Small-sample tests for stock return predictability with possibly non-stationary regressors and GARCH-type effects

Sermin Gungor* Richard Luger†
University of Western Ontario Laval University

August 31, 2018

*Department of Economics, University of Western Ontario, London, Ontario N6A 5C2, Canada. *E-mail address:* sgungor@uwo.ca

†Corresponding author. Department of Finance, Insurance and Real Estate, Laval University, Quebec City, Quebec G1V 0A6, Canada. *E-mail address:* richard.luger@fsa.ulaval.ca

Small-sample tests for stock return predictability with possibly non-stationary regressors and GARCH-type effects

Abstract: We develop a simulation-based procedure to test for stock return predictability with multiple regressors. The process governing the regressors is left completely free and the test procedure remains valid in small samples even in the presence of non-normalities and GARCH-type effects in the stock returns. The usefulness of the new procedure is demonstrated in a simulation study and by examining the ability of a group of financial variables to predict excess stock returns. We find some evidence of predictability during the period 1948–2014, driven entirely by the term spread. This empirical evidence, however, is much weaker over subsamples.

JEL classification: C12; C32; G14

Keywords: Stock returns; Predictive regression; Multiple predictors; Unit roots; Conditional heteroskedasticity; Robust inference

1 Introduction

A long-standing question in finance is whether stock returns can be predicted by economic and financial variables. The null hypothesis of no predictability is typically examined in the context of an ordinary least squares (OLS) regression of stock returns onto the lagged value of the predictor variable under study. A common finding of such predictive regressions is that the t-statistic often appears significant when compared to the conventional critical values for the t-test. In this case, a researcher might conclude that the financial variable in question has the ability to predict stock returns.

This inference relies on traditional asymptotic theory, which implies that the t-statistic follows the standard normal distribution in large samples. Yet the large-sample theory provides a poor approximation to the finite-sample distribution of the t-statistic when there is feedback from returns to future values of the regressor and the regressor variable itself is persistent over time (Mankiw and Shapiro, 1986; Stambaugh, 1999). The problem in this case is that the OLS estimator is biased and the t-test procedure rejects the null hypothesis much too often even in fairly large samples. The most prominent financial variables explored in the stock return predictability literature include the dividend-price ratio, the earnings-price ratio, the book-to-market ratio, and various interest rates and interest rate spreads. Given the empirical evidence of feedback and the highly persistent nature of these variables, one can seriously doubt any statistical evidence suggesting their predictive ability based on the conventional t-test.

A number of econometric solutions have been proposed to address the inference issues with predictive regressions. These include procedures based on local-to-unity asymptotics

that provide better approximations to the sampling distribution of the t-statistic when the predictor is nearly integrated (Campbell and Yogo, 2006; Cavanagh et al., 1995; Torous et al., 2004). Another strand of the predictive regression literature has proposed procedures that attempt to estimate and correct the bias of the OLS estimator (Amihud and Hurvich, 2004; Amihud et al., 2009; Lewellen, 2004; Polk et al., 2006; Stambaugh, 1999). What is common to all of these approaches is that they depend on a very specific model for the regressor (*i.e.*, a linear autoregressive model) and their behaviour under departures from that assumption is an open question.

In sharp contrast, the sign and signed rank tests of Campbell and Dufour (1997) are exact (*i.e.*, level-correct) without any modelling assumptions whatsoever for the regressor variable.[1] These Lagrange multiplier-type tests are far more general than most competing procedures based on autoregressive and local-to-unity assumptions. For example, they allow the regressor process to be subject to structural breaks, time-varying parameters, or any other unmodelled non-linearities – all of which may give the appearance of unit-root behaviour. Furthermore, the sign and signed rank tests do not impose any parametric assumptions on the distribution of stock return innovations. This setup allows for non-normalities and conditional heteroskedasticity (*e.g.*, GARCH or stochastic volatility) effects in the stock returns. It is well known that financial asset returns are typically characterized by heavy tails in both their conditional and unconditional distributions, and by time-varying conditional volatility (Cont, 2001). In stock return prediction tests, these stylized facts are a clear and present motivation for the use of sign and signed rank tests. Indeed, results from classical finite-

[1]The sign tests provide an assessment of whether the *median* of the response variable is predictable. Upon a further symmetry assumption, the signed rank tests yield a test of *mean* predictability.

sample non-parametric statistics show that such tests are the *only* tests which yield valid inference when one wishes to remain completely agnostic about distribution heterogeneities (Lehmann and Stein, 1949). Furthermore, the simulation results in Campbell and Dufour (1997) show that their non-parametric tests can be more powerful than the size-corrected t-test, especially in heavy-tailed settings.

A practical limitation of the sign and signed rank tests, however, is that they are developed for the single-predictor case only. In this paper, we extend the ideas of Campbell and Dufour (1997) to obtain small-sample tests for stock return predictability in the presence of multiple predictors.[2] The null hypothesis of no predictability implies that the innovations to the dependent variable should be orthogonal to *all* past information. This means that the dependent variable should not be predictable using any lagged regressors. The problem then consists of combining the predictability tests for each considered regressor in such a way that controls the overall significance level of the procedure.

Westfall and Young (1993) explain in great detail how bootstrap methods can be used to solve the multiple testing problem that occurs when considering a set of null hypotheses simultaneously. In this spirit, we propose a simulation-based procedure for controlling the joint significance of stock return predictability tests with multiple regressors. We achieve this by exploiting the technique of Monte Carlo tests (Barnard, 1963; Birnbaum, 1974; Dwass, 1957) to obtain provably exact randomized analogues of the Campbell and Dufour (1997) tests. See Dufour and Khalaf (2001) for a survey of Monte Carlo test techniques.

Observe that the problems of the single-predictor setting are compounded by the presence

[2]Liu and Maynard (2007) extend the Campbell and Dufour (1997) single-predictor tests to a long-horizon setting.

of multiple regressors, since there can be feedback from the return innovations to future values of all the regressors and each of these regressors is potentially highly persistent. So not surprisingly, the standard Wald test suffers from the same over-rejection problem as the t-statistic in the single-predictor model. Amihud et al. (2009) propose a multi-predictor augmented regression method (mARM) to correct the bias of the Wald test. They show that estimating and correcting the bias yields a Wald test statistic with size closer to the nominal level than "plain vanilla" OLS and bootstrapping. The mARM approach assumes that the predictors follow a vector autoregressive (VAR) model, which is both Gaussian and stationary. Under those strict stationarity conditions, the Amihud et al. (2009) method works well but its performance deteriorates as the persistence of the regressors approaches the non-stationary boundary.

Other methods that have been proposed for multiple predictor testing include the extended instrumental variables (IVX) procedure of Kostakis et al. (2015), the subsampling approach of Wolf (2000), the jackknife of Zhu (2013), and the robust bootstrap and subsampling methods of Camponovo et al. (2012). Just like the mARM of Amihud et al. (2009), all of these methods heavily depend on the assumption that the predictors follow a linear VAR model. On the contrary, the methods we propose cover a much wider class of applications by leaving completely free the joint process governing the regressors. In fact, the developed Monte Carlo test procedure inherits all the properties of the original Campbell and Dufour (1997) distribution-free tests (*e.g.*, robustness to non-normalities and GARCH-type effects in the stock returns) in addition to being free of modelling assumptions on the regressors. What makes this possible is that: (i) the basic building block of the Campbell-Dufour test statistics is a sign function applied to products involving the outcome variable and the re-

gressors; and (ii) the null distribution of these signs does not depend on the regressors. Our simulation experiments further reveal that the proposed non-parametric Monte Carlo tests can be more powerful than the size-corrected Wald, mARM, and IVX tests.

Our final contribution is empirical. We apply the developed procedure to test the predictability of the excess returns on the S&P 500 value-weighted index using six widely used predictors: the dividend-price ratio, the earnings-price ratio, the book-to-market ratio, the default yield, the term spread, and the short rate. We use monthly data for the 67-year sample period 1948–2014. In addition to the full sample, we also perform the analysis over fixed 10-year and 20-year subsamples and 20-year rolling-window subsamples. The standard Wald test overwhelmingly rejects the joint null hypothesis of no predictability, but this evidence is questionable given the highly persistent and endogenous nature of the employed predictors. With the new test procedure we find more trustworthy, albeit weak, evidence of stock return predictability in the full-sample period. Tests of the marginal significance reveal that among the six regressors it is only the term spread that has predictive ability for monthly excess stock returns. The takeaway message is that while the new joint tests reveal some evidence of stock return predictability, this evidence is weak and entirely driven by the term spread. The predictability evidence turns out to be even weaker over the monthly subsamples. Our results suggest that test power depends more on the span of the data rather than the number of observations.

The current paper is organized as follows. Section 2 establishes the statistical framework and Section 3 develops the small-sample predictability tests based on signs and ranks. We begin by assuming provisionally that the intercept value in the predictive regression model is known. In this context, we show some key results about the finite-sample distribution

of test statistics that pinpoint the predictive ability of individual regressors. We also show how to combine these marginal statistics to obtain a test of the joint null hypothesis of no predictability. Then we drop the assumption of a known intercept. For this case, we adopt a two-stage maximized Monte Carlo method (Dufour, 2006) to deal with the nuisance intercept parameter. Section 4 presents the results of simulation experiments in which the performance of the new test procedure is compared to the standard Wald test, the mARM-based Wald test of Amihud et al. (2009), and the IVX-estimated "persistence-robust" Wald test of Kostakis et al. (2015). Section 5 presents the empirical application to U.S. equity data and Section 6 offers some concluding remarks. The Appendix contains the proofs of the formal propositions.

2 Predictive regression model

Consider a stock return (or excess stock return) r_t at time t and a finite $K \times 1$ vector of variables $\boldsymbol{x}_{t-1} = (x_{1,t-1}, ..., x_{K,t-1})'$ observed at $t-1$ that could have the ability to predict r_t. The complete model specification involves the random variables $r_1, ..., r_T$, $\boldsymbol{x}_0, \boldsymbol{x}_1, ...\boldsymbol{x}_{T-1}$, and the corresponding information vectors $\mathcal{I}_t = (\boldsymbol{x}_0', \boldsymbol{x}_1', ..., \boldsymbol{x}_t', r_1, ..., r_t)'$, defined for $t = 0, 1, ..., T-1$, with the convention that $\mathcal{I}_0 = \boldsymbol{x}_0$. Specifically, we consider the predictive regression model

$$r_t = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_{t-1} + \varepsilon_t, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_K)'$ is $K \times 1$ vector comprising the parameters of interest. The null hypothesis of no predictability is formally stated as

$$H_0 : \boldsymbol{\beta} = \mathbf{0}$$

which is to be tested against the two-sided alternative $\boldsymbol{\beta} \neq \mathbf{0}$, the right-sided alternative $\boldsymbol{\beta} > \mathbf{0}$, or the left-sided alternative $\boldsymbol{\beta} < \mathbf{0}$. Observe that H_0 is a joint hypothesis, so a rejection signifies that one or more variables in $\boldsymbol{x}_{t-1}$ have the ability to predict returns. For any number K of predictors, our approach will keep under control the probability of rejecting at least one $H_{0,i} : \beta_i = 0$, $i = 1, ..., K$, when the complete H_0 is true.

For one group of tests, we merely assume that the distribution of ε_t in (1) has a conditional median equal to zero such that

$$\Pr(\varepsilon_t > 0 \,|\, \mathcal{I}_{t-1}) = \Pr(\varepsilon_t < 0 \,|\, \mathcal{I}_{t-1}) = 1/2 \tag{2}$$

and we also develop tests under the stronger assumption:

$$\varepsilon_t \text{ is continuously and symmetrically distributed about zero, given } \mathcal{I}_{t-1}. \tag{3}$$

The tests derived under (2) should thus be interpreted as tests of whether the conditional median of r_t is predictable using $\boldsymbol{x}_{t-1}$. So if we let $Q_\tau(r_t \,|\, \mathcal{I}_{t-1})$ denote the τth conditional quantile of r_t, then the first group of tests provides an assessment of $\boldsymbol{\beta}_{0.5} = \mathbf{0}$ in the context

of the predictive quantile regression

$$Q_\tau(r_t \,|\, \mathcal{I}_{t-1}) = \beta_{0,\tau} + \boldsymbol{\beta}_\tau' \boldsymbol{x}_{t-1},$$

when $\tau = 0.5$ (the median).[3] It is easy to see that (3) implies (2), but not *vice versa*. Observe also that when ε_t has a well-defined first moment, then, under H_0 and (3), the conditional mean and median (point of symmetry) of r_t both equal β_0 (Randles and Wolfe, 1979, Remark 1.3.11). In this case, the tests that rest on (3) yield an assessment of H_0 in the context of

$$E(r_t \,|\, \mathcal{I}_{t-1}) = \beta_0 + \boldsymbol{\beta}' \boldsymbol{x}_{t-1},$$

which corresponds to the usual predictive mean regression. Note that OLS-based procedures can only be justified by assuming that ε_t has well-defined first and second moments, while here no such moment assumptions are needed.

In addition to heavy tails and other non-normalities, this setup allows for GARCH-type effects of unknown form in the conditional distribution of returns. For example, a wide class of GARCH and stochastic volatility models take the form $\varepsilon_t = \sigma_t \eta_t$, where the return innovations $\{\eta_t\}$ are independent and identically distributed (i.i.d.) according to a symmetric distribution (*e.g.*, normal or Student-t). Such specifications are fully compatible with (3) as long as the random variable $\sigma_t^2 > 0$ capturing conditional heteroskedasticity is a measurable function of $\mathcal{I}_{t-1}$. Of course, a far wider class with asymmetric innovations can

[3] Cenesizoglu and Timmermann (2008), Maynard et al. (2010), and Lee (2016) consider the more general case of predictive quantile regressions defined for any quantile level $\tau \in (0, 1)$. The quantile predictability methods used by these authors rely on asymptotic inference under more restrictive parametric assumptions for the predictor process. Moreover, those methods can only handle a single predictor whereas our median predictability tests allow for multiple predictors.

be entertained under (2). Here the process governing the dynamics of σ_t^2 over time need not even be stationary, which allows for integrated GARCH-type effects; see Coudin and Dufour (2009) for more discussion on this point.

To discuss some of the issues with testing H_0, it is instructive to complement (1) with a VAR model for the predictor variables so the entire system becomes

$$\begin{aligned} r_t &= \beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_{t-1} + \sigma_t \eta_t, \\ \boldsymbol{x}_t &= \boldsymbol{\mu} + \boldsymbol{\Phi}\boldsymbol{x}_{t-1} + \boldsymbol{v}_t, \end{aligned} \tag{4}$$

where the contemporaneous covariance matrix of $\boldsymbol{\epsilon}_t = (\eta_t, \boldsymbol{v}_t')'$ is given by

$$\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \begin{pmatrix} 1 & \boldsymbol{\sigma}_{\eta \boldsymbol{v}}' \\ \boldsymbol{\sigma}_{\eta \boldsymbol{v}} & \boldsymbol{\Sigma}_{\boldsymbol{v}} \end{pmatrix}.$$

If we define $\boldsymbol{Y} = (r_1, ..., r_T)'$ and $\boldsymbol{X} = [\boldsymbol{\iota}, \boldsymbol{X}_1, ..., \boldsymbol{X}_K]$, where $\boldsymbol{\iota}$ is a column vector of ones and $\boldsymbol{X}_i = (x_{i,0}, ..., x_{i,T-1})'$, $i = 1, ..., K$, then the predictive regression in (4) can be written as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{e}$. Here the parameters are stacked in $\boldsymbol{\gamma} = [\beta_0, \boldsymbol{\beta}']'$. The OLS estimate is $\hat{\boldsymbol{\gamma}} = [\hat{\beta}_0, \hat{\boldsymbol{\beta}}']' = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ and the usual Wald statistic for testing H_0 is computed as

$$\text{Wald} = \hat{\boldsymbol{\beta}}' \big(\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) \big)^{-1} \hat{\boldsymbol{\beta}},$$

where $\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})$ is read from the estimated covariance matrix $\hat{s}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $\hat{s}^2$ is the estimated residual variance. Note that the computation of the Wald statistic does not require any information from the VAR part of (4), just like our approach.

The standard practice is to compare the computed value of the Wald test statistic to the

critical values of its asymptotic $\chi^2(K)$ distribution. This procedure, however, may reject the null of no predictability much too often, even with fairly large samples. The problem largely originates from $\boldsymbol{\sigma}_{\eta\boldsymbol{v}} \neq \mathbf{0}$, in which case there is feedback from return innovations that may affect future values of the regressors, even though the innovations and the regressors are contemporaneously uncorrelated. In this case, the OLS estimator is biased and the sampling distribution of the Wald statistic differs from the $\chi^2(K)$ distribution. The overrejection problem is further exacerbated when the regressors are persistent, *i.e.*, as the eigenvalues of $\boldsymbol{\Phi}$ become large in absolute value. This problem is well known especially when $K = 1$ (Mankiw and Shapiro, 1986; Stambaugh, 1999) in which case the square of the standard t-statistic corresponds to the Wald statistic.

Amihud et al. (2009) focus on the small-sample bias of the OLS estimates $\hat{\boldsymbol{\beta}}$ in the multiple-predictor regression context. Their multi-predictor augmented (by a VAR model) regression method (mARM) is an iterative procedure which yields a reduced bias estimator of $\boldsymbol{\beta}$ in (4). The mARM-based estimator is used to form a bias-corrected Wald statistic and this statistic is then compared to the usual asymptotic $\chi^2(K)$ distribution. The mARM approach to predictability testing is developed in the specific context of the linear system in (4) assuming $\boldsymbol{\epsilon}_t \sim$ i.i.d. $N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ with σ_t constant over time (conditional homoskedasticity), and that all the eigenvalues of the VAR persistence matrix $\boldsymbol{\Phi}$ are less than 1 in absolute value (stationarity of the regressors).

Another prominent approach to multiple predictability testing that has been developed in the context of a system like (4) is the IVX procedure of Kostakis et al. (2015), which promises robustness to the regressors' degree of persistence. The idea is to construct instrumental variables (IVs) whose persistence is explicitly controlled according to $\boldsymbol{z}_t = \boldsymbol{\Phi}_{Tz}\boldsymbol{z}_{t-1} + \Delta\boldsymbol{x}_t$,

where by choice of $\boldsymbol{\Phi}_{Tz}$ the persistence of $\boldsymbol{z}_t$ is between that of $\boldsymbol{x}_t$ (levels data) and $\Delta\boldsymbol{x}_t$ (first differences); Kostakis et al. (2015) provide guidance for the choice of $\boldsymbol{\Phi}_{Tz}$. In this way, the problems arising from the unknown $\boldsymbol{\Phi}$ matrix of the original regressors in (4) can be avoided. With the constructed IVs, one then performs a standard IV estimation of $\boldsymbol{\beta}$. The resulting estimate along with a Newey-West estimate of the long-run covariance matrix yields the IVX-estimated Wald statistic, which follows the $\chi^2(K)$ distribution asymptotically. Of course, there are several regularity assumptions needed for this result to hold.

What distinguishes our approach is that: (i) besides the minimal assumptions in (2) and (3), there are no restrictions on the distribution of ε_t; (ii) conditional heteroskedasticity of unknown form is allowed; (iii) there are no restrictions on the data-generating process for $\boldsymbol{x}_t$; and (iv) the probability of rejecting the null when it is true (a Type I error) is kept under control no matter the sample size.

3 Small-sample predictability tests

Our approach is based on sign and signed rank statistics defined for each considered regressor. Let $s[z] = 1$ when $z \geq 0$, and $s[z] = 0$ when $z < 0$, and consider a non-parametric analogue of the t-statistic given by the following sign statistic:

$$S_i(b) = \sum_{t=1}^{T} s\big[(r_t - b)g_{i,t-1}\big], \tag{5}$$

where $\{g_{i,t}\}_{t=0}^{T-1} = \{g_{i,t}(\mathcal{I}_t)\}_{t=0}^{T-1}$ is a sequence of measurable functions of the information vector $\mathcal{I}_t$. We specify $g_{i,t} = g_t(x_{i,0}, ..., x_{i,t})$ so that $S_i(b)$ pinpoints the predictive ability of

x_i, $i = 1, ..., K$. The sign statistic in (5) belongs to a broader class of linear signed rank statistics defined by

$$SR_i(b) = \sum_{t=1}^{T} s\big[(r_t - b)g_{i,t-1}\big]\varphi\big(R_t^+(b)\big), \tag{6}$$

where $R_t^+(b)$ is the rank of $|r_t - b|$ when $|r_1 - b|, ..., |r_T - b|$ are placed in ascending order. Observe that $R_1^+(b), ..., R_T^+(b)$ is an arrangement of the first T positive integers $1, 2, ..., T$. A general class of statistics is then defined from the set of scores $\varphi(t)$, $t = 1, ..., T$, such that $0 \leq \varphi(1) \leq ... \leq \varphi(T)$ with $\varphi(T) > 0$. The sign statistic (5) is obtained from the constant score function $\varphi(t) = 1$. Another familiar member of this class is the following Wilcoxon signed rank statistic:

$$W_i(b) = \sum_{t=1}^{T} s\big[(r_t - b)g_{i,t-1}\big]R_t^+(b), \tag{7}$$

which is obtained with $\varphi(t) = t$, for $t = 1, ..., T$.

The motivation for using sign-based inference methods comes from the Lehmann and Stein (1949) impossibility theorem. This result shows that a test with level α given a finite number of observations in the presence of heteroskedasticity of unknown form must be a sign test, or, more precisely, its level must be equal to α conditional on the absolute values of the observations. For more on this result, see Pratt and Gibbons (1981, p. 218) and Dufour (2003).

When β_0 is known, it is natural to complete the definitions of the test statistics in (5) and (6) by setting $b = \beta_0$ and $g_{i,t} = x_{i,t}$. This way the conditional median of $(r_t - \beta_0)x_{i,t-1}$ depends on $\beta_i x_{i,t-1}^2$. So, under the alternative hypothesis that $\beta_i \neq 0$, the power of $S_i(\beta_0)$ and $SR_i(\beta_0)$ will increase with the magnitude of $|\beta_i|$. For the more realistic case of an unknown β_0 (developed in §3.2), we use a two-stage procedure which proceeds by: (i) building a

confidence interval for β_0; and (ii) maximizing the p-value of the test statistic over this confidence interval. When inference proceeds in this fashion, a straightforward extension of the arguments in Campbell and Dufour (1997) suggests that better power can be achieved by setting

$$g_{i,t} = x_{it} - \hat{m}_{it}, \text{ for } i = 1, ..., K, \ t = 0, ..., T-1,$$

where $\hat{m}_{it} = \text{median}\{x_{i0}, ..., x_{it}\}$ only depends on observations up to time t (so that g_{it} is a function of $\mathcal{I}_t$).

3.1 Inference when β_0 is known

Suppose for a moment that the value of β_0 in model (1) is known. The following proposition characterizes the exact distribution of $S_i(\beta_0)$ and $SR_i(\beta_0)$ in this case. Here we let $\tilde{\varepsilon}_1, ..., \tilde{\varepsilon}_T$ be independent median-zero random variables such that $\Pr(\tilde{\varepsilon}_t > 0) = \Pr(\tilde{\varepsilon}_t < 0) = 1/2$, for $t = 1, ..., T$; and $B_1, ..., B_T$ are independent Bernoulli variables such that $\Pr(B_t = 1) = \Pr(B_t = 0) = 1/2$, for $t = 1, ..., T$. From here on, the symbol "$\stackrel{d}{=}$" is used to denote an equality in distribution.

Proposition 1. *Suppose model (1) holds and let* $\{g_{i,t}\}_{t=0}^{T-1} = \{g_{i,t}(\mathcal{I}_t)\}_{t=0}^{T-1}$, $i = 1, ..., K$, *be sequences of measurable functions of* $\mathcal{I}_t$ *such that* $\Pr(g_{i,t} = 0) = 0$, *for all* t.

(i) If H_0 *and Assumption (2) are satisfied, then, given* $g_{i,0}, ..., g_{i,T-1}$, *the sign statistic* $S_i(\beta_0)$ *defined by (5) is such that*

$$S_i(\beta_0) \stackrel{d}{=} \sum_{t=1}^{T} s\big[\tilde{\varepsilon}_t g_{i,t-1}\big] \stackrel{d}{=} \sum_{t=1}^{T} B_t,$$

for each $i = 1, ..., K$.

(ii) If H_0 *and Assumption (3) are satisfied, then, given* $g_{i,0}, ..., g_{i,T-1}$ *and* $|r_1 - \beta_0|, ..., |r_T - \beta_0|$*, the signed rank statistic* $SR_i(\beta_0)$ *defined by (6) is such that*

$$SR_i(\beta_0) \stackrel{d}{=} \sum_{t=1}^{T} s\big[\tilde{\varepsilon}_t g_{i,t-1}\big]\varphi\big(R_t^+(\beta_0)\big) \stackrel{d}{=} \sum_{t=1}^{T} B_t \varphi(t),$$

for each $i = 1, ..., K$

This proposition shows that the null distribution of $S_i(\beta_0)$ is independent of $g_{i,0}, ..., g_{i,T-1}$, for $i = 1, ..., K$, while the null distribution of $SR_i(\beta_0)$ is independent of $g_{i,0}, ..., g_{i,T-1}$ and $|r_1 - \beta_0|, ..., |r_T - \beta_0|$, for $i = 1, ..., K$. This is the key property that allows us to construct *conditional* tests that account for the dependence among a joint collection of test statistics, where the individual statistics comprising the collection are defined for $i = 1, ..., K$.

Indeed, part (i) of Proposition 1 shows that the statistic $S_i(\beta_0)$ follows a binomial distribution $\text{Bi}(T, 1/2)$ under the null hypothesis. As T grows large, the binomial distribution of $S_i(\beta_0)$ can be approximated by a normal with mean $T/2$ and variance $T/4$, *i.e.*,

$$S_i^*(\beta_0) = \frac{S_i(\beta_0) - T/2}{\sqrt{T/4}} \to N(0, 1) \text{ as } T \to \infty.$$

More generally, standard results found in Randles and Wolfe (1979, §10.2) show that under the conditions of Proposition 1, the standardized linear signed rank statistic

$$SR_i^*(\beta_0) = \left[SR_i(\beta_0) - \frac{1}{2}\sum_{t=1}^{T}\varphi(t)\right] \Bigg/ \sqrt{\frac{1}{4}\sum_{t=1}^{T}\varphi^2(t)}$$

has a limiting standard normal distribution. If we let $\Phi(\cdot)$ denote the standard normal cumulative distribution function, the associated p-values can be defined as: $p_i^S(\beta_0) = 2\big(1 - \Phi(|S_i^*(\beta_0)|)\big)$ and $p_i^{SR}(\beta_0) = 2\big(1 - \Phi(|SR_i^*(\beta_0)|)\big)$ for a two-sided alternative; $p_i^S(\beta_0) = 1 - \Phi(S_i^*(\beta_0))$ and $p_i^{SR}(\beta_0) = 1 - \Phi(SR_i^*(\beta_0))$ for a right-sided alternative; $p_i^S(\beta_0) = \Phi(S_i^*(\beta_0))$ and $p_i^{SR}(\beta_0) = \Phi(SR_i^*(\beta_0))$ for a left-sided alternative. We carry on assuming that H_0 is tested against a two-sided alternative. (For left- and right-sided alternatives, simply use the appropriate p-value as defined above.)

Test statistics like $S_i(\beta_0)$ in (5) and $SR_i(\beta_0)$ in (6) will have power to detect the predictive ability of x_i. In order to obtain power against all x_i's, we consider two methods of combining the marginal p-values associated with each individual test statistic. The first method rejects H_0 when at least one of the individual p-values is sufficiently small. Specifically, if we let $\mathcal{S}$ refer to either the S or SR statistic and define

$$p_{min}^{\mathcal{S}}(\beta_0) = \min\left\{p_1^{\mathcal{S}}(\beta_0), ..., p_K^{\mathcal{S}}(\beta_0)\right\} \text{ and } \mathcal{S}_{min}(\beta_0) = 1 - p_{min}^{\mathcal{S}}(\beta_0), \tag{8}$$

then we reject H_0 when $p_{min}^{\mathcal{S}}(\beta_0)$ is small, or, equivalently, when $\mathcal{S}_{min}(\beta_0)$ is large. The intuition here is that the null hypothesis of no predictability should be rejected if at least one of the individual p-values is significant. This method of combining tests was suggested by Tippett (1931) and Wilkinson (1951).

The second combination method we consider – originally suggested by Fisher (1932) and Pearson (1933) – is based on the product of the individual p-values:

$$p_{\times}^{\mathcal{S}}(\beta_0) = \prod_{i=1}^{K} p_i^{\mathcal{S}}(\beta_0) \text{ and } \mathcal{S}_{\times}(\beta_0) = 1 - p_{\times}^{\mathcal{S}}(\beta_0), \tag{9}$$

which may provide more information about departures from H_0 compared to using only the minimum p-value. Indeed, the product of several p-values may indicate a rejection of the joint null hypothesis even though the individual p-values may not be small enough to be significant on their own. For further discussion and recent examples of the test combination technique, see Folks (1984), Westfall and Young (1993), Dufour et al. (2015) and Gungor and Luger (2015).

The p-values $p_1^{\mathcal{S}}(\beta_0), ..., p_K^{\mathcal{S}}(\beta_0)$ are obviously not statistically independent and may in fact have a very complex dependence structure. Nevertheless, if we choose the individual levels α_i such that $\sum_{i=1}^{K} \alpha_i = \alpha$, then, by Bonferroni's inequality, we have

$$\Pr\left(\bigcup_{i=1}^{K} p_i^{\mathcal{S}}(\beta_0) \leq \alpha_i\right) \leq \alpha,$$

such that the *induced* test, which consists of rejecting H_0 when any of the individual tests rejects, has level α.[4] For example, if we set each individual level at α/K, then the overall probability of committing a Type I error does not exceed α. Such p-value adjustments, however, yield a test lacking in power as K grows; see Savin (1984) for a survey discussion of these issues.

In order to resolve the multiple comparison issue, we propose a Monte Carlo (MC) test procedure based on the combination of the individual p-values (either through the minimum or the product rule). The idea is to treat the combination like any other *pivotal* statistic for the purpose of MC resampling (Barnard, 1963; Birnbaum, 1974; Dwass, 1957). This approach bears resemblance to a double bootstrap scheme (cf. MacKinnon, 2009) which

[4]Here we follow the terminology in Lehmann and Romano (2005, Ch. 3) and say that a test of H_0 has *size* α if Pr(Rejecting $H_0 \,|\, H_0$ true) $= \alpha$, and that it has *level* α if Pr(Rejecting $H_0 \,|\, H_0$ true) $\leq \alpha$.

is typically quite expensive computationally as it requires a second layer of simulations to obtain the p-value of the combined (first-level) bootstrap p-values. Here though we only require a single layer of simulations, since the individual p-values are available analytically. A remarkable feature of the MC test combination procedure is that it remains exact even if the individual p-values based on $\Phi(\cdot)$ may only be approximate.[5] Indeed, the MC procedure implicitly accounts for the fact that the individual p-values may not be exact and yields an overall p-value for the combined statistic which itself is exact.

Let $\widetilde{S}_i(\beta_0) = \sum_{t=1}^{T} s\big[\tilde{\varepsilon}_t g_{i,t-1}\big]$, which is recognized as the artificial sign statistic in part (i) of Proposition 1. This statistic yields $\tilde{p}_i^S(\beta_0) = 2\big(1 - \Phi(|\widetilde{S}_i^*(\beta_0)|)\big)$, where $\widetilde{S}_i^*(\beta_0) = (\widetilde{S}_i(\beta_0) - T/2)/\sqrt{T/4}$. Similarly, let $\widetilde{SR}_i(\beta_0) = \sum_{t=1}^{T} s\big[\tilde{\varepsilon}_t g_{i,t-1}\big]\varphi\big(R_t^+(\beta_0)\big)$ denote the artificial signed rank statistic in part (ii) of Proposition 1, with standardized version $\widetilde{SR}_i^*(\beta_0) = \Big[\widetilde{SR}_i(\beta_0) - \frac{1}{2}\sum_{t=1}^{T}\varphi(t)\Big] \Big/ \sqrt{\frac{1}{4}\sum_{t=1}^{T}\varphi^2(t)}$ and corresponding p-value $\tilde{p}_i^{SR}(\beta_0) = 2\big(1 - \Phi(|\widetilde{SR}_i^*(\beta_0)|)\big)$. The following proposition is the key result for the joint predictability tests.

Proposition 2. *Suppose model (1) holds and let* $\{g_{i,t}\}_{t=0}^{T-1} = \{g_{i,t}(\mathcal{I}_t)\}_{t=0}^{T-1}$, $i = 1, ..., K$, *be sequences of measurable functions of* $\mathcal{I}_t$ *such that* $\Pr(g_{i,t} = 0) = 0$, *for all* t.

(i) If H_0 *and Assumption (2) are satisfied, then, given* $g_{i,0}, ..., g_{i,T-1}$, $i = 1, ..., K$, *the* $S_{min}(\beta_0)$ *and* $S_{\times}(\beta_0)$ *statistics defined as in (8) and (9) are such that*

$$S_{min}(\beta_0) \overset{d}{=} \widetilde{S}_{min}(\beta_0) = 1 - \min\big\{\tilde{p}_1^S(\beta_0), ..., \tilde{p}_K^S(\beta_0)\big\},$$

$$S_{\times}(\beta_0) \overset{d}{=} \widetilde{S}_{\times}(\beta_0) = 1 - \prod_{i=1}^{K} \tilde{p}_i^S(\beta_0).$$

[5] Recall that an exact p-value has a standard uniform distribution.

(ii) If H_0 and Assumption (3) are satisfied, then, given $g_{i,0}, ..., g_{i,T-1}$, $i = 1, ..., K$, and $|r_1 - \beta_0|, ..., |r_t - \beta_0|$, the $SR_{min}(\beta_0)$ and $SR_{\times}(\beta_0)$ statistics defined as in (8) and (9) are such that

$$SR_{min}(\beta_0) \overset{d}{=} \widetilde{SR}_{min}(\beta_0) = 1 - \min\left\{\tilde{p}_1^{SR}(\beta_0), ..., \tilde{p}_K^{SR}(\beta_0)\right\},$$

$$SR_{\times}(\beta_0) \overset{d}{=} \widetilde{SR}_{\times}(\beta_0) = 1 - \prod_{i=1}^{K} \tilde{p}_i^{SR}(\beta_0).$$

Note that the simulated terms $\tilde{\varepsilon}_t$ (used to obtain the building blocks $\widetilde{S}_i(\beta_0)$ and $\widetilde{SR}_i(\beta_0)$, $i = 1, ..., K$) may simply be set as i.i.d. according to *any* continuous median-zero distribution like the standard normal, for example. An important remark is that the same values of $\tilde{\varepsilon}_1, ..., \tilde{\varepsilon}_T$ serve to compute all the simulated statistics. For instance, the same value of $\tilde{\varepsilon}_t$ is used to compute all the time-t values $s\left[\tilde{\varepsilon}_t g_{1,t-1}\right], ..., s\left[\tilde{\varepsilon}_t g_{K,t-1}\right]$ appearing in the definitions of $\widetilde{S}_i(\beta_0)$ and $\widetilde{SR}_i(\beta_0)$, for $i = 1, ..., K$. This requirement is needed to preserve the contemporaneous dependence among the individual statistics.

Let $\mathcal{S}_{\bullet}(\beta_0)$ denote any one of the combined statistics $S_{min}(\beta_0)$, $S_{\times}(\beta_0)$, $SR_{min}(\beta_0)$, or $SR_{\times}(\beta_0)$ featured in Proposition 2. In principle, critical values for the combined statistics could be found from the conditional distribution of $\mathcal{S}_{\bullet}(\beta_0)$ derived from the 2^T equally likely possibilities represented by $\widetilde{\mathcal{S}}_{\bullet}(\beta_0)$. Determination of this distribution from a complete enumeration of all possible realizations is obviously impractical. The MC test technique circumvents this problem while still yielding an exact p-value for $\mathcal{S}_{\bullet}(\beta_0)$.

The MC test proceeds by generating $M - 1$ artificial statistics $\widetilde{\mathcal{S}}_{\bullet 1}(\beta_0), ..., \widetilde{\mathcal{S}}_{\bullet M-1}(\beta_0)$, each one according to Proposition 2. Note that the distribution of these statistics is discrete,

meaning that ties can occur among the resampled values. A test with size α can nevertheless be obtained by applying the following tie-breaking rule (Dufour, 2006). Draw M i.i.d. variates U_m, $m = 1, ..., M$, from the standard uniform distribution $U(0, 1)$, randomly pair the U and $\mathcal{S}_\bullet(\beta_0)$ statistics (actual and artificial), and compute the lexicographic rank of $\big(\mathcal{S}_\bullet(\beta_0), U_M\big)$ according to

$$\tilde{R}_M[\mathcal{S}_\bullet(\beta_0)] = 1 + \sum_{j=1}^{M-1} \mathbb{I}\left[\mathcal{S}_\bullet(\beta_0) > \widetilde{\mathcal{S}}_{\bullet j}(\beta_0)\right] + \sum_{j=1}^{M-1} \mathbb{I}\left[\mathcal{S}_\bullet(\beta_0) = \widetilde{\mathcal{S}}_{\bullet j}(\beta_0)\right] \times \mathbb{I}\big[U_M > U_j\big], \quad (10)$$

where $\mathbb{I}[A]$ is the indicator function of event A.

Upon recognizing that the pairs $\big(\widetilde{\mathcal{S}}_{\bullet 1}(\beta_0), U_1\big), ..., \big(\widetilde{\mathcal{S}}_{\bullet M-1}(\beta_0), U_{M-1}\big), \big(\mathcal{S}_\bullet(\beta_0), U_M\big)$ are *exchangeable* under the conditions of Proposition 2, we then know from Lemma 2.3 in Dufour (2006) that the lexicographic ranks are uniformly distributed over the integers $1, ..., M$; *i.e.*, $\Pr\big(\tilde{R}_M[\mathcal{S}_\bullet(\beta_0)] = j\big) = 1/M$, for $j = 1, ..., M$. So the MC p-value can be defined as

$$\tilde{p}_M\big[\mathcal{S}_\bullet(\beta_0)\big] = \frac{M - \tilde{R}_M[\mathcal{S}_\bullet(\beta_0)] + 1}{M}, \quad (11)$$

where $\tilde{R}_M[\mathcal{S}_\bullet(\beta_0)]$ is the rank of $\big(\mathcal{S}_\bullet(\beta_0), U_M\big)$, computed according to (10). If αM is an integer, then the critical region $\tilde{p}_M\big[\mathcal{S}_\bullet(\beta_0)\big] \leq \alpha$ has exactly size α in the sense that

$$\Pr\Big(\tilde{p}_M\big[\mathcal{S}_\bullet(\beta_0)\big] \leq \alpha\Big) = \alpha, \quad (12)$$

under the conditions of Proposition 2.

3.2 Inference when β_0 is unknown

A straightforward way of dealing with an unknown β_0 is to replace it by the estimate $\check{\beta}_0 = \text{median}\{r_1, ..., r_T\}$, and to base inference on $\tilde{p}_M\left[\mathcal{S}_\bullet(\check{\beta}_0)\right]$. These MC p-values based on the aligned sign and signed rank statistics are quite natural, so we will examine their size and power properties in the simulation study. However, we do not have any theoretical results to justify their use (either in finite samples or asymptotically) and it seems quite doubtful that such a theory is even possible given the generality of our statistical framework. To simplify the notation, we will use S^m_{min}, $S^m_\times$, W^m_{min}, $W^m_\times$ to refer to these plug-in (median-estimate) tests based on (5) and (7).

In order to obtain tests that remain truly exact even when β_0 is unknown, we adopt a two-stage *maximized p-value* approach (Dufour, 2006). The first stage consists of establishing a set of admissible values for the nuisance parameter. Next, the p-value of the test statistic is maximized over this set. The idea of this two-stage approach can be understood by viewing the null hypothesis as a union of point null hypotheses:

$$H_0 : \bigcup_{b \in \mathcal{B}} H_0(b), \tag{13}$$

where $H_0(b) : \boldsymbol{\beta} = \mathbf{0}, \beta_0 = b$. Here $\mathcal{B} \subseteq \mathbb{R}$ denotes a set of admissible values for β_0 that are compatible with H_0. The expression in (13) makes clear that β_0 is a nuisance parameter in the present context, since it is not pinned down to a specific value under H_0. In order to test such a hypothesis, which contains several distributions, we can appeal to a *minimax* argument stated as: "reject the null hypothesis whenever, for all admissible values of the nuisance parameter under the null, the corresponding point null hypothesis is rejected"

(Savin, 1984).

With any of the signed rank test statistics, this would mean maximizing the MC p-value $\tilde{p}_M\big[\mathcal{S}_\bullet(b)\big]$ over $b \in \mathcal{B}$. The rationale is that

$$\sup_{b\in\mathcal{B}} \tilde{p}_M\big[\mathcal{S}_\bullet(b)\big] \leq \alpha \Longrightarrow \tilde{p}_M\big[\mathcal{S}_\bullet(\beta_0)\big] \leq \alpha,$$

where the latter is the MC p-value of the test statistic based on the true parameter value. Moreover, $\Pr\big(\tilde{p}_M[\mathcal{S}_\bullet(b)] \leq \alpha\big) = \alpha$ under $H_0(b)$ and for all $b \in \mathcal{B}$. So if αM is an integer, it then follows that

$$\Pr\left(\sup_{b\in\mathcal{B}} \tilde{p}_M\big[\mathcal{S}_\bullet(b)\big] \leq \alpha\right) \leq \alpha$$

under the conditions of Proposition 2. The decision rule in this case would be to reject H_0 if the maximized p-value is $\leq \alpha$. Otherwise, accept H_0 since there is not sufficient evidence to reject it. Note that this test has *level* α, meaning it is conservative.

Campbell and Dufour (1997), and more recently Beaulieu et al. (2007), suggest replacing the first-stage $\mathcal{B}$ appearing in (13) by an exact confidence interval for β_0 which is valid at least under the null hypothesis. This can be interpreted as plugging in an estimator of the (perhaps unknown) set of admissible β_0-values.[6] Let $CI_{\beta_0}(\alpha_1)$ denote a confidence interval for β_0 with level $1-\alpha_1$, *i.e.*, such that $\Pr\big(\beta_0 \in CI_{\beta_0}(\alpha_1)\big) \geq 1-\alpha_1$ under H_0. In the following proposition, $CI_{\beta_0}(\alpha_1)$ is assumed to be valid in this sense either under H_0 and Assumption (2) for part (i); or under H_0 and Assumption (3) for part (ii). Given the desired level $\alpha = \alpha_1 + \alpha_2$, it is further assumed that the choice of M ensures that $\alpha_2 M$ is an integer.

[6]Note also that this is the main idea of the Bonferroni methods frequently used to deal with nuisance parameters in predictive regressions; see, for example, Cavanagh et al. (1995) and Campbell and Yogo (2006).

Proposition 3. *Suppose model (1) holds and let* $\{g_{i,t}\}_{t=0}^{T-1} = \{g_{i,t}(\mathcal{I}_t)\}_{t=0}^{T-1}$, $i = 1, ..., K$, *be sequences of measurable functions of* $\mathcal{I}_t$ *such that* $\Pr(g_{i,t} = 0) = 0$, *for all* t.

(i) If H_0 *and Assumption (2) are satisfied, then, given* $g_{i,0}, ..., g_{i,T-1}$, $i = 1, ..., K$, *the critical region* $\sup_{b \in CI_{\beta_0}(\alpha_1)} \tilde{p}_M\big[S_\bullet(b)\big] \leq \alpha_2$ *is such that*

$$\Pr\left(\sup_{b \in CI_{\beta_0}(\alpha_1)} \tilde{p}_M\big[S_\bullet(b)\big] \leq \alpha_2\right) \leq \alpha_1 + \alpha_2,$$

where $\tilde{p}_M[S_\bullet(b)]$ *is the MC p-value of the combined sign statistics computed as in (11).*

(ii) If H_0 *and Assumption (3) are satisfied, then, given* $g_{i,0}, ..., g_{i,T-1}$, $i = 1, ..., K$, *and* $|r_1 - b|, ..., |r_T - b|$, *for* $b \in CI_{\beta_0}(\alpha_1)$, *the critical region* $\sup_{b \in CI_{\beta_0}(\alpha_1)} \tilde{p}_M\big[SR_\bullet(b)\big] \leq \alpha_2$ *is such that*

$$\Pr\left(\sup_{b \in CI_{\beta_0}(\alpha_1)} \tilde{p}_M\big[SR_\bullet(b)\big] \leq \alpha_2\right) \leq \alpha_1 + \alpha_2,$$

where $\tilde{p}_M[SR_\bullet(b)]$ *is the MC p-value of the combined linear signed rank statistics computed according to (11).*

The first-stage confidence intervals appearing in this proposition can be obtained by considering the special cases of (5) and (7) in which $g_{i,t} = 1$. Specifically, the exact confidence interval for β_0 used in part (i) of Proposition 3 is constructed from the order statistics $r_{(1)}, ..., r_{(T)}$. If we choose δ such that $\Pr(B \leq \delta) = \alpha_1/2 = \Pr(B \geq T - \delta)$, where B follows a binomial distribution $Bi(T, 1/2)$, then $[r_{(\delta+1)}, r_{(T-\delta)}]$ is a $(1-\alpha_1)100\%$ confidence interval for β_0 which is valid under H_0 and Assumption (2). This confidence interval simply reports all the values b that are not rejected by the sign test $\sum_{t=1}^{T} s[r_t - b]$ of the hypothesis that $r_1, ..., r_T$

are random variables each with a distribution whose median equals b; see Pratt and Gibbons (1981, pp. 92–96) and Hettmansperger (1984, pp. 12–15) for details. If the sample size is large enough (> 20), a normal approximation can be used to find δ as $\delta \doteq T/2 - z_{\alpha_1/2}\sqrt{T/4}$, where $z_{\alpha_1/2}$ is the upper $\alpha_1/2$ percentile of the standard normal distribution.

When the innovations ε_t are further assumed symmetric as in (3), a tighter confidence interval for β_0 can be obtained by inverting a Wilcoxon signed rank test $\mathcal{W} = \sum_{t=1}^{T} s[r_t - b]R_t^+(b)$. This confidence interval [used in part (ii) of Proposition 3] is easily constructed from the $\mathcal{N} = T(T+1)/2$ Walsh averages $(r_i + r_j)/2$, $1 \leq i \leq j \leq T$. If $\omega_{(1)}, ..., \omega_{(\mathcal{N})}$ are the ordered Walsh averages and $\Pr(\mathcal{W} \leq \delta) = \alpha_1/2 = \Pr(\mathcal{W} \geq \mathcal{N} - \delta)$, then $[\omega_{(\delta+1)}, \omega_{(\mathcal{N}-\delta)}]$ is the $(1-\alpha_1)100\%$ confidence interval for β_0 based on the $\mathcal{W}$ test. The distribution of the Wilcoxon variate has been tabulated for various values of T; see, for example, Wilcoxon et al. (1970). As before, the normal approximation can be used to find

$$\delta \doteq \frac{T(T+1)}{2} - z_{\alpha_1/2}\sqrt{\frac{T(T+1)(2T+1)}{24}},$$

which works well even in small samples; see Hettmansperger (1984, pp. 38–41) for further details. In what follows, we use the maximized p-values over first-stage confidence intervals based on the normal approximation. Note that the sample median $\check{\beta}_0$ is always an element of $CI_{\beta_0}(\alpha_1)$, whether this confidence interval is constructed by inverting the sign test or the Wilcoxon signed rank test.

4 Simulation results

This section presents the results of some simulation experiments designed to examine the performance of the proposed tests for stock return predictability. Here we simply use S_{min}, $S_{\times}$, W_{min}, $W_{\times}$ to refer to the two-stage MC tests, implemented with the sign statistic in (5) and the Wilcoxon signed rank statistic in (7) according to Proposition 3. The tests are performed at the nominal $\alpha = 5\%$ significance level with $M - 1 = 99$ MC samples. We compute S_{min}, $S_{\times}$, W_{min}, and $W_{\times}$ by grid search.[7] Given a desired level α, Proposition 3 shows that there is a trade-off between the width of the first-stage confidence interval $CI_{\beta_0}(\alpha_1)$ and the significance level $\alpha_2 = \alpha - \alpha_1$ of the second-stage tests based on the elements of $CI_{\beta_0}(\alpha_1)$. While the choice of α_1, α_2 has no effect on the overall level (as long as $\alpha_1 + \alpha_2 = \alpha$), it does matter for power. Campbell and Dufour (1997) show that it is better to take a wider confidence interval for β_0 in order to have a tighter critical value in the second stage. We therefore carry on with the testing strategy represented by $\alpha_1 = 1\%$, $\alpha_2 = 4\%$.

The data-generating process is the system in (4) comprising the predictive regression model with GARCH-type effects and two potential predictors, themselves governed by a VAR model. The complete specification is given as

$$\begin{aligned} r_t &= \beta_0 + \beta_1 x_{1,t-1} + \beta_2 x_{2,t-1} + \sigma_t \eta_t, \\ x_{1,t} &= \mu_1 + \phi_{11} x_{1,t-1} + \phi_{12} x_{2,t-1} + v_{1t}, \\ x_{2,t} &= \mu_2 + \phi_{21} x_{1,t-1} + \phi_{22} x_{2,t-1} + v_{2t}, \end{aligned} \tag{14}$$

[7] In practical applications the search for the maximal p-value can be stopped and the null hypothesis can no longer be rejected at level α as soon as a grid point yields a non-rejection. For instance, if $\tilde{p}_M[S_{min}(\check{\beta}_0)] > \alpha_2$ then $\sup_{b \in CI_{\beta_0}(\alpha_1)} \tilde{p}_M\big[S_{min}(b)\big] > \alpha_2$ and H_0 is not significant at the overall level α.

for $t = 1, ..., T$, after a burn-in period of length 1000. The vector $\boldsymbol{\epsilon}_t = (\eta_t, v_{1t}, v_{2t})'$ is i.i.d. $N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ with

$$\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \begin{bmatrix} 1 & \rho_{x_1 r} & 0 \\ \rho_{x_1 r} & 1 & \rho_{x_1 x_2} \\ 0 & \rho_{x_1 x_2} & 1 \end{bmatrix}.$$

The parameter $\rho_{x_1 r} = \text{Corr}(v_{1t}, \eta_t)$ controls the strength of direct feedback from η_t to future values of the regressors appearing on the right-hand side of the predictive regression in (14), and $\rho_{x_1 x_2} = \text{Corr}(v_{1t}, v_{2t})$ is the innovation correlation between the two predictor variables. When $\rho_{x_1 r} \neq 0$, the return innovation η_t affects $x_{1,t}$; and when $\rho_{x_1 r} \neq 0$, $\rho_{x_1 x_2} \neq 0$, the term η_t affects both $x_{1,t}$ and $x_{2,t}$ (*i.e.*, the regressors used to explain r_{t+1}).

We consider three specifications for the volatility of returns. The first is a standard GARCH model (Bollerslev, 1986) of the form

$$\sigma_t^2 = a_0 + a_1 \sigma_{t-1}^2 \eta_{t-1}^2 + b_1 \sigma_{t-1}^2,$$

where positive and negative values of η_{t-1} have the same effect on σ_t^2. Next, we consider the exponential GARCH (EGARCH) model of Nelson (1991) that allows for an asymmetric response of volatility to shocks:

$$\log(\sigma_t^2) = a_0 + h(\eta_{t-1}) + b_1 \log(\sigma_{t-1}^2),$$

where $h(\eta_{t-1}) = a_1 \eta_{t-1} + c_1(|\eta_{t-1}| - E|\eta_{t-1}|)$. With this specification σ_t^2 depends on both the size and the sign of η_{t-1}. Indeed, the function $h(\eta_{t-1})$ has slope $c_1 + a_1$ when $\eta_{t-1} \geq 0$,

and slope $c_1 - a_1$ when $\eta_{t-1} < 0$. So when $a_1 < 0$ the EGARCH captures the usual *leverage effect*, whereby a negative return tends to increase future volatility by more than a positive return of the same magnitude (Black, 1976; Christie, 1982; French et al., 1987).

Finally we consider a stochastic volatility (SV) model *à la* Harvey and Shephard (1996) given as

$$\log(\sigma_{t+1}^2) = \theta \log(\sigma_t^2) + u_t,$$

with $u_t = \rho_{x_1 h} v_{1t} + \sqrt{1 - \rho_{x_1 h}^2} z_t$ and where z_t is i.i.d. N(0,1), independently of $\boldsymbol{\epsilon}_t$. The timing of the volatility innovation term u_t appearing in this equation ensures that $\sigma_t \eta_t$ in (14) is a martingale difference; see Harvey and Shephard (1996) and Yu (2005) for more on this point.

The parameter $\rho_{x_1 h} = \text{Corr}(v_{1t}, u_t)$ allows for a leverage effect via what we call here a "volatility feedback effect" in the following way.[8] When $\rho_{x_1 h} \neq 0$, the variation in u_t has a feedback effect on $x_{1,t}$, since v_{1t} and u_t are correlated.[9] And when both $\rho_{x_1,r} \neq 0$ and $\rho_{x_1 h} \neq 0$, then η_t can exert a leverage effect on σ_{t+1}^2. Moreover, these effects operate through both v_{1t} and v_{2t} whenever $\rho_{x_1 x_2} \neq 0$. Note also that the feedback from either η_t or u_t to future values of the regressors will persist over time whenever the ϕ's are non-zero.

With the monthly data (presented in the next section), we tried estimating the model with a particle filter to get a sense of the empirically relevant values for the volatility feedback parameter, $\rho_{x_1 h}$. Even with just a single predictor it proved well-nigh impossible to obtain

[8] The *volatility feedback effect* usually refers to a causal effect from volatility to prices: if volatility is priced, an anticipated increase in volatility will raise the required rate of return, and necessitate an immediate stock price decline in order to allow for higher future returns (French et al., 1987).

[9] Note that this mechanism gives rise to outer quantile predictability (Cenesizoglu and Timmermann, 2008; Lee, 2016; Maynard et al., 2010), since variation in v_{1t} also affects σ_{t+1}^2 and hence the conditional quantiles of r_{t+1}.

reliable estimates owing to the model's complexity and the limited sample size ($T = 804$). Our second-best solution was to estimate the correlation parameters using proxies for σ_t^2, obtained from the sum of squared daily returns (a realized variance measure) and from standard GARCH and EGARCH filters. The resulting estimates for $\rho_{x_1 h}$ were 0.12, 0.28, and 0.73, respectively, with the dividend-price ratio as predictor.

The intercept of the predictive regression in (14) is set as $\beta_0 = 0$, but this is not assumed known and the procedures in §3.2 dealing with an unknown β_0 are applied. The GARCH and EGARCH parameters are set to closely match the estimates obtained from the monthly excess stock returns on the S&P value-weighted index. These values are $a_0 = 0.01$, $a_1 = 0.12$, $b_1 = 0.84$ for the GARCH; and $a_0 = -0.67$, $a_1 = -0.12$, $c_1 = 0.21$, $b_1 = 0.90$ for the EGARCH. The parameters of the VAR component in (14) are set as $\mu_1 = \mu_2 = 0$, $\phi_{12} = \phi_{21} = 0$, and all the other parameters are varied to examine their effects.

Specifically, we provide results for selected subsets of the cases for which $\phi_{11} = \phi_{22} = 0.90$, 0.99; $\rho_{x_1 r} = 0$, -0.90; $\theta = 0.90$, 0.99; $\rho_{x_1 h} = 0$, 0.25, 0.5, 0.7; $\rho_{x_1 x_2} = 0$, -0.1, 0.1; and $T = 120$, 240, 720. These sample sizes correspond to 10, 20, and 60 years of monthly returns, respectively, which closely matches our empirical application. In each case, the reported results are based on 1000 replications of the data-generating configuration.

The size results are shown in Table 1 for the GARCH cases and in Table 2 for the SV cases. Here we use the standard Wald test, the Amihud et al. (2009) mARM-based Wald test, and the Kostakis et al. (2015) IVX-estimated Wald test as benchmarks for comparison purposes. The main takeaways from the size experiments are summarized as follows.

1. When there are no feedback effects whatsoever ($\rho_{x_1 r} = 0$, $\rho_{x_1 h} = 0$), all the tests respect

the nominal 5% level constraint fairly well. This holds under the SV and GARCH specifications. The Wald test, however, is sensitive to the presence of feedback ($\rho_{x_1 r} = -0.90$). The over-rejection problem becomes apparent when $\phi_{11} = 0.99$, especially with the smaller sample sizes. The mARM and IVX tests do much better than the Wald test with empirical sizes closer to 5% with this type of (direct) feedback.

2. The presence of SV without a volatility feedback effect ($\theta = 0.99$, $\rho_{x_1 h} = 0$) does not cause major size distortions for the Wald, mARM, and IVX tests. These findings agree with Table 1 and those reported by Kostakis et al. (2015, §2.2) who also found that their IVX test exhibits no size distortions in the presence of pure GARCH effects. However, when SV is accompanied by a volatility feedback effect ($\rho_{x_1 h} \neq 0$), the Wald, mARM, and IVX tests can over-reject. Table 2 shows how fast these over-rejections grow with $\rho_{x_1 h}$. Why does IVX over-reject? Because IVX is a *linear* filtering method for the regressors $\boldsymbol{x}_t$ which cannot attenuate the *non-linear* effects of volatility feedback.

3. Not surprisingly then, a direct return feedback effect coupled with an indirect volatility feedback effect ($\rho_{x_1 r} = -0.90$, $\rho_{x_1 h} \neq 0$) makes matters potentially worse for the Wald-mARM-IVX group of tests, especially when the time-series persistence increases. For instance, when $\phi_{11} = 0.99$ and $\theta = 0.99$ (last set of results in Table 2), the empirical size of these tests can be four times larger than the nominal level. Observe also that the over-rejection problem is exacerbated as the sample size increases from 240 to 720.

4. Even the plug-in S^m_{min}, $S^m_{\times}$, W^m_{min}, $W^m_{\times}$ tests are influenced by the volatility feedback effect, although to a lesser extent than the Wald-mARM-IVX group. For example, with $\phi_{11} = 0.99$ and $\theta = 0.99$ (last set of results in Table 2), the empirical size of S^m_{min}

and $S_{\times}^{m}$ reaches 11% when $T = 720$. The W_{min}^{m} and $W_{\times}^{m}$ tests seem to do much better.

5. The empirical size of the S_{min}, $S_{\times}$, W_{min}, $W_{\times}$ tests (based on a first-stage confidence interval for β_0) is seen to remain close to zero in Tables 1 and 2. This concurs with the fact that these tests are the only ones that are completely robust (*i.e.*, invariant) to: (i) the strength of direct feedback and regressor persistence; and (ii) the strength of volatility feedback and the persistence of stochastic volatility. Indeed, the empirical size of these conservative tests is always less than the nominal 5% significance level, in accordance with the developed theory.

The power comparisons in Tables 3 and 4 use data generated according to (14) with $\beta_0 = \mu_1 = \mu_2 = 0$; $\phi_{12} = \phi_{21} = 0$; $\phi_{11} = \phi_{22} = \theta = 0.90$; $\rho_{x_1 r} = -0.90$; the other parameter values and the sample sizes are listed in the tables. Given the size distortions seen in Tables 1 and 2, the power results for the Wald, mARM, IVX, and even the S_{min}^{m}, $S_{\times}^{m}$, W_{min}^{m}, $W_{\times}^{m}$ tests are based on size-corrected critical values. Such adjustments were not applied to the S_{min}, $S_{\times}$, W_{min}, $W_{\times}$ tests because the probability of a Type I error with these tests is $\leq \alpha$. The main findings that emerge from Tables 3 and 4 can be summarized as follows.

1. Under GARCH dynamics (Table 3), the Wald-mARM-IVX group has the best power. From the much smaller β_1, β_2 values we see that EGARCH makes return predictability far easier to detect compared to the standard GARCH. To understand why, consider the signal-to-noise ratio β_1/σ_t^2. With the standard GARCH, positive and negative return shocks have the same effect on β_1/σ_t^2. In contrast, return shocks have an asymmetric effect under EGARCH, whereby $\eta_{t-1} > 0$ lowers σ_t^2 and hence increases β_1/σ_t^2. And from Table 3 we see that this positive effect outstrips the opposing effect of negative

return shocks.

2. Under SV dynamics (Table 4), the plug-in S^m_{min}, $S^m_{\times}$, W^m_{min}, $W^m_{\times}$ tests have the best power among all the tests. These findings corroborate the conventional wisdom that non-parametric tests can perform well, particularly in the presence of (conditional) heteroskedasticity. Comparing the results with $\rho_{x_1 h} = 0.5$ against those with $\rho_{x_1 h} = 0$, we see that the volatility feedback effect lowers the power of all the tests. Of course, power improves across the board as the sample size increases.

3. The power of the S_{min}, $S_{\times}$, W_{min}, $W_{\times}$ tests (based on a first-stage confidence interval for β_0) can be quite low when T is small. But with $T = 240$, 720 they can outperform the Wald, mARM, and IVX tests by a surprisingly wide margin. This is seen to occur in Table 4.

4. An examination of the S_{min}, $S_{\times}$, W_{min}, $W_{\times}$ tests in Tables 3 and 4 when $\text{sign}(\beta_1) = \text{sign}(\beta_2)$ reveals that they tend to suffer when $\rho_{x_1 x_2} = -0.1$ and to benefit when $\rho_{x_1 x_2} = 0.1$. This occurs because β_1 and β_2 share the same sign, so changes in $x_{1,t-1}$ and $x_{2,t-1}$ tend to move the conditional median of r_t (given by $\beta_0 + \beta_1 x_{1,t-1} + \beta_2 x_{2,t-1}$) in the same direction when $\rho_{x_1 x_2}$ is positive, whereas the two predictors have opposing effects on r_t's conditional median when $\rho_{x_1 x_2}$ is negative.

5. Comparing the $\mathcal{S}_{min}$ and $\mathcal{S}_{\times}$ tests when $\beta_1 = \beta_2$, we see that the signed rank tests tend to perform better if they are combined using the product rule rather than via the minimum p-value.

5 Empirical results

To further illustrate the new test procedure, we examine the predictability of monthly excess stock returns (r) using U.S. data. Our empirical investigation uses six predictors that are widely used in the stock return predictability literature: the log dividend-price ratio (d/p), the log earnings-price ratio (e/p), the book-to-market ratio (btm), the default yield spread (dfy), the term spread (tms), and the short rate (tbl).[10] These data are in fact a subset of those used by Goyal and Welch (2008) and updated through the year 2014. Here we consider monthly data obtained from Amit Goyal's website for the 67-year time span from January 1948 to December 2014. Several studies find evidence of structural breaks in stock return predictive regressions over the postwar era. In particular, Paye and Timmermann (2006), Rapach and Wohar (2006), Lettau and Van Nieuwerburgh (2008), and Gonzalo and Pitarakis (2012) estimate predictive regression models over different periods and find that the in-sample predictive ability of financial variables can vary markedly over time. Following theses authors, we address the issue of possible predictive instabilities by also running the tests over fixed 10- and 20-year subsamples[11] and 20-year rolling window subsamples.[12]

Figure 1 shows the time series of monthly excess returns and Figure 2 shows the monthly time series of the six considered predictors. The predictors appear very persistent with a tendency to wander off for long periods. The notable exception is the term spread in panel (e), which seems to be far more mean-reverting than the five other predictors. This can also be ascertained from Table 5 which reports some summary statistics (mean, standard

[10] A full description of these data is found in Goyal and Welch (2008).

[11] The last subsamples of fixed length (Jan 2008–Dec 2014 and Jan 1988–Dec 2014) are slightly shorter than the previous ones.

[12] Note that the subsample analysis is subject to the problem of multiple comparisons, since the predictability tests are performed on each subsample separately.

deviation, first-order autocorrelation) and the correlations among the variables. The autocorrelation of excess stock returns is near zero, whereas the predictors are highly persistent with autocorrelations close to one.

Table 6 summarizes some of the distributional properties of the monthly excess stock returns for the full sample and fixed subsamples. The reported statistics include the mean, standard deviation, skewness, kurtosis, Jarque-Bera normality test statistic, first-order autocorrelation, and the Ljung-Box portmanteau test statistic $Q^2(k)$ for squared returns using $k = 6$ and $k = 12$ lags. The latter statistic is used to detect serial dependence in the volatility of excess returns. Besides the well-known Jarque-Bera joint test, we assess the normality of the excess return distribution with the D'Agostino (1970) test for skewness and the Anscombe and Glynn (1983) test for kurtosis. Both of these test statistics are approximately normally distributed when the population data follows a normal distribution. In Table 6, bold entries indicate statistical significance at the 10% level.

Over the full sample period, there is some evidence of negative skewness in the return data. The evidence from the 10- and 20-year subsamples, however, indicates that returns are symmetrically distributed. The monthly stock returns tend to be heavy-tailed, both in the full sample and the subsamples. Finally, the Ljung-Box tests clearly indicate the presence of conditional heteroskedasticity at the monthly frequency. These findings are completely in line with the huge body of literature that documents GARCH-type or stochastic volatility effects in stock returns (cf. Cont, 2001).

Following Amihud et al. (2009), we report in Table 7 the parameter estimates from the system of equations in (4) along with Newey-West standard errors in parentheses. The first row shows the one-month ahead predictive regression results. The remaining rows display

the parameter estimates of the VAR model, estimated using equation-by-equation OLS. The entries in bold represent cases of significance at the 5% level. According to the Newey-West adjusted t-statistics, only the short rate appears to be a significant predictor of stock returns. Looking at the own persistence estimates in the VAR block, we see that the predictors are highly persistent with autoregressive coefficients between 0.918 and 1.007.

Table 8 shows the estimated residual correlations from model (4). The first column shows the residual correlations between the stock returns and the predictors, and these correlations give an indication of the strength of direct feedback. Not surprisingly, financial ratios such as d/p, e/p, b/m are highly and negatively correlated with stock returns since the stock price appears in the denominator of these ratios. Observe also that the conditional correlations in Table 8 are much higher than their unconditional counterparts in Table 5. The results in Tables 7 and 8 are in line with the literature that highlights the persistence and endogeneity of the usual predictors appearing in stock return predictive regressions. Indeed, the evidence of high persistence seen in the VAR block of Table 7 combined with the strong residual correlations revealed by Table 8 is an early warning that the Wald test may not be reliable.

Table 9 reports the results for the sign and signed rank joint predictability tests along with the standard Wald test (in the last column). For the full-sample period, the Wald test strongly rejects the null hypothesis with a p-value essentially equal to zero. The plug-in S^m_{min} and W^m_{min} tests tend to agree with the Wald test with p-values $\leq 4\%$. On the contrary, the two-stage tests do not reject the null of no predictability over the full 67-year sample.

The two-stage tests are theoretically guaranteed to be level-correct, but they tend to be less powerful owing to their conservative nature. On the other hand, the plug-in tests lack a proper theory but the simulation evidence in Section 4 shows that they have relatively minor

size distortions and better power. If we are willing to take a leap of faith with the plug-in tests, then the results suggest that there is *some* evidence of stock return predictability at the monthly frequency.

The subsample analysis reveals an even more nuanced picture. Indeed the plug-in tests find some evidence of predictability in the 20-year subperiods ($T = 240$), but hardly any with the 10-year subperiods ($T = 120$) and the two-stage test clearly indicate non-rejections. The Wald test continues to consistently reject the null hypothesis of no predictability, except in the 10-year subperiod from January 1988 to December 1997. One could argue that the non-rejection by the two-stage tests is due to their lower power in smaller samples. On the other hand, it seems more unlikely that the disagreement between the plug-in tests and the Wald test is due to power differences. Indeed, the simulation evidence presented earlier shows that the power of the plug-in tests can be similar or higher than the Wald test.

More striking yet are the 20-year rolling-window predictability test results in Figure 3. Here the solid black lines show the p-values of the sign and signed rank tests based on the minimum p-value rule, the solid grey lines show the p-values of those tests based on the product of the individual p-values, the dashed grey lines show the p-values of the Wald test, and finally the horizontal dotted line indicates the nominal 5% significance level. Figure 3 clearly shows the tendency of the Wald test to almost always reject the null, except during the most recent times.[13] In sharp contrast, the wild fluctuations in the p-values of the proposed tests point to rejections only infrequently. Most of these rejections occur during the very early period 1968–1972. The plug-in tests also tend to reject the null hypothesis in

[13]More specifically, at the 5% level, the Wald test rejects the null 503 times out of the 564 monthly rolling regressions.

the late 1980s to early 1990s, however this is not supported by the two-stage tests.[14]

In addition to the evidence of joint predictability over the full 67-year sample period, one may also want to know which predictors drive these results. So next we investigate the source of the predictability by evaluating the marginal p-value of each predictor in a univariate regression setup. The reported marginal p-values in Table 10 reveal that among the six regressors it is only the term spread (tms) that has predictive ability for the monthly excess stock returns. Another way to see this is from Table 11. When the term spread is excluded from the information set and the joint tests are conducted with the five remaining regressors – the cases with $K = 5$ in Table 11 – the p-values of the joint predictability tests cease to reject the null hypothesis.

The evidence uncovered here about the predictive ability of the term spread agrees with the findings of Fama and French (1989), Fama (1990), Schwert (1990), Campbell and Thompson (2008), and Rapach et al. (2016). In particular, Fama and French (1989) argue that the term spread captures cyclical variation in expected returns because of its covariation with short-term business cycle fluctuations. Estrella and Hardouvelis (1991) also show a strong association between future changes in real economic activity and the term spread.

The takeaway message from our empirical application is that although the new tests uncover some evidence of stock return predictability at the monthly frequency, this evidence is not consistently supported by all the tests. Moreover, the predictability evidence is entirely driven by the term spread and it does not consistently hold up over subsamples. As our

[14] As Lettau and Van Nieuwerburgh (2008) explain, changes in predictive ability could be due to changes in the steady-state growth rate of economic fundamentals resulting from permanent technological innovations and/or changes in the expected return caused by, for example, improved risk sharing, changes in stock market participation, changes in the tax code, or lower macroeconomic volatility.

working paper shows, similar results were obtained at the quarterly frequency.[15] Taken together, these results suggest that there is indeed a predictable component in excess stock returns, but one that only holds over a long time span.[16]

6 Concluding remarks

Investigations of stock return predictability have to contend with several problems that can undermine the reliability of statistical inference in "small" samples. Chief among these is that typically there is feedback from returns to future values of the regressors and these endogenous regressors are highly persistent over time. In such circumstances, OLS yields biased estimates and standard testing procedures may reject the null hypothesis of no predictability much too often. This over-rejection problem can be further exacerbated by the presence of time-varying conditional non-normalities and other stock return distribution heterogeneities, like stochastic volatility feedback effects. Indeed, the standard Wald test and even the Amihud et al. (2009) bias-corrected Wald test and the Kostakis et al. (2015) IVX-estimated "persistence-robust" Wald test can fail substantially to control test size under such conditions.

In this paper, we have developed a small-sample testing procedure that is invariant to all these sources of size distortions in predictability testing. Furthermore, the proposed tests display good power properties under a variety of data-generating configurations. This is

[15] See the Bank of Canada Staff Working Paper 2017-10, which can be found at www.bankofcanada.ca/2017/03/staff-working-paper-2017-10/.

[16] It is interesting to note the similarity of this finding with Shiller and Perron (1985) who show that the power of random walk tests depends more on the span of the data rather than the number of observations. Observe also that the random walk hypothesis is a special case of the present framework. For instance, to test whether, say, p_t follows a random walk, simply recast (1) as $p_t - p_{t-1} = \beta_0 + \beta_1 p_{t-1} + \sigma_t \varepsilon_t$ and apply the Campbell and Dufour (1997) sign and signed rank tests.

achieved with tests based on signs and signed ranks for each considered regressor and by using Monte Carlo resampling techniques to combine the marginal p-values in a way that controls the joint significance level. The Lehmann and Stein (1949) impossibility theorem shows that such sign-based tests are the *only* ones that yield valid inference in the presence of non-normalities and heteroskedasticity of unknown form. Another interesting feature of the proposed test procedure is that no modelling assumptions whatsoever are made for the regressor variables. This means that the predictors may exhibit any degree of persistence and may be subject to unmodelled structural breaks, time-varying parameters, or any other non-linearities.

Finally, we note that the developed procedure could be extended to test for long-horizon stock return predictability with multiple regressors by following the approach in Liu and Maynard (2007). Their idea is based on a rearrangement of the predictive regression, whereby the regression of a long-horizon return on a single-period predictor is replaced by a regression of a one-period return on a long-horizon regressor. We leave this extension for future research.

Acknowledgments

We thank a guest co-editor and two anonymous referees for several useful comments. We are also grateful for the comments by Gregory Bauer, Antonio Diez de los Rios, as well as seminar and conference participants at the Bank of Canada, the 2015 Computational Financial Econometrics Conference, the 2016 CIREQ Econometrics Conference in Honor of Jean-Marie Dufour, the 2016 International Association for Applied Econometrics Conference, the 56e Congrès annuel de la société canadienne de science économique, the 50th Annual Conference

of the Canadian Economics Association, the 69th European Meeting of the Econometric Society, and the 20th OxMetrics User Conference. All remaining errors and omissions are our own. The second author gratefully acknowledges financial support by the Social Sciences and Humanities Research Council of Canada.

Appendix: Proofs

Proof of Proposition 1: (i) Suppose model (1) holds, and let $\{g_{i,t}\}_{t=0}^{T-1} = \{g_{i,t}(\mathcal{I}_t)\}_{t=0}^{T-1}$, $i = 1, ..., K$, be sequences of measurable functions of $\mathcal{I}_t$ such that $\Pr(g_{i,t} = 0) = 0$, for all t. Let $s_{i,t} = s\big[(r_t - \beta_0)g_{i,t-1}\big]$ and consider the distribution of the vector

$$\big(s_{i,1}, ..., s_{i,T-1}, s_{i,T}\big).$$

Conditional on $\mathcal{I}_{T-1} = (\boldsymbol{x}_0', \boldsymbol{x}_1', ..., \boldsymbol{x}_{T-1}', r_1, ..., r_{T-1})'$, the variables $s_{i,1}, ..., s_{i,T-1}, g_{i,T-1}$ are fixed. So under H_0 and given $\mathcal{I}_{T-1}$, we have that $s\big[(r_T - \beta_0)g_{i,T-1}\big] \overset{d}{=} s\big[\varepsilon_T g_{i,T-1}\big]$. The assumption in (2) that the distribution of ε_T has a conditional median equal to zero further implies that $s\big[\varepsilon_T g_{i,T-1}\big] \overset{d}{=} s\big[\tilde{\varepsilon}_T g_{i,T-1}\big] \overset{d}{=} B_T$, where $\tilde{\varepsilon}_T$ is any random variable such that $\Pr(\tilde{\varepsilon}_T > 0) = \Pr(\tilde{\varepsilon}_T < 0) = 1/2$ and B_T is a Bernoulli variable such that $\Pr(B_T = 0) = \Pr(B_T = 1) = 1/2$. It follows that if H_0 and Assumption (2) are satisfied, then, given $\mathcal{I}_{T-1}$, we have

$$\big(s_{i,1}, ..., s_{i,T-1}, s_{i,T}\big) \overset{d}{=} \big(s_{i,1}, ..., s_{i,T-1}, s[\tilde{\varepsilon}_T g_{i,T-1}]\big) \overset{d}{=} \big(s_{i,1}, ..., s_{i,T-1}, B_T\big).$$

Applying the same argument recursively to $\big(s_{i,1},...,s_{i,\tau},B_T\big)$ for $\tau = T-1,...,1$, we find that

$$\big(s_{i,1},...,s_{i,T-1},s_{i,T}\big) \stackrel{d}{=} \big(s[\tilde{\varepsilon}_1 g_{i,0}],...,s[\tilde{\varepsilon}_{T-1} g_{i,T-2}],s[\tilde{\varepsilon}_T g_{i,T-1}]\big) \stackrel{d}{=} \big(B_1,...,B_{T-1},B_T\big),$$

where $\tilde{\varepsilon}_1,...,\tilde{\varepsilon}_T$ are mutually independent random variables such that $\Pr(\tilde{\varepsilon}_t > 0) = \Pr(\tilde{\varepsilon}_t < 0) = 1/2$; and $B_1,...,B_T$ are mutually independent Bernoulli variables on $\{0,1\}$ with $\Pr(B_t = 0) = \Pr(B_t = 1) = 1/2$. Thus the distribution of $(s_{i,1},...,s_{i,T})$ is independent of $g_{i,0},...,g_{i,T-1}$. Furthermore we have that $S_i(\beta_0) \stackrel{d}{=} \sum_{t=1}^{T} s\big[\tilde{\varepsilon}_t g_{i,t-1}\big] \stackrel{d}{=} \sum_{t=1}^{T} B_t$, for each $i = 1,...,K$, since $\mathbf{X} \stackrel{d}{=} \mathbf{Y} \Rightarrow f(\mathbf{X}) \stackrel{d}{=} f(\mathbf{Y})$ for any measurable function $f(\cdot)$ defined on the common support of $\mathbf{X}$ and $\mathbf{Y}$ (Randles and Wolfe, 1979, Theorem 1.3.7).

(ii) Define d_t to be the position of the integer t in the realization of the vector $\big(R_1^+(\beta_0),...,R_T^+(\beta_0)\big)$, $t = 1,...,T$. Thus

$$\sum_{t=1}^{T} s_{i,t}\varphi\big(R_t^+(\beta_0)\big) = \sum_{t=1}^{T} s_{i,d_t}\varphi(t)$$

and

$$\sum_{t=1}^{T} s[\tilde{\varepsilon}_t g_{i,0}]\varphi\big(R_t^+(\beta_0)\big) = \sum_{t=1}^{T} s[\tilde{\varepsilon}_{d_t} g_{i,t-1}]\varphi(t).$$

Under the conditions of part (i), we have that

$$\big(s_{i,1},...,s_{i,T}\big) \stackrel{d}{=} \big(s_{i,d_1},...,s_{i,d_T}\big) \stackrel{d}{=} \big(s[\tilde{\varepsilon}_{d_1} g_{i,0}],...,s[\tilde{\varepsilon}_{d_T} g_{i,T-1}]\big) \stackrel{d}{=} \big(B_1,...,B_T\big),$$

since i.i.d. random variables are *exchangeable*. The symmetry assumption in (3) further implies that $s_{i,t}$ is independent of $|r_t - \beta_0|$ and thus of $R_t^+(\beta_0)$ and $\varphi(R_t^+(\beta_0))$ (Randles

and Wolfe, 1979, Lemma 2.4.2). Moreover, this fact applies to each of the T mutually independent elements of $(s_{i,1}, ..., s_{i,T})$. We also have, conditional on $|r_1 - \beta_0|, ..., |r_T - \beta_0|$, that the vector of scores $\big(\varphi(R_1^+(\beta_0)), ..., \varphi(R_T^+(\beta_0))\big)$ is a fixed permutation of $\varphi(1), ..., \varphi(T)$. So given $g_{i,0}, ..., g_{i,T-1}$ and $|r_1 - \beta_0|, ..., |r_T - \beta_0|$, it follows that

$$\sum_{t=1}^{T} s_{i,t}\varphi\big(R_t^+(\beta_0)\big) \stackrel{d}{=} \sum_{t=1}^{T} s_{i,d_t}\varphi(t) \stackrel{d}{=} \sum_{t=1}^{T} s[\tilde{\varepsilon}_{d_t} g_{i,t-1}]\varphi(t) \stackrel{d}{=} \sum_{t=1}^{T} B_t\varphi(t)$$

and hence

$$SR_i(\beta_0) \stackrel{d}{=} \sum_{t=1}^{T} s[\tilde{\varepsilon}_t g_{i,0}]\varphi\big(R_t^+(\beta_0)\big) \stackrel{d}{=} \sum_{t=1}^{T} B_t\varphi(t).$$

Proof of Proposition 2: From the proof of Proposition 1 it is easy to see that

$$\begin{bmatrix} s_{1,1}, & ..., & s_{1,T-1}, & s_{1,T} \\ \vdots & & \vdots & \vdots \\ s_{K,1}, & ..., & s_{K,T-1}, & s_{K,T} \end{bmatrix} \stackrel{d}{=} \begin{bmatrix} s\big[\tilde{\varepsilon}_1 g_{1,0}\big], & ..., & s\big[\tilde{\varepsilon}_{T-1} g_{1,T-2}\big], & s\big[\tilde{\varepsilon}_T g_{1,T-1}\big] \\ \vdots & & \vdots & \vdots \\ s\big[\tilde{\varepsilon}_1 g_{K,0}\big], & ..., & s\big[\tilde{\varepsilon}_{T-1} g_{K,T-2}\big], & s\big[\tilde{\varepsilon}_T g_{K,T-1}\big] \end{bmatrix},$$

given $g_{i,0}, ..., g_{i,T-1}$, for $i = 1, ..., K$. This equality in distribution between random matrices further implies that

$$\begin{bmatrix} S_1(\beta_0) \\ \vdots \\ S_K(\beta_0) \end{bmatrix} \stackrel{d}{=} \begin{bmatrix} \widetilde{S}_1(\beta_0) \\ \vdots \\ \widetilde{S}_K(\beta_0) \end{bmatrix} \text{ and } \begin{bmatrix} p_1(\beta_0) \\ \vdots \\ p_K(\beta_0) \end{bmatrix} \stackrel{d}{=} \begin{bmatrix} \tilde{p}_1(\beta_0) \\ \vdots \\ \tilde{p}_K(\beta_0) \end{bmatrix},$$

from which part (i) follows. The proof of part (ii) is identical, except that it begins from

$$\begin{bmatrix} s_{1,1}\varphi\big(R_1^+(\beta_0)\big), & ..., & s_{1,T}\varphi\big(R_T^+(\beta_0)\big) \\ \vdots & & \vdots \\ s_{K,1}\varphi\big(R_1^+(\beta_0)\big), & ..., & s_{K,T}\varphi\big(R_T^+(\beta_0)\big) \end{bmatrix} \stackrel{d}{=} \begin{bmatrix} s\big[\tilde{\varepsilon}_1 g_{1,0}\big]\varphi\big(R_1^+(\beta_0)\big), & ..., & s\big[\tilde{\varepsilon}_T g_{1,T-1}\big]\varphi\big(R_T^+(\beta_0)\big) \\ \vdots & & \vdots \\ s\big[\tilde{\varepsilon}_1 g_{K,0}\big]\varphi\big(R_1^+(\beta_0)\big), & ..., & s\big[\tilde{\varepsilon}_T g_{K,T-1}\big]\varphi\big(R_T^+(\beta_0)\big) \end{bmatrix},$$

which holds conditionally on $g_{i,0}, ..., g_{i,T-1}$, $i = 1, ..., K$, and $|r_1 - \beta_0|, ..., |r_T - \beta_0|$.

Proof of Proposition 3: The proof here closely follows that of Campbell and Dufour (1997, Proposition 2). We will begin by establishing part (i) for the $S_\bullet$ statistic. All the probability statements made here are conditional on $g_{i,0}, ..., g_{i,T-1}$, $i = 1, ..., K$. We wish to show that $\Pr\left(\sup_{b \in CI_{\beta_0}(\alpha_1)} \tilde{p}_M[S_\bullet(b)] \leq \alpha_2\right) \leq \alpha_1 + \alpha_2$ under the conditions of Proposition 3. This will be true if $\Pr(A) \leq \alpha_1 + \alpha_2$, where A is the event $\tilde{p}_M[S_\bullet(b)] \leq \alpha_2$ for all $b \in CI_{\beta_0}(\alpha_1)$. Define the set $I = \{b : b \in CI_{\beta_0}(\alpha_1) \text{ and } \tilde{p}_M[S_\bullet(b)] > \alpha_2\}$. Then, via Bonferroni's inequality, we have that

$$\begin{aligned} \Pr(\beta_0 \in I) &= 1 - \Pr\big(\beta_0 \notin CI_{\beta_0}(\alpha_1) \text{ or } \tilde{p}_M\big[S_\bullet(\beta_0) \leq \alpha_2\big]\big) \\ &\geq 1 - \Pr(\beta_0 \notin CI_{\beta_0}(\alpha_1)) - \Pr\big(\tilde{p}_M\big[S_\bullet(\beta_0) \leq \alpha_2\big]\big) \\ &\geq 1 - \alpha_1 - \alpha_2, \end{aligned}$$

since $\Pr(\beta_0 \in CI_{\beta_0}(\alpha_1)) \geq 1 - \alpha_1$ by definition of the first-stage confidence interval for β_0, and $\Pr\big(\tilde{p}_M\big[S_\bullet(\beta_0) \leq \alpha_2\big]\big) = \alpha_2$ from (12). Observe that $\Pr(A) = \Pr(B^c)$, where B is the

event $\tilde{p}_M[S_\bullet(b)] > \alpha_2$ for some $b \in CI_{\beta_0}(\alpha_1)$. Note also that $\beta_0 \in I \Rightarrow B$. Hence

$$\Pr(B) \geq \Pr(\beta_0 \in I) \geq 1 - \alpha_1 - \alpha_2,$$

which implies the desired result: $\Pr(A) \leq \alpha_1 + \alpha_2$.

The proof of part (ii) for the $SR_\bullet$ statistic is identical except that the probability statements are conditional on $g_{i,0}, ..., g_{i,T-1}$, $i = 1, ..., K$, and $|r_1 - b|, ..., |r_T - b|$, for $b \in CI_{\beta_0}(\alpha_1)$.

References

Amihud, Y. and C. Hurvich (2004). Predictive regression: A reduced-bias estimation method. *Journal of Financial and Quantitative Analysis 39*, 813–841.

Amihud, Y., C. Hurvich, and Y. Wang (2009). Multiple-predictor regressions: Hypothesis testing. *Review of Financial Studies 22*, 413–434.

Anscombe, F. and W. Glynn (1983). Distribution of the kurtosis statistic b_2 for normal samples. *Biometrika 70*, 227–234.

Barnard, G. (1963). Comment on 'The spectral analysis of point processes' by M.S. Bartlett. *Journal of the Royal Statistical Society (Series B) 25*, 294.

Beaulieu, M.-C., J.-M. Dufour, and L. Khalaf (2007). Multivariate tests of mean-variance efficiency with possibly non-Gaussian errors: An exact simulation-based approach. *Journal of Business and Economic Statistics 25*, 398–410.

Birnbaum, Z. (1974). Computers and unconventional test statistics. In F. Proschan and R. Serfling (Eds.), *Reliability and Biometry*, pp. 441–458. SIAM, Philadelphia.

Black, F. (1976). The pricing of commodity contracts. *Journal of Financial Economics 3*, 167–179.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics 31*, 307–327.

Campbell, B. and J.-M. Dufour (1997). Exact nonparametric tests of orthogonality and random walk in the presence of a drift parameter. *International Economic Review 38*, 151–173.

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

Campbell, J. and M. Yogo (2006). Efficient tests of stock return predictability. *Journal of Financial Economics 81*, 27–60.

Camponovo, L., O. Scaillet, and F. Trojani (2012). Predictive regression and robust hypothesis testing: Predictability hidden by anomalous observations. *SSRN Working Paper*.

Cavanagh, C., G. Elliott, and J. Stock (1995). Inference in models with nearly integrated regressors. *Econometric Theory 11*, 1131–1147.

Cenesizoglu, T. and A. Timmermann (2008). Is the distribution of stock returns predictable? *SSRN Working Paper*.

Christie, A. (1982). The stochastic behavior of common stock variances: Value, leverage and interest rate effects. *Journal of Financial Economics 10*, 407–432.

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance 1*, 223–236.

Coudin, E. and J.-M. Dufour (2009). Finite-sample distribution-free inference in linear median regressions under heteroscedasticity and non-linear dependence of unknown form. *Econometrics Journal 12*, S19–S49 (Tenth Anniversary Special Issue).

D'Agostino, R. (1970). Transformation to normality of the null distribution of g_1. *Biometrika 57*, 679–681.

Dufour, J.-M. (2003). Identification, weak instruments, and statistical inference in econometrics. *Canadian Journal of Economics 36*, 767–808.

Dufour, J.-M. (2006). Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. *Journal of Econometrics 133*, 443–477.

Dufour, J.-M. and L. Khalaf (2001). Monte Carlo test methods in econometrics. In B. Baltagi (Ed.), *A Companion to Theoretical Econometrics*, pp. 494–510. Basil Blackwell, Oxford, UK.

Dufour, J.-M., L. Khalaf, and M. Voia (2015). Finite-sample resampling-based combined hypothesis tests, with applications to serial correlation and predictability. *Communications in Statistics – Simulation and Computation, 44*, 2329–2347.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics 28*, 181–187.

Estrella, A. and G. Hardouvelis (1991). The term structure as a predictor of real economic activity. *Journal of Finance 46*, 555–576.

Fama, E. and K. French (1989). Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics 25*, 23–49.

Fama, E. F. (1990). Stock returns, expected returns, and real activity. *Journal of Finance 45*, 1089–1108.

Fisher, R. (1932). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

Folks, J. (1984). Combination of independent tests. In P. Krishnaiah and P. Sen (Eds.), *Handbook of Statistics 4: Nonparametric Methods*, pp. 113–121. North-Holland, Amsterdam.

French, K., W. Schwert, and R. Stambaugh (1987). Expected stock returns and volatility. *Journal of Financial Economics 19*, 3–29.

Gonzalo, J. and J.-Y. Pitarakis (2012). Regime-specific predictability in predictive regressions. *Journal of Business and Economic Statistics 30*, 229–241.

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

Gungor, S. and R. Luger (2015). Bootstrap tests of mean-variance efficiency with multi-

ple portfolio groupings. *L'Actualité économique 91*, 35–65 (Special issue in English on *Identification, Simulation, and Finite–Sample Inference*).

Harvey, A. and N. Shephard (1996). Estimation of an asymmetric stochastic volatility model for asset returns. *Journal of Business and Economic Statistics 14*, 429–434.

Hettmansperger, T. (1984). *Statistical Inference Based on Ranks*. Wiley, New York.

Kostakis, A., T. Magdalinos, and M. Stamatogiannis (2015). Robust econometric inference for stock return predictability. *Review of Financial Studies 28*, 1506–1553.

Lee, J. (2016). Predictive quantile regression with persistent covariates: IVX-QR approach. *Journal of Econometrics 192*, 105–118.

Lehmann, E. and J. Romano (2005). *Testing Statistical Hypotheses, Third Edition*. Springer, New York.

Lehmann, E. and C. Stein (1949). On the theory of some non-parametric hypotheses. *Annals of Mathematical Statistics 20*, 28–45.

Lettau, M. and S. Van Nieuwerburgh (2008). Reconciling the return predictability evidence. *Review of Financial Studies 21*, 1607–1652.

Lewellen, J. (2004). Predicting returns with financial ratios. *Journal of Financial Economics 74*, 209–235.

Liu, W. and A. Maynard (2007). A new application of exact nonparametric methods to long-horizon predictability tests. *Studies in Nonlinear Dynamics and Econometrics 11*, Article 7.

MacKinnon, J. (2009). Bootstrap hypothesis testing. In D. Belsley and J. Kontoghiorghes (Eds.), *Handbook of Computational Econometrics*, pp. 183–213. Wiley.

Mankiw, G. and M. Shapiro (1986). Do we reject too often?: Small sample properties of tests of rational expectations models. *Economics Letters 20*, 139–145.

Maynard, A., K. Shimotsu, and Y. Wang (2010). Inference in predictive quantile regressions. *University of Guelph Working Paper*.

Nelson, D. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica 59*, 347–370.

Paye, B. and A. Timmermann (2006). Instability of return prediction models. *Journal of Empirical Finance 13*, 274–315.

Pearson, K. (1933). On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika 25*, 379–410.

Polk, C., S. Thompson, and T. Vuolteenaho (2006). Cross-sectional forecasts of the equity premium. *Journal of Financial Economics 81*, 101–141.

Pratt, J. and J. Gibbons (1981). *Concepts of Nonparametric Theory*. Springer-Verlag, New York.

Randles, R. and D. Wolfe (1979). *Introduction to the Theory of Nonparametric Statistics*. Wiley, New York.

Rapach, D., M. Ringgenberg, and G. Zhou (2016). Short interest and aggregate stock returns. *Journal of Financial Economics 121*, 46–65.

Rapach, D. and M. Wohar (2006). Structural breaks and predictive regression models of aggregate U.S. stock returns. *Journal of Financial Econometrics 4*, 238–274.

Savin, N. (1984). Multiple hypothesis testing. In Z. Griliches and M. Intriligator (Eds.), *Handbook of Econometrics*, pp. 827–879. North-Holland, Amsterdam.

Schwert, G. (1990). Stock returns and real activity: A century of evidence. *Journal of Finance 45*, 1237–1257.

Shiller, R. and P. Perron (1985). Testing the random walk hypothesis: Power versus frequency of obervation. *Economics Letters 18*, 381–386.

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

Tippett, L. (1931). *The Methods of Statistics*. Williams and Norgate, London.

Torous, W., R. Valkanov, and S. Yan (2004). On predicting stock returns with nearly integrated explanatory variables. *Journal of Business 77*, 937–966.

Westfall, P. and S. Young (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley, New York.

Wilcoxon, F., S. Katti, and R. Wilcox (1970). Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test. In H. Harter and D. Owen (Eds.), *Selected Tables in Mathematical Statistics*, pp. 827–879. Institute of Mathematical Statistics, Providence, Rhode Island.

Wilkinson, B. (1951). A statistical consideration in psychological research. *Psychology Bulletin 48*, 156–158.

Wolf, M. (2000). Stock returns and dividend yields revisited: A new way to look at an old problem. *Journal of Business and Economic Statistics 18*, 18–30.

Yu, J. (2005). On leverage in a stochastic volatility model. *Journal of Econometrics 127*, 165–178.

Zhu, M. (2013). Jackknife for bias reduction in predictive regressions. *Journal of Financial Econometrics 11*, 193–220.

Table 1: Empirical size of predictability tests: GARCH dynamics

ϕ_{11}	$\rho_{x_1 r}$	T	Wald	mARM	IVX	S^m_{min}	$S^m_\times$	W^m_{min}	$W^m_\times$	S_{min}	$S_\times$	W_{min}	$W_\times$
GARCH													
0.90	0	120	6.6	5.4	6.8	6.0	5.6	5.8	5.1	0.2	0.1	0.8	1.2
		240	5.8	5.5	5.6	4.9	4.8	4.3	4.5	0.3	0.2	0.5	0.2
0.90	-0.9	120	8.4	5.9	4.8	5.0	5.7	5.0	4.0	0.1	0.2	0.4	0.7
		240	8.1	9.6	5.3	4.0	4.8	6.2	4.7	0.4	0.4	1.1	0.8
0.99	-0.9	120	20.0	11.0	6.8	4.9	3.6	4.0	3.6	0.1	0.1	0.3	0.2
		240	16.5	9.0	5.8	3.9	3.5	3.7	3.8	0.0	0.0	0.5	0.3
EGARCH													
0.90	0	120	6.1	5.7	6.3	5.5	5.7	5.9	5.1	0.1	0.1	0.8	1.2
		240	6.0	5.4	5.9	5.1	4.8	4.8	4.2	0.2	0.1	0.5	0.2
0.90	-0.9	120	8.8	7.3	5.0	5.4	5.2	4.9	4.4	0.2	0.2	0.5	0.5
		240	8.3	9.5	4.9	4.0	4.4	5.8	5.0	0.3	0.4	1.1	0.6
0.99	-0.9	120	20.3	11.5	7.4	4.8	3.8	4.2	3.3	0.1	0.1	0.3	0.4
		240	17.0	9.3	5.9	4.0	3.2	3.9	4.1	0.3	0.4	0.5	0.2

Notes: This table reports the empirical size (in percentage) of the standard Wald test, the Amihud et al. (2009) mARM-based Wald test, the Kostakis et al. (2015) IVX-estimated Wald test, and the proposed MC signed-rank tests with $M - 1 = 99$ for a given nominal level $\alpha = 5\%$.

Table 2: Empirical size of predictability tests: stochastic volatility dynamics

ϕ_{11}	$\rho_{x_1 r}$	θ	$\rho_{x_1 h}$	T	Wald	mARM	IVX	S^m_{min}	$S^m_\times$	W^m_{min}	$W^m_\times$	S_{min}	$S_\times$	W_{min}	$W_\times$
0.90	0	0.90	0	240	7.4	7.1	6.7	5.1	4.8	5.0	4.7	0.2	0.1	0.6	0.3
				720	6.0	6.2	6.6	5.0	4.1	6.1	5.9	0.2	0.1	0.5	0.5
0.90	0	0.90	0.25	240	8.6	8.4	8.3	5.5	5.0	4.7	4.4	0.1	0.1	0.4	0.3
				720	8.2	8.2	7.7	4.9	4.6	6.0	6.0	0.1	0.2	0.5	0.5
			0.50	240	11.8	11.1	11.6	5.9	5.5	4.9	4.5	0.2	0.1	0.5	0.3
				720	13.8	13.8	13.4	5.5	5.8	6.0	5.8	0.3	0.3	0.7	0.5
			0.70	240	15.7	14.6	14.9	7.0	6.9	5.6	4.7	0.4	0.2	0.4	0.3
				720	20.9	20.8	20.1	7.0	7.5	5.5	6.1	0.2	0.3	0.8	0.7
0.90	-0.90	0.90	0.25	240	5.9	5.8	3.8	4.6	5.1	4.9	5.0	0.2	0.3	0.5	0.4
				720	5.7	7.4	4.9	5.0	5.5	4.4	4.6	0.5	0.5	1.1	0.9
			0.50	240	7.3	7.8	6.1	6.0	5.5	5.0	5.5	0.3	0.2	0.7	0.7
				720	9.3	10.6	8.2	5.8	5.3	4.5	4.7	0.3	0.4	1.3	0.9
			0.70	240	12.2	11.4	9.7	7.2	7.2	5.9	5.9	0.2	0.2	0.7	0.6
				720	15.3	15.2	13.1	6.9	6.3	4.3	5.5	0.1	0.4	1.3	0.7
0.99	-0.90	0.99	0.25	240	8.5	7.6	5.8	5.8	5.8	5.7	4.9	0.1	0.2	0.6	0.6
				720	7.9	7.8	6.7	7.2	7.6	5.7	4.4	0.3	0.1	0.6	0.7
			0.50	240	14.2	11.6	9.8	6.2	6.1	4.8	4.8	0.1	0.3	0.8	0.7
				720	13.1	10.6	11.5	8.9	9.0	5.2	5.0	0.5	0.4	0.6	0.6
			0.70	240	21.5	17.7	16.8	7.0	8.5	4.8	4.4	0.2	0.2	0.7	0.6
				720	21.7	16.6	18.5	10.9	11.4	5.3	4.8	0.3	0.5	0.8	0.8

Notes: See Table 1.

Table 3: Power comparison of predictability tests: GARCH dynamics, $T = 240$

$\rho_{x_1x_2}$	β_1	β_2	Wald	mARM	IVX	S^m_{min}	$S^m_\times$	W^m_{min}	$W^m_\times$	S_{min}	$S_\times$	W_{min}	$W_\times$
GARCH													
0.0	-0.2	0.0	99.6	99.6	99.4	97.6	96.6	98.6	98.7	82.5	81.8	95.3	95.8
	-0.1	-0.1	99.2	99.6	99.1	82.1	83.8	87.5	93.5	49.6	60.2	76.7	82.5
-0.1	-0.2	0.0	99.6	99.7	99.5	97.7	97.1	99.0	99.1	83.0	82.4	95.2	95.1
	-0.1	-0.1	99.3	99.6	98.6	75.2	80.9	86.7	91.2	41.6	51.7	70.2	76.2
0.1	-0.2	0.0	99.7	99.8	99.4	96.7	95.1	98.9	98.9	83.2	81.3	95.1	94.8
	-0.1	-0.1	99.6	99.7	99.4	85.9	87.2	93.3	96.4	56.7	68.6	81.4	86.9
EGARCH													
0.0	-0.01	0.0	99.8	99.9	99.7	97.4	96.7	99.4	99.2	87.0	86.4	97.2	97.6
	-0.005	-0.005	99.0	99.5	98.1	78.0	78.8	86.4	89.2	46.0	56.4	71.5	78.0
-0.1	-0.01	0.0	99.8	99.9	99.7	97.2	96.0	99.4	99.5	86.7	87.1	96.9	97.3
	-0.005	-0.005	98.1	99.1	96.4	67.4	71.6	80.7	85.3	35.7	45.8	64.3	71.1
0.1	-0.01	0.0	99.8	99.9	99.7	96.9	96.3	99.5	99.4	86.8	86.4	96.9	96.9
	-0.005	-0.005	99.2	99.7	98.7	80.7	84.5	91.6	94.6	55.4	64.9	77.8	83.6

Notes: This table reports the power (in percentage) of the standard Wald test, the Amihud et al. (2009) mARM-based Wald test, the Kostakis et al. (2015) IVX-estimated Wald test, and the proposed MC signed-rank tests with $M - 1 = 99$ for a given nominal level $\alpha = 5\%$. All the tests are size-adjusted to ensure they respect the 5% level constraint, expect for the conservative S_{min}, $S_\times$, W_{min}, $W_\times$ tests which are exact.

Table 4: Power comparison of predictability tests: stochastic volatility dynamics

$\rho_{x_1 h}$	$\rho_{x_1 x_2}$	T	β_1	β_2	Wald	mARM	IVX	S^m_{min}	$S^m_\times$	W^m_{min}	$W^m_\times$	S_{min}	$S_\times$	W_{min}	$W_\times$
0.0	0.0	240	-0.2	0.0	48.7	57.1	49.8	90.5	90.9	85.9	83.7	69.6	70.2	66.8	66.3
			-0.1	-0.1	30.0	35.5	30.2	65.1	69.4	57.7	58.6	35.8	42.5	33.8	39.7
0.0	0.0	720	-0.2	0.0	82.9	85.9	81.7	100.0	100.0	100.0	99.9	99.9	99.8	99.6	99.6
			-0.1	-0.1	61.7	64.3	59.1	99.0	99.7	96.7	98.1	95.6	98.5	89.4	94.6
0.0	-0.1	240	-0.2	0.0	48.6	58.3	49.0	90.2	91.8	85.8	83.6	71.4	70.7	66.1	66.9
			-0.1	-0.1	25.8	34.4	27.2	57.6	63.7	52.6	52.1	30.5	37.0	27.8	34.3
0.0	-0.1	720	-0.2	0.0	83.5	85.8	82.1	100.0	100.0	100.0	100.0	99.8	99.7	99.5	99.4
			-0.1	-0.1	58.6	60.2	56.7	98.5	99.1	96.5	97.0	91.8	96.4	82.1	89.6
0.0	0.1	240	-0.2	0.0	48.2	56.9	50.7	89.4	89.4	83.8	84.9	70.8	70.8	68.3	66.3
			-0.1	-0.1	31.2	37.9	35.1	66.8	73.1	60.4	68.3	41.9	49.0	38.6	46.0
0.0	0.1	720	-0.2	0.0	82.2	84.9	81.5	100.0	100.0	99.9	99.8	99.9	99.9	99.4	99.2
			-0.1	-0.1	64.0	65.8	62.7	99.8	99.9	98.5	99.3	97.4	99.0	93.4	96.0
0.5	0.0	240	-0.2	0.0	39.5	51.6	39.2	85.3	85.5	80.6	80.8	59.9	61.2	61.1	61.4
			-0.1	-0.1	25.0	31.7	24.7	59.5	64.1	52.2	57.1	31.9	39.5	31.2	37.0
0.5	0.0	720	-0.2	0.0	71.5	77.7	69.3	100.0	100.0	99.8	99.8	99.5	99.5	99.0	98.8
			-0.1	-0.1	47.5	53.3	45.6	98.3	99.0	96.3	97.6	93.9	96.9	87.9	93.2
0.5	-0.1	240	-0.2	0.0	38.7	52.0	40.7	83.2	85.5	80.0	80.4	60.3	62.2	61.7	62.2
			-0.1	-0.1	22.7	30.8	23.6	51.0	57.3	47.5	49.6	27.6	33.4	26.5	32.7
0.5	-0.1	720	-0.2	0.0	72.4	77.9	71.0	100.0	100.0	99.9	99.9	99.1	99.2	99.0	98.8
			-0.1	-0.1	44.7	50.1	42.8	97.0	98.1	93.9	95.9	90.4	95.3	81.8	89.4
0.5	0.1	240	-0.2	0.0	41.0	51.5	40.8	83.2	85.5	77.5	80.1	59.3	61.8	60.3	61.0
			-0.1	-0.1	28.1	34.5	27.4	61.1	70.0	52.8	62.6	36.6	45.5	36.3	43.6
0.5	0.1	720	-0.2	0.0	71.6	76.7	69.1	99.9	99.9	99.8	99.7	99.1	99.1	98.9	98.7
			-0.1	-0.1	51.0	56.2	49.4	99.5	99.6	97.4	98.9	95.9	98.4	92.0	95.4

Notes: See Table 3.

Table 5: Summary statistics of employed variables

	r	d/p	e/p	b/m	dfy	tms	tbl
			Summary statistics				
Mean	0.01	-3.48	-2.75	0.54	0.01	0.02	0.04
Std dev	0.04	0.44	0.45	0.25	0.00	0.01	0.03
Autocorr.	0.03	1.00	0.99	0.99	0.97	0.96	0.99
			Correlation matrix				
r	1						
d/p	-0.003	1					
e/p	-0.009	0.780	1				
b/m	-0.024	0.884	0.816	1			
dfy	0.029	0.122	-0.028	0.255	1		
tms	0.051	-0.260	-0.361	-0.313	0.270	1	
tbl	-0.047	0.264	0.349	0.444	0.444	-0.421	1

Notes: This table presents the mean, standard deviation, first-order autocorrelation, and the correlations among the variables over the full-sample period from January 1948 to December 2014. The employed variables include excess returns (r), log dividend-price ratio (d/p), log earnings-price ratio (e/p), book-to-market ratio (b/m), default yield spread (dfy), term spread (tms), and short rate (tbl).

Table 6: Statistical properties of excess stock returns

	Mean	Std dev	Skewness	Kurtosis	JB	Autocorr.	$Q^2(6)$	$Q^2(12)$
67-year period								
Jan 1948 – Dec 2014	0.01	0.04	**-0.43**	**4.62**	**113.23**	0.04	**41.31**	**52.59**
10-year subperiods								
Jan 1948 – Dec 1957	0.01	0.04	-0.15	2.52	1.60	-0.03	4.41	11.84
Jan 1958 – Dec 1967	0.01	0.03	-0.54	**3.76**	**8.68**	0.09	**16.38**	17.03
Jan 1968 – Dec 1977	0.00	0.05	0.28	**4.12**	**7.82**	0.03	**23.92**	**28.93**
Jan 1978 – Dec 1987	0.01	0.05	**-0.68**	**5.90**	**51.24**	0.05	0.17	6.56
Jan 1988 – Dec 1997	0.01	0.03	-0.13	3.41	1.21	-0.17	**14.49**	**23.33**
Jan 1998 – Dec 2007	0	0.04	-0.53	**3.76**	**8.48**	0.02	**14.31**	**19.45**
Jan 2008 – Dec 2014	0.01	0.05	**-0.79**	**4.15**	**13.34**	0.19	**21.99**	**23.93**
20-year subperiods								
Jan 1948 – Dec 1967	0.01	0.04	-0.27	3.05	3.02	0.02	**11.81**	17.96
Jan 1968 – Dec 1987	0.00	0.05	-0.24	**5.04**	**43.90**	0.04	5.12	12.96
Jan 1988 – Dec 2014	0.01	0.04	**-0.59**	**4.17**	**37.13**	0.04	**49.51**	**54.55**

Notes: This table reports on the statistical properties of monthly excess returns from January 1948 to December 2014. In addition to the full 67-year sample, 10-year and 20-year subsamples are also considered. In each period, the sample skewness and kurtosis are tested against normally distributed data using the D'Agostino (1970) test and the Anscombe and Glynn (1983) test, respectively. JB refers to the Jarque-Bera normality test based on both the sample skewness and kurtosis. Finally, $Q^2(6)$ and $Q^2(12)$ are the Ljung-Box test statistics with 6 and 12 lags to test for serial dependence in return volatility. Bold face numbers indicate statistical significance at the nominal 10% level.

Table 7: Parameter estimates

	const.	d/p_{t-1}	e/p_{t-1}	b/m_{t-1}	dfy_{t-1}	tms_{t-1}	tbl_{t-1}	Adj. R^2
r_t	**0.099**	0.015	0.011	-0.025	0.712	0.096	**-0.163**	0.020
	(0.038)	(0.009)	(0.008)	(0.019)	(0.826)	(0.157)	(0.082)	
d/p_t	-0.062	**0.980**	0.005	0.019	-0.684	-0.095	0.055	0.990
	(0.040)	(0.010)	(0.008)	(0.020)	(0.854)	(0.164)	(0.087)	
e/p_t	**-0.237**	-0.028	**0.962**	**0.108**	**-5.519**	**0.842**	0.338	0.984
	(0.069)	(0.020)	(0.031)	(0.033)	(1.634)	(0.285)	(0.143)	
b/m_t	**0.099**	-0.004	0.000	**1.000**	0.712	-0.027	0.071	0.988
	(0.025)	(0.005)	(0.004)	(0.012)	(0.446)	(0.096)	(0.048)	
dfy_t	0.000	0.000	0.000	0.000	**0.971**	-0.004	0.003	0.945
	(0.001)	(0.000)	(0.000)	(0.001)	(0.035)	(0.005)	(0.003)	
tms_t	0.003	0.001	0.000	-0.002	**0.228**	**0.918**	-0.013	0.920
	(0.003)	(0.001)	(0.001)	(0.002)	(0.086)	(0.020)	(0.008)	
tbl_t	0.099	-0.001	0.000	0.003	**0.712**	**0.054**	**1.007**	0.983
	(0.004)	(0.001)	(0.001)	(0.002)	(0.101)	(0.027)	(0.010)	

Notes: This table presents the OLS parameter estimates from the multipredictor model over the sample period from January 1948 to December 2014. The predictors of excess stock returns (r) are log dividend-price ratio (d/p), log earnings-price ratio (e/p), book-to-market (b/m), default yield spread (dfy), term spread (tms), and short rate (tbl). The numbers in parentheses are the Newey-West adjusted standard deviations. Bold face numbers indicate significant t-statistics at 5% level.

Table 8: Residual correlation matrix

	r	d/p	e/p	b/m	dfy	tms	tbl
r	1						
d/p	-0.988	1					
e/p	-0.644	0.634	1				
b/m	-0.749	0.734	0.538	1			
dfy	-0.035	0.038	-0.150	-0.042	1		
tms	0.035	-0.031	0.019	-0.051	0.066	1	
tbl	-0.132	0.122	0.114	0.174	-0.262	-0.762	1

Notes: This table presents estimated correlation between the innovations of returns and the predictor variables over the sample period from January 1948 to December 2014. The predictors of excess stock returns (r) are log dividend-price ratio (d/p), log earnings-price ratio (e/p), book-to-market (b/m), default yield spread (dfy), term spread (tms), and short rate (tbl).

Table 9: Joint predictability tests using all six predictors

	S^m_{min}	$S^m_{\times}$	W^m_{min}	$W^m_{\times}$	S_{min}	$S_{\times}$	W_{min}	$W_{\times}$	Wald
67-year period									
Jan 1948 – Dec 2014	**0.04**	0.06	**0.02**	0.17	0.34	0.39	0.09	0.26	**0.00**
10-year subperiods									
Jan 1948 – Dec 1957	0.06	0.09	0.18	0.10	0.85	0.89	0.84	0.65	0.06
Jan 1958 – Dec 1967	0.10	0.17	**0.04**	**0.02**	0.92	0.95	0.97	0.98	0.07
Jan 1968 – Dec 1977	0.25	0.09	0.09	**0.02**	0.33	0.41	0.17	0.20	**0.00**
Jan 1978 – Dec 1987	0.33	0.24	0.22	0.17	0.56	0.56	0.29	0.43	**0.00**
Jan 1988 – Dec 1997	0.16	0.15	0.34	0.11	0.77	0.63	0.91	0.93	0.49
Jan 1998 – Dec 2007	0.06	0.07	0.34	0.30	0.95	0.92	0.88	0.79	**0.03**
Jan 2008 – Dec 2014	0.58	0.79	0.67	0.66	0.82	0.94	0.86	0.86	**0.02**
20-year subperiods									
Jan 1948 – Dec 1967	0.06	0.05	**0.02**	**0.01**	0.49	0.44	0.21	0.45	**0.04**
Jan 1968 – Dec 1987	0.21	0.08	**0.03**	0.06	0.55	0.55	0.29	0.17	**0.00**
Jan 1988 – Dec 2014	0.45	0.23	0.16	0.07	0.86	0.97	0.87	0.91	0.06

Notes: This table reports the p-values of the proposed non-parametric tests and the standard Wald test. The variables are defined at the monthly frequency from January 1948 to December 2014. Bold face numbers indicate joint significance at the nominal 5% level.

Table 10. Marginal p-values of each predictor in a univariate regression setup

	d/p_{t-1}	e/p_{t-1}	b/m_{t-1}	dfy_{t-1}	tms_{t-1}	tbl_{t-1}
S^m	0.41	0.48	0.08	1.00	**0.01**	0.27
W^m	0.58	0.78	0.40	0.49	**0.01**	0.20
S	0.99	0.99	0.82	0.97	0.08	0.99
W	1.00	0.98	0.95	1.00	**0.04**	0.99
Wald	**0.03**	0.11	0.33	0.61	0.05	**0.01**

Notes: This table shows the marginal p-values for each predictor obtained with the proposed tests and the standard Wald test. Bold face numbers indicate statistical significance at the nominal 5% level.

Table 11. Joint predictability tests with and without the term spread

	S^m_{min}	$S^m_\times$	W^m_{min}	$W^m_\times$	S_{min}	$S_\times$	W_{min}	$W_\times$
$K=6$	**0.04**	0.06	**0.02**	0.17	0.34	0.34	0.09	0.26
$K=5$	0.20	0.25	0.59	0.57	0.55	0.64	0.81	0.81

Notes: This table shows the p-values of the joint sign and signed rank tests. When $K=6$, the tests are based on all six predictors. The cases with $K=5$ is when the term spread (tms) is excluded and the joint predictability tests are performed with the remaining 5 predictors (d/p, e/p, b/m, dfy, tbl). Bold face numbers indicate joint significance at the nominal 5% level.

Figure 1: Monthly excess stock returns

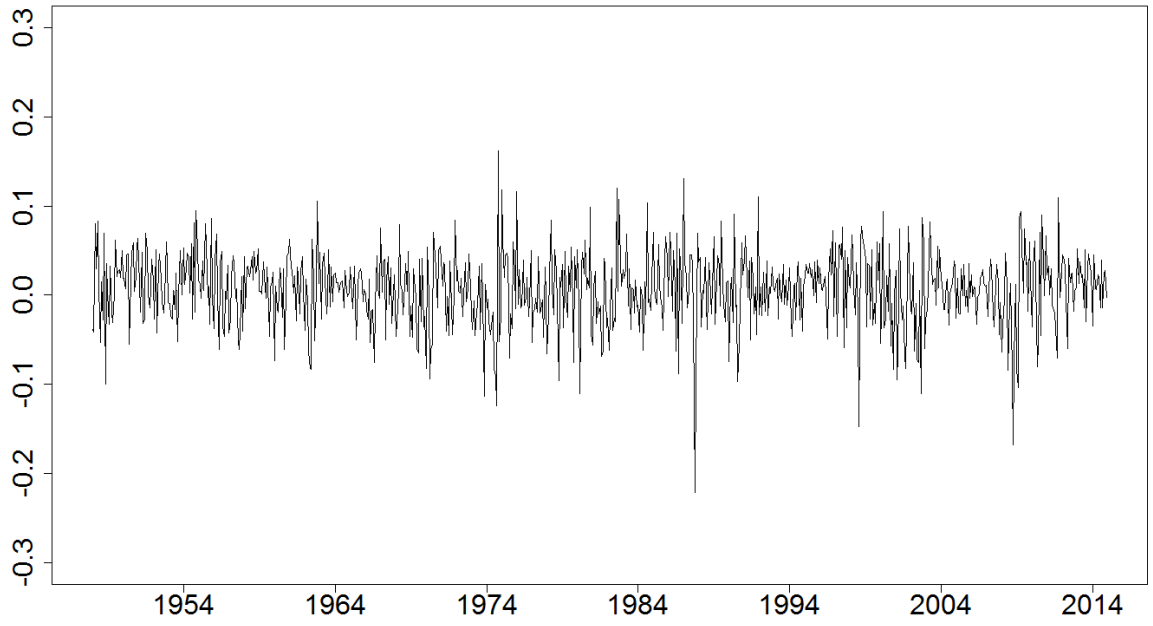

Figure 1 shows the monthly time series of excess returns on the S&P value-weighted index over the period from January 1948 to December 2014.

Figure 2: Monthly predictors

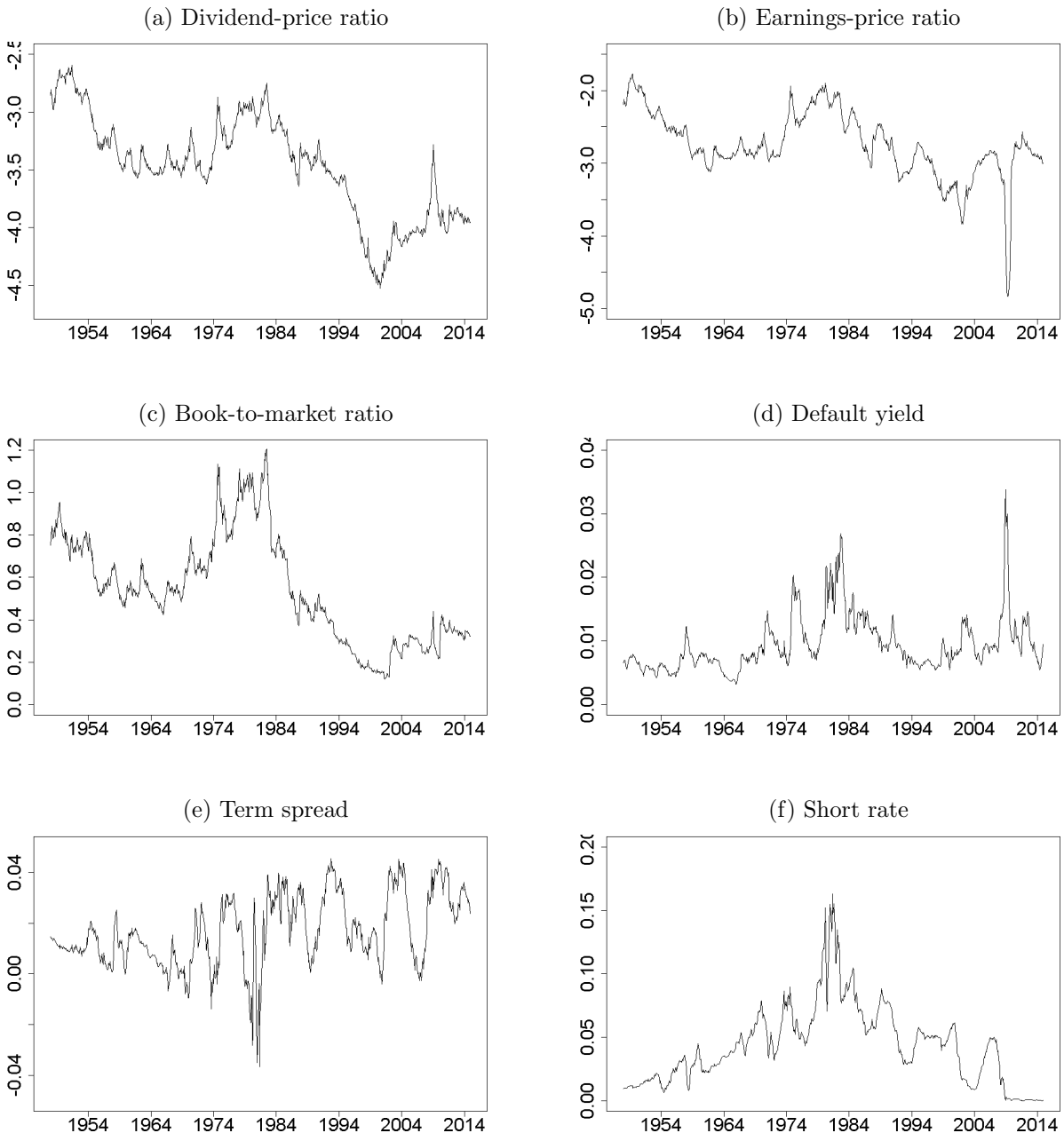

Figure 2 shows the monthly time series of the six predictors over the period from January 1948 to December 2014. Panels (a)–(f) show the log dividend-price ratio (d/p), the log earnings-price ratio (e/p), the book-to-market ratio (b/m), the default yield (dfy), the term spread (tms), and the short rate (tbl), respectively.

Figure 3: Rolling-window predictability tests with monthly excess returns

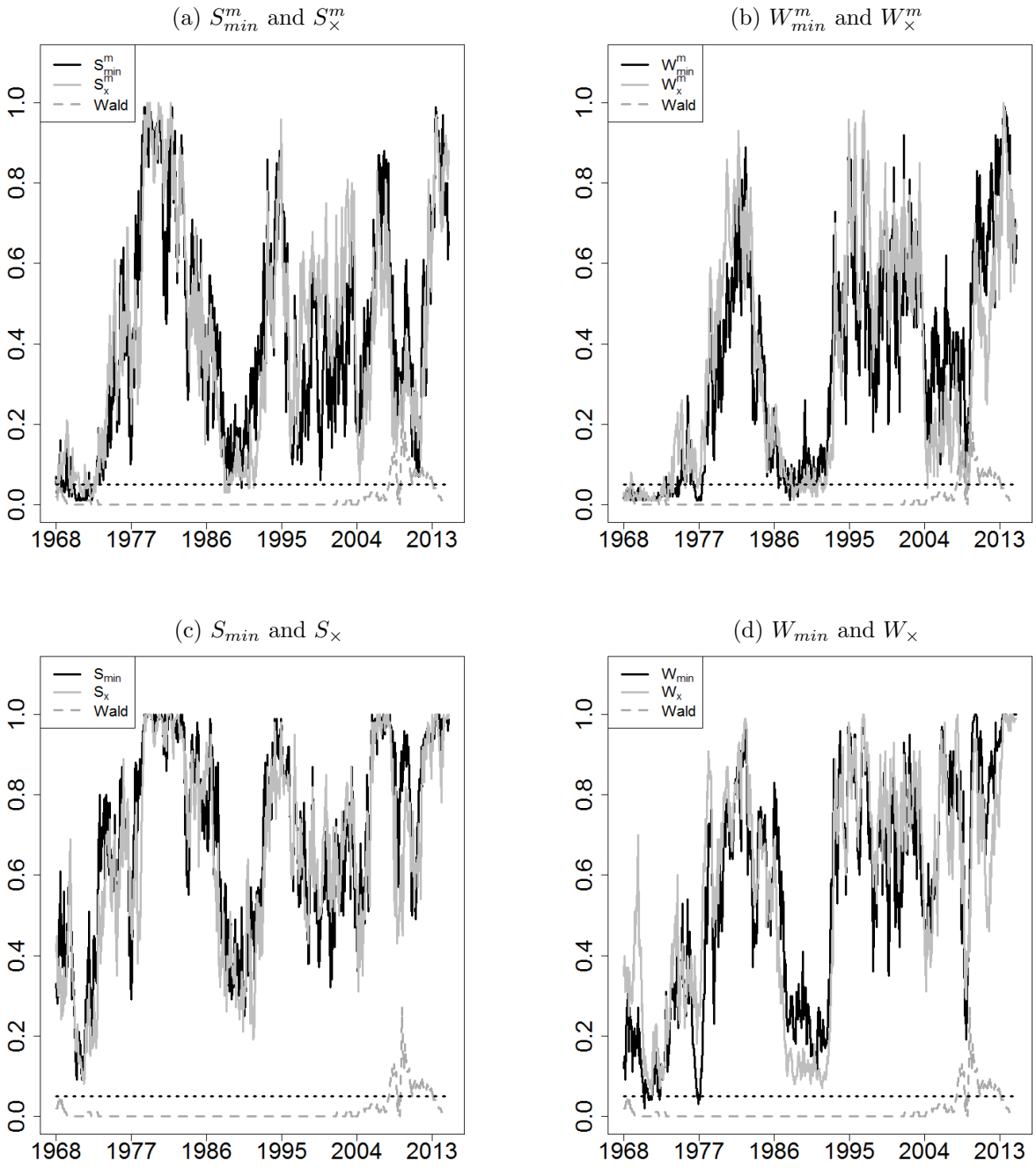

Figure 3 shows the p-values of the proposed sign and signed rank tests and the benchmark Wald test using a 240-month (20-year) rolling window. The solid black line indicates the tests based on the minimum p-value, the solid grey line is for the tests based on the product of the p-values, the dashed grey line is for the Wald test, and finally the horizontal dotted line shows the nominal 5% significance level.