Taking Stock of Long-Horizon Predictability Tests: Are Factor Returns Predictable?

Alexandros Kostakis[*], Tassos Magdalinos[†] and Michalis P. Stamatogiannis[‡]

13 June 2022

Abstract

This study provides a critical assessment of long-horizon return predictability tests using highly persistent regressors. We show that the commonly used statistics are typically oversized, leading to spurious inference. Instead, we propose a Wald statistic, which accommodates multiple predictors of (unknown) arbitrary persistence degree within the I(0)-I(1) range. The test statistic, based on an adaptation of the IVX procedure to a long-horizon regression framework, is shown to have a standard chi-squared asymptotic distribution (regardless of the stochastic properties of the regressors used as predictors) and to exhibit excellent finite-sample size and power properties. Employing this test statistic, we find evidence of predictability for "old" and "new" pricing factors with monthly returns, but this becomes weaker as the predictive horizon increases. The predictability evidence substantially weakens with annual data. Overall, we question the incremental value of using long-horizon predictive regressions.

1 Introduction

Stock return predictability has an important impact on the theory and practice of all aspects of modern finance. If returns are predictable, then risk premia, and hence the cost of capital become time-varying, investors may engage in strategic asset allocation, and conditional asset pricing models are bound to explain better than unconditional models the time-series and cross-sectional properties of stock returns.

The empirical asset pricing literature keeps "discovering" significant predictors of market returns (see Welch and Goyal, 2008; Rapach and Zhou, 2013; and the references therein). These conclusions rely on the statistical inference derived from predictive regressions, where the lagged value of a financial variable is used as predictor of next-period stock returns. Given the low R^2 of these regressions, the marginal significance of the coefficient estimates for most of the predictors, and the questionable validity of standard t-tests in the presence of highly persistent and endogenous regressors (see, inter alia, Stambaugh 1986, 1999; Cavanagh et al., 1995; Elliott, 1998; Campbell and Yogo, 2006; Kostakis et al., 2015, hereafter KMS; Demetrescu et al. 2020), the literature often resorts to long-horizon predictive regressions (see the pioneering studies of Fama and French, 1988, 1989; Campbell and Shiller, 1988). Long-horizon predictive regressions typically yield R^2s

[*] Accounting and Finance Group, University of Liverpool Management School & Alliance Manchester Business School.

[†] Corresponding Author. Department of Economics, University of Southampton.

[‡] Accounting and Finance Group. University of Liverpool Management School.

that increase with the assumed horizon and highly significant coefficient estimates when Newey and West (1987) or Hansen and Hodrick (1980) standard errors are employed to account for using overlapping observations.

Nevertheless, the validity of inference using long-horizon predictive regressions has also been questioned (see Nelson and Kim, 1993; Goetzmann and Jorion, 1993; Boudoukh and Richardson, 1994; Ang and Bekaert, 2007; Boudoukh et al., 2008). Moreover, the increasing number of predictors being discovered in prior studies unavoidably raises concerns regarding data mining and inadvertent p-hacking (see the critiques of Ferson et al., 2003, and Harvey, 2017). Therefore, it is time to take stock and re-examine the issue of long-horizon return predictability in a comprehensive manner. This is the aim of this study.

Motivated by the very strong persistence of the commonly used predictors, a series of studies have modelled these variables as local-to-unity processes à la Phillips (1987), whose autoregressive root converges to unity at rate n (see Campbell and Yogo, 2006; Jansson and Moreira, 2006). Based on this modelling innovation, Valkanov (2003), Torous et al. (2004), Rossi (2007), and Hjalmarsson (2011) have devised Bonferroni-type confidence interval procedures for inference in long-horizon predictive regressions. Such procedures may exhibit good properties when the regressor is at least as persistent as a local-to-unity process but typically they become asymptotically invalid if the predictor is less persistent than a near-I(1) process. Since the exact degree of predictor persistence remains unknown, this assumption is rather restrictive, leading to a substantially undersized test statistic as we deviate from a (near) nonstationary data generation mechanism. Equally importantly, because of its reliance on confidence interval construction for the local to unity parameter, Bonferroni-type methods are typically restricted to the case of a single predictor and cannot accommodate multivariate predictive regression systems that are of particular empirical interest.

To overcome the limitations described above, we propose a Wald test statistic that provides valid inference for long-horizon predictive regressions, in the presence of potentially endogenous regressors with arbitrary persistence properties covering the entire I(0)-I(1) spectrum. In particular, our methodology extends the IVX procedure of Phillips and Magdalinos (2009) to long-horizon multivariate predictive regression systems, encompassing as a special case the short-horizon setup of KMS. The key idea of the presented methodology is to construct an instrumental variable whose degree of persistence we explicitly control. In this way, the inference problems arising due to the uncertainty regarding the persistence of the original regressor are avoided. Using the constructed instrument, we then perform a standard instrumental variable estimation. The derived estimator asymptotically follows a mixed normal distribution, and hence the corresponding Wald statistic asymptotically follows a chi-squared distribution under the null, considerably simplifying inference.

The proposed test statistic presents a number of advantages in comparison to the previously suggested approaches. First, as already mentioned, it does not require a priori knowledge of the exact time series properties of the employed predictors. In fact, it accommodates regressors with very general time series characteristics, varying from purely stationary to purely nonstationary processes, including all intermediate persistence regimes. Second, it can be used to conduct joint predictability tests in multivariate predictive regression systems, rather than univariate predictability tests only. Third, it is much simpler to implement in comparison to Bonferroni-type tests, which require computing critical values for each case in hand. Fourth, it can be used to test hypotheses for any set of linear restrictions, not just the null of no predictability. Last, we show that this test statistic exhibits excellent finite-sample properties for a very large range of empirically relevant parameter values, leading to valid inference. A related IVX-based method for long-horizon regressions is developped by Demetrescu, Rodrigues and Taylor (2022) who augment the regression before applying IVX instrumentation, leading to different predictability tests.

Equipped with this correctly sized test statistic, we examine whether factor returns are predictable. To provide comprehensive evidence, we consider both "old" and "new" factors, beyond the well-examined market portfolio, which carry significant premia and have been proposed in the empirical asset pricing literature to risk-adjust returns. In particular we examine the size and value factors of Fama and French (1993), hereafter FF1993, the momentum factor of Carhart (1997), the profitability and investment factors of Fama and French (2015), hereafter FF2015, as well as the corresponding size, profitability and investment factors of Hou et al. (2015), hereafter HXZ.

Instead of "searching" for predictors, a practice that would raise a valid p-hacking criticism, we examine the predictability of factor returns using a small set of six financial variables: dividend-price ratio, earnings-price ratio, book-to-market value ratio, default yield spread, T-bill rate, and term spread. Not only these variables have been extensively used as predictors of market returns during the last three decades (see, inter alia, Keim and Stambaugh, 1986; Campbell and Shiller, 1988; Fama and French, 1988, 1989; Kothari and Shanken, 1997; Lamont, 1998; Pontiff and Schall, 1998), but their combinations have also been commonly used to span the state space of the economy (see, for example, Campbell et al., 2003; Campbell and Vuolteenaho, 2004; Petkova, 2006; Maio and Santa-Clara, 2012).

Our empirical analysis yields a number of interesting conclusions. Regarding univariate predictability tests with monthly returns, using the correctly sized Wald test statistic, we find evidence in favor of in-sample predictability for the market portfolio at short horizons via the earnings-price ratio, book-to-market value ratio, T-bill rate and term spread. However, this evidence becomes weaker, not stronger, as the predictive horizon increases, and disappears when annual returns are employed. This conclusion is in stark contrast with the findings of prior studies, which relying on Newey-West or Hansen-Hodrick standard errors, concluded that market returns are highly significantly predictable at long horizons via price-scaled ratios and term structure variables. We confirm that this spurious inference would arise in our sample period too.

Beyond the market portfolio, we find evidence that the default yield spread and the book-to-market value ratio can significantly predict the returns of the FF1993 size and value factors. Again, this evidence becomes weaker as the horizon increases or when annual returns are used. Interestingly, we find that the default yield spread can also significantly predict momentum returns, drawing a link between the premium that this factor yields and the credit conditions in the economy. Regarding the recently proposed factors, we find almost no evidence that the premia earned by the FF2015 profitability and investment factors are predictably time-varying via the state variables we employ. To the contrary, the profitability factor of HXZ, which is constructed in a different way than the factor of FF2015, is indeed predictable via the earnings-price ratio, the default yield spread, and the T-bill rate. Overall, different predictors contain predictive ability over different factors. To the extent that these factors mimic different sources of risk, one could argue that different state variables are necessary to capture these alternative dimensions of risk.

Our multivariate predictability tests reveal some interesting patterns too. Confirming the arguments of Ang and Bekaert (2007), we find that combinations which include the dividend-price ratio and the T-bill rate possess in-sample predictability with respect to market returns, but this evidence is significant only at short horizons. Moreover, we find only weak evidence of predictability by the examined combinations of predictors with respect to the returns of the FF1993 size and value factors. To the contrary, we report robust and significant predictability for the returns of the momentum factor and the HXZ profitability factor. Under the common assumption that these variables are good proxies for the state of the economy, this evidence shows that the momentum and profitability premia can be interpreted as compensation for exposure to macroeconomic risks. Nevertheless, we show that no single combination of these predictors can accurately capture the time-variation in the returns of *all* factors.

Taken together, our study provides a critical assessment of the long-horizon stock return predictability literature. Long-horizon predictability tests were initially perceived as a tool to confirm that market returns are significantly predictable, overcoming the marginal significance and the low explanatory power of short-horizon regressions. The underlying assumption supporting this practice was that long-horizon returns are less noisy, and hence these tests would be more powerful (see Campbell, 2001; Rapach and Wohar, 2005; for interesting discussions). However, it turns out that, in the presence of strongly persistent variables, which have been predominantly used as predictors, the highly significant evidence reported in favor of predictability is mostly spurious due to use of severely oversized test statistics. In fact, using a correctly sized test statistic, the significance of long-horizon predictability mostly disappears. Equally importantly, our simulation analysis shows that the correctly sized test statistics become *less*, not more, powerful as the predictive horizon increases, deteriorating the quality of inference at long horizons. Hence, our study questions the incremental value of conducting statistical inference using long-horizon predictive regressions instead of the actual data generating process. The related paper of Demetrescu, Rodrigues and Taylor (2022) reaches similar conclusions.

The rest of the study is organized as follows. Section 2 formally presents the predictive regression setup, introduces the proposed IVX estimator and IVX-Wald statistic for long-horizon return predictability tests and presents the asymptotic theory of estimation and inference. Section 3 contains an extensive simulation analysis, documenting the finite-sample properties of the IVX-Wald test in comparison to other commonly used predictability tests. Section 4 contains our empirical application regrading the predictability of factor returns, whereas Section 5 concludes.

2 Econometric Analysis of Long Horizon Regressions

2.1 Predictive Regression Setup

We consider the following multivariate system of predictive regressions with regressors exhibiting an arbitrary degree of persistence:

$$\begin{aligned} y_t &= \mu + A x_{t-1} + \varepsilon_t, & (1) \\ x_t &= R_n x_{t-1} + u_t, & (2) \end{aligned}$$

where A is an $m \times r$ coefficient matrix for $t \in \{1, ..., n\}$, where n denotes the sample size. The vector of predictor variables x_t in (2), initialised at $x_0 = 0$ for simplicity, exhibits a degree of persistence induced by the autoregressive matrix in (2) according to the following assumption.

Assumption P. *The autoregressive matrix R_n in (2) satisfies*

$$C_n := \kappa_n \left(R_n - I_r \right) \to C \quad as \ n \to \infty \tag{3}$$

for some $r \times r$ matrix C satisfying $\|C\| < \infty$ and some sequence $(\kappa_n)_{n \in \mathbb{N}}$ of positive numbers. The regressor x_t in (2) belongs to one of the following classes:
(i) *Near-nonstationary regressors, if (3) holds with $\kappa_n / n \to \kappa \in (0, \infty]$.*
(ii) *Near-stationary regressors, if (3) holds with $\kappa_n / n \to 0$, $\kappa_n \to \infty$ and C a negative stable matrix (i.e. all eigenvalues of C have negative real part)*
(iii) *Stationary regressors, if (3) holds with $\kappa_n = 1$ and $R = I_r + C$ has spectral radius $\rho(R) < 1$.*

The classes P(i)-P(iii) include predictor variables with very general time series characteristics, varying from purely stationary to purely nonstationary processes and accommodating all intermediate persistence regimes. It is worth noting that the above data generation environment represents a major generalisation of that in KMS and Phillips and Magdalinos (2009), since the (severely restrictive) diagonality assumption on C is replaced by assumptions on the spectrum of C. These assumptions are minimal for cases P(ii) and P(iii): P(iii) is the standard (necessary and sufficient) stability requirement for autoregressive processes, whereas the condition of P(ii) gives rise to regressors with near-stationary characteristics; see the discussion following Assumption N and Lemma 2.1 of Magdalinos and Phillips (2020). In addition, negative stability of C in P(ii) is necessary and sufficient for the existence of a matrix-valued improper integral of the form $\int_0^\infty e^{rC} \Omega e^{rC'} dr$ (for some positive definite matrix Ω) that arises as the probability limit of the sample moment matrix $n^{-1} \kappa_n^{-1} \sum_{t=1}^n x_t x_t'$ under P(ii); see the definition of V_C below. In sum, the parametrisation in (3) builds on the asymptotic development of Magdalinos and Phillips (2020) and accommodates a much wider class of unrestricted VAR(1) regressors x_t than KMS with a single unknown persistence degree within the $I(0)$-$I(1)$ range.

A standard assumption in the stock return predictability literature is to assume that the innovations ε_t of the predictive equation (1) are uncorrelated, while allowing for correlation in the innovations of the predictor sequence u_t in the form of a stationary linear process. The dependence structure of the innovations is formally presented in the following Assumption, designed to include both conditional homoskedastic and covariance stationary GARCH innovation processes.

Assumption INNOV. *Let $\epsilon_t = (\varepsilon_t', e_t')'$, with ε_t as in (1), denote an $\mathbb{R}^{m+r}$-valued martingale difference sequence with respect to the natural filtration $\mathcal{F}_t = \sigma(\epsilon_t, \epsilon_{t-1}, ...)$ satisfying*

$$E_{\mathcal{F}_{t-1}}(\epsilon_t \epsilon_t') = \Sigma_t \ \ a.s. \ \ and \ \ \sup_{t \in \mathbb{Z}} E \|\epsilon_t\|^{2\nu} < \infty \tag{4}$$

for some $\nu > 1$, where Σ_t is a positive definite matrix. Let u_t in (2) be a stationary linear process

$$u_t = \sum_{j=0}^{\infty} F_j e_{t-j}, \tag{5}$$

where $(F_j)_{j \geq 0}$ is a sequence of constant matrices such that $F_u(1) := \sum_{j=0}^{\infty} F_j$ has full rank and $F_0 = I_r$. We maintain one of the following assumptions:
(i) *$\Sigma_t = \Sigma_\epsilon$ for all t and $\sum_{j=0}^{\infty} \|F_j\| < \infty$.*
(ii) *The process $(\epsilon_t)_{t \in \mathbb{Z}}$ satisfies (4), with $\nu = 2$ and $(\Sigma_t)_{t \in \mathbb{Z}}$ being a stationary ergodic process. The process $(\varepsilon_t)_{t \in \mathbb{Z}}$ in (1) admits the following stationary vec-GARCH(p, q) representation:*

$$\varepsilon_t = H_t^{1/2} \eta_t, \quad \mathrm{vec}(H_t) = \overline{\varphi} + \sum_{i=1}^{q} A_i \mathrm{vech}(\varepsilon_{t-i}\varepsilon_{t-i}') + \sum_{k=1}^{p} B_k \mathrm{vech}(H_{t-k}), \tag{6}$$

where $(\eta_t)_{t \in \mathbb{Z}}$ is an $\mathcal{F}_t$-adapted sequence of i.i.d. $(0, I_m)$ random vectors, $\overline{\varphi}$ is a constant vector, A_i, B_k are symmetric positive semidefinite matrices for all i, k, and the spectral radius of the matrix $\Gamma = \sum_{i=1}^{q} A_i + \sum_{k=1}^{p} B_k$ satisfies $\rho(\Gamma) < 1$. The sequence $(F_j)_{j \geq 0}$ in (5) satisfies $\sum_{j=0}^{\infty} j \|F_j\| < \infty$.

Assumption INNOV(i) imposes conditional homoskedasticity on the martingale difference sequence ϵ_t and short-memory on the linear process (5). Assumption INNOV(ii) accounts for conditionally heteroskedastic ϵ_t with finite fourth-order moments of a very general form: the vec-GARCH process in (6) is the most general multivariate GARCH specification (see Chapter 11 of Francq and Zakoian, 2010).

Following standard notational convention, we define the short-run and long-run covariance matrices associated with the innovations ε_t and u_t in (1), (2) as follows: $\Sigma_{\varepsilon\varepsilon} = E(\varepsilon_t \varepsilon_t')$, $\Sigma_{\varepsilon u} = E(\varepsilon_t u_t')$, $\Sigma_{uu} = E(u_t u_t')$, $\Omega_{uu} = \sum_{h=-\infty}^{\infty} E(u_t u_{t-h}')$, $\Lambda_{u\varepsilon} = \sum_{h=1}^{\infty} E(u_t \varepsilon_{t-h}')$ and $\Omega_{\varepsilon u} = \Sigma_{\varepsilon u} + \Lambda_{u\varepsilon}'$. Note that $\Omega_{\varepsilon u}$ is only a one-sided long run covariance matrix because ε_t is an uncorrelated sequence by Assumption INNOV. For the same reason, the long-run covariance of the ε_t sequence is equal to the short-run covariance $\Sigma_{\varepsilon\varepsilon}$. Denoting by $\hat{\varepsilon}_t$ the OLS residuals from (1) and by $\hat{u}_t$ the OLS residuals from (2), the above covariance matrices can be estimated in a standard way: $\hat{\Sigma}_{\varepsilon\varepsilon} = n^{-1} \sum_{t=1}^{n} \hat{\varepsilon}_t \hat{\varepsilon}_t'$, $\hat{\Sigma}_{\varepsilon u} = n^{-1} \sum_{t=1}^{n} \hat{\varepsilon}_t \hat{u}_t'$ and $\hat{\Sigma}_{uu} = n^{-1} \sum_{t=1}^{n} \hat{u}_t \hat{u}_t'$. Accommodating autocorrelation in u_t that takes the general form (5) requires non-parametric estimation of the long-run covariance matrices: letting M_n be a bandwidth parameter satisfying $M_n \to \infty$ and $M_n/\sqrt{n} \to 0$ as $n \to \infty$, we employ the usual Newey-West type estimators

$$\left[\hat{\Lambda}_{uu}, \hat{\Lambda}_{u\varepsilon}\right] = \frac{1}{n} \sum_{h=1}^{M_n} \left(1 - \frac{h}{M_n + 1}\right) \left[\sum_{t=h+1}^{n} \hat{u}_t \hat{u}_{t-h}', \sum_{t=h+1}^{n} \hat{u}_t \hat{\varepsilon}_{t-h}'\right] \tag{7}$$

$\hat{\Omega}_{uu} = \hat{\Sigma}_{uu} + \hat{\Lambda}_{uu} + \hat{\Lambda}_{uu}'$ and $\hat{\Omega}_{\varepsilon u} = \hat{\Sigma}_{\varepsilon u} + \hat{\Lambda}_{u\varepsilon}'$. Under the full generality of Assumption INNOV, we provide robust inference for the matrix of coefficients A that is invariant to the predictor variables belonging to classes P(i)-P(iii).

2.2 Long-Horizon Predictive Regressions

Inference based on regression estimators from (1), i.e., estimators derived from regressing y_t on x_{t-1} and an intercept, is said to apply in the *short-horizon.* An issue of substantial empirical interest concerns inference in *long-horizon* predictive regressions, i.e., inference based on estimators derived from regressing a K-period accumulation of y_t on x_{t-1} and an intercept, while the true data generating process (DGP) continues to be given by (1). In particular, denoting $y_t(K) = \sum_{i=0}^{K-1} y_{t+i}$, long-horizon estimates are derived from the fitted regression:

$$y_t\left(K\right) = \mu_f + A_K x_{t-1} + \eta_{f,t} \quad t \in \{1, ..., n-K+1\} \tag{8}$$

for a pre-determined horizon value K, when the true relationship between y_t and x_t is given by (1). For brevity, we introduce the notation

$$v_t\left(l\right) := \sum_{i=0}^{l-1} v_{t+i} \quad \text{for } t \in \{1, ..., n-l+1\} \tag{9}$$

for any sequence $\left(v_t\right)_{t\geq 1}$ and denote $n_K := n - K + 1$, $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$.

It is clear that the accumulation of predicted variables on the left side of (8) generates additional correlations that are not present in short-horizon regressions and affect the stochastic properties of long-horizon estimators. A standard result on partitioned regression yields that least squares estimation of A_K from the regression (8) is equivalent to least squares estimation of A_K from the regression:

$$y_t\left(K\right) - \bar{y}_{n_K}\left(K\right) = A_K\left(x_{t-1} - \bar{x}_{n_K-1}\right) + \vartheta_t \quad t \in \{1, ..., n_K\}. \tag{10}$$

Let $\underline{Y}\left(K\right) = \left[y_1'\left(K\right) - \bar{y}_{n_K}'\left(K\right), ..., y_{n_K}'\left(K\right) - \bar{y}_{n_K}'\left(K\right)\right]$, $\bar{y}_{n_K}\left(K\right) = n_K^{-1}\sum_{t=1}^{n_K} y_t\left(K\right)$, $\underline{X}\left(K\right) = \left[x_0'\left(K\right) - \bar{x}_{n_K-1}'\left(K\right), ..., x_{n_K-1}'\left(K\right) - \bar{x}_{n_K-1}'\left(K\right)\right]'$ and $\bar{x}_{n_K-1}\left(K\right) = n_K^{-1}\sum_{t=1}^{n_K} x_{t-1}\left(K\right)$; denoting by $\bar{y}_{n_K}$ and $\bar{x}_{n_K-1}$ the usual sample means of y_t and x_{t-1} based on the first n_K observations, the OLS estimator of A_K in (8)/(10) is given by

$$\hat{A}_K^{OLS} = \underline{Y}\left(K\right)'\underline{X}_{n_K-1}\left(\underline{X}_{n_K-1}'\underline{X}_{n_K-1}\right)^{-1}. \tag{11}$$

The additional correlations generated by the accumulation of predicted variables in (8) induce a least squares bias that fails to vanish asymptotically. The magnitude of this asymptotic bias depends on horizon K. The following assumption controls the growth rate of the horizon parameter K relative to the sample size n.

Assumption H. *The horizon K may be a fixed integer or a sequence $\left(K_n\right)_{n\in\mathbb{N}}$ that increases to infinity slower than the sample size n: $K_n/n \to 0$ as $n \to \infty$.*

The following result derives an explicit expression for the asymptotic bias/inconsistency of the least squares estimator in (11) as a function of the horizon parameter K and the regressor persistence degree κ_n.

Proposition 1. *Consider the model (1)–(3) under Assumptions P, INNOV and H. The OLS estimator (11) generated by the long-horizon regression (8) has the following asymptotic behaviour as $n \to \infty$:* **(i)** *Under P(i)-P(ii), $\left(K \wedge \kappa_n\right)^{-1}\hat{A}_K^{OLS} = A\left(K \wedge \kappa_n\right)^{-1}\sum_{i=0}^{K-1} R_n^i + o_p\left(1\right)$. In particular, $K^{-1}\hat{A}_K^{OLS} \to_p A$ when $K/\kappa_n \to 0$ and $\kappa_n^{-1}\hat{A}_K^{OLS} \to_p -AC^{-1}$ when $K/\kappa_n \to \infty$;* **(ii)** *Under P(iii), $\hat{A}_K^{OLS} \to_p A\sum_{i=0}^{K-1}\Gamma_{x_0}\left(i\right)\Gamma_{x_0}^{-1}\left(0\right)$ for $K \in \mathbb{N}\cup\{\infty\}$, where $x_{0,t} = \sum_{j=0}^{\infty} R^j u_{t-j}$, with $R = I_r + C$, denotes the stationary version of the process x_t, and $\Gamma_{x_0}\left(\cdot\right)$ denotes the autocovariance function of $x_{0,t}$.*

Proposition 1 shows that the long-horizon OLS estimator (11) is inconsistent for all horizons $K > 1$ and provides, to our knowledge, the first general representation of the asymptotic bias/inconsistency arising in long-horizon least squares regression. The form of the asymptotic bias depends on the relative magnitude of the horizon parameter K and the regressor persistence degree κ_n: when κ_n dominates K (as will be the case for predictor variables in the unit root and local-to-unity persistence regimes P(i) and P(ii) by Assumption H), the long-horizon OLS estimator (11) estimates $K \times A$ instead of A in large samples; when K dominates κ_n and $\kappa_n \to \infty$, $\hat{A}_K^{OLS}$ estimates $-\kappa_n A C_n^{-1} = A\left(I_r - R_n\right)^{-1}$ instead of A in large samples. In both cases, the distance between the true parameter A and the value estimated by $\hat{A}_K^{OLS}$ diverges to infinity with the sample size when

$K \to \infty$. The least squares asymptotic bias is less severe for increasing horizons in the stationary case P(iii) since $\sum_{i=0}^{\infty} \Gamma_{x_0}(i) < \infty$.

Proposition 1 also provides an insight into the failure of standard hypothesis testing procedures based on the long-horizon OLS estimator. Classical procedures for testing the null hypothesis $A = 0$, such as the Wald test (or the t-test for individual significance), are based on computing a statistical distance between an estimator of A and 0 and rejecting the null hypothesis when this distance is large. However, $\hat{A}_K^{OLS}$ estimates $K \times A$ instead of A for (local-to-) unit root regressors, so the resulting hypothesis test computes the distance between $K \times A$ and 0, leading to over-rejections that become increasingly severe as the horizon K increases; when $K \to \infty$, the probability of Type I error increases to 1. For the standard t-statistic (t_{OLS}), the pattern of monotonically increasing size with the regression horizon is illustrated by the simulated empirical size results of Table 1.

The limitations of employing a test statistic based on an estimator that is consistent only at a single point of the parameter space, even if this point coincides with the restriction imposed by the null hypothesis, can be further illustrated by the asymptotic invalidity of confidence intervals based on that test statistic. By Proposition 1, the standard t-statistic satisfies $|t_{OLS}| \to \infty$ when $K \to \infty$ under Assumptions P(i)-P(ii), so the standard asymptotic confidence interval for A based on t_{OLS} would be $(-\infty, \infty)$ for all $A \neq 0$ (even for very small values of A). Put differently, when rejecting the null hypothesis $A = 0$, a predictability test should also reject very small values of A with the correct probability of Type I error: for example if $|A| = 10^{-6}$, a reasonable statistical decision rule should conclude that there is no predictability. However, when $K \to \infty$ with the sample size in the above scenario, both the standard long-horizon OLS estimator in (19) and the associated t-statistic would diverge to $+\infty$ if $A = 10^{-6}$ and to $-\infty$ if $A = -10^{-6}$, giving rise to a probability of Type I error increasing to 1 with n for a two-sided rejection region or to a probability of Type I error increasing to 1 with n when $A = 10^{-6}$ and decreasing to 0 when $A = -10^{-6}$ for a one-sided rejection region. This irregularity is a consequence of the inconsistency of the standard OLS estimator in (11): a test based on a consistent estimator would continue to reject values of A very close to the null hypothesis up to the point when departures from the null hypothesis reach the Pitman local alternative (defined by the consistency rate of the estimator on which the test statistic is based) with correct probability of Type I error. In view of the above, the starting point of our analysis will be to obtain a corrected version of the OLS estimator in (11), see the estimator in (12) below, that achieves consistency along the entire parameter space of A.

One way to proceed would be to employ a deterministic K-dependent correction to the estimator in (11). However, Proposition 1 shows that the validity of such corrections is conditional upon a priori knowledge of the predictor variables' persistence degree κ_n and of the relative magnitude of κ_n and the horizon K. Since the persistence degree is unknown, a correction of this type is not feasible along the classes P(i)-P(iii) of predictor variables. A similar situation applies to the t-statistic: a straightforward calculation, shows that[1] $t_{OLS} = O_p\left(K^{-1/2}(K \wedge \kappa_n)\left(\frac{K}{\kappa_n} \vee 1\right)^{1/2}\right)$ under the null hypothesis $A = 0$. As a result, deterministic rescaling such as $t_{SCALED} = K^{-1/2} t_{OLS}$, considered inter alia by Hjalmarsson (2011), works for nonstationary predictors but may distort inference when the predictor exhibits (near-) stationary characteristics.

To obtain a consistent estimator for long-horizon regressions without the complications arising from deterministic scalings that are dependent on the unknown stochastic properties of the predictor variables, we introduce a stochastic modification to the OLS estimator in (11):

$$\tilde{A}_K^{MOLS} = \underline{Y}(K)' \underline{X}_{n_K-1} \left[\underline{X}(K)' \underline{X}_{n_K-1}\right]^{-1}. \tag{12}$$

In effect, the modification in (12) amounts to adjusting the regression signal matrix from $\underline{X}_{n_K-1}' \underline{X}_{n_K-1}$ to $\underline{X}(K)' \underline{X}_{n_K-1}$; this stochastic adjustment produces a consistent estimator of A along the entire parameter space, as opposed to $\hat{A}_K^{OLS}$ which is only consistent when $A = 0$.

Despite the consistency of $\tilde{A}_K^{MOLS}$, the limit distribution of $\tilde{A}_K^{MOLS} - A$ (even under the suitable

[1] Using the consistency of $K^{-1}\hat{A}_K^{OLS}$ by Proposition 1 and Lemma A4 in the Appedix, it is easy to show that the sample mean of the squared residuals from (8) is of order $O_p(K)$.

normalisation) will not be mixed Gaussian in the case of unit root and local-to-unity regressors, and will, in all cases, depend on the nuisance parameter C in (3). Since consistent estimation of C is not possible, tests that are used to conduct long-horizon inference on A are valid only when the rate of regressor persistence is assumed to be known, i.e., when there is *a priori* knowledge that x_t belongs to one of the persistence classes P(i)-P(iii) above.

However, given the high degree of persistence that characterizes most of the popular predictors, one cannot possess such knowledge, casting doubt on the validity of inference. Therefore, an alternative inference procedure that is robust to the persistence properties of the predictor variables is desirable. The IVX framework of Phillips and Magdalinos (2009), which has been adapted by KMS for short-horizon return predictability tests, achieves the required robustness and provides valid inference for predictors with arbitrary persistence properties.

Our study extends the IVX methodology in a way that delivers robust inference in long-horizon predictive regression systems when there is no a priori knowledge of whether x_t belongs to class P(i), P(ii) or P(iii). The key idea of this methodology is to construct an instrumental variable whose degree of persistence we explicitly control. In this way, the inference problems arising due to the uncertainty regarding the persistence of the original regressor are avoided. Using the constructed instrument, one then performs a standard instrumental variable estimation. The derived estimator asymptotically follows a mixed normal distribution, and hence the corresponding Wald statistic asymptotically follows a chi-squared distribution under the null, considerably simplifying inference.

Specifically, we devise near-stationary instruments belonging to the class P(ii) by differencing the regressor x_t and constructing a new process according to an artificial autoregressive matrix with a specified degree of persistence. Despite the fact that the difference $\Delta x_t = u_t + \frac{C_n}{\kappa_n} x_{t-1}$ is not an innovation unless the regressor belongs to the class of $I(1)$ processes (P(i) with $C = 0$), it behaves asymptotically as an innovation after linear filtering by a matrix consisting of near-stationary roots of the type P(ii). Choosing an artificial matrix,

$$R_{nz} = I_r + \frac{C_z}{\kappa_{nz}}, \quad \kappa_{nz} \to \infty, \ \kappa_{nz}/n \to 0 \text{ and } C_z < 0, \tag{13}$$

IVX instruments $\tilde{z}_t$ are constructed as a first-order autoregressive process with autoregressive matrix R_{nz} and innovations Δx_t,

$$\tilde{z}_t = R_{nz}\tilde{z}_{t-1} + \Delta x_t, \tag{14}$$

initialized at $\tilde{z}_0 = 0$.

Given a consistent OLS estimator such as $\tilde{A}_K^{MOLS}$, the IVX estimator is constructed as a feasible instrumental variables estimator that replaces the regressor x_t by the instrument $\tilde{z}_t$ in (12) in a standard way:

$$\tilde{A}_K^{IVX} = \underline{Y}(K)' \tilde{Z}_{n_K-1} \left[\underline{X}(K)' \tilde{Z}_{n_K-1}\right]^{-1}, \tag{15}$$

where $\tilde{Z}_{n_K-1} = \left[\tilde{z}_0', ..., \tilde{z}_{n_K-1}'\right]'$. Theorem 1 below shows that the normalised and centred IVX estimator in (15) is asymptotically mixed Gaussian under all empirically relevant persistence regimes P(i)-P(iii) for the predictor variables, implying a standard chi-squared limit distribution for the associated IVX-Wald test statistic (Theorem 2). The distributional invariance of the IVX-Wald test statistic to the stochastic properties of the regressors makes it suitable for general application.

In addition, the consistency of the IVX estimator over the entire parameter space allows testing general hypotheses on the parameter matrix A; this is in contrast to testing procedures based on the inconsistent OLS estimator (11), whether applied directly or combined with Bonferroni confidence interval construction for C, where the inconsistency of this estimator limits the range of testable hypotheses to the null $A = 0$. This is a particularly important limitation in multivariate predictive regression models, where the joint null hypothesis $A = 0$ cannot be used to test the individual significance of a predictor in the presence of other predictors. As a result, in addition to providing robust inference to the regressors' stochastic properties, the IVX-Wald test extends the range of testable hypotheses in multivariate long-horizon regression models.

2.3 IVX Asymptotic Inference in Long-Horizon Regressions

In this section, we present the asymptotic properties of the IVX estimator (15) and the corresponding IVX-Wald test statistic that delivers robust inference for long-horizon return predictability tests. Writing out the centred long-horizon IVX estimator $\tilde{A}_K^{IVX}$ in (15)

$$\tilde{A}_K^{IVX} - A = \sum_{t=1}^{n_K} \left[\varepsilon_t(K) - \bar{\varepsilon}_{n_K}(K)\right] \tilde{z}_{t-1}' \left[\underline{X}(K)' \tilde{Z}_{n_K-1}\right]^{-1}, \tag{16}$$

we observe that asymptotic mixed normality of $\tilde{A}_K^{IVX} - A$ requires establishing a central limit theorem for the sample covariance $\sum_{t=1}^{n_K} \varepsilon_t(K) \tilde{z}_{t-1}'$, under suitable normalisation. Unlike the short-horizon case, autocorrelation in the $\varepsilon_t(K)$ sequence when $K > 1$ implies that the above sample moment is not a martingale array. Under Assumption H, however, it is possible to obtain a martingale approximation of $\sum_{t=1}^{n_K} \varepsilon_t(K) \tilde{z}_{t-1}'$ as we now show. Changing the order of summation we obtain

$$\begin{aligned}
\sum_{t=1}^{n_K} \varepsilon_t(K) \tilde{z}_{t-1}' &= \sum_{t=1}^{n_K} \sum_{i=t}^{t+K-1} \varepsilon_i \tilde{z}_{t-1}' \\
&= \sum_{i=1}^{K-1} \varepsilon_i \sum_{t=1}^{i} \tilde{z}_{t-1}' + \sum_{i=K}^{n_K-1} \varepsilon_i \sum_{t=i-K+1}^{i} \tilde{z}_{t-1}' + \sum_{i=n_K}^{n} \varepsilon_i \sum_{t=i-K+1}^{n_K} \tilde{z}_{t-1}' \\
&= \sum_{i=1}^{K-1} \varepsilon_i \sum_{t=1}^{i} \tilde{z}_{t-1}' + \sum_{i=0}^{n-2K} \varepsilon_{i+K} \tilde{z}_i'(K) + \sum_{i=0}^{K-1} \varepsilon_{n-i} \sum_{t=0}^{i} \tilde{z}_{n_K-t-1}'.
\end{aligned} \tag{17}$$

All terms on the right of (17) are matrix valued martingale arrays and, since under Assumption H $n - 2K$ dominates K, the leading term of (17) will be the second term $\sum_{i=0}^{n-2K} \varepsilon_{i+K} \tilde{z}_i'(K)$. After vectorisation, this term becomes a martingale array with conditional variance matrix given by

$$\tilde{V}_n = \sum_{i=0}^{n-2K} \left[\tilde{z}_i(K) \tilde{z}_i'(K) \otimes H_{i+K}\right], \tag{18}$$

where $H_t = E_{\mathcal{F}_{t-1}}(\varepsilon_t \varepsilon_t')$. Denoting the vectorised version of the numerator of (16) by

$$\phi_n = (n\varkappa_n)^{-1/2} \sum_{t=1}^{n_K} \left\{\tilde{z}_{t-1} \otimes \left[\varepsilon_t(K) - \bar{\varepsilon}_{n_K}(K)\right]\right\} \tag{19}$$

where $\varkappa_n = (\kappa_n \wedge \kappa_{nz}) \left[K \wedge (\kappa_n \wedge \kappa_{nz})\right] \left[K \wedge (\kappa_n \vee \kappa_{nz})\right]$, Lemma 1 below formally establishes the martingale approximation discussed in (17) (see the asymptotic equivalence in (20)) and derives the limit distribution of ϕ_n by applying a martingale central limit theorem to the leading term of (20). We denote by $\mathbf{1}\left\{\frac{\kappa_{nz}}{\kappa_n} \to 0\right\}$ the indicator function that takes value 1 if $\frac{\kappa_{nz}}{\kappa_n} \to 0$ and 0 otherwise.

Lemma 1. *Under Assumptions P, INNOV and H,*

$$\phi_n = (n\varkappa_n)^{-1/2} \sum_{i=0}^{n-2K} \left[\tilde{z}_i(K) \otimes \varepsilon_{i+K}\right] + o_p(1) \Rightarrow N(0, V) \tag{20}$$

where $V = V_{\tilde{z}} \otimes \Sigma_{\varepsilon\varepsilon}$ *under Assumption INNOV(i) or Assumption INNOV(ii) with* $\kappa_n \to \infty$ *or* $K \to \infty$ *with* $V_{\tilde{z}}$ *given by:* **(i)** $V_{\tilde{z}} = V_{C_z} \mathbf{1}\left\{\frac{\kappa_{nz}}{\kappa_n} \to 0\right\} + V_C \mathbf{1}\left\{\frac{\kappa_n}{\kappa_{nz}} \to 0\right\}$ *when* $K/(\kappa_n \wedge \kappa_{nz}) \to 0$; **(ii)** $V_{\tilde{z}} = C_z^{-1} \Omega_{uu} \left(C_z^{-1}\right)' \mathbf{1}\left\{\frac{\kappa_{nz}}{\kappa_n} \to 0\right\} + C^{-1} \Omega_{uu} \left(C^{-1}\right)' \mathbf{1}\left\{\frac{\kappa_n}{\kappa_{nz}} \to 0\right\}$ *when* $K/(\kappa_n \wedge \kappa_{nz}) \to \infty$ *and*

$K/(\kappa_n \vee \kappa_{nz}) \to 0$; **(iii)** $V_{\tilde{z}} = 2C_z^{-1}V_C\left(C_z^{-1}\right)' \mathbf{1}\left\{\frac{\kappa_{nz}}{\kappa_n} \to 0\right\} + 2C^{-1}V_{C_z}\left(C^{-1}\right)'\mathbf{1}\left\{\frac{\kappa_n}{\kappa_{nz}} \to 0, \kappa_n \to \infty\right\}$ *when* $K/(\kappa_n \vee \kappa_{nz}) \to \infty$, *where* $V_C = \int_0^\infty e^{rC}\Omega_{uu}e^{rC'}dr$ *and* $V_{C_z} = \int_0^\infty e^{rC_z}\Omega_{uu}e^{rC_z'}dr$. *Under Assumption INNOV(ii) with* $\kappa_n = 1$ *and* K *fixed,* $V = W_{0,K} = \sum_{j,l=0}^{K-1} E\left(x_{0,j}x_{0,l}' \otimes \varepsilon_K\varepsilon_K'\right)$ *where* $x_{0,t} = \sum_{j=0}^{\infty} R^j u_{t-j}$ *with* $R = I_r + C$ *is a stationary process.*

Lemma 1 establishes a Gaussian asymptotic distribution for the "numerator" ϕ_n of the centred IVX estimator in (16) under all persistence regimes of Assumption P, including (near) nonstationarity. Since $\Omega_{uu} > 0$ and C and C_z are negative stable matrices, the matrices V_C and V_{C_z} are well-defined and positive definite. Consequently, $V_{\tilde{z}}$ and the asymptotic covariance matrix V in (20) are positive definite. It is worth noting that the asymptotic covariance matrix V in (20) admits a convenient signal/noise covariance factorisation $V_{\tilde{z}} \otimes \Sigma_{\varepsilon\varepsilon}$ in all but one cases of Lemma 1, the exception being the combination of a stationary regressor satisfying P(iii), a conditionally heteroskedastic innovation process ε_t and a fixed horizon K. In this case, the IVX instrument $\tilde{z}_i(K)$ can be approximated by the regressor $x_i(K)$ in ϕ_n, so the stationarity of the regressor process and the finite horizon fail to eliminate the GARCH effects present in the innovation process.

The next result derives the limit distribution of the "denominator" of (16) and establishes the asymptotic relevance condition for the IVX instrumentation in the long-horizon regression case.

Lemma 2. *Under Assumptions P, INNOV, H,* $n^{-1}(K \wedge \kappa_n)^{-1}(\kappa_n \wedge \kappa_{nz})^{-1}\underline{X}(K)'\tilde{Z}_{n_K-1} \Rightarrow \Psi$, *where the limit matrix is given by:* **(i)** $\Psi = -\tilde{\Psi}_{uu}C_z^{-1}\mathbf{1}\{\kappa_{nz}/\kappa_n \to 0\} + V_C\mathbf{1}\{\kappa_n/\kappa_{nz} \to 0\}$ *when* $K/\kappa_n \to 0$; **(ii)** $\Psi = -V_C\left(C_z^{-1}\right)'\mathbf{1}\{\kappa_{nz}/\kappa_n \to 0\} - C^{-1}V_C\mathbf{1}\{\kappa_n/\kappa_{nz} \to 0\}$ *when* $K/\kappa_n \to \infty$ *and* $\kappa_n \to \infty$; **(iii)** $\Psi = \sum_{i=0}^{K-1}\Gamma_{x_0}(i)$ *with* $\Gamma_{x_0}(k) = E\left(x_{0t}x_{0t-k}'\right)$ *when* $\kappa_n = 1$, *where* V_C, V_{C_z} *and* $x_{0,t}$ *are defined in Lemma 1,* $\tilde{\Psi}_{uu} = \Omega_{uu} + \int_0^1 \underline{J}_C(t)\,dB_u(t)' + \int_0^1 \underline{J}_C(t)\,\underline{J}_C(t)'\,dtC'$ *under Assumption P(i) with* $C = 0$ *when* $\kappa_n/n \to \infty$, *and* $\tilde{\Psi}_{uu} = \Omega_{uu} + V_CC'$ *under Assumption P(ii).*

As in the short horizon case, the limit distribution of the normalised signal matrix is random in the near-I(1) case P(i) and constant in P(ii) and P(iii). By Lemma 3.1(v) of Magdalinos and Phillips (2020), $\det\left(\tilde{\Psi}_{uu}\right) > 0$ *a.s.* under P(i) and P(ii), so Lemma 2 above guarantees the asymptotic relevance condition of the instrumental variable estimator in (15).

By combining Lemma 1 and Lemma 2, we conclude that the long-horizon IVX estimator in (15)/(16) has a mixed Gaussian asymptotic distribution under all persistence regimes of Assumption P, given by the theorem below.

Theorem 1. *Consider the model (1)-(3) under Assumptions P, INNOV and H and the estimator* $\tilde{A}_K^{IVX}$ *in (15). Denoting* $\lambda_n = (K \wedge \kappa_n)(\kappa_n \wedge \kappa_{nz})^{1/2}\left[K \wedge (\kappa_n \wedge \kappa_{nz})\right]^{-1/2}\left[K \wedge (\kappa_n \vee \kappa_{nz})\right]^{-1/2}$,

$$\sqrt{n}\lambda_n vec\left(\tilde{A}_K^{IVX} - A\right) \Rightarrow MN\left(0, \left[\left(\Psi^{-1}\right)' V_{\tilde{z}}\Psi^{-1}\right] \otimes \Sigma_{\varepsilon\varepsilon}\right) \quad \text{as } n \to \infty$$

under Assumption INNOV(i) or under Assumption INNOV(ii) with $\kappa_n \to \infty$ *or* $K \to \infty$ *with* $V_{\tilde{z}}$ *and* Ψ *defined in Lemmas 1 and 2. Under Assumption INNOV(ii) with* $\kappa_n = 1$ *and fixed* K,

$$\sqrt{n}vec\left(\tilde{A}_K^{IVX} - A\right) \Rightarrow N\left(0, \left(\left(\Psi^{-1}\right)' \otimes I_m\right) W_{0,K}\left(\Psi^{-1} \otimes I_m\right)\right)$$

where $W_{0,K}$ *is defined in Lemma 1.*

When the horizon parameter is dominated by the persistence degree of both the regressor and the instrument, $K/(\kappa_n \wedge \kappa_{nz}) \to 0$, $\lambda_n = (\kappa_n \wedge \kappa_{nz})^{1/2}$ and the asymptotic behaviour of the long-horizon IVX estimator is identical to that of its short-horizon counterpart (Theorem A of KMS). On the other hand, when the regressor is not highly persistent, κ_n may be dominated by the horizon parameter K; if, in addition, K is dominated by κ_{nz} (as it is likely, since we set $\kappa_{nz} = n^{0.95}$), the horizon parameter appears in the denominator of the normalisation λ_n: $\lambda_n = \kappa_n/\sqrt{K}$ when $K/\kappa_n \to \infty$ and $K/\kappa_{nz} \to 0$. In this case, the consistency rate of the IVX estimator and, hence, the power of the IVX-Wald test decreases with the horizon K, a feature that explains our power simulation results in Section 3.

The asymptotic mixed normality property of the long-horizon IVX estimator implies that linear restrictions on the coefficients A generated by the system of predictive equations (1) can be tested by a standard Wald test based on the IVX estimator for all persistence scenaria conforming to the classes P(i)–P(iii). In particular, we consider a set of linear restrictions

$$H_0 : H\text{vec}\left(A\right) = h, \tag{21}$$

where H is a known $q \times mr$ matrix with rank q and h is a known vector. Since, by Theorem 1, the asymptotic variance of $H\text{vec}\Big(\tilde{A}_K^{IVX} - A\Big)$ can be estimated by $Q_{H,K} = H\{[\underline{X}'\left(K\right)P_{\tilde{Z},K}\underline{X}\left(K\right)]^{-1} \otimes \hat{\Sigma}_{\varepsilon\varepsilon}\}H'$ where $P_{\tilde{Z},K} = \tilde{Z}_{n-K}\left[\tilde{Z}\left(K\right)'\tilde{Z}\left(K\right)\right]^{-1}\tilde{Z}'_{n-K}$ and $\tilde{Z}\left(K\right) = \left[\tilde{z}'_0\left(K\right), ..., \tilde{z}'_{n_K-1}\left(K\right)\right]'$, then the IVX-Wald test statistic

$$W_K^{IVX} = \left[H\text{vec}\left(\tilde{A}_K^{IVX}\right) - h\right]' Q_{H,K}^{-1}\left[H\text{vec}\left(\tilde{A}_K^{IVX}\right) - h\right], \tag{22}$$

has a standard chi-squared limit distribution under H_0 in (21).

We actually propose and empirically implement an asymptotically equivalent modification of the above statistic, which possesses better finite-sample properties in the presence of an intercept in (1) and (8):

$$\tilde{W}_K^{IVX} = \left[H\text{vec}\left(\tilde{A}_K^{IVX}\right) - h\right]' \tilde{Q}_{H,K}^{-1}\left[H\text{vec}\left(\tilde{A}_K^{IVX}\right) - h\right], \tag{23}$$

where $\tilde{Q}_{H,K} = H[(\tilde{Z}'_{n-K}\underline{X}\left(K\right))^{-1} \otimes I_m]\mathbb{M}_K[(\underline{X}\left(K\right)'\tilde{Z}_{n-K})^{-1} \otimes I_m]H'$, $\mathbb{M}_K = \tilde{Z}\left(K\right)'\tilde{Z}\left(K\right) \otimes \hat{\Sigma}_{\varepsilon\varepsilon} - n_K\bar{z}_{n_K-1}\left(K\right)\bar{z}'_{n_K-1}\left(K\right) \otimes \hat{\Omega}_{FM}$, $\bar{z}_{n_K-1}\left(K\right) = n_K^{-1}\sum_{t=1}^{n_K}\tilde{z}_{t-1}\left(K\right)$ and $\hat{\Omega}_{FM} = \hat{\Sigma}_{\varepsilon\varepsilon} - \hat{\Omega}_{\varepsilon u}\hat{\Omega}_{uu}^{-1}\hat{\Omega}'_{\varepsilon u}$, $\hat{\Sigma}_{\varepsilon\varepsilon}$ is the standard OLS estimator and $\hat{\Omega}_{\varepsilon u}$ and $\hat{\Omega}_{uu}$ are long run covariance estimators defined below (7). We refer to KMS (p. 1516 and Remark A(2) in p. 1549) for a justification of the better finite sample properties of the modified statistic in (23) in the short-horizon case; the same argument essentially applies to the long-horizon case.

Theorem 2. *Consider the model (1)-(3) under Assumption H. Then, the IVX-Wald statistics in (22) and (23) for testing (21) are asymptotically equivalent and satisfy* $\tilde{W}_K^{IVX} \Rightarrow \chi^2\left(q\right)$ *as* $n \to \infty$ *under* H_0 *for the following classes of predictor processes* x_t *in (2):* **(i)** *P(i)-P(iii) under Assumption INNOV(i);* **(ii)** *P(i)-P(iii) under Assumption INNOV(ii) when* $K \to \infty$*;* **(iii)** *P(i)-P(ii) under Assumption INNOV(ii) when the horizon parameter* K *is fixed.*

Theorem 2 proposes a hypothesis testing procedure in long-horizon predictive regressions with the following advantages: (a) the procedure accommodates a very large class of predictors, ranging from purely stationary to unit root processes and including all intermediate persistence regimes; (b) the proposed IVX-Wald statistic may be used to test the predictive power of vector-valued regressors; (c) the proposed IVX-Wald statistic may be used to test general linear hypotheses on the matrix parameter A, which allows us to assess the predictive power of a subset of regressors (in the presence of other potential predictors). The possibility to conduct predictability tests for subsets of regressors is a consequence of employing an long-horizon IVX estimator that is consistent over the entire parameter space of A rather than just at the point $A = 0$ as is usually the case for long-horizon predictability tests (see the discussion following Proposition 1).

The only combination of Assumptions P and INNOV not covered by Theorem 2 is that of a purely stationary regressor satisfying P(iii), a conditionally heteroskedastic innovation process ε_t and a fixed horizon K. In this case, the long-horizon IVX estimator becomes asymptotically equivalent to its OLS counterpart which implies that the standard t or Wald statistics are asymptotically invalid; a White-type correction for the regression residuals (that accounts for the presence of GARCH effects in ε_t): see the discussion in p.1515-1516 in KMS and Theorem 4.4. of Magdalinos (2020). Applying such a heteroskedasticity correction to the IVX-Wald statistic will extend the validity of Theorem 2(iii) to the entire range of Assumption P.

We conclude the section with a discussion of the implementation of the IVX-Wald test in (23) with regards to the choice of instrument $\tilde{z}_t$ in (13) or, equivalently, the choice of R_{nz} in (13). Normalising $C_z = -I_r$, this reduces to a choice of the instrument persistence parameter κ_{nz}. It is well documented in the (short-horizon) IVX literature that there is a size-power trade-off in choosing κ_{nz} when the regressor process is near-I(1): a correctly sized critical region for the IVX-Wald statistic requires $\kappa_{nz}/n \to 0$ whereas the statistic's divergence rate under the alternative hypothesis increases with κ_{nz}. This trade-off continues to apply in the long-horizon case, as Theorem 1 shows: for a near-I(1) regressor satisfying $K/\kappa_{nz} \to 0$ the normalisation of the long-horizon IVX estimator $\tilde{A}_K^{IVX}$ in Theorem 1 is $\sqrt{n\kappa_{nz}}$. Our selection of κ_{nz} follows the same approach as KMS: (i) we correct for the finite sample effects on the size of the IVX-Wald test by employing the statistic (23); (ii) following the analysis of the remainder term arising from the intercept in Remark A(2) of KMS, we note that these finite sample effects are most prominent when x_t is a unit root process with innovations u_t highly correlated with the innovations ε_t of (1); (iii) we conduct Monte Carlo simulations for the size of the test in (23) under the worst scenario described in (ii) and select the largest value of κ_{nz} that keeps the empirical size of (23) sufficiently close to the nominal size. We find that the value $\kappa_{nz} = n^{0.95}$ employed in the short-horizon IVX-Wald test of KMS extends κ_{nz} as far as possible in the direction of the $O(n)$ threshold while maintaining size control. We, therefore, generate the instrument $\tilde{z}_t$ by selecting $C_z = -I_r$ and $\kappa_{nz} = n^{0.95}$ both for the Monte Carlo exercise and for the empirical implementation of our procedure.

3 Finite-Sample Properties of Long-Horizon Test Statistics

This section conducts an extensive Monte Carlo simulation, presenting the finite-sample properties of various test statistics that have been commonly used in long-horizon predictability studies. We consider the OLS t-statistic (t_{OLS}), its counterpart scaled by the square root of the predictive horizon K (t_{SCALED}), the t-statistic with Newey-West standard errors (t_{NW}), the t-statistic with Hansen-Hodrick standard errors (t_{HH}), the t-statistic with Hodrick (1992) standard errors (t_{HOD}) and the Bonferroni test statistic (t_{BONF}) proposed by Hjalmarsson (2011).[2] We compare the finite-sample properties of these test statistics with the properties of the long-horizon IVX-Wald statistic (W_{IVX}). This analysis covers a wide range of values for the parameters that determine the properties of these test statistics. Hence, this analysis can serve as a guide for the suitability of each test statistic for the particular combination of predictor(s) and equity factor in hand.

3.1 Univariate Long-Horizon Predictive Regressions

Starting with the univariate predictive regression setup, we use the following DGP, where y_t and x_t are scalars: (1) with $\varepsilon_t \sim NID(0,1)$ and (2) with $R_n = 1 + C/n$, $u_t = \phi u_{t-1} + e_t$ and $e_t \sim NID(0,1)$. The system is initialized at $x_0 = 0$. The IVX estimator and the corresponding Wald statistic are invariant to the value of μ, so we opt for $\mu = 0$. Equipped with the simulated data, we estimate the corresponding long-horizon predictive regressions, as in equation (8), for various horizons K. We consider two-sided tests with nominal size 5% for all statistics, corresponding to the null hypothesis of no predictability, i.e., that the slope coefficient of the predictive regressor in the DGP is equal to zero, $H_0 : A = 0$ in (1).[3]

To examine the power of each statistic, we consider the following sequence of alternatives:

$$A = \frac{b}{n}\sqrt{1-\delta^2} \text{ for } b \in \{0,\ 2,\ ..,\ 12,\ 16,\ ..,\ 32,\ 40,\ 60,\ 100,\ 500,\ 1,000\}, \tag{24}$$

with $b = 0$ corresponding to the size of each test and $\delta = E(\varepsilon_t u_t)$ denoting the contemporaneous correlation coefficient between ε_t and u_t, which measures the degree of predictor's "endogeneity".

[2] Following common practice in the literature, we use K lags to compute the corresponding Newey-West and Hansen-Hodrick standard errors.

[3] Under this null hypothesis, the slope coefficient of the long-horizon predictive regression (A_K) should also be equal to zero for all horizons K.

To examine the empirical size of the test statistics, each simulation experiment reports rejection rates of the null hypothesis using $10,000$ repetitions. For their power properties, each simulation experiment uses $1,000$ repetitions. We report results for various sample sizes n with relevant predictive horizons K, alternative values for the local-to-unity parameter C, different degrees of endogeneity δ as well as alternative autocorrelation coefficient ϕ of the innovation of the autoregression in (2).

Different sample sizes n correspond to the use of monthly, quarterly or annual data in empirical tests. The use of a broad but empirically relevant range of values for the local-to-unity parameter C can reveal how the degree of regressor persistence affects the finite-sample properties of these long-horizon test statistics. Different degrees of endogeneity δ correspond to different combinations of predictors and equity factors. For example, in case price-scaled financial ratios are used to predict market returns, then, by construction, δ is bound to be close to 1 in absolute value; to the contrary, term structure variables exhibit a much lower degree of endogeneity. Since the magnitude of the Stambaugh bias is primarily determined by the interaction of the predictor's persistence and its endogeneity, we consider various combinations of these parameter values to examine how this interaction affects the properties of the test statistics as the predictive horizon increases.

3.1.1 Size Properties

Table 1 presents the size of the tests for $n = 1,000$ and $\phi = 0$, whereas Tables IA.1 and IA.2 in the Internet Appendix report the tests' sizes for $n = 100$ and $n = 500$, respectively. In addition, Tables IA.3-IA.5 show the corresponding simulation results when $\phi = 0.25$.

A number of important conclusions can be drawn from these simulation results. First, we confirm that the commonly used tests (t_{OLS}, t_{NW}, t_{HH}) become severely oversized as the predictive horizon increases. Their overrejection becomes extreme in the case of predictors that are both highly persistent and exhibit a very strong degree of endogeneity. The oversizing of these test statistics in long-horizon predictive regressions is much more severe than the oversizing caused by the Stambaugh bias in the short-horizon setup. In other words, this bias is propagated as the predictive horizon increases, leading to severe overrejection of the null hypothesis.

Second, our simulation results show that the oversizing of these test statistics appears even in the cases where the Stambaugh bias would not be a concern. In particular, if the degree of endogeneity δ is equal to zero, the size of t_{OLS}, t_{NW}, and t_{HH}, appears to be correct in the short-horizon setup, even for unit root predictors. However, as the predictive horizon increases, these tests substantially overreject the null hypothesis relative to their nominal size. Hence, using overlapping observations in long-horizon predictive regressions can have a dramatic impact on the size properties of these test statistics. This evidence yields the conclusion that t_{OLS}, t_{NW}, and t_{HH} could lead to spurious inference due to oversizing for all persistent predictors, as long as the predictive horizon is sufficiently long, regardless of their degree of endogeneity with respect to factor returns. These conclusions hold across the examined sample sizes.

Third, scaling t_{OLS} by the square root of the predictive horizon to compute t_{SCALED} can only partially address its severe oversizing due to the overlapping nature of observations in long-horizon regressions. On the one hand, this approach obviously cannot address the oversizing due to the Stambaugh bias that carries over to longer horizons. On the other hand, this adjustment may actually lead to undersizing as the horizon increases, because t_{OLS} increases at a rate lower than the square root of the horizon when the predictor is less persistent than a (near-) unit root process.

Another conclusion of our analysis refers to the performance of t_{HOD}, which is suggested to have good size properties at long horizons (see Ang and Bekaert, 2007; and Wei and Wright, 2013). We find that an increase in the predictive horizon has only a mild effect on the size of t_{HOD}. Hence, as expected by the construction of the test, the size of t_{HOD} is not severely affected by the overlapping nature of observations in long-horizon predictive regressions. However, this test statistic can still exhibit severe oversizing due to the Stambaugh bias. In the case of highly persistent and endogenous predictors, t_{HOD} tends to overreject across all predictive horizons considered. This finding holds regardless of the sample size. As a result, even though t_{HOD} overall performs better than t_{NW} or t_{HH}, it can still lead to spurious inference in both short- and long-horizon predictive regressions.

The previous finding highlights the dramatic impact of the predictor's time series properties on tests' size. The uncertainty surrounding the exact type of persistence of commonly used predictors necessarily raises doubts about the appropriateness of traditional test statistics, including t_{HOD}.

This uncertainty motivates the use of W_{IVX}, which yield inference that is robust to the exact type of predictor's persistence. More generally, the poor size properties of the commonly used test statistics motivates the examination of the properties of alternative test statistics that are devised to deal with arbitrarily persistent and endogenous predictors.

To this end, t_{BONF} exhibits correct size, regardless of the horizon length, when the predictor follows a unit root or a near-unit root process. This is not surprising because this test statistic is constructed under the assumption that the predictor follows a local-to-unity process. However, as we deviate from the unit root, t_{BONF} seems to be somewhat undersized; this feature becomes more pronounced when the horizon is substantially long. Hence, t_{BONF} is bound to be very conservative at long horizons. We consider further this feature when we subsequently examine the power properties of the test.

Last but not least, W_{IVX} exhibits very good size properties regardless of the predictor's persistence and degree of endogeneity. Its empirical size is very close to the nominal 5% level even when the predictor follows a unit root or a near-unit root process. These good size properties remain intact when we consider sufficiently long predictive horizons. Moreover, W_{IVX} retains its very good size control across the examined sample sizes as well as in the presence of autocorrelation in the residuals of the autoregression. Hence, we conclude that W_{IVX} is affected neither by the Stambaugh bias nor by the overlapping nature of observations in long horizons, and it can yield inference that is robust to the exact type of predictor's persistence.

3.1.2 Power Properties

Next, we compare the finite-sample power properties of t_{HOD}, t_{BONF}, and W_{IVX}. In particular, for each of these three tests, we plot the rejection rate of the null hypothesis $H_0 : A = 0$, as the true value of the slope coefficient A in (1) increases according to the sequence in (24). Figure 1 plots these rejection rates when the sample size is $n = 1,000$ and the predictive horizon is $K = 10$.

We find that both t_{BONF} and W_{IVX} are powerful test statistics across all cases examined. Their rejection rates monotonically and rapidly increase, as the true value of the slope coefficient deviates from zero. In relative terms, t_{BONF} is particularly powerful when the predictor is a unit root process ($C = 0$) and exhibits a strong degree of endogeneity ($\delta = -0.99$). To the contrary, W_{IVX} becomes more powerful than t_{BONF} as the predictor becomes slightly less persistent. This finding is a consequence of the fact that t_{BONF} becomes somewhat undersized as we deviate from the unit root. The power of W_{IVX} also appears to be very similar to that of t_{HOD} when the empirical size of the latter is close to its nominal 5% level. In fact, in the case of no endogeneity ($\delta = 0$), where t_{HOD} is correctly sized, the power plots of these three test statistics appear to be almost indistinguishable across the three local-to-unity parameter values that we consider.

We derive very similar conclusions regarding the absolute and relative power properties of these test statistics when we alternatively consider a much longer predictive horizon, such as $K = 50$. The corresponding power plots are illustrated in Figure IA.1. Figures IA.2 and IA.3 yield very similar patterns, using sample size $n = 100$ and predictive horizons $K = 3$ and $K = 5$, respectively.

Our simulation setup also allows us to shed light on an important issue surrounding the use of long-horizon predictive regressions. An implicit argument in the prior literature is that long-horizon returns are less noisy, and hence long-horizon predictability tests would be more powerful (see Campbell, 2001). We examine this argument by comparing the power properties of the correctly-sized test statistics across different predictive horizons.

Figure IA.4 presents the power plots for W_{IVX} with sample size $n = 1,000$ and horizons $K = 1, 50$, and 100. It is evident that the power of this test statistic *decreases* as the predictive horizon increases. This is true for all local-to-unity parameter values C and degrees of endogeneity δ considered. Figure IA.5 similarly illustrates the inverse relationship between the power of W_{IVX} and the length of the predictive horizon using sample size $n = 100$.

Is this inverse relationship between power and horizon a feature of W_{IVX} only? To answer this question, we repeat this analysis for t_{HOD} and t_{BONF}. Figures IA.6 and IA.7 show their power plots when $n = 1,000$ and $K = 1, 50$, and 100, whereas Figures IA.8 and IA.9 present the corresponding power plots when $n = 100$ and $K = 1, 5$, and 10. We find the exact same inverse relationship that we reported for W_{IVX}; the power of t_{HOD} and t_{BONF} decreases as the predictive horizon increases, in all cases considered.

In conclusion, the argument that long-horizon predictability tests are more powerful than the short-horizon ones seems to be a misperception. This is most likely driven by the spurious evidence of strong long-horizon predictability reported in prior studies, due to the use of test statistics that become severely oversized as the horizon increases. To the contrary, when a correctly-sized test statistic is employed, we show that long-horizon predictive regressions actually lead to power *loss*. Taken together, our simulation analysis casts doubt on the incremental benefit of conducting long-horizon predictability tests.[4]

3.2 Conditionally Heteroskedastic DGP

We have also performed a simulation analysis using a GARCH(1,1) process for the innovations of the predictive regression model. The simulation setup is formally presented in the Internet Appendix. Tables IA.9, IA.10, and IA.11 present the sizes of the various test statistics for sample size $n = 100$, $n = 500$, and $n = 1,000$, respectively.

The main conclusions we derived using a homoskedastic DGP carry through in the case of heteroskedasticity. The rejection rate of t_{OLS}, t_{NW}, and t_{HH} monotonically increases as the predictive horizon increases. As long as the predictor is persistent, these tests become severely oversized at long horizons, regardless of the predictor's degree of endogeneity. The size properties of t_{HOD} also remain similar to the ones reported in the case of homoskedasticity. This test is substantially oversized when the predictor exhibits both very high degree of persistence and strong endogeneity. We also find that t_{BONF} never becomes oversized. To the contrary, as we deviate from the unit root case, this test statistic becomes undersized at sufficiently long horizons. Equally importantly, this simulation analysis confirms that W_{IVX} retains its good size control under heteroskedasticity.[5]

3.3 Multivariate Long-Horizon Predictive Regressions

Empirical predictability tests are commonly conducted in the presence of multiple persistent predictors. Nevertheless, the properties of the commonly used test statistics are not well understood in the context of multivariate predictive regressions, especially when overlapping observations are employed. The subsequent analysis fills this gap.

We conduct Monte Carlo simulations using the following DGP that accommodates two regressors:

$$y_t = \mu + A' x_{t-1} + \varepsilon_t,\ x_t = (I_2 + C/n)\, x_{t-1} + u_t,\ u_t = \Phi u_{t-1} + e_t, \tag{25}$$

with $\Phi = diag\,(\phi_1, \phi_2)$, $\zeta_t = (\varepsilon_t, u_t')'$, $\Sigma = E\left(\zeta_t \zeta_t'\right)$ and ε_t and e_t being zero-mean Gaussian with covariance structure determined by Σ.

Equipped with this DGP, we consider three cases, which correspond to empirically relevant combinations of values for the local-to-unity parameters C, the residuals' autocorrelation coefficients in Φ, and the covariance matrix Σ. Specifically, Case I assumes that $C = diag(0, -5)$, whereas the values for Φ and Σ are estimated using log excess market returns as the regressand, the earnings-price ratio as the first predictor, and the T-bill rate as the second predictor. Case II assumes that $C = diag(0, -5)$, whereas the parameter values for Φ and Σ are estimated using log excess market returns, the dividend-price ratio, and the T-bill rate, respectively. Last, Case III assumes that $C = diag(0, -10)$, and the parameter values for Φ and Σ are estimated using log excess market returns, the earnings-price ratio, and the default spread, respectively. It can be confirmed by the descriptive statistics of these predictors, which are presented in the subsequent Section, that these three cases give rise to different combinations regarding the predictors' degree of endogeneity (δ)

[4] We have also examined the finite-sample properties of the test statistics for the case where the correlation coefficient (δ) between ε_t and u_t is positive. Tables IA.6, IA.7, and IA.8 present the corresponding size properties for $n = 100$, $n = 500$, and $n = 1,000$, respectively. These simulation results yield similar conclusions to the ones derived from the benchmark setup and confirm the excellent size properties of W_{IVX}. To shed further light on the impact of endogeneity, Figures IA.10, IA.11, and IA.12 illustrate the size of the various test statistics for sample size $n = 100$ and predictive horizons $K = 3$, $K = 5$, and $K = 10$, respectively. It is evident that the finite-sample size of W_{IVX} is not sensitive to the degree of endogeneity, whereas this affects the size of commonly used test statistics.

[5] We have also investigated the finite sample properties of the examined statistics under the possibility of conditional heteroskedasticity in the error term of the autoregressive part of the system. The corresponding DGP is formally presented in the Internet Appendix and Tables IA.12, IA.13, and IA.14 present the size properties of the test statistics for alternative sample sizes n. In sum, this simulation analysis yields similar conclusions to the ones derived from the benchmark DGP and confirms the very good size control of W_{IVX}.

and the autocorrelation coefficient (ϕ) in the residuals of the corresponding autoregression.

3.3.1 Size Properties

Table 2 presents the rejection rates of the test statistics for sample size $n = 1,000$. The corresponding parameter values have been estimated using monthly data. Table IA.15 presents the corresponding simulation results for $n = 100$, using annual data to estimate the relevant parameter values. We present the rejection rates of the statistics with respect to three different tests. Panel A tests the joint null hypothesis $H_0 : A = (0,0)$, i.e., that the slope coefficients of the predictors are jointly equal to zero. For this joint test we compute the empirical size of Wald statistics with alternative covariance estimation methods. We consider the standard least squares Wald statistic scaled by the length of the predictive horizon (W_{SCALED}) as well as Wald statistics with Newey-West (W_{NW}), Hansen-Hodrick (W_{HH}), and Hodrick (W_{HOD}) covariances. We also present the empirical size of the IVX-Wald statistic (W_{IVX}). Panel B tests the null hypothesis $H_0 : A_1 = 0$, i.e., that the slope coefficient of the first predictor is equal to zero, whereas Panel C tests the null hypothesis $H_0 : A_2 = 0$, i.e., that the slope coefficient of the second predictor is equal to zero. The latter two hypotheses are of interest because, in the presence of multiple predictors, one may wish to examine the individual significance of each predictor, not just their joint predictive ability. For these tests, apart from W_{IVX}, we also report the sizes for t_{SCALED}, t_{NW}, t_{HH}, and t_{HOD}.

A number of conclusions arise from this analysis. W_{NW} and W_{HH} have very poor size properties. Their oversizing becomes extreme as the predictive horizon increases due to the effect of overlapping observations. Hence, in the presence of a price-scaled ratio, which is highly persistent and exhibits a strong degree of endogeneity, these test statistics are bound to spuriously reject the joint null of no predictability, even when the second predictor is not endogenous and is less persistent. In fact, we observe that in the presence of two persistent regressors, the oversizing of the joint test is exacerbated at long horizons relative to the corresponding cases with a single predictor.

With respect to tests of individual significance, the empirical size of t_{NW} and t_{HH} follows the patterns we reported above in the case of univariate predictive regressions. The rejection rate of these statistics monotonically increases as the predictive horizon increases. This oversizing becomes more pronounced in the case of endogenous predictors, such as the price-scaled ratios (see Panel B), but the horizon effect leads to overrejection even when the predictor's degree of endogeneity is very low (see Panel C).

Moreover, we find that the simple adjustment of the OLS Wald statistic by the length of the predictive horizon to compute W_{SCALED} partially cancels out the horizon effect on the empirical size of the former. Nevertheless, this adjustment cannot address the oversizing that arises due to the Stambaugh bias and carries over to longer horizons. A direct implication of the reported rejection rates for W_{SCALED} is that the standard, non-scaled OLS Wald statistic is bound to almost certainly reject the joint null of no predictability, if the predictive horizon is sufficiently long. Hence, inference based on this test statistic would certainly be spurious. In addition, the properties of t_{SCALED} for tests of individual significance are very similar to the ones reported above in the univariate predictive setup.

Regarding W_{HOD}, we observe that when a price-scaled ratio is included as one of the predictors, joint tests of no predictability are typically oversized. This oversizing appears already in the single-period ($K = 1$) predictive regression and increases with the length of the horizon. However, the horizon effect is much less detrimental for W_{HOD} relative to W_{NW} and W_{HH}. In fact, the empirical size of W_{HOD} is very similar to the one of W_{SCALED}. With respect to tests of individual significance using t_{HOD}, the emerging patterns are again similar to the ones reported in the case of univariate predictive regressions. In sum, the presence of a persistent and endogenous predictor raises concerns regarding the suitability of W_{HOD} for joint predictability tests. To the contrary, W_{IVX} exhibits an empirical size that is very close to the nominal 5% level across the various cases considered, regardless of the horizon length. Hence, we conclude that W_{IVX} exhibits the best finite-sample size properties for both joint and individual significance tests in the context of multivariate long-horizon predictive regressions.

3.3.2 Power Properties

Given the very good size properties of W_{IVX}, we use this bivariate regression setup to examine the finite-sample power of this statistic. In particular, Figure 2 plots the rejection rate of the joint null hypothesis $H_0 : A = (0, 0)$ for each of the three combinations of parameter values described above (Cases I-III), using sample size $n = 1,000$ and horizons $K = 1, 10, 50$, and 100. Figure IA.13 presents the corresponding power plots for $n = 100$ and horizons $K = 1, 3, 5$, and 10. The left column reports the rejection rate as the true value of the slope coefficient of the first regressor (A_1) deviates from zero, whereas the right column presents the corresponding rejection rate when the true value of the second regressor's slope coefficient (A_2) is different from zero.

Both figures show that W_{IVX} is a powerful statistic for joint hypothesis tests. Its rejection rate rapidly increases as the slope coefficient of either of the two regressors increases. W_{IVX} is particularly powerful when the true value of the slope coefficient of the unit root predictor differs from zero. We get a very similar picture across the three cases we consider. Hence, the power properties of W_{IVX} are not materially affected by the predictors' degree of endogeneity or the autocorrelation in the residuals of their autoregression. Importantly, these power plots confirm, in the context of joint tests, that an increase in the horizon length typically leads to power *loss*. This is true across the various combinations of parameter values and for different sample sizes.

4 Are Factor Returns Predictable?

This Section examines whether equity factor returns are predictable. We consider both "old" and "new" factors that have been proposed in the empirical asset pricing literature, beyond the standard market portfolio. These factors are typically found to carry significant full-sample premia. Though there is still an active debate whether these premia provide compensation for exposure to a macroeconomic/fundamental source of risk or they capture systematic mispricing, including them in multi-factor models to risk-adjust returns essentially assumes that they are risk factors. Therefore, examining whether these factor returns are predictable sheds further light on whether risk premia, and hence discount rates, are time-varying in a predictable way, extending the analysis to these alternative dimensions of risk.

4.1 Data

In addition to the excess returns of the market portfolio (MKT), we consider the returns of the size (SMB) and value (HML) factors that comprise the 3-factor model of FF1993. Furthermore, we examine the momentum (MOM) factor that is included in the 4-factor model of Carhart (1997). We also use the profitability (RMW) and investment (CMA) factors, which have been devised by FF2015 in their extended 5-factor model. For completeness, we additionally consider the size (ME), investment (IA), and profitability (ROE) factors of the recently proposed 4-factor model by HXZ.

Following the convention in the predictability literature, returns are logarithmized. The sample period for MKT, SMB, HML, and MOM is January 1927-December 2017, whereas the sample period for RMW and CMA is July 1963-December 2017. Returns for these factors are sourced from Kenneth French's online data library. The sample period for ME, IA, and ROE is January 1967-December 2016; these factor returns have been kindly provided by Kewei Hou.

We use a number of financial variables that have been extensively used in the prior literature as predictors of market returns as well as proxies for business cycle conditions. In particular, we report results for the following variables: dividend-price ratio (d/p), earnings-price ratio (e/p), book-to-market value ratio (b/m), default yield spread (dfy), T-bill rate (tbl), and term spread (tms). This is a subset of the predictors considered in Welch and Goyal (2008). The sample period for the predictors is January 1927-December 2017, with the exception of the term structure variables (tbl and tms), which, following Campbell and Yogo (2006), we use in predictability tests post-1952. Data are sourced from Amit Goyal's website.

Table 3 presents the descriptive statistics for the monthly dataset. Panel A reports the average, standard deviation and correlations of the factor returns. Panel B provides information regarding the properties of the employed predictors. In particular, it reports the degree of correlation ($\hat{\delta}$) between the residuals from the univariate predictive regression of each predictor on each factor's returns, as in equation (1), and the residuals from the predictor's AR(1) model, as in equation (2). For each predictor, it also reports the corresponding AR(1) coefficient estimate ($\widehat{R_n}$) as well as the autocorrelation coefficient estimate ($\hat{\phi}$) for the residuals of the predictor's autoregression. Table

IA. 16 presents the corresponding descriptive statistics for the annual dataset.

The descriptive statistics show that most of these factors yield significant premia. As expected, we find that there is a very strong correlation between SMB and ME as well as between CMA and IA. To the contrary, the correlation between RMW and ROE is not particularly high due to the different way that these factors have been constructed.

Regarding the properties of the employed predictors, we confirm that they are highly persistent. In fact, their estimated AR(1) coefficient is particularly close to unity. This feature gives rise to uncertainty regarding the predictors' exact type of persistence or even their order of integration. Given the crucial impact that the predictors' time series characteristics have on the properties of the predictability test statistics, this uncertainty further motivates the use of W_{IVX}, which is robust to different types of persistence.

Panel B of Table 3 shows that, apart from being highly persistent, most of these predictors exhibit a non-negligible degree of endogeneity with respect to the factor returns. This feature is most pronounced in the case of price-scaled ratios with respect to MKT, where a highly negative correlation between their residuals appears by construction, but it is also evident in various combinations involving SMB, HML, MOM, CMA, and IA. Hence, the Stambaugh bias could undermine the validity of inference in standard predictability tests for most of the factors we examine here, not just MKT. Moreover, it is evident from Panel B that the residuals from each predictor's AR(1) model typically exhibit a considerable degree of autocorrelation.

The predictability literature keeps discovering new predictors (see Rapach et al., 2016, for a recent successful attempt). A priori, it cannot be excluded that different variables could predict the returns of different factors (see, for example, the evidence in Cooper and Maio, 2019). However, this practice has raised serious concerns about the reliability of inference due to data mining (Ferson et al., 2003; Ferson et al., 2008) and the corresponding search for predictors (Harvey et al., 2016). This practice may inadvertently lead to p-hacking, if critical values are not adjusted for multiple hypothesis testing (see Harvey, 2017). To sidestep these criticisms, we rely on a parsimonious set of six predictors, which have been commonly used to span the state space of the economy.

4.2 Univariate Predictive Regressions

This Section presents the results from univariate predictive regressions of each of the six financial variables (d/p, e/p, b/m, dfy, tbl, tms) on each of the pricing factors considered (MKT, SMB, HML, MOM, RMW, CMA, ME, IA, and ROE).

4.2.1 Market Factor

We firstly examine whether monthly excess market returns are predictable. Panel A of Table 4 reports the least squares estimate ($\hat{A}_K$) of the slope coefficient from regression (8) for horizons $K = 1, 12, 36$, and 60 months, as well as the values of the statistics t_{SCALED}, t_{NW}, t_{HH}, t_{HOD}, and W_{IVX}, testing the null hypothesis of no predictability, i.e., that the slope coefficient of the predictive regressor in the DGP is equal to zero, $H_0 : A = 0$ in (1).

The reported results lead to the following observations. First, we confirm prior findings that relying upon t_{NW} or t_{HH}, d/p appears to be a very strong predictor of excess market returns as the horizon increases. Similar is the pattern using t_{OLS}, whereas this horizon effect is to a large extent neutralized using t_{SCALED}. In view of the simulation analysis presented in Section 3, we conclude that these statistics lead to spurious inference, since d/p is the prototype of an extremely highly persistent and almost perfectly endogenous predictor, causing t_{NW} and t_{HH} to be severely oversized at long horizons. t_{HOD} seems to be less affected by the horizon effect, though its value still increases monotonically as the horizon increases. In sharp contrast with the inference based on t_{NW} or t_{HH}, W_{IVX} suggests that d/p is not a significant predictor of excess market returns at the 5% level or lower, regardless of the horizon length.

Second, using each of the other two price-scaled ratios, e/p and b/m, we find again that t_{NW} and t_{HH} are typically higher in long-horizon regressions relative to the single-period regression. Most interestingly, using either t_{NW} or t_{HH}, b/m is found to be a very strong predictor of excess market returns at long horizons. Similar is the conclusion for b/m using t_{HOD}. Regarding e/p, t_{NW} and t_{HOD} yield strong evidence of predictability at $K = 12$ and $K = 36$ months. Since both e/p and b/m are both highly persistent and strongly endogenous, inference based on t_{NW}, t_{HH}, or t_{HOD} can be spurious. Hence, we resort to W_{IVX}, which is shown to be correctly sized. We find that excess market returns are indeed predictable at the 5% significance level via e/p and b/m. However, this evidence becomes weaker, not stronger as the horizon increases. This finding may be a result of the lower power that the correctly sized W_{IVX} exhibits as the horizon increases.

The third observation refers to dfy. t_{NW} and t_{HH} tend to increase as the horizon increases, even though dfy is substantially less endogenous than the price-scaled ratios with respect to excess market returns. Nevertheless, its very high degree of persistence can still lead to spurious inference in favor of predictability at sufficiently long horizons. To the contrary, W_{IVX} shows that dfy is not a significant predictor. Similar is the conclusion we reach using t_{HOD}, whose oversizing is much less pronounced in this case because dfy exhibits a low degree of endogeneity.

Relying on W_{IVX} to examine the predictive ability of tbl in the post-1952 period, we find significant evidence in favor of predictability only when $K = 1$ month. The rest of the test statistics yield similar inference, even though t_{NW} and t_{HH} would also support predictability at the 10% level or lower at $K = 12$ and $K = 36$ months. Last but not least, W_{IVX} indicates that tms is a significant predictor of excess market returns across the presented horizons. Very similar is the inference derived using t_{HOD}. Interestingly, t_{NW} and t_{HH} exhibit again a horizon effect, with the slope coefficient of tms being significant at the 1% level at $K = 36$ months. On the other hand, using t_{SCALED} would lead to the conclusion of no predictability at sufficiently long horizons.

Panel A of Table IA.17 presents the corresponding results for annual excess market returns. The most striking finding is that, using the correctly sized W_{IVX}, we would not reject the null of no predictability at the 5% significance level across the horizons considered for any of these six predictors. One would derive a similar conclusion using t_{HOD}, with a few exceptions of marginal significance. To the contrary, using t_{NW} or t_{HH}, one would find significant predictability, especially at $K = 3$ years, for all predictors apart from dfy. Most notably, the "strong predictive ability" of d/p at the 3- and 5-year horizon, which has been long regarded as a "stylized fact" (see Cochrane 1999; 2008; 2011; Campbell, 2000), seems to be an artefact of the severe oversizing that characterizes t_{NW} and t_{HH} at long horizons in the presence of highly persistent and endogenous regressors. Though we cannot exclude the possibility that other variables could prove more successful, the evidence based on these commonly used state variables casts doubt on the conventional wisdom that the market premium is predictable at the annual frequency.

4.2.2 FF1993 Factors

Next, we examine whether the FF1993 factors are predictable. Size and value premia are often claimed to reflect compensation for exposure to macroeconomic risk factors. Since the employed predictors have been commonly used as business cycle proxies, it is logical to hypothesize that they should be able to capture the time-variation in SMB and HML returns. Nevertheless, there is only limited empirical work on this issue (see Ferson and Harvey, 1999; Pontiff and Schall, 1999; Cohen et al., 2003; Stivers and Sun, 2010; Gulen et al., 2011). Panel B of Table 4 reports the results from predictability tests using monthly SMB returns, whereas Panel C presents the corresponding results using monthly HML returns.

Using W_{IVX}, we find strong evidence of predictability for SMB returns via b/m and dfy. However, this evidence becomes weaker as the predictive horizon increases. As mentioned above, this finding is most likely driven by the loss of power for W_{IVX} as the predictive horizon increases. To the contrary, we find no evidence of predictability for e/p, tbl, and tms. It is worth noting that most of the examined test statistics yield qualitatively similar inference. This is because these predictors' degree of endogeneity with respect to SMB returns is very low. Hence, the conventional test statistics would not be substantially oversized, as the magnitude of the Stambaugh bias and the impact of overlapping observations at long horizons would be rather limited.

Regarding HML returns, the evidence in favor of predictability is weaker. In particular, using the proposed IVX-Wald statistic, we find b/m to be a significant predictor at $K = 1$ and $K = 12$ months, whereas dfy significantly predicts HML returns only when $K = 1$ month. To the contrary, relying on t_{NW} or t_{HH}, one would conclude that HML returns are strongly predictable at long horizons, with tbl emerging as significant predictor. On the other hand, t_{HOD} appears to be very conservative, with no predictor being significant at the 5% level, regardless of the horizon length. Taken together, if SMB and HML premia compensate investors for being exposed to macroeconomic risks, it is only b/m and dfy that capture the time-variation in these premia, confirming their validity as business cycle proxies.

Panels B and C of Table IA.17 present the results from predictability tests for annual SMB and HML returns, respectively. We find that b/m and dfy are strongly significant predictors of annual SMB returns, but this evidence is weaker beyond the 3-year horizon. None of the rest predictors is found to be significant across the examined horizons. Regarding HML, the weak predictability evidence we reported using monthly returns now entirely disappears. Based on W_{IVX}, we find that none of these variables can predict annual HML returns. t_{HOD} yields the same conclusion. To the contrary, t_{NW} and t_{HH} would spuriously provide support for predictability via b/m and tbl at $K = 5$ years. Despite the fact that HML yields an economically and statistically significant premium at the annual frequency, we find no evidence that this premium is time-varying in a predictable manner via the set of state variables that we employ in this study.

4.2.3 Momentum Factor

The debate on whether MOM mimics an underlying risk factor or its premium reflects systematic mispricing remains unsettled. In fact, there is no universally accepted theoretical background to hypothesize that MOM returns should be predictable via business cycle proxies. Nevertheless, given the magnitude of the momentum premium, it is worth examining whether it is predictably time-varying; such a finding would have important implications for factor investing.

Panel A of Table IA.18 reports the results from predictability tests for monthly MOM returns. Using W_{IVX}, we find strong evidence that dfy is a significant predictor of MOM returns across the examined horizons. We find that d/p and b/m can also significantly predict MOM returns at short horizons, but this relationship is insignificant at $K = 36$ and $K = 60$ months. It is interesting to observe that, in some cases, the examined test statistics lead to qualitatively different inference regarding the predictability of MOM returns. In fact, using t_{HOD}, no predictor appears to be significant at the 5% level, regardless of the horizon length. Equally importantly, t_{NW} and t_{HH} would not reveal significant predictability for any variable at $K = 1$ month. This disagreement highlights the importance of using a test statistic with good finite-sample properties, such as W_{IVX}.[6]

Panel A of Table IA.19 reports the corresponding results for annual MOM returns. Using W_{IVX}, we find significant evidence in favor of predictability via dfy at $K = 1$ and $K = 2$ years. There is also evidence that e/p and tbl are significant predictors of annual MOM returns at short horizons. Interestingly, t_{NW}, t_{HH} and t_{HOD} point towards no predictability across the variables considered, regardless of the horizon length. In sum, we find that it is only dfy that can consistently predict both monthly and annual MOM returns, providing a link between the premium that this factor bears and the credit conditions at the aggregate level.

4.2.4 FF2015 Factors

FF2015 have recently proposed two more pricing factors (RMW and CMA) to explain the cross-section of expected stock returns, extending their original 3-factor model to a 5-factor model. If RMW and CMA reflect systematic sources of risk and they carry significant premia, then it is legitimate to ask whether these factor returns are predictable too. Panel B of Table IA.18 reports the results from predictability tests for monthly RMW returns, whereas Panel C reports the corresponding results for monthly CMA returns.

Regarding the profitability factor, using W_{IVX}, we find no evidence whatsoever that its monthly returns are predictable via any of the employed state variables. The rest of the test statistics lead to a very similar conclusion. This agreement across the various test statistics is presumably due to the fact that the degree of endogeneity of the employed regressors with respect to monthly RMW returns is quite low (see Panel B of Table 3), and hence the Stambaugh bias becomes negligible.

Similarly, we find no evidence that the monthly returns of the investment factor are predictable at the 5% significance level when W_{IVX} is considered. The only exception is when tms is used as a predictor and $K = 60$ months. To the contrary, t_{NW}, t_{HH}, and t_{HOD} provide some evidence in favor of predictability via tbl.

[6] In the commonly considered case of price-scaled ratios being regressed on excess market returns, an upward bias appears in the estimated slope coefficient because $\hat{\delta} < 0$. To the contrary, when $\hat{\delta} > 0$, as it is the case for most of the examined regressors with respect to monthly MOM returns (see Panel B of Table 3), the estimated slope coefficient would be biased downwards (see Stambaugh, 1986). This *downward* Stambaugh bias could result in t-tests failing to reject the null when the alternative is true.

Panels B and C of Table IA.19 present the corresponding results for annual RMW and CMA returns. Using W_{IVX}, we find no evidence that the returns of these factors are predictable, regardless of the horizon considered. t_{HOD} points to the same conclusion, whereas t_{NW} and t_{HH} indicate tbl as a significant predictor at long horizons. Overall, the results based on the correctly sized IVX-Wald statistic show that neither monthly nor annual RMW or CMA returns are significantly predictable via the examined regressors. To the extent that these variables span the state space of the economy, we can conclude that RMW and CMA premia do not reflect time-varying macroeconomic risk premia.

4.2.5 HXZ Factors

Concurrently with FF2015, HXZ proposed an alternative set of pricing factors based on Tobin's q-theory of investment. Table IA.20 reports the results from predictability tests for monthly ME (Panel A), IA (Panel B), and ROE (Panel C) returns. ME returns are highly correlated with SMB returns. Nevertheless, since ME is only available post-1967, these tests could be viewed as a subsample predictability analysis for the size factor. We find almost no evidence of predictability for monthly ME returns. Using W_{IVX}, we only find dfy to be a significant predictor at the 5% level when $K = 1$ month. To the contrary, b/m, which was found to significantly predict monthly SMB returns in the full sample period, appears now to be insignificant across the examined horizons.

Similarly, IA returns are strongly, but not perfectly, correlated with CMA returns. Hence, the evidence from predictability tests for IA returns is very similar to the one reported for CMA returns in Table IA.18. The only exception is that tbl is now found to be a significant predictor of monthly IA returns. Based on W_{IVX}, we also find significant evidence of predictability for monthly ROE returns via e/p, dfy, and tbl. However, this evidence becomes weaker as the predictive horizon increases. To the contrary, using t_{NW} or t_{HH}, one would erroneously conclude that ROE returns are strongly significantly predictable at long horizons.

Table IA.21 presents the results from the predictability tests using annual ME, IA, and ROE returns. On the basis of W_{IVX}, we find no evidence of predictability for ME returns, regardless of the horizon length. Similarly, we find no evidence of predictability for IA returns; the only exception is when tms is used as predictor and $K = 5$ years. To the contrary, annual ROE returns are found to be significantly predictable via e/p, dfy, and tbl. This finding is consistent with the results reported for monthly ROE returns. However, these predictive relationships become insignificant when $K = 3$ or $K = 5$ years.

4.3 Multivariate Predictive Regressions

In this Section, we consider multivariate predictive regressions using combinations of the predictors employed in the univariate analysis presented above. Multivariate predictive regressions are of particular interest for a number of reasons. In tests of the semi-strong form of market efficiency, one may be interested in testing the joint predictive ability of a set of information variables rather than the individual significance of each of them separately. In addition, multivariate predictive regressions can be used to select conditioning variables for the specification of conditional asset pricing models (see Petkova and Zhang, 2005; Cooper and Maio, 2019). More formally, the state space of the economy may be more appropriately spanned by a set of state variables rather than a single one. The VAR models that are commonly used in the intertemporal asset pricing and asset allocation literature typically include multiple state variables (see Campbell et al., 2003; Campbell and Vuolteenaho, 2004; Petkova, 2006; Maio and Santa-Clara, 2012), corresponding to multivariate predictive regressions for stock returns.

Unfortunately, most of the test statistics that have been proposed to deal with highly persistent regressors are developed within a univariate setup (see Valkanov, 2003; Torous et al., 2004; Campbell and Yogo, 2006; Rossi, 2007; Hjalmarsson, 2011) and cannot be applied to the multivariate framework, resulting in a methodological gap. Filling this gap, our test statistic can accommodate multiple predictors of arbitrary persistence, exhibiting very good finite-sample properties.

We report results for six combinations of the employed predictors. Motivated by the present-value model of Ang and Bekaert (2007), Combination I uses d/p and tbl, whereas Combination II includes e/p, instead of d/p, together with tbl. Combination III is motivated by the bivariate regression by Lamont (1998), using d/p and e/p as predictors. Combination IV corresponds to the trivariate regression of Ang and Bekaert (2007), including d/p, e/p, and tbl. Combination V, which consists of e/p, b/m, and tms resembles the state vector in the VAR model of Campbell and Vuolteenaho (2004), using b/m instead of the value spread. Combination VI includes the four "information variables" of Ferson and Schadt (1996) and Petkova (2006), namely d/p, tbl, tms, and dfy. Using these pre-specified combinations of the predictors, we sidestep the data mining concerns that would arise if we were instead "searching" among their numerous permutations.

4.3.1 Market Factor

Panel A of Table 5 reports the results from multivariate predictability tests using monthly excess market returns, leading to a series of interesting conclusions. First, using W_{IVX}, we find evidence of joint predictability for some of the combinations considered. However, the evidence based on W_{IVX} is much weaker relative to what W_{NW}, W_{HH}, and W_{HOD} indicate. This pattern reflects the severe oversizing that mainly characterizes W_{NW} and W_{HH}, but also W_{HOD} in the presence of highly persistent and endogenous regressors, such as the price-scaled ratios.

Second, using W_{IVX}, the evidence for predictability mostly disappears as the horizon increases. We find that none of the reported combinations yields joint significance when $K = 36$ or $K = 60$ months. To an extent, this feature may be driven by the loss of power of the test statistic as the horizon increases. To the contrary, W_{NW} and W_{HH} indicate that excess market returns are strongly predictable at long horizons. This feature highlights how spurious inference due to these tests' oversizing at long horizons has been misinterpreted by prior studies as strong evidence in favor of predictability.

The third conclusion refers to the combinations that are jointly significant using the correctly sized IVX-Wald statistic. In line with Ang and Bekaert (2007), we find that d/p and tbl are jointly significant predictors in the post-1952 period, but only when $K = 1$ month. Interestingly, e/p and tbl are also found to be jointly significant at short horizons. To the contrary, the Lamont regression (d/p & e/p) does not yield joint significance. The only other combination that seems to contain robust and significant predictive ability at short horizons is Combination VI (d/p, tbl, tms & dfy).

Panel A of Table IA.22 reports the corresponding results using annual excess market returns. The evidence in favor of predictability becomes even weaker relative to the monthly results. In particular, using W_{IVX}, we find that neither of these six combinations yields joint predictability beyond $K = 1$ year. Similar is the conclusion using W_{HOD}. To the contrary, W_{NW} and W_{HH} spuriously indicate that excess market returns are strongly predictable, especially at long horizons. Taken together, our results show that excess market returns are significantly predictable at short horizons via the combinations of variables considered by Ferson and Schadt (1996), Petkova (2006), and Ang and Bekaert (2007), but using a correctly sized test statistic, this evidence almost entirely disappears as the horizon increases.

4.3.2 FF1993 Factors

Panels B and C of Table 5 examine whether monthly SMB and HML returns, respectively, are jointly predictable by the combinations of variables that we consider in this Section. Interestingly, we find very little evidence in favor of predictability. In particular, on the basis of W_{IVX}, we find that Combination VI (d/p, tbl, tms & dfy) can jointly predict monthly SMB returns only when $K = 1$ month. None of the six combinations contains significant predictive ability for SMB returns beyond the horizon of $K = 12$ months. Similarly, there is no evidence of long-horizon predictability for monthly HML returns. It is only the combinations of d/p & tbl and e/p & tbl that significantly predict HML returns when $K = 1$ month.

Panels B and C of Table IA.22 report the corresponding results for annual SMB and HML returns, respectively. The results are striking. Regardless of the horizon length, none of the combinations considered yields significant predictability when either W_{IVX} or W_{HOD} is used as test statistic. To the contrary, W_{NW} and W_{HH} spuriously indicate that HML returns, in particular, are strongly predictable when $K = 5$ years. In conclusion, the very weak in-sample predictability that we find for SMB and HML returns questions the reliability of conditional Fama-French models with multiple conditioning variables.

4.3.3 Momentum Factor

We next examine MOM returns. Panel A of Table IA.23 reports the results for monthly MOM returns, whereas Panel A of Table IA.24 present the corresponding results for annual MOM returns. Based on W_{IVX}, we find strong evidence of predictability at short horizons. In the case of monthly returns, all six combinations yield joint significance at the 5% level when $K = 1$ month. However, this evidence becomes substantially weaker or even disappears as the predictive horizon increases. Similarly, all combinations, apart from Combination VI, yield joint significance for annual MOM returns when $K = 1$ year. This evidence disappears as the horizon increases, particularly when $K > 2$ years, highlighting again the loss of power at long horizons.

These results reveal that the momentum premium is predictably time-varying via the state variables that are commonly used in the empirical asset pricing literature. In fact, this evidence is much more robust relative to the evidence for the in-sample predictability of the market premium, which has attracted most of the attention in prior studies. Apart from drawing a clear link between business cycle conditions and the time-variation in momentum returns, our findings show that, within a factor investing setup, timing momentum returns may prove to be more reliable than timing market, size, or value returns.

4.3.4 FF2015 Factors

We also conduct joint predictability tests with respect to RMW and CMA returns. The results for monthly RMW returns are reported in Panel B of Table IA.23; the corresponding results for annual RMW returns are presented in Panel B of Table IA.24. Using W_{IVX}, we find no evidence of predictability across the six combinations of predictors considered, regardless of the horizon length. This conclusion holds true for both monthly and annual RMW returns. W_{HOD} yields the same inference, whereas W_{NW} and W_{HH} spuriously indicate that RMW returns are predictable at long horizons. We conclude that the profitability premium, as captured by the RMW factor of FF2015, is not predictably time-varying through the commonly used state variables.

Panel C of Table IA.23 contains the results from joint predictability tests with respect to monthly CMA returns. Relying on W_{IVX}, we find significant evidence in favor of predictability for most of the combinations considered, which remains robust up to $K = 36$ months. Similar is the evidence based on W_{HOD}. On the other hand, the corresponding results for annual CMA returns, which are reported in Panel C of Table IA.24, show no evidence of predictability across the six combinations considered. These results highlight that the inference on predictability also depends on the frequency of the factor returns used, undermining the robustness of conclusions based solely on monthly or annual returns.

4.3.5 HXZ Factors

Last, we run joint predictability tests for the HXZ factors. Results are presented in Table IA.25 for monthly returns and in Table IA.26 for annual returns. Regarding the ME factor, W_{IVX} provides no evidence of joint predictability for any of the six combinations examined, regardless of the length of the predictive horizon. This is true for both monthly and annual ME returns. As with SMB, this evidence questions the common assumption that the size premium is related to the business cycle conditions, which are commonly proxied by the employed state variables.

On the other hand, using W_{IVX}, we find significant evidence that monthly IA returns are predictable via combinations that involve d/p, e/p, and tbl, particularly at short horizons. However, these predictability relationships become weaker or even insignificant when we examine annual IA returns. In fact, none of the examined combinations yields joint predictability when $K = 1$ year.

Among the factors proposed by HXZ, the strongest evidence in favor of predictability is reported for ROE. In particular, W_{IVX} indicates significant joint predictability at short horizons for all six combinations, with respect to both monthly and annual ROE returns. However, this evidence becomes weaker as the predictive horizon increases. This finding is in stark contrast with the corresponding results for the profitability factor (RMW) of FF2015, for which no joint predictability is reported. This is due to the different way that these two profitability factors have been constructed and the relatively low correlation that their returns exhibit.

Taken together, our results support the argument that the investment and profitability premia of HXZ are related to the state of the economy and that they are predictably time-varying, at least at the monthly frequency. Hence, it is worth considering conditional versions of this 4-factor model, as they may exhibit superior pricing ability relative to its unconditional version.[7]

[7] We have also conducted multivariate predictability tests using a general-to-specific approach. Specifically, for each factor return and a given horizon, we initially estimate a predictive regression including all six predictors. Then, we drop the predictor with the lowest individual W_{IVX} value, as long as this does not exceed the 10% chi-squared

5 Conclusion

This study provides a critical assessment of long-horizon return predictability tests via persistent regressors. We conduct an extensive simulation analysis to show that, in the presence of a sufficiently persistent regressor, the test statistics using Newey-West or Hansen-Hodrick standard errors become severely oversized as the horizon increases, leading to spurious inference. Moreover, the test statistic with Hodrick standard errors also tends to overreject the null of no predictability in the presence of highly persistent and endogenous regressors, such as the price-scaled ratios that have been predominantly used in the literature during the last three decades. Hence, we side with prior critical views, which cast doubt on the conventional wisdom that market returns are strongly significantly predictable at long horizons via price-scaled ratios or term structure variables.

Whereas a number of alternative testing methodologies have been proposed to conduct valid inference, they all have certain limitations, either because they make strong assumptions about the exact time series properties of the predictors or because they cannot accommodate multiple predictors. As a remedy, we propose a simple IVX-Wald statistic, which accommodates multiple predictors, exhibits excellent finite-sample properties regardless of the predictive horizon's length, and is robust to a wide range of regressor persistence types.

Employing the proposed test statistic and a small set of variables that have been commonly used as proxies for business cycle conditions, we find evidence of predictability for "old" and "new" pricing factors with monthly returns. However, this evidence becomes weaker, not stronger, as the predictive horizon increases and disappears for most of the factors with annual returns. This is particularly true for market returns, the predictability of which has been long debated in the literature. Interestingly, however, we find robust and significant predictability for the returns of the momentum factor as well as the profitability factor of HXZ. This evidence provides a link between the macroeconomy and the premia that these factors yield.

In conclusion, our study questions the incremental value of using long-horizon predictive regressions, for the additional reason that correctly sized test statistics are *less*, not more powerful as the horizon increases. Hence, we argue that it is preferable to conduct inference relying on the actual data generating process, rather than resorting to long-horizon predictive regressions. To a large extent, this conclusion is consistent with the revisionary statement of Cochrane (2017, p. 490): "After a long controversy, I think it is fair to say that long-horizon regressions are most important for showing the *economic* rather than *statistical* significance of forecasting regressions. The number of nonoverlapping observations declines as the horizon lengthens, so larger standard errors make up for larger coefficients, and there is not really a huge statistical advantage either way".

References

[1] Ang, A., Bekaert, G., 2007. Stock return predictability: Is it there?. Review of Financial Studies 20, 651-707.

[2] Boudoukh, J., Richardson, M.P., 1994. The statistics of long-horizon regressions revisited. Mathematical Finance 4, 103-119.

[3] Boudoukh, J., Richardson, M.P., Whitelaw, R.F., 2008. The myth of long-horizon predictability. Review of Financial Studies 21, 1577-1605.

[4] Campbell, J.Y., 2000. Asset pricing at the millenium. Journal of Finance 55, 1515-1567.

[5] Campbell, J.Y., 2001. Why long horizons? A study of power against persistent alternatives. Journal of Empirical Finance 8, 459-491.

critical value. We repeat this procedure until all remaining predictors are inidividually significant at the 10% level or only one predictor remains. Table IA.27 presents the results from this predictor selection procedure for monthly factor returns, whereas Table IA.28 presents the corresponding results for annual factor returns.

[6] Campbell, J.Y, Shiller, R.J., 1988. Stock prices, earnings, and expected dividends. Journal of Finance 43, 661-676.

[7] Campbell, J.Y., Vuolteenaho, T., 2004. Bad beta, good beta. American Economic Review 94, 1249-1275.

[8] Campbell, J.Y., Yogo, M., 2006. Efficient tests of stock return predictability. Journal of Financial Economics 81, 27-60.

[9] Campbell, J.Y., Chan, Y.L., Viceira, L.M., 2003. A multivariate model of strategic asset allocation. Journal of Financial Economics 67, 41-80.

[10] Carhart, M.M., 1997. On Persistence in Mutual Fund Performance. Journal of Finance 52, 57-82.

[11] Cavanagh, C., Elliott, G., Stock, J.H., 1995. Inference in models with nearly integrated regressors. Econometric Theory 11, 1131-1147.

[12] Cochrane, J.H., 1999. New facts in finance. Economic Perspectives 23, 36-58.

[13] Cochrane, J.H., 2008. Financial markets and the real economy. In: Mehra, R. (ed.), Handbook of the Equity Premium, Elsevier.

[14] Cochrane, J.H., 2011. Presidential address: Discount rates. Journal of Finance 66, 1047-1108.

[15] Cochrane, J.H., 2017. Return forecasts and time-varying risk premiums. In: Cochrane, J.H., Moskowitz, T.J. (eds.), The Fama Portfolio, University of Chicago Press.

[16] Cohen, R.B., Polk, C., Vuolteenaho, T., 2003. The value spread. Journal of Finance 58, 609-641.

[17] Cooper, I., Maio, P., 2019. New evidence on conditional factor models. Journal of Financial and Quantitative Analysis 54, 1975-2016.

[18] Demetrescu, M., Georgiev, I., Rodrigues, P.M.M. and Taylor, A.M.R., 2020. Testing for episodic predictability in stock returns. Journal of Econometrics, forthcoming.

[19] Demetrescu, M., Rodrigues, P.M.M. and Taylor, A.M.R., 2022. Transformed regression-based long-horizon predictability tests. Journal of Econometrics, forthcoming.

[20] Elliott, G., 1998. On the robustness of cointegration methods when regressors almost have unit roots. Econometrica 66, 149-158.

[21] Fama, E.F., French, K.R., 1988. Dividend yields and expected stock returns. Journal of Financial Economics 22, 3-24.

[22] Fama, E.F., French, K.R., 1989. Business conditions and expected returns on stocks and bonds. Journal of Financial Economics 25, 23-49.

[23] Fama, E.F., French, K.R., 1993. Common risk factors in the returns of stocks and bonds. Journal of Financial Economics 33, 3-56.

[24] Fama, E.F., French, K.R., 2015. A five-factor asset pricing model. Journal of Financial Economics 116, 1-22.

[25] Ferson, W.E., Harvey, C.R., 1999. Conditioning variables and the cross section of stock returns. Journal of Finance 54, 1325-1360.

[26] Ferson, W.E., Schadt, R.W., 1996. Measuring fund strategy and performance in changing economic conditions. Journal of Finance 51, 425-461.

[27] Ferson, W.E., Sarkissian, S., Simin, T.T., 2003. Spurious regressions in financial economics?. Journal of Finance 58, 1393-1413.

[28] Ferson, W.E., Sarkissian, S., Simin, T.T., 2008. Asset pricing models with conditional betas and alphas: The effects of data snooping and spurious regression. Journal of Financial and Quantitative Analysis 43, 331-354.

[29] Francq, C., Zakoian, JM., 2010. GARCH Models: stucture, statistical inference and financial applications. Wiley.

[30] Goetzmann, W.N., Jorion, P., 1993. Testing the predictive power of dividend yields. Journal of Finance 48, 663-679.

[31] Gulen, H., Xing, Y., Zhang, L., 2011. Value versus growth: Time-varying expected stock returns. Financial Management 40, 381-407.

[32] Hansen, L., Hodrick, R.J., 1980. Forward exchange rates as optimal predictors of future spot rates: An econometric analysis. Journal of Political Economy 88, 829–853.

[33] Harvey, C.R., 2017. Presidential address: The scientific outlook in financial economics. Journal of Finance 72, 1399-1440.

[34] Harvey, C.R., Yan, L., Zhu, H., 2016. ... and the cross-section of expected returns. Review of Financial Studies 29, 5-68.

[35] Hjalmarsson, E., 2011. New methods for inference in long-horizon regressions. Journal of Financial and Quantitative Analysis 46, 815-839.

[36] Hodrick, R.J., 1992. Dividend yields and expected stock returns: Alternative procedures for inference and measurement. Review of Financial Studies 5, 357–386.

[37] Hou, K., Xue, C., Zhang, L., 2015. Digesting anomalies: An investment approach. Review of Financial Studies 28, 650-705.

[38] Jansson, M., Moreira, M.J., 2006. Optimal inference in regression models with nearly integrated regressors. Econometrica 74, 681-714.

[39] Keim, D.B., Stambaugh, R.F., 1986. Predicting returns in the stock and the bond markets. Journal of Financial Economics 17, 357-390.

[40] Kostakis, A., Magdalinos, A., Stamatogiannis, M.P., 2015. Robust econometric inference for stock return predictability. Review of Financial Studies 28, 1506-1553.

[41] Kothari, S., Shanken, J., 1997. Book-to-market, dividend yield, and expected market returns: a time-series analysis. Journal of Financial Economics 44, 169–203.

[42] Lamont, O., 1998. Earnings and expected returns, 1998. Journal of Finance 53, 1563-1587.

[43] Maio, P., Santa-Clara, P., 2012. Multifactor models and their consistency with the ICAPM. Journal of Financial Economics 106, 586-613.

[44] Magdalinos, T., 2020. Least squares and IVX limit theory in systems of predictive regressions with GARCH innovations. Working paper.

[45] Magdalinos, T and Phillips, P.C.B. (2020). Econometric inference in matrix vicinities of unity and stationarity. Working paper.

[46] Nelson, C.R., Kim, M.J., 1993. Predictable stock returns: The role of small sample bias. Journal of Finance 48, 641-661.

[47] Newey, W., West, K., 1987. A simple, positive definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–708.

[48] Petkova, R., 2006. Do the fama-french factors proxy for innovations in predictive variables?. Journal of Finance 61, 581-612.

[49] Petkova, R., Zhang, L., 2005. Is value riskier than growth?. Journal of Financial Economics 78, 187-202.

[50] Phillips, P.C.B., 1987. Time series regression with a unit root. Econometrica 55, 277-302.

[51] Phillips, P.C.B., Magdalinos, T., 2009. Econometric inference in the vicinity of unity. CoFie Working Paper No 7, Singapore Management University.

[52] Pontiff, J., Schall, L.D., 1998. Book-to-market ratios as predictors of market returns. Journal of Financial Economics 49, 141-160.

[53] Rapach, D., Wohar, M.E., 2005. Valuation ratios and long-horizon stock price predictability. Journal of Applied Econometrics 20, 327-344.

[54] Rapach, D., Zhou, G., 2013. Forecasting stock returns. In: Elliott, G., Timmermann, A. (eds.), Handbook of Economic Forecasting, Vol. 2A, Elsevier.

[55] Rapach, D., Riggenberg, M.C., Zhou, G., 2016. Short interest and aggregate stock returns. Journal of Financial Economics 121, 46-65.

[56] Rossi, B., 2007. Expectations hypotheses tests at long horizons. The Econometrics Journal 10, 1-26.

[57] Stambaugh, R.F., 1986. Bias in regressions with lagged stochastic regressors. Working Paper, University of Chicago.

[58] Stambaugh, R.F., 1999. Predictive regressions. Journal of Financial Economics 54, 375-421.

[59] Stivers, C., Sun, L., 2010. Cross-sectional return dispersion and time variation in value and momentum premiums. Journal of Financial and Quantitative Analysis 45, 987-1014.

[60] Torous, W., Valkanov, R., Yan, S., 2004. On predicting stock returns with nearly integrated explanatory variables. Journal of Business 77, 937-966.

[61] Valkanov, R., 2003. Long-horizon regressions: Theoretical results and applications. Journal of Financial Economics 68, 201-232.

[62] Wei, M., Wright, J.H., 2013. Reverse regressions and long-horizon forecasting. Journal of Applied Econometrics 28, 353-371.

[63] Welch, I., Goyal, A., 2008. A comprehensive look at the empirical performance of equity premium prediction. Review of Financial Studies 21, 1455-1508.

Table 1

Finite−sample (n=1,000) sizes with no autocorrelation in the residuals of the autoregression

This table presents finite−sample sizes testing the null hypothesis $H_0: A = 0$ versus the alternative $H_1: A \neq 0$ in (1) when there is no autocorrelation in the residuals of the autoregressive equation (2). The reported rejection rates for each test correspond to a 5% nominal size and they are based on the Monte Carlo simulation described in Section 3.1 with 10,000 repetitions and sample size $n = 1,000$. t_{OLS} denotes the t-statistic from an ordinary least squares (OLS) K −horizon predictive regression. t_{SCALED} is the OLS t-statistic scaled by the square root of the predictive horizon K. t_{NW} denotes the t-statistic computed with Newey-West (1987) standard errors, t_{HH} is the t-statistic computed with Hansen-Hodrick (1980) standard errors, and t_{HOD} is the corresponding t-statistic computed with Hodrick (1992) standard errors. t_{BONF} is the Bonferroni test statistic of Hjalmarsson (2011). W_{IVX} denotes the IVX-Wald test statistic defined in (23). Results are reported for different degrees of correlation between the residuals of regressions (1) and (2), $\delta = -0.99, -0.5,\ 0$, different local-to-unity parameters, $C = 0, -5, -10, -50$, corresponding to the autoregressive root R_n, and different predictive horizons $K = 1, 10, 50, 100$.

$\delta = -0.99$	C	R_n	K	t_{OLS}	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	t_{BONF}	W_{IVX}
	0	1	1	0.306	0.306	0.310	0.312	0.303	0.046	0.056
			10	0.874	0.315	0.484	0.352	0.307	0.036	0.049
			50	0.948	0.341	0.595	0.499	0.317	0.028	0.045
			100	0.963	0.365	0.676	0.638	0.335	0.022	0.038
	-5	0.995	1	0.119	0.119	0.121	0.122	0.119	0.031	0.057
			10	0.668	0.116	0.227	0.149	0.117	0.031	0.057
			50	0.856	0.118	0.334	0.278	0.132	0.023	0.048
			100	0.900	0.113	0.434	0.441	0.152	0.014	0.041
	-10	0.990	1	0.087	0.087	0.088	0.088	0.085	0.029	0.054
			10	0.599	0.084	0.176	0.113	0.087	0.028	0.054
			50	0.811	0.069	0.270	0.236	0.095	0.021	0.046
			100	0.867	0.058	0.357	0.389	0.110	0.010	0.042
	-50	0.950	1	0.056	0.056	0.057	0.059	0.054	0.024	0.048
			10	0.507	0.039	0.125	0.083	0.057	0.020	0.048
			50	0.710	0.007	0.187	0.215	0.064	0.007	0.041
			100	0.753	0.002	0.260	0.361	0.079	0.001	0.036
$\delta = -0.5$	C	R_n	K	t_{OLS}	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	t_{BONF}	W_{IVX}
	0	1	1	0.109	0.109	0.112	0.114	0.107	0.034	0.053
			10	0.638	0.112	0.211	0.140	0.113	0.035	0.050
			50	0.836	0.121	0.290	0.241	0.125	0.039	0.044
			100	0.880	0.123	0.361	0.353	0.145	0.048	0.038
	-5	0.995	1	0.068	0.068	0.070	0.071	0.068	0.024	0.053
			10	0.562	0.068	0.145	0.092	0.069	0.023	0.051
			50	0.792	0.067	0.218	0.179	0.078	0.024	0.050
			100	0.843	0.061	0.290	0.297	0.092	0.026	0.044
	-10	0.990	1	0.057	0.057	0.060	0.061	0.057	0.022	0.050
			10	0.541	0.055	0.128	0.081	0.057	0.021	0.048
			50	0.777	0.046	0.201	0.167	0.065	0.019	0.045
			100	0.827	0.039	0.269	0.277	0.078	0.018	0.044
	-50	0.950	1	0.051	0.051	0.052	0.053	0.049	0.019	0.049
			10	0.501	0.033	0.109	0.066	0.050	0.013	0.048
			50	0.695	0.007	0.154	0.148	0.053	0.004	0.043
			100	0.735	0.001	0.179	0.241	0.063	0.001	0.041
$\delta = 0$	C	R_n	K	t_{OLS}	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	t_{BONF}	W_{IVX}
	0	1	1	0.049	0.049	0.051	0.052	0.049	0.049	0.051
			10	0.539	0.048	0.115	0.066	0.051	0.046	0.048
			50	0.791	0.052	0.182	0.144	0.064	0.048	0.047
			100	0.848	0.053	0.253	0.252	0.079	0.058	0.043
	-5	0.995	1	0.053	0.053	0.053	0.054	0.052	0.051	0.050
			10	0.532	0.049	0.118	0.070	0.051	0.045	0.049
			50	0.774	0.047	0.180	0.146	0.058	0.040	0.048
			100	0.827	0.040	0.248	0.253	0.067	0.041	0.042
	-10	0.990	1	0.050	0.050	0.050	0.051	0.049	0.047	0.049
			10	0.531	0.047	0.118	0.070	0.050	0.041	0.049
			50	0.764	0.040	0.177	0.146	0.055	0.035	0.045
			100	0.823	0.028	0.239	0.243	0.064	0.030	0.044
	-50	0.950	1	0.047	0.047	0.048	0.051	0.047	0.044	0.049
			10	0.498	0.031	0.107	0.064	0.045	0.028	0.045
			50	0.692	0.006	0.138	0.128	0.052	0.006	0.046
			100	0.726	0.001	0.158	0.205	0.057	0.001	0.044

Table 2

Finite−sample (n=1,000) sizes for predictive systems with two persistent regressors

This table presents finite−sample sizes for tests based on the predictive system in equation (25) with two persistent regressors. The reported rejection rates for each test correspond to a 5% nominal size and they are computed using the Monte Carlo simulation described in Section 3.3 with 10,000 repetitions and sample size $n = 1{,}000$. Panel A reports the rejection rates for joint tests of the null hypothesis H_0: $A \equiv (A_1\ A_2) = 0_{1x2}$, i.e., that the slope coefficients of both regressors are equal to zero. W_{SCALED} refers to the Wald statistic computed from an ordinary least squares (OLS) regression scaled by the predictive horizon K. W_{NW} denotes the Wald statistic computed with Newey-West (1987) standard errors. W_{HH} represents the Wald statistic computed with Hansen-Hodrick (1980) standard errors. W_{HOD} refers to the corresponding Wald statistic computed with Hodrick (1992) standard errors. W_{IVX} denotes the IVX-Wald test statistic defined in (23). Panel B reports the rejection rates for tests under the null hypothesis H_0: $A_1 = 0$, i.e., that the coefficient of the first persistent regressor is equal to zero. Panel C reports the corresponding rejection rates for tests under the null hypothesis H_0: $A_2 = 0$, i.e., that the coefficient of the second persistent regressor is equal to zero. t_{SCALED} is the OLS t-statistic scaled by the square root of the predictive horizon K. t_{NW} denotes the t-statistic computed with Newey-West standard errors, t_{HH} is the t-statistic computed with Hansen-Hodrick standard errors, and t_{HOD} is the corresponding t-statistic computed with Hodrick standard errors. Results are reported for three different combinations of the relevant parameter values (C, Φ and Σ). diag(C) provides the local-to-unity parameters of the regressors employed in each case. For all cases considered data of monthly log excess market return (MKT) is employed for the regressand, whereas for each case monthly data for a combination of two regressors (Predictors) is used. For each case, the estimated autocorrelation coefficients ($\phi's$) in the residuals of the autoregressive equations are reported (diag(Φ)) as well as the degrees of correlation ($\delta's$) between the ε_t and u_t with matrix Σ given in (25). For each case, the combination of predictors employed for the estimation of the simulation parameters along with their values are:

	Predictors	Data period	diag(C)	diag(Φ)	Σ
Case I	dividend-price ratio, T-bill rate	1952-2017	(0, -5)	(0.0640, 0.3379)	$\begin{pmatrix} 1 & -0.9827 & -0.1251 \\ -0.9827 & 1 & 0.3379 \\ -0.1251 & 0.3379 & 1 \end{pmatrix}$
Case II	earnings-price ratio, default yield spread	1927-2017	(0, -5)	(0.2741, 0.2167)	$\begin{pmatrix} 1 & -0.7596 & -0.2787 \\ -0.7596 & 1 & 0.1246 \\ -0.2787 & 0.1246 & 1 \end{pmatrix}$
Case III	earnings-price ratio, T-bill rate	1952-2017	(0, -10)	(0.3538, 0.3379)	$\begin{pmatrix} 1 & -0.6156 & -0.1282 \\ -0.6156 & 1 & 0.1583 \\ -0.1282 & 0.1583 & 1 \end{pmatrix}$

		Panel A: H_0: $A_1 = A_2 = 0$					Panel B: H_0: $A_1 = 0$					Panel C: H_0: $A_2 = 0$				
Case I	K	W_{SCALED}	W_{NW}	W_{HH}	W_{HOD}	W_{IVX}	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}
	1	0.123	0.130	0.133	0.124	0.060	0.165	0.166	0.169	0.163	0.072	0.075	0.077	0.079	0.074	0.057
	10	0.130	0.289	0.187	0.130	0.061	0.167	0.284	0.201	0.165	0.069	0.074	0.153	0.101	0.074	0.054
	50	0.142	0.433	0.377	0.143	0.052	0.172	0.373	0.319	0.176	0.058	0.067	0.220	0.184	0.072	0.051
	100	0.146	0.560	0.555	0.163	0.043	0.171	0.454	0.452	0.189	0.044	0.056	0.297	0.305	0.073	0.047
Case II	K	W_{SCALED}	W_{NW}	W_{HH}	W_{HOD}	W_{IVX}	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}
	1	0.273	0.279	0.285	0.266	0.079	0.373	0.378	0.381	0.371	0.105	0.116	0.117	0.118	0.115	0.064
	10	0.281	0.504	0.355	0.273	0.076	0.378	0.540	0.417	0.369	0.099	0.113	0.206	0.139	0.110	0.062
	50	0.314	0.658	0.575	0.280	0.064	0.395	0.635	0.561	0.378	0.078	0.102	0.276	0.230	0.092	0.059
	100	0.347	0.764	0.730	0.303	0.053	0.411	0.705	0.680	0.380	0.056	0.087	0.337	0.328	0.081	0.062
Case III	K	W_{SCALED}	W_{NW}	W_{HH}	W_{HOD}	W_{IVX}	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}
	1	0.170	0.177	0.180	0.167	0.066	0.217	0.222	0.226	0.217	0.073	0.082	0.085	0.086	0.081	0.062
	10	0.174	0.350	0.234	0.167	0.065	0.219	0.354	0.253	0.216	0.069	0.078	0.165	0.107	0.080	0.061
	50	0.181	0.498	0.436	0.181	0.055	0.230	0.451	0.385	0.228	0.057	0.064	0.228	0.192	0.073	0.054
	100	0.187	0.620	0.603	0.201	0.047	0.236	0.531	0.516	0.244	0.046	0.049	0.293	0.304	0.069	0.053

Table 3
Descriptive statistics, monthly data

This table presents descriptive statistics for monthly equity factor returns (Panel A) and predictive regressors (Panel B). Panel A contains the correlation matrix, mean, standard deviation and autoregressive coefficient of monthly equity factor log returns during the period 1927-2017. MKT, SMB, and HML denote the excess market, size, and value factors, respectively, from the Fama-French 3-factor model. MOM stands for the momentum factor. RMW and CMA denote the profitability and investment factors, respectively, from the Fama-French 5-factor model. ME, IA, and ROE stand for the size, investment, and profitability factors, respectively, from the Hou-Xue-Zhang (HXZ) 4-factor model. RMW and CMA are available post 1964. ME, IA, and ROE are available for the period 1967-2016. Panel B presents for each predictive regressor: (i) the autoregressive coefficient $\widehat{R_n}$, which is estimated from the autoregressive equation (2), (ii) the autocorrelation coefficient $\hat{\phi}$ for the residuals of the autoregressive equation (2), and (iii) the correlation coefficient $\hat{\delta}$ between the residuals of the univariate predictive regression model (1) for each equity factor and the corresponding autoregressive equation (2). These statistics are reported for the following predictive regressors: log dividend-price ratio (d/p), log earnings-price ratio (e/p), book-to-market value ratio (b/m), default yield spread (dfy), T-bill rate (tbl), and term spread (tms). The sample period for tbl and tms is 1952-2017.

Panel A: Equity Factor Returns

	Correlations of Factor Returns											
	MKT	SMB	HML	MOM	RMW	CMA	ME	IA	ROE	Mean (%)	St. Dev. (%)	AR coeff.
MKT	1.00									0.51	5.35	0.113
SMB	0.31	1.00								0.17	3.13	0.055
HML	0.21	0.10	1.00							0.32	3.39	0.202
MOM	-0.33	-0.15	-0.41	1.00						0.53	5.22	0.145
RMW	-0.23	-0.40	0.07	0.11	1.00					0.22	2.24	0.144
CMA	-0.38	-0.16	0.69	-0.01	-0.05	1.00				0.27	2.00	0.136
ME	0.27	0.95	-0.04	-0.03	-0.37	-0.05	1.00			0.26	3.06	0.041
IA	-0.38	-0.26	0.67	0.03	0.10	0.91	-0.15	1.00		0.39	1.87	0.132
ROE	-0.20	-0.37	-0.14	0.52	0.66	-0.09	-0.31	0.04	1.00	0.51	2.56	0.134

Panel B: Predictive Regressors

	Correlations of Residuals ($\hat{\delta}'s$)										
	MKT	SMB	HML	MOM	RMW	CMA	ME	IA	ROE	$\widehat{R_n}$	$\hat{\phi}$
d/p	-0.97	-0.23	-0.22	0.36	0.16	0.35	-0.15	0.33	0.17	1.000	0.10
e/p	-0.75	-0.21	-0.15	0.29	0.08	0.25	-0.10	0.25	0.07	1.000	0.27
b/m	-0.81	-0.18	-0.32	0.39	0.03	0.20	-0.11	0.19	0.04	0.997	0.18
dfy	-0.27	-0.23	-0.27	0.26	0.04	0.03	-0.07	0.03	0.08	0.993	0.22
tbl	-0.13	-0.03	-0.07	0.08	0.05	-0.05	-0.01	0.00	0.07	0.997	0.34
tms	0.06	0.11	0.10	-0.13	-0.09	0.08	0.10	0.03	-0.17	0.983	0.11

Table 4

Univariate predictive regressions for monthly MKT, SMB, and HML factor returns

This table presents the results of univariate K −horizon predictive regressions for monthly equity factor returns, as in equation (8), testing the null hypothesis $H_0: A = 0$ versus the alternative $H_1: A \neq 0$ in (1). The sample period is 1927-2017. MKT, SMB, and HML denote the excess market, size, and value factors, respectively, from the Fama-French 3-factor model. Results are presented for each of the following predictive regressors: log dividend-price ratio (d/p), log earnings-price ratio (e/p), book-to-market value ratio (b/m), default yield spread (dfy), T-bill rate (tbl), and term spread (tms). The sample period for tbl and tms is 1952-2017. Results are reported for the predictive horizons $K = 1, 12, 36, 60$. $\hat{A}_K$ denotes the ordinary least squares (OLS) slope coefficient estimate from the corresponding K −horizon regression model. t_{SCALED} is the OLS t-statistic scaled by the square root of the predictive horizon K. t_{NW} denotes the t-statistic computed with Newey-West (1987) standard errors, t_{HH} is the t-statistic computed with Hansen-Hodrick (1980) standard errors, and t_{HOD} is the corresponding t-statistic computed with Hodrick (1992) standard errors. W_{IVX} denotes the IVX-Wald test statistic defined in (23). * and ** indicate significance at the 5% and 1% level, respectively.

Predictor		Panel A: MKT						Panel B: SMB						Panel C: HML					
d/p	K	$\hat{A}_K$	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}	$\hat{A}_K$	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}	$\hat{A}_K$	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}
	1	0.00	1.40	0.96	0.89	1.04	1.84	0.00	1.30	1.07	1.12	1.03	2.20	0.00	1.19	0.67	0.61	0.76	1.29
	12	0.08	1.61	1.75	1.47	1.42	3.06	0.04	1.69	1.87	1.53	1.33	4.59*	0.03	1.11	0.96	0.80	0.82	1.46
	36	0.21	1.61	2.57*	2.48*	1.69	2.93	0.08	0.87	0.97	0.78	0.91	1.73	0.06	0.87	0.99	0.89	0.66	0.46
	60	0.34	1.68	3.40**	3.42**	1.85	2.57	0.07	0.45	0.56	0.51	0.55	0.39	0.11	0.94	1.09	0.98	0.79	0.52
e/p	K	$\hat{A}_K$	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}	$\hat{A}_K$	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}	$\hat{A}_K$	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}
	1	0.01	1.94	1.72	1.61	1.90	4.22*	0.00	-0.10	-0.10	-0.10	-0.09	0.01	0.00	1.36	1.05	0.97	1.16	1.72
	12	0.09	1.75	2.09*	1.87	2.24*	4.57*	0.01	0.46	0.62	0.51	0.49	0.30	0.03	0.86	1.00	0.88	0.99	0.85
	36	0.20	1.35	1.96*	1.87	2.03*	3.21	0.02	0.22	0.32	0.25	0.33	0.11	0.05	0.67	0.93	0.83	0.77	0.33
	60	0.25	1.11	1.62	1.58	1.71	2.12	0.01	0.05	0.06	0.05	0.06	0.00	0.07	0.56	0.95	0.87	0.71	0.21
b/m	K	$\hat{A}_K$	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}	$\hat{A}_K$	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}	$\hat{A}_K$	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}
	1	0.01	2.16*	1.12	1.00	1.25	3.90*	0.01	2.60**	1.95	2.00*	1.90	7.00**	0.01	3.15**	1.25	1.10	1.43	8.79**
	12	0.19	2.43*	2.93**	2.53*	1.84	6.12*	0.15	3.43**	3.60**	2.96**	2.20*	13.07**	0.11	2.15*	1.99*	1.68	1.41	5.05*
	36	0.46	2.03*	2.51*	2.13*	2.25*	3.94*	0.28	1.79	2.24*	1.85	1.96*	6.05*	0.18	1.48	1.80	1.64	1.24	1.76
	60	0.60	1.75	2.60*	2.21*	2.18*	2.71	0.25	0.95	1.41	1.32	1.29	2.45	0.25	1.31	2.00*	1.85	1.28	1.37
dfy	K	$\hat{A}_K$	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}	$\hat{A}_K$	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}	$\hat{A}_K$	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}
	1	0.15	0.65	0.23	0.21	0.27	0.20	0.60	4.43**	2.37*	2.36*	2.25*	18.83**	0.35	2.37*	0.72	0.63	0.82	5.12*
	12	2.26	0.73	0.64	0.61	0.41	0.39	5.93	3.61**	4.67**	4.24**	1.99*	14.25**	2.96	1.52	1.27	1.09	0.69	2.75
	36	5.97	0.67	1.19	1.16	0.55	0.32	12.89	2.27*	3.94**	3.38**	1.88	9.45**	3.08	0.64	1.15	1.36	0.37	0.36
	60	12.44	0.91	2.04*	1.71	0.90	0.74	14.88	1.55	3.15**	2.95**	1.69	5.60*	5.38	0.72	1.53	1.73	0.52	0.50
tbl	K	$\hat{A}_K$	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}	$\hat{A}_K$	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}	$\hat{A}_K$	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}
	1	-0.12	-2.40*	-2.20*	-2.17*	-2.19*	5.21*	-0.02	-0.48	-0.54	-0.54	-0.55	0.14	0.06	1.85	1.63	1.55	1.69	3.66
	12	-0.96	-1.49	-1.84	-1.84	-1.58	2.44	0.08	0.20	0.31	0.27	0.26	0.10	0.38	0.85	0.98	0.89	1.00	1.32
	36	-2.05	-1.14	-2.06*	-1.63	-1.29	1.28	0.53	0.35	0.70	0.59	0.60	0.36	1.52	1.30	2.21*	2.23*	1.51	2.46
	60	-2.23	-0.77	-1.11	-0.96	-0.90	0.67	0.22	0.10	0.17	0.16	0.16	0.10	2.56	1.40	3.98**	3.78**	1.64	2.53
tms	K	$\hat{A}_K$	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}	$\hat{A}_K$	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}	$\hat{A}_K$	t_{SCALED}	t_{NW}	t_{HH}	t_{HOD}	W_{IVX}
	1	0.23	2.16*	1.96*	1.90	2.02*	4.39*	0.09	1.17	1.18	1.16	1.21	1.81	-0.08	-1.13	-0.97	-0.89	-1.05	0.83
	12	2.65	1.90	2.56*	2.52*	2.09*	4.98*	0.25	0.27	0.36	0.30	0.34	0.20	0.34	0.34	0.38	0.33	0.44	0.64
	36	5.98	1.60	4.08**	3.65**	2.05*	4.44*	-2.22	-0.68	-1.23	-1.19	-1.24	1.31	-1.08	-0.42	-0.77	-0.80	-0.63	0.06
	60	8.28	1.39	2.46*	2.19*	2.07*	4.76*	-4.09	-0.73	-1.60	-2.92**	-1.52	2.68	-3.14	-0.78	-1.94	-2.12*	-1.35	0.99

Table 5

Multivariate predictive regressions for monthly MKT, SMB, and HML factor returns

This table presents the results of multivariate K −horizon predictive regressions for monthly equity factor returns, as in equation (8), testing the joint null hypothesis $H_0: A = 0_{1xr}$ in (1). MKT, SMB, and HML denote the excess market, size, and value factors, respectively, from the Fama-French 3-factor model. Results are presented for 6 combinations of predictive regressors. Combination I uses d/p and tbl. Combination II employs e/p and tbl. Combination III involves d/p and e/p, whereas Combination IV uses d/p, e/p, and tbl. Combination V utilizes e/p, b/m, and tms. Combination VI employs d/p, tbl, tms, and dfy. The sample period for combinations that involve tbl or tms is 1952-2017. For Combination III, the sample period is 1927-2017. Results are reported for the predictive horizons $K = 1, 12, 36, 60$. W_{SCALED} refers to the Wald statistic computed from an ordinary least squares K −horizon predictive regression scaled by K. W_{NW} denotes the Wald statistic computed with Newey-West (1987) standard errors. W_{HH} represents the Wald statistic computed with Hansen-Hodrick (1980) standard errors. W_{HOD} refers to the corresponding Wald statistic computed with Hodrick (1992) standard errors. W_{IVX} denotes the IVX-Wald test statistic defined in (23). * and ** indicate significance at the 5% and 1% level, respectively.

		Panel A: MKT					Panel B: SMB					Panel C: HML				
Combination I	K	W_{SCALED}	W_{NW}	W_{HH}	W_{HOD}	W_{IVX}	W_{SCALED}	W_{NW}	W_{HH}	W_{HOD}	W_{IVX}	W_{SCALED}	W_{NW}	W_{HH}	W_{HOD}	W_{IVX}
d/p & tbl	1	14.56**	13.19**	13.04**	12.52**	6.22*	1.05	0.67	0.69	0.66	1.63	4.53	3.01	2.73	3.20	6.26*
	12	10.50**	14.63**	10.80**	10.85**	4.32	0.64	0.84	0.59	0.43	1.70	0.74	0.96	0.79	0.99	1.80
	36	7.83*	22.37**	25.87**	6.89*	2.36	0.16	0.73	0.63	0.40	0.48	1.70	4.95	5.25	2.34	3.23
	60	4.50	93.38**	–	4.76	1.46	0.13	0.68	0.99	0.21	0.75	1.98	15.87**	14.80**	2.82	3.51
Combination II	K	W_{SCALED}	W_{NW}	W_{HH}	W_{HOD}	W_{IVX}	W_{SCALED}	W_{NW}	W_{HH}	W_{HOD}	W_{IVX}	W_{SCALED}	W_{NW}	W_{HH}	W_{HOD}	W_{IVX}
e/p & tbl	1	12.89**	11.54**	11.36**	11.25**	11.25**	0.40	0.42	0.42	0.43	0.36	5.54	4.44	4.06	4.64	6.32*
	12	7.03*	10.19**	8.40*	8.47*	7.71*	0.12	0.17	0.12	0.12	0.31	1.30	1.39	1.10	1.75	2.96
	36	4.27	12.98**	13.48**	5.64	4.48	0.12	0.56	0.44	0.36	0.36	1.99	6.69*	7.04*	2.74	3.47
	60	1.59	5.55	4.62	2.86	2.22	0.01	0.03	0.03	0.04	0.13	2.60	18.60**	15.48**	3.76	4.02
Combination III	K	W_{SCALED}	W_{NW}	W_{HH}	W_{HOD}	W_{IVX}	W_{SCALED}	W_{NW}	W_{HH}	W_{HOD}	W_{IVX}	W_{SCALED}	W_{NW}	W_{HH}	W_{HOD}	W_{IVX}
d/p & e/p	1	3.76	2.97	2.60	3.61	3.38	3.93	2.29	2.38	2.17	4.34	1.95	1.11	0.96	1.35	1.84
	12	3.33	5.84	4.55	5.07	3.79	4.06	5.70	3.79	2.16	6.14*	1.23	1.12	0.82	1.06	1.48
	36	2.73	7.92*	7.37*	4.51	2.98	1.08	1.51	1.07	1.02	2.76	0.76	0.97	0.88	0.65	0.47
	60	2.90	12.75**	23.28**	3.92	2.63	0.34	0.76	0.68	0.54	1.06	0.88	1.42	1.56	0.71	0.57
Combination IV	K	W_{SCALED}	W_{NW}	W_{HH}	W_{HOD}	W_{IVX}	W_{SCALED}	W_{NW}	W_{HH}	W_{HOD}	W_{IVX}	W_{SCALED}	W_{NW}	W_{HH}	W_{HOD}	W_{IVX}
d/p, e/p & tbl	1	15.44**	14.96**	15.18**	13.71**	6.26	1.11	0.79	0.92	0.72	1.69	5.55	4.51	4.10	4.73	6.19
	12	10.72*	14.82**	10.80*	11.02*	4.26	0.73	1.94	1.61	0.69	1.84	1.57	4.01	3.12	2.96	2.89
	36	7.87*	23.90**	38.26**	6.89	2.66	0.19	0.77	0.66	0.59	0.71	2.08	9.20*	10.14*	3.58	3.25
	60	4.61	110.20**	–	5.51	3.65	0.30	1.45	1.70	2.03	3.12	2.84	29.94**	34.09**	5.49	3.88
Combination V	K	W_{SCALED}	W_{NW}	W_{HH}	W_{HOD}	W_{IVX}	W_{SCALED}	W_{NW}	W_{HH}	W_{HOD}	W_{IVX}	W_{SCALED}	W_{NW}	W_{HH}	W_{HOD}	W_{IVX}
e/p, b/m & tms	1	8.09*	7.24	6.88	7.47	6.58	4.70	5.35	6.58	4.57	5.03	3.33	2.43	2.14	2.73	2.66
	12	7.11	12.44**	9.38*	9.25*	7.88*	4.06	8.17*	5.51	5.03	5.48	2.50	6.20	4.59	4.35	4.03
	36	5.22	27.02**	26.56**	6.81	6.01	1.37	3.53	2.53	3.38	3.17	3.10	6.45	7.97*	6.15	4.89
	60	3.17	12.93**	10.05*	5.36	5.79	1.05	3.41	36.22**	3.10	3.17	3.65	32.36**	41.31**	7.14	5.79
Combination VI	K	W_{SCALED}	W_{NW}	W_{HH}	W_{HOD}	W_{IVX}	W_{SCALED}	W_{NW}	W_{HH}	W_{HOD}	W_{IVX}	W_{SCALED}	W_{NW}	W_{HH}	W_{HOD}	W_{IVX}
d/p, tbl, tms & dfy	1	17.12**	14.76**	15.13**	13.45**	10.36*	8.97	9.69*	11.11*	8.59	9.85*	8.18	3.57	3.14	3.86	8.62
	12	13.79**	19.93**	14.53**	12.92*	10.00*	2.93	9.36	7.53	3.92	4.74	1.50	4.19	3.71	2.69	4.23
	36	11.09*	39.59**	42.35**	9.43	7.63	0.95	7.51	12.08*	4.00	3.05	1.80	6.21	6.02	2.45	3.54
	60	7.74	69.58**	188.90**	9.44	7.99	0.82	3.89	–	3.65	3.60	2.03	20.70**	19.32**	3.55	3.45

Figure 1

Power plots for sample size $n = 1,000$ and predictive horizon $K = 10$

This figure shows the rejection rates for tests of the null hypothesis $H_0: A = 0$ versus the alternative $H_1: A \neq 0$ in (1), as the true value of A increases, using a K −horizon predictive regression model with $K = 10$. The reported rejection rates for each test with 5% nominal size (horizontal line) are based on the Monte Carlo simulation described in Section 3.1 with 1,000 repetitions, sample size $n = 1{,}000$, and no autocorrelation in the residuals of the autoregressive equation (2). The solid curve ($Wald_{IVX}$) shows the rejection rate for the IVX-Wald test statistic defined in (23). The dashed curve (t_{Bonf}) illustrates the rejection rate for the Bonferroni test statistic and the dotted one (t_{scaled}) the scaled, by $\sqrt{K}$, t-statistic of Hjalmarsson (2011). The dash-dot curve (t_{Hod}) shows the rejection rate for the t-statistic computed with Hodrick (1992) standard errors. Power plots are presented for different combinations of the local-to-unity parameter, $C = 0, -5, -10, -50, -100, -500$, and the degree of correlation between the residuals of regressions (1) and (2), $\delta = -0.99, -0.5,\ 0$.

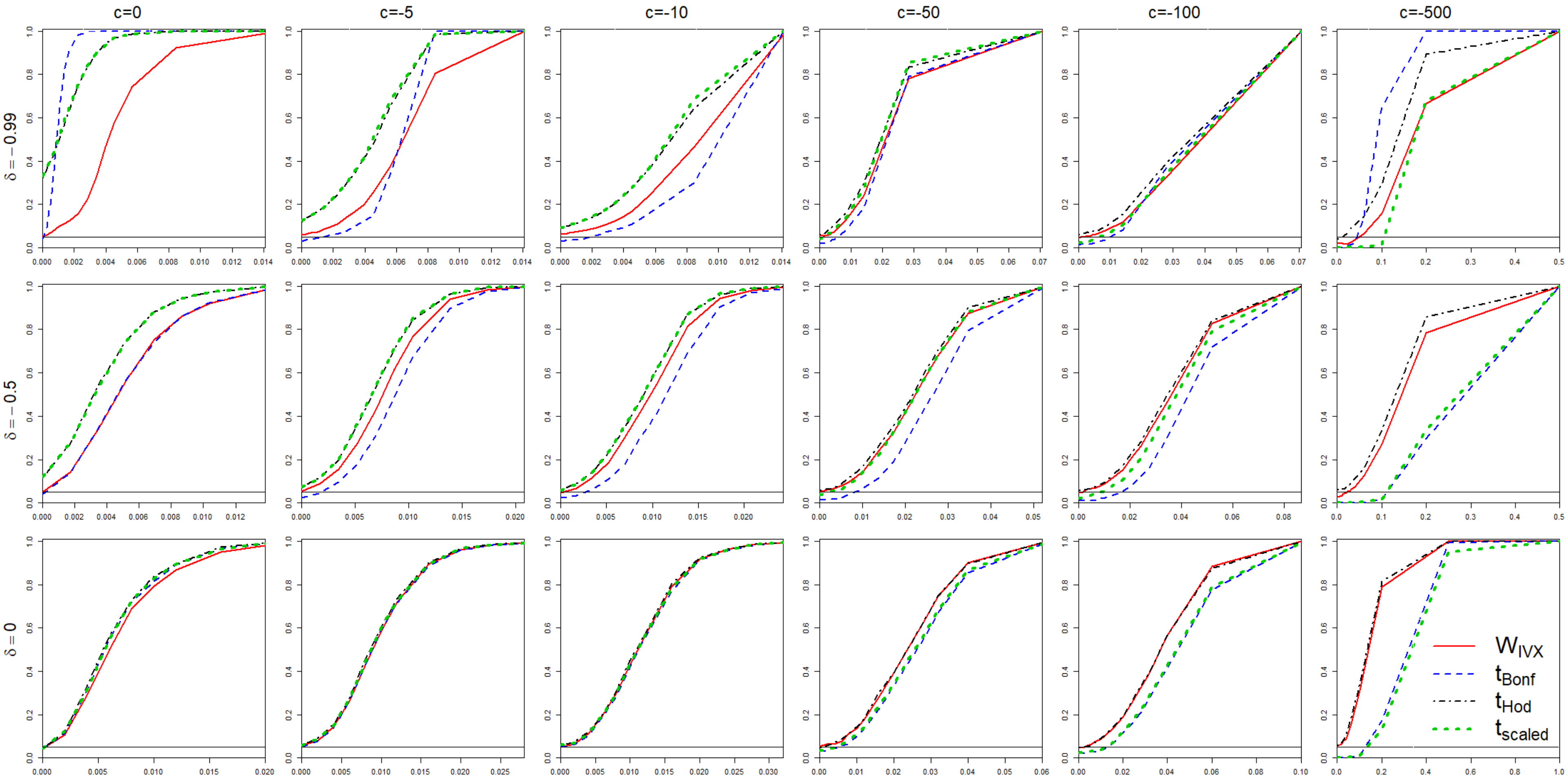
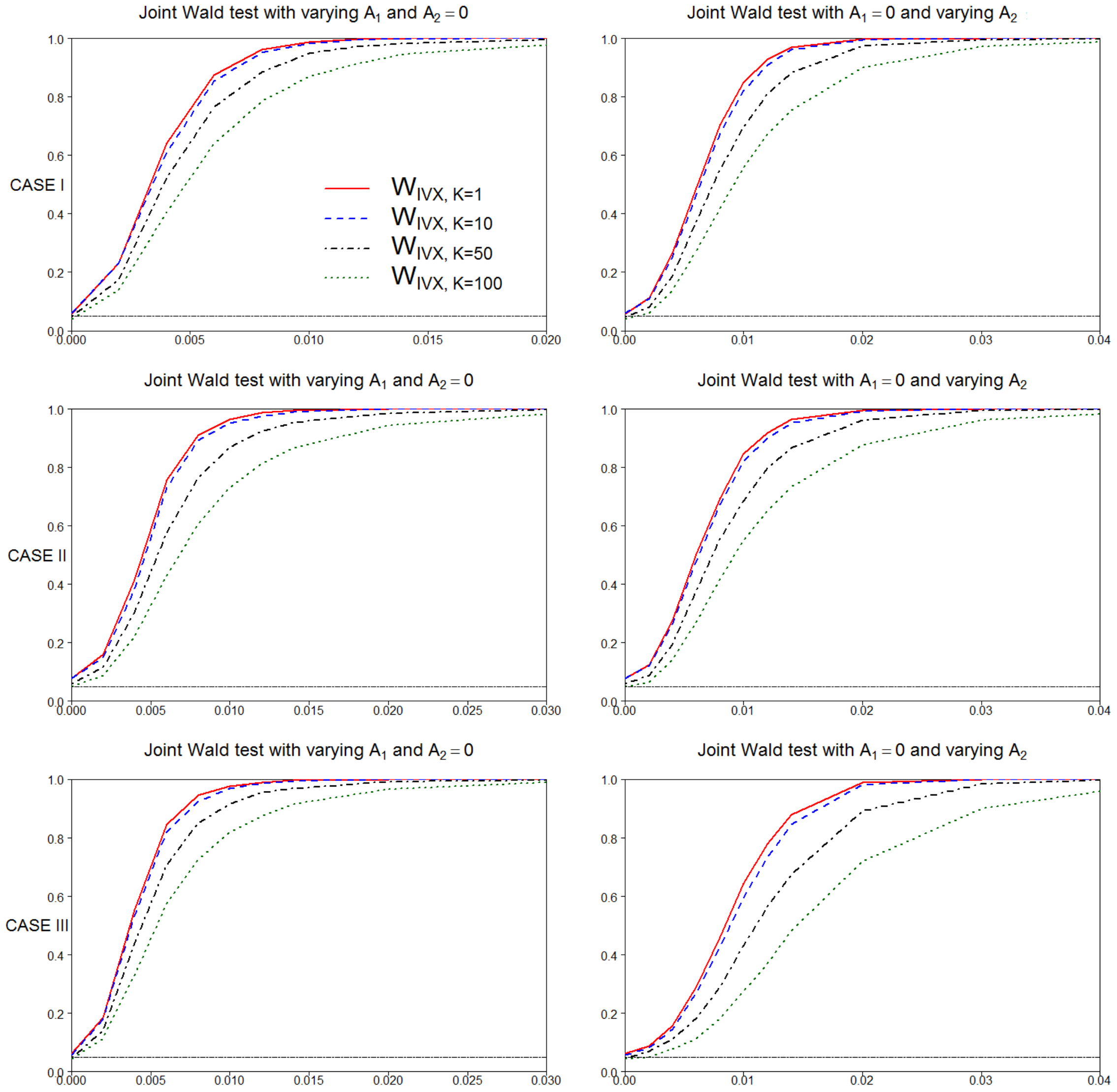


Figure 2
Power plots of the joint IVX-Wald test statistic with two regressors and sample size $n = 1,000$

This figure shows the rejection rates for joint tests based on the predictive system in equation (25) with two persistent regressors, using the IVX-Wald test statistic defined in (23) with 5% nominal size (horizontal line). The reported rejection rates are computed using the Monte Carlo simulation described in Section 3.3, with 1,000 repetitions and sample size $n = 1{,}000$, under the null hypothesis H_0: $A = 0_{1x2}$ in (25), i.e., that the slope coefficients of both regressors are equal to zero, as the true value of each regression coefficient A_i increases. In each row, the left panel illustrates the rejection rate of the test statistic as the true value of the first regressor's coefficient (A_1) increases whereas the second regressor's coefficient remains equal to zero ($A_2 = 0$). The right panel illustrates the corresponding rejection rate as the true value of the second regressor's coefficient (A_2) increases whereas the first regressor's coefficient remains equal to zero ($A_1 = 0$). Results are reported for three different combinations of the relevant parameter values (C, Φ and Σ). diag(C) provides the local-to-unity parameters of the regressors employed in each case. For all cases considered data of monthly log excess market return (MKT) is employed for the regressand, whereas for each case monthly data for a combination of two regressors (Predictors) is used. For each case, the estimated autocorrelation coefficients ($\phi's$) in the residuals of the autoregressive equations are reported (diag(Φ)) as well as the degrees of correlation ($\delta's$) between between ε_t and u_t of with matrix Σ given in (25). For each case, the combination of predictors employed for the estimation of the simulation parameters along with their values are:

	Predictors	Data period	diag(C)	diag(Φ)	Σ
Case I	dividend-price ratio, T-bill rate	1952-2017	(0, -5)	(0.0640, 0.3379)	$\begin{pmatrix} 1 & -0.9827 & -0.1251 \\ -0.9827 & 1 & 0.3379 \\ -0.1251 & 0.3379 & 1 \end{pmatrix}$
Case II	earnings-price ratio, default yield spread	1927-2017	(0, -5)	(0.2741, 0.2167)	$\begin{pmatrix} 1 & -0.7596 & -0.2787 \\ -0.7596 & 1 & 0.1246 \\ -0.2787 & 0.1246 & 1 \end{pmatrix}$
Case III	earnings-price ratio, T-bill rate	1952-2017	(0, -10)	(0.3538, 0.3379)	$\begin{pmatrix} 1 & -0.6156 & -0.1282 \\ -0.6156 & 1 & 0.1583 \\ -0.1282 & 0.1583 & 1 \end{pmatrix}$

These power plots are illustrated for different predictive horizons K. The solid curve corresponds to $K = 1$, the dashed curve corresponds to $K = 10$, the dash-dot curve corresponds to $K = 50$, and the dotted curve corresponds to $K = 100$.


Joint Wald test with varying A_1 and $A_2 = 0$
Joint Wald test with $A_1 = 0$ and varying A_2
CASE I
CASE II
CASE III
$W_{IVX, K=1}$
$W_{IVX, K=10}$
$W_{IVX, K=50}$
$W_{IVX, K=100}$