BANK OF CANADA
BANQUE DU CANADA

Staff Working Paper/Document de travail du personnel 2017-10

Small-Sample Tests for Stock Return Predictability with Possibly Non-Stationary Regressors and GARCH-Type Effects

by Sermin Gungor and Richard Luger

Bank of Canada Staff Working Paper 2017-10

March 2017

Small-Sample Tests for Stock Return Predictability with Possibly Non-Stationary Regressors and GARCH-Type Effects

by

Sermin Gungor[1] and Richard Luger[2]

[1]Financial Markets Department
Bank of Canada
Ottawa, Ontario, Canada K1A 0G9
sgungor@bankofcanada.ca

[2]Département de finance, assurance et immobilier
Université Laval
Québec, Quebec, Canada G1V 0A6
richard.luger@fsa.ulaval.ca

ISSN 1701-9397

Acknowledgements

We would like to thank Gregory Bauer, Antonio Diez de los Rios and seminar participants at the Bank of Canada, 2015 Computational Financial Econometrics Conference, and 2016 International Association for Applied Econometrics Conference for valuable comments. The views expressed in this paper are those of the authors. No responsibility for them should be attributed to the Bank of Canada.

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 both in a simulation study and by examining the ability of a group of financial variables to predict excess stock returns. We find robust evidence of predictability during the period 1948–2014, driven entirely by the term spread. This empirical evidence, however, is much weaker over subsamples.

Bank topics: Econometric and statistical methods; Asset pricing; Financial markets
JEL codes: C12; C32; G14

Résumé

Nous développons une méthode de simulation pour tester la prévisibilité du rendement des actions à l'aide de multiples variables de régression. Le processus déterminant les variables de régression n'est aucunement restreint et la méthode de simulation reste valide à distance finie même en présence de distributions autres que la loi normale et d'effets GARCH sur le rendement des actions. L'utilité de la nouvelle méthode est démontrée à la fois dans une étude de simulation et par l'examen de la capacité d'un ensemble de variables financières à prévoir le rendement excédentaire des actions. Nous observons, pour la période 1948-2014, des signes probants de prévisibilité qui s'expliquent entièrement par l'écart de taux. Toutefois, ces résultats empiriques sont beaucoup plus faibles dans le cas des sous-échantillons.

Sujets : Méthodes économétriques et statistiques; Évaluation des actifs; Marchés financiers
Codes JEL : C12, C32, G14

Non-Technical Summary

A long-standing question in finance is whether asset returns can be predicted by economic and financial variables. This question has important and broad economic implications. However, the robustness of the evidence on asset return predictability remains controversial. A common practice in the literature is to estimate an ordinary least squares (OLS) regression of asset returns on the lagged values of the predictor variable under study. Such predictive regressions are then evaluated using a t-test, which often appears significant when compared to traditional critical values. As a result, the prevailing tone in the literature is that asset returns are predictable using financial and economic variables.

Common features of the predictability regressions are the feedback from returns to the future values of the predictor variable and the persistent behavior of the predictor variable. The problem in this case is that the t-statistic often rejects the null hypothesis of no predictability much too often. This problem has generated substantial interest in both econometrics and empirical finance, and a number of econometric solutions have been proposed. All of these proposed approaches, however, depend on a very specific model for the predictor variable.

In sharp contrast, in this study, we propose a simulation-based procedure without any modelling assumptions being imposed on the predictor variable. In addition, this new procedure does not impose any parametric assumptions on the distribution of stock return innovations, and it can be applied for hypothesis testing in predictive regressions for multiple-predictor models.

Our simulation experiments reveal that the proposed simulation-based test procedure

has the correct rejection rate (size) and can be more powerful than the extant tests. We apply the developed procedure to test the predictability of S&P 500 value-weighted index using six widely used predictors, i.e., the dividend-price ratio, the earnings-price ratio, the book-to-market ratio, the default yield, the term spread, and the short rate. Our empirical application indicates robust evidence of stock return predictability. The takeaway message is that, among the six predictors, only the term spread has predictive ability for excess returns.

1 Introduction

A long-standing question in finance is whether asset 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 asset 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 asset 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 is persistent (Mankiw and Shapiro, 1986; Stambaugh, 1999). The problem in this case is that 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 without any modelling assumptions whatsoever for the regressor variable. 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 for structural breaks, time-varying parameters, and other unmodelled non-linearities in the regressor process 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-sample non-parametric statistics show that such tests are the *only* tests that yield valid inference when one wishes to remain completely agnostic about distribution heterogeneities (Lehmann and Stein, 1949). Furthermore, the non-parametric tests of Campbell and Dufour (1997) can be more powerful

than the size-corrected t-test.

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.[1] The economic motivation underlying predictive regression is controversial. The efficient markets hypothesis view argues that predictability of asset returns indicates inefficiencies in the capital markets. The alternative view interprets return predictability as consistent with an efficient capital market where the returns reflect time-varying expected returns. Regardless of the interpretation, asset predictability should be evaluated based on *all* past information. 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 overall 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 of multiple regressors, since there can be feedback from the return innovations to future

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

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 (2014), 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. Our simulation experiments further reveal that the proposed non-parametric Monte Carlo test procedure can be more powerful than the size-corrected Wald, mARM, and IVX test procedures.

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 both monthly and quarterly 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. Using the new test procedure, we find more trustworthy evidence of stock return predictability at both the monthly and quarterly frequencies in the full-sample period. Tests of the marginal significance reveal that among the six regressors, only the term spread has predictive ability for both monthly and quarterly excess stock returns. The takeaway message is that while the new joint tests reveal robust evidence of stock return predictability, this evidence is entirely driven by the term spread. This empirical evidence, however, turns out to be much weaker over the monthly and quarterly subsamples. These results suggest that test power depends more on the span of the data rather than the number of observations.

The 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 in period t and a $K \times 1$ 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. In our framework, there are no restrictions on the number K of potential predictors.

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

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

and we also develop tests under the stronger assumption:

$$\varepsilon_t \text{ is symmetric around zero, given } \mathcal{I}_{t-1}. \tag{3}$$

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 provide an assessment of H_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).[2] 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

[2]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 methods used in those papers, however, can only handle a single predictor, whereas our tests allow for multiple predictors.

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. It is interesting to note that OLS-based procedures can only be justified by assuming that ε_t has well-defined first and second moments, while no such moment assumptions are needed here.

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 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 $\sigma_t > 0$ capturing conditional heteroskedasticity is a measurable function of $\mathcal{I}_{t-1}$. Of course, a far wider class with asymmetric innovations can be entertained under (2). Here, the process governing the dynamics of σ_t 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

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

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 $s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and 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 $\boldsymbol{\sigma}_{\varepsilon \boldsymbol{v}} \neq \boldsymbol{0}$, in which case there is feedback from 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 that yields a reduced-biased 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 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. 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 $g_{i,t} = g_{i,t}(\mathcal{I}_t)$, $t = 0, ..., T-1$, 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 (which amounts to considering a test based on the signs 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 $g_{i,t} = x_{i,t}$ owing to power considerations. Indeed, this choice makes the median of $(r_t - \beta_0)x_{i,t-1}$ a function of $\beta_i x_{i,t-1}^2$ under the alternative hypothesis that $\beta_i \neq 0$. The power of $S_i(\beta_0)$ and $SR_i(\beta_0)$ will therefore tend to increase as the magnitude of β_i increases. 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. From now on, let the symbol "$\stackrel{d}{=}$" stand for the equality in distribution.

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

(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 the additional 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 $i = 1, ..., K$

Here $\tilde{\varepsilon}_1, ..., \tilde{\varepsilon}_T$ *are independent median-zero random variables, i.e.,* $\Pr(\tilde{\varepsilon}_t \geq 0) = \Pr(\tilde{\varepsilon}_t < 0) = 1/2$ *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* $t = 1, ..., T$.

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 predictive ability in the direction of x_i. To obtain power against all x_is, 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

$$p^{\mathcal{S}}_{min}(\beta_0) = \min\left\{p^{\mathcal{S}}_1(\beta_0), ..., p^{\mathcal{S}}_K(\beta_0)\right\} \text{ and } \mathcal{S}_{min}(\beta_0) = 1 - p^{\mathcal{S}}_{min}(\beta_0), \tag{8}$$

then we reject H_0 when $p^{\mathcal{S}}_{min}(\beta_0)$ 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:

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

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 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.[3] 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.

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 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.[4] Indeed, the MC procedure implicitly

[3]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) $\leq \alpha$.

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

accounts for the fact that the individual p-values may not be exact and yields an overall p-value for the combined statistic that itself is exact. The key is the following result, which is an immediate corollary of the first proposition.

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

(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*

$$\begin{aligned} S_{min}(\beta_0) &\overset{d}{=} \widetilde{S}_{min}(\beta_0) = 1 - \min\left\{\tilde{p}_1^S(\beta_0), ..., \tilde{p}_K^S(\beta_0)\right\}, \\ S_{\times}(\beta_0) &\overset{d}{=} \widetilde{S}_{\times}(\beta_0) = 1 - \prod_{i=1}^{K} \tilde{p}_i^S(\beta_0), \end{aligned}$$

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

$$\begin{aligned} \tilde{p}_i^S(\beta_0) &= 2\big(1 - \Phi(|\widetilde{S}_i^*(\beta_0)|)\big), \\ \widetilde{S}_i^*(\beta_0) &= \frac{\widetilde{S}_i(\beta_0) - T/2}{\sqrt{T/4}}, \\ \widetilde{S}_i(\beta_0) &= \sum_{t=1}^{T} s\big[\tilde{\varepsilon}_t g_{i,t-1}\big]. \end{aligned}$$

(ii) If H_0 *and the additional 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

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

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

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

$$\begin{aligned}
\tilde{p}_i^{SR}(\beta_0) &= 2\big(1 - \Phi(|\widetilde{SR}_i^{*}(\beta_0)|)\big), \\
\widetilde{SR}_i^{*}(\beta_0) &= \left[\widetilde{SR}_i(\beta_0) - \frac{1}{2}\sum_{t=1}^{T}\varphi(t)\right] \Bigg/ \sqrt{\frac{1}{4}\sum_{t=1}^{T}\varphi^2(t)}, \\
\widetilde{SR}_i(\beta_0) &= \sum_{t=1}^{T} s\big[\tilde{\varepsilon}_t g_{i,t-1}\big]\varphi\big(R_t^{+}(\beta_0)\big).
\end{aligned}$$

Here again $\tilde{\varepsilon}_1, ..., \tilde{\varepsilon}_T$ *are independent median-zero random variables, i.e.,* $\Pr(\tilde{\varepsilon}_t \geq 0) = \Pr(\tilde{\varepsilon}_t < 0) = 1/2$ *for* $t = 1, ..., T$*; and* $B_1, ..., B_T$ *are independent Bernoulli variables such that* $\Pr(B_t = 1) = \Pr(B_t = 0) = 1/2$, $t = 1, ..., T$.

This proposition shows how to obtain the building blocks $\widetilde{S}_i(\beta_0)$ and $\widetilde{SR}_i(\beta_0)$, for $i = 1, ..., K$. Note that the simulated terms $\tilde{\varepsilon}_t$ may simply be set as i.i.d. according to *any* continuous median-zero distribution like the standard normal, for example. An important remark about these results 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 $s\big[\tilde{\varepsilon}_t g_{1,t-1}\big], ..., s\big[\tilde{\varepsilon}_t g_{K,t-1}\big]$ 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 indi-

vidual 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.,

$\Pr\left(\tilde{R}_M[\mathcal{S}_\bullet(\beta_0)] = j\right) = 1/M$, for $j = 1, ..., M$. So the MC p-value can be defined as

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

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

$$\Pr\left(\tilde{p}_M\left[\mathcal{S}_\bullet(\beta_0)\right] \leq \alpha\right) = \alpha, \tag{12}$$

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 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).

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. 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 that 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.[5] Let $CI_{\beta_0}(\alpha_1)$ denote a confidence interval for β_0 with level $1-\alpha_1$, i.e., such that $\Pr\left(\beta_0 \in CI_{\beta_0}(\alpha_1)\right) \geq 1-\alpha_1$ under H_0. From Bonferroni's inequality, we then have the following results (see the Appendix for the proof).

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

(i) Suppose further that H_0 *and Assumption (2) are satisfied, and* $CI_{\beta_0}(\alpha_1)$ *is a confidence interval for* β_0 *such that* $\Pr\left(\beta_0 \in CI_{\beta_0}(\alpha_1)\right) \geq 1-\alpha_1$ *under* H_0 *and Assumption (2). If* αM *is an integer, then, given* $g_{i,0}, ..., g_{i,T-1}$, $i = 1, ..., K$, *the critical region* $\sup_{b \in CI_{\beta_0}(\alpha_1)} \tilde{p}_M\left[S_\bullet(b)\right] \leq \alpha_2$ *is such that*

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

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

(ii) Suppose furthermore that H_0 *and Assumption (3) are satisfied, and* $CI_{\beta_0}(\alpha_1)$ *is a confidence interval for* β_0 *such that* $\Pr\left(\beta_0 \in CI_{\beta_0}(\alpha_1)\right) \geq 1-\alpha_1$ *under* H_0 *and Assumption (3). If* αM *is an integer, then, given* $g_{i,0}, ..., g_{i,T-1}$, $i = 1, ..., K$, *and* $|r_1 - b|, ..., |r_T - b|$,

[5]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).

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, p. 92–96) and Hettmansperger (1984, p. 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 though, 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, p. 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 simulation experiments 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. A useful remark for practical applications is that 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 $\tilde{p}_M[S_{min}(\check{\beta}_0)] \leq \alpha_2$ then $\sup_{b \in CI_{\beta_0}(\alpha_1)} \tilde{p}_M\big[S_{min}(b)\big] \leq \alpha_2$ and H_0 is not significant at the overall level α.

The data-generating process is the system in (4) comprising the predictive regression model with two potential predictors governed by a VAR model. For convenience, the complete specification is given here as

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

for $t = 1, ..., T$, and the recursion is started with $(x_{1,0}, x_{2,0})' = (\mu_1, \mu_2)' + (v_{1,0}, v_{2,0})'$. The vectors $\boldsymbol{\epsilon}_t = (\eta_t, v_{1t}, v_{2t})'$ are i.i.d. over time according to a multivariate normal or Student-t distribution with 3 degrees of freedom. In both cases, the multivariate innovation distribution (denoted by $\mathcal{D}$) has mean zero and contemporaneous scale matrix given by

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

The parameter $\rho_{x_1 r}$ controls the strength of feedback from η_t to future values of the regressors appearing on the right-hand side of the predictive regression in (14), and $\rho_{x_1 x_2}$ is the innovation correlation between the two predictor variables. 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 parameters of the VAR component in (14) are set as $\mu_1 = \mu_2 = 0$, $\phi_{12} = \phi_{21} = 0$, and the other parameters are varied to examine their effects.

We consider two cases for the conditional variance of returns. In the i.i.d. case, we have $\sigma_t = 1$ so the predictor variables may only affect the conditional location (mean and median) of r_t. Next we allow for stochastic volatility effects of the form $\sigma_t = \exp(x_{2,t-1}/100)$. The exponential function guarantees that $\sigma_t > 0$ and the division by 100 is simply a scaling factor. In this conditional heteroskedasticity ("het") case, one of the predictor variables has predictive ability for the volatility (uncertainty) of returns, even though it may not predict mean returns. To see the plausibility of this specification, consider the "het" model under $\beta_1 = \beta_2 = 0$ (the null hypothesis). The τth conditional quantile of r_t is then given by

$$Q_\tau(r_t \,|\, \mathcal{I}_{t-1}) = \beta_{0,\tau} + Q_\tau(\varepsilon_t \,|\, \mathcal{I}_{t-1}),$$

where $Q_\tau(\varepsilon_t \,|\, \mathcal{I}_{t-1}) = \exp(x_{2,t-1}/100) Q_\tau(\eta_t)$ and $Q_\tau(\eta_t)$ is the quantile of the innovation η_t at a given quantile level $\tau \in (0, 1)$. In this case, the conditional median of r_t equals β_0 no matter the predictors, since $Q_{0.5}(\varepsilon_t \,|\, \mathcal{I}_{t-1}) = Q_{0.5}(\eta_t) = 0$ under (2). Furthermore, $Q_{0.5}(r_t \,|\, \mathcal{I}_{t-1}) = E(r_t \,|\, \mathcal{I}_{t-1}) = 0$ under (3). Note, however, that $x_{2,t-1}$ influences the outer quantiles of r_t since $|Q_\tau(\varepsilon_t \,|\, \mathcal{I}_{t-1})| = \exp(x_{2,t-1}/100)|Q_\tau(\eta_t)|$ is an increasing function of $|\tau - 0.5|$.[6] This setup is motivated by the empirical evidence in Cenesizoglu and Timmermann (2008), Maynard et al. (2010), and Lee (2016) who find that many economic variables have far greater predictive ability for the outer quantiles of the return distribution compared to the centre of the return distribution.

We provide results for selected subsets of the cases for which $\phi = 0.95$, 0.99, 1; $\rho_{x_1 r} = 0$,

[6]Here $x_{2,t-1}$ has predictive ability for the conditional variance of r_t, which is close to the notion of Granger causality *in volatility* (Granger et al., 1986). In turn, such a mechanism implies the (outer) quantile predictability.

-0.9, -0.99; $\rho_{x_1x_2} = 0$, -0.1, 0.1; and $T = 100$, 200. In each case, the reported results are based on 1000 simulation replications of the data-generating configuration.

Given a desired level α, Proposition 3 shows that there is a tradeoff 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. Table 1 illustrates this tradeoff for different testing strategies with α_1, α_2 taking values $1, ..., 4$ such that $\alpha = 5\%$. These results for the MC signed-rank tests S_{min}, $S_\times$, W_{min}, $W_\times$ were obtained by generating data according to (14) with $\beta_0 = \mu_1 = \mu_2 = 0$; $\phi_{12} = \phi_{21} = 0$; $\phi_{11} = \phi_{22} = 0.95$; $\rho_{x_1r} = -0.9$; $\rho_{x_1x_2} = 0$. The sample size is $T = 200$, the innovations $\boldsymbol{\epsilon}_t$ are i.i.d. according to a trivariate Student-t distribution with 3 degrees of freedom, and $\sigma_t = 1$.

The results in Table 1 suggest 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. Indeed, by shrinking α_1, there is a clear gain in power from 20 to nearly 30 percentage points depending on the test. Similar results were obtained by Campbell and Dufour (1997) in a single predictor context. We therefore carry on with the testing strategy represented by $\alpha_1 = 1\%$, $\alpha_2 = 4\%$.

The size and power results are presented in Tables 2 and 3, respectively. 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 in Table 2 are summarized as follows.

1. All the tests respect the nominal 5% level constraint in the i.i.d. case with no feedback ($\rho_{x1r} = 0$). When departing from that case, however, the Wald, mARM, and IVX tests

tend to over-reject. This problem is abundantly clear in the presence of conditional heteroskedasticty (the "het" cases) where the Wald, mARM, and IVX tests have rejection rates anywhere between 20% and 30%. Note also that doubling the sample size from $T = 100$ to $T = 200$ does not improve matters for the Wald, mARM, and IVX tests in these conditional heteroskedasticty cases.

2. The over-rejection problem with the Wald test tends to be exacerbated when: (i) there is an increase in the strength of feedback from $\rho_{x_1 r} = 0$ to $\rho_{x_1 r} = -0.9$ to $\rho_{x_1 r} = -0.99$; and (ii) there is an increase in regressor persistence from $\phi_{11} = 0.95$ to $\phi_{11} = 0.99$ to $\phi_{11} = 1$. The mARM-based Wald test appears more robust than the Wald test to the presence of feedback, but nevertheless tends to become oversized when the regressor persistence reaches $\phi_{11} = 0.99$ and $\phi_{11} = 1$. The IVX test is relatively more robust than the mARM test to increases in regressor persistence. Interestingly, the empirical size of the Wald, mARM, and IVX tests is almost the same whether the innovations are i.i.d normal or i.i.d. t_3.

3. The plug-in S^m_{min}, $S^m_{\times}$, W^m_{min}, $W^m_{\times}$ tests appear well behaved with empirical size close to the nominal 5% level in the i.i.d. cases regardless of the innovation distribution (normal or t_3), the regressor persistence (ϕ_{11}), or the strength of feedback ($\rho_{x_1 r}$). This is perhaps not surprising given that the sample median is a consistent estimator of β_0 in these i.i.d. cases (cf. Mizera and Wellner, 1998). In the "het" cases, however, the plug-in tests display slight over-rejections. For instance, S^m_{min} and $S^m_{\times}$ tests have empirical size close to 10%. Obviously this is nowhere near as bad as the behaviour of the Wald, mARM, and IVX tests in the "het" cases.

4. The S_{min}, $S_{\times}$, W_{min}, $W_{\times}$ tests based on a first-stage confidence interval for β_0 are the only tests that are completely robust (i.e., invariant) to the strength of feedback, the degree of regressor persistence, and the presence of non-normalities and conditional heteroskedasticity. Indeed, the empirical size of these conservative tests is always strictly less than the nominal 5% significance level, in accordance with the developed theory.

Given the size distortions seen in Table 2, the power results for the Wald, mARM, IVX, and even the S^m_{min}, $S^m_{\times}$, W^m_{min}, $W^m_{\times}$ tests are based on size-corrected critical values. Such adjustments were not applied to the S_{min}, $S_{\times}$, W_{min}, $W_{\times}$ tests, since the probability of a Type I error with these tests is $\leq \alpha$. The main findings that emerge from Table 3 can be summarized as follows.

1. The Wald, mARM, and IVX tests have the best power among all the tests in the i.i.d. settings, with the mARM test outperforming the Wald and IVX tests. Of course, power improves for all the tests as the sample size increases. It is interesting to observe that the power of the Wald, mARM, and IVX tests remains about the same whether the innovations are i.i.d. normal or i.i.d. t_3. On the contrary, the power of the signed rank tests improves dramatically as the innovation tail-heaviness increases.

2. Indeed, the signed rank tests do much better than the Wald, mARM, and IVX tests in the "het" cases. In fact, we see that the presence of conditional heteroskedasticity diminishes the relative power of the Wald, mARM, and IVX tests. It is remarkable that even the conservative S_{min}, $S_{\times}$, W_{min}, $W_{\times}$ tests outperform the Wald-mARM-IVX group, often by a wide margin. The conventional wisdom that non-parametric tests

perform well in the presence of heavy tails is thus corroborated. Notice as well that the power performance improves considerably when the plug-in S^m_{min}, $S^m_\times$, W^m_{min}, $W^m_\times$ tests are used instead of the two-stage tests.

3. An examination of the S_{min}, $S_\times$, W_{min}, $W_\times$ tests shows that they tend to suffer when the correlation between regressor innovations ($\rho_{x_1x_2}$) becomes negative and to benefit when this correlation increases. Comparing the $\mathcal{S}_{min}$ and $\mathcal{S}_\times$ tests, we see that when $\beta_1 = \beta_2$ the signed rank tests perform somewhat better if they are combined using the product of the marginal p-values rather than the minimum p-value. This can be gleaned at once from the last line of Table 3, for instance.

5 Empirical results

To further illustrate the new test procedure, we examine the predictability of excess stock returns using U.S. data. Our empirical investigation uses six predictors that are widely used in the stock return predictability literature: (i) the dividend-price ratio, (ii) the earnings-price ratio, (iii) the book-to-market ratio, (iv) the default yield spread, (v) the term spread, and (vi) the short rate. These data are in fact a subset of those used by Welch and Goyal (2008) and updated through the year 2014. Here we consider monthly and quarterly data obtained from Amit Goyal's website for the 67-year time span from January 1948 to December 2014. Earlier studies point out that the predictive power of the employed variables may not be robust over time (Lettau and Ludvigson, 2001; Welch and Goyal, 2008). So in addition to the entire 67-year period, we also examine return predictability over fixed 10-year and 20-year

subsamples[7] and 20-year rolling window subsamples. A brief description of the variables used is given below; see Welch and Goyal (2008) for full details.

- *Excess returns:* Excess stock return (r) is defined as the rate of return on the S&P 500 value-weighted index minus the 3-month Treasury bill rate.

- *Dividend-price ratio:* The d/p predictor is the natural logarithm of the dividend-price ratio, where dividends are 12-month moving sums of dividends paid on the S&P 500 index.

- *Earnings-price ratio:* The e/p predictor is the natural logarithm of earnings-price ratio, where earnings are 12-month moving sums of earnings on the S&P 500 Index.

- *Book-to-market ratio:* The *btm* predictor is the ratio of book value to market value for the Dow Jones Industrial Average.

- *Default yield spread:* The *dfy* predictor is the difference between BAA and AAA-rated corporate bond yields.

- *Term spread:* The *tms* predictor is defined as the difference between the long-term yield on government bonds and the 3-month Treasury bill rate.

- *Short rate:* The *tbl* predictor is the short-term interest rate, taken as the 3-month Treasury bill rate.

Figure 1 plots the time-series of excess returns at the monthly (panel a) and quarterly (panel b) frequencies. The monthly time-series plots of the six predictors are presented in

[7]The last fixed subsample is slightly shorter due to the time span covered by the data.

Figure 2 and the quarterly ones are shown in Figure 3. 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 panels (a) and (b) of Table 4, which report some summary statistics (mean, standard deviation, first-order autocorrelation) and the correlations among the variables at the monthly and quarterly frequencies. The autocorrelation of excess stock returns is near zero, whereas the predictors are highly persistent, with autocorrelations close to one. Even at the quarterly frequency, the predictors remain very persistent. Indeed, it is only the default yield (*dfy*) and the term spread (*tms*) that appear somewhat less persistent, with autocorrelation coefficients 0.88 and 0.84, respectively. The autocorrelation of the other predictors is at least 0.95 at either frequency.

Table 5 summarizes some of the distributional properties of the monthly and quarterly 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 5, bold entries indicate statistical significance at the 10% level.

Over the full sample period, there is some evidence of negative skewness in both the monthly and quarterly return data. In the 10-year and 20-year subsamples, however, the

evidence inidicates that returns are symmetrically distributed. The monthly stock returns tend to be heavy-tailed, both in the full sample and the subsamples. In contrast, the quarterly returns exhibit relatively less kurtosis. Finally, the Ljung-Box tests clearly indicate the presence of conditional heteroskedasticity at the monthly frequency and less so at the quarterly 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 6 the parameter estimates from the system of equations in (4) along with Newey-West standard errors in parentheses. The first column in panels (a) and (b) show the one-month and one-quarter ahead predictive regression results, respectively. The remaining columns 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. First, notice that based on the Newey-West adjusted t-statistics, only the short rate appears to be a significant predictor of stock returns. Looking at the persistence estimates along the main diagonal in panel (a), we see that the predictors are highly persistent with autoregressive coefficients between 0.918 and 1.007. They appear relatively less persistent at the quarterly frequency, with autocorrelation coefficients between 0.768 and 0.997 in panel (b).

Table 7 shows the estimated residual correlations from model (4). The first column showing the residual correlations between the stock returns and the predictors gives an indication of the strength of feedback. Not surprisingly, financial ratios such as d/p, e/p, b/m are highly and negatively correlated with returns since the stock price appears in the denominator of these ratios. Observe also that the conditional correlations in Table 7 are much higher than their unconditional counterparts in Table 4. The results in Tables 6

and 7 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 along the main diagonal in Table 6 combined with the strong residual correlations shown in Table 7 is an early warning that the Wald test may not be reliable.

Table 8 reports the results from the application of the developed sign and signed rank 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 p-values essentially equal to zero, both at the monthly and quarterly frequency. The plug-in S^m_{min} and W^m_{min} tests tend to agree with the Wald test with p-values $\leq 8\%$. On the contrary, the two-stage tests disagree with those rejections, except for the W_{min} test, which has p-values $\leq 9\%$ using the full 67-year sample. The subsample analysis reveals a more nuanced picture. Indeed, the plug-in tests show some evidence of predictability in the 20-year subperiods, but hardly any with the 10-year subperiods, and the two-stage tests 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.

More striking yet are the 20-year rolling-window predictability test results in Figures 4 and 5. 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 indicate 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 the horizontal dotted line shows the nominal 5% significance level. The figures clearly show the tendency of the Wald test to almost always reject the null, except

during the most recent times.[8] 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 the late 1980s to early 1990s; however, this is not supported by the two-stage tests.

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 9 reveal that, among the six regressors, it is only the term spread (*tms*) that has predictive ability for both monthly and quarterly excess stock returns. Another way to see this is from Table 10. 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 10–the p-values of the joint predictability tests cease to reject the null hypothesis.

The evidence uncovered here about the strong 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 robust evidence of stock return predictability at both the monthly and quarterly

[8]More specifically, at the 5% level, the Wald test rejects the null 503 times (155 times) out of the 564 monthly (188 quarterly) rolling regressions.

frequencies, this evidence is entirely driven by the term spread. Moreover, the empirical evidence of predictive ability does not consistently hold up over subsamples. 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.[9]

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 GARCH-type or stochastic volatility 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 in controlling test size under such conditions.

In this paper, we have developed a small-sample testing procedure that is truly robust (i.e., 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 config-

[9]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.

urations. This is 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 overall significance level. It is important to remember that 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 remarkable 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.

Appendix: Proofs

Proof of Proposition 1: (i) Suppose model (1) holds, and let $g_{i,t} = g_{i,t}(\mathcal{I}_t)$, $i = 1, ..., K$, $t = 0, ..., T-1$, be a sequence of measurable functions of $\mathcal{I}_t$ such that $\Pr(g_{i,t} = 0) = 0$ for $i = 1, ..., K$ and $t = 0, ..., T-1$. 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 $\Pr(\tilde{\varepsilon}_T \geq 0) = \Pr(\tilde{\varepsilon}_T < 0) = 1/2$ 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

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

where $\tilde{\varepsilon}_1,...,\tilde{\varepsilon}_T$ are mutually independent random variables such that $\Pr(\tilde{\varepsilon}_t \geq 0) = \Pr(\tilde{\varepsilon}_t < 0) = 1/2$; 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 $S_i(\beta_0) \overset{d}{=} \sum_{t=1}^{T} s\big[\tilde{\varepsilon}_t g_{i,t-1}\big] \overset{d}{=} \sum_{t=1}^{T} B_t$, for each $i = 1,...,K$, since $\mathbf{X} \overset{d}{=} \mathbf{Y} \Rightarrow f(\mathbf{X}) \overset{d}{=} f(\mathbf{Y})$ 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

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

Conditional on $|r_1 - \beta_0|,...,|r_T - \beta_0|$, the vector of scores $\big(\varphi(R_1^+(\beta_0)),...,\varphi(R_T^+(\beta_0))\big)$ is a fixed permutation of $\big(\varphi(1),...,\varphi(T)\big)$. So under the conditions of part (i) and given $|r_1 -$

$\beta_0|, ..., |r_T - \beta_0|$, we have that

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

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 $(s_{i,1}, ..., s_{i,T})$. Therefore, it is also the case unconditionally 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 $i = 1, ..., K$, since the distribution of $\sum_{t=1}^{T} B_t\varphi(t)$ does not depend on $|r_1 - \beta_0|, ..., |r_T - \beta_0|$.

Proof of Proposition 2: Follows immediately upon recognizing that the elements of $\big\{p_1^S(\beta_0), ..., p_K^S(\beta_0)\big\}$ and $\big\{p_1^{SR}(\beta_0), ..., p_K^{SR}(\beta_0)\big\}$ are pivotal, i.e., free of nuisance parameters.

Proof of Proposition 3: The proof 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 $\Pr\Big(\sup_{b \in CI_{\beta_0}(\alpha_1)} \tilde{p}_M[S_\bullet(b)] \leq \alpha_2\Big) \leq \alpha_1 + \alpha_2$ 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

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

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 $\Pr\left(\tilde{p}_M\left[S_\bullet(\beta_0) \leq \alpha_2\right]\right) = \alpha_2$ 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 approach. *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. *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.

Campbell, B. and J.-M. Dufour (1997). Exact non-parametric 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*.

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. *The Econometrics Journal 12*, S19–S49.

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 in econometrics. *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. *The 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. *The 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.

Granger, C. W. J., R. Robins, and R. F. Engle (1986). Wholesale and retail prices: Bivariate time series modeling with forecastable error variances. In D. Belsley and E. Kuh (Eds.), *Model reliability*, pp. 1–17. MIT Press Cambridge, MA.

Gungor, S. and R. Luger (2015). Bootstrap tests of mean-variance efficiency with multiple portfolio groupings. *L'Actualité économique*, Special issue (in English) on *Identification, Simulation, and Finite–Sample Inference*, forthcoming.

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. Ludvigson (2001). Consumption, aggregate wealth, and expect stock returns. *Journal of Finance 56*, 815–849.

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

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

Mizera, I. and J. Wellner (1998). Necessary and sufficient conditions for weak consistency of the median of independent but not identically distributed random variables. *Annals of Statistics 26*, 672–691.

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.

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. *The 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.

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

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.

Zhu, M. (2014). Jackknife for bias reduction in predictive regressions. *Journal of Financial Econometrics 11*, 193–220.

Table 1: Power comparison of different testing strategies

β_1	β_2	α_1	α_2	S_{min}	$S_\times$	W_{min}	$W_\times$
-0.1	0.0	0.01	0.04	32.1	33.1	47.9	47.9
		0.02	0.03	31.2	33.7	46.4	47.0
		0.03	0.02	24.5	25.3	39.6	40.6
		0.04	0.01	12.6	12.1	20.8	21.6
-0.1	-0.1	0.01	0.04	47.4	54.6	65.4	70.8
		0.02	0.03	43.7	52.2	61.0	66.4
		0.03	0.02	37.4	44.4	53.0	60.8
		0.04	0.01	23.7	31.5	37.2	45.5

Notes: This table reports the power (in percentages) of the proposed MC signed-rank tests S_{min}, $S_\times$, W_{min}, $W_\times$ with $M = 100$ under various choices for α_1, α_2 such that $\alpha_1 + \alpha_2 = 5\%$. The data are generated according to (14) with $\beta_0 = \mu_1 = \mu_2 = 0$; $\phi_{12} = \phi_{21} = 0$; $\phi_{11} = \phi_{22} = 0.95$; $\rho_{x_1 r} = -0.9$; $\rho_{x_1 x_2} = 0$. The sample size is $T = 200$, the innovations $\boldsymbol{\epsilon}_t$ are i.i.d. according to a trivariate Student-t distribution with 3 degrees of freedom, and $\sigma_t = 1$ so the returns are conditionally homoskedastic. The reported results are based on 1000 simulation replications of each data-generating configuration.

Table 2: Empirical size of predictability tests

ϕ_{11}	$\rho_{x_1 r}$	σ_t	$\mathcal{D}$	T	Wald	mARM	IVX	S^m_{min}	$S^m_\times$	W^m_{min}	$W^m_\times$	S_{min}	$S_\times$	W_{min}	$W_\times$
0.95	0	iid	N	100	5.2	4.8	5.2	5.4	4.5	4.7	4.3	0.1	0.0	0.4	0.4
				200	4.7	4.5	4.5	4.2	4.2	4.1	3.4	0.3	0.2	0.6	0.5
			t_3	100	5.5	3.7	4.7	4.2	4.0	4.2	3.9	0.0	0.1	0.2	0.1
				200	5.0	4.2	4.6	4.1	4.2	4.7	4.8	0.0	0.0	0.5	0.6
		het	N	100	22.9	21.3	21.7	8.2	7.3	5.5	5.3	0.3	0.3	0.3	0.4
				200	25.4	24.4	24.9	8.7	9.7	6.5	6.3	0.3	0.2	1.1	1.0
0.95	-0.9	iid	N	100	14.2	7.1	6.4	4.8	4.2	5.4	4.3	0.1	0.0	0.3	0.5
				200	10.1	6.8	6.9	5.6	5.0	5.4	4.2	0.4	0.1	0.9	0.6
			t_3	100	14.3	6.2	7.5	5.3	5.2	5.8	5.4	0.1	0.0	0.4	0.4
				200	9.8	7.5	6.6	5.2	4.4	4.0	3.8	0.0	0.0	0.1	0.1
		het	N	100	25.4	19.9	22.4	9.3	10.1	8.4	8.6	0.2	0.3	0.9	0.7
				200	26.5	25.4	25.2	10.1	10.1	6.3	6.4	0.3	0.2	0.8	0.9
0.95	-0.99	iid	N	100	15.8	6.7	7.3	5.8	4.5	5.7	4.8	0.0	0.0	0.5	0.4
				200	9.9	6.2	5.9	4.7	4.5	4.7	4.2	0.1	0.2	0.6	0.3
			t_3	100	13.2	5.9	7.4	5.2	4.5	5.7	5.3	0.1	0.1	0.3	0.3
				200	10.7	6.2	6.1	5.8	6.0	6.2	5.2	0.2	0.1	0.6	0.4
		het	N	100	27.0	19.6	22.2	8.2	8.1	6.9	6.3	0.1	0.4	0.6	0.7
				200	27.9	24.2	24.4	10.0	8.8	7.2	6.8	0.3	0.4	1.2	1.0
0.99	-0.99	iid	N	100	23.4	10.2	7.4	4.8	4.8	5.6	5.0	0.0	0.0	0.0	0.1
				200	18.9	5.1	6.8	4.8	4.2	5.7	4.4	0.0	0.0	0.6	0.4
			t_3	100	22.2	9.3	7.2	4.4	3.5	5.3	4.1	0.0	0.1	0.3	0.2
				200	19.2	6.8	7.5	5.2	6.0	5.4	4.8	0.1	0.1	0.3	0.3
		het	N	100	30.3	20.7	21.7	8.7	8.5	6.4	5.6	0.0	0.0	0.4	0.5
				200	29.6	23.3	25.9	10.8	9.7	7.5	6.5	0.5	0.9	1.0	1.0
1.00	-0.99	iid	N	100	28.1	13.6	7.4	5.1	4.1	5.0	4.3	0.0	0.0	0.2	0.1
				200	27.3	9.8	6.4	5.2	4.0	5.2	4.9	0.2	0.1	0.8	0.8
			t_3	100	27.4	12.6	7.2	4.5	4.0	4.6	3.2	0.0	0.0	0.2	0.1
				200	29.1	10.7	7.2	5.0	5.2	5.4	5.7	0.2	0.1	0.4	0.5
		het	N	100	31.0	23.5	21.8	7.4	7.9	5.2	5.4	0.1	0.1	0.2	0.2
				200	30.8	24.8	25.8	10.2	8.4	7.2	5.7	0.2	0.5	0.7	0.4

Notes: This table reports the empirical size (in percentages) 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 $M = 100$ for a given nominal level $\alpha = 5\%$. The data are generated according to (14) with $\beta_0 = \mu_1 = \mu_2 = 0$; $\beta_1 = \beta_2 = 0$ (the null hypothesis); $\phi_{12} = \phi_{21} = 0$; $\phi_{22} = 0.95$; $\rho_{x_1 x_2} = 0$; the other parameter values and the sample sizes are listed in columns 1–5. The innovations $\boldsymbol{\epsilon}_t$ in (14) are i.i.d. according to either a trivariate normal distribution (N) or a Student-t distribution with 3 degrees of freedom (t_3). The "iid" case corresponds to $\sigma_t = 1$, while "het" refers to the conditional heteroskedasticity case obtained with $\sigma_t = \exp(x_{2,t-1}/100)$ in (14). The reported results are based on 1000 simulation replications of each data-generating configuration.

Table 3: Size-adjusted power comparison of predictability tests

$\rho_{x_1x_2}$	σ_t	$\mathcal{D}$	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	iid	N	100	-0.1	0.0	24.2	48.3	31.9	19.9	19.2	25.2	28.3	2.0	3.1	7.4	8.1
				-0.1	-0.1	58.5	76.7	62.6	28.2	29.4	38.8	44.5	7.0	9.0	16.9	18.8
			200	-0.1	0.0	65.8	84.1	70.8	39.7	40.3	57.2	60.4	14.3	15.2	30.6	32.2
				-0.1	-0.1	91.9	95.5	91.5	54.2	57.5	71.9	77.7	27.5	32.9	49.0	55.8
		t_3	100	-0.1	0.0	23.4	51.4	31.2	28.4	27.9	36.0	35.5	6.3	6.9	13.5	14.9
				-0.1	-0.1	56.6	74.4	59.2	39.5	42.6	50.0	51.8	10.8	12.9	22.9	24.4
			200	-0.1	0.0	66.4	84.6	70.1	61.2	65.1	73.4	73.7	32.1	33.1	47.9	47.9
				-0.1	-0.1	93.0	96.7	93.6	76.5	82.3	86.9	88.5	47.4	54.6	65.4	70.8
0	het	N	100	-0.2	0.0	24.2	29.4	25.3	48.8	40.9	44.9	42.2	24.3	26.3	28.3	29.9
				-0.2	-0.2	33.0	37.5	34.6	42.9	38.1	45.9	46.7	22.3	26.4	29.8	32.0
			200	-0.2	0.0	26.7	30.4	27.1	75.6	80.2	76.2	76.9	64.9	65.7	60.0	60.5
				-0.2	-0.2	37.4	39.9	37.8	69.2	77.9	76.1	80.7	61.1	66.5	60.9	63.4
-0.1	het	N	100	-0.2	0.0	21.7	27.8	23.4	48.2	48.7	49.2	48.5	26.7	28.0	30.6	31.5
				-0.2	-0.2	30.0	34.7	31.2	41.2	42.9	45.5	47.1	21.4	26.1	28.3	30.8
			200	-0.2	0.0	24.7	29.8	24.3	81.7	76.2	75.3	74.5	62.9	63.2	58.7	58.8
				-0.2	-0.2	34.1	37.7	33.1	70.1	64.9	71.7	73.5	54.2	60.0	54.9	60.1
0.1	het	N	100	-0.2	0.0	22.9	28.7	24.3	47.6	49.1	50.6	51.3	25.2	26.8	29.9	30.8
				-0.2	-0.2	35.8	40.9	37.1	46.6	47.9	55.2	57.0	26.0	29.3	31.4	34.3
			200	-0.2	0.0	25.2	30.1	25.8	76.7	75.1	75.0	70.9	64.2	65.0	60.2	60.4
				-0.2	-0.2	34.9	39.2	36.0	70.3	74.3	77.2	75.5	62.6	67.5	64.8	68.2

Notes: This table reports the power (in percentages) 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 $M = 100$ 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. The data are generated according to (14) with $\beta_0 = \mu_1 = \mu_2 = 0$; $\phi_{12} = \phi_{21} = 0$; $\phi_{11} = \phi_{22} = 0.95$; $\rho_{x_1r} = -0.9$; the other parameter values and the sample sizes are listed in columns 1–6. The innovations $\boldsymbol{\epsilon}_t$ in (14) are i.i.d. according to either a trivariate normal distribution (N) or a Student-t distribution with 3 degrees of freedom (t_3). The "iid" case corresponds to $\sigma_t = 1$, while "het" refers to the conditional heteroskedasticity case obtained with $\sigma_t = \exp(x_{2,t-1}/100)$ in (14). The reported results are based on 1000 simulation replications of each data-generating configuration.

Table 4: Summary statistics of employed variables

	r	d/p	e/p	b/m	dfy	tms	tbl
Panel (a) Monthly data							
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
Panel (b) Quarterly data							
Summary statistics							
Mean	0.03	-3.48	-2.75	0.55	0.01	0.02	0.04
Std dev	0.08	0.44	0.46	0.25	0.04	0.01	0.03
Autocorr.	0.09	0.98	0.95	0.98	0.88	0.84	0.95
Correlation matrix							
R	1						
d/p	-0.019	1					
e/p	0.001	0.772	1				
b/m	-0.043	0.886	0.791	1			
dfy	-0.010	0.121	-0.029	0.273	1		
tms	0.088	-0.268	-0.354	-0.305	0.257	1	
tbl	-0.063	0.270	0.352	0.437	0.437	-0.423	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), dividend-price ratio (d/p), earnings-price ratio (e/p), book-to-market ratio (b/m), default yield spread (dfy), term spread (tms), and short rate (tbl). Panels (a) and (b) report the results with monthly and quarterly data, respectively.

Table 5: Statistical properties of excess stock returns

	Mean	Std dev	Skewness	Kurtosis	JB	Autocorr.	$Q^2(6)$	$Q^2(12)$
Panel (a) Monthly excess returns								
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**
Panel (b) Quarterly excess returns								
67-year period								
1948Q1 – 2014Q4	0.02	0.79	**-0.59**	**3.93**	**25.29**	0.10	**15.23**	**24.36**
20-year subperiods								
Jan 1948 – Dec 1967	0.03	0.07	-0.71	3.95	**9.67**	0.14	4.81	11.03
Jan 1968 – Dec 1987	0.01	0.09	-0.38	3.67	3.47	0.13	6.49	12.38
Jan 1988 – Dec 2014	0.02	0.08	-0.57	3.61	**7.50**	0.05	**13.12**	15.73

Notes: This table reports on the statistical properties of monthly and quarterly 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 6: 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
Panel (a) Monthly data								
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)	
Panel (b) Quarterly data								
r_t	0.253	0.045	0.019	-0.052	1.826	0.343	**-0.457**	0.043
	(0.141)	(0.027)	(0.027)	(0.070)	(2.520)	(0.455)	(0.233)	
d/p_t	-0.154	**0.937**	0.027	0.038	-1.673	-0.347	0.124	0.966
	(0.154)	(0.029)	(0.028)	(0.075)	(2.624)	(0.475)	(0.253)	
e/p_t	**-0.957**	-0.087	**0.822**	**0.441**	**-15.990**	**2.356**	0.848	0.918
	(0.341)	(0.068)	(0.119)	(0.163)	(6.358)	(1.102)	(0.588)	
b/m_t	**0.253**	-0.010	-0.006	**0.997**	1.826	-0.401	0.103	0.961
	(0.076)	(0.016)	(0.011)	(0.039)	(1.765)	(0.316)	(0.138)	
dfy_t	-0.002	-0.001	0.001	0.001	**0.842**	0.007	**0.015**	0.792
	(0.004)	(0.001)	(0.001)	(0.002)	(0.079)	(0.012)	(0.007)	
tms_t	0.009	0.003	-0.002	-0.007	**0.466**	**0.768**	-0.003	0.723
	(0.010)	(0.002)	(0.002)	(0.005)	(0.145)	(0.055)	(0.021)	
tbl_t	**0.253**	-0.003	0.001	0.008	**1.826**	0.129	**0.973**	0.907
	(0.013)	(0.003)	(0.002)	(0.007)	(0.180)	(0.072)	(0.024)	

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 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 the 5% level.

Table 7: Residual correlation matrix

	r	d/p	e/p	b/m	dfy	tms	tbl
Panel (a) Monthly data							
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
Panel (b) Quarterly data							
r	1						
d/p	-0.974	1					
e/p	-0.376	0.376	1				
b/m	-0.795	0.781	0.378	1			
dfy	-0.137	0.136	-0.296	-0.006	1		
tms	0.080	-0.049	-0.011	-0.063	0.101	1	
tbl	-0.127	0.090	0.139	0.178	-0.324	-0.837	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 8: 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
Panel (a) Monthly excess returns									
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
Panel (b) Quarterly excess returns									
67-year period									
1948Q1 – 2014Q4	0.08	0.22	**0.02**	0.06	0.22	0.47	**0.05**	0.19	**0.00**
20-year subperiods									
1948Q1 – 1967Q4	**0.01**	0.07	**0.01**	**0.03**	0.26	0.36	0.07	0.27	**0.02**
1968Q1 – 1987Q4	0.11	0.13	**0.04**	0.09	0.85	0.79	0.48	0.27	**0.00**
1988Q1 – 2014Q4	**0.03**	**0.03**	0.17	0.26	0.95	0.95	0.75	0.8	0.07

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 and the quarterly frequency from January 1948 to December 2014. Bold face numbers indicate joint significance at the nominal 5% level.

Table 9. Marginal p-values of each predictor in a univariate regression setup

	$\mathrm{d/p}_{t-1}$	$\mathrm{e/p}_{t-1}$	$\mathrm{b/m}_{t-1}$	dfy_{t-1}	tms_{t-1}	tbl_{t-1}
Panel (a) Monthly excess returns						
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**
Panel (b) Quarterly excess returns						
S^m	0.47	0.88	0.22	0.75	**0.01**	0.51
W^m	0.22	0.40	0.83	0.90	**0.01**	0.29
S	1.00	1.00	1.00	0.90	0.08	0.95
W	1.00	0.99	1.00	1.00	**0.02**	1.00
Wald	**0.03**	0.16	0.23	0.55	0.08	**0.03**

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 10. 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}	$W_{\times}$
Panel (a) Monthly excess returns								
$K = 6$	**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
Panel (b) Quarterly excess returns								
$K = 6$	0.08	0.22	**0.02**	0.06	0.22	0.47	0.05	0.19
$K = 5$	0.64	0.77	0.64	0.59	0.94	0.9	0.97	0.97

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 $K = 5$ are 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: Excess stock returns

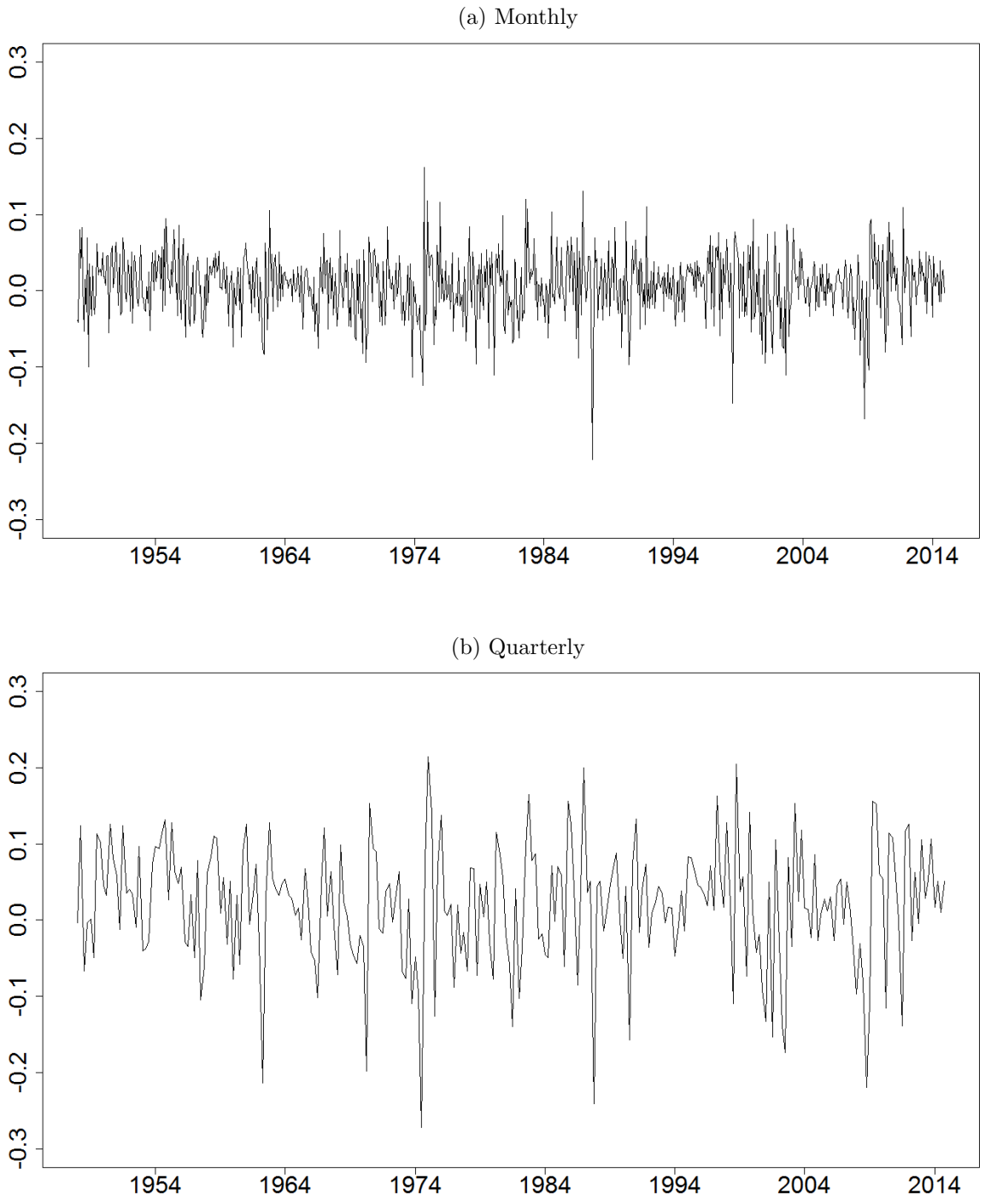

Figure 1 shows the monthly (panel a) and quarterly (panel b) time series of excess returns on the S&P value-weighted index over the period from January 1948 to December 2014.

Figure 2: Monthly predictors

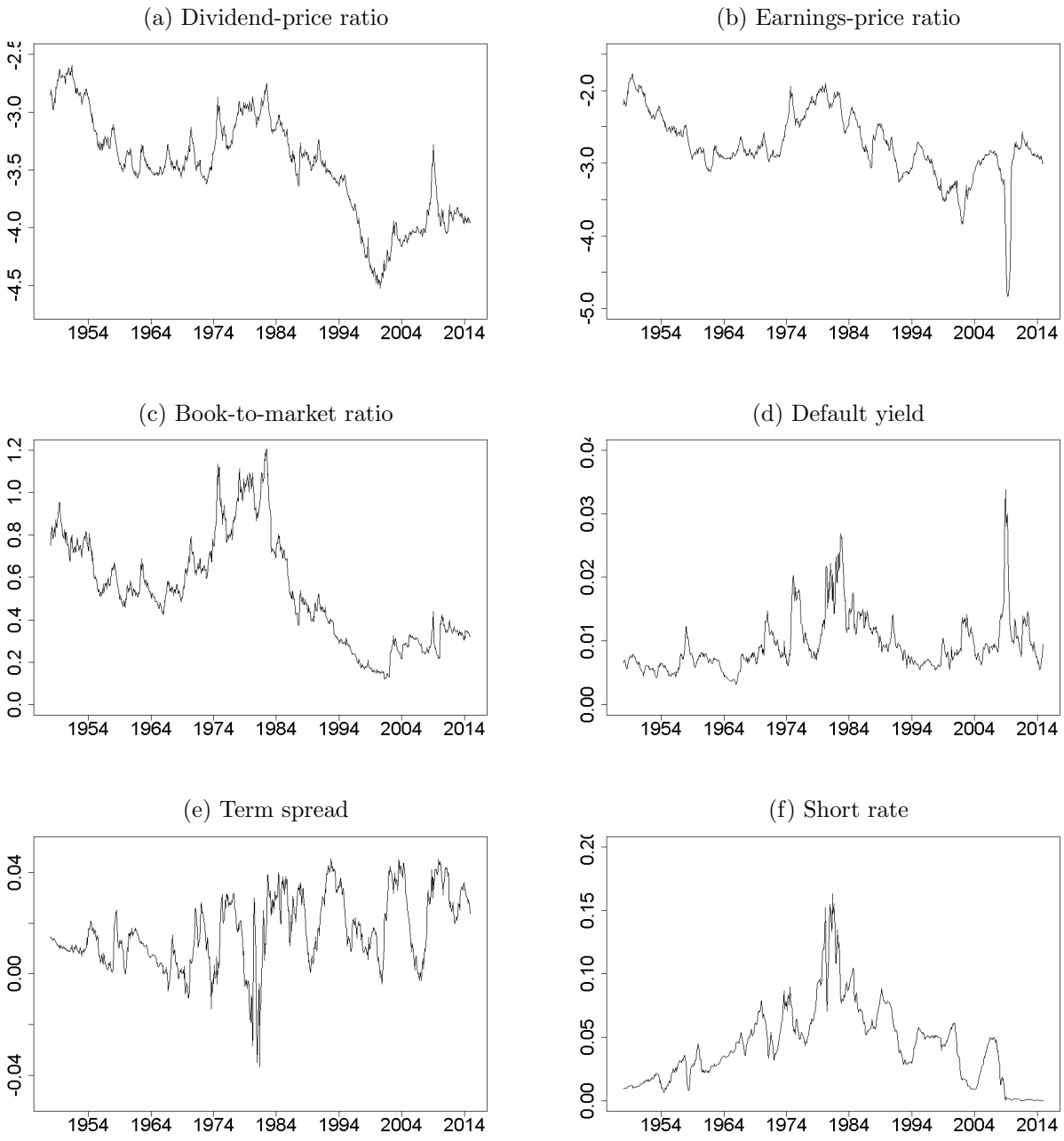

Figure 2 shows the monthly time series of the six predictors over the period from January 1948 to December 2014. Panels (a)–(f) show the dividend price ratio (d/p), the earnings-price ratio (e/p), the book-to-market ratio (b/m), the default yield (dfy), the term spread (tms), and the short rate (tbl),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

Figure 3: Quarterly predictors

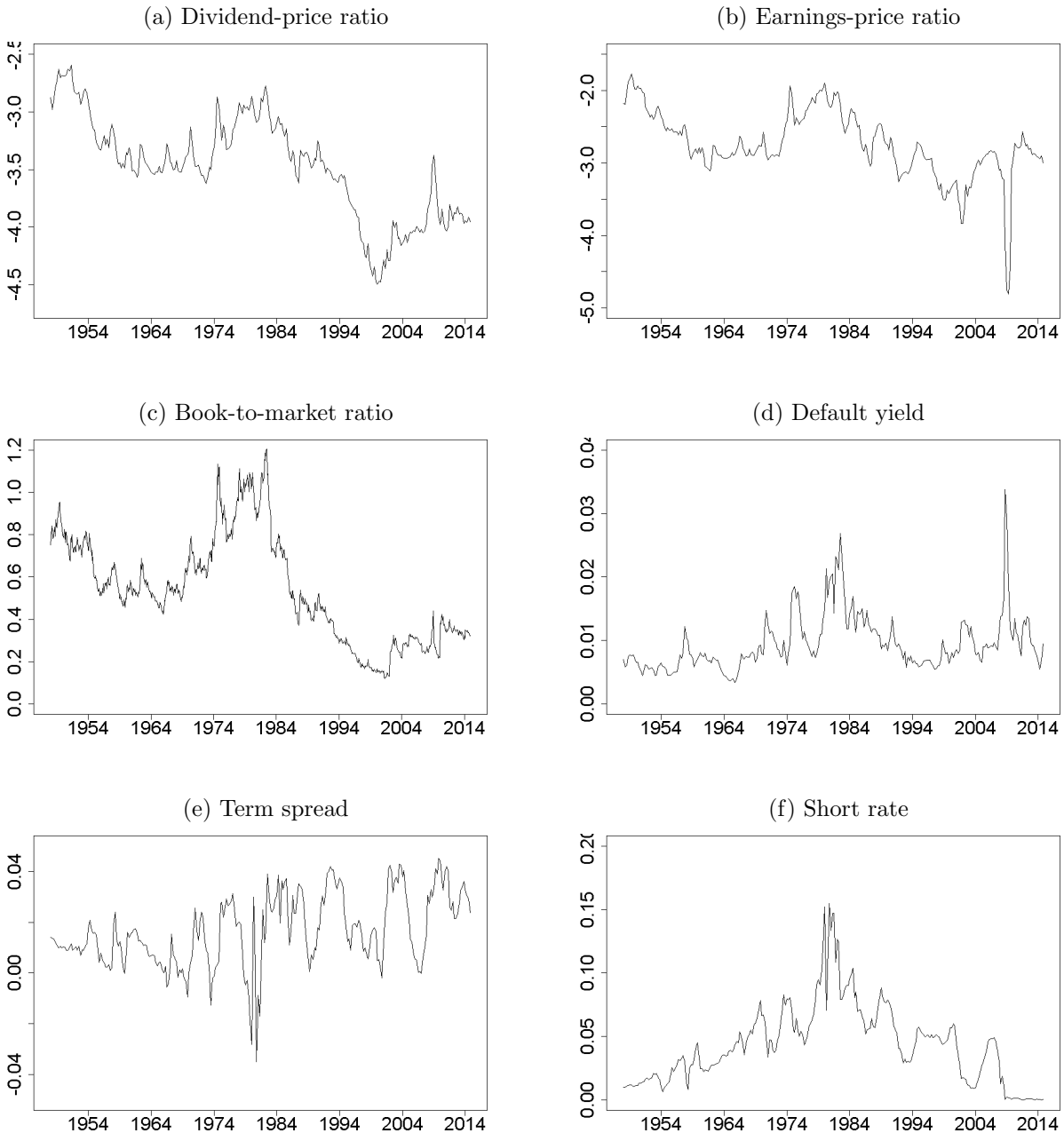

Figure 3 shows the quarterly time series of the six predictors over the period 1948Q1–2014Q4. Panels (a)–(f) show the dividend price ratio (d/p), the earnings-price ratio (e/p), the book-to-market ratio (b/m), the default yield (dfy), the term spread (tms), and the short rate (tbl), respectively.

Figure 4: Rolling-window predictability tests with monthly excess returns

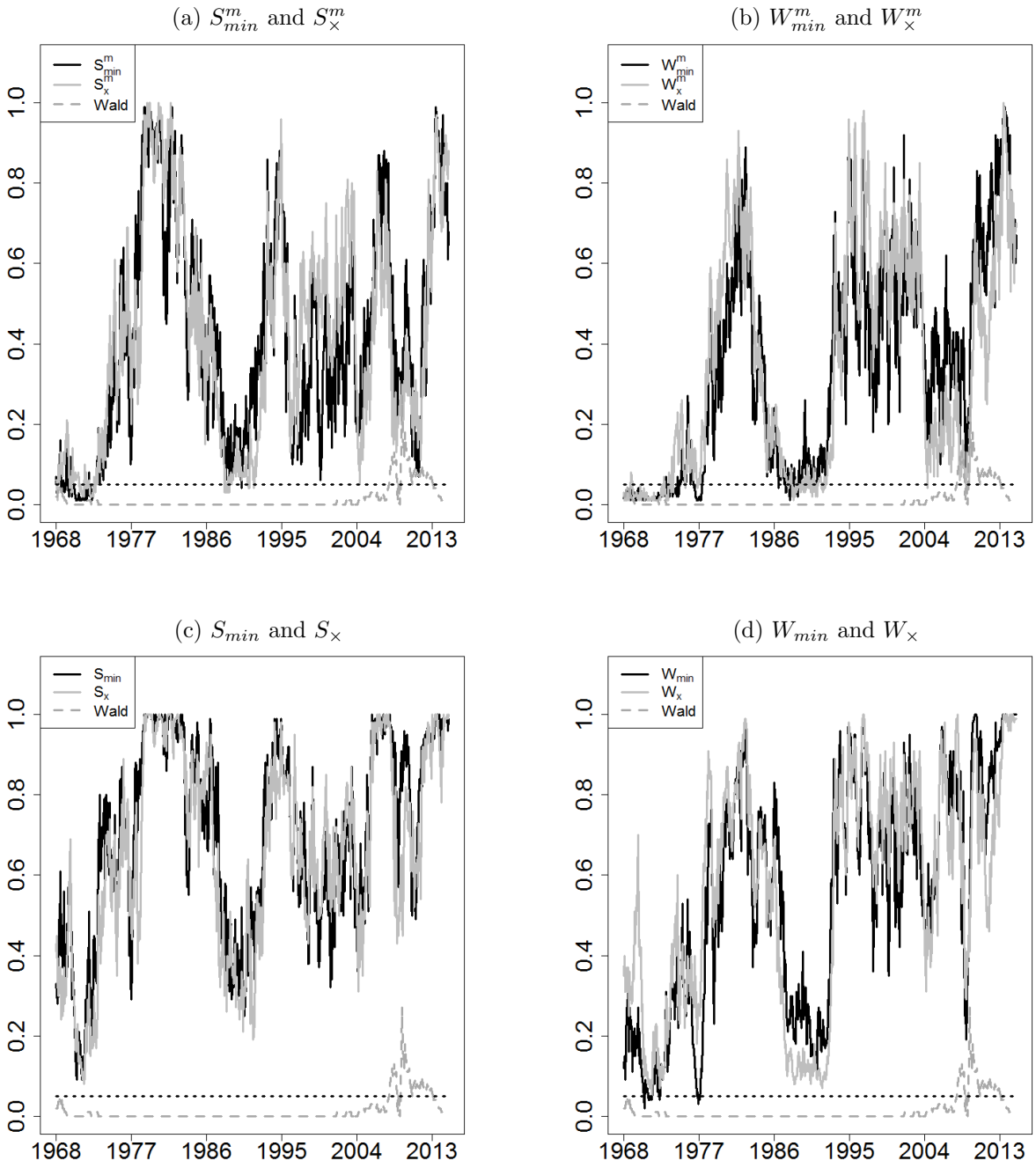

Figure 4 shows the p-values of the proposed sign and signed rank tests and the benchmark Wald test using a 240-month (20-year) rolling window. The solid black line indicates the tests based on the minimum p-value, the solid grey line is for the tests based on the product of the p-values, the dashed grey line is for the Wald test, and finally the horizontal dotted line shows the nominal 5% significance level.

Figure 5: Rolling-window predictability tests with quarterly excess returns

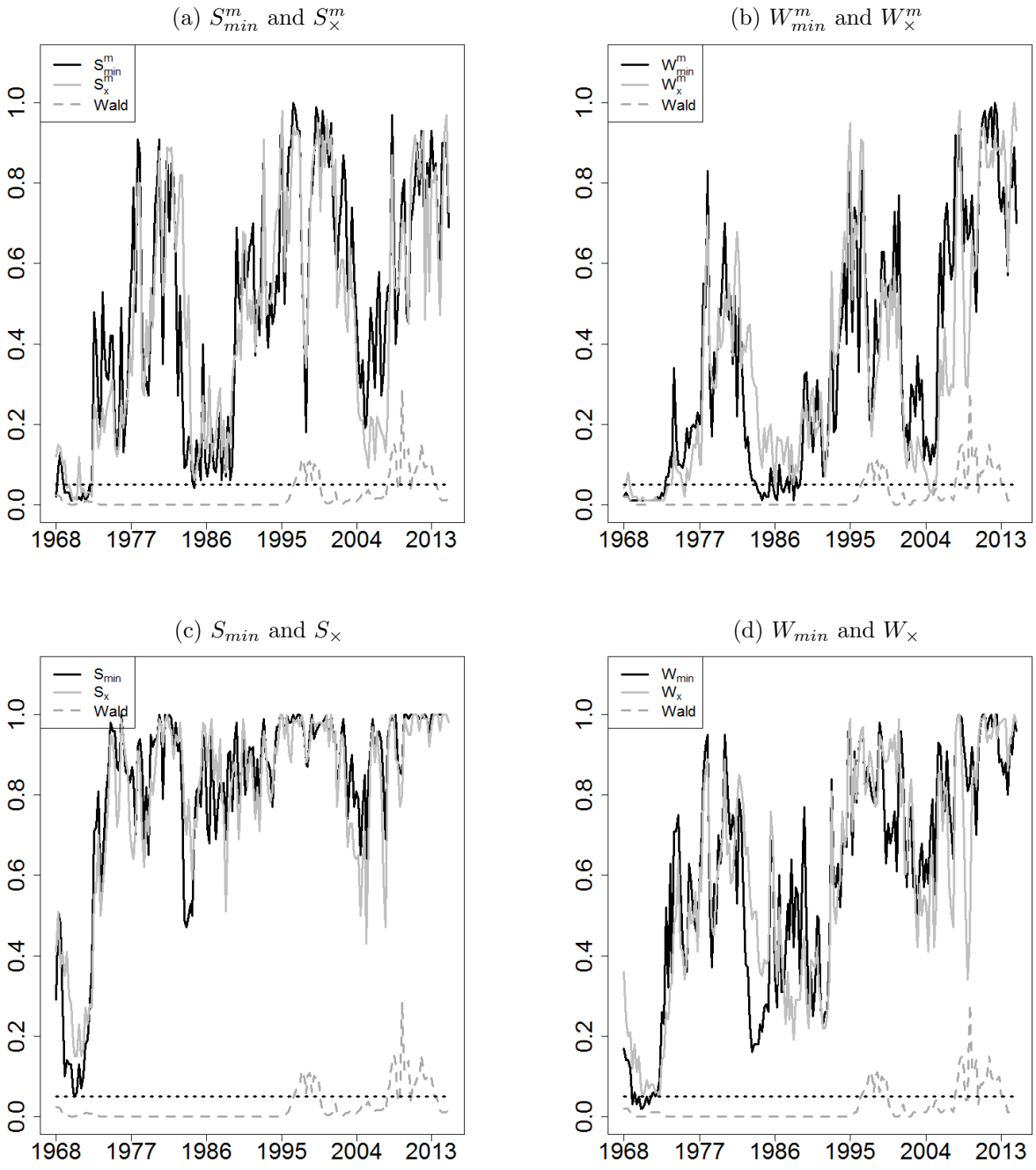

Figure 5 shows the p-values of the proposed sign and signed rank tests and the benchmark Wald test using an 80-quarter (20-year) rolling window. The solid black line indicates the tests based on the minimum p-value, the solid grey line is for the tests based on the product of the p-values, the dashed grey line is for the Wald test, and finally the horizontal dotted line shows the nominal 5% significance level.