A new application of exact nonparametric methods to long-horizon predictability tests

Wei Liu∗
Department of Economics
University of Toronto

Alex Maynard
School of Business and Economics
Wilfrid Laurier University

August 29, 2006

Abstract

Empirical results from long-horizon regression tests have been influential in the finance literature. Yet, it has come to be understood that traditional long-horizon tests may be unreliable in finite samples when regressors are persistent and when the horizon is long relative to sample size. Recent research has provided valid alternative inference procedures in long-horizon regression in the case for which the regressor follows a near-unit root autoregressive process. However, in small samples, such processes may sometimes be difficult to distinguish with confidence from other persistent data generating processes, such as those displaying long-memory or structural breaks. In this paper, we demonstrate a simple means by which existing nonparametric sign and signed rank tests may be applied to provide exact inference in long-horizon predictive tests, without requiring any modeling assumptions on the regressor. Employing this robust approach, we find evidence of stock return predictability at moderate horizons using short-term interest rates, but little evidence of either short or long-run predictability using dividend-price ratios.

JEL Classification: C12,C22,G14.

Keywords: sign test, signed rank test, long-horizon regression, predictive regression, efficient market hypothesis, structural breaks, long-memory.

∗Correspondence may be sent to W. Liu (leo.liu@utoronto.ca) or A. Maynard (amaynard@wlu.ca). We thank B. Campbell, J.-M. Dufour, R. Luger, L. Kilian, Z. Xiao, E. Zivot, seminar participants at the University of Washington and conference participants at the 2005 CIREQ Time Series Conference in Montréal for useful discussion and comments. We are also grateful for useful comments and suggestions from two anonymous referees. Maynard thanks the SSHRC for research funding. We take full responsibility for all remaining errors.
1 Introduction

Predictability testing has many important applications in finance, including tests of stock return predictability (Fama and French, 1988; Campbell and Shiller, 1988; Campbell, Lo and MacKinley, 1997), the expectations hypothesis of the term structure (Lanne, 2000), and the forward rate unbiasedness hypothesis (Fama, 1984). A common feature of predictability regressions is the persistent or near-nonstationary behavior of the regressor. This leads to well known problems of size distortion in predictability testing (Mankiw and Shapiro, 1986), and has generated substantial interest in both econometrics and empirical finance (Cavanagh, Elliott and Stock, 1995; Stambaugh, 1999; Campbell and Yogo, 2006; Jansson and Moreira, 2006, for example).

While many econometric solutions have been proposed to address this problem, the sign and signed rank tests of Campbell and Dufour (1995, 1997) have a number of unique and appealing aspects to them. First, they are the only tests we are aware of that provide correct size without any modeling assumptions whatsoever on the regressor. This provides for even greater generality than most competing procedures based on local-to-unity modeling assumptions. For example, it allows for both unmodeled structural breaks and long-memory, two alternatives that may in practice be difficult to distinguish from near-unit roots. Secondly, while common size corrections provide improved asymptotic approximations to finite sample behavior, the sign and signed rank tests provide for exact finite sample inference under weak conditions.

One practical limitation of finite sample sign and signed rank tests is that they require white noise assumptions on the dependent variable under the null hypothesis. This assumption is satisfied in simple one-period predictability tests, since the dependent variable is by assumption unpredictable under the null hypothesis. However, it rules out the direct application of these robust tests to long-horizon predictability regressions, which test for predictability in long-horizon returns cumulated over several periods. Yet, long-horizon regressions have been particularly influential in the empirical literature (Fama and French, 1988; Campbell and Shiller, 1988; Mark, 1995; Campbell et al., 1997), especially as they typically show much stronger evidence of stock return predictability. By suggesting a simple means of applying sign and signed rank tests in a long-horizon setting, we hope to widen the scope for empirical application of nonparametric predictive testing.\footnote{Campbell and Galbraith (1993) and Campbell and Dufour (1997) employ sign and signed rank statistics to test the expectation hypothesis of the term structure. In their case, they use 3 month horizons with monthly data, but split the data into three separate non-overlapping sub-samples, as in the sample-splitting technique described in the paragraph below. Other sign test applications have used return horizons that match the sampling frequency. For example, Maynard (2006) employs sign tests to provide robust tests of forward rate unbiasedness and Wu and Zhang (1997) employ them to test the sign of the spot return/forward premium relationship.}
The reason that sign tests cannot be directly applied to long-horizon regressions is that the return horizon in these regressions (e.g., 4 years) typically exceeds the sampling frequency (e.g., 1 month). Thus, the returns on the left-hand side (LHS) overlap for multiple periods thereby violating the required white noise assumptions. In fact, when the horizon length is long relative to the sampling frequency, this also causes additional problems for statistical inference in predictive regression. While HAC standard errors may be sufficient to correct inference at moderate horizons, the standard asymptotics start to break down at longer horizons (Richardson and Stock, 1989; Valkanov, 2003).

When the horizon length is moderate (or fixed) relative to the sample size, one elegant solution to this problem, suggested briefly in Campbell and Dufour (1995), would be to employ the sample splitting methods of Dufour and Torres (1998). In this approach, the original sequence of overlapping $k$-period returns would be divided into $k$ sub-samples of independent (or at least non-overlapping) $k$ period returns. The sign/signed rank test could then be applied to each sub-sample, with bounds procedures used to pool the individual test results. Although not yet fully explored, this approach seems likely to work quite well for moderate or fixed horizon lengths. On the other hand, its suitability under the long-horizon assumptions of Richardson and Stock (1989) and Valkanov (2003) is somewhat less clear, since these assumptions would imply fixed sub-sample sizes.

In this paper, we propose a simple alternative approach by which nonparametric sign and signed rank tests may be applied to provide exact inference under very weak assumptions in long-horizon predictive tests. Our empirical strategy is motivated by a rearrangement of the predictive regression considered earlier in the finance literature (Jegadeesh, 1991; Cochrane, 1991). Jegadeesh (1991) and Cochrane (1991) show that the regression of a long-horizon return on a single period predictor may be replaced by a regression of a one period return on a long-horizon regressor without fundamentally altering the interpretation of the null hypothesis. The advantage of replacing a long-horizon LHS variable with a long-horizon right-hand side (RHS) variable is that we recover the white noise assumption on the LHS variable under the null hypothesis. On the other hand, this further adds to the persistence in the RHS variable and Hodrick (1992) finds that the regression based predictability tests using this rearrangement still show substantial size distortion.

We propose the application of long-horizon sign and signed rank tests based on this alternative specification of Jegadeesh (1991) and Cochrane (1991). In

---

2 Under long-horizon assumptions the number of horizons and thus the number of sub-samples required for sample-splitting techniques is a fixed fraction of sample size. See Section 3.3 for further discussion.
fact, the benefits of this alternative long-horizon specification are substantially greater for the sign/signed rank tests than they are for the regression tests. This is because these tests are particularly sensitive to dependence in the LHS variable, but insensitive to the dependence in the RHS variable. Thus, while the sign/signed rank tests of Campbell and Dufour (1995, 1997) cannot be directly applied to the original form of the long-horizon regression, they provide exact finite sample inference using the rearranged version of this regression.

Before turning to the empirical study, we provide an extensive Monte Carlo comparison of the sign and signed rank test procedures to the sup-bound version of Valkanov (2003)’s long-horizon predictive test.\(^3\) Our interest focuses on the robustness of the procedures to alternative forms of persistence used to characterize the regressor, such as long-memory/fractional integration and various types of structural break processes, that may in practice be difficult to distinguish from the near-unit root autoregressive models upon which many predictive tests are commonly based. Although the performance of long-horizon tests has been well-studied in the local-to-unity case, we know of no prior evidence regarding their robustness under these alternative modeling specifications. Thus, in addition to providing a natural basis of comparison to the sign and signed rank tests, we believe these results may be of practical interest in themselves.

The traditional long-horizon tests using robust standard errors were found to severely over-reject in all of the data generating processes considered. The Valkanov (2003) sup-bound test, which is designed for autoregressive/near-unit root regressors, turned out to be fairly robust in more general specifications. While it modestly over-rejected in a few specific cases, it tended to be fairly conservative under less persistent parameterizations of the data generating process for the predictors. Overall, the sign and signed rank tests provided the most consistent accuracy in terms of size across the full set of models considered. This good performance results from the specification-free properties and exact finite sample distributions of these nonparametric tests.

\(^3\)Another version of the Valkanov (2003) test imposes the assumptions of the Gordon Growth Model to obtain a plug-in estimate of the local-to-unity parameter. Since, in general, the local-to-unity parameter cannot be consistently estimated using time series data, we did not include it in our simulation study. More recently, Rossi (2003) provides novel inference procedures in long-horizon expectation tests, which base the decision rule on Bonferroni type confidence bounds. Since Rossi’s methods also employ a local-to-unity setting, it would be interesting to consider their performance under alternative specifications. We leave this for future work. Many other predictive test procedures have been developed for short or fixed-horizon tests with persistent regressors (Cavanagh et al., 1995; Maynard and Shimotsu, 2004; Jansson and Moreira, 2006, e.g.) in the short-horizon context, but Valkanov (2003) and Rossi (2003) provide the only explicit long-horizon test procedures that we are aware of.
By construction, these nonparametric tests, which use only information on signs and/or signed ranks, may in some cases be less powerful than their parametric counterparts. Therefore, knowing their performance in terms of power relative to parametric approaches is of importance. In the long horizon case, we found that the power of the sign and signed rank tests were comparable in many cases to that of the parametric sup-bound test. The ranking of the two approaches varied with the specification of the model and the choice of model parameters.

We employ our new approach to provide exact inference in long-horizon stock return predictability tests employing both the one-month treasury bill and the dividend-price ratio as predictors, with return horizons ranging from one-month to four years. Earlier influential work (Fama and French, 1988; Campbell and Shiller, 1988; Campbell et al., 1997) found strong evidence of stock return predictability using both series, particularly for the dividend-price ratio in long-horizons. However, both predictors are highly persistent, and predetermined but not exogenous, suggesting the possibility of size distortion even in short-horizons. Consequently, a recent literature has been devoted to the application of appropriately sized stock return predictability tests at both short (Stambaugh, 1999; Lewellen, 2004; Wolf, 2000; Campbell and Yogo, 2006, e.g.) and long (Valkanov, 2003; Rossi, 2003; Torous, Valkanov and Yan, 2005, e.g.) horizons. Nonetheless, these studies still impose autoregressive or local-to-unity modeling assumptions on the predictor, ruling out interesting possibilities such as structural breaks or long-memory behavior. Ours is the first empirical study we know of to conduct valid long-horizon predictability tests without imposing any assumptions on the behavior of the regressors. Using this robust approach, we confirm the existing evidence of return predictability using the treasury bill at short to medium horizons, but find no significant evidence of predictability at either short or long-horizons employing the dividend-price ratio as a predictor.

The rest of this paper is structured as follows. Section 2 introduces long-horizon predictability tests, explains their size distortion, and discusses the corrections currently available in an autoregressive/near-unit root context. Section 3 presents and motivates the robust approach considered here. Simulations are provided in Section 4 and Section 5 employs our new approach to provide robust tests for long-horizon predictability in stock returns. Section 6 concludes.
The traditional long-horizon framework

In what follows, \( y_{t+1} \) denotes a single period financial return, such as a stock or exchange rate return, and \( x_t \) is a predetermined predictor such as an interest rate or dividend-price ratio. Predictive tests are commonly implemented via either the coefficient restriction \( \beta = 0 \) in standard predictive regressions of the form

\[
y_{t+1} = \alpha + \beta x_t + \varepsilon_{1,t+1},
\]

or alternatively, as the restriction \( \beta(k) = 0 \) in long-horizon regressions of the form

\[
y_{t+k} = \alpha(k) + \beta(k)x_t + \varepsilon_{1,t+k}, \quad \text{where}
\]

\[
y_{t+k} = y_{t+1} + \ldots + y_{t+k}
\]

defines the \( k \)-period return and residual, \( \varepsilon_{1,t+k} \), satisfies

\[
\varepsilon_{1,t+k} = \varepsilon_{1,t+1} + \ldots + \varepsilon_{1,t+k},
\]

when \( \beta(k) = 0 \).

Since the hypothesis of interest restricts only the relation between \( y_{t+k} \) and \( x_t \) (i.e. \( \beta(k) = 0 \)), it is in principle unnecessary to model \( x_t \) so long as it behaves in a sufficiently stationary manner. In practice, however, the observed behavior of \( x_t \) is often quite persistent, in which case the model choice for \( x_t \) can affect the statistical behavior of \( \hat{\beta} \) (or \( \hat{\beta}(k) \)).

There are a number of possible modeling strategies that can be used to capture the persistence in \( x_t \). Most common is the autoregressive framework of the type given by

\[
(1 - \phi L) b(L)(x_t - \mu) = \varepsilon_{2,t},
\]

where \( \phi \) represents the largest autoregressive root, \( L \) denotes the lag operator, and the largest root of the polynomial \( b(L) \) is strictly less than one. To complete this form of the model, the residuals may then be modeled as the white noise process,

\[
\varepsilon_t = \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix} \sim W N(0, \Sigma)
\]

\[
\Sigma = \begin{pmatrix} \sigma_{11}^2 & \delta \sigma_{11} \sigma_{22} \\ \delta \sigma_{11} \sigma_{22} & \sigma_{22}^2 \end{pmatrix}
\]

\(^4\)For given specifications of \( x_t \), \( \beta \) and \( \beta(k) \) can be explicitly related (Rossi, 2003).
with cross-correlation $\delta$ unrestricted under the null hypothesis.

In this model, the parameter $\phi$ plays the key role in determining the persistence of the process. For fixed $\phi$, $x_t$ is integrated of order zero (I(0)) for $\phi < 1$ and integrated of order one (I(1)) for $\phi = 1$. The large sample behavior of many estimators and test statistics is known to depend critically on this distinction. On the other hand, it is often difficult in practice to distinguish with confidence between these two alternatives, and large sample approximations based on either assumption may often be misleading when $\phi$ is close, but not equal, to one. This near-unit root case is arguably better approximated by the local-to-unity model, in which the largest root is modeled as

$$\phi = 1 + \frac{c}{T},$$

where $c = 0$ corresponds to the unit root, small negative values of $c$ correspond to near-unit roots, and $c << 0$ approximates a stationary process. Although seldom taken literally as a data generating process, large sample theory based on this model is often more accurate in small sample sizes, than those based on a fixed value of $\phi$.

### 2.1 Two sources of size distortion

There are two well-documented sources of size distortion that may arise in long-horizon regression of the type specified in (2). The first source of size distortion arises from the fact that many predictors, such as dividend and earning price ratios, interest rates, and forward premia, are highly persistent, but only predetermined, rather than fully exogenous. Since this problem is common to both the short and long-horizon regressions we will focus our discussion on (1). In the context of (8), (1) has the form of a cointegrating regression when $c = 0$ and describes cointegration between near I(1) variables for $c < 0$. As in the cointegrating regression, when $\delta \neq 0$, the regression coefficient $\hat{\beta}$ remains consistent, but has a nonstandard limiting distribution. Likewise, standard test-statistics based on this parameter, do not have their usual limiting distribution (Cavanagh et al., 1995, for example). Critical values based on the standard normal distribution are valid only for $\delta = 0$ and/or $c << 0$. Otherwise, the use of standard critical values is well known to generate size distortion, with rejection rates of up to nearly 30% for a 5% nominal test in a model with intercept and reaching above 50% in a model with trend (Cavanagh et al., 1995; Mankiw and Shapiro, 1986).

In the long-horizon context, these problems are further complicated by the fact that the error term in (1) is no-longer serially uncorrelated even under the
null hypothesis, in which case it follows an MA(k-1), as seen from (4). If k is small relative to T, the OLS estimator \( \hat{\beta}(k) \) will yield a consistent estimate of \( \beta(k) \). Hence assuming stationary \( x_t \) and fixed k, the usual asymptotic theory with HAC standard errors (Hansen and Hodrick, 1980; Newey and West, 1987; Andrews, 1991, for example), still provides for appropriate statistical inference, at least in large sample. However, for the sample sizes encountered in practice, this fixed-k theory often provides a poor approximation to the sampling distribution (Richardson and Stock, 1989; Valkanov, 2003). As these authors show, a better asymptotic approximation is obtained by modeling the horizon length as a fixed proportion of sample size, i.e.

\[
k = \lambda T, \quad \text{for } 0 < \lambda < 1.
\]

Under these assumptions, the behavior of the regression coefficient \( \hat{\beta}(k) \) is quite different. Since k increases linearly in sample size, the residual \( \varepsilon_{1,t+k}^k \) behaves like an I(1) process, with \( \frac{1}{\sqrt{T}} \varepsilon_{1,t+k}^k \) satisfying a functional central limit theorem (Valkanov, 2003). Under the null of \( \beta(k) = 0 \), \( y_{t+k}^k = \alpha(k) + \varepsilon_{1,t+k}^k \) behaves similarly. Therefore, when \( x_t \) is also persistent, so that (8) applies, and the null \( \beta(k) = 0 \) holds, Valkanov (2003) shows that (2) exhibits the basic characteristics of a spurious regression (Granger and Newbold, 1974; Phillips, 1986), in which \( y_{t+k}^k \) is integrated, but not cointegrated with the near unit root regressor \( x_t \). In particular, the t-statistic of the standard long-horizon regression \( t_{\beta} \), say) diverges at rate \( \sqrt{T} \), giving rise to the same false significance observed in spurious regression. In other words, rejection rates under the null hypothesis approach one, for large T (and hence large k).

### 2.2 Corrections based on autoregressive/near unit root models

A large number of modified testing procedures have been proposed to correct inference in (1) in the context of the autoregressive specification of the type given by (5). The principal difficulty encountered in the local-to-unity model (3) is the dependence of the limit distribution on the local-to-unity parameter \( c \), a parameter which cannot be consistently estimated in a time series context. Cavanagh et al. (1995) suggest a solution to this problem based on bounds procedures, in which the rejection region is chosen conservatively by choosing

---

5Another problem related to long-horizon predictability is investigated by Rossi (2005). In particular, she showed that, when variables used in the predictive regression are persistent, then even the correct model may be beaten out of sample by simple alternatives, such as the random walk, due to the resulting estimation error.
the largest (in absolute value) critical values across a range of possible values for $c$. The range for $c$ may be based on a first stage confidence interval for $c$ as in Stock (1991) (a Bonferroni Bound) or may simply include all plausible values of $c$ (a sup bound). Cavanagh et al. (1995) also suggest a finite sample adjustment to reduce the conservative nature of the Bonferroni test, with which it performs quite well.

Recognizing that $y_t$ is stationary under the null hypothesis but has a near-unit root component under the alternative hypothesis, a second approach has been to reinterpret tests of $\beta = 0$ as stationarity tests on $y_t$ (Wright, 2000; Lanne, 2002). Other solutions include finite sample size corrections (Stambaugh, 1999; Lewellen, 2004), augmented regression methods (Amihud and Hurvich, 2004), and resampling methods (Wolf, 2000, e.g.). Most recently, in an important new development, Jansson and Moreira (2006) develop a test for $\beta = 0$ in (1) which has correct size and optimal power conditional on a set of sufficient statistics for $\phi$.

While the above mentioned procedures have been developed primarily for the short-horizon regression (1), many of them can be extended to longer-horizons under the assumption that the horizon length $k$ is a fixed value independent of sample size. For example, by adjusting the standard errors, Torous et al. (2005) extend the bounds procedure of Cavanagh et al. (1995) to the long-horizon case. Likewise, Jansson and Moreira (2006) show how their procedure may be extended to allow for more general stationary error processes for $\varepsilon_{1,t+1}$, which, for fixed values of $k$, includes moving average errors of the type exhibited by $\varepsilon_{1,t+k}^k$. Such extensions, may be expected to work well when horizon length is moderate relative to sample size, but the results of Richardson and Stock (1989) and Valkanov (2003) suggest some caution regarding the use of fixed-$k$ asymptotic approximations in longer-horizon regressions.

In contrast to the numerous econometric procedures designed to improve inference in short-horizon regression, we know of only a few methods explicitly designed for long-horizon regressions and remaining valid under general conditions, including those specified in (2). Valkanov (2003) proposes tests based on a rescaled t-statistic $t_{\beta}/\sqrt{T}$, where the dependence of its distribution on $c$ may be handled by sup bounds as in Cavanagh et al. (1995). More recently, Rossi (2003) shows that under certain standard assumptions the coefficient $\beta(k)$ in the long-horizon regression can be approximated by $\beta(k) \approx k\beta(1) \left( \frac{e^{\lambda} - 1}{\lambda} \right)$ and uses this to provide a confidence interval on $\beta(k)$ based on a confidence interval for $\beta(1)$ and a confidence interval for $c$. 
3 Robust long-horizon tests

Before introducing the robust tests in Section 3.2 and their application to long-horizon tests in Section 3.3, in Section 3.1 we first discuss a number of alternate models for the persistent behavior in $x_t$, which we employ later in the simulations of Section 4 to investigate the robustness of various test procedures.

3.1 Alternative models for persistent regressors

Although, the local-to-unity model in (5) and (8) provides a very general model for $x_t$, allowing simultaneously for stationary, unit root, and near-unit root regressors, it remains just one of several possible modeling strategies for capturing persistent behavior. Other commonly employed specifications, such as the long memory/fractionally integrated model

$$(1 - L)^d (x_t - \mu_x) = \varepsilon_{2,t} \quad 0 < d < 1$$

where $0 < d < 1$ and various models incorporating structural breaks, may also induce the high persistence features displayed by the forecasting variables. Indeed, the possibilities available for modeling structural breaks are quite rich. They may be modeled as historical breaks in the parameters governing (5), as in

$$(1 - \phi_j L) b_j(L)(x_t - \mu_j) = \varepsilon_{2,t}, \quad \text{for } \tau_{j-1} \leq t \leq \tau_j$$

where $\tau_0 = 1$ and the $\tau_j$ for $j = 1, \ldots, J$ denote the break dates. Alternatively, the breaks can be endogenized as in the Markov switching models (Hamilton, 1989; Diebold and Inoue, 2001)

$$(1 - \phi_{st} L) b_{st}(L)(x_t - \mu_{st}) = \varepsilon_{2,t}, \quad \text{where}$$

$$p_{t,ij} = P(s_t = j|s_{t-1} = i)$$

for $i, j \in \{1, \ldots, J\}$ define the (possibly time-varying) transition probabilities between states and $s_t$ denotes the state at time $t$. Likewise, in the stochastic permanent break (or STOP-BREAK) model (Engle and Smith, 1999; Diebold

\footnote{The fractionally integrated (I(d)) process, for $0 < d < 1$, in which shocks decay hyperbolically provides an intermediate case between the I(1) process, in which shocks are fully persistence and I(0) process, in which shocks decay exponentially. It has stationary long-memory for $d < 0.5$ and is nonstationary for $d > 0.5$. See Baillie (1996) for an excellent survey on long-memory modeling in econometrics.}
and Inoue, 2001)

\[ x_t = \mu_t + \varepsilon_{2,t}, \quad \mu_t = \mu_{t-1} + q_{t-1} \varepsilon_{2,t-1}, \]  

(14)

the random coefficient \( q_{t-1} \) depends on the size of last period’s shock, \( \varepsilon_{2,t-1} \), as in

\[ q_{t-1} = \frac{\varepsilon_{2,t-1}^2}{\gamma_T + \varepsilon_{2,t-1}^2}, \]  

(15)

so that small shocks are primarily temporary, whereas large shocks are permanent and act like breaks.

There is plenty of empirical evidence to suggest that these models may be taken seriously as possible alternatives to the simpler autoregressive framework in (12). For instance, evidence of long-memory has been reported in several financial series. Baillie and Bollerslev (1994) found the monthly forward premium to be well described by an ARFIMA(2,d,0) model; Shea (1991) uncovered long-memory in interest rate spreads and some interest rates in levels; Backus and Zin (1993) documented long-memory evidence in short rate returns. Likewise Perron (1989, 1997) documented evidence of structural breaks in many economic time series. In the financial literature, Timmermann (2001) uncovered structural breaks in the US dividend processes, Atkins and Rakoz (2005) showed that real interest rates in Canada and the U.S. can be well represented by a stochastic break model, Choi and Zivot (2006) and Sakoulis and Zivot (2002) find structural breaks in the forward premium, and Zhou (2005) provided evidence of regime switching in foreign exchange rate data.

Economic or financial theory often offers few a priori guidelines regarding the underlying model for \( x_t \). Moreover, predictive tests concern only the relation between \( y_{t+1} \) and \( x_t \) and not the behavior of \( x_t \) itself. Thus, even in cases where theory does suggest a priori restrictions on the behavior of \( x_t \), the imposition of these restrictions implicitly transforms the predictive test into a test of a joint null hypothesis involving both the non-predictability of \( y_{t+1} \) and the model’s implications for \( x_t \).

While it is generally possible to distinguish empirically between these various models in large sample, it may not always possible to do so with confidence in smaller samples. Theoretical results (Faust 1996 and 1999) suggest certain limits on our ability to infer low frequency behavior in finite sample. Like-

\(^7\)Choi and Zivot (2006), found evidence of both long-memory and structural breaks in the forward premium.

\(^8\)Numerous other empirical studies report evidence on structural breaks in financial time series (Bliss and Smith, 1998; Clemente, Montaño and Reyes, 2003; Repach and Wohar, 2005; Koćenda, 2005, for example).
wise, as the number of breaks \((J)\) grows larger, data generated from models such as \((\Pi)\) increasingly resemble unit root processes, which may be interpreted as processes with a break in every period (Engle and Smith, 1999, e.g.). The more recent literature has also emphasized the near observational equivalence of long-memory/fractionally integrated models and certain structural break processes. Diebold and Inoue (2001) demonstrated that, with an appropriate choice of break probability, both the Markov switching model \((\Pi)\) and the stochastic permanent break model in \((\Pi)\) display long-memory behavior. A number of other studies have demonstrated similar difficulty in distinguishing between long-memory and structural break models (Granger and Hyung, 2004; Gourieroux and Jasiak, 2001; Smith, 2002; Hwang, 2000, for example). Distinguishing \(I(1)\) or near-\(I(1)\) processes from fractionally integrated processes is in principle more straightforward. However, practical difficulties may be encountered here as well. For instance, traditional unit root and stationarity tests often have poor power against long memory alternatives (Diebold and Rudebusch, 1991; Hassler and Wolters, 1994; Lee and Schmidt, 1996) and semi-parametric estimators of the long-memory parameter \(d\), such as the GPH (Geweke and Porter-Hudak, 1983), may be subject to serious bias in finite sample, particularly when the autoregressive persistence is strong (Agiaxloglou, Newbold and Wohar, 1993). These distinctions can matter for both short and long-horizon tests since the limit distribution upon which many such tests are based depend on the autoregressive/local-to-unity modeling assumptions. If \(x_t\) is fractionally integrated, than \((\Pi)\) is fractionally cointegrated and alternative asymptotics are well-known to apply. Lee (2006) provides asymptotics for long-horizon regression when \(x_t\) is fractionally integrated. Inference in long-horizon regression when the regressor obeys the various structural break models discussed above has not to our knowledge been studied. In the sub-sections below, we suggest a simple mechanism by which existing sign and signed rank tests may be used to provide long-horizon predictability tests that remain valid irrespective of the underlying data generating process. We first provide a brief description of existing sign and signed rank tests procedures and then describe how they may be employed in a long-horizon context.

### 3.2 Finite sample sign and signed rank tests

Campbell and Dufour (1995, 1997) propose nonparametric sign and signed rank predictability tests that allow for exact inference without assumption on \(x_t\) thus avoiding the size problems in \((\Pi)\) that plague standard regression tests. Following their notation, we define the time \(t\)-information set
\[ I_t = \sigma(y_t, x_t, y_{t-1}, x_{t-1}, \ldots) \] and let \( g_t \) denote any \( I_t \) measurable conditioning variable (i.e. \( g_t \) is a function of \( y_{t-j} \) and \( x_{t-j} \) for \( j \geq 0 \)). Further, we assume a constant unconditional median for \( y_t \), which we denote by \( b_0 \). Then, using \( g_t \) as the conditioning variable, and slightly generalizing the notation of Campbell and Dufour (1995, 1997) to allow for truncations at the beginning and end of sample, infeasible versions of the sign (\( S_T \)) and signed rank (\( SR_T \)) statistics are respectively defined by

\[
S_T (y_{t+1}, g_t, r, s) = \sum_{t=1+r}^{T-s} u((y_{t+1} - b_0) g_t), \quad \text{and} \quad (16)
\]

\[
SR_T (y_{t+1}, g_t, r, s) = \sum_{t=1+r}^{T-s} u((y_{t+1} - b_0) g_t) R_{t+1}^{+}(b_0), \quad (17)
\]

where \( u(z) \) is a sign indicator function equal to one if \( z \geq 0 \) and zero otherwise, \( R_{t+1}^{+}(b_0) = \sum_{j=1+r}^{T-s} u(|y_{t+1} - b_0| - |y_j - b_0|) \) denotes the rank and \( r \) and \( s \) denote the number of data points truncated at the beginning and end of sample respectively. In the short horizon context one sets \( r = 0 \) and \( s = 1 \).

Campbell and Dufour (1995, 1997) show both statistics to have exact finite sample null distributions under weak assumptions. For the sign test they impose only a constant median \((b_0)\) on \( y_t \) and the independence of \( y_{t+1} \) with respect to \( I_t \) under the null hypothesis. Following the results of Coudin and Dufour (2003, Proposition 3.2), the latter assumption, which implies \( \beta = 0 \) in (11), can be further weakened to require only the following mediangale difference sequence assumption\(^9\)

\[
P(y_{t+1} - b_0 > 0 | I_t) = P(y_{t+1} - b_0 < 0 | I_t) = 0.5, \quad (18)
\]

which as noted by Coudin and Dufour (2003), allows for conditional heteroskedasticity of a general nature. Under these relatively weak assumptions \( S_T \) was shown to have an exact binomial distribution with \( \tilde{T} = T - r - s \) trials and probability of success 0.5. The signed rank test \((SR_T)\) requires the slightly stronger condition that \( y_t \) is continuously distributed and symmetric about \( b_0 \). Then, under the null distribution that \( y_t \) is independent of \( g_t \), it also has an exact finite sample distribution given by the Wilcoxon signed-rank variate of size \( \tilde{T} \), \( W_{\tilde{T}} = \sum_{t=1}^{\tilde{T}} t B_t \), where \( B_t \) are independent Bernoulli random variables with \( P[B_t = 0] = P[B_t = 1] = 0.5 \), \( t = 1, \ldots, \tilde{T} \).

In practical applications the median \( b_0 \) is generally unknown. Campbell and Dufour (1997) combine an exact first-stage confidence interval on \( b_0 \), together

with a Bonferroni bound to provide feasible two-stage versions of sign test and
signed rank test whose rejection rates never exceed the nominal level. They
also consider a simpler plug-in approach, in which $b_0$ is replaced by the sample
median of $y_{t+1}$. This approach is no longer exact in finite sample, but relies
only on the consistency of the sample median for its asymptotic validity. Thus
it may be expected to have reliable size in practice, as found by Campbell
and Dufour (1997) and expanded upon in our simulations below. Furthermore
the power is improved by centering $g_t$ about a sequence of estimated medians.
So long as these sample medians are estimated using only the sub-sample
$g_1, \ldots, g_t$, and are thus predetermined, the finite sample size properties are
unaffected.

3.3 Application to short and long-horizon predictability
tests

The application of the sign and signed rank tests to the short-horizon pre-
dictability testing problem is straightforward. One simply defines $g_t$ as the
value of $x_t$ centered about its median, as described above. The mediangale or
independence assumption of $y_{t+1}$ with respect to $I_t$ differ somewhat from the
martingale difference sequence assumption typically imposed on $\varepsilon_{1,t+1}$ in the
short-horizon regression (1). Nevertheless they conform with the basic notion
that the location of $y_{t+1}$ should be unpredictable under the null hypothesis.
In particular, if $x_t$ has no predictive content for $y_{t+1} - b_0$, it should also have
no predictive content for either its sign or signed rank.

In this case, the sign statistic simply counts the number of times that $x_t$
(e.g. the dividend-price ratio) predicts $y_{t+1} - b_0$ (e.g. the centered stock return)
with the correct sign. Under the null hypothesis this should happen with
probability 0.5, yielding a roughly equal number of correct and incorrectly
signed predictions. On the other hand, if $x_t$ has positive (resp. negative)
predictive content for $y_{t+1}$, as implied by $\beta > 0$ ($\beta < 0$ resp.), then there
should be a significant majority of correct (incorrect) sign predictions. The
signed rank tests operates in a similar fashion, except that rather than treating
all observations equally, it attaches more weight to larger returns.

Both the sign and signed rank tests, which are not based on the parametric
regression procedure, have some very attractive features in the context of the
predictive testing. In particular, they do not require any model specification
for the forecasting variable $x_t$. In other words, they are valid under all possi-
ble specifications for $x_t$, including not only the standard autoregressive, unit
root, and local-to-unity specifications typically considered in the predictive re-
gression literature, but also the long-memory (10) and various structural break
processes (11, 12, & 14) discussed in Section 3.1. Therefore they are unaffected by the persistence and residual cross-correlation that distort the finite sample distributions of standard regression based statistics. Moreover, since the sign and signed rank tests have exact size in finite sample, they do not rely on large sample approximation. Other advantages of the sign test in financial applications include robustness to both outliers and conditional heteroskedasticity of a very general nature.\footnote{On the other hand, potential drawbacks of the nonparametric predictability tests are (i) that they test a conditional median restriction rather than a conditional mean restriction and (ii) that information may be discarded by employing only signs or sign-ranks leading in some cases to a reduction of power. More recently, ongoing work by Luger (2006) notes that Bonferroni versions of these tests may lose power when $x_t$ is persistent due to increased uncertainty regarding the location of the median. As a substitute for the Bonferroni approach, he suggests a permutation method that overcomes this short-coming.}

Such features would be equally, if not more, desirable in the long-horizon predictive regression context, in which the size distortion inherent in standard regression tests is considerably worse. Unfortunately, application of these exact tests to long-horizon predictive tests is somewhat less straightforward and we are aware of no existing empirical long-horizon applications to date.

One of the few assumptions required for the exact finite sample tests described above is that the dependent variable, $y_{t+1}$ must be either a mediangale difference sequence with respect to $I_t$ (the sign test) or it must be fully independent of $I_t$ (the signed rank test). These assumptions are reasonably imposed on the one-period returns $y_{t+1}$ employed in the short-horizon regression (1), which are unpredictable under the null hypothesis that $\beta = 0$. However, this is no longer the case when working with the multi-period returns $y_{t+k}$ typically employed in the long-horizon regression. When $\beta(k) = 0$ in (2) the long-horizon dependent variable is given by $y_{t+k}^k = \alpha(k) + \varepsilon_{1,t+k}$, where $\varepsilon_{1,t+k} = \varepsilon_{1,t+1} + \ldots + \varepsilon_{1,t+k}$ follows a moving average process of order $k - 1$ if $\varepsilon_{1,t}$ is serially uncorrelated. This clearly violates the required mediangale difference sequence and independence assumptions for $k > 1$ and thus the sign and signed rank tests cannot be directly applied to test $\beta(k) = 0$ in (2).

One solution to this problem suggested briefly in Campbell and Dufour (1995) is to employ the type of sample splitting procedures discussed in Dufour and Torres (1998). Briefly stated, in this approach the MA($k-1$) structure of the multi-period returns is used to divide the original dependent return series into $k$ sub-samples of the form $\{y_{j+(s+1)k}\}_{s=1}^T$ for $j = 1, \ldots, k$. The approximately $T/k$ returns within each sub-sample are then mutually independent, allowing for the separate application of the sign/signed rank test to each sub-sample. The overall test would then reject at level $\alpha$ if any one
of the $k$ tests rejected at level $\alpha/k$. This procedure, suggested in a regression context by Dufour and Torres (1998) constitutes a valid conservative inference procedure.

This is a clever approach that should be expected to work well for fixed and/or moderate horizon returns. On the other hand, it is not clear how well the test would work under the long-horizon assumptions in (9), in which each sub-sample size is fixed at approximately $1/\lambda$. When the sub-sample size is too small it may be difficult to pick appropriate critical values for a discrete distributions, such as the binomial, especially for a test of size $\alpha/k$. This makes practical implementation difficult. Likewise, the power implications of fixed (or very small) sub-sample sizes for the test procedure are uncertain.

We suggest an even simpler approach by which the Campbell and Dufour (1995, 1997) sign and signed rank tests may be employed to provide exact long-horizon predictive tests based on an alternative test restriction motivated by the rearrangement of (2) employed previously in the finance literature for regression tests. This rearrangement, which was only partially successful in a regression context, works quite well conjunction with sign and signed rank tests.

To be concrete, instead of employing the sign and signed rank methods to test (2) directly, we instead follow an approach similar to that of Jegadeesh (1991) and Cochrane (1991) who base their test of $\beta(k) = 0$ on a simple rearrangement of (2) under the null hypothesis, that avoids the serial correlation in the residuals. Define a long-horizon version of the regressor $x_t$ as

$$x_t^k = x_{t-k+1} + x_{t-k+2} + \ldots + x_t.$$  

(19)

The intuition behind their procedure follows from the fact that

$$\beta(k) = \text{cov} \left( y_{t+k}, x_t \right) / \text{var} \left( x_t \right).$$

When $x_t$ is stationary, the long-horizon non-predictability restriction $\beta(k) = 0$ is equivalent to the orthogonality condition $\text{cov} \left( y_{t+k}, x_t \right) = 0$ and

$$\text{cov} \left( y_{t+k}^k, x_t \right) = \text{cov} \left( y_{t+1}, x_t^k \right),$$

(20)

where the latter covariance is the numerator of the slope coefficient $\gamma(k)$ in the regression of $y_{t+1}$ on $x_t^k$.

$$y_{t+1} = \gamma_0(k) + \gamma(k)x_t^k + v_{t+1}.$$  

(21)

Thus, the restriction of the null hypothesis, $\beta(k) = 0$, in (2) is equivalent to
the null restriction that $\gamma(k) = 0$ in (21) and Jegadeesh (1991) and Cochrane (1991) therefore suggest testing $\beta(k) = 0$ via the restriction $\gamma(k) = 0$ in (21). Although, the original intuition in (20) requires $x_t$ stationary, even when $x_t$ is nonstationary, as e.g. in model (5) with $\phi = 1$, it can be seen by the law of iterated expectations that the null hypothesis $E_t y_{t+1} = \gamma_0(k)$ (i.e. $\gamma(k) = 0$, $E_t v_{t+1} = 0$ in (21)) still implies $E_t \tilde{y}_{t+k}^k = \alpha(k)$ (i.e. $\beta(k) = 0$, $E_t \tilde{\varepsilon}_{1,t+k}^k = 0$ in (2)), for $\alpha(k) = k\gamma_0(k)$. Thus, rejections of the latter imply rejections of the former.

Since $\gamma(k) \neq \beta(k)$ under the alternative ($\beta(k) \neq 0$) the two tests may yield different power. Yet, in a careful simulation study, Hodrick (1992) reports that the power differences between the two testing procedures are minor, with the specification in (21) often offering a slight improvement.

The principle advantage of this specification is that the new residual $v_{t+1}$ no longer follows an $MA(k-1)$ process, leaving the dependent variable $y_{t+1}$ serially uncorrelated under the null hypothesis. Nevertheless, in a regression context, this rearrangement cannot fully solve the long-horizon inference problem. The simplification of the residual and dependent variable comes at the cost of increasing the persistence in the regressor. In fact, when $k$ is large relative to sample size, as in (2), then the behavior of $x_t^k$ may approximate that of an $I(1)$ process even for $x_t$ i.i.d. More realistically, if $x_t$ itself has a near-unit root as in (3) than for large $k$ the behavior of $x_t^k$ may resemble a near-$I(2)$ process. Perhaps for these reasons, Hodrick (1992) still reports substantial size distortion in tests based on (21).

However, this alternate specification turns out to be much more advantageous in the case of the sign and signed rank tests. The reasoning is as follows. The regression test is potentially distorted by both the persistence in the residuals, as generated by a long-horizon dependent variable in (2), and persistence in the regressor, as generated (or reinforced) by the long-horizon regressor in (21). By contrast, while the sign test is invalidated by the serially correlated residuals in (2), it retains the correct finite sample size, no matter how persistent the regressor $x_t^k$ in (2). Thus there is far more to gain by replacing the long-horizon dependent variable $y_{t+k}^k$ by the long-horizon regressor $x_t^k$.

Our testing strategy may by this point have become apparent. Rather than basing our sign and signed rank tests on the statistics $S_T(y_{t+k}^k, x_{t}^k, 0, k)$ and $SR_T(y_{t+k}^k, x_{t}^k, 0, k)$ and employing the same dependent variable and predictor as in (2), we base our test instead on

$$S_T^k \equiv S_T(y_{t+1}, x_{t}^{k*}, k-1, 1)$$

$$SR_T^k \equiv SR_T(y_{t+1}, x_{t}^{k*}, k-1, 1)$$
where \( x_{t}^{k} \equiv x_{t}^{k} - \text{med}(x_{t}^{k}) \) is the value of \( x_{t}^{k} \) centered about the sample median of \( x_{1}^{k}, \ldots, x_{t}^{k} \). Centering of this type is known to improve test power (Campbell and Dufour, 1997), but does not affect test size since \( \text{med}(x_{t}^{k}) \) is predetermined.

While \( S_T(y_{t+k}^k, x_t^*, 0, k) \) and \( SR_T(y_{t+k}^k, x_t^*, 0, k) \) have unknown and potentially complicated distributions due to the dependence in \( y_{t+k}^k \), \( S_T^k \) and \( SR_T^k \) satisfy the conditions of Campbell and Dufour (1995), and thus their distributions are exactly described by the binomial distribution and Wilcoxon signed-rank variate, respectively. Thus using definitions of the sign and signed rank test we may directly apply the methods of Campbell and Dufour (1997) to provide correct size in finite sample. For ease of reference we state this formally in the remarks below.

**Remark 1** Let \((y_{t+1}, x_t^{k*})\) satisfy (18) and \(P(y_{t+1} - b_0 = 0) = P(x_t^{k*} = 0) = 0, \) for \( t = 1, \ldots, T \). Then \( S_T^k \sim \text{Bi}(\tilde{T}, 0.5) \), where \( \text{Bi}(\tilde{T}, p) \) denotes the binomial distribution with \( \tilde{T} = T - k \) trials and probability of success \( p \).

**Remark 2** Let \((y_{t+1}, x_t^{k*})\) satisfy the assumptions:

\[
\begin{align*}
&y_{t+1} \text{ is independent of } I_t, \text{ for } t = 1, \ldots, T, \text{ and} \\
&y_1, \ldots, y_{t+1} \text{ have continuous distributions symmetric about } b_0, \quad (24a) \\
&P(x_t^{k*} = 0) = 0, \text{ for } t = 1, \ldots, T \\
&\text{(24b)} \\
&\text{(24c)}
\end{align*}
\]

Then \( SR_T^k \sim W_{\tilde{T}} = \sum_{t=1}^{T} tB_t, \) for \( B_t \sim \text{i.i.d. Bi}(1, 0.5) \), where \( W_{\tilde{T}} \) denotes a Wilcoxon signed rank variate of size \( \tilde{T} \).

Upon noting that \( g_t = x_t^{k*} \) remains \( I_t \) measurable, the remarks follow as immediate applications of Campbell and Dufour (1995, Propositions 1 & 2) and Coudin and Dufour (2003, Proposition 3.2). The assumptions on the sign test are quite weak. Nevertheless, symmetry of the conditional distribution would be required to ensure the exact equivalence of the conditional mean and median restrictions, \( E_t y_{t+1} = 0 \) and \( \text{median}(y_{t+1} | I_t) = 0 \), as well as the equivalence between \( \text{median}(y_{t+1} | I_t) = 0 \), as in (18), and \( \text{median}(y_{t+k}^k | I_t) = 0 \).
the assumptions of the above remarks the signs, $a((y_{t+1} - b_0)x_{k^*}^t), t = k + 1, k + 2, \ldots, T$, form a sequence of independent Bernoulli random variables with probability of success 0.5, regardless of the dependence in $x_t^k$. This thus provides a novel practical application of the sign and signed rank tests, which, to our knowledge, has not been suggested in the previous literature. Moreover, it yields a long-horizon test whose robustness properties are unmatched by existing procedures.

4 Simulation Study

In this section we compare the simulation performance (both size and power) of several long-horizon predictive tests across a number of data generating processes that could potentially be employed to model the persistent behavior commonly observed in predictive regressors. In addition to standard regression tests, we consider the sup version of Valkanov’s re-scaled bounds test ($V_{sup}$) and the long-horizon implementations of the sign ($S_T^k$) and signed rank tests ($SR_T^k$) tests suggested in Section 3.3 above.

4.1 Simulation models

Because a variety of specifications may induce similar looking persistent behavior, we consider several different models for $x_t$ in assessing the performance of the long-horizon tests described below. The DGPs used in this paper are the AR(1) model (5), the long-memory/I(d) model (10), the historical break model (11), the modified STOP-BREAK model (14) and the Markov switching model (12). Table 1 summarizes the model specifications (columns 1-2) and parameter values (column 3) used to simulate $x_t$. Diebold and Inoue (2001) show that under certain parameterizations, the modified STOP-BREAK model and Markov switching model generates data that resembles a long-memory series. The degree of apparent fractional integration ($d$) in the break series is in both cases controlled by single parameter, which, to make the comparison explicit, they also denote by $d$. To facilitate interpretation, we adopt the same parameterization and notation in our simulations. Although, $d$ refers to different parameters in each model, it represents a similar degree of persistence in the sense of Diebold and Inoue (2001).\textsuperscript{13}

\textsuperscript{13}Since the standard Markov switching model generates an $I(0)$ process, the modified version tends to generate processes that appear similar to $I(d)$ processes. To generate highly persistent Markov switching series we incorporate an AR(1) process into the simple Markov switching model. Indeed, this Markov switching-AR(1) process can be treated as a stochastic historical break model.
Table 1: Summary of simulation DGPs and parameters used for \( x_t \)

<table>
<thead>
<tr>
<th>Model</th>
<th>DGP</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(1)</td>
<td>( x_t = \phi x_{t-1} + \varepsilon_{2,t} )</td>
<td>( \phi = {0.9, 0.95, 0.99} )</td>
</tr>
<tr>
<td>I(d)</td>
<td>( x_t = (1-L)^d \varepsilon_{2,t} = \sum_{j=0}^{\infty} \psi_j \varepsilon_{2,t-j} )</td>
<td>( d = {0.4, 0.7, 0.9} )</td>
</tr>
<tr>
<td>Historical</td>
<td>( x_t = \nu_t )</td>
<td></td>
</tr>
<tr>
<td>break</td>
<td>( x_t = \mu + \phi \nu_t )</td>
<td>( \phi = {0.5, 0.9, 0.99} )</td>
</tr>
<tr>
<td>STOP-BREAK</td>
<td>( x_t = \mu_t + \varepsilon_{2,t} )</td>
<td>( \gamma_T = 0.999T^{-0.3} )</td>
</tr>
<tr>
<td>Markov switching</td>
<td>( x_t = \mu_i + \phi \nu_{i-1} + \varepsilon_{2,t} )</td>
<td>( \theta = 2d )</td>
</tr>
<tr>
<td></td>
<td>( P(s_t = i</td>
<td>s_{t-1} = i) = 1 - 0.999T^{-2\theta} )</td>
</tr>
</tbody>
</table>

Columns 1 and 2 provide the model types and precise specifications for simulations. Column 3 lists the parameter values used in size comparisons in Tables 2 and 3. The residual process is specified by \( (\varepsilon_{1,t}, \varepsilon_{2,t})' \sim N(0, \Sigma), \Sigma = \begin{pmatrix} 1 & \delta \\ \delta & 1 \end{pmatrix} \). Since the finite sample behavior of the modified STOP-BREAK model and the modified Markov switching model will resemble I(d) processes, we also provide the corresponding values for the order of fractional integration, \( d \). \( 1 \) is the indicator function.

As in Valkanov (2003) and Rossi (2003), the short-horizon returns \( y_{t+1} \) are generated from (1), with \( \alpha = 0 \) and the long-horizon returns \( y_{t+k} \) are defined as in (3). The innovations are drawn from the joint normal distribution as

\[
(\varepsilon_{1,t}, \varepsilon_{2,t})' \sim \text{i.i.d.} N(0, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \delta \\ \delta & 1 \end{pmatrix}
\]

in which \( \delta \) measures the residual cross-correlation. In all cases, our simulation analysis is conducted with sample size 200 and 2000 replications.

### 4.2 Implementation of long-horizon tests

Our primary interest focuses on the long-horizon sign and signed rank tests \( S_T^k \) and \( SP_T^k \) defined in (22) and (23) respectively. Since \( b_0 \) is unknown, these tests are not immediately implementable. As discussed in Section 3.2, one may either employ the Campbell and Dufour (1997) Bonferroni bound to maintain exact inference or use the sample median \( \text{med}_T^m \) \( (y_t) \) as a plug-in estimator of \( b_0 \). Although the Bonferroni approach is more appealing from a theoretical perspective, simulations in the short-horizon autoregressive case suggest that this approach can be fairly conservative in practice, whereas the plug-in approach,
while not exact in finite sample, appears to yield good size (Campbell and Dufour, 1997, Tables 2-4). We therefore employ the plug-in version of the tests in the simulations discussed below.

We compare these tests to the long-horizon testing method introduced by Valkanov (2003). This test is based on the rescaled t-statistic \( T^{-1/2} t_\beta \), and its limiting distribution is deduced in Valkanov (2003) under the local-to-unity model \( (5-8) \). Critical values for this distribution are obtained by simulation following the same steps as described in Valkanov (2003). However, these depend on three nuisance parameters: \( \lambda, \delta, \) and \( c \). The first two have natural plug-in estimators, but the third cannot be consistently estimated. Valkanov (2003) suggests a sup-bound test in which the rescaled test-statistic is compared to a grid of critical values across a pre-specified range for \( c \). We adopt the same range, \( c \in [-10, 0] \), as Valkanov (2003). The overall test rejects only if a rejection is obtained for all \( c \) in this range. We refer to this method as \( V^{\sup} \).

To put the results in the proper perspective, we also make some comparisons to the standard long-horizon t-test in \( (2) \) employing Newey-West standard errors \( (t_\beta^{NW}) \) and to the standard t-test \( (t_\gamma) \) in the alternate long-horizon regression in \( (21) \).

### 4.3 Size analysis

In this section we report size comparisons for the long-horizon tests discussed above using a sample size of \( T = 200 \) and a horizon length of \( k = 24 \). Tables 2 and 3 present rejection rates under the null hypothesis \( (\beta(k) = 0) \) using a five percent nominal size for each of the models and parameters values listed in Table 1. In both tables, the value of the residual cross-correlation parameter \( \delta \), which controls the degree of endogeneity, is varied across the rows (see column 2), while the relevant persistence parameter is varied across the columns. For the autoregressive, historical break and Markov switching processes given in Table 2 the persistence of \( x_t \) is varied via the autoregressive parameter \( \phi \), while for the long-memory and STOP-BREAK models of Table 3 we vary the long-memory parameter \( d \).

---

14 We employ a normal approximation to the exact critical values. With \( T=200 \), the normal approximation and the exact distribution yield very similar critical values. This is not a large sample approximation but rather a convenient approximation to the finite sample distribution and thus the quality of the approximation depends on the sample size but not on the stochastic properties of the data.

15 They are based on 20000 replications using simulation sample sizes of 1000 observations.

16 A similar sup-bound approach was suggested in Cavanagh et al. (1995) for the short-horizon case.
As reported elsewhere, the standard long-horizon test with robust standard errors ($t_{NW}^β$) shown in the top panel of both tables suffers from substantial over-rejection. This reflects size distortion due to both the long-horizon returns and the persistence of the regressor, which is predetermined, but not exogenous. The problem can be quite severe, with rejection rates in the long-memory and STOP-BREAK models reaching above sixty-percent (see Table 3, panel 1).

Table 2: Size comparisons for AR(1), historical break and Markov switching models.

<table>
<thead>
<tr>
<th>Tests</th>
<th>AR(1)</th>
<th>Historical break</th>
<th>Markov switching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\phi/\delta$</td>
<td>$\delta/\phi$</td>
<td>$\phi/\delta$</td>
</tr>
<tr>
<td></td>
<td>$\beta/\phi$</td>
<td>$\beta/\phi$</td>
<td>$\beta/\phi$</td>
</tr>
<tr>
<td></td>
<td>$\beta/\phi$</td>
<td>$\beta/\phi$</td>
<td>$\beta/\phi$</td>
</tr>
<tr>
<td></td>
<td>$\beta/\phi$</td>
<td>$\beta/\phi$</td>
<td>$\beta/\phi$</td>
</tr>
<tr>
<td></td>
<td>$\beta/\phi$</td>
<td>$\beta/\phi$</td>
<td>$\beta/\phi$</td>
</tr>
<tr>
<td>$t_{NW}$</td>
<td>0.2435 0.2705 0.3070</td>
<td>0.2160 0.2505 0.2835</td>
<td>0.2240 0.1165 0.1115</td>
</tr>
<tr>
<td>$t_{NW}$</td>
<td>0.3280 0.3565 0.5390</td>
<td>0.2380 0.2900 0.4335</td>
<td>0.2220 0.2045 0.3705</td>
</tr>
<tr>
<td>$t_{NW}$</td>
<td>0.3085 0.3645 0.5160</td>
<td>0.2480 0.2800 0.4375</td>
<td>0.2050 0.2370 0.4045</td>
</tr>
<tr>
<td>$t_{\gamma}$</td>
<td>0.0575 0.0625 0.0580</td>
<td>0.0490 0.0535 0.0585</td>
<td>0.0590 0.0590 0.0605</td>
</tr>
<tr>
<td>$t_{\gamma}$</td>
<td>0.0725 0.0695 0.1600</td>
<td>0.0510 0.0655 0.1175</td>
<td>0.0550 0.0715 0.0955</td>
</tr>
<tr>
<td>$t_{\gamma}$</td>
<td>0.0705 0.0775 0.1400</td>
<td>0.0490 0.0595 0.1260</td>
<td>0.0540 0.0535 0.1010</td>
</tr>
<tr>
<td>$S_{kT}$</td>
<td>0.0385 0.0425 0.0350</td>
<td>0.0190 0.0250 0.0300</td>
<td>0.0415 0.0495 0.0495</td>
</tr>
<tr>
<td>$S_{kT}$</td>
<td>0.0420 0.0335 0.0345</td>
<td>0.0205 0.0200 0.0275</td>
<td>0.0380 0.0400 0.0300</td>
</tr>
<tr>
<td>$S_{kT}$</td>
<td>0.0415 0.0490 0.0445</td>
<td>0.0160 0.0300 0.0340</td>
<td>0.0405 0.0450 0.0510</td>
</tr>
<tr>
<td>$S_{kT}$</td>
<td>0.0405 0.0375 0.0320</td>
<td>0.0295 0.0235 0.0330</td>
<td>0.0420 0.0450 0.0450</td>
</tr>
<tr>
<td>$S_{kT}$</td>
<td>0.0450 0.0375 0.0415</td>
<td>0.0250 0.0285 0.0335</td>
<td>0.0420 0.0455 0.0340</td>
</tr>
<tr>
<td>$S_{kT}$</td>
<td>0.0505 0.0490 0.0500</td>
<td>0.0265 0.0320 0.0425</td>
<td>0.0430 0.0485 0.0470</td>
</tr>
<tr>
<td>$V_{sup}$</td>
<td>0.0075 0.0235 0.0535</td>
<td>0.0605 0.0620 0.0580</td>
<td>0.0115 0.0000 0.0000</td>
</tr>
<tr>
<td>$V_{sup}$</td>
<td>0.0285 0.0245 0.2025</td>
<td>0.0790 0.1000 0.3360</td>
<td>0.0295 0.0000 0.0085</td>
</tr>
<tr>
<td>$V_{sup}$</td>
<td>0.0250 0.0230 0.0200</td>
<td>0.0680 0.0955 0.0415</td>
<td>0.0225 0.0345 0.0130</td>
</tr>
</tbody>
</table>

The table entries show rejection rates under the null hypothesis for a nominal 5% test, where $x_t$ is generated respectively by the AR(1), historical break and Markov switching processes as described in Table 1 and $y_t$ is given by (1), with $\beta = 0$. Horizon length is set to $k = 24$. The test procedures are described in Section 4.

The $t_{\gamma}$ statistic based on the specification in (21), shown in the second panel, eliminates the size distortion due to long-horizon returns. In both tables, we notice a substantial drop in rejection rates when we switch from $t_{NW}^β$ (panel 1) to $t_{\gamma}$ (panel 2). However, it does not eliminate the size-distortion due to the persistence in $x_t$ and still over-rejects when this persistence is strong. Moreover, the MA$(k)$ structure in $x_t^k$ can aggravate the size problem. For example, for $k = 24$ the rejection rates can exceed twice the nominal size for stationary long-memory ($d = 0.4$, column 3, panel 2, in Table 3), whereas, in results not shown, size distortion is not present for $d = 0.4$ and $k = 1$. By contrast, the long-horizon sign ($S_{kT}^k$, third panel) and signed rank ($S_{R_k}^k$, fourth panel) tests, which are also based on the specification in (21), perform quite well in all of the models under consideration. The tests are slightly conservative in a number of cases, but never over-reject.\footnote{Due to the plug-in parameters for the median the test is not exact.}
Table 3: Size comparisons for long-memory and STOP-BREAK models

<table>
<thead>
<tr>
<th>Tests</th>
<th>$\delta/d$</th>
<th>(\text{Long-memory}^2)</th>
<th>(\text{STOP-BREAK}^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.4 0.7 0.9</td>
<td>0.4 0.7 0.9</td>
</tr>
<tr>
<td>$t_{NW}^\beta$</td>
<td>-0.3</td>
<td>0.2060 0.3055 0.2975</td>
<td>0.2400 0.3035 0.3235</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3850 0.5280 0.6030</td>
<td>0.5090 0.5460 0.6135</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3950 0.5255 0.6055</td>
<td>0.5140 0.5725 0.6010</td>
</tr>
<tr>
<td>$t_\gamma$</td>
<td>-0.9</td>
<td>0.1090 0.1625 0.1935</td>
<td>0.1850 0.1685 0.2055</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1095 0.1590 0.2020</td>
<td>0.1840 0.1870 0.1970</td>
</tr>
<tr>
<td>$S_k^\beta$</td>
<td>-0.3</td>
<td>0.0400 0.0210 0.0235</td>
<td>0.0270 0.0210 0.0205</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0370 0.0285 0.0215</td>
<td>0.0225 0.0210 0.0240</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0460 0.0455 0.0420</td>
<td>0.0375 0.0410 0.0345</td>
</tr>
<tr>
<td>$SR_k^J$</td>
<td>-0.9</td>
<td>0.0395 0.0265 0.0295</td>
<td>0.0330 0.0310 0.0275</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0450 0.0325 0.0290</td>
<td>0.0295 0.0245 0.0355</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0460 0.0475 0.0470</td>
<td>0.0340 0.0415 0.0465</td>
</tr>
<tr>
<td>$V_{sup}$</td>
<td>-0.9</td>
<td>0.0000 0.0145 0.0250</td>
<td>0.0020 0.0230 0.0460</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0005 0.0060 0.0235</td>
<td>0.0000 0.0105 0.0335</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0045 0.0055 0.0255</td>
<td>0.0015 0.0180 0.0320</td>
</tr>
</tbody>
</table>

The table entries show rejection rates under the null hypothesis for a nominal 5% test, where \(x_t\) is generated by the long-memory/fractionally integrated and STOP-BREAK processes respectively as described in Table 1 and \(y_t\) is given by (1), with \(\beta = 0\). Horizon length is set to \(k = 24\). The test procedures are described in Section 4.

Finally, the bottom panel of both tables show the rejection rates for the Valkanov (2003) \(V_{sup}\) test. Unlike the standard regression tests discussed above, this test has the correct asymptotic distribution in the presence of both local-to-unity persistence and long-horizon returns. On the other hand, its behavior under structural break and long-memory models has not been previously studied. Thus, in addition to providing a basis of comparison, the simulation results on \(V_{sup}\) may provide potentially useful information. Not surprisingly, the \(V_{sup}\) does not over-reject under the local-to-unity model considered in columns 3-5 of Table 2. This test, which is explicitly based on local-to-unity asymptotics, also shows considerable robustness to the alternate models for \(x_t\) shown in the remainder of Table 2 and 3. This is a very encouraging finding. Nevertheless, the size of the long-horizon sign and signed rank tests are generally more accurate. The \(V_{sup}\) tends to modestly over-reject in the presence of a historical structural break (see Table 2 panel 5, column 7), with rejection rates up to ten percent. For the other processes, it more often has a tendency to under-reject, particularly when \(x_t\) is less persistent. In a few cases, the test can become extremely conservative, for example, when \(x_t\)
follows a Markov switching process as shown in Table 2, Columns 9-11. This is primarily due to the conservative nature of the bounds procedure. However, because the local-to-unity parameter cannot be consistently estimated in general, it is not possible to replace this by a consistent plug-in parameter as in the case of the sign and signed rank tests.

4.4 Power comparison

Assigning non-zero values to $\beta$, we evaluate the power of each testing approach and their comparisons are presented in Figures 1-5. These figures show power functions for a right sided alternative ($\beta > 0$) under the autoregressive, long-memory, historical break, STOP-BREAK, and Markov switching models respectively. We fix $\delta$ at 0.9 and allow $\beta$ to vary. For each DGP, we consider a highly persistent (panel A) and a moderately persistent series (panel B). In each figure, three power functions are plotted. The two lines with circles and triangles represent the power functions of the sign test and signed rank tests respectively. The line with the solid diamonds corresponds to the power function of $V_{sup}$. The two regression tests, $t_{NW}^\beta$ and $t_\gamma$, are omitted on account of their poor size performance. We again set $k = 24$. In work excluded for space considerations, we also compared power at horizons $k = 1, 3$ and 12. The relative power comparison between nonparametric and $V_{sup}$ tests at these horizons were qualitatively similar. However, we also observed that the power of both tests was noticeably larger for smaller values of $k$.\footnote{Torous et al. (2005) also document an increase in power at shorter horizons for their size adjusted predictive test. Likewise, Kilian (1999) notes that the size-adjusted power of predictive tests does not generally increase with horizon length in linear models, although it may if the underlying model is non-linear (Kilian and Taylor, 2003). These issues may warrant further study and are the subject of our ongoing work.}

No uniformly most powerful test has been proposed in the context of long-horizon regression.\footnote{The Jansson and Moreira (2006) test is conditionally uniformly most powerful in short-horizon predictive regression but has yet to be adapted to long-horizon regression assumptions, in which the horizon length grows with sample size.} The $V_{sup}$ test may suffer power loss due to the conservative nature of the bounds procedure, whereas the nonparametric sign and signed rank tests may not use the full information in the data set. In practice, neither test clearly dominates. The $V_{sup}$ test tends to yield better power in more persistent cases, as evidenced by panel A of the figures. However, the order is reversed for slightly more moderate persistence levels, with the nonparametric tests often showing better power in panel B of the figures. For example in Figure 1, with an autoregressive parameter of $\phi = 0.99$, the $V_{sup}$ test has better power in panel A, while in panel B for $\phi = 0.95$ the power curves cross. The power comparison also varies with the specification of $x_t$.\footnote{The Jansson and Moreira (2006) test is conditionally uniformly most powerful in short-horizon predictive regression but has yet to be adapted to long-horizon regression assumptions, in which the horizon length grows with sample size.}
For example, the relative power of the $V_{sup}$ test is particularly high in both the deterministic historical break (Figure 3) and STOP-BREAK (Figure 4) models, while the nonparametric tests do relatively well in the Markov switching case (Figure 5).

Figure 1: The power functions when $x_t$ follows an autoregressive process with $k = 24$ and $\delta = 0.9$ (panel A: $\phi = 0.99$, panel B: $\phi = 0.95$).

5 Robust long horizon stock return predictability tests

A large volume of research has been devoted to the stock return predictability problem in the last two decades, but without yielding a clear cut conclusion. While the formidable forecasting powers of stock return predictors such as the dividend-price and earnings price ratio were documented using standard asymptotic approaches in the original research on this topic (Fama and French, 1988; Campbell and Shiller, 1988; Campbell et al., 1997, e.g.), the evidence of predictability disappeared in some recent studies using relatively more reliable statistical inference procedures (Torous et al., 2005; Valkanov, 2003, e.g.). However, see Lewellen (2004) for an alternate conclusion.
Figure 2: The power functions when $x_t$ follows a fractionally integrated process with $k = 24$ and $\delta = 0.9$ (panel A: $d=0.9$, panel B: $d=0.4$).

Figure 3: The power functions when $x_t$ follows a historical break process with $k = 24$, $\delta = 0.9$ and $\phi = 0.9$ (panel A: $\mu = 7$, panel B: $\mu = 3$).
Figure 4: The power functions when $x_t$ follows a STOP-BREAK process with $k = 24$, $\delta = 0.9$ (panel A: $d=0.9$, panel B: $d=0.4$).

Figure 5: The power functions when $x_t$ follows a Markov switching process with $k = 24$, $\delta = 0.9$, $\phi = 0.9$ and $p_{00} = p_{11} = 1 - 0.999T^{-0.6}$ (panel A: $\mu = 7$, panel B: $\mu = 3$).
rank tests to the stock return predictability problem. We first describe the data and then present the test results.

5.1 Data

We employ real stock returns as the dependent variable \((y_t)\) using monthly log returns of the CRSP value-weighted (VW) market portfolio, corrected for inflation using the CPI. We consider two commonly employed persistent predictors \((x_t)\), the (log) dividend-price ratio and the one-month treasury bill rate. Following common practice (Campbell et al., 1997, Chapter 7) the dividend-price ratio is defined as the sum of the dividends over the past twelve months divided by the current price.\(^{21}\) The short-term rate is given by the annualized one-month treasury bill rate, also obtained from the CRSP dataset. In addition, following Campbell et al. (1997), we also consider the stochastically detrended short-term interest rate as an alternative forecasting variable.\(^{22}\) The full sample period runs from January 1927 to December 2003. Following common practice, we also break this into two sub-periods, January 1927-December 1951 and January 1952-December 2003.

5.2 Persistence and endogeneity

In addition to the horizon length \(k\), the extent of the size distortion in standard long-horizon predictive tests is known to depend on both the degree of the persistence and the strength of the endogeneity or residual cross-correlation. In the autoregressive near-unit root model \((5-8)\) these are determined respectively by the size of the parameters \(\phi\) (or \(c\)) and \(\delta\). Even if we are not sure whether this is the correct model for \(x_t\) these parameters may still provide a rough indication of the potential for size distortion. Table \(4\) shows the estimated value of \(\delta\) (columns 2 & 7), the Augmented Dickey-Fuller (ADF) t-statistic (columns 3 & 8) and the Stock (1991) 95% confidence interval for the largest autoregressive root \((\phi)\) based on an inversion of the ADF test (columns 4-5 & 9-10). The lag-length \((q)\) shown in columns 6 and 11 is chosen by the Ng and Perron (2001) MIC criteria.\(^{23}\) Both the dividend-price ratio (columns 2-6) and the treasury bill (columns 7-11) are found to be highly persistent with confidence intervals on the largest root including one in every case. The

\(^{21}\)Denoting the dividend and stock price by \(D_t\) and \(P_t\) respectively, the log-dividend-price ratio is defined as \(ln \left( \sum_{j=0}^{11} D_{t-j} / P_t \right)\).

\(^{22}\)Denoting the one-month treasury bill as \(i_t\), the stochastically detrended treasury bill is defined as \(i_t - \frac{1}{12} \sum_{j=0}^{11} i_{t-j}\).

\(^{23}\)We use the one-period return in calculating \(\delta\).
estimated endogeneity parameter $\hat{\delta}$ is also large and negative for the dividend-price ratio. This may not be surprising since the calculation of both the stock returns and dividend-price ratios depends on the stock price. However, in conjunction with the strong persistence in $x_t$, it suggests serious size distortion even in short-horizon regression tests. By contrast, the treasury bill rate shows relatively modest residual correlation and tests using this predictor may not be subject to as strong size distortion, at least at short-horizons. Time series plots of the dividend-price ratio and treasury-bill are presented in Figure 6.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Dividend price ratio</th>
<th>One-month treasury bill rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\delta}$ $t_{ADF}$ $\phi$ $\phi$ $q$</td>
<td>$\hat{\delta}$ $t_{ADF}$ $\phi$ $\phi$ $q$</td>
</tr>
<tr>
<td>1927-2003</td>
<td>-0.9312 -1.9701 0.9845 1.0036 1</td>
<td>-0.1129 -1.9719 0.9845 1.0036 5</td>
</tr>
<tr>
<td>1927-1951</td>
<td>-0.9262 -2.5815 0.9274 1.0067 3</td>
<td>-0.0014 -2.2581 0.9413 1.0094 4</td>
</tr>
<tr>
<td>1952-2003</td>
<td>-0.8858 -1.3969 0.9864 1.0067 1</td>
<td>-0.2011 -2.1126 0.9745 1.0049 5</td>
</tr>
</tbody>
</table>

Columns 2 and 7 show the estimate of the residual correlation parameter $\delta$ in (7). Columns 3 and 8 show the Augmented Dickey-Fuller test statistic and Columns 4 and 9 and 5 and 10 respectively show the lower ($\phi$) and upper ($\bar{\phi}$) bounds for the (Stock, 1991) ninety-five percent confidence interval on the value of the largest root ($\psi$) in (5). The number of lags ($q$) is chosen using the Ng and Perron (2001) MIC criteria.

Figure 6: Time series plots of the predictors (panel A: Log dividend-price ratio, panel B: One-month Treasury Bill rate).
5.3 Test results

The test outcomes using the dividend-price ratio, one-month treasury bill rate and the stochastically detrended treasury bill rate as predictors are presented in Tables 5, 6 and 7 respectively. The top panel of each table shows the full sample results (1927-2003), while the next two panels show the results for two split samples, (1927-1951 and 1952-2003). The data is sampled monthly and we employ returns horizons of $k = 1, 3, 12, 24, 36,$ and $48$ months, as listed in the top row of each table. The tables show the results from all five tests employed in the simulations above. For purposes of comparison, we show $p$-values for both the standard long-horizon test $t_\gamma^{NW}$ and the $t$-statistic $t_\gamma$ in (21), despite their potential for size distortion. We also include $p$-values for the long-horizon sign ($S_T^\gamma$) and signed rank ($SR_T^\gamma$) tests and the $V_{sup}$ test. The implied sign of the estimated predictive relation is denoted by the $+/−$ in the superscript to the right of each $p$-value.

Table 5: Tests of stock return predictability using the log dividend-price ratio

<table>
<thead>
<tr>
<th>Sample year</th>
<th>Test</th>
<th>1</th>
<th>3</th>
<th>12</th>
<th>24</th>
<th>36</th>
<th>48</th>
</tr>
</thead>
<tbody>
<tr>
<td>1927</td>
<td>$t_\gamma^{NW}$</td>
<td>0.1753+</td>
<td>0.1651+</td>
<td>0.0519+</td>
<td>0.0232+</td>
<td>0.0209+</td>
<td>0.0356+</td>
</tr>
<tr>
<td>to</td>
<td>$t_\gamma$</td>
<td>0.1753+</td>
<td>0.1246+</td>
<td>0.0956+</td>
<td>0.1015+</td>
<td>0.1123+</td>
<td>0.1164+</td>
</tr>
<tr>
<td>2003</td>
<td>$S_T^\gamma$</td>
<td>0.8953−</td>
<td>0.8434−</td>
<td>0.7658−</td>
<td>0.8156−</td>
<td>0.7628−</td>
<td>0.7612−</td>
</tr>
<tr>
<td></td>
<td>$SR_T^\gamma$</td>
<td>0.6210+</td>
<td>0.5282+</td>
<td>0.5792+</td>
<td>0.5023+</td>
<td>0.3951+</td>
<td>0.1323+</td>
</tr>
<tr>
<td></td>
<td>$V_{sup}$</td>
<td>0.9893+</td>
<td>0.6261+</td>
<td>0.6247+</td>
<td>0.6225+</td>
<td>0.7511+</td>
<td>0.9368+</td>
</tr>
<tr>
<td>1927</td>
<td>$t_\gamma^{NW}$</td>
<td>0.3098+</td>
<td>0.3217+</td>
<td>0.0546+</td>
<td>0.0001+</td>
<td>0.0014+</td>
<td>0.0000+</td>
</tr>
<tr>
<td>to</td>
<td>$t_\gamma$</td>
<td>0.3098+</td>
<td>0.1870+</td>
<td>0.1037+</td>
<td>0.0209+</td>
<td>0.0153+</td>
<td>0.0092+</td>
</tr>
<tr>
<td>1951</td>
<td>$S_T^\gamma$</td>
<td>0.5637−</td>
<td>0.4870−</td>
<td>0.2167−</td>
<td>0.4347−</td>
<td>0.4992−</td>
<td>0.6599−</td>
</tr>
<tr>
<td></td>
<td>$SR_T^\gamma$</td>
<td>0.4203−</td>
<td>0.5067−</td>
<td>0.1869−</td>
<td>0.7374−</td>
<td>0.8699−</td>
<td>0.5831−</td>
</tr>
<tr>
<td></td>
<td>$V_{sup}$</td>
<td>0.8225+</td>
<td>0.9630+</td>
<td>0.9863+</td>
<td>0.7594+</td>
<td>0.6863+</td>
<td>0.6325+</td>
</tr>
<tr>
<td>1952</td>
<td>$t_\gamma^{NW}$</td>
<td>0.0877+</td>
<td>0.0452+</td>
<td>0.0226+</td>
<td>0.0381+</td>
<td>0.0448+</td>
<td>0.0689+</td>
</tr>
<tr>
<td>to</td>
<td>$t_\gamma$</td>
<td>0.0877+</td>
<td>0.0669+</td>
<td>0.0488+</td>
<td>0.0570+</td>
<td>0.1650+</td>
<td>0.3014+</td>
</tr>
<tr>
<td>2003</td>
<td>$S_T^\gamma$</td>
<td>0.9362−</td>
<td>0.9361−</td>
<td>0.4428−</td>
<td>0.9675−</td>
<td>0.8368−</td>
<td>0.8351−</td>
</tr>
<tr>
<td></td>
<td>$SR_T^\gamma$</td>
<td>0.4984−</td>
<td>0.5291−</td>
<td>0.1459−</td>
<td>0.2510−</td>
<td>0.1966−</td>
<td>0.2101−</td>
</tr>
<tr>
<td></td>
<td>$V_{sup}$</td>
<td>0.6521+</td>
<td>0.2636+</td>
<td>0.3471+</td>
<td>0.4291+</td>
<td>0.6105+</td>
<td>0.9943+</td>
</tr>
</tbody>
</table>

The table entries provide two-sided $p$-values for the standard long-horizon regression ($t_\gamma^{NW}$), the regression test based on (21) ($t_\gamma$), the long-horizon sign ($S_T^\gamma$) and signed rank ($SR_T^\gamma$) tests and the Valkanov (2003)’s long horizon sup-bound test ($V_{sup}$). The implied sign of the estimated predictive relation is denoted by the $+/−$ next to each entry. The column headings give the horizon length ($k$) measured in months. Here $y_t$ denotes the log real stock return and $x_t$ is the annually adjusted log dividend-price ratio.

---

The $p$-value of $V_{sup}$ is obtained by selecting the largest two-sided empirical $p$-value (defined as twice the one-sided $p$-value) among the simulated distributions under all possible values of $c$. Thus, it is conservative in general. Somewhat smaller $p$-values are obtained by Valkanov (2003) when imposing the Gordon Growth Model.
Table 6: Tests of stock return predictability using the one month treasury bill rate

<table>
<thead>
<tr>
<th>Sample</th>
<th>Test</th>
<th>1</th>
<th>3</th>
<th>12</th>
<th>24</th>
<th>36</th>
<th>48</th>
</tr>
</thead>
<tbody>
<tr>
<td>1927</td>
<td>$t_{\beta}^{NW}$</td>
<td>0.4031</td>
<td>0.3511</td>
<td>0.6739</td>
<td>0.8573</td>
<td>0.7370</td>
<td>0.8407</td>
</tr>
<tr>
<td>to $t_{\gamma}$</td>
<td>0.4031</td>
<td>0.3953</td>
<td>0.6693</td>
<td>0.9002</td>
<td>0.8284</td>
<td>0.8521</td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>$S_{\beta}^{\phi}$</td>
<td>0.4298</td>
<td>0.2628</td>
<td>0.2748</td>
<td>0.8156</td>
<td>0.9732</td>
<td>0.7612</td>
</tr>
<tr>
<td></td>
<td>$SR_{\beta}^{\phi}$</td>
<td>0.9349</td>
<td>0.9052</td>
<td>0.8212</td>
<td>0.5554</td>
<td>0.4440</td>
<td>0.2439</td>
</tr>
<tr>
<td></td>
<td>$V_{sup}$</td>
<td>0.3625</td>
<td>0.3011</td>
<td>0.6535</td>
<td>0.8529</td>
<td>0.7817</td>
<td>0.8413</td>
</tr>
</tbody>
</table>

The table entries provide two-sided p-values for the standard long-horizon regression ($t_{\beta}^{NW}$), the regression test based on (21) ($t_{\gamma}$), the long-horizon sign ($S_{\beta}^{\phi}$) and signed rank ($SR_{\beta}^{\phi}$) tests and the Valkanov (2003)’s long horizon sup-bound test ($V_{sup}$). The implied sign of the estimated predictive relation is denoted by the +/− next to each entry. The column headings give the horizon length (k) measured in months. Here $y_t$ denotes the log real stock return and $x_t$ is the one-month treasury bill rate.

Results using the dividend-price ratio are shown in Table 5. The standard regression test $t_{\beta}^{NW}$ (top row of each panel) with robust standard errors shows strong evidence of predictability, particularly at long-horizons. The t-statistic $t_{\gamma}$ (second row) from the rearranged regression shows somewhat more modest evidence of predictability. However, both tests likely suffer substantial size distortion, especially in light of the large values of $\delta$ and $\phi$ shown in Table 4. This distortion can be particularly severe at the longer horizons for which the evidence of predictability appears strongest (Valkanov, 2003). By contrast, the robust sign and signed rank tests, whose p-values are shown in the next two rows, do not suffer from size distortion. Nor, do they show any evidence of return predictability using the dividend-price ratio. Moreover, the $V_{sup}$ test, which essentially corrects the critical values for $t_{\beta}^{NW}$, also fails to reject at the five percent level in all cases.
We next turn to the predictability tests employing the treasury bill as predictor in Tables 6 and 7. The results depend somewhat on whether the interest rate is specified in levels (Table 6) or is stochastically detrended (Table 7). Using the level, none of the five tests shows evidence of predictability during the full sample. However, some of the robust tests, particularly the signed rank test, show evidence of predictability in the later sub-sample (bottom panel). This evidence seems particularly strong at the shorter horizons. These tests also offer marginal evidence of predictability (but of the opposite sign) at short-horizons in the 1927-1951 period. It is interesting to note that the evidence of predictability in Table 6 is considerably stronger using the signed rank test than employing the traditional regression tests.

Using the stochastically detrended treasury bill rate in Table 7, both the robust tests (particularly the signed rank test) and the traditional regression tests show evidence of return predictability at short-horizons in both the full and later sub-samples, but not the earlier sub-sample. The stochastic detrending procedure reduces the persistence in the treasury bill, suggesting that the standard tests may in this case be more reliable when the horizon length is

<table>
<thead>
<tr>
<th>Sample</th>
<th>Test</th>
<th>1</th>
<th>3</th>
<th>12</th>
<th>24</th>
<th>36</th>
<th>48</th>
</tr>
</thead>
<tbody>
<tr>
<td>1927</td>
<td>$t_{\beta}^+$</td>
<td>0.1699</td>
<td>0.0311</td>
<td>0.1224</td>
<td>0.8578</td>
<td>0.2616</td>
<td>0.5696</td>
</tr>
<tr>
<td>to $t_s$</td>
<td>0.1699</td>
<td>0.0480</td>
<td>0.0593</td>
<td>0.8079</td>
<td>0.5733</td>
<td>0.6428</td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>$S_{T_k}^+$</td>
<td>0.1671</td>
<td>0.0096</td>
<td>0.6670</td>
<td>0.1938</td>
<td>0.7122</td>
<td>0.8131</td>
</tr>
<tr>
<td>$SR_{T_k}^+$</td>
<td>0.0697</td>
<td>0.0079</td>
<td>0.0293</td>
<td>0.7314</td>
<td>0.4373</td>
<td>0.6641</td>
<td></td>
</tr>
<tr>
<td>$V_{sup}$</td>
<td>0.1484</td>
<td>0.0400</td>
<td>0.2157</td>
<td>0.8775</td>
<td>0.7203</td>
<td>0.8389</td>
<td></td>
</tr>
</tbody>
</table>

The table entries provide two-sided p-values for the standard long-horizon regression ($t_{\beta}^{NW}$), the regression test based on ($t_s$), the long-horizon sign ($S_{T_k}^+$) and signed rank ($SR_{T_k}^+$) tests and the Valkanov (2003)’s long horizon sup-bound test ($V_{sup}$). The implied sign of the estimated predictive relation is denoted by the $+/-$ next to each entry. The column headings give the horizon length (k) measured in months. Here $y_t$ denotes the log real stock return and $x_t$ is the stochastically detrended one-month treasury bill rate (see footnote 22).
small. Thus perhaps it is not surprising that the tests are more closely in agreement.

In summary, even before turning to the robust tests, the evidence of return predictability using the dividend-price ratio is called into question by the high levels of persistence and endogeneity shown in Table 4, which suggest severe size distortion, especially in conjunction with the use of long-horizon returns. Using the robust sign and signed rank tests suggested here we find substantially less evidence of predictability than in traditional, but possibly misleading regression tests. The traditional regression evidence of predictability using interest rates is not as striking as it is using the dividend-price ratio. Nonetheless, this evidence appears much more robust. In fact, the robust tests, particularly the signed rank test, often showed stronger evidence of predictability than the original regression tests. This clearly confirms the existing evidence in favor of return predictability using short-term treasury bills.

6 Conclusion

This paper has suggested a simple way to provide robust finite sample inference for long-horizon predictive tests that brings together the sign and signed rank tests of Campbell and Dufour (1995, 1997) with an alternative specification of the predictive regression considered earlier in the finance literature (Jegadeesh, 1991; Cochrane, 1991). While both steps are straightforward given the existing literature, it is only by combining the two that we arrive at a simple robust approach to long-horizon predictive testing. The standard sign and signed rank tests are not valid in a long-horizon context under the original formulation, since long-horizons induce residual serial correlation. On the other hand, the alternative specification, which removes serial correlation in the residual at the expense of more persistent regressors, does not succeed in curing the size distortion often found in traditional predictive regression tests. It works much better when applied to the sign and signed rank tests, which are distorted by residual correlation, but insensitive to persistent regressors.

Our simulation results suggest that the test compares well to other existing long-horizon tests. Applying these tests to the long-horizon stock return predictability problem in finance, we confirm the robustness of the short-term interest rate as a stock return predictor, but find little evidence of predictability using the dividend-price ratio.
References


Atkins, F. and C. Rakoz (2005), Stochastic permanent breaks in the nominal interest rate and the inflation rate, working paper, University of Calgary.


Clemente, J., A. Montañés and M. Reyes (2003), Structural breaks, inflation and interest rate: Evidence for the G7 countries, working paper, University of Zaragoza.


Coudin, E. and J.-M. Dufour (2003), Finite distribution-free inference in linear median regressions under heteroskedasticity and nonlinear dependence of unknown form, mimeo, CREST and Université de Montréal.


Hwang, S. (2000), Structural breaks, persistence, and long memory, mimeo, Cass Business School, City University.


Luger, R. (2006), Exact split-sample permutation tests of orthogonality and random walk in the presence of a drift parameter, working paper, Department of Economics, Emory University.


Maynard, A. and K. Shimotsu (2004), Covariance-based orthogonality tests for regressors with unknown persistence, working paper, Queen’s University.


Rossi, B. (2003), Expectation hypotheses tests at long horizons, working paper, Department of Economics, Duke University.


Sakoulis, G. and E. Zivot (2002), Time-variation and structural change in the forward discount: Implications for the forward rate unbiasedness hypothesis, working paper, Department of Economics, University of Washington.

Smith, A. (2002), Why regime switching creates the illusion of long memory, working paper, Department of Agricultural and Resource Economics, University of California, Davis.


