with the unsigned portfolio weight to arrive at the signed portfolio weight raw_w_{t-1}. This is the simplest way to let the data speak on the direction of the trades. The unscaled portfolio return is this weight multiplied by the market excess return in period t. Note that this procedure uses nothing of the predictive coefficient other than its sign.

The expanding-window unscaled portfolio volatility $\bar{\sigma}^x_{t-1}$ is the unscaled portfolio's historical return volatility up to $t-1$. The scaled portfolio weight is $raw\_w_{t-1}\frac{0.05}{\bar{\sigma}^x_{t-1}}$. We further truncate the weight on the market at -3 and 3 to avoid unrealistic positions.[9] The portfolio's construction involves no look-ahead bias, and the scaling can be adjusted according to an investor's risk appetite. We refer to these portfolios as forecast-agnostic (FA) portfolios, which are robust to estimation errors in regression coefficients.

Table 1 computes forecast-agnostic αs for the predictor variables that Goyal et al. (2024) consider. For comparison, the table also in-sample statistics, and out-of-sample R^2 statistics, which we compute in four different ways.

The last two columns of Table 1 tabulate alphas and their individual t-statistics of the FA portfolios for the 46 predictors. Two points are worth noting. First, the alphas have a positive mean and median across the predictors. Second, while the OOS R^2 can sometimes lead to catastrophic results (for example, `vrp`), negative alphas are small in magnitude. Ta-

[9]Since overly high leverages are unattainable in practice, such truncation is always (though perhaps implicitly) present when trading in the real world.

ble 2 further evaluates the overall statistical significance of these alphas. Column "Panel 1926" conducts a portfolio-month panel regression of these portfolio returns on the contemporaneous market returns and a constant. It shows that the average alpha across the portfolio-month observations, as captured by the coefficient on the constant, is 16.1 basis points per month. The associated t-statistic of 4.57 indicates that the alpha is highly statistically significant at the 1% level. Column "Panel 1946" starts the evaluation from 1946 and conveys a similar message. Column "TS Avg 1926" creates a meta portfolio averaging across all available FA portfolios and evaluates its alpha with a time-series regression. Compared to the previous two rows, it underweights observations in cross sections in which more predictors are available. This average alpha appears higher at 36 basis points per month but comes with a lower t-statistic of 2.87. This is owing to volatile portfolios returns at the beginning of the sample, which, as fewer predictors were available then, are overweighted in this method. The last column "TS Avg 1946" excludes this period and behaves similarly to its panel counterpart.

Second, the alphas of the FA portfolios are, perhaps surprisingly, dissimilar to the OOS R^2s. Table 3 show that, even with the same burn-in period of 20 years, the alphas have correlations in the 0.2 range with the OOS R^2s in Table 1. This point is further illustrated in Figure 2, which shows that sorting the OOS R^2s does not sort the FA alphas. The FA alphas more positively correlate with the in-sample R^2s, though at 0.46, this correlation is not high.

The FA alphas paint a starkly different picture from that drawn from the OOS R^2s. The OOS R^2s tell us that the 46 predictors on average fail to forecast stock market returns for investors in real time, even if they employ the predictors at the inception of their availability. The FA alphas tell us that the very same set of predictors add significant value to such investors. This difference is important, as it leads to completely different evaluations of the market return predictability literature as a whole. Below, we provide a thorough explanation for it.

4 Why is it possible for a predictor to have a negative OOS R^2 and a positive alpha?

FA alphas and OOS R^2s evaluate the same set of predictors differently because they ask different things from those predictors. Positive FA alphas require the predictor to be genuine—it needs to be a part of market return—and, in our implementation, that investors know which direction to trade. Positive OOS R^2s further require investors to know how much they will make in the trade. This additional self-awareness requirement turns out to be too high given the amount of available data.

To see where this self-awareness requirement comes from, note first that the R^2 computation involves explicit return forecasts and therefore the estimation of an regression coefficient on the predictor in real time. This coefficient is a representation of the predictor's historical performance: a time-series regression of the market excess return mkt_t on the predictor x_{t-1} and a constant yields a coefficient of $\beta^x = \frac{1}{T}\sum_t \frac{(x_{t-1}-\bar{x})}{\frac{1}{T}\sum_t (x_{t-1}-\bar{x})^2} mkt_t = \frac{1}{T}\sum_t w_{t-1} mkt_t$ on x. Here, $\bar{x}$ is the mean of the predictor x in this sample, and T is the total number of periods. Note that β^x is the mean excess return of a portfolio that rebalances between positions in the aggregate stock market (which yields an excess return of mkt_t) and the risk-free bond (which yields an excess return of 0). The month-t portfolio weight on the stock market $w_{t-1} = \frac{(x_{t-1}-\bar{x})}{\frac{1}{T}\sum_t (x_{t-1}-\bar{x})^2}$ is i) linear in the predictor's value x_{t-1}, ii) on average 0, so that the portfolio is on average market neutral in this sample, and iii) scaled by $\frac{1}{T}\sum_t (x_{t-1}-\bar{x})^2$, so that the portfolio on average has unit exposure to x (i.e., $\frac{1}{T}\sum_t \frac{(x_{t-1}-\bar{x})}{\frac{1}{T}\sum_t (x_{t-1}-\bar{x})^2} x_{t-1} = 1$) in this sample.[10]

Using this coefficient to form a quantitative forecast of the market excess return implicitly assumes that the performance of the predictor's coefficient portfolio is sufficiently stable so that its past performance reliably predicts its future performance at any point in history.

[10]If x does not have an economically interpretable scale itself, the scaling in step iii) has no economic interpretation either. Our FA portfolio is designed to mimic this coefficient portfolio, except that we replace the in-sample mean with the expanding mean for implementability, adjust the scaling in iii) for economic interpretability, and impose a position constraint that is inevitable in reality.

This is a potentially paradoxical assumption, as it may imply too high a T-statistic, Sharpe ratio, and R^2. In the most extreme case, if we want the expanding-window coefficient to equal the in-sample coefficient so that the OOS R^2 equals the in-sample R^2, we need the coefficient portfolio's return volatility to be 0, its Sharpe ratio to be infinite, the predictor's in-sample T-statistic to be infinite, and the in-sample R^2 to be 1. Realistically, given the in-sample R^2 and T-statistics we observe, using (without any adjustment) this expanding-window regression coefficient to compute a quantitative forecast is likely counterproductive, and the OOS R^2s are likely poor even if the predictors' are genuine. We formally illustrate this point in the simulation section below.

Our FA portfolio, on the other hand, requires no quantitative prediction of the predictor's performance.[11] The size of the bet on the market is based on a constant risk appetite (5% in our example) and does not change with the past returns of the predictor's coefficient portfolio. This is consistent with the construction of cross sectional factors (e.g., HML, SMB, MOM, etc.), in which the stock weights also do not change with the past factor returns. Removing the step of predicting the predictor's performance eliminates potential errors in that step. The improvement is large, as the amount of available data is small (again, more on this in the simulation section). Therefore, the FA portfolio alphas appear better than the OOS R^2s overall. A well-behaved FA portfolio does depend on the stability of the predictor's mean and volatility over time, and, if you consider the predictors altogether, such stability appears better supported by the data.

To more directly illustrate why the FA portfolio alphas can behave so differently from the OOS R^2s, we use `vrp` as an example, because it has an extreme negative OOS R^2 of -19.51%. We plot in Figure 1 the univariate expanding-window regression coefficients of the stock market excess return on the predictors and a constant. The predictors are signed so that the full-sample coefficients are positive, and scaled to have a standard deviation of 1 to make the coefficients comparable. As explained above, the difference between the OOS R^2 and the in-sample R^2 occurs because these coefficients are not constant. As the sample

[11] If one is evaluating predictors *individually*, it is not obvious why such quantitative prediction is beneficial in the first place.

size grows over time, the coefficients become increasingly stable towards the right end of the plot, resembling the full-sample coefficients, which are the last observation in each series. Hence, the largest deviations occur at the beginning of the sample. From this figure, it is immediately apparent that vrp is the predictor with the most persistent large deviations in the coefficients. It is precisely the predictor with the lowest OOS R^2.

It is worth emphasizing that we do not imply that the metric of OOS R^2s used in Goyal and Welch (2008) is inappropriate. On the contrary, we believe it is a good, consistent, and general metric. We do want to point out that there is more to this metric. Importantly, poor OOS R^2s alone do not imply that the corresponding predictor cannot be successfully employed in real time trading. The FA portfolio we propose is a simple approach that both complements the OOS R^2s and fills an obvious gap between the practices of two literatures studying the time-series and the cross-sections of stock returns.

Another point worth reiterating is that the FA portfolios are not new to the literature. In fact, Goyal et al. (2024) consider similar portfolios in their Table 4. Despite considering similar portfolios, they arrived at the opposite conclusion from ours. In fact, they conclude that "the investment performance was poor" for 20 out of the 29 new predictors they individually evaluated.[12] The reason behind the discrepancy is that Goyal et al. (2024) asked a different economic question: can these predictors individually generate a trading strategy that outperforms the stock market? The question is appropriate if investors choose one of the predictors or the market to invest in. However, as Goyal et al. (2024) acknowledge, the FA portfolios or similar strategies can be viewed as "overlays" on top of the market and can be invested along with it. From that perspective, the appropriate benchmark is whether these portfolios add alphas.

[12]For 8 out of the other 9 predictors, the evaluation does not appear favorable, even though the author did not use the quoted sentence. For the old predictors in Goyal and Welch (2008), Goyal et al. (2024)'s investment performance evaluation was overall unfavorable.

5 Simulation

5.1 An out-of-sample statistic with more power: responding to Campbell and Thompson (2005)

As discussed in the previous sections, requiring an OOS R^2 to be the same as the IS R^2 leads to requirements that the predictor's in-sample T-statistic is infinite and that its in-sample R^2 is 1. These are clearly unrealistic, because we observe an average in-sample absolute T-statistic of 1.06 and an average in-sample R^2 of 32 basis points. However, is it too much to require the OOS R^2 to be positive on average? What typical OOS R^2 values should one expect, given the observed amount of in-sample predictability? To answer these questions, we conduct a simulation exercise.

Our simulations consist of 47 monthly time series: stock market excess returns and 46 predictors, with the same start and end months as the real data used in Table 1. In Panel A of Table 4, the predictors follow AR(1) processes with a mean of zero and AR(1) coefficients estimated from the real predictor data. Both the estimation and the simulation are done at the same frequency as the real data—monthly, quarterly, or annual. The innovations of the predictors are i.i.d. normal shocks with a volatility of 30 basis points multiplied by the square root of one minus the square of the AR(1) coefficients. The value of 30 basis points is the population volatility of all the predictors. It governs the strength of predictability in our simulations and is set to roughly match real data. In the first 3 columns with the header "no predictability," the simulated returns equal the full-sample average monthly stock market excess return (58 basis points) plus i.i.d. normal shocks with a mean of zero and a volatility equal to the full-sample volatility of monthly stock market excess returns (4.78%). In the next 3 columns, the simulated returns equal the aforementioned two terms plus 46 additional terms, each being the product of one of the 46 predictors and a beta coefficient randomly drawn from a uniform distribution between -1 and 1.

In the "no predictability" simulations, all predictors are spurious. In the "with predictability" runs, they are all genuine, even though predictors with a beta farther from zero are more important than those close to zero. Panel B presents similar sets of simulations, except that the beta coefficients and the predictors' volatility are estimated from data rather than randomly drawn and manually set. The innovation terms in both the predictors and the returns have time-varying conditional variance estimated with a GARCH(1,1) model on the data.

In each simulation, we compute 5 statistics across the 46 simulated predictors: i) the average absolute value of Newey-West T-statistics, ii) the average in-sample R^2, iii) the average OOS R^2, iv) the average OOS R^2 after applying the positive-sign restriction from Campbell and Thompson (2008), and v) the average alphas of the FA portfolios. The quantities are computed using exactly the same approach as in Tables 1 and 2. We report the 5th, 50th, and 95th percentiles across 1000 simulations in the first 6 columns of Table 4, and report the corresponding value computed from real data in the last column.

In the with-predictability simulations of Panel A, the medians of the average T-statistics and average in-sample R^2 roughly match those computed from real data. This is by design—the dispersion of the simulated predictors is set to match these two moments, which are computed to capture the amount of in-sample predictability. In the next two rows, we see that the medians of the average OOS R^2 and the average CTOOS R^2 are both negative at -0.22% and -0.10%, even when the predictability is genuine. This shows that, given the amount of in-sample predictability that we observe in real data, we likely won't observe positive OOS R^2s. In Panel B, the message is similar. Here, the average T-statistics and average in-sample R^2 appear somewhat higher than in real data, even though the predictors' dispersion and their regression coefficients are estimated. This is likely because the GARCH(1,1) model does not fully capture the heteroskedasticity structure in the data. We nonetheless observe negative median average OOS R^2s, with or without the positive-sign restriction.

These negative medians reflect a known power issue of the OOS R^2s. Notably, it was pointed out by Campbell and Thompson (2005) in an early draft of their 2008 paper. Goyal and Welch (2008) respond to this critique by arguing that the OOS R^2s should not be used alone but rather as a conditional diagnostic. Specifically, they should be examined only when the in-sample diagnostics are significant. Here, even though the predictability is genuine, the average in-sample T-statistics are only 1.34 and 1.67. As the predictability is too weak for the in-sample T-statistics to reliably detect, it is not surprising that the OOS R^2s look poor. We further demonstrate this point in Table 5, which shows that if the strength of predictability increases, the average T-statistics rise above the conventional threshold of 2, and the average OOS R^2s become positive.

While we agree with Goyal and Welch (2008)'s response, we provide a simpler response to Campbell and Thompson (2005) in which an OOS statistic is used on a standalone basis. In the last rows of Panels A and B, we compute the median average alphas of the FA portfolios. When the predictors are genuine, these medians are positive and close to the observed values in real data. Furthermore, even the 5th percentiles across the simulations are positive, showing good power of this statistic in detecting the genuine predictability scenario. When the predictors are spurious, we see that the average alphas are roughly centered around 0 as expected, and the 95th percentiles are not much higher than the 5th percentile values among the with predictability simulations. Even if we impose a 95% type-I error rate on the null of no predictability, we can achieve a power that is not much lower than 95% for the alternative hypothesis of real predictability if we compute the average alpha of the FA portfolios.

Goyal and Welch (2008) make an important point, which is that look-ahead bias in parameter estimation is especially important in a time-series setting and should be carefully removed before a predictor can be fairly evaluated. Out-of-sample diagnostics address this point by removing the look-ahead bias—both OOS R^2s and FA alphas are such diagnostics. Using the average FA alphas as an example, we show that such statistics can and should be used independently of any in-sample statistic.

5.2 How much data and how strong a predictor do we need?

An important economic difference between the FA alphas and the OOS R^2s is whether they try to distinguish between strong and weak predictors. The FA portfolios allocate to each predictor an equal risk budget of 5% per month. This budget does not change with the predictor's past performance. On the other hand, OOS R^2s use forecasts computed from expanding-window regression coefficients, which become more disperse if the predictor historically does well. We show that when return predictability is as weak as the data suggest, giving more dispersion to historically stronger predictors is counterproductive. However, this is not generally true. In fact, with i) more data or ii) stronger predictability, one can more reliably distinguish stronger predictors from weaker ones. We demonstrate this point by conducting our simulations with longer samples and stronger predictability.

In these simulations, we form an alternative portfolio for each predictor using a similar approach to that in Campbell and Thompson (2008). Specifically, we first compute the expanding-window coefficients estimated by regressing stock market excess returns on the demeaned predictor (with its expanding-window mean) and the constant. We then take the product of the demeaned predictor and its coefficient, and multiply this product by 200. The multiplier of 200 comes from $\frac{1}{2\times0.05^2}$. Here, 2 is risk aversion, 0.05 is roughly the monthly volatility of the stock market, and the multiplier converts expected market returns to positions on the market for a risk-averse investor optimizing across the market and the risk free asset. The risk aversion of 2 is chosen such that the investor would on average choose to place roughly 100% weight on the market. The resulting portfolio weight is then winsorized at 3 and -3, as in our FA portfolios. We refer to these portfolios as forecast-driven (FD) portfolios.

As FA and FD portfolios may have different exposures to the predictors, we cannot directly compare their returns or their CAPM alphas. We therefore compute the two-factor alpha the FD portfolios controlling for the market and the FA portfolios, as well as the 2-factor FA alphas controlling for the market and the FD portfolios. We continue to compute the OOS and CTOOS R^2s.

Table 5 reports simulation results with 2-30 times as much data as the original simulation in Panel A of Table 4, although the 20 year estimation periods and the CRSP starting point are fixed. We see that, with 5 times the data, the OOS R^2s are comfortably positive. The average T-statistics at that point is 2.35. With 20 times the data, the FD portfolios start to add alphas on top of the FA portfolios, and the average T-statistics is about 4.41. This point is visually illustrated by the top panel of Figure 3.

In Panel B, we simulate with 10 times the data and stronger predictability. That is, the dispersion of the predictors increases from 30 to 180 basis points across the columns. Here, we see that with 3 times stronger predictability, FD portfolios start to add more alphas on top of the FA portfolios than the other way around. The average t-statistic is around 7.47 at that point. This point is visually illustrated by the lower panel of Figure 3.

Overall, we show that, with a moderate in-sample T-statistic of around 2.3, we can expect to see positive OOS R^2s on average. The practice of chasing past performance (by estimating the expanding-window coefficients) can be productive with more data or stronger predictability, although that requires T-statistics that are rather high.

6 The pre- and post-sample alphas

Table 6 shows that post-publication, the alphas are essentially zero. So investors should not trade on predictors obtained from academic papers. This, however, can be either because too many investors trade on those papers and "destroy" the return predictability (a la Mclean and Pontiff (2016)) or because the statistical patterns are spurious. Thanks to the data extension of Goyal et al. (2024) and the monthly resolution of the GFD data, we can compute the pre-sample alphas of the FA portfolios. Such pre-sample alphas are valuable because they reflect statistical bias but not post-publication arbitrage activities. Column (1) shows that the average pre-sample alphas are in fact somewhat higher than the in-sample alphas. This speaks against the notion that the predictability patterns are owing to statistical bias, as they work pre-sample (which is out-of-sample). The post-sample alphas are positive but insignificant, suggesting that post-publication arbitraging activities do destroy the alphas. Hence, a marginal investor should not trade on predictors in published papers, but may consider trading on predictors she privately discovers with a similar standard as in published papers.

If a predictor forecasts stock market excess returns via predictable resolution of risks, then, if investors' attitudes toward such risks do not change, the predictability should not decline post-publication. Column (2) tests this prediction by showing the post-sample alphas for predictors that are published with a rational explanation. It shows that the post-sample alphas are also not statistically distinguishable from zero (although higher than the alphas computed with all predictors). In contrast, they are significantly different from the pre-sample alphas at 1% level. One possible explanation is that investors who do not care about the modeled risk become the marginal investors post-publication. Another explanation is that the proposed risk-based explanation is incorrect in the first place.

7 Obtaining higher OOS R^2s—the value of statistical and economic constraints

One important point made by Campbell and Thompson (2008) is that constraints help improve predictors' out-of-sample performance. In particular, equity premium forecasts below 0 are theoretically hard to justify and should therefore be truncated at zero. Comparing the columns CT20 and OOS20 in Table 1 confirms that implementing this constraint substantially improves the OOS R^2s, although the mean and median of these R^2s remain negative even with this constraint.

A critical assumption behind this constraint is that the predictability arises from predictable resolution of risks. The large post-publication decline shown in the previous section casts this assumption in doubt. It is possible that the non-negative constraint improves forecasting performance for the predictors not because the negative forecasts are economically implausible, but rather because extreme forecasts, whether too high or too low, are statistically implausible. We then implement a symmetric constraint, winsorizing from above return forecasts exceeding 2 times the expanding-window mean of market excess returns. Table 7 shows that this positive-end constraint alone improves the median OOS R^2 from -0.25% to -0.09%.[13] This is not far from the improvement to a median of -0.03% produced by the non-negativity constraint. Implementing both constraints further increases this median to -0.01% (and the mean to 0.03%). The improvement from this additional constraint appears appreciable but insufficient.

A less appreciated constraint proposed in Campbell and Thompson (2008) is that imposed by an economic model. Unlike the non-negativity constraint, such a constraint is not uniformly applicable. Even when they apply, the constraints are often not straightforward to obtain. We extend the work in Campbell and Thompson (2008) and apply the economic model-based constraints to 8 predictors.

The most straightforward case in this table is `rsvix`. In the model of Martin (2017), its predictive coefficient is simply 1/12 to convert annual forecasts to monthly units (and the coefficient of the constant term is 0). This constraint can then be applied to implied volatility-based predictors similar to `rsvix`, specifically `vp`, `impvar`, and `vrp`. This is because implied volatility indices such as the VIX differ from Martin (2017)'s SVIX only in weights across options and are similar in their levels and variability. Even though these three papers do not have a model that directly provides the coefficient, one can impose the same constraint as in Martin (2017) based on their similarity. As `impvar` appears to be in monthly units, the predictor's coefficient is 1 and that on the constant term is 0. `Vrp` and `vp` are in annualized volatility percentage-point units. We convert them back to fractional monthly variance units and impose a coefficient of 1. We further constrain the coefficient on the constant term to be the expanding-window mean of the squared log market excess returns to be consistent with SVIX.

[13]The mean OOS R^2 remains poor at -0.57%, though slightly improved from -0.62% without any constraint. This is entirely owing to vrp, in which an extreme negative (rather than positive) predictor value tanked its forecasting performance.

Another set of predictors on which we can impose economic constraints are the valuation ratios. Campbell and Thompson (2008) use a drifting steady-state valuation model to express the expected stock market return next period as the sum of the level of dividend yield and the average dividend growth, estimated as the expanding-window average of the aggregate plowback ratio multiplied by the expanding-window average of aggregate return on equity. We combine this relation with equation 5.87 in Campbell (2017), adding half of the expanding-window mean of the squared log market excess returns:

$$\frac{1}{12}(d_t/p_{t-1} + (1 - \overline{de}_t)\overline{roe}_t) + \frac{1}{2}\overline{\sigma^2}_t.$$

This imposes a coefficient of 1/12 on the predictor dy along with a constant of $\frac{1}{12}(1 - \overline{de}_t)\overline{roe}_t + \frac{1}{2}\overline{\sigma^2}_t$. The same constraint can be placed on the predictor `dp`. Following Campbell and Thompson (2008), the equity premium forecasts based on ep and bm are:

$$\frac{1}{12}(ep_t\overline{de}_t + (1 - \overline{de}_t)\overline{roe}_t) + \frac{1}{2}\overline{\sigma^2}_t,$$

and

$$\frac{1}{12}(1 + \overline{de}_t(bm_t - 1))\overline{roe}_t + \frac{1}{2}\overline{\sigma^2}_t,$$

respectively. They impose coefficients of $\frac{\overline{de}_t}{12}$ and $\frac{\overline{de}_t\overline{roe}_t}{12}$ on the predictors, and a coefficient of $\frac{1}{12}(1 - \overline{de}_t)\overline{roe}_t + \frac{1}{2}\overline{\sigma^2}_t$ on the constant term.

Table 8 shows that the economically motivated constraints improve the predictors' forecasting performance overall. In our view, the value of an economic model in this return forecasting exercise is not having a strong prior about the corresponding predictor—that could in fact be counterproductive. Rather, it is to fulfill the self-awareness requirement in the forecasting exercise. That is, to inform investors about the coefficient on the predictor even without any historical data. When the amount of data is insufficient to detect a potentially weak predictor—which is generally the case in our return-predicting setting—this approach is especially valuable.

The literature has shown considerable interest in statistical methods that generate better OOS R^2s. Examples include a simple average across forecasts (Rapach et al. (2010)), shrinking toward a conservative lower bound (Li et al. (2025)), isolating the fundamental component (Lin et al. (2025)), and varying the strength of the shrinkage according to the predictor (Goyal et al. (2024), Kelly et al. (2024)). Even though we push the adoption of FA portfolio alphas, we contribute to the return-forecasting literature by showing that economically motivated coefficient values are an additional set of possible shrinkage targets.

8 Conclusion

We create forecast-agnostic (FA) portfolios for predictors of aggregate stock market returns. These portfolios target a given volatility level and avoid numerical return forecasts, thus becoming robust to errors in estimated regression coefficients. We show that, given the low strength of predictability we observe, the FA alphas are more powerful statistics than OOS R^2, the classic performance metric used by the literature to evaluate the performance of time-series predictors. Circumventing the estimation of regression coefficients allows the FA portfolios to perform better than the forecast-driven trading strategies proposed in Campbell and Thompson (2008), even when the sample were longer than what is available. Making the model less complex than linear regression adds value to investors who time the market.

References

Bachelier, L. (1900). *Théorie de la spéculation*. PhD thesis, Faculté des Sciences de Paris.

Baks, K. P., Metrick, A., and Wachter, J. (2001). Should investors avoid all actively managed mutual funds? a study in bayesian performance evaluation. *Journal of Finance*, 56(1):45–86.

Campbell, J. (2017). *Financial decisions and markets: A course in asset pricing*. Princeton University Press, Princeton, NJ.

Campbell, J. and Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies*, 21(4):1509–1531.

Campbell, J. Y. and Shiller, R. J. (1988). The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies*, 1(3):195–228.

Campbell, J. Y. and Thompson, S. B. (2005). Predicting the Equity Premium Out of Sample: Can Anything Beat the Historical Average?

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

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3–56.

Fama, E. F. and MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, 81(3):607–636.

Goyal, A. and Welch, I. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4):1455–1508.

Goyal, A., Welch, I., and Zafirov, A. (2024). A comprehensive 2022 look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 37(11):3490–3557.

Kandel, S. and Stambaugh, R. F. (1996a). On the Predictability of Stock Returns: An Asset-Allocation Perspective. *The Journal of Finance*, 51(2):385–424. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.1996.tb02689.x.

Kandel, S. and Stambaugh, R. F. (1996b). On the predictability of stock returns: An asset allocation perspective. *Journal of Finance*, 51:385–424.

Kelly, B., Malamud, S., and Zhou, K. (2024). The Virtue of Complexity in Return Prediction. *The Journal of Finance*, 79(1):459–503. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.13298.

Li, K., Li, Y., Lyu, C., and Yu, J. (2025). How to Dominate the Historical Average. *The Review of Financial Studies*, 38(10):3086–3116.

Lin, Y., Wu, C., Zhou, G., and Zhu, S. (2025). Market Return Decomposition and Equity Risk Premium Predictability.

Malkiel, B. G. (2003). *A random walk down Wall Street*. W. W. Norton and Company, Inc., New York, NY.

Martin, I. (2017). What is the expected return on the market? *Quarterly Journal of Economics*, 132(1):367–433.

Mclean, R. D. and Pontiff, J. (2016). Does academic research destroy stock return predictability? *Journal of Finance*, 71(1):5–32.

Rapach, D. E., Strauss, J. K., and Zhou, G. (2010). Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy. *The Review of Financial Studies*, 23(2):821–862.

Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics*, 54:375–421.

Wachter, J. A. and Warusawitharana, M. (2009). Predictable returns and asset allocation: Should a skeptical investor time the market? *Journal of Econometrics*, 148:162–178.

Table 1: R^2 in forecasting monthly stock market excess returns

	In-sample				Out-of-sample R^2				FA	
Predictor	Start	End	t	IS R^2	CT20	OOS20	CT10	CT20 Post-46	Alpha	Alpha-t
vp	1990-01	2021-12	1.64	1.94	4.72	4.72	-0.46	4.72	0.01	0.05
impvar	1996-01	2023-08	0.63	0.28	0.49	0.49	-2.89	0.49	0.00	-0.01
vrp	1990-01	2023-12	0.23	0.10	-1.00	-19.51	-0.22	-1.00	-0.11	-0.25
lzrt	1926-01	2024-11	0.96	0.23	0.02	-0.11	0.04	0.02	0.14	1.29
ogap	1926-01	2024-11	-0.98	0.40	0.72	0.72	0.28	0.72	0.08	3.73
wtexas	1926-01	2024-11	-1.71	0.34	0.36	0.15	0.32	0.36	0.15	0.95
sntm	1975-07	2023-11	0.22	0.01	-0.41	-0.41	-0.67	-0.41	-0.10	-0.51
ndrbl	1958-02	2024-11	-1.50	0.40	-0.58	-0.84	-0.08	-0.58	-0.07	-0.35
skvw	1926-07	2024-11	0.00	0.00	-0.58	-0.64	-0.46	-0.58	-0.04	-0.56
tail	1926-07	2024-11	0.59	0.03	-0.16	-0.16	-0.29	-0.16	-0.14	-1.15
fbm	1936-06	2024-11	-0.17	0.01	-0.89	-0.89	-0.84	-0.89	0.05	0.40
dtoy	1926-01	2024-11	-0.06	0.00	-0.19	-0.19	-0.23	-0.19	0.05	0.83
dtoat	1926-01	2024-11	-0.75	0.15	0.15	0.15	0.11	0.15	0.06	1.06
ygap	1953-04	2024-11	0.54	0.06	-0.87	-1.35	-0.78	-0.87	0.10	0.33
rdsp	1926-07	2024-11	1.04	0.63	-1.08	-1.21	-1.62	-1.08	-0.04	-1.22
rsvix	1996-01	2023-08	0.62	0.39	-0.02	-0.02	-3.17	-0.02	-0.63	-0.90
tchi	1951-01	2024-11	1.49	0.37	0.14	-0.02	-0.23	0.14	0.58	2.49
avgcor	1926-03	2024-11	1.20	0.27	0.37	0.37	-0.02	0.37	0.08	0.73
shtint	1973-01	2024-11	-1.33	0.32	-0.07	-0.07	-0.79	-0.07	-0.14	-0.49
disag	1981-12	2024-11	-0.09	0.00	-0.48	-0.90	-0.51	-0.48	-0.35	-0.99
ntis	1926-12	2024-11	-1.54	0.48	-0.54	-0.54	-0.68	-0.54	-0.01	-0.25
tby	1920-01	2024-11	-2.01	0.34	0.61	0.45	-0.66	0.48	0.33	2.12
dp	1871-01	2024-11	0.52	0.06	-0.09	-0.59	-0.09	-0.20	0.22	1.12
dy	1871-02	2024-11	0.87	0.21	0.02	-0.56	0.02	-0.57	0.14	0.69
ep	1871-01	2024-11	1.60	0.15	0.10	0.07	0.10	0.09	0.26	1.41
de	1871-01	2024-11	0.08	0.00	0.01	-0.52	0.01	-0.09	0.60	2.61
svar	1885-02	2024-11	0.21	0.02	-0.06	-1.62	-0.06	-0.52	0.24	1.61
lty	1919-01	2024-11	-1.73	0.21	0.51	-0.07	-0.59	0.54	0.41	1.90
ltr	1926-01	2024-11	1.48	0.18	-0.47	-0.49	-0.30	-0.47	0.10	0.60
tms	1920-01	2024-11	1.24	0.13	0.26	0.24	-0.32	0.09	0.10	0.79
dfy	1919-01	2024-11	0.53	0.20	-0.35	-0.39	-0.29	-0.02	-0.04	-0.96
dfr	1926-01	2024-11	0.54	0.06	-0.25	-0.32	-0.21	-0.25	0.08	0.60
infl	1913-02	2024-11	-1.59	0.28	0.44	0.45	0.12	0.56	0.15	1.78
bm	1921-03	2024-11	1.17	0.54	-0.86	-1.15	-0.53	-1.38	0.03	0.59
pce	1953-12	2024-11	-3.73	1.57	1.09	0.92	1.18	1.09	0.65	2.89
govik	1947-03	2024-11	1.76	0.24	-0.34	-0.71	-0.35	-0.34	0.12	1.11
crdstd	1990-06	2024-11	-1.15	0.64	-1.04	-2.41	-0.28	-1.04	-0.25	-0.97
ik	1947-03	2024-11	-3.76	1.49	0.88	0.77	0.85	0.88	0.56	2.58
cay	1952-03	2024-11	0.24	0.01	-0.63	-0.87	-0.87	-0.63	0.20	0.82
skew	1951-12	2020-11	0.28	0.01	-0.29	-0.43	-0.58	-0.29	0.24	1.32
accrul	1965-12	2024-11	0.05	0.00	-0.78	-0.81	-0.95	-0.78	-0.02	-0.06
cfacc	1965-12	2024-11	-0.84	0.10	-0.70	-0.70	-0.76	-0.70	0.20	1.04
gpce	1947-12	2024-11	-3.39	1.26	0.17	0.16	0.67	0.17	0.24	1.50
gip	1926-12	2024-11	-0.34	0.03	0.06	0.06	-1.07	0.06	-0.01	-0.38
house	1929-12	2024-11	0.38	0.03	0.11	0.11	-0.12	0.11	0.07	1.13
eqis	1927-12	2024-11	-2.32	0.70	0.27	0.15	-0.15	0.27	0.09	1.69
Mean				0.32	-0.03	-0.62	-0.40	-0.06	0.10	
Median				0.21	-0.06	-0.25	-0.29	-0.12	0.08	

This table shows performance statistics of the 46 predictors in Goyal et al. (2024) forecasting monthly market returns in excess of the risk-free rate. Column CT20 reports out-of-sample R^2 imposing the Campbell and Thompson (2008) non-negativity constraint and starting the evaluation sample 20 years after the inception of the predictor or Jan 1926, whichever comes later. Column OOS20 removes the non-negativity constraint. Column CT10 starts the evaluation sample 10 years after the inception of the predictors. CT20 Post-46 starts the evaluation sample 20 years after the inception of the predictor or Jan 1946, whichever comes later. Column Alpha reports the alphas of the FA portfolios, and the next column t reports their associated t-statistics. The t-statistics are computed with Newey-West standard errors following Goyal et al. (2024). The R^2s and the alphas are monthly and in percentage units.

Table 2: Average alphas of FA portfolios

	(1) Panel 26	(2) Panel 46	(3) TS Avg 26	(4) TS Avg 46
mkt	-0.139***	-0.099***	-0.285***	-0.097***
	[-5.00]	[-7.92]	[-3.00]	[-7.63]
Alpha	0.161***	0.112***	0.360***	0.118***
	[4.57]	[4.40]	[2.87]	[4.48]
Sample size	35,395	33,670	1,187	947

Column (1) and (2) conduct the following predictor-monthly level panel regression: $r_{i,t}^{FA} = \alpha + \beta mkt_t + \epsilon_{i,t}$. Here, $r_{i,t}^{FA}$ is the FA portfolio return for predictor i in month t, and mkt_t is the stock market return in excess of the risk-free rate in month t. Column (1) starts the evaluation 20 years after the inception of the predictor or Jan 1926, whichever comes later. Column (2) uses Jan 1946 rather than Jan 1926. Columns (3) and (4) conduct the monthly time-series regression: $avg\_r_t^{FA} = \alpha + \beta mkt_t + \epsilon_t$, in which $avg\_r_t^{FA}$ is the average of $r_{i,t}^{FA}$ across i for month t. In columns (3) and (4), the $r_{i,t}^{FA}$ values included in the calculation of $avg\_r_t^{FA}$ are the same as in columns (1) and (2), respectively. The t-statistics in columns (1)-(2) and (3)-(4) are computed with Driscoll-Kraay and Newey-West standard errors, respectively, following Goyal et al. (2024).

Table 3: Correlation among evaluation metrics for aggregate stock market return predictors

	abs(t)	IS R^2	CT20	CT20 Post-46	OOS20	Alpha
abs(t)	1.00					
IS R^2	0.79	1.00				
CT20	0.40	0.65	1.00			
CT20 Post-46	0.40	0.65	0.99	1.00		
OOS20	0.25	0.27	0.46	0.46	1.00	
Alpha	0.46	0.24	0.25	0.22	0.18	1.00

This table shows the correlations among different statistics across the 46 predictors in Table 1.

Table 4: Simulation results

Percentiles	No predictability 5	50	95	With predictability 5	50	95	Data —
	A: Random betas						
Avg Abs T	0.627	0.805	0.988	1.029	1.344	1.796	1.068
Avg IS R^2	0.071	0.112	0.172	0.174	0.288	0.509	0.324
Avg OOS R^2	-0.568	-0.382	-0.092	-0.677	-0.222	0.971	-0.620
Avg CT OOS	-0.515	-0.333	-0.046	-0.531	-0.095	0.941	-0.027
Avg alpha	-0.073	-0.004	0.056	0.026	0.112	0.207	0.097
	B: Estimated betas & conditional variance						
Avg Abs T	0.629	0.813	1.033	1.252	1.665	2.220	1.068
Avg IS R^2	0.092	0.156	0.265	0.345	0.617	1.105	0.324
Avg OOS R^2	-0.996	-0.586	-0.266	-1.558	-0.242	1.622	-0.620
Avg CT OOS	-0.758	-0.467	-0.157	-1.026	-0.012	1.550	-0.027
Avg alpha	-0.072	-0.006	0.055	0.053	0.139	0.245	0.097

This table shows the 5th, 50th, and 95th percentiles of 5 statistics across 1000 simulations, as well as the corresponding values computed from real data. The 5 statistics are i) the average absolute value of Newey-West t-statistics, ii) the average in-sample R^2, iii) the average OOS R^2, iv) the average OOS R^2 after applying the positive-sign restriction from CT (2008), and v) the average alphas of the FA portfolios. Each simulation consists of 47 monthly time series representing stock market excess returns and 46 predictors with the same start and end months as the real data. In Panel A, the predictors follow AR(1) processes with a mean of zero and AR(1) coefficients estimated from the real predictor data. The innovations of the predictors are i.i.d. normal shocks with a volatility of 30 basis points multiplied by the square root of one minus the square of the AR(1) coefficients. In the first 3 columns with the header "no predictability," the simulated returns equal a constant of 58 basis points plus i.i.d. normal shocks with a mean of zero and a volatility of 4.78%. In the next 3 columns, the simulated returns equal the aforementioned two terms plus 46 additional terms, each being the product of one of the 46 predictors and a beta coefficient randomly drawn from a uniform distribution between -1 and 1. Panel B presents similar two sets of simulations, except that the beta coefficients and the predictors' volatility are estimated from data rather than randomly drawn and manually set. The innovation terms in both the predictors and the returns have time-varying conditional variance estimated with a GARCH(1,1) model on the data.

Table 5: Simulation results with longer samples and stronger predictability

	A: Longer sample					
Sample length	1X	2X	5X	10X	20X	30X
Avg Abs t	1.344	1.670	2.347	3.191	4.405	5.318
Avg IS R^2	0.288	0.213	0.162	0.144	0.135	0.131
Avg OOS R^2	-0.222	-0.090	0.009	0.058	0.085	0.096
Avg CT OOS	-0.095	-0.025	0.034	0.072	0.092	0.101
Avg FA 2F alpha	0.0457	0.0325	0.0234	0.0189	0.0148	0.0125
Avg FD 2F alpha	-0.0194	-0.0080	-0.0016	-0.0001	0.0003	0.0004
	B: Stronger predictability with 10X sample length					
Predictability	1X	2X	3X	4X	5X	6X
Avg Abs t	3.191	5.672	7.466	8.674	9.488	10.035
Avg IS R^2	0.144	0.475	0.874	1.245	1.558	1.806
Avg OOS R^2	0.058	0.399	0.802	1.181	1.494	1.738
Avg CT OOS	0.072	0.406	0.738	1.008	1.203	1.339
Avg FA 2F alpha	0.0189	0.0156	0.0166	0.0230	0.0342	0.0475
Avg FD 2F alpha	-0.0001	0.0087	0.0220	0.0387	0.0568	0.0757

This table shows the median percentiles of 6 statistics across 1000 simulations. The 6 statistics are i) the average absolute value of Newey-West t-statistics, ii) the average in-sample R^2, iii) the average OOS R^2, iv) the average OOS R^2 after applying the positive-sign restriction from CT (2008), v) the average 2-factor alphas of the FA portfolios on top of the market and the FD portfolios, and vi) the average 2-factor FD alphas of the portfolios on top of the market and the FA portfolios. Each simulation consists of 47 monthly time series representing stock market excess returns and 46 predictors with the same start and end months as the real data. The predictors follow AR(1) processes with a mean of zero and AR(1) coefficients estimated from the real predictor data. The innovation of the predictors are iid normal shocks with a volatility of 30 basis points multiplied by the square root of one minus the square of the AR(1) coefficients. The simulated returns equal the aforementioned two terms plus 46 additional terms, each being the product of one of the 46 predictors and a beta coefficient randomly drawn from a uniform distribution between -1 and 1. Panel A reports results simulated with 1-30 times the length of the original data. Panel B reports results simulated with 10 times the data and 1-6 times the dispersion of the predictors.

Table 6: Pre- and Post-sample behaviors of FA portfolios

	(1) all	(2) risk
pre-sample alpha	0.315***	0.421***
	[2.97]	[3.05]
in-sample alpha	0.201***	0.191***
	[4.86]	[4.19]
post-sample alpha	0.007	0.041
	[0.16]	[0.92]
pre-sample × mkt	-0.193*	-0.251*
	[-1.68]	[-1.79]
in-sample × mkt	-0.136***	-0.132***
	[-4.69]	[-4.01]
post-sample × mkt	-0.103***	-0.132***
	[-5.15]	[-6.27]
Sample size	37,325	31,355
p(pre - in)	0.251	0.072
p(post - in)	0.002	0.022
p(pre - post)	0.007	0.009

This table shows results of regressions of all FA portfolio returns on 3 dummy variables indicating whether the observation is before, within, or after the authors' original sample, and their interaction with the contemporaneous market returns. The coefficients on the 3 dummy variables extract the FA alphas pre-, in-, and post-sample. Evaluation begins 20 years after the inception of the predictors, and possibly before 1926. Data are monthly from 1891 to 2024. T-statistics computed with standard errors clustered at the monthly level are reported in square brackets.

Table 7: R^2 in forecasting monthly stock market excess returns–winsorizing extreme positive values

	Constraint			
	None	Negative	Positive	Both
vp	4.72	4.72	1.76	1.76
impvar	0.49	0.49	0.49	0.49
vrp	-19.51	-1.00	-18.73	-0.22
lzrt	-0.11	0.02	-0.03	0.10
ogap	0.72	0.72	0.71	0.71
wtexas	0.15	0.36	0.28	0.49
sntm	-0.41	-0.41	-0.41	-0.41
ndrbl	-0.84	-0.58	-0.59	-0.32
skvw	-0.64	-0.58	-0.54	-0.48
tail	-0.16	-0.16	0.01	0.01
fbm	-0.89	-0.89	-0.58	-0.58
dtoy	-0.19	-0.19	-0.13	-0.13
dtoat	0.15	0.15	0.15	0.15
ygap	-1.35	-0.87	-0.96	-0.48
rdsp	-1.21	-1.08	-1.13	-1.00
rsvix	-0.02	-0.02	-0.02	-0.02
tchi	-0.02	0.14	-0.02	0.14
avgcor	0.37	0.37	0.41	0.41
shtint	-0.07	-0.07	-0.05	-0.05
disag	-0.90	-0.48	-0.90	-0.48
ntis	-0.54	-0.54	0.16	0.16
tby	0.45	0.61	0.53	0.69
dp	-0.59	-0.09	-0.59	-0.08
dy	-0.56	0.02	-0.91	-0.32
ep	0.07	0.10	-0.01	0.03
de	-0.52	0.01	-0.52	0.01
svar	-1.62	-0.06	-1.47	0.09
lty	-0.07	0.51	0.15	0.73
ltr	-0.49	-0.47	-0.13	-0.11
tms	0.24	0.26	0.66	0.68
dfy	-0.39	-0.35	-0.37	-0.33
dfr	-0.32	-0.25	-0.16	-0.09
infl	0.45	0.44	0.46	0.45
bm	-1.15	-0.86	-0.64	-0.35
pce	0.92	1.09	0.75	0.92
govik	-0.71	-0.34	-0.71	-0.34
crdstd	-2.41	-1.04	-2.30	-0.93
ik	0.77	0.88	0.76	0.87
cay	-0.87	-0.63	-0.58	-0.33
skew	-0.43	-0.29	-0.46	-0.33
accrul	-0.81	-0.78	-0.78	-0.75
cfacc	-0.70	-0.70	-0.69	-0.69
gpce	0.16	0.17	0.30	0.31
gip	0.06	0.06	0.08	0.08
house	0.11	0.11	0.11	0.11
eqis	0.15	0.27	0.55	0.67
Mean	-0.62	-0.03	-0.57	0.03
Median	-0.25	-0.06	-0.09	-0.01

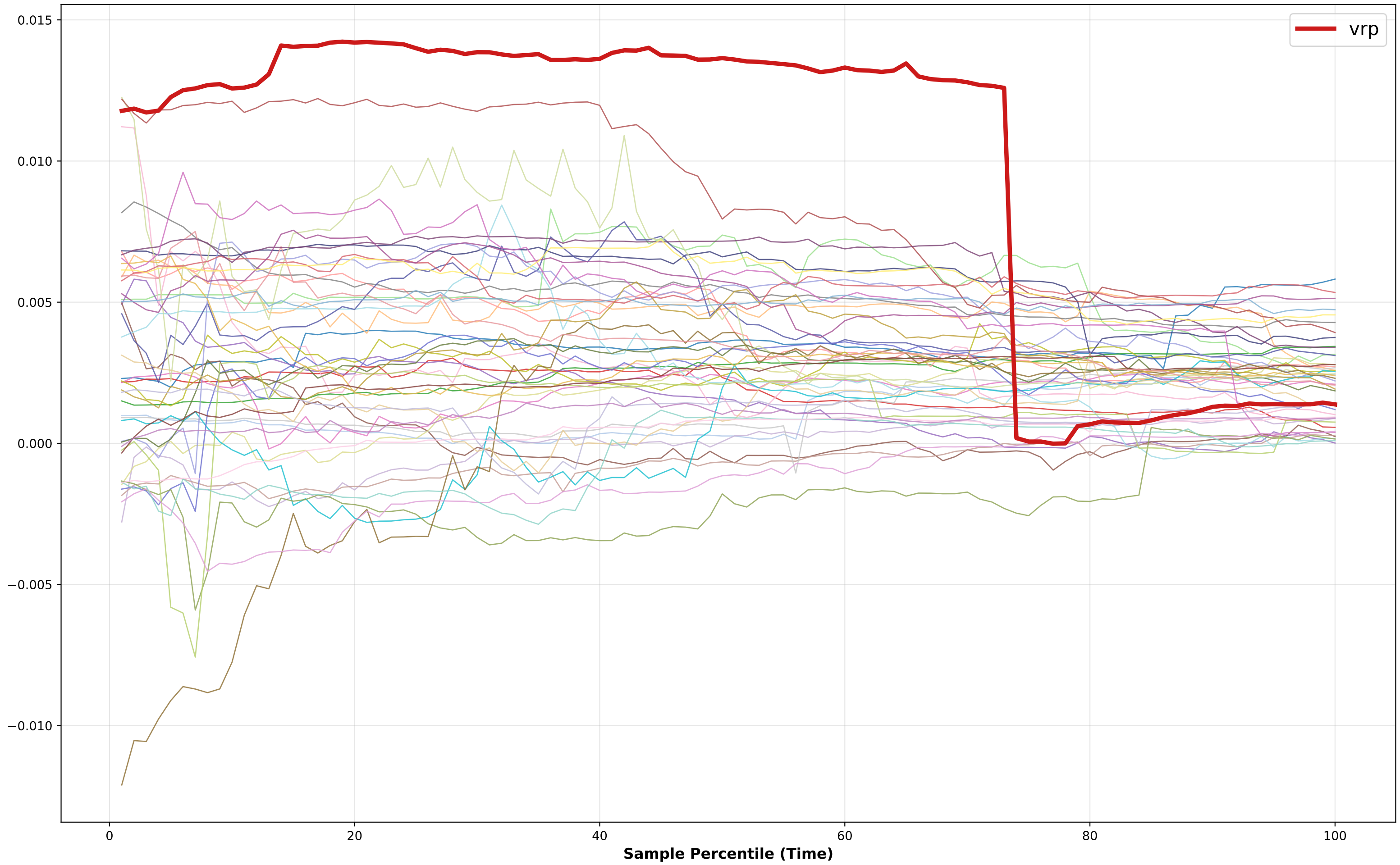
This table shows performance statistics of the 46 predictors in Goyal et al. (2024) forecasting monthly market returns in excess of the risk-free rate. Column “None” reports out-of-sample R^2 on forecasts computed with freely estimated expanding-window coefficients and is equivalent to OOS20 in Table 1. Column “Negative” imposes the Campbell and Thompson (2008) non-negativity constraint and is equivalent to CT20 in Table 1. Column “Positive” winsorizes from above forecasts exceeding 2 times the expanding-window mean of the market excess returns. Column “Both” imposes both constraints from the previous two columns. The R^2s evaluation starts 20 years after the inception of the predictor or Jan 1926, whichever comes later. All R^2s are in percentage units.

Table 8: R^2 in forecasting monthly stock market excess returns–imposing economic constraint

	Constraint				
	None	Negative	Positive	Both	Economic
`vp`	4.72	4.72	1.76	1.76	1.19
`impvar`	0.49	0.49	0.49	0.49	0.94
`vrp`	-19.51	-1.00	-18.73	-0.22	0.47
`rsvix`	-0.02	-0.02	-0.02	-0.02	0.26
`dp`	-0.59	-0.09	-0.59	-0.08	0.44
`dy`	-0.56	0.02	-0.91	-0.32	0.45
`ep`	0.07	0.10	-0.01	0.03	0.32
`bm`	-1.15	-0.86	-0.64	-0.35	0.24
Mean	-2.07	0.42	-2.33	0.16	0.54
Median	-0.29	0.00	-0.30	-0.05	0.45

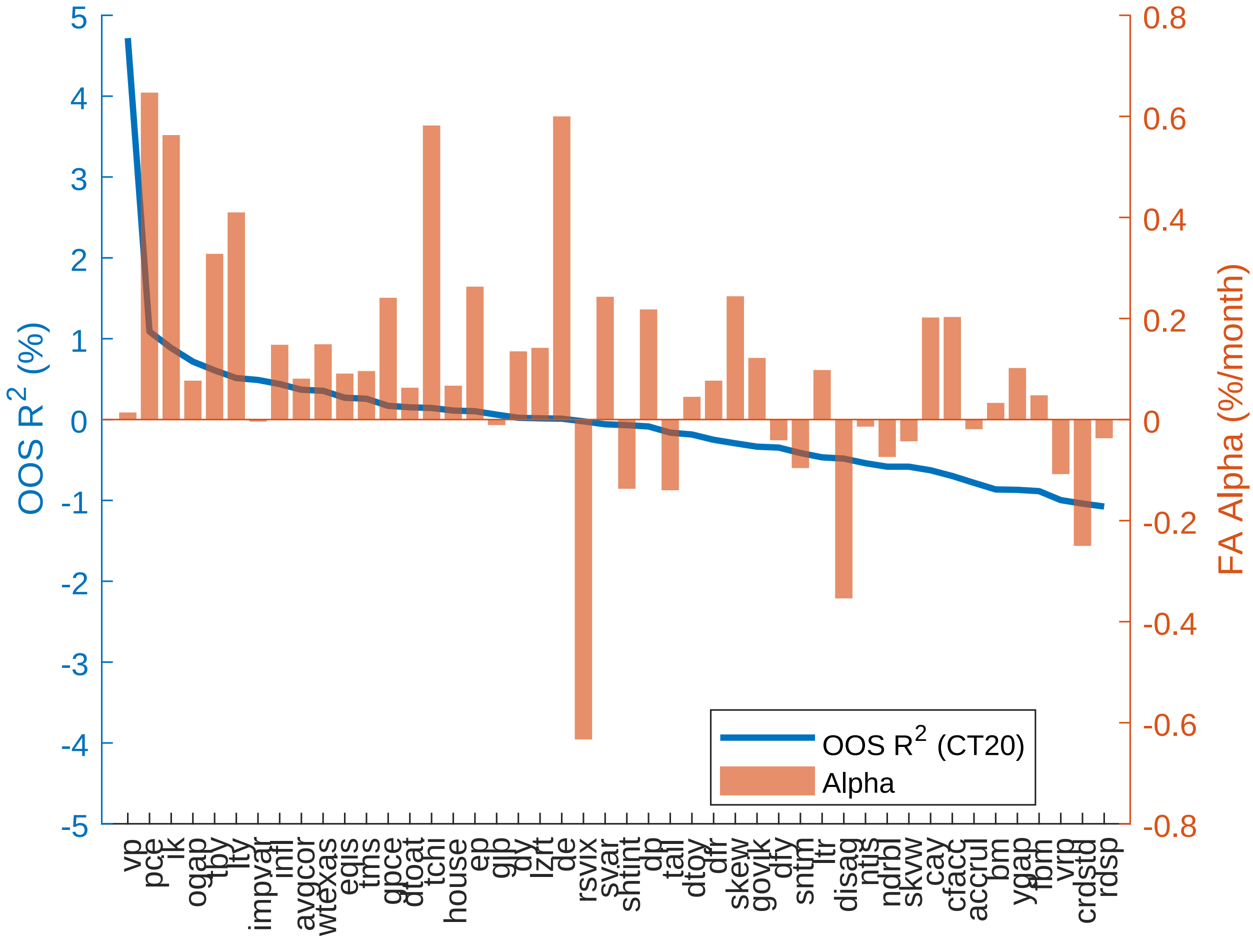
This table shows performance statistics of the 8 predictors in Goyal et al. (2024) forecasting monthly market returns in excess of the risk-free rate. Column "None" reports out-of-sample R^2 on forecasts computed with freely estimated expanding-window coefficients. Column "Negative" imposes the Campbell and Thompson (2008) non-negativity constraint. Column "Positive" winsorizes from above forecasts exceeding 2 times the expanding-window mean of the market excess returns. Column "Both" imposes both constraints from the previous two columns. Column "Economic" imposes economic constraints specifically applicable to the signal. The R^2s evaluation starts 20 years after the inception of the predictor or Jan 1926, whichever comes later. All R^2s are in percentage units.

Figure 1: Expanding window coefficients for the 46 predictors



This figure plots the expanding-window regression coefficients of the aggregate stock market excess returns on the 46 predictors in Table 1. The x-axis is the percentile of the sample length of the predictors, with 1 corresponding to the first percentile of the sample, and 100 corresponding to the end of the sample. The coefficients for `vrp` are highlighted.

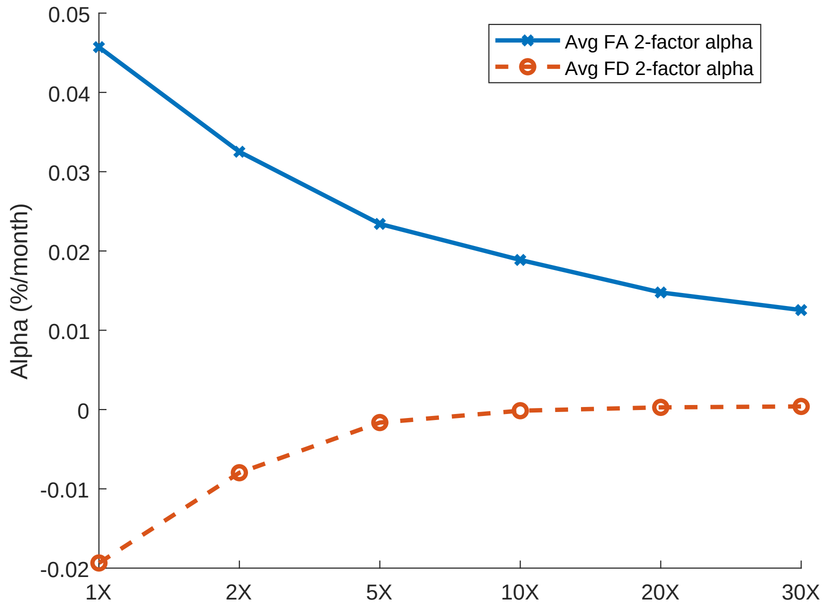
Figure 2: OOS R^2s and FA alphas across predictors



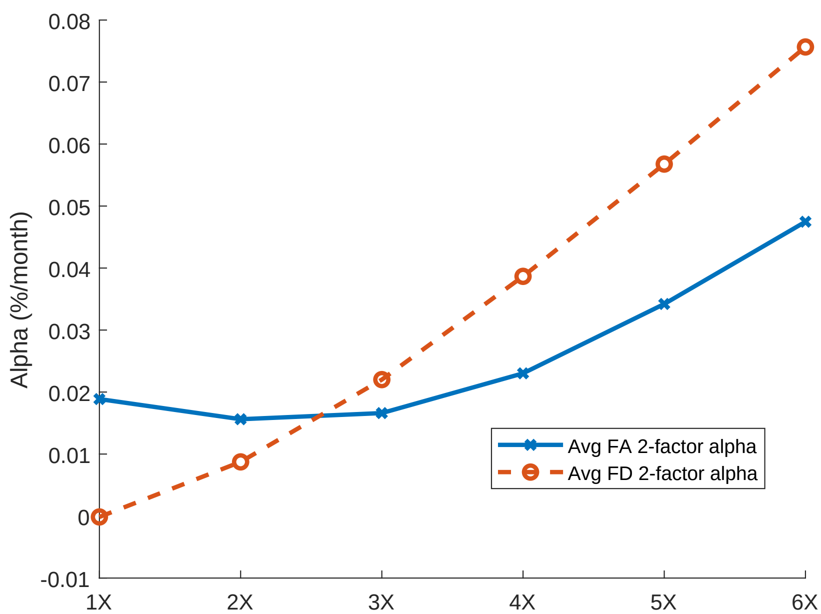
This figure plots OOS R^2s with the Campbell Thompson non-negativity constraint and forecast-agnostic alphas across the 46 predictors in Table 1. The OOS R^2s declines monotonically from left to right.

Figure 3: Average FA and FD two-factor alphas

(a) Longer samples



(b) Stronger predictability with 10X sample length



This figure plots the medians of the average (across predictors) FA and FD alphas controlling for each other and the market excess returns across 1000 simulations. Panel (a) conducts the simulations with increasingly longer samples, whereas Panel (b) conducts simulations with 10 times the sample length and increasingly stronger predictability than observed.

Appendix

A Data

We obtain data on stock returns and on predictor variables from Amit Goyal's website at `https://sites.google.com/view/agoyal145`, and specifically the 2024 vintage. For the purpose of pre-sample testing, we extend the stock market return series back to 1871 using data from Global Financial Data (GFD), which are ultimately based on Standard & Poor's. These return data are available at a monthly frequency. We construct predictor variables using data that would have been available to the investor at that time. This includes `sntm` and `fbm` which embed an estimation in their construction.

B FA alphas with alternative volatility targets

Table B1 calculates panel and time-series alphas of the FA portfolios as in Columns (1) and (3) of Table 2, except they have different volatility targets from 5% per month. The t-statistics remain stable across columns, demonstrating that the FA alphas are highly statistically significant regardless of the specific choice of the volatility target.

Table B1: FA alphas with alternative volatility targets

	2.5%	5%	10%	15%	20%	25%	50%
	A: Panel alphas						
mkt	-0.082***	-0.139***	-0.194***	-0.225***	-0.249***	-0.268***	-0.309***
	[-4.32]	[-5.00]	[-5.44]	[-5.92]	[-6.29]	[-6.45]	[-6.72]
Alpha	0.088***	0.161***	0.232***	0.262***	0.281***	0.295***	0.326***
	[4.51]	[4.57]	[4.79]	[4.78]	[4.73]	[4.63]	[4.32]
Sample size	35,395	35,395	35,395	35,395	35,395	35,395	35,395
	B: Time series alphas						
mkt	-0.181***	-0.285***	-0.360***	-0.374***	-0.369***	-0.371***	-0.400***
	[-2.65]	[-3.00]	[-3.10]	[-3.05]	[-2.88]	[-2.79]	[-2.89]
Alpha	0.203***	0.360***	0.482***	0.527***	0.546***	0.563***	0.622***
	[2.86]	[2.87]	[3.14]	[3.20]	[3.12]	[3.05]	[3.06]
Sample size	1,187	1,187	1,187	1,187	1,187	1,187	1,187

Panel A conducts the following predictor-monthly level panel regression: $r_{i,t}^{FA} = \alpha + \beta mkt_t + \epsilon_{i,t}$. Here, $r_{i,t}^{FA}$ is the FA portfolio return for predictor i in month t, and mkt_t is the stock market return in excess of the risk-free rate in month t. Panel B conduct the monthly time-series regression: $avg\_r_t^{FA} = \alpha + \beta mkt_t + \epsilon_t$, in which $avg\_r_t^{FA}$ is the average of $r_{i,t}^{FA}$ across i for month t. In Panel B, the $r_{i,t}^{FA}$ values included in the calculation of $avg\_r_t^{FA}$ are the same as in Panel A. All regressions start the evaluation 20 years after the inception of the predictor or Jan 1926, whichever comes later. The FA alphas are constructed with difference volatility targets, indicated in the column headers. The t-statistics in Panels A and B are computed with Driscoll-Kraay and Newey-West standard errors, respectively, following Goyal et al. (2024).