Journal of Banking and Finance 84 (2017) 188–201

Contents lists available at ScienceDirect

Journal of Banking and Finance

journal homepage: www.elsevier.com/locate/jbf

Out-of-sample equity premium predictability and sample split–invariant inference

Gueorgui I. Kolev[a,*], Rasa Karapandza[b]

[a] Department of Economics, Middlesex University Business School, The Burroughs, London NW4 4BT, United Kingdom
[b] Department of Finance, Accounting & Real Estate, EBS Business School, Gustav-Stresemann-Ring 3, 65189 Wiesbaden, Germany

ARTICLE INFO

Article history:
Received 11 May 2013
Accepted 11 July 2016
Available online 21 October 2016

JEL Classification:
G12
G14
G17
C22
C53

Keywords:
Equity premium predictability
Out-of-sample inference
Sample split choice
Bootstrap

ABSTRACT

For a comprehensive set of 21 equity premium predictors we find extreme variation in out-of-sample predictability results depending on the choice of the sample split date. To resolve this issue we propose reporting in graphical form the out-of-sample predictability criteria for *every* possible sample split, and two out-of-sample tests that are invariant to the sample split choice. We provide Monte Carlo evidence that our bootstrap-based inference is valid. The in-sample, and the sample split invariant out-of-sample mean and maximum tests that we propose, are in broad agreement. Finally we demonstrate how one can construct sample split invariant out-of-sample predictability tests that simultaneously control for data mining across many variables.

1. Introduction

The question of whether asset returns are predictable is important not only from the theoretical (asset-pricing) perspective but also from the practical (market-timing) perspective. An important concern is whether in-sample or instead out-of-sample econometric methods should be used to assess the predictability of returns. According to Ashley et al. (1980, p. 1149), "the out-of-sample forecasting performance" provides "the best information" and should therefore be preferred. More recently, Inoue and Kilian (2004) argue that out-of-sample tests are less able to reject a false null hypothesis; that loss of power is due to splitting the finite sample into an in-sample estimation period and an out-of-sample evaluation period, although the authors acknowledge there is a role for out-of-sample methods in choosing the best (though possibly misspecified) forecasting model among the few competitors. Hence most attention in the recent literature on returns predictability has focused on out-of-sample forecasting methods and inference (see, among others, Goyal and Welch, 2003; Rapach and Wohar, 2006; Campbell and Thompson, 2008; Kolev, 2008; Welch and Goyal, 2008; Rapach et al., 2010).[1]

Out-of-sample methods involve splitting the available data into an in-sample estimation period, which is used to produce an initial set of regression estimates, and an out-of-sample forecast evaluation period over which forecasts are generated and then both evaluated (in terms of some specified criterion of goodness) and compared with results from competing models West (2006, p. 106). The natural question that arises in this context is just *how* the sample should be split into these two periods. This paper considers a set of 21 predictors (including those used in the influential paper of Welch and Goyal, 2008). We demonstrate that some

[1] Apart from equity premium predictability, another interesting and active field of finance addresses whether mean-variance optimization improves on the naive $1/N$ portfolio allocation rule on an *out*-of-sample basis (DeMiguel et al., 2009; Kirby and Ostdiek, 2012). The issues we raise as well as our proposed inferential methods are applicable also to this problem—provided the expected portfolio return under the null (e.g., with naive $1/N$ portfolio weights) and the expected portfolio return under the alternative (e.g., with mean-variance–optimal weights) can be written as a regression function of the returns of the underlying assets constituting the portfolio. Britten-Jones (1999) shows how mean-variance optimization can be recast as a regression problem.

* Corresponding author.
E-mail addresses: joro.kolev@gmail.com (G.I. Kolev), rasa.karapandza@ebs.edu (R. Karapandza).

http://dx.doi.org/10.1016/j.jbankfin.2016.07.017

such splits yield results indicating that returns are *not* predictable whereas other splits lead to the opposite conclusion. That is: for any given predictor and any given data set, the derived predictability of returns is sensitive to where the sample is split between the estimation and forecast evaluation subsamples.

To address this problem and resolve contradictory findings, we propose two simple (but computationally intensive) methods that do not suffer from this dependence on the choice of split date. The first approach is to report in graphical form the out-of-sample predictability results for *every* possible sample split. Thus we report the *p*-values for the Clark and West (2007) mean squared prediction error–adjusted (MSPE-adj) statistic for every possible sample split, where the sample split date τ falls within the interval $[\text{int}(.05T),\ T - \text{int}(.05T)]$; here int(·) denotes the argument's integer part. It follows that neither the in-sample estimation period nor the out-of-sample evaluation period ever contains less than 5% of the total number T of observations. Computationally, we determine the *p*-value by generating 9999 bootstrap samples under the null of no predictability and then calculating the *t*-statistic associated with the MSPE-adj for each τ; let us call this with $\mathbb{t}(\text{MSPE} - \text{adj})_\tau$ to emphasize that one such statistic goes with each individual sample split indexed by τ. Then the *p*-value for each sample split τ is the fraction of bootstrap samples for which the bootstrap *t*-statistic is larger than the *t*-statistic calculated from the original data for the corresponding τ.

The second approach is to calculate, again across all possible sample splits, some summary statistic of an out-of-sample predictability criterion and then, via a bootstrap procedure, to determine (for inferential purposes) the *distribution* of that statistic under the null hypothesis of no return predictability. Thus we calculate $\text{mean}_\tau\{\mathbb{t}(\text{MSPE} - \text{adj})_\tau\}$ and $\max_\tau\{\mathbb{t}(\text{MSPE} - \text{adj})_\tau\}$, the mean and the maximum taken over τ. In this way we capture the entire sequence of statistics $\mathbb{t}(\text{MSPE} - \text{adj})_\tau$ for each τ. We evaluate the 15 predictors considered in Welch and Goyal (2008) and six additional, behavioral predictors (that end up failing to predict the equity premium). We show that *some* of the traditional forecasting variables work very well for *some* sample splits. The satisfactory (albeit episodic) performance of traditional predictors is in contrast to the widespread failure of behavioral predictors. The hypothesis that investor sentiment predicts future stock returns is a plausible one, and it is not unreasonable to suppose that such sentiment can be measured (howsoever imperfectly) by behavioral factors. Yet of all the behavioral sentiment variables we examine, only one is predictive of the equity premium: *Equity Share in New Issues*.

We can summarize as follows the issues at hand and our paper's contributions.

1. Scholars who are interested in such general questions as "Are stock returns predictable?" and who want to use out-of-sample methods should not take the sample split date as given. We propose two methods of dealing with this sample split problem: (i) report the results for each possible sample split; or (ii) calculate a statistic that is invariant to the chosen sample split within any given set of possible sample splits.
2. We document that conclusions about stock market predictability when using out-of-sample methods are strongly dependent on the choice of a sample split date. In fact, a researcher may derive evidence supporting (or refuting) predictability simply by adjusting the sample split date. We also show that traditional predictors of stock returns exhibit satisfactory performance often but not for all sample splits; in contrast, behavioral predictors hardly ever exhibit satisfactory performance.
3. As soon as one allows any possible sample split to be chosen (e.g., via the mean or maximum of results based on various forecasting criteria taken across all possible sample splits), most debates over competing in- or out-of sample methods and splits become moot. Results from using in-sample methods are in broad agreement with those from using sample split–invariant out-of-sample methods.
4. We show how to construct out-of-sample predictability tests that (i) are sample split invariant and (ii) control for data mining. Using this out-of-sample, sample split independent joint test of predictive power of 21 predictors we reject the null hypothesis of no predictability - contrary to results of Welch and Goyal (2008).
5. We provide Monte Carlo evidence to support the validity of our bootstrap-based inference.

Four other works are closely related to this paper. Hubrich and West (2010) and Clark and McCracken (2012) propose taking the maxima of various statistics for simultaneously judging whether a small set of alternative models that nest a benchmark model improve upon the benchmark model's MSPE. After completing this paper, we became aware of independent and contemporaneous work by Rossi and Inoue (2012) and Hansen and Timmermann (2012). Both of these papers examine in great detail the theoretical econometric properties of the sample split problem. Rossi and Inoue derive the theoretical distribution of (general) sample split–invariant mean and maximum tests; Hansen and Timmermann propose a "minimum *p*-value" approach to sample split–invariant inference.

For nested model comparisons, such as our paper's asset pricing application, Rossi and Inoue (2012) propose taking either the mean or the maximum over all possible sample splits of the Clark and McCracken (2001) ENC-NEW test statistic. Rossi and Inoue characterize and tabulate the distributions of their mean and maximum tests. We show via Monte Carlo experiments that their tabulated null distributions poorly approximate the true null distributions for asset pricing applications such as ours. Take, for example, our Monte Carlo simulation where calibration is based on time-series properties of the dividend-to-price ratio (with 1200 time-series observations). At the 5% nominal significance level, the Rossi and Inoue (2012) *mean* test statistic has empirical size of 9.81%; similarly, at the 5% nominal significance level their *maximum* test statistic has empirical size of 13.25%. So for a predictor with the time series properties of the dividend-to-price ratio their tests over-reject the true null even for samples as large as 1200 observations.

Hence our paper differs from both Rossi and Inoue (2012) and Hansen and Timmermann (2012) along several dimensions. First, we employ bootstrap techniques so that we can evaluate empirically the null distributions of the mean and maximum tests. Doing so renders our testing procedure robust not only to the nearly nonstationary behavior of the predictors (since the autoregressive parameters of the predictive variables are approximately 1) but also to the high correlation between innovations of the predictive variables and innovations of the predicted term (e.g., returns). Note that those high correlations and also predictor nonstationarity are characteristic of real-world financial data. Second, we evaluate comprehensively the out-of-sample forecasting performance of 21 equity premium predictors; we find that it is possible to predict the equity premium and also show that the distributions of test statistics are not pivotal. Hence, we argue that any sample split independent test based on a theoretical distribution that is *not* a function of the predictor's autoregressive parameter and of the correlation between innovations of predictor and predictand (like those derived by Rossi and Inoue and by Hansen and Timmermann) will not work well in practice. Third, we show that our bootstrap procedure allows one to control for data-mining issues by evaluating the *joint* forecasting ability of a set of predictive vari-

Table 1
Summary statistics.

Variable	Mean	Standard deviation	Min.	Max.	T
Equity Premuim, (Rm − Rf)	0.0062	0.0559	−0.2877	0.4162	1020
Variables downloaded from Amit Goyal's web site					
Dividend to Price Ratio	−3.3239	0.4511	−4.524	−1.8732	1021
Dividend Yield	−3.3197	0.4494	−4.5313	−1.9129	1020
Book to Market Ratio	0.5868	0.2654	0.1205	2.0285	1021
Earnings to Price Ratio	−2.7141	0.4255	−4.8365	−1.775	1021
Dividend Payout Ratio	−0.6097	0.3229	−1.2247	1.3795	1021
Treasury Bill Rate	0.0366	0.0306	0.0001	0.163	1021
Long Term Yield	0.053	0.028	0.0182	0.1482	1021
Long Term Return	0.0047	0.0239	−0.1124	0.1523	1020
Term Spread	0.0163	0.0131	−0.0365	0.0455	1021
Default Yield Spread	0.0114	0.0071	0.0032	0.0564	1021
Default Return Spread	0.0003	0.0132	−0.0975	0.0737	1020
Inflation Lagged 2 Months	0.0024	0.0053	−0.0208	0.0574	1021
Net Equity Expansion	0.0191	0.0246	−0.0575	0.1732	1009
Stock Variance	0.0025	0.005	0.0001	0.0558	1021
Cross Sectional Premium	0.0004	0.0024	−0.0042	0.0077	788
Variables downloaded from Jeffrey Wurgler's web site					
Dividend Premium	−2.2399	16.2884	−50.23	32.9	600
Number of IPOs	26.2516	23.6031	0	122	612
Average First-day IPO Returns	16.3867	20.0237	−28.8	119.1	612
NYSE Share Turnover	0.5084	0.3665	0.105	1.738	636
Closed-end Fund Discount	8.9577	7.4353	−10.91	25.28	548
Equity Share in New Issues	0.1827	0.1092	0.0167	0.6349	636

ables. Such control is not feasible under the approaches adopted by Rossi and Inoue (2012) and Hansen and Timmermann (2012).

Several limitations of this study are worth mentioning. First, we study only single predictors; that is, we do not study combinations of them (as in Rapach et al., 2010; Welch and Goyal, 2008), including combinations based on principal components (e.g., Neely et al., 2014). Neither do we follow Welch and Goyal by looking at rolling regressions. Also, recent research indicates that the accuracy of traditional predictors (e.g., ratio of dividend to price) can be improved by enlarging the predictive information set with the information implied by derivative markets (Binsbergen et al., 2011; Golez, 2014; Kostakis et al., 2011). To keep the number of variables that we study tractable we do not include these refinements in our study. Finally, we limit ourselves to standard ordinary least-squares (OLS) regressions that incorporate neither the economic gains nor the utility gains of potential investors. All these extensions should be tractable when pursued within the framework described here, and they constitute a fruitful research agenda.

2. Methodology and data

Following much of the extant literature, we estimate OLS bivariate predictive regressions. In particular, we regress the equity premium—constructed as the return on the S&P 500 index (including dividends distributions) minus the risk-free rate, $R_t \equiv (R_m - R_f)_t$—on a constant and on a lagged value of a predictor

$$R_t = \beta_0 + \beta_1 X_{t-1} + u_t. \tag{1}$$

The predictor X is one of the variables listed in Table 1 (in Section 2.5), depending on the specification. The βs are population parameters (to be estimated), and u is a disturbance term.

2.1. In-sample predictability

The in-sample predictive ability of X is assessed via the t-statistic corresponding to b_1 , the OLS estimate of β_1 in Eq. (1). Under the null hypothesis that X_{t-1} is uncorrelated with R_t, the expected returns are constant and $\beta_1 = 0$ (the sign of β_1 is typically suggested by theory). The tables presented in this paper are agnostic concerning whether the alternative hypothesis is one- or two-sided and simply report t-statistics associated with the estimates. The estimated slopes for all variables are in the direction predicted by theory and so, roughly speaking, one can consider a t-statistic whose absolute value exceeds 1.65 to be significant either at the 10% significance level for the two-sided alternative or at the 5% significance level for the one-sided alternative.

When predictive regressions employ a highly persistent predictor whose innovations are correlated with those in the predictand, severe small-sample biases may occur (Mankiw and Shapiro, 1986; Nelson and Kim, 1993; Stambaugh, 1986; 1999). To test whether the in-sample results could be an artifact of this small-sample bias, we follow the bias correction methodology of Amihud and Hurvich (2004). The model is defined over the whole sample $t = 1, 2, \ldots, T$,

$$R_t = \beta_0 + \beta_1 X_{t-1} + u_t, \tag{2}$$

$$X_t = \mu + \rho X_{t-1} + w_t; \tag{3}$$

where the disturbances (u_t, w_t) are serially independently and identically distributed as bivariate normal, and the autoregressive coefficient in Eq. (3) is less than 1.

We shall use superscript c to denote a bias-corrected estimator. First, we estimate Eq. (3) to obtain the OLS estimator r of ρ. We can then use r to compute the bias-corrected estimator of ρ as follows:

$$r^c = r + \frac{1+3r}{T} + \frac{3(1+3r)}{T^2}. \tag{4}$$

This bias-corrected estimator r^c is, in turn, used to compute the corrected residuals $\hat{w}_t^c$ of Eq. (3):

$$\hat{w}_t^c = X_t - (m + r^c X_{t-1}),$$

where m is the OLS estimator of μ.[2]

Next, we run an auxiliary regression of R_t on an intercept, X_{t-1} and $\hat{w}_t^c$. In this auxiliary regression, let b_1^c (resp., f^c) be the OLS estimator of the slope parameter on X_{t-1} (resp., on $\hat{w}_t^c$). Here b_1^c is the bias-corrected estimator of β_1, our variable of interest.

[2] The choice of estimator m does not affect the predictive regression slope's bias.

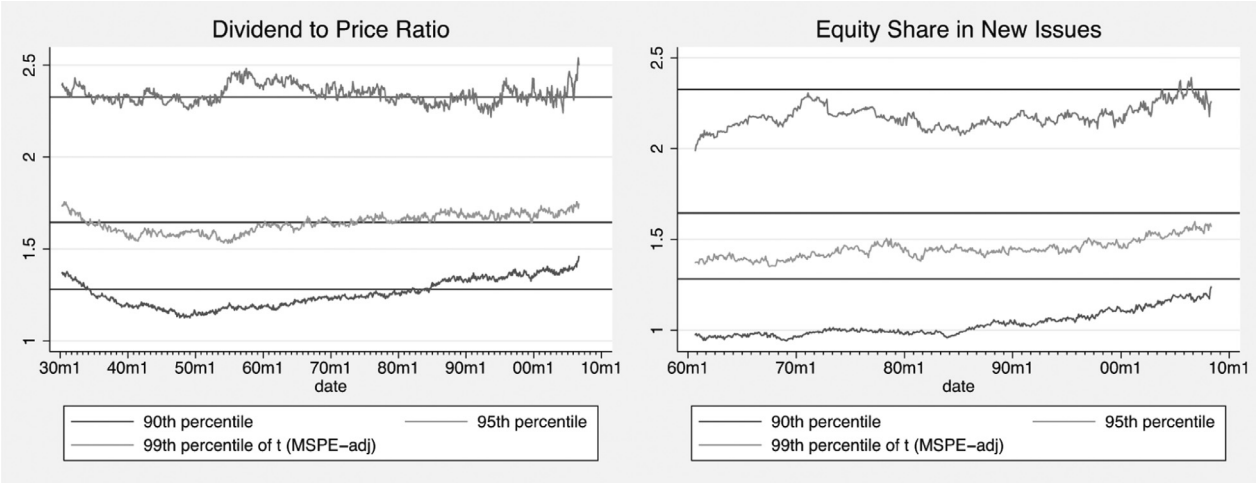

Fig. 1. Plots of the 90th, 95th, and 99th percentiles of the null distribution of $\mathbb{t}(\text{MSPE-adj})_\tau$ statistic versus τ (the start of the out-of-sample forecasting exercise), together with straight horizontal lines at 1.2815, 1.6448 and 2.3263 that mark the respective percentiles in the standard normal distribution, when the Dividend to Price Ratio (left) and the Equity Share in New Issues (right) are used as a predictor of the Equity Premium.

Finally, to conduct inferences based on β_1 we need the bias-corrected *standard error* (SE) of b_1^c, which is given by

$$[\text{SE}^c(b_1^c)]^2 = [f^c]^2 * [1 + 3/T + 9/T^2]^2 * [\text{SE}(r)]^2 + [\text{SE}(b_1^c)]^2, \quad (5)$$

where $\text{SE}(r)$ denotes the usual OLS standard error of r produced by any regression package and $\text{SE}(b_1^c)$ denotes the usual OLS standard error of b_1^c, which comes as a direct output from the auxiliary regression of R_t on an intercept, X_{t-1} and $\hat{w}_t^c$.

2.2. Out-of-sample predictability: fixed sample split date

We generate out-of-sample predictions by the following recursive scheme. First we split the available sample into an in-sample estimation period and an out-of-sample evaluation period. Let T denote the total number of observations and let dates run from 1 to T inclusive. As described in the Introduction, we fix a date $\tau \in [\text{int}(.05T),\ T - \text{int}(.05T)]$. We use the first period, $[1, \tau - 1]$, for the in-sample estimation and use the second period, $[\tau, T]$, for making out-of-sample predictions. For time t, we use $R_{\text{pn},t} = b_{0,t-1}$ to denote the null model prediction and $R_{\text{pa},t} = b_{0,t-1} + b_{1,t-1}X_{t-1}$ to denote the alternative model prediction. Here "pn" stands for "prediction with the null imposed" (i.e., when b_1 is constrained to be 0), and "pa" stands for "prediction under the alternative" (i.e., using Eq. (1)).

The b-values are estimated by OLS using data no more recent than one period *before* the forecast is made. Thus the first prediction under Eq. (1) is $R_{\text{pa},\tau} = b_{0,\tau-1} + b_{1,\tau-1}X_{\tau-1}$, where each b is estimated using only data points from $t = 1$ through $t = \tau - 1$. The second prediction is $R_{\text{pa},\tau+1} = b_{0,\tau} + b_{1,\tau}X_\tau$, where the b-values are estimated using only data points from the 1st period though the τth period. The last prediction is $R_{\text{pa},T} = b_{0,T-1} + b_{1,T-1}X_{T-1}$; here each b is estimated using only data points from period 1 though period $T - 1$.

In this way we obtain, for each fixed τ, a sequence of predictions under the null model and also a sequence of predictions under the alternative model. As an informal measure of the predictive regression's out-of-sample performance, we calculate the out-of-sample R-squared of Campbell and Thompson (2008):

$$R^2_{\text{os}}(\tau) = 1 - \frac{\sum_{t=\tau,\ldots,T}(R_t - R_{\text{pa},t})^2}{\sum_{t=\tau,\ldots,T}(R_t - R_{\text{pn},t})^2} \quad (6)$$

(here "os" stands for "out-of-sample").

We formally test the null hypothesis—that Eq. (1) does *not* improve on the historical average return—by employing the Clark and West (2007) mean squared prediction error–adjusted statistic:

$$\text{MSPE} - \text{adj}(\tau) = \frac{\sum_{t=\tau,\ldots,T}\{(R_t - R_{\text{pn},t})^2 - [(R_t - R_{\text{pa},t})^2 - (R_{\text{pn},t} - R_{\text{pa},t})^2]\}}{\sum_{t=\tau,\ldots,T} 1}. \quad (7)$$

Clark and West (2007) observe that under the null that $\beta_1 = 0$, the alternative model in Eq. (1) estimates additional parameters whose population values are 0, and that the estimation induces additional noise. Hence under the null hypothesis the MSPE of the alternative model is expected to be larger than the null model's. These authors propose an adjustment to the alternative model's MSPE. In Eq. (7), the term in brackets is the *adjusted* MSPE of the alternative model. We calculate the t-statistic associated with Eq. (7); following Clark and West, we denote it $\mathbb{t}(\text{MSPE} - \text{adj})_\tau$. We regress the quantity in braces in Eq. (7) on a constant; the t-statistic from this regression, $\mathbb{t}(\text{MSPE} - \text{adj})_\tau$, is calculated for each sample split τ.

Clark and West (2007, p. 298) justify the approximate normality of their $\mathbb{t}(\text{MSPE} - \text{adj})_\tau$ statistic by observing that its null distribution obeys the following inequalities across a large set of simulations: (90thpercentile) $\leq$ 1.282 $\leq$ (95thpercentile); and (95th-percentile) $\leq$ 1.645 $\leq$ (99thpercentile). In these expressions, the percentiles refer to the distribution of the $\mathbb{t}(\text{MSPE} - \text{adj})_\tau$ statistic under the null of no predictability (i.e., the t-ratio associated with MSPE-adj). Indeed, these inequalities are usually satisfied in our bootstrap simulations. Fig. 1 (left) shows the predictor for which the previous inequalities are *most* often violated, which is the dividend-to-price ratio. This figure plots the 90th, 95th and 99th percentiles of the bootstrap null distribution of $\mathbb{t}(\text{MSPE} - \text{adj})_\tau$ for each sample split date τ; the horizontal lines at 1.2815, 1.6448, and 2.3263 mark the respective percentiles in the standard normal distribution. All violations of these inequalities are slight, so is reasonable to infer that predictability is approximately normal.

More troublesome is that, for the other predictive variables, the corresponding plots look like Fig. 1 (right, for the Equity Share in New Issues' predictor).[3] For these 20 predictors, the inequalities in Clark and West (2007, p. 298) are satisfied yet the approximately normal inference is usually too conservative. In particular: for most of the sample splits τ and most of the predictive variables, the 90th percentile of the bootstrap null distribution of $\mathbb{t}(\text{MSPE} - \text{adj})_\tau$

[3] We present only some of the plots in order to conserve space; all plots are available from the authors upon request.

is slightly less than 1—instead of about 1.282 (the 90th percentile in the standard normal distribution).

We therefore eschew the normal approximation of Clark and West (2007) and instead use bootstrapping to calculate the $t(\text{MSPE}-\text{adj})_\tau$ statistic's p-value at each sample split. The p-value for $t(\text{MSPE}-\text{adj})_\tau$, which we shall denote $\text{pv}(\tau)$, is calculated by drawing 9999 bootstrap samples under the null hypothesis of no predictability and, for each draw, calculating $t(\text{MSPE}-\text{adj})_\tau$. Then $\text{pv}(\tau)$ is the fraction of times that this statistic calculated from the null distribution bootstrap sample is greater than the same statistic calculated from the original data.

After calculating $R^2_{\text{os}}(\tau)$ and $\text{pv}(\tau)$ for each possible sample split date τ, we can construct a graph that plots those two calculated terms against τ. This graph contains all the information needed to assess how well an investor making decisions in real time would have done, on an out-of-sample basis, by using the predictor in question starting at sample split date τ.

2.3. Out-of-sample predictability: invariant to sample split date

If the aim is to distill into a single indicator whether the equity premium is predictable on an out-of-sample basis, then one can construct a summary statistic of the distribution of $t(\text{MSPE}-\text{adj})_\tau$ across the sample split dates τ. For instance, Roy's (1953) union-intersection principle dictates that we do *not* reject the null of no predictability if none of the $t(\text{MSPE}-\text{adj})_\tau$ statistics reject it. We shall therefore reject that null only if at least one of the $t(\text{MSPE}-\text{adj})_\tau$ rejects the null for at least one τ. Formally, that approach is equivalent to using

$$\max_\tau\{t(\text{MSPE}-\text{adj})_\tau\} \tag{8}$$

as a test statistic; alternatively, we could use

$$\text{mean}_\tau\{t(\text{MSPE}-\text{adj})_\tau\} \tag{9}$$

as a test statistic.

There is no clear basis ex ante for preferring one of Eqs. (8) and (9) over the other, since the null distribution is not known for either test statistic. Hence we determine the null distribution of both the maximum and the mean statistics just displayed by generating 9999 bootstrap samples under the null of no predictability. After computing the two statistics for each of these bootstrap samples, we calculate the p-value of the maximum (resp. mean) test statistic by counting how often the bootstrap maximum (resp. mean) statistic is larger than the maximum (resp. mean) statistic calculated from the original data. As before, the bootstrap samples are generated while imposing the null hypothesis of no predictability.

2.4. Bootstrap procedure: description

To generate our 9999 bootstrap samples under the null hypothesis of no predictability, we follow Kilian (1999), Mark (1995), Rapach and Wohar (2006), and Welch and Goyal (2008). (For a general treatment of bootstrap hypothesis testing, see MacKinnon, 2009—especially its section on the residual bootstrap.) We use the original data to estimate Eqs. (2) and (3) by ordinary least squares and then store the residuals ($\hat{u}_t$ and $\hat{w}_t$) for resampling.[4] Next we use the original data to estimate Eqs. (2) and (3) via OLS while imposing the null hypothesis of no predictability ($\beta_1 = 0$); the resulting restricted estimates are denoted $\tilde{\beta}_0$, $\hat{\mu}$, and $\hat{\rho}$ and are stored for later use to generate bootstrap data under the null.

The sample is restricted in that X_t is available only for times $t = 0, \ldots, T$ and R_t is available only for times $t = 1, \ldots, T$. To initiate the recursion in Eq. (3), we randomize (with equal probability) over dates $t = 0, \ldots, T$, denote the draw t^0, and set $X^b_0 = X_{t^0}$. Then we randomize again with equal probability but now with replacement over dates $t = 1, \ldots, T$; we use t^* to signify a single draw from this randomization. For one bootstrap round we generate T such draws. Then we set $u^b_t = \hat{u}_{t^*}$ and $w^b_t = \hat{w}_{t^*}$ for $t = 1, \ldots, T$, thereby drawing (with replacement) the residuals $\hat{u}_t$ and $\hat{w}_t$ as a pair that are matched by t to preserve their cross correlation. For each bootstrap round, we generate $R^b_t = \tilde{\beta}_0 + u^b_t$ and $X^b_t = \hat{\mu} + \hat{\rho}X^b_{t-1} + w^b_t$ for $t = 1, \ldots, T$. Finally, the 9999 bootstrap samples generated under the null of no predictability are obtained by following the same procedure another 9999 times. We estimate the unrestricted model in Eqs. (2) and (3) for each of the 9999 bootstrap-generated data sets on R^b_t and on X^b_t and then calculate, for each set, the statistics described previously: $R^2_{\text{os}}(\tau)$ and $t(\text{MSPE}-\text{adj})_\tau$ for each τ as well as $\max_\tau\{t(\text{MSPE}-\text{adj})_\tau\}$ and $\text{mean}_\tau\{t(\text{MSPE}-\text{adj})_\tau\}$. These 9999 replicates of each are used to estimate each statistic's distribution under the null hypothesis of no predictability. For example, we evaluate the p-value of the maximum statistic by checking for how many of the 9999 bootstrap samples $\max_\tau\{t(\text{MSPE}-\text{adj})_\tau\}$ is larger than $\max_\tau\{t(\text{MSPE}-\text{adj})_\tau\}$ calculated for the original data.

2.5. Data

The equity premium measure that we use, $R_t \equiv (R_m - R_f)_t$, is based on monthly returns on the S&P 500 index, including dividends. The end-of-month values are a series provided by the Center for Research in Security Prices for the period January 1926 to December 2010; we subtract the risk-free rate, defined as the contemporaneous 1-month US Treasury bill (T-bill) rate.

We study the predictive performance of 21 variables. Of these, the first 15 are from Welch and Goyal (2008) and are downloaded from Amit Goyal's website. The remaining six are *behavioral* predictors; they are downloaded from Jeffrey Wurgler's Web page.

Summary statistics for the equity premium and for all of the predictors are given in Table 1.

3. Results

3.1. In-sample predictability

Table 2 reports the in-sample regression results. The predictand is always the equity premium, and the predictor variable is named at the start of each row. We estimate Eq. (1) by OLS; the b_1 value, the t-statistic for that value, and the R-squared from estimating Eq. (1) are given in (respectively) the first, second, and third data columns of the table.

The fourth column reports the bias-corrected estimator b^c_1 of β_1 from the predictive regression, calculated as explained in Section 2 (cf. Amihud and Hurvich, 2004), and the fifth column reports the t^c-statistic ($= b^c_1/[\text{SE}^c(b^c_1)]$). The values in the sixth column are for r, the OLS estimate of the autoregressive parameter ρ in Eq. (3); the seventh column gives r^c, the bias-corrected estimator of ρ. The table's last column reports f, an unbiased estimator of $[\text{Cov}(u_t, w_t)]/[\text{Var}(w_t)]$ (Amihud and Hurvich, 2004, Lemma 1).

An examination of Table 2 reveals that, even after bias correction, many variables remain significant in-sample predictors of the equity premium. At the same time, not even a liberal cut-off point for "significant" (e.g., 1.28) and carrying out a one-sided test at the 10% level are enough to make the following variables significantly predictive: Dividend Payout Ratio, Default Return Spread, Inflation Lagged 2 Months, Stock Variance, Dividend Premium,

[4] Rapach and Wohar (2006, p. 237) resample the restricted model residuals (i.e., the residuals when $\beta_1 = 0$); Welch and Goyal (2008, p. 1462) resample the unrestricted residuals. According to MacKinnon (2009, p. 195), "there might be a slight advantage in terms of power if we were to use unrestricted rather than restricted residuals." We therefore opt to resample the unrestricted residuals.

Table 2
In-sample predictive regressions estimates, and bias corrected in-sample estimates.

	b_1	t−stat	R−sq	b_1^c	t^c−stat	r	r^c	Cov(u_t, w_t)/Var w_t
Dividend to Price Ratio	0.0087	2.26	0.0050	0.0050	1.28	0.99	1.00	−0.96
Dividend Yield	0.0101	2.59	0.0066	0.0098	2.51	0.99	1.00	−0.08
Book to Market Ratio	0.0196	2.98	0.0087	0.0157	2.38	0.98	0.99	−1.01
Earnings to Price Ratio	0.0088	2.13	0.0044	0.0063	1.54	0.99	0.99	−0.62
Dividend Payout Ratio	0.0018	0.34	0.0001	0.0015	0.28	0.99	1.00	−0.08
Treasury Bill Rate	−0.0966	−1.69	0.0028	−0.1011	−1.77	0.99	1.00	−1.14
Long Term Yield	−0.0750	−1.20	0.0014	−0.0862	−1.38	1.00	1.00	−2.86
Long Term Return	0.1153	1.57	0.0024	0.1155	1.57	0.04	0.04	0.25
Term Spread	0.1823	1.37	0.0018	0.1823	1.37	0.96	0.96	0.01
Default Yield Spread	0.4105	1.67	0.0027	0.3733	1.52	0.98	0.98	−9.67
Default Return Spread	0.1378	1.03	0.0011	0.1382	1.04	−0.12	−0.12	0.57
Inflation Lagged 2 Months	−0.3491	−1.06	0.0011	−0.3503	−1.07	0.55	0.55	−0.46
Net Equity Expansion	−0.1455	−2.03	0.0041	−0.1461	−2.04	0.97	0.98	−0.16
Stock Variance	−0.2039	−0.58	0.0003	−0.2142	−0.61	0.62	0.63	−3.65
Cross Sectional Premium	2.1014	3.03	0.0115	2.0595	2.96	0.98	0.98	−3.44
Dividend Premium	0.0000	0.07	0.0000	−0.0000	−0.17	0.98	0.99	−0.00
Number of IPOs	−0.0000	−0.38	0.0002	−0.0000	−0.35	0.86	0.87	0.00
Average First-day IPO Returns	0.0000	0.33	0.0002	0.0000	0.36	0.67	0.68	0.00
NYSE Share Turnover	−0.0022	−0.46	0.0003	−0.0023	−0.48	0.97	0.98	−0.02
Closed-End Fund Discount	0.0001	0.20	0.0001	0.0001	0.26	0.96	0.97	0.00
Equity Share in New Issues	−0.0402	−2.58	0.0104	−0.0404	−2.59	0.69	0.69	−0.02

Number of IPOs, Average First-day IPO Returns, NYSE Share Turnover, and Closed-end Fund Discount.

Predictors that appear to be significant at better than the 5% level (i.e., even when we consider the test to be two-sided) are Dividend Yield, Book to Market Ratio, Net Equity Expansion, Cross Sectional Premium, and Equity Share in New Issues; all of these variables have bias-corrected t-statistics greater than 2 in absolute value (fifth column of Table 2). With the lone exception of Equity Share in New Issues, all of the predictors that have a bias-adjusted t-statistic greater than 2 also have an autoregressive root greater than 0.97—that is, fairly close to 1. Curiously enough, the bias corrections make a big difference only for Dividend to Price Ratio and Earnings to Price Ratio and matter somewhat for Book to Market Ratio; note that these three are exactly the variables by which such corrections were motivated in the first place. The bias correction has little effect on the other 18 predictors.

The behavioral variables—Dividend Premium, Number of IPOs, Average First-day IPO Returns, NYSE Share Turnover and Closed-end Fund Discount —are not statistically significant predictors of the equity premium when judged by the in-sample criterion (with or without bias corrections).

Overall, we have evidence that a large number of variables are statistically significant predictors of the equity premium, even after bias corrections are applied.

3.2. Out-of-sample inference about predictability: invariant to sample split date

Table 3 presents, side by side, the in-sample results and the out-of-sample (split sample–invariant) results on predictability. Column [1] gives the bias-corrected in-sample t-statistic, and column [2] gives the probability that a standard normal variable is larger than the absolute value of that t^c-statistic. Column [3] reports $\text{mean}_\tau\{\mathbb{t}(\text{MSPE}-\text{adj})_\tau\}$, where the mean is computed over τ; column [4] gives the bootstrap-determined p-value of the *previous* column's statistic. Analogously, column [5] reports $\max_\tau\{\mathbb{t}(\text{MSPE}-\text{adj})_\tau\}$ with the maximum taken over τ; its bootstrap-determined p-value is given in column [6].

The following observations can be made about the results reported in Table 3.

- *None* of the variables that are in-sample insignificant at the 10% level are significant at the 10% level in the two out-of-sample tests. Results of the in-sample bias-corrected test and of the out-of-sample, sample split–invariant tests are in broad agreement.
- All but one of the predictors appear to be somewhat "less significant" when judged by the two out-of-sample tests.[5]
- Dividend to Price Ratio is the only variable that appears "more significant" when judged by the two out-of-sample criteria than by the in-sample criterion.
- The two out-of-sample criteria agree in a rough sense. In particular, their p-values are usually within a multiple of 2.
- There are five variables for which the in-sample bias-corrected t-statistic exceeds 2 in absolute value (fifth column of Table 2): Dividend Yield, Book to Market Ratio, Net Equity Expansion, Cross Sectional Premium, and Equity Share in New Issues. Of these, only Net Equity Expansion is insignificant out-of-sample. Each of the other four variables is significant at the 5% level by at least one of the two out-of-sample criteria.
- Every variable shown to be "very insignificant" in-sample (here, having the in-sample bias-corrected t-statistic's p-value exceed 15%) is shown to be "even more insignificant" by the two out-of-sample statistics (i.e., the p-values for the two out-of-sample statistics exceed 30%).

Overall, we do not find much disagreement between in-sample and out-of-sample predictability criteria—provided the latter are invariant to the choice of sample split date. The bias-corrected in-sample t-test as well as the mean and the maximum out-of-sample tests all tell much the same story as regards whether the equity premium is or is not reliably predicted by a given variable.

3.3. Bootstrap null distribution: selected percentiles of mean and maximum statistics

For our proposed out-of-sample $\text{mean}_\tau\{\mathbb{t}(\text{MSPE}-\text{adj})_\tau\}$ and $\max_\tau\{\mathbb{t}(\text{MSPE}-\text{adj})_\tau\}$ sample split–invariant tests, p-values are a function of the test statistic's observed value and also of its

[5] A test statistic is either significant or insignificant at the chosen (a priori) significance level (e.g., 5%). It follows that such modifiers as "less" or "more" or "very" (in)significant are, strictly speaking, abuses of statistical terminology. Such wording serves as shorthand for a longer statement; for example, a claim that some predictor is "less significant" judged by test X than by test Y is supposed to mean that if the null hypothesis were true, the probability of observing as large or larger Y as actually observed is smaller than the probability of observing as large or larger X as actually observed.

Table 3
The in-sample bias corrected t-statistic (column 1) is followed by its p-value (column 2). The out-of-sample sample split invariant $\text{mean}_\tau\,\mathbb{t}(\text{MSPE}-\text{adj})_\tau$ (column 3) is followed by its bootstrap determined p-value (column 4). The out-of-sample sample split invariant $\max_\tau\,\mathbb{t}(\text{MSPE}-\text{adj})_\tau$ (column 5) is followed by its bootstrap determined p-value (column 6).

	t^c	$\leftarrow$ p-value	$\text{Mean}_\tau\,\mathbb{t}$	$\leftarrow$ p-value	$\text{Max}_\tau\,\mathbb{t}$	$\leftarrow$ p-value
Dividend to Price Ratio	1.2817	0.1000	1.3838	0.0451	2.7530	0.0514
Dividend Yield	2.5106	0.0060	1.4412	0.0197	2.8287	0.0129
Book to Market Ratio	2.3800	0.0087	0.7450	0.1233	2.4635	0.0609
Earnings to Price Ratio	1.5410	0.0617	0.7922	0.1135	2.1455	0.1170
Dividend Payout Ratio	0.2827	0.3887	−1.3445	0.9333	−0.0266	0.8496
Treasury Bill Rate	−1.7667	0.0386	0.8038	0.0903	1.7881	0.1380
Long Term Yield	−1.3796	0.0839	0.7718	0.0989	1.6793	0.1728
Long Term Return	1.5738	0.0578	0.7897	0.1029	1.4374	0.2240
Term Spread	1.3675	0.0857	0.6290	0.1273	1.4914	0.2154
Default Yield Spread	1.5188	0.0644	0.1110	0.2893	1.4789	0.2297
Default Return Spread	1.0364	0.1500	−0.0611	0.3528	0.2500	0.7138
Inflation Lagged 2 Months	−1.0671	0.1430	−0.4874	0.5523	0.5439	0.5930
Net Equity Expansion	−2.0350	0.0209	−0.3316	0.4658	1.2130	0.3260
Stock Variance	−0.6112	0.2705	−0.7976	0.7147	−0.3131	0.9179
Cross Sectional Premium	2.9554	0.0016	1.4509	0.0205	2.4760	0.0297
Dividend Premium	−0.1682	0.4332	0.2934	0.2151	1.3407	0.2717
Number of IPOs	−0.3466	0.3644	0.0124	0.3341	1.1921	0.3188
Average First-Day IPO Returns	0.3578	0.3603	−0.0266	0.3351	1.3230	0.2596
NYSE Share Turnover	−0.4836	0.3143	−0.4910	0.5476	0.0106	0.8267
Closed-End Fund Discount	0.2571	0.3985	0.1304	0.2727	1.1134	0.3626
Equity Share in New Issues	−2.5853	0.0049	1.3748	0.0285	1.9659	0.0929

Table 4
The 90th, 95th, and the 99th percentiles of the bootstrap null distribution for our $\text{mean}_\tau[\mathbb{t}(\text{MSPE}-\text{adj})_\tau]$ and $\max_\tau[\mathbb{t}(\text{MSPE}-\text{adj})_\tau]$ sample split invariant statistics, together with the OLS estimate $\hat{\rho}$ of the autoregressive parameter in Eq. (3) and the correlation between the residuals in Eqs. (2) and (3).

	Mean$\mathbb{t}$90	Mean$\mathbb{t}$95	Mean$\mathbb{t}$99	Max$\mathbb{t}$90	Max$\mathbb{t}$95	Max$\mathbb{t}$99	$\hat{\rho}$	$\text{Corr}(\hat{u}_t, \hat{w}_t)$
Dividend to Price Ratio	1.0194	1.3464	1.9006	2.4700	2.7584	3.3726	0.99	−0.98
Dividend Yield	0.7398	1.0694	1.7019	1.9404	2.2846	2.9325	0.99	−0.08
Book to Market Ratio	0.8566	1.2002	1.8243	2.2068	2.5456	3.1729	0.98	−0.83
Earnings to Price Ratio	0.8513	1.1801	1.7814	2.2284	2.5810	3.1719	0.99	−0.84
Dividend Payout Ratio	0.7748	1.1038	1.7740	2.0050	2.3316	2.9493	0.99	−0.01
Treasury Bill Rate	0.7506	1.0681	1.7385	1.9486	2.3087	2.9246	0.99	−0.03
Long Term Yield	0.7632	1.1166	1.7847	2.0014	2.3532	3.0198	0.98	−0.12
Long Term Return	0.8055	1.1507	1.7940	1.8994	2.2129	2.7914	0.04	0.13
Term Spread	0.7614	1.0887	1.7607	1.9315	2.2600	2.8933	0.96	−0.04
Default Yield Spread	0.7615	1.1121	1.7419	1.9691	2.2842	2.8938	0.98	−0.25
Default Return Spread	0.7815	1.1261	1.7596	1.9044	2.2201	2.8275	−0.15	0.11
Inflation Lagged 2 Months	0.8093	1.1607	1.7556	1.9101	2.2208	2.8156	0.52	−0.03
Net Equity Expansion	0.7761	1.1037	1.7784	1.9469	2.3134	2.9186	0.97	−0.06
Stock Variance	0.8330	1.1597	1.7575	1.9730	2.2711	2.8478	0.60	−0.36
Cross Sectional Premium	0.7797	1.0876	1.7025	1.9584	2.2737	2.8716	0.98	−0.03
Dividend Premium	0.7507	1.0860	1.7068	1.9992	2.3322	2.9943	0.96	−0.27
Number of IPOs	0.7866	1.1126	1.7594	1.9368	2.2736	2.9246	0.89	0.09
Average First-day IPO Returns	0.8000	1.1293	1.7643	1.9031	2.2460	2.8650	0.61	0.15
NYSE Share Turnover	0.7674	1.1086	1.7464	1.9773	2.3083	2.9197	0.96	−0.02
Closed-End Fund Discount	0.7598	1.1011	1.7463	1.9704	2.2945	2.9526	0.97	0.13
Equity Share in New Issues	0.8175	1.1506	1.7459	1.9185	2.2423	2.8515	0.63	−0.07

bootstrap determined null distribution. As a result, Table 3 is not informative regarding the distributions of the mean and maximum statistics under the null hypothesis of no predictability. Yet we are interested in such questions as: What do the null distributions of these two statistics look like? Are the mean and maximum statistics pivotal, or do they depend on the parameters of the (bootstrap) data-generating process?

Suppose we found a single null distribution for the mean statistic across all 21 predictors and also a single null distribution for the maximum statistic across those same predictors. In that case, our proposed statistics would be pivotal and independent of the parameters used to generate the data. If, on the contrary, we found the null distributions to be considerably different across the 21 predictors, the implication would be that the two statistics are not pivotal and so *are* sensitive to those data-generating parameters.

Table 4 reports the 90th, 95th, and 99th percentiles of the bootstrap determined null distribution of the mean and maximum statistics. We can make the following remarks regarding the null distributions of the sample split–invariant $\text{mean}_\tau\{\mathbb{t}(\text{MSPE}-\text{adj})_\tau\}$ and $\max_\tau\{\mathbb{t}(\text{MSPE}-\text{adj})_\tau\}$ statistics reported in that table.

It seems reasonable to suppose that the $\text{mean}_\tau\{\mathbb{t}(\text{MSPE}-\text{adj})_\tau\}$ and $\max_\tau\{\mathbb{t}(\text{MSPE}-\text{adj})_\tau\}$ statistics are pivotal because their null distributions do not depend strongly on the parameters of the bootstrap data-generating process. The percentiles under the null hypothesis of no predictability across the 21 focal predictors are similar, and they are not more dissimilar across predictive variables than are the percentiles of $\mathbb{t}(\text{MSPE}-\text{adj})_\tau$. Under conditional homoskedasticity, $\mathbb{t}(\text{MSPE}-\text{adj})_\tau$ is known to be pivotal with respect to the parameters of the data-generating process and for the type of forecasting regressions considered here (Clark and McCracken, 2001).[6]

[6] The null distribution of the $\mathbb{t}(\text{MSPE}-\text{adj})_\tau$ statistic does depend on the constant to which the ratio of evaluation data points to estimation data points con-

However, the bootstrap distributions generated under the null of no predictability depend to some extent on the correlation between the error term in Eq. (2) and that in Eq. (3). More specifically: the higher the correlation, the higher the percentiles of the null distribution for the corresponding predictor. Thus the previous point's conjecture might just as well turn out to be false.

If we are comfortable (for exploratory purposes) with the rough sense in which $\mathbb{t}(\text{MSPE} - \text{adj})_\tau$ is approximately normal for a fixed sample split, then we can propose the following rough criterion: if the data-based $\text{mean}_\tau\{\mathbb{t}(\text{MSPE} - \text{adj})_\tau\}$ is greater than 1 (resp., 1.5, 2), then the bootstrap sample split–invariant test would likely reject the null of no predictability at the 10% (resp., 5%, 1%) level of significance. Similarly, we can say that if the data-based $\max_\tau\{\mathbb{t}(\text{MSPE} - \text{adj})_\tau\}$ is greater than 2 (resp., 2.5, 3) then the sample split–invariant bootstrap test would probably reject the null of no predictability at the 10% (resp., 5%, 1%) level of significance. Even a rule of this approximate nature could save some programming and simulation time, though it must be used with an eye toward the in-sample results. From Table 3 we see that the in-sample test of predictability for a given variable is extremely informative about the complete out-of-sample simulation outcomes.

3.4. Summarizing out-of-sample predictability in graphical form

Fig. 2 displays complete information on a real-time investor's investment outcomes as a function of the sample split. Each subfigure within Fig. 2 presents the out-of-sample R-squared and the bootstrap-determined p-value for the MSPE-adj t-statistic for each sample split date $\tau \in [\text{int}(.05T),\ T - \text{int}(.05T)]$. For each τ we bootstrap the $\mathbb{t}(\text{MSPE} - \text{adj})_\tau$ statistic, rather than the MSPE-adj(τ) statistic, because the former is pivotal and the latter is not. MacKinnon (2009) emphasizes that one should bootstrap pivotal quantities whenever possible because doing so yields an asymptotic refinement: the error in rejection probability committed by the pivot-based *bootstrap* test is of lower order in the sample size than is the error in rejection probability of the *asymptotic* test based on the same pivot (Beran, 1988; Hall, 1992).

The 21 sub-figures show how well an investor (starting in, say, January 2000; the actual years plotted vary depending on data availability) would have done by using each of our 21 focal predictors—as compared with using the recursive mean—to forecast the equity premium out-of-sample. The following list offers brief comments about the out-of-sample predictive success of each variable as a function of the sample split date τ. (When discussing *overall* in-sample and sample split–invariant out-of-sample inference about predictability, we refer always to the results in Table 3. We discuss the predictors in Fig. 2 consecutively from left to right, and down Fig. 2.)

The (log of the) **Dividend to Price Ratio** is a reasonably accurate out-of-sample predictor of the equity premium. It loses predictive power around year 1973, but it regains power in the late 1990s and at the start of the new millennium. Then, from about year 2002, it is once again unable to outperform the recursive mean. The sample split–invariant mean and maximum tests show that the dividend-to-price ratio outperforms the recursive mean overall. This is the *only* variable for which our out-of-sample tests reject the null of no predictability at better significance levels than does the in-sample test. The out-of-sample R-squared is negative for the most part; it becomes positive only for a short period around year 2000. The (log of the) **Dividend Yield** displays the same predictive pattern as Dividend to Price Ratio. All tests—in-sample and sample split–invariant out-of-sample—show this predictor to be significant at better than the 2% significance level.

The **Book to Market Ratio** loses predictive power around January 1950 and regains it only for a short period around the start of the new millennium. Although the in-sample tests of predictive power for Book to Market Ratio and Dividend Yield have similar p-values, the out-of-sample tests yield conflicting results for these two variables; thus, for most of the sample splits, the book-to-market ratio would *not* have been an accurate predictor for an investor making decisions in real time. However, that ratio would have been helpful to an investor starting to time the market a few months before or after January 2000.

The (log of the) **Earnings to Price Ratio** lost out-of-sample predictive power shortly after year 1950 and has never regained it. As expected, both the mean and the maximum sample split–invariant out-of-sample tests show this variable to be only marginally significant. The (log of the) **Dividend Payout Ratio** has never been an accurate out-of-sample predictor of the equity premium for the simple reason that it consistently underperforms the recursive mean benchmark.

The **Treasury Bill Rate** outperforms the recursive mean benchmark out-of-sample until the start of the 1970s but never thereafter. This predictor is shown to be marginally significant by the sample split–invariant out-of-sample tests. The **Long Term Yield** exhibits forecasting patterns strongly similar to the T-bill rate, which suggests that these two variables capture the same information regarding the economy's future state. We expected to see a dramatic difference between the two since the Great Recession started; their continued similarity is puzzling given that short-term US debt has in recent years come to resemble money (Cochrane, 2011), an asset that differs from long-term debt.

The **Long Term Return** outperforms the recursive mean from the mid-1950s until the mid-1970s. The **Term Spread** outperforms the recursive mean until the start of the 1970s. This variable's sample split–invariant mean statistic has a p-value of .14 and its maximum statistic has a p-value of .21. It is probably fair to interpret Fig. 2 as indicating that the term spread is actually a better out-of-sample predictor than our two out-of-sample tests would suggest.

The **Default Yield Spread** is an accurate out-of-sample predictor of the equity premium from the mid-1950s to the mid-1960s. Thereafter, it fails to outperform the recursive mean benchmark. The **Default Return Spread** exhibits stable but unimpressive out-of-sample performance. The **Lagged Inflation** has never been an accurate out-of-sample predictor of the equity premium.

The **Net Equity Expansion** is one of the few variables that the in-sample test shows to be a significant predictor of the equity premium. However, all of the out-of-sample evidence points to unimpressive out-of-sample predictive performance when compared with the recursive mean. This is the variable on which in-sample tests and sample split–invariant out-of-sample tests disagreed the most.

The **Stock Variance** has never been an accurate out-of-sample predictor of the equity premium.

The **Cross Sectional Premium** (Polk et al., 2006) is an excellent predictor according to both in-sample and out-of-sample tests. Note, however, that data for the cross-sectional premium is not available for recent years; hence we cannot say how this variable would have performed during the last decade or so.

The (log) **Dividend Premium** has outperformed the recursive mean only sporadically: around the mid-1960s and around January 2000. The **Number of Initial Public Offerings** nearly outperforms the recursive mean in late 1960s. In general, this variable seems *not* to be an accurate out-of-sample predictor of the equity premium. The **Average First-day IPO Returns** comes close to

verges as the sample size grows to infinity. The assumption that this ratio is constant is not supported by our calculations for each possible sample split. Of course, statistical behavior as the sample size approaches infinity is an abstraction with no clear meaning in practice; our sample is always finite. Hence the only question is whether or not this abstraction yields an accurate approximation for the samples that are typically available.

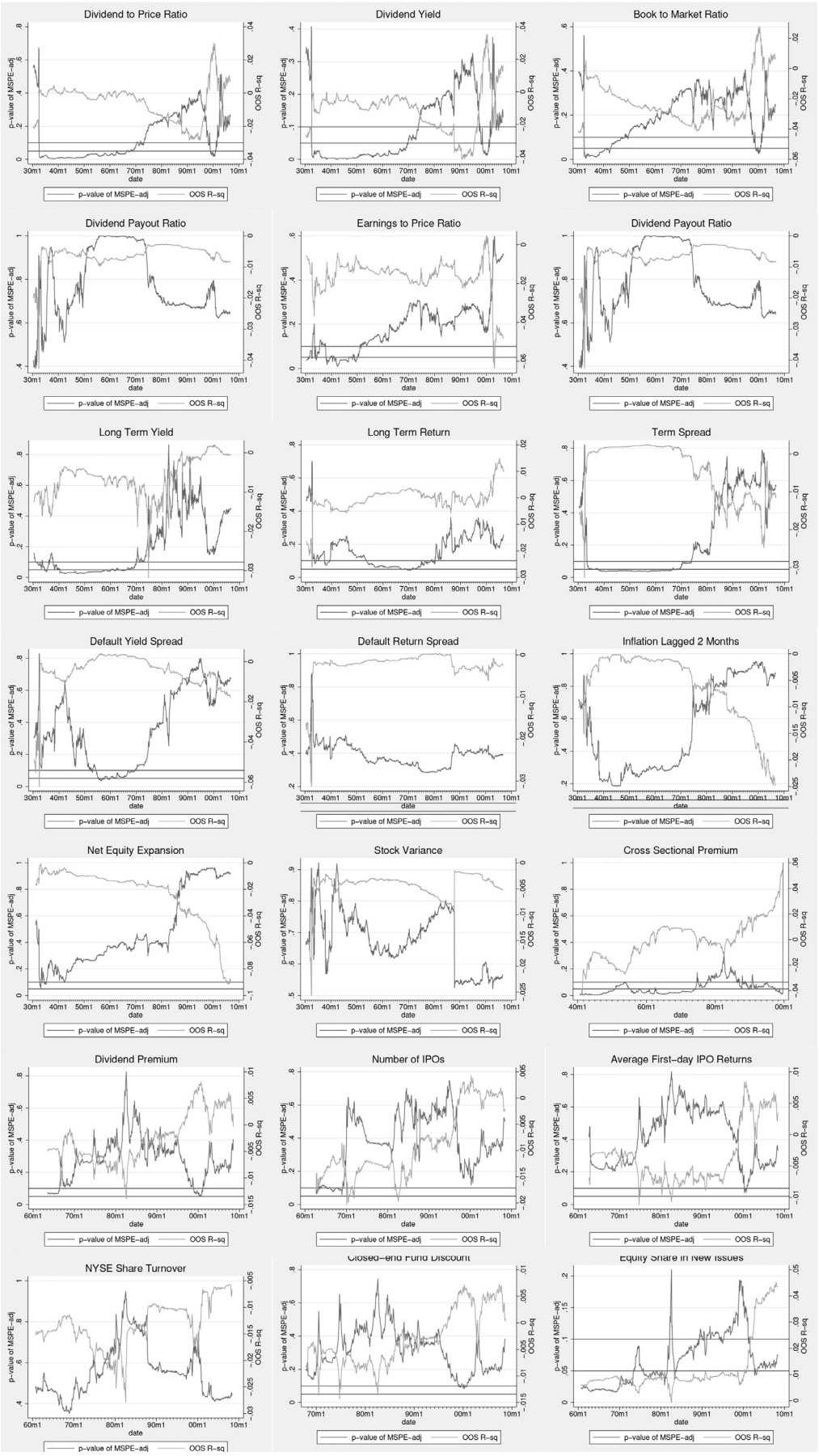

Fig. 2. Plots of the *p*-value of the $\mathbb{t}$(MSPE-adj) statistic and the out-of-sample R-squared versus τ (the start of the out-of-sample forecasting exercise).

outperforming the recursive mean around January 2000, but overall it is not an accurate out-of-sample predictor. The **NYSE Share Turnover** is not an accurate predictor of the equity premium on an out-of-sample basis. The **Closed-end Fund Discount** is a poor out-of-sample predictor of the equity premium, which is surprising in light of how much attention this variable has received as an indicator of sentiment (Lee et al., 1991; Neal and Wheatley, 1998; Zweig, 1973). The **Equity Share in New Issues** is a reliable predictor of the equity premium when judged by any in-sample or out-of-sample criterion. It loses predictive power for most of the 1990s but regains it in later years.

We can summarize the preceding analysis in two main points, as follow. First, there are periods during which the equity premium is difficult to predict and so a forecaster can hardly do any better than simply using the recursive mean. From the mid-1970s until the mid-1990s, for example, reliable predictors are hard to find; in fact, the only variables of any use in this time span are the Cross Sectional Premium and the Equity Share in New Issues. (Even so, there are a number of sample split dates τ during this period for which those two predictors *fail* to outperform the recursive mean.) This finding has implications for "combination" forecast methods such as those proposed by Rapach et al. (2010). Such a method works for most sample split dates because it "irons out" parameter instability across models and model uncertainty. Yet there seem always to be some sample split dates for which good predictors are hard to find, and it is an open question whether forecast combination can deliver superior performance when few (a fortiori none) of the constituent predictors can improve on the recursive mean benchmark.[7]

Second, the observed predictive patterns are fairly similar across variables derived from related economic intuition. Predictors reflecting economic fundamentals, i.e., dividend/price, dividend yield, book/market, earnings/price) exhibit strikingly similar forecasting patterns. The "interest rate" variables (T-bill rate, long-term yield, long-term return, term spread, default return spread, default yield spread, and inflation) likewise exhibit closely similar forecasting patterns. Another group of variables that exhibit similar predictive patterns includes net equity expansion, the number of IPOs, and the equity share in new issues. In short: the graphs plotted in Fig. 2 confirm our expectation that economically similar variables display similar patterns of predictive strength and weakness, versus the recursive mean, as a function of the chosen sample split date.

3.5. Out-of-sample inference about predictability: invariant to sample split date and robust to data mining

When testing the ability of financial variables to predict stock returns, data mining is a serious concern.[8] Lo and MacKinlay (1990) and Foster et al. (1997) stress this issue for in-sample tests of security returns predictability. Until recently, out-of-sample tests have been viewed as a viable preventive against data mining. However, Inoue and Kilian (2004) and also Rapach and Wohar (2006) argue that data mining should be of concern also in out-of-sample tests of predictability, and especially when a large number of predictive variables is considered. They suggest addressing this problem by using corrected critical values obtained via a bootstrap procedure.

More recently, Hubrich and West (2010) and Clark and McCracken (2012) propose taking the maxima of various statistics for simultaneously judging whether a small set of alternative models that nest within a benchmark model improve upon that benchmark model's MSPE. Clark and McCracken (2012) propose a fixed-regressor "wild" bootstrap procedure for evaluating the sampled distributions of the maximum statistics they study.

In this paper we consider only the variables surveyed by Welch and Goyal (2008) plus a few behavioral predictors. Although we do not use data-mining methods to detect viable predictors, data mining still could be a concern given that we study so many (21) variables. That is, their sheer number makes it more likely that one or more exhibit, just by chance, a statistically significant association with the predictand. Inoue and Kilian (2004); Rapach and Wohar (2006), and Clark and McCracken (2012) present ideas that are relevant to our own paper's data-mining issues.

So far, our bootstrap procedure has assumed that the predictive power of each variable is tested separately. That we examine 21 possible predictors actually increases the chances of coming to a wrong conclusion. Therefore, when testing predictability we control for data mining by applying to our test statistic the ideas first proposed by Inoue and Kilian (2004) and also used by Rapach and Wohar (2006).

For this purpose, we start by specifying the null hypothesis as $H_0 : \beta_1^j = 0$ for all j, where $j = 1, \ldots, 21$ indexes the variables being tested for predictive power. We specify the alternative hypothesis as $H_1 : \beta_1^j \neq 0$ for some j, where β_1^j is the slope in Eq. (2) when the predictive variable is X_t^j.

As in-sample test statistics, we use the maximum and the mean of the square of the in-sample t–statistic for testing that the slope is 0 across the variables of interest for a two sided test. Note that in such bivariate regressions the *square* of that in-sample t-statistic is numerically equivalent to the F-statistic from testing for whether that slope is 0. In other words, we use both $\max_{j=1,\ldots,21}\{t^2_{\hat{\beta}_1^j}\}$, where $t^2_{\hat{\beta}_1^j}$ is the square of the t-statistic corresponding to $\hat{\beta}_1^j$ (we will call this statistic *max-t-squared*), and $\text{mean}_{j=1,\ldots,21}\{t^2_{\hat{\beta}_1^j}\}$ (or *mean-t-squared*). We use the *square* of the t-statistic because this is a two-sided test.

Our two out-of-sample test statistics are as follows.

1. $\text{mean}_{j=1,\ldots,21}\{\text{mean}_\tau \mathbb{t}^j(\text{MSPE} - \text{adj})_\tau\}$, also known as the *double mean*. In words, we take the mean across the 21 variables of the sample split–invariant mean statistic; this is a "double" mean because we first average across sample split dates and then average across variables.
2. $\max_{j=1,\ldots,21}\{\max_\tau \mathbb{t}^j(\text{MSPE} - \text{adj})_\tau\}$, or the *double max*. Here we take the maximum across the 21 variables of the sample split–invariant max statistic; it is a "double" max because we first take the maximum across sample split dates and then across variables.

Inoue and Kilian (2004) derive asymptotic distributions for their maximum in-sample and out-of-sample statistics under the null hypothesis of no predictability. But since the limiting distributions are generally data dependent, these authors recommend that bootstrap procedures be used in practice. Our bootstrap method (described in Section 2.4) is used—while imposing the null hypothesis $H_0 : \beta_1^j = 0$ for all $j = 1, \ldots, 21$ in the bootstrap data-generating process—to determine the null distributions of our mean-t-squared, max-t-squared, double mean, and double max statistics.

The results are reported in Table 5. The in-sample data-mining–robust test based on mean-t-squared ($\text{mean}_{j=1,\ldots,21}\{t^2_{\hat{\beta}_1^j}\}$) rejects

[7] Rapach et al. (2010) start their out-of sample forecasting exercises in the first quarters of 1965, 1976, and 2000. One can infer from the preceding analysis of each variable that the first quarter of 1976 is the most challenging split date; nonetheless, significant individual predictors can be found even for that choice. Not surprisingly, these authors find that particular sample split to generate the weakest (though still significant) result. Hence an intriguing question is whether a forecast combination technique could deliver superior performance for a date on which *none* of the individual predictors could.

[8] We are grateful to an anonymous referee for proposing tests that are not only sample split–invariant but also robust to data mining.

Table 5
The mean-t-squared ($\text{mean}_{j\in 1,\ldots,21} t^2_{\hat{\beta}_1^j}$), max-t-squared ($\max_{j\in 1,\ldots,21} t^2_{\hat{\beta}_1^j}$), double mean $\text{mean}_{j\in 1,\ldots,21}[\text{mean}_\tau \mathbb{t}^j(\text{MSPE}-\text{adj})_\tau]$ and double max $\max_{j\in 1,\ldots,21}[\max_\tau \mathbb{t}^j(\text{MSPE}-\text{adj})_\tau]$ test statistics (column 1), their bootstrap determined p-value (column 2) and the 90th, 95th and 99th percentiles of the bootstrap determined null distribution of the respective statistic (last three columns).

	Statistic	p-value	90 percentile	95 percentile	99 percentile
Mean-t-squared	2.8473	0.0000	1.4553	1.6007	1.9365
Max-t-squared	9.1607	0.0647	8.3271	9.6241	12.6491
Double mean	0.3231	0.0003	−0.0639	0.0041	0.1532
Double max	2.8287	0.1627	3.0051	3.2067	3.7371

the hypothesis of no predictability at any standard significance level, and the max-t-squared statistic ($\max_{j=1,\ldots,21}\{t^2_{\hat{\beta}_1^j}\}$) rejects the same hypothesis at the 6% significance level. Similarly, the out-of-sample data-mining–robust double mean test rejects the null hypothesis (of no predictability) at any level and the double max test *fails* to reject the null at standard significance levels (yet since the p-value of .16 is relatively low, there is some evidence against the null).

In sum, three of our four predictability tests that control for data mining *reject* the null hypothesis of no predictability. This conclusion is evidently not driven by the distinction between in-sample and out-of-sample testing, and it runs counter to Welch and Goyal's (2008) claim of no predictability.

This extension of our bootstrap method to sample split–invariant out-of-sample inference that is also robust to data mining demonstrates the flexibility of our method—especially as compared to the tests proposed by Rossi and Inoue (2012) and Hansen and Timmermann (2012).

4. Monte Carlo experiments: validity of the bootstrap procedure

In this section we use Monte Carlo experiments to study the validity of our bootstrap procedure. We also examine how accurate is the distribution characterized and tabulated by Rossi and Inoue (2012). Throughout, we assume that the sample split date τ falls within the interval $[\text{int}(.15T),\ T-\text{int}(.15T)]$. This choice reflects our intention to compare this test procedure to the one described by Rossi and Inoue (2012), who do not tabulate critical values for $\tau \in [\text{int}(.05T),\ T-\text{int}(.05T)]$.

When studying the 21 predictors we use $\tau \in [\text{int}(.05T),\ T-\text{int}(.05T)]$ because (a) the predictor with the *smallest* sample size has 548 observations and (b) traditional predictors such as the dividend-to-price ratio have about 1020 observations available. We consider $\text{int}(.05\times 548)=27$ to be a sufficient sample size for deriving an initial set of parameter estimates or reliable out-of-sample averages. We remark that the Rossi and Inoue (2012) distributions being tabulated only for certain intervals renders their method less attractive because an empiricist would be constrained by those particular tabulations; our bootstrap method does not suffer from that limitation.

All our Monte Carlo experiments are calibrated to the moments of the actual data. We start by estimating the systems in Eqs. (2) and (3) for a given predictor (e.g., dividend/price). We estimate the system parameters via ordinary least squares and then store them. Let $\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\mu},\ \hat{\rho},\ \text{Std}(\hat{u}_t),\ \text{Std}(\hat{w}_t)$ and $\text{Corr}(\hat{u}_t,\hat{w}_t)$ be the respective unrestricted estimates, including the parameters and the residuals. Let $\tilde{\beta}_0$ be the restricted estimate with $\beta_1=0$ imposed (i.e., the unconditional average of the equity premium).

For each Monte Carlo round we initiate the recursion in Eq. (3) by drawing with equal probability from the actual sample path of X_t and setting X_0^m equal to that value (here the superscript $m=1,2,\ldots$ identifies the Monte Carlo round). Then we generate $X_t^m=\hat{\mu}+\hat{\rho}X_{t-1}^m+w_t^m$ for $t=1,\ldots,T$. For each t we generate a bivariate normal vector of true errors $[u_t^m, w_t^m]$ that is serially independent and identically distributed (i.i.d.) and whose covariance matrix is calibrated to have the same parameters as the covariance matrix of the residual vector $[\hat{u}_t,\hat{w}_t]$. Thus $\text{Std}(\hat{u}_t)=\text{Std}(u_t^m)$, $\text{Std}(\hat{w}_t)=\text{Std}(w_t^m)$, and $\text{Corr}(\hat{u}_t,\hat{w}_t)=\text{Corr}(u_t^m,w_t^m)$.

If we generate Monte Carlo data under the null hypothesis, then $R_t^m=\tilde{\beta}_0+u_t^m$; that is, the equity premium is just the sample's average equity premium plus the error term. If we generate data under the alternative hypothesis, then $R_t^m=\hat{\beta}_0+\hat{\beta}_1X_{t-1}^m+u_t^m$.

In this section our Monte Carlo experiments are calibrated to the moments of Eqs. (2) and (3), where the dividend-to-price ratio plays the role of X_t. In particular, $\text{Corr}(u_t^m,w_t^m)=-0.9768$, $\text{Std}(u_t^m)=.0557$, $\text{Std}(w_t^m)=.0565$, and $\hat{\rho}=.9931$. Recall that, in theory, the dividend-to-price ratio cannot be nonstationary in the population $\rho<1$. Since the price is simply the discounted sum of future dividends, it follows that the dividend-to-price ratio cannot just drift away (up or down) to infinity; the dividend and the price series must be co-integrated. Yet when Eqs. (2) and (3) are calibrated to the dividend/price ratio, their respective systems will—in *small* samples—be rather ill behaved and "nonstationary looking". Perhaps it would be more accurate to say "in *finite* samples" given that $T=1020$ is not really that small.

In Experiment 1, we generate 300 Monte Carlo paths of length $T=1020$ under the null hypothesis (of no return predictability) that $R_t^m=.0062+u_t^m,\ X_t^m=-.0240+.9931X_{t-1}^m+w_t^m$, and $\text{Corr}(u_t^m,w_t^m)=-0.9768$. Then, for each of those Monte Carlo paths we carry out 300 bootstrap replications as described in Section 2.4. For each sample path and across the 300 replications we determine the 90th, 95th, and 99th percentiles—in the bootstrap distribution generated under the null—of the statistics $\text{mean}_\tau\{\mathbb{t}(\text{MSPE}-\text{adj})_\tau\}$ and $\max_\tau\{\mathbb{t}(\text{MSPE}-\text{adj})_\tau\}$. In other words, for each Monte Carlo path we repeat exactly the same bootstrap procedure described in Section 2.4 and applied to the 21 predictors.

Now, if the calculated $\max_\tau\{\mathbb{t}(\text{MSPE}-\text{adj})_\tau\}$ statistic for the Monte Carlo path is larger than the 90th percentile of the bootstrap distribution of that statistic, then we record rejection at the 10% significance level. We proceed analogously for the 5% and 1% significance levels and then proceed likewise for $\text{mean}_\tau\{\mathbb{t}(\text{MSPE}-\text{adj})_\tau\}$. The number of Monte Carlo and bootstrap replications is fairly low because the computational burden quickly becomes an obstacle (recall that each calculation must be performed 300×300 times. We experimented with various numbers of bootstrap replications (100, 190, 300, 999, 2999, 9999) and assembled the results as in Table 3. Overall we find that more than 999 replications seldom yield non-negligible differences but that fewer than 300 replications nearly always yield erratic results. The ideal scenario would involve performing 999 bootstrap replications on 999 Monte Carlo paths rather than 300 by 300.

Table 6 shows that our bootstrap tests are slightly oversized but still fairly accurate. The mean test has better size than the maximum test. At the 10% nominal significance level, the mean (resp.

Table 6
Panel A contains the Monte Carlo determined actual size of our mean and maximum tests at the stated significance level. For each of 300 Monte Carlo paths, sample size $T = 1020$, the given statistic is calculated, say $\text{mean}_\tau\,\mathbb{t}$. Then for this given Monte Carlo path 300 bootstrap replications are used to calculate, e.g., the 90th percentile of the $\text{mean}_\tau\,\mathbb{t}$ null bootstrap determined distribution. If the original $\text{mean}_\tau\,\mathbb{t}$ exceeds the 90th percentile of the bootstrap null distribution for this Monte Carlo path, the test rejects at the 10% significance level, and the average of this rejection indicator is calculated across the 300 Monte Carlo paths. Similar procedure is followed for the max statistic and for the other significance levels. Panel B: reports the average and the standard deviations of say 90th percentile in the null bootstrap distribution of the statistics across the 300 Monte Carlo rounds. Note that for each Monte Carlo round the 90th percentile of say the mean statistic is a *number* and only this number is used as a cut off point to determine the rejection of the test. Panel B simply reports the average of these numbers across Monte Carlo rounds, in other words, the numbers in Panel B are *not used* to determine rejection in Panel A at this stage. $\tau \in [\text{int}(.15T), T - \text{int}(.15T)]$.

Panel A				Panel B			
Rejection rate at				Bootstrap critical values			
significance level	10%	5%	1%		90 percentile	95 percentile	99 percentile
$\text{Mean}_\tau\,\mathbb{t}$	11.00%	6.67%	1.67%	Mean (respective percentile)	1.0060	1.3650	2.0200
				Standard deviation (respective percentile)	0.1100	0.1270	0.2020
$\text{Max}_\tau\,\mathbb{t}$	13.67%	8.00%	2.67%	Mean (respective percentile)	2.1320	2.5010	3.1560
				Standard deviation (respective percentile)	0.1510	0.1650	0.2100

Table 7
Panel C contains the Monte Carlo determined actual size of our mean and maximum tests, and of the mean L and max L tests of Rossi and Inoue (2012) at the stated significance level. For each of 10,000 Monte Carlo paths of size $T = 1020$ the given statistic is calculated, say mean L (Rossi and Inoue, 2012, eq.(11) p. 436) and max L (Rossi and Inoue, 2012, eq.(10) p.436). Then for this given Monte Carlo path the calculated statistic is compared to the respective percentile of the null distribution. If the original statistic exceeds the 90th percentile of the null distribution for this Monte Carlo path, the test rejects at 10% significance level, and the average of this rejection indicator is calculated across the 10,000 Monte Carlo paths. For the mean L and max L tests the critical values of the null distribution are taken from (Rossi and Inoue, 2012, Table 2(b) p.438). For $\text{mean}_\tau\,\mathbb{t}$ and $\text{max}_\tau\,\mathbb{t}$ the critical values are taken from Experiment 1, Table 6 Panel B mean(respective percentile). Panel D: repeats the average of say 90th percentile in the null bootstrap distribution of the statistics across the 300 Monte Carlo rounds in Experiment 1, and this average percentile is used as a critical value in all of the 10,000 Monte Carlo rounds here. $\tau \in [\text{int}(.15T), T - \text{int}(.15T)]$.

Panel C				Panel D			
Rejection rate at				Theoretical critical values			
significance level	10%	5%	1%		90 percentile	95 percentile	99 percentile
Mean L	16.60%	9.81%	2.76%	Mean L	0.8620	1.4560	2.8620
Max L	21.39%	13.25%	4.23%	Max L	2.0430	3.0640	5.6200
Rejection rate at				Bootstrap critical values			
significance level	10%	5%	1%		90 percentile	95 percentile	99 percentile
$\text{Mean}_\tau\,\mathbb{t}$	11.12%	5.52%	1.07%	Mean (respective percentile)	1.0060	1.3650	2.0200
$\text{Max}_\tau\,\mathbb{t}$	12.40%	6.42%	1.35%	Mean (respective percentile)	2.1320	2.5010	3.1560

max) test's actual size is 11% (resp. 13.67%). At the 5% nominal significance level, the mean (resp. max) test's actual size is 6.67% (resp. 8.00%).

In Experiment 2 (under the null hypothesis of no return predictability) we generate 10,000 Monte Carlo paths in the same way as in Experiment 1 but independent of the paths in that previous experiment. Here, however, we do not generate the bootstrap distributions for each Monte Carlo round and instead simply use as critical values the averages reported in panel B of Table 6. For example, from Experiment 1 we have the 90th percentile (a number) of the bootstrap distribution of (say) $\text{mean}_\tau\{\mathbb{t}(\text{MSPE} - \text{adj})_\tau\}$ for each of the 300 Monte Carlo rounds. We take the average of these 300 numbers (1.0060) and report that value in panel B of Table 6 and also in panel D of Table 7. Then, for each of the 10,000 Monte Carlo rounds, we reject the null hypothesis of Experiment 2 (at the 10% significance level) if the calculated $\text{mean}_\tau\{\mathbb{t}(\text{MSPE} - \text{adj})_\tau\}$ statistic exceeds 1.0060. Finally, the *rejection rate* is the average number of rejections across the 10,000 paths.

Table 7 shows that the size of our bootstrap tests is close to the nominal significance level. In particular, that size is more accurate than in Table 6; this result suggests that the oversized Table 6 tests were due in part to an insufficient number of Monte Carlo and bootstrap replications. Here again, our mean bootstrap test has better size than our maximum test. At the 5% nominal significance level, the mean (resp. max) bootstrap test's actual size is 5.52% (resp. 6.42%); at the 1% nominal significance level, the mean (resp. max) test's actual size is 1.07% (resp. 1.35%).

In Table 7 we can also see that the tests proposed by Rossi and Inoue (2012) are grossly oversized and perform worse than our tests. For example: at the 5% nominal significance level, their *mean* L test (Rossi and Inoue, 2012, eq. (11)) has actual size of 9.81% and their *max* L test (Rossi and Inoue, 2012, eq. (10)) has actual size of 13.25%; at the 1% nominal significance level the corresponding values are 2.76% and (respectively) 4.23%.[9]

In Experiment 3, we generate 300 Monte Carlo paths of length $T = 1020$ under the *alternative* hypothesis (of return predictability) that $R_t^m = .0353 + .0087X_{t-1}^m + u_t^m$ calibrated to the sample OLS estimates for the dividend-to-price ratio, where $X_t^m = -.0240 + .9931X_{t-1}^m + w_t^m$ and $\text{Corr}(u_t^m, w_t^m) = -0.9768$. Then we repeat the procedure from Experiment 1. Experiment 3 establishes that our tests are consistent in that they do reject the incorrect null hypothesis. The rejection rates in Table 8 are equivalent to the empirical power of our test under the alternative hypothesis that $\beta_1 = .0087$.

In Experiment 4, we generate 10,000 Monte Carlo paths of length $T = 1020$ under the alternative hypothesis (of return predictability) that $R_t^m = .0353 + .0087X_{t-1}^m + u_t^m$ before repeating the procedure from Experiment 2. Experiment 4 also confirms that our

[9] Rossi and Inoue (2012) characterize the distributions of mean L and max L statistics and then tabulate those distributions in their Table 2(b) for the case of a recursive forecasting scheme. The Monte Carlo experiments that we conduct, e.g., panel C of Table 7, indicate that the distribution tabulated by Rossi and Inoue is a poor approximation for asset pricing applications. A high autoregressive parameter of the predictor (e.g., $\rho > .9$) shifts the distributions of mean L and max L sharply to the right, even for large sample sizes such as 1020 time-series observations. The problem is even worse for combinations where ρ is close to unity and the absolute value of $\text{Corr}(u_t^m, w_t^m)$ is high. Yet in asset pricing applications, ρ near 1 and large $\text{Corr}(u_t^m, w_t^m)$ are more the rule than the exception. All traditional predictors (dividend/price, dividend yield, book/market, earnings/price, etc.) have such time-series properties.

Table 8
Description of Table 6 applies with the only difference being that the Monte Carlo paths of size $T = 1020$ here are generated under the alternative hypothesis: $R_t^m = .0353 + .0087X_{t-1}^m + u_t^m$. Notice that if the test procedure is consistent it must reject the incorrect null hypothesis.

Panel A				Panel B			
Rejection rate at				Bootstrap critical values			
significance level	10%	5%	1%		90 percentile	95 percentile	99 percentile
$\text{Mean}_\tau\,\mathbb{t}$	84.67%	66.67%	24.00%	Mean (respective percentile)	1.0020	1.3600	2.0100
				Standard deviation (respective percentile)	0.1120	0.1370	0.1980
$\text{Max}_\tau\,\mathbb{t}$	77.67%	59.00%	22.67%	Mean (respective percentile)	2.1210	2.4950	3.1700
				Standard deviation (respective percentile)	0.1600	0.1630	0.2130

Table 9
Description of Table 7 applies with the only difference being that the Monte Carlo paths of size $T = 1020$ here are generated under the alternative hypothesis: $R_t^m = .0353 + .0087X_{t-1}^m + u_t^m$. Notice that if the test procedure is consistent it must reject the incorrect null hypothesis.

Panel C				Panel D			
Rejection rate at				Theoretical critical values			
significance level	10%	5%	1%		90 percentile	95 percentile	99 percentile
Mean L	98.27%	93.39%	62.61%	Mean L	0.8620	1.4560	2.8620
Max L	99.05%	94.26%	63.71%	Max L	2.0430	3.0640	5.6200
		Bootstrap					
Rejection rate at				Bootstrap critical values			
significance level	10%	5%	1%		90 percentile	95 percentile	99 percentile
$\text{Mean}_\tau\,\mathbb{t}$	86.71%	66.46%	21.92%	Mean (respective percentile)	1.0060	1.3650	2.0200
				Standard deviation (respective percentile)	0.1100	0.1270	0.2020
$\text{Max}_\tau\,\mathbb{t}$	79.37%	58.20%	20.00%	Mean (respective percentile)	2.1320	2.5010	3.1560
				Standard deviation (respective percentile)	0.1510	0.1650	0.2100

tests are consistent. The mean L and max L tests of Rossi and Inoue (2012) appear to have greater power; yet because their tests are considerably oversized as seen in Table 7, such a comparison is misleading. One can easily construct a test with 100% power: if one calculates any test statistic but rejects all resulting values, then that test's power will always be 100%.

5. Conclusion

The recent literature on equity premium predictability has focused on out-of-sample evaluation methods. These methods require the researcher to choose a sample split date that divides the data into an in-sample estimation period and an out-of-sample evaluation period. Because the sample split date is always treated as a priori given, reported out-of-sample predictability results are germane only to the particular date used by the researcher.

In this paper we document that conclusions about predictability in this context are strongly dependent on the choice of a sample split date. In fact, this dependence is so prevailing that—with respect to proven in-sample equity premium predictors—a researcher may derive evidence supporting (or refuting) predictability simply by adjusting the sample split date. Hence there are many sample splits indicating strong out-of-sample predictability but also many other splits that indicate no evidence of such predictability.

We describe two ways of addressing this unfortunate state of affairs. The first approach is simply to plot (graph) the out-of-sample predictability results for every possible sample split. In this paper we plot the (bootstrap-determined) p-value of the MSPE-adj statistic, as well as the out-of-sample R-squared, against each possible sample split date $\tau \in [\text{int}(.05T),\ T - \text{int}(.05T)]$, where T is the total number of observations.

The second approach is to calculate the maximum and the mean of the set of the t-statistics associated with the MSPE-adj for each possible sample split. We let $\mathbb{t}(\text{MSPE} - \text{adj})_\tau$ denote the t-statistic associated with the MSPE-adj, and we take both the maximum and the mean of that term calculated across each possible sample split date τ. We determine the null distribution of the mean and the maximum statistics via a bootstrap procedure that imposes the null hypothesis of no predictability. In this second approach we distill the complete set of $\mathbb{t}(\text{MSPE} - \text{adj})_\tau$ values for each $\tau \in [\text{int}(.05T),\ T - \text{int}(.05T)]$ into two numbers: the mean and the maximum of the set. In this way we produce two tests of out-of-sample predictability—one each based on the mean and the maximum—that are *invariant* to the choice of a sample split date. We also provide Monte Carlo evidence that our bootstrap approach to inference is valid.

Each of these proposed approaches has advantages. The graphical approach transmits all the relevant information on out-of-sample performance, and the human eye can absorb this information quickly. The second approach yields a general test of out-of-sample predicability that has the advantage of being easily compared to the test of in-sample predictability.

We apply each approach to a comprehensive set of 21 equity premium predictors. We find occasionally impressive out-of-sample predictability for most of the traditional variables. That is to say, many investors making decisions in real time could have benefitted from forecasts given by traditional predictors on an out-of-sample basis. We find that results from the in-sample test of predictability agree, by and large, with our two proposed out-of-sample tests—which are invariant to the sample split date. Finally, we extend these results by demonstrating how to construct out-of-sample tests of predictability that are not only sample split invariant but also robust to data mining.

The most important conclusions to be drawn from this work can be summarized as follows. On the one hand, there are widely varying results reported for out-of-sample predictability tests that differ only in the chosen sample split date. On the other hand, there are but minor disagreements between in-sample predictability test results and results from our proposed (mean and maximum) sample split–invariant out-of-sample predictability tests. All three of these tests tell the same story about predictability: when split-invariant tests are used, the equity premium is well forecast by only few traditional predictors on an out-of-sample basis.

Acknowledgments

We would like to thank Ronald W. Anderson, Rene Garcia, Benjamin Golez, and Abraham Lioui as well as two anonymous referees for detailed comments and suggestions. Remaining errors are ours.

References

Amihud, Y., Hurvich, C., 2004. Predictive regressions: a reduced-bias estimation method. J. Financ. Quant. Anal. 39, 813–841.

Ashley, R., Granger, C.W.J., Schmalensee, R., 1980. Advertising and aggregate consumption: an analysis of causality. Econometrica 48 (5), 1149–1167.

Beran, R., 1988. Prepivoting test statistics: a bootstrap view of asymptotic refinements. J. Am. Stat. Assoc. 83, 687–697.

Binsbergen, J.H.v., Hueskes, W., Koijen, R.S.J., Vrugt, E.B., 2011. Equity Yields. NBER Working Paper 17416.

Britten-Jones, M., 1999. The sampling error in estimates of mean-variance efficient portfolio weights. J. Finance 54 (2), 655–671.

Campbell, J.Y., Thompson, S.B., 2008. Predicting excess stock returns out of sample: can anything beat the historical average? Rev. Financ. Stud. 21, 1509–1531.

Clark, T.E., McCracken, M.W., 2001. Tests of equal forecast accuracy and encompassing for nested models. J Econom 105, 85–110.

Clark, T.E., McCracken, M.W., 2012. Reality checks and comparisons of nested predictive models.. J. Bus. Econ. Statis. 30 (1), 53–66.

Clark, T.E., West, K.D., 2007. Approximately normal tests for equal predictive accuracy in nested models. J. Econ., Elsevier 127 (1), 291–311.

Cochrane, J.H., 2011. Understanding policy in the great recession: some unpleasant fiscal arithmetic. Eur. Econ. Rev. 55, 2–30.

DeMiguel, V., Garlappi, L., Uppal, R., 2009. Optimal versus naive diversification: how inefficient is the 1/n portfolio strategy? Rev. Financ. Stud. 22 (5), 1915–1953.

Foster, F.D., Smith, T., Whaley, R.E., 1997. Assessing goodness of fit of asset pricing models: the distribution of the maximal r2. J. Finance 52 (2), 591–607.

Golez, B., 2014. Expected returns and dividend growth rates implied by derivative markets. Rev. Financ. Stud. 27 (3), 790–822.

Goyal, A., Welch, I., 2003. Predicting the equity premium with dividend ratios. Manage. Sci. 49, 639–654.

Hall, P., 1992. The Bootstrap and Edgeworth Expansion. Springer-Verlag, New York.

Hansen, P.R., Timmermann, A., 2012. Choice of Sample Split in Out-of-Sample Forecast Evaluation. EUI Working Paper ECO 2012/10

Hubrich, K., West, K.D., 2010. Forecast evaluation of small nested model sets. J. Appl. Econ. 25 (4), 574–594.

Inoue, A., Kilian, L., 2004. In-sample or out-of-sample tests of predictability? Which one should we use? Econ. Rev. 23, 371–402.

Kilian, L., 1999. Exchange rates and monetary fundamentals: what do we learn from long-horizon regressions? J. Appl. Econ. 14, 491–510.

Kirby, C., Ostdiek, B., 2012. Its all in the timing: simple active portfolio strategies that outperform naive diversification. J. Financ. Quant. Anal.. Accepted Manuscript, Published online: 20 January 2012

Kolev, G.I., 2008. Forecasting aggregate stock returns using the number of initial public offerings as a predictor. Econ. Bull. 7 (13), 1–8.

Kostakis, A., Panigirtzoglou, N., Skiadopoulos, G., 2011. Market timing with option-implied distributions: a forward-looking approach. Managem. Sci.. Published online before print May 16, 2011

Lee, C., Shleifer, A., Thaler, R.H., 1991. Investor sentiment and the closed-end fund puzzle. J. Finance 46, 75–109.

Lo, A.W., MacKinlay, A.C., 1990. Data-snooping biases in tests of financial asset pricing models. Rev. Financ. Stud. 3 (3), 431–467.

MacKinnon, J.G., 2009. Bootstrap Hypothesis Testing. In: Belsley, D.A., Erricos, J.K. (Eds.), Handbook of Computational Econometrics, pp. 183–213.

Mankiw, G., Shapiro, M., 1986. Do we reject too often? small-sample properties of tests of rational expectations models. Econ. Lett. 20, 139–145.

Mark, N.C., 1995. Exchange rates and fundamentals: evidence on long-horizon predictability. Am. Econ. Rev. 85 (1), 201–218.

Neal, R., Wheatley, S., 1998. Do measures of investor sentiment predict stock returns? J. Financ. Quant. Anal. 34, 523–547.

Neely, C.J., Rapach, D.E., Tu, J., Zhou, G., 2014. Forecasting the equity risk premium: the role of technical indicators. Manage. Sci. 60, 1772–1791.

Nelson, C.R., Kim, M., 1993. Predictable stock returns: the role of small-sample bias. J. Finance 48, 641–661.

Polk, C., Thompson, S., Vuolteenaho, T., 2006. Cross-sectional forecasts of the equity premium. J. Financ. Econ. 81, 101–141.

Rapach, D.E., Strauss, J.K., Zhou, G., 2010. Out-of-sample equity premium prediction: combination forecasts and links to the real economy. Rev. Financ. Stud. 23 (2), 821–862.

Rapach, D.E., Wohar, M.E., 2006. In-sample vs. out-of-sample tests of stock return predictability in the context of data mining. J. Emp. Finance, Elsevier 13 (March (2)), 231–247.

Rossi, B., Inoue, A., 2012. Out-of-sample forecast tests robust to the choice of window size. J. Bus. Econ. Statis. 30 (3), 432–453.

Roy, S.N., 1953. On a heuristic method of test construction and its use in multivariate analysis. Ann. Math. Statis. 24, 220–238.

Stambaugh, R.F., 1986. Bias in Regression with Lagged Stochastic Regressors. CRSP Working Paper No. 156. University of Chicago.

Stambaugh, R.F., 1999. Predictive regression. J. Financ. Econ. 54, 375–421.

Welch, I., Goyal, A., 2008. A comprehensive look at the empirical performance of equity premium prediction. Rev. Financ. Stud. 21, 1455–1508.

West, K.D., 2006. Forecast Evaluation. In: Graham, E., Granger, C.W.J., Timmermann, A. (Eds.), Handbook of Economic Forecasting, Vol. 1, pp. 99–134.

Zweig, M.E., 1973. An investor expectations stock price predictive model using closed-end fund premiums. J. Finance 28, 67–87.