Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Financial Economics 81 (2006) 27–60
JOURNAL OF Financial ECONOMICS
www.elsevier.com/locate/jfec

Efficient tests of stock return predictability ☆

John Y. Campbell[a], Motohiro Yogo[b,*]

[a]*Department of Economics, Harvard University, Cambridge, MA 02138, USA*
[b]*Finance Department, The Wharton School, University of Pennsylvania, 3620 Locust Walk, Philadelphia, PA 19104, USA*

Received 27 January 2004; received in revised form 30 March 2005; accepted 18 May 2005
Available online 18 January 2006

Abstract

Conventional tests of the predictability of stock returns could be invalid, that is reject the null too frequently, when the predictor variable is persistent and its innovations are highly correlated with returns. We develop a pretest to determine whether the conventional *t*-test leads to invalid inference and an efficient test of predictability that corrects this problem. Although the conventional *t*-test is invalid for the dividend–price and smoothed earnings–price ratios, our test finds evidence for predictability. We also find evidence for predictability with the short rate and the long-short yield spread, for which the conventional *t*-test leads to valid inference.

JEL classification: C12; C22; G1

Keywords: Bonferroni test; Dividend yield; Predictability; Stock returns; Unit root

1. Introduction

Numerous studies in the last two decades have asked whether stock returns can be predicted by financial variables such as the dividend–price ratio, the earnings–price ratio,

☆We have benefited from comments and discussions by Andrew Ang, Geert Bekaert, Jean-Marie Dufour, Markku Lanne, Marcelo Moreira, Robert Shiller, Robert Stambaugh, James Stock, Mark Watson, the referees, and seminar participants at Harvard, the Econometric Society Australasian Meeting 2002, the NBER Summer Institute 2003, and the CIRANO-CIREQ Financial Econometrics Conference 2004.
*Corresponding author.
E-mail address: yogo@wharton.upenn.edu (M. Yogo).

0304-405X/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.jfineco.2005.05.008

and various measures of the interest rate.[1] The econometric method used in a typical study is an ordinary least squares (OLS) regression of stock returns onto the lag of the financial variable. The main finding of such regressions is that the t-statistic is typically greater than two and sometimes greater than three. Using conventional critical values for the t-test, one would conclude that there is strong evidence for the predictability of returns.

This statistical inference of course relies on first-order asymptotic distribution theory, where the autoregressive root of the predictor variable is modeled as a fixed constant less than one. First-order asymptotics implies that the t-statistic is approximately standard normal in large samples. However, both simulation and analytical studies have shown that the large-sample theory provides a poor approximation to the actual finite-sample distribution of test statistics when the predictor variable is persistent and its innovations are highly correlated with returns (Elliott and Stock, 1994; Mankiw and Shapiro, 1986; Stambaugh, 1999).

To be concrete, suppose the log dividend–price ratio is used to predict returns. Even if we were to know on prior grounds that the dividend–price ratio is stationary, a time-series plot (more formally, a unit root test) shows that it is highly persistent, much like a nonstationary process. Since first-order asymptotics fails when the regressor is nonstationary, it provides a poor approximation in finite samples when the regressor is persistent. Elliott and Stock (1994, Table 1) provide Monte Carlo evidence which suggests that the size distortion of the one-sided t-test is approximately 20 percentage points for plausible parameter values and sample sizes in the dividend–price ratio regression.[2] They propose an alternative asymptotic framework in which the regressor is modeled as having a local-to-unit root, an autoregressive root that is within $1/T$-neighborhood of one where T denotes the sample size. Local-to-unity asymptotics provides an accurate approximation to the finite-sample distribution of test statistics when the predictor variable is persistent.

These econometric problems have led some recent papers to reexamine (and even cast serious doubt on) the evidence for predictability using tests that are valid even if the predictor variable is highly persistent or contains a unit root. Torous et al. (2004) develop a test procedure, extending the work of Richardson and Stock (1989) and Cavanagh et al. (1995), and find evidence for predictability at short horizons but not at long horizons. By testing the stationarity of long-horizon returns, Lanne (2002) concludes that stock returns cannot be predicted by a highly persistent predictor variable. Building on the finite-sample theory of Stambaugh (1999), Lewellen (2004) finds some evidence for predictability with valuation ratios.

A difficulty with understanding the rather large literature on predictability is the sheer variety of test procedures that have been proposed, which have led to different conclusions about the predictability of returns. The first contribution of this paper is to provide an understanding of the various test procedures and their empirical implications within the unifying framework of statistical optimality theory. When the degree of persistence of the predictor variable is known, there is a uniformly most powerful (UMP) test conditional on

[1]See, for example, Campbell (1987), Campbell and Shiller (1988), Fama and French (1988, 1989), Fama and Schwert (1977), Hodrick (1992), and Keim and Stambaugh (1986). The focus of these papers, as well as this one, is classical hypothesis testing. Other approaches include out-of-sample forecasting (Goyal and Welch, 2003) and Bayesian inference (Kothari and Shanken, 1997; Stambaugh, 1999).

[2]We report their result for the one-sided t-test at the 10% level when the sample size is 100, the regressor follows an AR(1) with an autoregressive coefficient of 0.975, and the correlation between the innovations to the dependent variable and the regressor is -0.9.

an ancillary statistic. Although the degree of persistence is not known in practice, this provides a useful benchmark for thinking about the relative power advantages of the various test procedures. In particular, Lewellen's (2004) test is UMP when the predictor variable contains a unit root.

Our second contribution is to propose a new Bonferroni test, based on the infeasible UMP test, that has three desirable properties for empirical work. First, the test can be implemented with standard regression methods, and inference can be made through an intuitive graphical output. Second, the test is asymptotically valid under fairly general assumptions on the dynamics of the predictor variable (i.e., a finite-order autoregression with the largest root less than, equal to, or even greater than one) and on the distribution of the innovations (i.e., even heteroskedastic). Finally, the test is more efficient than previously proposed tests in the sense of Pitman efficiency (i.e., requires fewer observations for inference at the same level of power); in particular, it is more powerful than the Bonferroni t-test of Cavanagh et al. (1995).

The intuition for our approach, similar to that underlying the work by Lewellen (2004) and Torous et al. (2004), is as follows. A regression of stock returns onto a lagged financial variable has low power because stock returns are extremely noisy. If we can eliminate some of this noise, we can increase the power of the test. When the innovations to returns and the predictor variable are correlated, we can subtract off the part of the innovation to the predictor variable that is correlated with returns to obtain a less noisy dependent variable for our regression. Of course, this procedure requires us to measure the innovation to the predictor variable. When the predictor variable is highly persistent, it is possible to do so in a way that retains power advantages over the conventional regression.

Although tests derived under local-to-unity asymptotics, such as Cavanagh et al. (1995) or the one proposed in this paper, lead to valid inference, they can be somewhat more difficult to implement than the conventional t-test. A researcher might therefore be interested in knowing when the conventional t-test leads to valid inference. Our third contribution is to develop a simple pretest based on the confidence interval for the largest autoregressive root of the predictor variable. If the confidence interval indicates that the predictor variable is sufficiently stationary, for a given level of correlation between the innovations to returns and the predictor variable, one can proceed with inference based on the t-test with conventional critical values.

Our final contribution is empirical. We apply our methods to annual, quarterly, and monthly U.S. data, looking first at dividend–price and smoothed earnings–price ratios. Using the pretest, we find that these valuation ratios are sufficiently persistent for the conventional t-test to be misleading (Stambaugh, 1999). Using our test that is robust to the persistence problem, we find that the earnings–price ratio reliably predicts returns at all frequencies in the sample period 1926–2002. The dividend–price ratio also predicts returns at annual frequency, but we cannot reject the null hypothesis at quarterly and monthly frequencies.

In the post-1952 sample, we find that the dividend–price ratio predicts returns at all frequencies if its largest autoregressive root is less than or equal to one. However, since statistical tests do not reject an explosive root for the dividend–price ratio, we have evidence for return predictability only if we are willing to rule out an explosive root based on prior knowledge. This reconciles the "contradictory" findings by Torous et al. (2004, Table 3), who report that the dividend–price ratio does not predict monthly returns in the postwar sample, and Lewellen (2004, Table 2), who reports strong evidence for predictability.

Finally, we consider the short-term nominal interest rate and the long-short yield spread as predictor variables in the sample period 1952–2002. Our pretest indicates that the conventional t-test is valid for these interest rate variables since their innovations have low correlation with returns (Torous et al., 2004). Using either the conventional t-test or our more generally valid test procedure, we find strong evidence that these variables predict returns.

The rest of the paper is organized as follows. In Section 2, we review the predictive regressions model and discuss the UMP test of predictability when the degree of persistence of the predictor variable is known. In Section 3, we review local-to-unity asymptotics in the context of predictive regressions, then introduce the pretest for determining when the conventional t-test leads to valid inference. We also compare the asymptotic power and finite-sample size of various tests of predictability. We find that our Bonferroni test based on the UMP test has good power. In Section 4, we apply our test procedure to U.S. equity data and reexamine the empirical evidence for predictability. We reinterpret previous empirical studies within our unifying framework. Section 5 concludes. A separate note (Campbell and Yogo, 2005), available from the authors' webpages, provides self-contained user guides and tables necessary for implementing the econometric methods in this paper.

2. Predictive regressions

2.1. The regression model

Let r_t denote the excess stock return in period t, and let x_{t-1} denote a variable observed at $t-1$ which could have the ability to predict r_t. For instance, x_{t-1} could be the log dividend–price ratio at $t-1$. The regression model that we consider is

$$r_t = \alpha + \beta x_{t-1} + u_t, \tag{1}$$

$$x_t = \gamma + \rho x_{t-1} + e_t, \tag{2}$$

with observations $t = 1, \ldots, T$. The parameter β is the unknown coefficient of interest. We say that the variable x_{t-1} has the ability to predict returns if $\beta \neq 0$. The parameter ρ is the unknown degree of persistence in the variable x_t. If $|\rho| < 1$ and fixed, x_t is integrated of order zero, denoted as I(0). If $\rho = 1$, x_t is integrated of order one, denoted as I(1).

We assume that the innovations are independently and identically distributed (i.i.d.) normal with a known covariance matrix.

Assumption 1 (*Normality*). $w_t = (u_t, e_t)'$ is independently distributed $\mathbf{N}(0, \Sigma)$, where

$$\Sigma = \begin{bmatrix} \sigma_u^2 & \sigma_{ue} \\ \sigma_{ue} & \sigma_e^2 \end{bmatrix}$$

is known. x_0 is fixed and known.

This is a simplifying assumption that we maintain throughout the paper in order to facilitate discussion and to focus on the essence of the problem. It can be relaxed to more realistic distributional assumptions as demonstrated in Appendix A. We also assume that the correlation between the innovations, $\delta = \sigma_{ue}/(\sigma_u \sigma_e)$, is negative. This assumption is without loss of generality since the sign of β is unrestricted; redefining the predictor variable as $-x_t$ flips the signs of both β and δ.

The joint log likelihood for the regression model is given by

$$L(\beta,\rho,\alpha,\gamma) = -\frac{1}{1-\delta^2}\sum_{t=1}^{T}\left[\frac{(r_t-\alpha-\beta x_{t-1})^2}{\sigma_u^2} - 2\delta\frac{(r_t-\alpha-\beta x_{t-1})(x_t-\gamma-\rho x_{t-1})}{\sigma_u\sigma_e} + \frac{(x_t-\gamma-\rho x_{t-1})^2}{\sigma_e^2}\right], \tag{3}$$

up to a multiplicative constant of $1/2$ and an additive constant. The focus of this paper is the null hypothesis $\beta=\beta_0$. We consider two types of alternative hypotheses. The first is the *simple alternative* $\beta=\beta_1$, and the second is the one-sided *composite alternative* $\beta>\beta_0$. The hypothesis testing problem is complicated by the fact that ρ is an unknown nuisance parameter.

2.2. The t-test

One way to test the hypothesis of interest in the presence of the nuisance parameter ρ is through the maximum likelihood ratio test (LRT). Let $x_{t-1}^{\mu}=x_{t-1}-T^{-1}\sum_{t=1}^{T}x_{t-1}$ be the de-meaned predictor variable. Let $\widehat{\beta}$ be the OLS estimator of β, and let

$$t(\beta_0)=\frac{\widehat{\beta}-\beta_0}{\sigma_u(\sum_{t=1}^{T}x_{t-1}^{\mu 2})^{-1/2}} \tag{4}$$

be the associated t-statistic. The LRT rejects the null if

$$\max_{\beta,\rho,\alpha,\gamma} L(\beta,\rho,\alpha,\gamma) - \max_{\rho,\alpha,\gamma} L(\beta_0,\rho,\alpha,\gamma) = t(\beta_0)^2 > C, \tag{5}$$

for some constant C. (With a slight abuse of notation, we use C to denote a generic constant throughout the paper.) In other words, the LRT corresponds to the t-test.

Note that we would obtain the same test (5) starting from the marginal likelihood $L(\beta,\alpha)=-\sum_{t=1}^{T}(r_t-\alpha-\beta x_{t-1})^2$. The LRT can thus be interpreted as a test that ignores information contained in Eq. (2) of the regression model. Although the LRT is not derived from statistical optimality theory, it has desirable large-sample properties when x_t is I(0) (Cox and Hinkley, 1974, Chapter 9). For instance, the t-statistic is asymptotically pivotal, that is, its asymptotic distribution does not depend on the nuisance parameter ρ. The t-test is therefore a solution to the hypothesis testing problem when x_t is I(0) and ρ is unknown, provided that the large-sample approximation is sufficiently accurate.

2.3. The optimal test when ρ is known

To simplify the discussion, assume for the moment that $\alpha=\gamma=0$. Now suppose that ρ were known a priori. Since β is then the only unknown parameter, we denote the likelihood function (3) as $L(\beta)$. The Neyman–Pearson Lemma implies that the most powerful test against the simple alternative $\beta=\beta_1$ rejects the null if

$$\sigma_u^2(1-\delta^2)(L(\beta_1)-L(\beta_0)) = 2(\beta_1-\beta_0)\sum_{t=1}^{T}x_{t-1}[r_t-\beta_{ue}(x_t-\rho x_{t-1})] - (\beta_1^2-\beta_0^2)\sum_{t=1}^{T}x_{t-1}^2 > C, \tag{6}$$

where $\beta_{ue} = \sigma_{ue}/\sigma_e^2$. Since the optimal test statistic is a weighted sum of two minimal sufficient statistics with the weights depending on the alternative β_1, there is no UMP test.

However, the second statistic $\sum_{t=1}^{T} x_{t-1}^2$ is ancillary, that is, its distribution does not depend on β. Hence, it is natural to restrict ourselves to tests that condition on the ancillary statistic. Since the second term in Eq. (6) can then be treated as a "constant," the optimal *conditional* test rejects the null if

$$\sum_{t=1}^{T} x_{t-1}[r_t - \beta_{ue}(x_t - \rho x_{t-1})] > C, \tag{7}$$

for any alternative $\beta_1 > \beta_0$. Therefore, the optimal conditional test is UMP against one-sided alternatives when ρ is known. It is convenient to recenter and rescale test statistic (7) so that it has a standard normal distribution under the null. The UMP test can then be expressed as

$$\frac{\sum_{t=1}^{T} x_{t-1}[r_t - \beta_0 x_{t-1} - \beta_{ue}(x_t - \rho x_{t-1})]}{\sigma_u(1-\delta^2)^{1/2}(\sum_{t=1}^{T} x_{t-1}^2)^{1/2}} > C. \tag{8}$$

Note that this inequality is reversed for left-sided alternatives $\beta_1 < \beta_0$.

Now suppose that ρ is known, but α and γ are unknown nuisance parameters. Then within the class of *invariant tests*, the test based on the statistic

$$Q(\beta_0, \rho) = \frac{\sum_{t=1}^{T} x_{t-1}^{\mu}[r_t - \beta_0 x_{t-1} - \beta_{ue}(x_t - \rho x_{t-1})]}{\sigma_u(1-\delta^2)^{1/2}(\sum_{t=1}^{T} x_{t-1}^{\mu 2})^{1/2}} \tag{9}$$

is UMP conditional on the ancillary statistic $\sum_{t=1}^{T} x_{t-1}^{\mu 2}$. For simplicity, we refer to this statistic as the Q-statistic, and the (infeasible) test based on this statistic as the Q-test. Note that the only change from statistic (8) to (9) is that x_{t-1} has been replaced by its de-meaned counterpart x_{t-1}^{μ}.

The class of invariant tests refers to those tests whose test statistics do not change with additive shifts in r_t and x_t (see Lehmann 1986, Chapter 6). Or equivalently, the value of the test statistic is the same regardless of the values of α and γ. (The reader can verify that the value of the Q-statistic does not depend on α and γ.) The reason to restrict attention to invariant tests is that the magnitudes of α and γ depend on the units in which the variables are measured. For instance, there is an arbitrary scaling factor involved in computing the dividend–price ratio, which results in an arbitrary constant shifting the level of the log dividend–price ratio. Since we do not want inference to depend on the units in which the variables are measured, it is natural to restrict attention to invariant tests.

When $\beta_0 = 0$, $Q(\beta_0, \rho)$ is the t-statistic that results from regressing $r_t - \beta_{ue}(x_t - \rho x_{t-1})$ onto a constant and x_{t-1}. It collapses to the conventional t-statistic (4) when $\delta = 0$. Since $e_t + \gamma = x_t - \rho x_{t-1}$, knowledge of ρ allows us to subtract off the part of innovation to returns that is correlated with the innovation to the predictor variable, resulting in a more powerful test. If we let $\widehat{\rho}$ denote the OLS estimator of ρ, then the Q-statistic can also be written as

$$Q(\beta_0, \rho) = \frac{(\widehat{\beta} - \beta_0) - \beta_{ue}(\widehat{\rho} - \rho)}{\sigma_u(1-\delta^2)^{1/2}(\sum_{t=1}^{T} x_{t-1}^{\mu 2})^{-1/2}}. \tag{10}$$

Drawing on the work of Stambaugh (1999), Lewellen (2004) motivates the statistic by interpreting the term $\beta_{ue}(\widehat{\rho}-\rho)$ as the "finite-sample bias" of the OLS estimator. Assuming that $\rho \leqslant 1$, Lewellen tests the predictability of returns using the statistic $Q(\beta_0, 1)$.

3. Inference with a persistent regressor

Fig. 1 is a time-series plot of the log dividend–price ratio for the NYSE/AMEX value-weighted index and the log smoothed earnings–price ratio for the S&P 500 index at quarterly frequency. Following Campbell and Shiller (1988), earnings are smoothed by taking a backwards moving average over ten years. Both valuation ratios are persistent and even appear to be nonstationary, especially toward the end of the sample period. The 95% confidence intervals for ρ are $[0.957, 1.007]$ and $[0.939, 1.000]$ for the dividend–price ratio and the earnings–price ratio, respectively (see Panel A of Table 4).

The persistence of financial variables typically used to predict returns has important implications for inference about predictability. Even if the predictor variable is I(0), first-order asymptotics can be a poor approximation in finite samples when ρ is close to one because of the discontinuity in the asymptotic distribution at $\rho = 1$ (note that $\sigma_x^2 = \sigma_e^2/(1-\rho^2)$ diverges to infinity at $\rho = 1$). Inference based on first-order asymptotics could therefore be invalid due to size distortions. The solution is to base inference on more accurate approximations to the actual (unknown) sampling distribution of test statistics. There are two main approaches that have been used in the literature.

The first approach is the exact finite-sample theory under the assumption of normality (i.e., Assumption 1). This is the approach taken by Evans and Savin (1981, 1984) for

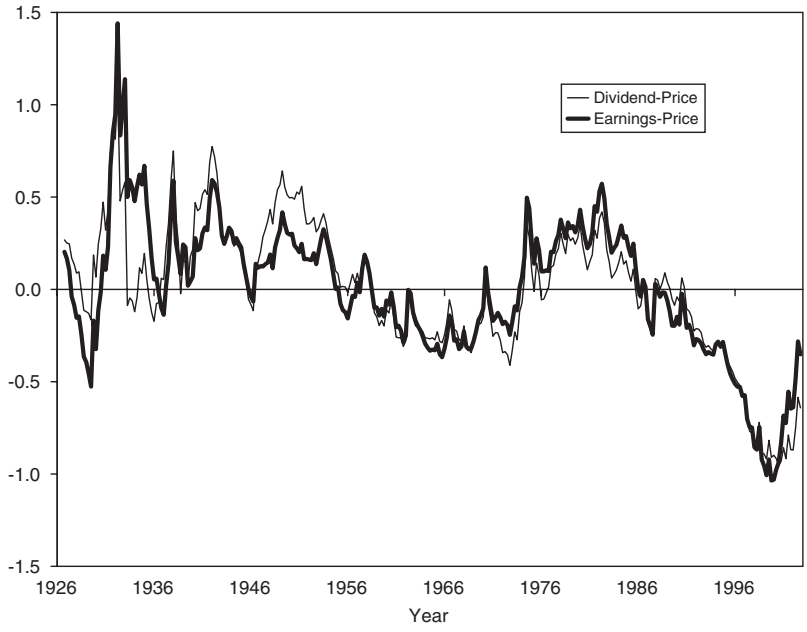

Fig. 1. Time-series plot of the valuation ratios. This figure plots the log dividend–price ratio for the CRSP value-weighted index and the log earnings–price ratio for the S&P 500. Earnings are smoothed by taking a 10-year moving average. The sample period is 1926:4–2002:4.

autoregression and Stambaugh (1999) for predictive regressions. The second approach is local-to-unity asymptotics, which has been applied successfully to approximate the finite-sample behavior of persistent time series in the unit root testing literature; see Stock (1994) for a survey and references. Local-to-unity asymptotics has been applied to the present context of predictive regressions by Elliott and Stock (1994), who derive the asymptotic distribution of the t-statistic. This has been extended to long-horizon t-tests by Torous et al. (2004).

This paper uses local-to-unity asymptotics. For our purposes, there are two practical advantages to local-to-unity asymptotics over the exact Gaussian theory. The first advantage is that the asymptotic distribution of test statistics does not depend on the sample size, so the critical values of the relevant test statistics do not have to be recomputed for each sample size. (Of course, we want to check that the large-sample approximations are accurate, which we do in Section 3.6.) The second advantage is that the asymptotic theory provides large-sample justification for our methods in empirically realistic settings that allow for short-run dynamics in the predictor variable and heteroskedasticity in the innovations.

Although local-to-unity asymptotics allows us to considerably relax the distributional assumptions, we continue to work in the text of the paper with the simple model (1) and (2) under the assumption of normality (i.e., Assumption 1) to keep the discussion simple. Appendix A works out the more general case when the predictor variable is a finite-order autoregression and the innovations are a martingale difference sequence with finite fourth moments.

3.1. Local-to-unity asymptotics

Local-to-unity asymptotics is an asymptotic framework where the largest autoregressive root is modeled as $\rho = 1 + c/T$ with c a fixed constant. Within this framework, the asymptotic distribution theory is not discontinuous when x_t is I(1) (i.e., $c = 0$). This device also allows x_t to be stationary but nearly integrated (i.e., $c<0$) or even explosive (i.e., $c>0$). For the rest of the paper, we assume that the true process for the predictor variable is given by Eq. (2), where $c = T(\rho - 1)$ is fixed as T becomes arbitrarily large.

An important feature of the nearly integrated case is that sample moments (e.g., mean and variance) of the process x_t do not converge to a constant probability limit. However, when appropriately scaled, these objects converge to functionals of a diffusion process. Let $(W_u(s), W_e(s))'$ be a two-dimensional Weiner process with correlation δ. Let $J_c(s)$ be the diffusion process defined by the stochastic differential equation $\mathrm{d}J_c(s) = cJ_c(s)\,\mathrm{d}s + \mathrm{d}W_e(s)$ with initial condition $J_c(0) = 0$. Let $J_c^{\mu}(s) = J_c(s) - \int J_c(r)\,\mathrm{d}r$, where integration is over $[0, 1]$ unless otherwise noted. Let $\Rightarrow$ denote weak convergence in the space $D[0, 1]$ of cadlag functions (see Billingsley, 1999, Chapter 3).

Under first-order asymptotics, the t-statistic (4) is asymptotically normal. Under local-to-unity asymptotics, the t-statistic has the null distribution

$$t(\beta_0) \Rightarrow \delta\frac{\tau_c}{\kappa_c} + (1 - \delta^2)^{1/2}Z, \tag{11}$$

where $\kappa_c = (\int J_c^{\mu}(s)^2\,\mathrm{d}s)^{1/2}$, $\tau_c = \int J_c^{\mu}(s)\,\mathrm{d}W_e(s)$, and Z is a standard normal random variable independent of $(W_e(s), J_c(s))$ (see Elliott and Stock, 1994). Note that the t-statistic is not asymptotically pivotal. That is, its asymptotic distribution depends on an

unknown nuisance parameter c through the random variable τ_c/κ_c, which makes the test infeasible.

The Q-statistic (9) is normal under the null. However, this test is also infeasible since it requires knowledge of ρ (or equivalently c) to compute the test statistic. Even if ρ were known, the statistic (9) also requires knowledge of the nuisance parameters in the covariance matrix Σ. However, a feasible version of the statistic that replaces the nuisance parameters in Σ with consistent estimators has the same asymptotic distribution. Therefore, there is no loss of generality in assuming knowledge of these parameters for the purposes of asymptotic theory.

3.2. Relation to first-order asymptotics and a simple pretest

In this section, we first discuss the relation between first-order and local-to-unity asymptotics. We then develop a simple pretest to determine whether inference based on first-order asymptotics is reliable.

In general, the asymptotic distribution of the t-statistic (11) is nonstandard because of its dependence on τ_c/κ_c. However, the t-statistic is standard normal in the special case $\delta = 0$. The t-statistic should therefore be approximately normal when $\delta \approx 0$. Likewise, the t-statistic should be approximately normal when $c \ll 0$ because first-order asymptotics is a satisfactory approximation when the predictor variable is stationary. Formally, Phillips (1987, Theorem 2) shows that $\tau_c/\kappa_c \Rightarrow \widetilde{Z}$ as $c \to -\infty$, where $\widetilde{Z}$ is a standard normal random variable independent of Z.

Fig. 2 is a plot of the asymptotic size of the nominal 5% one-sided t-test as a function of c and δ. More precisely, we plot

$$p(c, \delta; 0.05) = \Pr\left(\delta \frac{\tau_c}{\kappa_c} + (1 - \delta^2)^{1/2} Z > z_{0.05} \right), \tag{12}$$

where $z_{0.05} = 1.645$ denotes the 95th percentile of the standard normal distribution. The t-test that uses conventional critical values has approximately the correct size when δ is small in absolute value or c is large in absolute value.[3] The size distortion of the t-test peaks when $\delta = -1$ and $c \approx 1$. The size distortion arises from the fact that the distribution of τ_c/κ_c is skewed to the left, which causes the distribution of the t-statistic to be skewed to the right when $\delta < 0$. This causes a right-tailed t-test that uses conventional critical values to over-reject, and a left-tailed test to under-reject. When the predictor variable is a valuation ratio (e.g., the dividend–price ratio), $\delta \approx -1$ and the hypothesis of interest is $\beta = 0$ against the alternative $\beta > 0$. Thus, we might worry that the evidence for predictability is a consequence of size distortion.

In Table 1, we tabulate the values of $c \in (c_{\min}, c_{\max})$ for which the size of the right-tailed t-test exceeds 7.5%, for selected values of δ. For instance, when $\delta = -0.95$, the nominal 5% t-test has asymptotic size greater than 7.5% if $c \in (-79.318, 8.326)$. The table can be used to construct a pretest to determine whether inference based on the conventional t-test is sufficiently reliable.

Suppose a researcher is willing to tolerate an actual size of up to 7.5% for a nominal 5% test of predictability. To test the null hypothesis that the actual size exceeds 7.5%, we first

[3]The fact that the t-statistic is approximately normal for $c \gg 0$ corresponds to asymptotic results for explosive AR(1) with Gaussian errors. See Phillips (1987) for a discussion.

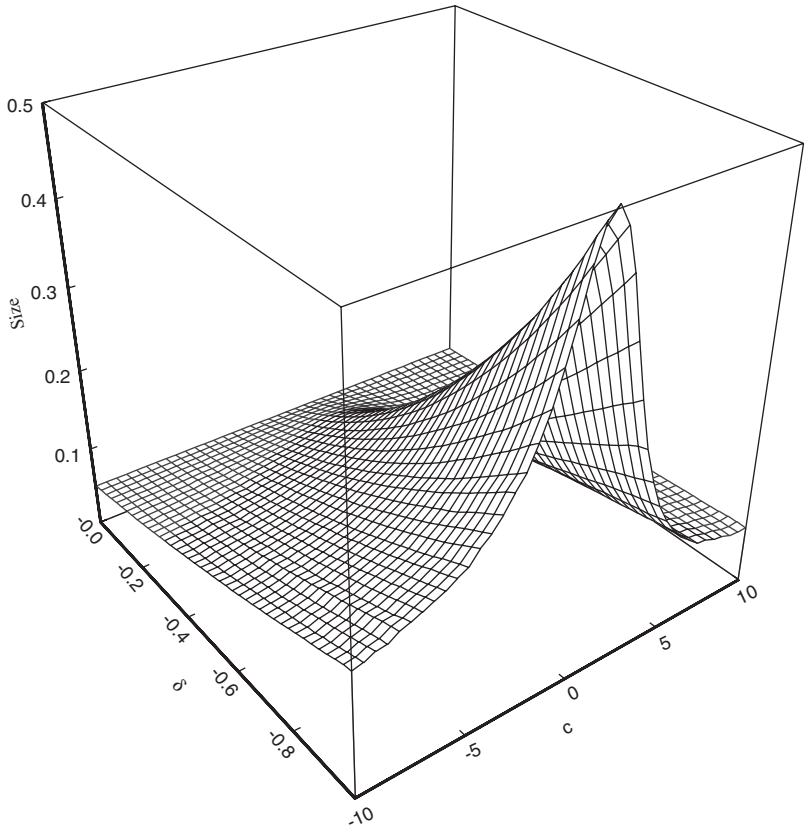

Fig. 2. Asymptotic size of the one-sided t-test at 5% significance. This figure plots the actual size of the nominal 5% t-test when the largest autoregressive root of the predictor variable is $\rho = 1 + c/T$. The null hypothesis is $\beta = \beta_0$ against the one-sided alternative $\beta > \beta_0$. δ is the correlation between the innovations to returns and the predictor variable. The dark shade indicates regions where the size is greater than 7.5%.

construct a $100(1-\alpha_1)\%$ confidence interval for c and estimate δ using the residuals from regressions (1) and (2).[4] We reject the null if the confidence interval for c lies strictly below (or above) the region of the parameter space $(c_{\min}, c_{\max})$ where size distortion is large. The relevant region $(c_{\min}, c_{\max})$ is determined by Table 1, using the value of δ that is closest to the estimated correlation $\widehat{\delta}$. As emphasized by Elliott and Stock (1994), the rejection of the unit root hypothesis $c = 0$ is not sufficient to assure that the size distortion is acceptably small. Asymptotically, this pretest has size α_1.

In our empirical application, we construct the confidence interval for c by applying the method of confidence belts as suggested by Stock (1991). The basic idea is to compute a unit root test statistic in the data and to use the known distribution of that statistic under the alternative to construct the confidence interval for c. A relatively accurate confidence interval can be constructed by using a relatively powerful unit root test (Elliott and Stock, 2001). We therefore use the Dickey–Fuller generalized least squares (DF-GLS) test of

[4]When the predictor variable is generalized to an AR(p), the residual is that of regression (23) in Appendix A.

Table 1
Parameters leading to size distortion of the one-sided t-test

δ	$c_{\min}$	$c_{\max}$	δ	$c_{\min}$	$c_{\max}$
−1.000	−83.088	8.537	−0.550	−28.527	6.301
−0.975	−81.259	8.516	−0.525	−27.255	6.175
−0.950	−79.318	8.326	−0.500	−25.942	6.028
−0.925	−76.404	8.173	−0.475	−23.013	5.868
−0.900	−69.788	7.977	−0.450	−19.515	5.646
−0.875	−68.460	7.930	−0.425	−17.701	5.435
−0.850	−63.277	7.856	−0.400	−14.809	5.277
−0.825	−59.563	7.766	−0.375	−13.436	5.111
−0.800	−58.806	7.683	−0.350	−11.884	4.898
−0.775	−57.618	7.585	−0.325	−10.457	4.682
−0.750	−51.399	7.514	−0.300	−8.630	4.412
−0.725	−50.764	7.406	−0.275	−6.824	4.184
−0.700	−42.267	7.131	−0.250	−5.395	3.934
−0.675	−41.515	6.929	−0.225	−4.431	3.656
−0.650	−40.720	6.820	−0.200	−3.248	3.306
−0.625	−36.148	6.697	−0.175	−1.952	2.800
−0.600	−33.899	6.557	−0.150	−0.614	2.136
−0.575	−31.478	6.419	−0.125	—	—

This table reports the regions of the parameter space where the actual size of the nominal 5% t-test is greater than 7.5%. The null hypothesis is $\beta = \beta_0$ against the alternative $\beta > \beta_0$. For a given δ, the size of the t-test is greater than 7.5% if $c \in (c_{\min}, c_{\max})$. Size is less than 7.5% for all c if $\delta \leqslant -0.125$.

Elliott et al. (1996), which is more powerful than the commonly used augmented Dickey–Fuller (ADF) test. The idea behind the DF-GLS test is that it exploits the knowledge $\rho \approx 1$ to obtain a more efficient estimate of the intercept γ.[5] We refer to Campbell and Yogo (2005) for a detailed description of how to construct the confidence interval for c using the DF-GLS statistic.

3.3. Making tests feasible by the Bonferroni method

As discussed in Section 3.1, both the t-test and the Q-test are infeasible since the procedures depend on an unknown nuisance parameter c, which cannot be estimated consistently. Intuitively, the degree of persistence, controlled by the parameter c, influences the distribution of test statistics that depend on the persistent predictor variable. This must be accounted for by adjusting either the critical values of the test (e.g., t-test) or the value of the test statistic itself (e.g., Q-test). Cavanagh et al. (1995) discuss several (sup-bound, Bonferroni, and Scheffe-type) methods of making tests that depend on c feasible.[6] Here, we focus on the Bonferroni method.

[5] A note of caution regarding the DF-GLS confidence interval is that the procedure might not be valid when $\rho \ll 1$ (since it is based on the assumption that $\rho \approx 1$). In practical terms, this method should not be used on variables that would not ordinarily be tested for an autoregressive unit root.

[6] These are standard parametric approaches to the problem. For a nonparametric approach, see Campbell and Dufour (1991, 1995).

To construct a Bonferroni confidence interval, we first construct a $100(1-\alpha_1)\%$ confidence interval for ρ, denoted as $C_\rho(\alpha_1)$. (We parameterize the degree of persistence by ρ rather than c since this is the more natural choice in the following.) For each value of ρ in the confidence interval, we then construct a $100(1-\alpha_2)\%$ confidence interval for β given ρ, denoted as $C_{\beta|\rho}(\alpha_2)$. A confidence interval that does not depend on ρ can be obtained by

$$C_\beta(\alpha) = \bigcup_{\rho \in C_\rho(\alpha_1)} C_{\beta|\rho}(\alpha_2). \tag{13}$$

By Bonferroni's inequality, this confidence interval has coverage of at least $100(1-\alpha)\%$, where $\alpha = \alpha_1 + \alpha_2$.

In principle, one can use any unit root test in the Bonferroni procedure to construct the confidence interval for ρ. Based on work in the unit root literature, reasonable choices are the ADF test and the DF-GLS test. The DF-GLS test has the advantage of being more powerful than the ADF test, resulting in a tighter confidence interval for ρ.

In the Bonferroni procedure, one can also use either the t-test or the Q-test to construct the confidence interval for β given ρ. We know that the Q-test is a more powerful test than the t-test when ρ is known. In fact, it is UMP conditional on an ancillary statistic in that situation. This means that the conditional confidence interval $C_{\beta|\rho}(\alpha_2)$ based on the Q-test is tighter than that based on the t-test at the true value of ρ. Without numerical analysis, however, it is not clear whether the Q-test retains its power advantages over the t-test at other values of ρ in the confidence interval $C_\rho(\alpha_1)$.

In practice, the choice of the particular tests in the Bonferroni procedure should be dictated by the issue of power. Cavanagh et al. (1995) propose a Bonferroni procedure based on the ADF test and the t-test. Torous et al. (2004) have applied this procedure to test for predictability in U.S. data. In this paper, we examine a Bonferroni procedure based on the DF-GLS test and the Q-test. While there is no rigorous justification for our choice, our Bonferroni procedure turns out to have better power properties, which we show in Section 3.5.

Because the Q-statistic is normally distributed, and the estimate of β declines linearly in ρ when δ is negative, the confidence interval for our Bonferroni Q-test is easy to compute. The Bonferroni confidence interval for β runs from the lower bound of the confidence interval for β, conditional on ρ equal to the upper bound of its confidence interval, to the upper bound of the confidence interval for β, conditional on ρ equal to the lower bound of its confidence interval. More formally, an equal-tailed α_2-level confidence interval for β given ρ is simply $C_{\beta|\rho}(\alpha_2) = [\underline{\beta}(\rho,\alpha_2), \overline{\beta}(\rho,\alpha_2)]$, where

$$\beta(\rho) = \frac{\sum_{t=1}^{T} x_{t-1}^{\mu}[r_t - \beta_{ue}(x_t - \rho x_{t-1})]}{\sum_{t=1}^{T} x_{t-1}^{\mu 2}}, \tag{14}$$

$$\underline{\beta}(\rho,\alpha_2) = \beta(\rho) - z_{\alpha_2/2}\sigma_u \left(\frac{1-\delta^2}{\sum_{t=1}^{T} x_{t-1}^{\mu 2}}\right)^{1/2}, \tag{15}$$

$$\overline{\beta}(\rho,\alpha_2) = \beta(\rho) + z_{\alpha_2/2}\sigma_u \left(\frac{1-\delta^2}{\sum_{t=1}^{T} x_{t-1}^{\mu 2}}\right)^{1/2}, \tag{16}$$

and $z_{\alpha_2/2}$ denotes the $1-\alpha_2/2$ quantile of the standard normal distribution. Let $C_\rho(\alpha_1) = [\underline{\rho}(\underline{\alpha}_1), \overline{\rho}(\overline{\alpha}_1)]$ denote the confidence interval for ρ, where $\underline{\alpha}_1 = \Pr(\rho < \underline{\rho}(\underline{\alpha}_1))$,

$\overline{\alpha}_1 = \Pr(\rho > \overline{\rho}(\overline{\alpha}_1))$, and $\alpha_1 = \underline{\alpha}_1 + \overline{\alpha}_1$. Then the Bonferroni confidence interval is given by

$$C_\beta(\alpha) = [\underline{\beta}(\overline{\rho}(\overline{\alpha}_1), \alpha_2), \overline{\beta}(\underline{\rho}(\underline{\alpha}_1), \alpha_2)]. \tag{17}$$

In Campbell and Yogo (2005), we lay out the step-by-step recipe for implementing this confidence interval in the empirically relevant case when the nuisance parameters (i.e., σ_u, δ, and β_{ue}) are not known.

3.4. A refinement of the Bonferroni method

The Bonferroni confidence interval can be conservative in the sense that the actual coverage rate of $C_\beta(\alpha)$ can be greater than $100(1-\alpha)\%$. This can be seen from the equality

$$\begin{aligned}\Pr(\beta \notin C_\beta(\alpha)) = {} & \Pr(\beta \notin C_\beta(\alpha) | \rho \in C_\rho(\alpha_1)) \Pr(\rho \in C_\rho(\alpha_1)) \\ & + \Pr(\beta \notin C_\beta(\alpha) | \rho \notin C_\rho(\alpha_1)) \Pr(\rho \notin C_\rho(\alpha_1)).\end{aligned}$$

Since $\Pr(\beta \notin C_\beta(\alpha) | \rho \notin C_\rho(\alpha_1))$ is unknown, the Bonferroni method bounds it by one as the worst case. In addition, the inequality $\Pr(\beta \notin C_\beta(\alpha) | \rho \in C_\rho(\alpha_1)) \leqslant \alpha_2$ is strict unless the conditional confidence intervals $C_{\beta|\rho}(\alpha_2)$ do not depend on ρ. Because these worst case conditions are unlikely to hold in practice, the inequality

$$\Pr(\beta \notin C_\beta(\alpha)) \leqslant \alpha_2(1-\alpha_1) + \alpha_1 \leqslant \alpha$$

is likely to be strict, resulting in a conservative confidence interval.

Cavanagh et al. (1995) therefore suggest a refinement of the Bonferroni method that makes it less conservative than the basic approach. The idea is to shrink the confidence interval for ρ so that the refined interval is a subset of the original (unrefined) interval. This consequently shrinks the Bonferroni confidence interval for β, achieving an exact test of the desired significance level. Call this significance level $\widetilde{\alpha}$, which we must now distinguish from $\alpha = \alpha_1 + \alpha_2$, the sum of the significance levels used for the confidence interval for ρ (denoted α_1) and the conditional confidence intervals for β (denoted α_2).

To construct a test with significance level $\widetilde{\alpha}$, we first fix α_2. Then, for each δ, we numerically search to find the $\overline{\alpha}_1$ such that

$$\Pr(\underline{\beta}(\overline{\rho}(\overline{\alpha}_1), \alpha_2) > \beta) \leqslant \widetilde{\alpha}/2 \tag{18}$$

holds for all values of c on a grid, with equality at some point on the grid. We then repeat the same procedure to find the $\underline{\alpha}_1$ such that

$$\Pr(\overline{\beta}(\underline{\rho}(\underline{\alpha}_1), \alpha_2) < \beta) \leqslant \widetilde{\alpha}/2. \tag{19}$$

We use these values $\overline{\alpha}_1$ and $\underline{\alpha}_1$ to construct a tighter confidence interval for ρ. The resulting one-sided Bonferroni test has exact size $\widetilde{\alpha}/2$ for some permissible value of c. The resulting two-sided test has size at most $\widetilde{\alpha}$ for all values of c.

In Table 2, we report the values of $\underline{\alpha}_1$ and $\overline{\alpha}_1$ for selected values of δ when $\widetilde{\alpha} = \alpha_2 = 0.10$, computed over the grid $c \in [-50, 5]$. The table can be used to construct a 10% Bonferroni confidence interval for β (equivalently, a 5% one-sided Q-test for predictability). Note that $\underline{\alpha}_1$ and $\overline{\alpha}_1$ are increasing in δ, so the Bonferroni inequality has more slack and the unrefined Bonferroni test is more conservative the smaller is δ in absolute value. In order to implement the Bonferroni test using Table 2, one needs the confidence belts for the DF-GLS statistic. Campbell and Yogo (2005, Tables 2–11) provide lookup tables that report the appropriate confidence interval for c, $C_c(\alpha_1) = [\underline{c}(\underline{\alpha}_1), \overline{c}(\overline{\alpha}_1)]$, given the values of the

Table 2
Significance level of the DF-GLS confidence interval for the Bonferroni Q-test

δ	$\underline{\alpha}_1$	$\overline{\alpha}_1$	δ	$\underline{\alpha}_1$	$\overline{\alpha}_1$
−0.999	0.050	0.055	−0.500	0.080	0.280
−0.975	0.055	0.080	−0.475	0.085	0.285
−0.950	0.055	0.100	−0.450	0.085	0.295
−0.925	0.055	0.115	−0.425	0.090	0.310
−0.900	0.060	0.130	−0.400	0.090	0.320
−0.875	0.060	0.140	−0.375	0.095	0.330
−0.850	0.060	0.150	−0.350	0.100	0.345
−0.825	0.060	0.160	−0.325	0.100	0.355
−0.800	0.065	0.170	−0.300	0.105	0.360
−0.775	0.065	0.180	−0.275	0.110	0.370
−0.750	0.065	0.190	−0.250	0.115	0.375
−0.725	0.065	0.195	−0.225	0.125	0.380
−0.700	0.070	0.205	−0.200	0.130	0.390
−0.675	0.070	0.215	−0.175	0.140	0.395
−0.650	0.070	0.225	−0.150	0.150	0.400
−0.625	0.075	0.230	−0.125	0.160	0.405
−0.600	0.075	0.240	−0.100	0.175	0.415
−0.575	0.075	0.250	−0.075	0.190	0.420
−0.550	0.080	0.260	−0.050	0.215	0.425
−0.525	0.080	0.270	−0.025	0.250	0.435

This table reports the significance level of the confidence interval for the largest autoregressive root ρ, computed by inverting the DF-GLS test, which sets the size of the one-sided Bonferroni Q-test to 5%. Using the notation in Eq. (17), the confidence interval $C_\rho(\alpha_1) = [\underline{\rho}(\underline{\alpha}_1), \overline{\rho}(\overline{\alpha}_1)]$ for ρ results in a 90% Bonferroni confidence interval $C_\beta(0.1)$ for β when $\alpha_2 = 0.1$.

DF-GLS statistic and δ. The confidence interval $C_\rho(\alpha_1) = 1 + C_c(\alpha_1)/T$ for ρ then results in a 10% Bonferroni confidence interval for β.

Our computational results indicate that in general the inequalities (18) and (19) are close to equalities when $c \approx 0$ and have more slack when $c \ll 0$. For right-tailed tests, the probability (18) can be as small as 4.0% for some values of c and δ. For left-tailed tests, the probability (19) can be as small as 1.2%. This suggests that even the adjusted Bonferroni Q-test is conservative (i.e., undersized) when $c < 5$. The assumption that the predictor variable is never explosive (i.e., $c \leqslant 0$) would allow us to further tighten the Bonferroni confidence interval. In our judgment, however, the magnitude of the resulting power gain is not sufficient to justify the loss of robustness against explosive roots. (The empirical relevance of allowing for explosive roots is discussed in Section 4.)

3.5. Power under local-to-unity asymptotics

Any reasonable test, such as the Bonferroni t-test, rejects alternatives that are a fixed distance from the null with probability one as the sample size becomes arbitrarily large. In practice, however, we have a finite sample and are interested in the relative efficiency of test procedures. A natural way to evaluate the power of tests in finite samples is to consider their ability to reject local alternatives.[7] When the predictor variable contains a

[7]See Lehmann (1999, Chapter 3) for a textbook treatment of local alternatives and relative efficiency.

local-to-unit root, OLS estimators $\widehat{\beta}$ and $\widehat{\rho}$ are consistent at the rate T (rather than the usual $\sqrt{T}$). We therefore consider a sequence of alternatives of the form $\beta = \beta_0 + b/T$ for some fixed constant b. The empirically relevant region of b for the dividend–price ratio, based on OLS estimates of β, appears to be the interval $[8, 10]$, depending on frequency of the data (annual to monthly). Details on the computation of the power functions are in Appendix B.

3.5.1. Power of infeasible tests

We first examine the power of the t-test and Q-test under local-to-unity asymptotics. Although these tests assume knowledge of c and are thus infeasible, their power functions provide benchmarks for assessing the power of feasible tests.

Fig. 3 plots the power functions for the t-test (using the appropriate critical value that depends on c) and the Q-test. Under local-to-unity asymptotics, power functions are not symmetric in b. We only report the power for right-tailed tests (i.e., $b>0$) since this is the region where the conventional t-test is size distorted (recall the discussion in Section 3.2). The results, however, are qualitatively similar for left-tailed tests (available from the authors on request). We consider various combinations of c (-2 and -20) and δ (-0.95 and -0.75), which are in the relevant region of the parameter space when the predictor variable is a valuation ratio (see Table 4). The variances are normalized as $\sigma_u^2 = \sigma_e^2 = 1$.

As expected, the power function for the Q-test dominates that for the t-test. In fact, the power function for the Q-test corresponds to the Gaussian power envelope for conditional tests when ρ is known. In other words, the Q-test has the maximum achievable power when ρ is known and Assumption 1 holds. The difference is especially large when $\delta = -0.95$. When the correlation between the innovations is large, there are large power gains from subtracting the part of the innovation to returns that is correlated with the innovation to the predictor variable.

To assess the importance of the power gain, we compute the Pitman efficiency, which is the ratio of the sample sizes at which two tests achieve the same level of power (e.g., 50%) along a sequence of local alternatives. Consider the case $c = -2$ and $\delta = -0.95$. To compute the Pitman efficiency of the t-test relative to the Q-test, note first that the t-test achieves 50% power when $b = 4.8$. On the other hand, the Q-test achieves 50% power when $b = 1.8$. Following the discussion in Stock (1994; p. 2775), the Pitman efficiency of the t-test relative to the Q-test is $4.8/1.8 \approx 2.7$. This means that to achieve 50% power, the t-test asymptotically requires 170% more observations than the Q-test.

3.5.2. Power of feasible tests

We now analyze the power properties of several feasible tests that have been proposed. Fig. 3 reports the power of the Bonferroni t-test (Cavanagh et al., 1995) and the Bonferroni Q-test.[8]

In all cases considered, the Bonferroni Q-test dominates the Bonferroni t-test. In fact, the power of the Bonferroni Q-test comes very close to that of the infeasible t-test. The power gains of the Bonferroni Q-test over the Bonferroni t-test are larger the closer is c to zero and the larger is δ in absolute value. When $c = -2$ and $\delta = -0.95$, the Pitman

[8]The refinement procedure described in Section 3.4 for the Bonferroni Q-test with DF-GLS is also applied to the Bonferroni t-test with ADF. The significance levels $\overline{\alpha}_1$ and $\underline{\alpha}_1$ used in constructing the ADF confidence interval for ρ are chosen to result in a 5% one-sided test for β, uniformly in $c \in [-50, 5]$.

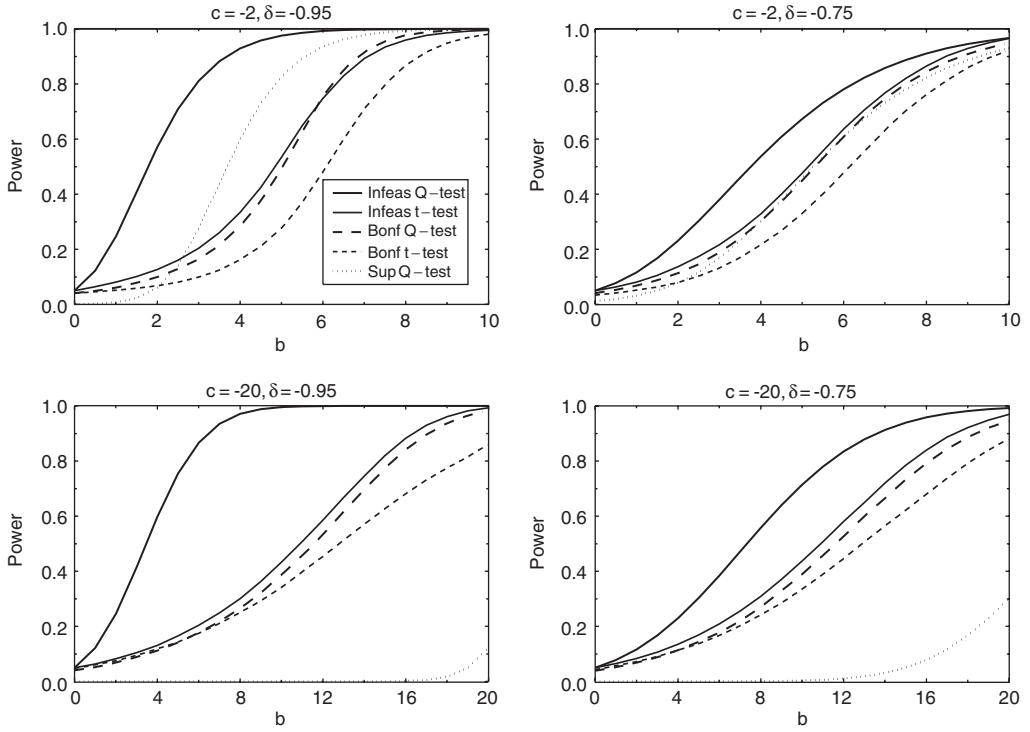

Fig. 3. Local asymptotic power of the Q-test and the t-test. This figure plots the power of the infeasible Q-test and t-test that assume knowledge of the local-to-unity parameter, the Bonferroni Q-test and t-test, and the sup-bound Q-test. The null hypothesis is $\beta = \beta_0$ against the local alternatives $b = T(\beta - \beta_0) > 0$. $c = \{-2, -20\}$ is the local-to-unity parameter, and $\delta = \{-0.95, -0.75\}$ is the correlation between the innovations to returns and the predictor variable.

efficiency is 1.24, which means that the Bonferroni t-test requires 24% more observations than the Bonferroni Q-test to achieve 50% power.

In addition to the Bonferroni tests, we also consider the power of Lewellen's (2004) test. In our notation (17), Lewellen's confidence interval corresponds to $[\underline{\beta}(1, \alpha_2), \overline{\beta}(1, \alpha_2)]$. Formally, this test can be interpreted as a sup-bound Q-test, that is, the $\overline{Q}$-test that sets ρ equal to the value that maximizes size. The value $\rho = 1$ maximizes size, provided that the parameter space is restricted to $\rho \leqslant 1$, since $Q(\beta_0, \rho)$ is decreasing in ρ when $\delta < 0$. By construction, the sup-bound Q-test is the most powerful test when $c = 0$. When $c = -2$ and $\delta = -0.95$, the sup-bound Q-test is undersized when b is small and has good power when $b \gg 0$. When $c = -2$ and $\delta = -0.75$, the power of the sup-bound Q-test is close to that of the Bonferroni Q-test. When $c = -20$, the sup-bound Q-test has very poor power.[9] In some sense, the comparison of the sup-bound Q-test with the Bonferroni tests is unfair because the size of the sup-bound test is greater than 5% when the true autoregressive root

[9] Lewellen (2004, Section 2.4) proposes a Bonferroni procedure to remedy the poor power of the sup-bound Q-test for low values of ρ. Although the particular procedure that he proposes does not have correct asymptotic size (see Cavanagh et al., 1995), it can be interpreted as a combination of the Bonferroni t-test and the sup-bound Q-test.

is explosive (i.e., $c>0$), while the Bonferroni tests have the correct size even in the presence of explosive roots.

We conclude that the Bonferroni Q-test has important power advantages over the other feasible tests. Against right-sided alternatives, it has better power than the Bonferroni t-test, especially when the predictor variable is highly persistent, and it has much better power than the sup-bound Q-test when the predictor variable is less persistent.

3.5.3. Where does the power gain come from?

The last section showed that our Bonferroni Q-test is more powerful than the Bonferroni t-test. In this section, we examine the sources of this power gain in detail. We focus our discussion of power to the case $\delta=-0.95$ since the results are similar when $\delta=-0.75$.

We first ask whether the power gain comes from the use of the DF-GLS test rather than the ADF test, or the Q-test rather than the t-test. To answer this question, we consider the following three tests:

1. A Bonferroni test based on the ADF test and the t-test.
2. A Bonferroni test based on the DF-GLS test and the t-test.
3. A Bonferroni test based on the DF-GLS test and the Q-test.

Tests 1 and 3 are the Bonferroni t-test and Q-test, respectively, whose power functions are discussed in the last section. Test 2 is a slight modification of the Bonferroni t-test, whose power function appears in an earlier version of this paper (Campbell and Yogo, 2002, Fig. 5). By comparing the power of tests 1 and 2, we quantify the marginal contribution to power coming from the DF-GLS test. By comparing the power of tests 2 and 3, we quantify the marginal contribution to power coming from the Q-test.

When $c=-2$ and $\delta=-0.95$, the Pitman efficiency of test 1 relative to test 2 is 1.03, which means that test 1 requires 3% more observations than test 2 to achieve 50% power. The Pitman efficiency of test 2 relative to test 3 is 1.20 (i.e., test 2 requires 20% more observations). This shows that when the predictor variable is highly persistent, the use of the Q-test rather than the t-test is a relatively important source of power gain for the Bonferroni Q-test.

When $c=-20$ and $\delta=-0.95$, the Pitman efficiency of test 1 relative to test 2 is 1.07 (i.e., test 1 requires 7% more observations). The Pitman efficiency of test 2 relative to test 3 is 1.03 (i.e., test 2 requires 3% more observations). This shows that when the predictor variable is less persistent, the use of the DF-GLS test rather than the ADF test is a relatively important source of power gain for the Bonferroni Q-test.

We now ask whether the refinement to the Bonferroni test, discussed in Section 3.4, is an important source of power. To answer this question, we recompute the power functions for the Bonferroni t-test and Q-test, reported in Fig. 3, without the refinement. Although these power functions are not directly reported here to conserve space, we summarize our findings.

When $c=-2$ and $\delta=-0.95$, there is essentially no difference in power between the unrefined and refined Bonferroni t-test. However, the Pitman efficiency of the unrefined relative to the refined Bonferroni Q-test is 1.62. When $c=-20$ and $\delta=-0.95$, the Pitman efficiency of the unrefined relative to the refined Bonferroni t-test is 1.23. For the Bonferroni Q-test, the corresponding Pitman efficiency is 1.55. This shows that the refinement is an especially important source of power gain for the Bonferroni Q-test. Since

the Q-test explicitly exploits information about the value of ρ, its confidence interval for β given ρ is very sensitive to ρ, resulting in a rather conservative Bonferroni test without the refinement.

3.6. Finite-sample rejection rates

The construction of the Bonferroni Q-test in Section 3.3 and the power comparisons of various tests in the previous section are based on local-to-unity asymptotics. In this section, we examine whether the asymptotic approximations are accurate in finite samples through Monte Carlo experiments.

Table 3 reports the finite-sample rejection rates for four tests of predictability: the conventional t-test, the Bonferroni t-test, the Bonferroni Q-test implemented as described in Campbell and Yogo (2005), and the sup-bound Q-test. All tests are evaluated at the 5% significance level, where the null hypothesis is $\beta = 0$ against the alternative $\beta > 0$. The rejection rates are based on 10,000 Monte Carlo draws of the sample path using the model (1)–(2), with the initial condition $x_0 = 0$. The nuisance parameters are normalized as $\alpha = \gamma = 0$ and $\sigma_u^2 = \sigma_e^2 = 1$. The innovations have correlation δ and are drawn from a bivariate normal distribution. We report results for three levels of persistence ($c = \{0, -2, -20\}$) and two levels of correlation ($\delta = \{-0.95, -0.75\}$). We consider fairly small sample sizes of 50, 100, and 250 since local-to-unity asymptotics are known to be very accurate for samples larger than 500 (e.g., see Chan, 1988).

The conventional t-test (using the critical value 1.645) has large size distortions, as reported in Elliott and Stock (1994) and Mankiw and Shapiro (1986). For instance, the

Table 3
Finite-sample rejection rates for tests of predictability

c	δ	Obs.	ρ	t-test	Bonf. t-test	Bonf. Q-test	Sup Q-test
0	−0.95	50	1.000	0.412	0.060	0.091	0.062
		100	1.000	0.418	0.055	0.062	0.059
		250	1.000	0.411	0.051	0.051	0.051
	−0.75	50	1.000	0.300	0.065	0.091	0.062
		100	1.000	0.294	0.057	0.063	0.055
		250	1.000	0.295	0.053	0.051	0.052
−2	−0.95	50	0.960	0.272	0.048	0.090	0.004
		100	0.980	0.283	0.047	0.064	0.002
		250	0.992	0.272	0.041	0.046	0.001
	−0.75	50	0.960	0.215	0.044	0.085	0.017
		100	0.980	0.208	0.039	0.061	0.015
		250	0.992	0.205	0.034	0.048	0.011
−20	−0.95	50	0.600	0.096	0.048	0.117	0.000
		100	0.800	0.102	0.050	0.059	0.000
		250	0.920	0.109	0.052	0.037	0.000
	−0.75	50	0.600	0.091	0.048	0.108	0.000
		100	0.800	0.088	0.046	0.051	0.000
		250	0.920	0.091	0.045	0.037	0.000

This table reports the finite-sample rejection rates of one-sided, right-tailed tests of predictability at the 5% significance level. From left to right, the tests are the conventional t-test, Bonferroni t-test, Bonferroni Q-test, and sup-bound Q-test. The rejection rates are based on 10,000 Monte Carlo draws of the sample path from the model (1)–(2), where the innovations are drawn from a bivariate normal distribution with correlation δ.

rejection probability is 27.2% when there are 250 observations, $\rho = 0.992$, and $\delta = -0.95$. On the other hand, the finite-sample rejection rate of the Bonferroni t-test is no greater than 6.5% for all values of ρ and δ considered, which is consistent with the findings reported in Cavanagh et al. (1995).

The Bonferroni Q-test has a finite-sample rejection rate no greater than 6.4% for all levels of ρ and δ considered, as long as the sample size is at least 100. The test does seem to have higher rejection rates when the sample size is as small as 50, especially when the degree of persistence is low (i.e., $c = -20$). Practically, this suggests caution in applying the Bonferroni Q-test in very small samples such as postwar annual data, although the test is satisfactory in sample sizes typically encountered in applications. The sup-bound Q-test is undersized when $c<0$, which translates into loss of power as discussed in the last section.

To check the robustness of our results, we repeat the Monte Carlo exercise under the assumption that the innovations are drawn from a t-distribution with five degrees of freedom. The excess kurtosis of this distribution is nine, chosen to approximate the fat tails in returns data; the estimated kurtosis is never greater than nine in annual, quarterly, or monthly data. The rejection rates are essentially the same as those in Table 3, implying robustness of the asymptotic theory to fat-tailed distributions. The results are available from the authors on request.

As an additional robustness check, we repeat the Monte Carlo exercise under different assumptions about the initial condition. With $c = -20$ and the initial condition $x_0 = \{-2, 2\}$, the Bonferroni Q-test is conservative in the sense that its rejection probability is lower than those reported in Table 3. With $c = \{-2, -20\}$ and the initial condition x_0 drawn from its unconditional distribution, the Bonferroni Q-test has a rejection probability that is slightly lower (at most 2% lower) than those reported in Table 3. To summarize, the Bonferroni Q-test has good finite-sample size under reasonable assumptions about the initial condition.

4. Predictability of stock returns

In this section, we implement our test of predictability on U.S. equity data. We then relate our findings to previous empirical findings in the literature.

4.1. Description of data

We use four different series of stock returns, dividend–price ratio, and earnings–price ratio. The first is annual S&P 500 index data (1871–2002) from Global Financial Data since 1926 and from Shiller (2000) before then. The other three series are annual, quarterly, and monthly NYSE/AMEX value-weighted index data (1926–2002) from the Center for Research in Security Prices (CRSP).

Following Campbell and Shiller (1988), the dividend–price ratio is computed as dividends over the past year divided by the current price, and the earnings–price ratio is computed as a moving average of earnings over the past ten years divided by the current price. Since earnings data are not available for the CRSP series, we instead use the corresponding earnings–price ratio from the S&P 500. Earnings are available at a quarterly frequency since 1935, and an annual frequency before then. Shiller (2000) constructs monthly earnings by linear extrapolation. We instead assign quarterly earnings to each month of the quarter since 1935 and annual earnings to each month of the year before then.

To compute excess returns of stocks over a risk-free return, we use the one-month T-bill rate for the monthly series and the three-month T-bill rate for the quarterly series. For the annual series, the risk-free return is the return from rolling over the three-month T-bill every quarter. Since 1926, the T-bill rates are from the CRSP Indices database. For our longer S&P 500 series, we augment this with U.S. commercial paper rates (New York City) from Macaulay (1938), available through NBER's webpage.

For the three CRSP series, we consider the subsample 1952–2002 in addition to the full sample. This allows us to add two additional predictor variables, the three-month T-bill rate and the long-short yield spread. Following Fama and French (1989), the long yield used in computing the yield spread is Moody's seasoned Aaa corporate bond yield. The short rate is the one-month T-bill rate. Although data are available before 1952, the nature of the interest rate is very different then due to the Fed's policy of pegging the interest rate. Following the usual convention, excess returns and the predictor variables are all in logs.

4.2. Persistence of predictor variables

In Table 4, we report the 95% confidence interval for the autoregressive root ρ (and the corresponding c) for the log dividend–price ratio (d–p), the log earnings–price ratio (e–p), the three-month T-bill rate (r_3), and the long-short yield spread (y–r_1). The confidence interval is computed by the method described in Section 3.2. The autoregressive lag length $p \in [1, \overline{p}]$ for the predictor variable is estimated by the Bayes information criterion (BIC). We set the maximum lag length $\overline{p}$ to four for annual, six for quarterly, and eight for monthly data. The estimated lag lengths are reported in the fourth column of Table 4.

All of the series are highly persistent, often containing a unit root in the confidence interval. An interesting feature of the confidence intervals for the valuation ratios (d–p and e–p) is that they are sensitive to whether the sample period includes data after 1994. The confidence interval for the subsample through 1994 (Panel B) is always less than that for the full sample through 2002 (Panel A). The source of this difference can be explained by Fig. 1, which is a time-series plot of the valuation ratios at quarterly frequency. Around 1994, these valuation ratios begin to drift down to historical lows, making the processes look more nonstationary. The least persistent series is the yield spread, whose confidence interval never contains a unit root.

The high persistence of these predictor variables suggests that first-order asymptotics, which implies that the t-statistic is approximately normal in large samples, could be misleading. As discussed in Section 3.2, whether conventional inference based on the t-test is reliable also depends on the correlation δ between the innovations to excess returns and the predictor variable. We report point estimates of δ in the fifth column of Table 4. As expected, the correlations for the valuation ratios are negative and large. This is because movements in stock returns and these valuation ratios mostly come from movements in the stock price. The large magnitude of δ suggests that inference based on the conventional t-test leads to large size distortions.

Suppose $\delta = -0.9$, which is roughly the relevant value for the valuation ratios. As reported in Table 1, the unknown persistence parameter c must be less than -70 for the size distortion of the t-test to be less than 2.5%. That corresponds to ρ less than 0.09 in annual data, less than 0.77 in quarterly data, and less than 0.92 in monthly data. More formally, we fail to reject the null hypothesis that the size distortion is greater than 2.5% using the pretest described in Section 3.2. For the interest rate variables (r_3 and y–r_1), δ is

Table 4
Estimates of the model parameters

Series	Obs.	Variable	p	δ	DF-GLS	95% CI: ρ	95% CI: c
Panel A: S&P 1880–2002, CRSP 1926–2002							
S&P 500	123	d–p	3	−0.845	−0.855	[0.949, 1.033]	[−6.107, 4.020]
		e–p	1	−0.962	−2.888	[0.768, 0.965]	[−28.262, −4.232]
Annual	77	d–p	1	−0.721	−1.033	[0.903, 1.050]	[−7.343, 3.781]
		e–p	1	−0.957	−2.229	[0.748, 1.000]	[−19.132, −0.027]
Quarterly	305	d–p	1	−0.942	−1.696	[0.957, 1.007]	[−13.081, 2.218]
		e–p	1	−0.986	−2.191	[0.939, 1.000]	[−18.670, 0.145]
Monthly	913	d–p	2	−0.950	−1.657	[0.986, 1.003]	[−12.683, 2.377]
		e–p	1	−0.987	−1.859	[0.984, 1.002]	[−14.797, 1.711]
Panel B: S&P 1880–1994, CRSP 1926–1994							
S&P 500	115	d–p	3	−0.835	−2.002	[0.854, 1.010]	[−16.391, 1.079]
		e–p	1	−0.958	−3.519	[0.663, 0.914]	[−38.471, −9.789]
Annual	69	d–p	1	−0.693	−2.081	[0.745, 1.010]	[−17.341, 0.690]
		e–p	1	−0.959	−2.859	[0.591, 0.940]	[−27.808, −4.074]
Quarterly	273	d–p	1	−0.941	−2.635	[0.910, 0.991]	[−24.579, −2.470]
		e–p	1	−0.988	−2.827	[0.900, 0.986]	[−27.322, −3.844]
Monthly	817	d–p	2	−0.948	−2.551	[0.971, 0.998]	[−23.419, −1.914]
		e–p	2	−0.983	−2.600	[0.970, 0.997]	[−24.105, −2.240]
Panel C: CRSP 1952–2002							
Annual	51	d–p	1	−0.749	−0.462	[0.917, 1.087]	[−4.131, 4.339]
		e–p	1	−0.955	−1.522	[0.773, 1.056]	[−11.354, 2.811]
		r_3	1	0.006	−1.762	[0.725, 1.040]	[−13.756, 1.984]
		y–r_1	1	−0.243	−3.121	[0.363, 0.878]	[−31.870, −6.100]
Quarterly	204	d–p	1	−0.977	−0.392	[0.981, 1.022]	[−3.844, 4.381]
		e–p	1	−0.980	−1.195	[0.958, 1.017]	[−8.478, 3.539]
		r_3	4	−0.095	−1.572	[0.941, 1.013]	[−11.825, 2.669]
		y–r_1	2	−0.100	−2.765	[0.869, 0.983]	[−26.375, −3.347]
Monthly	612	d–p	1	−0.967	−0.275	[0.994, 1.007]	[−3.365, 4.451]
		e–p	1	−0.982	−0.978	[0.989, 1.006]	[−6.950, 3.857]
		r_3	2	−0.071	−1.569	[0.981, 1.004]	[−11.801, 2.676]
		y–r_1	1	−0.066	−4.368	[0.911, 0.968]	[−54.471, −19.335]

This table reports estimates of the parameters for the predictive regression model. Returns are for the annual S&P 500 index and the annual, quarterly, and monthly CRSP value-weighted index. The predictor variables are the log dividend–price ratio (d–p), the log earnings–price ratio (e–p), the three-month T-bill rate (r_3), and the long-short yield spread (y–r_1). p is the estimated autoregressive lag length for the predictor variable, and δ is the estimated correlation between the innovations to returns and the predictor variable. The last two columns are the 95% confidence intervals for the largest autoregressive root (ρ) and the corresponding local-to-unity parameter (c) for each of the predictor variables, computed using the DF-GLS statistic.

much smaller. For these predictor variables, the pretest rejects the null hypothesis, which suggests that the conventional t-test leads to approximately valid inference.

4.3. Testing the predictability of returns

In this section, we construct valid confidence intervals for β through the Bonferroni Q-test to test the predictability of returns. In reporting our confidence interval for β, we scale it by $\widehat{\sigma}_e/\widehat{\sigma}_u$. In other words, we report the confidence interval for $\widetilde{\beta} = (\sigma_e/\sigma_u)\beta$ instead

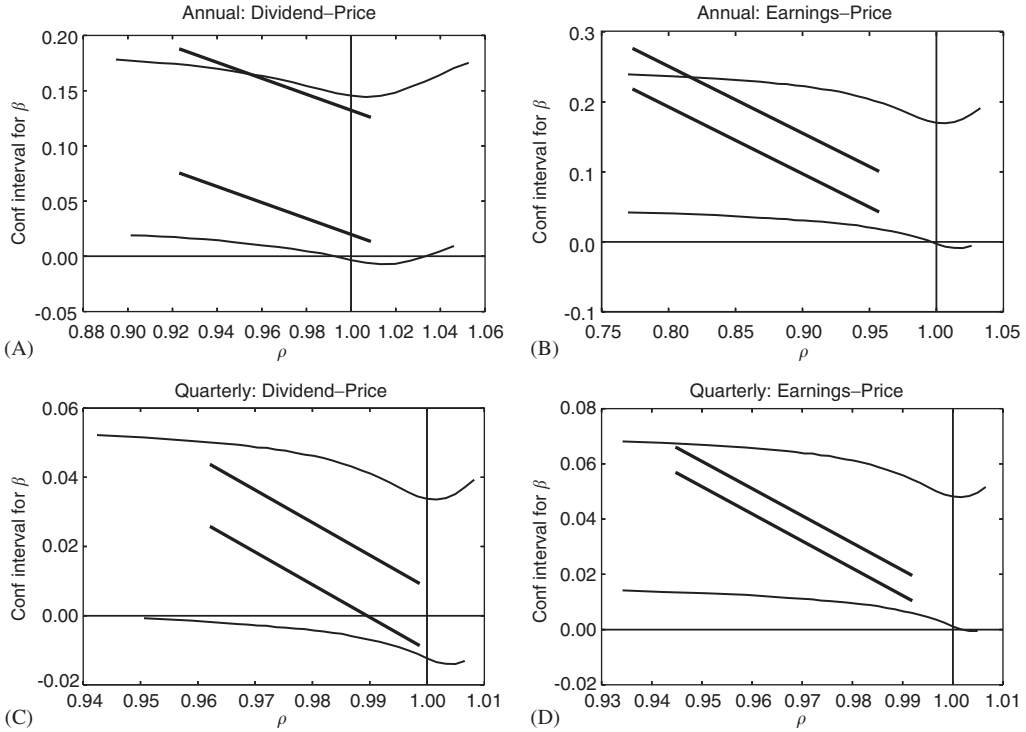

Fig. 4. Bonferroni confidence interval for the valuation ratios. This figure plots the 90% confidence interval for β over the confidence interval for ρ. The significance level for ρ is chosen to result in a 90% Bonferroni confidence interval for β. The thick (thin) line is the confidence interval for β computed by inverting the Q-test (t-test). Returns are for the annual and quarterly CRSP value-weighted index (1926–2002). The predictor variables are the log dividend–price ratio and the log earnings–price ratio.

of β. Although this normalization does not affect inference, it is a more natural way to report the empirical results for two reasons. First, $\widetilde{\beta}$ has a natural interpretation as the coefficient in Eq. (1) when the innovations are normalized to have unit variance (i.e., $\sigma_u^2 = \sigma_e^2 = 1$). Second, by the equality

$$\widetilde{\beta} = \frac{\sigma(\mathrm{E}_{t-1} r_t - \mathrm{E}_{t-2} r_t)}{\sigma(r_t - \mathrm{E}_{t-1} r_t)}, \tag{20}$$

$\widetilde{\beta}$ can be interpreted as the standard deviation of the change in expected returns relative to the standard deviation of the innovation to returns.

Our main findings can most easily be described by a graphical method. Campbell and Yogo (2005) provide a detailed description of the methodology. In Fig. 4, we plot the Bonferroni confidence interval, using the annual and quarterly CRSP series (1926–2002), when the predictor variable is the dividend–price ratio or the earnings–price ratio. The thick lines represent the confidence interval based on the Bonferroni Q-test, and the thin lines represent the confidence interval based on the Bonferroni t-test. Because of the asymmetry in the null distribution of the t-statistic, the confidence interval for ρ used for the right-tailed Bonferroni t-test differs from that used for the left-tailed test (see also

footnote 8). This explains why the length of the lower bound of the interval, corresponding to the right-tailed test, can differ from the upper bound, corresponding to the left-tailed test. The application of the Bonferroni Q-test is new, but the Bonferroni t-test has been applied previously by Torous et al. (2004). We report the latter for the purpose of comparison.

For the annual dividend–price ratio in Panel A, the Bonferroni confidence interval for β based on the Q-test lies strictly above zero. Hence, we can reject the null $\beta = 0$ against the alternative $\beta > 0$ at the 5% level. The Bonferroni confidence interval based on the t-test, however, includes $\beta = 0$. Hence, we cannot reject the null of no predictability using the Bonferroni t-test. This can be interpreted in light of the power comparisons in Fig. 3. From Table 4, $\widehat{\delta} = -0.721$ and the confidence interval for c is $[-7.343, 3.781]$. In this region of the parameter space, the Bonferroni Q-test is more powerful than the Bonferroni t-test against right-sided alternatives, resulting in a tighter confidence interval.

For the quarterly dividend–price ratio in Panel C, the evidence for predictability is weaker. In the relevant range of the confidence interval for ρ, the confidence interval for β contains zero for both the Bonferroni Q-test and t-test, although the confidence interval is again tighter for the Q-test. Using the Bonferroni Q-test, the confidence interval for β lies above zero when $\rho \leqslant 0.988$. This means that if the true ρ is less than 0.988, we can reject the null hypothesis $\beta = 0$ against the alternative $\beta > 0$ at the 5% level. On the other hand, if $\rho > 0.988$, the confidence interval includes $\beta = 0$, so we cannot reject the null. Since there is uncertainty over the true value of ρ, we cannot reject the null of no predictability.

In Panel B, we test for predictability in annual data using the earnings–price ratio as the predictor variable. We find that stock returns are predictable with the Bonferroni Q-test, but not with the Bonferroni t-test. In Panel D, we obtain the same results at the quarterly frequency. Again, the Bonferroni Q-test gives tighter confidence intervals due to better power, which is empirically relevant for detecting predictability.

In Fig. 5, we repeat the same exercise as Fig. 4, using the quarterly CRSP series in the subsample 1952–2002. We report the plots for all four of our predictor variables: (A) the dividend–price ratio, (B) the earnings–price ratio, (C) the T-bill rate, and (D) the yield spread.

For the dividend–price ratio, we find evidence for predictability if $\rho \leqslant 1.004$. This means that if we are willing to rule out explosive roots, confining attention to the area of the figure to the left of the vertical line at $\rho = 1$, we can conclude that returns are predictable with the dividend–price ratio. The confidence interval for ρ, however, includes explosive roots, so we cannot impose $\rho \leqslant 1$ without using prior information about the behavior of the dividend–price ratio.

The earnings–price ratio is a less successful predictor variable in this subsample. We find that ρ must be less than 0.997 before we can conclude that the earnings–price ratio predicts returns. Taking account of the uncertainty in the true value of ρ, we cannot reject the null hypothesis $\beta = 0$. The weaker evidence for predictability in the period since 1952 is partly due to the fact that the valuation ratios appear more persistent when restricted to this subsample. The confidence intervals therefore contain rather large values of ρ that were excluded in Fig. 4.

For the T-bill rate, the Bonferroni confidence interval for β lies strictly below zero for both the Q-test and the t-test over the entire confidence interval for ρ. For the yield spread, the evidence for predictability is similarly strong, with the confidence interval strictly above zero over the entire range of ρ. The power advantage of the Bonferroni Q-test over the

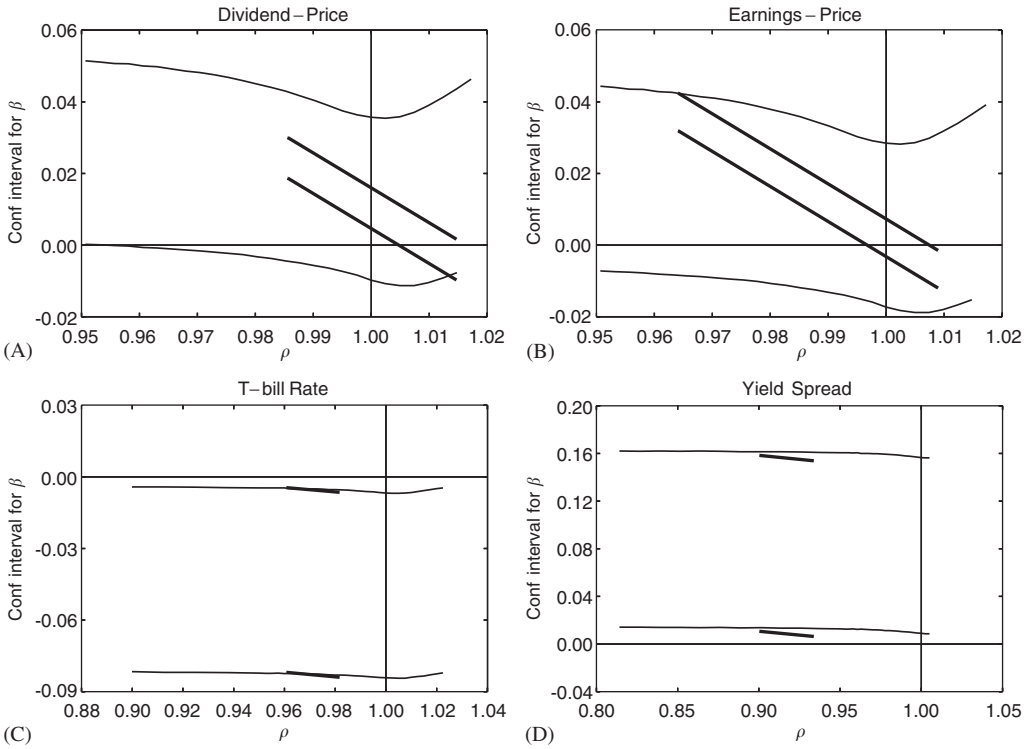

Fig. 5. Bonferroni confidence interval for the post-1952 sample. This figure plots the 90% confidence interval for β over the confidence interval for ρ. The significance level for ρ is chosen to result in a 90% Bonferroni confidence interval for β. The thick (thin) line is the confidence interval for β computed by inverting the Q-test (t-test). Returns are for the quarterly CRSP value-weighted index (1952–2002). The predictor variables are the log dividend–price ratio, the log earnings–price ratio, the three-month T-bill rate, and the long-short yield spread.

Bonferroni t-test is small when δ is small in absolute value, so these tests result in very similar confidence intervals.

In Table 5, we report the complete set of results in tabular form. In the fifth column of the table, we report the 90% Bonferroni confidence intervals for β using the t-test. In the sixth column, we report the 90% Bonferroni confidence interval using the Q-test. In terms of Figs. 4–5, we simply report the minimum and maximum values of β for each of the tests.

Focusing first on the full-sample results in Panel A, the Bonferroni Q-test rejects the null of no predictability for the earnings–price ratio (e–p) at all frequencies. For the dividend–price ratio (d–p), we fail to reject the null except for the annual CRSP series. Using the Bonferroni t-test, we always fail to reject the null due to its poor power relative to the Bonferroni Q-test.

In the subsample through 1994, reported in Panel B, the results are qualitatively similar. In particular, the Bonferroni Q-test finds predictability with the earnings–price ratio at all frequencies. Interestingly, the Bonferroni t-test also finds predictability in this subsample, although the lower bound of the confidence interval is lower than that for the Bonferroni Q-test whenever the null hypothesis is rejected. In this subsample, the evidence for

Table 5
Tests of predictability

Series	Variable	t-stat	$\widehat{\beta}$	90% CI: β		Low CI β
				t-test	Q-test	($\rho = 1$)
Panel A: S&P 1880–2002, CRSP 1926–2002						
S&P 500	d–p	1.967	0.093	[−0.040, 0.136]	[−0.033, 0.114]	−0.017
	e–p	2.762	0.131	[−0.003, 0.189]	**[0.042, 0.224]**	−0.023
Annual	d–p	2.534	0.125	[−0.007, 0.178]	**[0.014, 0.188]**	0.020
	e–p	2.770	0.169	[−0.009, 0.240]	**[0.042, 0.277]**	0.002
Quarterly	d–p	2.060	0.034	[−0.014, 0.052]	[−0.009, 0.044]	−0.010
	e–p	2.908	0.049	[−0.001, 0.068]	**[0.010, 0.066]**	0.002
Monthly	d–p	1.706	0.009	[−0.006, 0.014]	[−0.005, 0.010]	−0.005
	e–p	2.662	0.014	[−0.001, 0.019]	**[0.002, 0.018]**	0.001
Panel B: S&P 1880–1994, CRSP 1926–1994						
S&P 500	d–p	2.233	0.141	[−0.035, 0.217]	[−0.048, 0.183]	−0.081
	e–p	3.321	0.196	**[0.062, 0.272]**	**[0.093, 0.325]**	−0.030
Annual	d–p	2.993	0.212	**[0.025, 0.304]**	**[0.056, 0.332]**	0.011
	e–p	3.409	0.279	**[0.048, 0.380]**	**[0.126, 0.448]**	0.012
Quarterly	d–p	2.304	0.053	[−0.004, 0.083]	[−0.006, 0.076]	−0.027
	e–p	3.506	0.079	**[0.018, 0.107]**	**[0.027, 0.109]**	0.005
Monthly	d–p	1.790	0.013	[−0.004, 0.022]	[−0.007, 0.017]	−0.013
	e–p	3.185	0.022	**[0.002, 0.030]**	**[0.005, 0.028]**	0.000
Panel C: CRSP 1952–2002						
Annual	d–p	2.289	0.124	[−0.023, 0.178]	[−0.007, 0.183]	0.020
	e–p	1.733	0.114	[−0.078, 0.178]	[−0.031, 0.229]	−0.025
	r_3	−1.143	−0.095	[−0.229, 0.045]	[−0.231, 0.042]	—
	y–r_1	1.124	0.136	[−0.087, 0.324]	[−0.075, 0.359]	−0.156
Quarterly	d–p	2.236	0.036	[−0.011, 0.051]	[−0.010, 0.030]	0.005
	e–p	1.777	0.029	[−0.019, 0.044]	[−0.012, 0.042]	−0.003
	r_3	−1.766	−0.042	**[−0.084, −0.004]**	**[−0.084, −0.004]**	−0.086
	y–r_1	1.991	0.090	**[0.009, 0.162]**	**[0.006, 0.158]**	−0.002
Monthly	d–p	2.259	0.012	[−0.004, 0.017]	[−0.004, 0.010]	0.001
	e–p	1.754	0.009	[−0.006, 0.014]	[−0.004, 0.012]	−0.001
	r_3	−2.431	−0.017	**[−0.030, −0.006]**	**[−0.030, −0.006]**	−0.030
	y–r_1	2.963	0.047	**[0.020, 0.072]**	**[0.020, 0.072]**	0.016

This table reports statistics used to infer the predictability of returns. Returns are for the annual S&P 500 index and the annual, quarterly, and monthly CRSP value-weighted index. The predictor variables are the log dividend–price ratio (d–p), the log earnings–price ratio (e–p), the three-month T-bill rate (r_3), and the long-short yield spread (y–r_1). The third and fourth columns report the t-statistic and the point estimate $\widehat{\beta}$ from an OLS regression of returns onto the predictor variable. The next two columns report the 90% Bonferroni confidence intervals for β using the t-test and Q-test, respectively. Confidence intervals that reject the null are in bold. The final column reports the lower bound of the confidence interval for β based on the Q-test at $\rho = 1$.

predictability is sufficiently strong that a relatively inefficient test can also find predictability.

In Panel C, we report the results for the subsample since 1952. In this subsample, we cannot reject the null hypothesis for the valuation ratios (d–p and e–p). For the T-bill rate and the yield spread (r_3 and $y-r_1$), however, we reject the null hypothesis except at annual frequency. For the interest rate variables, the correlation δ is sufficiently small that conventional inference based on the t-test leads to approximately valid inference. This is

confirmed in Panel C, where inference based on the conventional t-test agrees with that based on the Bonferroni Q-test.

As we have seen in Fig. 5, the weak evidence for predictability in this subsample arises from the fact that the confidence intervals for ρ contain explosive roots. If we could obtain tighter confidence intervals for ρ that exclude these values, the lower bound of the confidence intervals for β would rise, strengthening the evidence for predictability. In the last column of Table 5, we report the lower bound of the confidence interval for β at $\rho = 1$. This corresponds to Lewellen's (2004) sup-bound Q-test, which restricts the parameter space to $\rho \leqslant 1$. In terms of Figs. 4–5, this is equivalent to discarding the region of the plots where $\rho > 1$. Under this restriction, the lower bound of the confidence interval for the dividend–price ratio lies above zero at all frequencies. The dividend–price ratio therefore predicts returns in the subsample since 1952 provided that its autoregressive root is not explosive, consistent with Lewellen's findings.

In Table 4, we report that the estimated autoregressive lag length for the predictor variable is one for most of our series. Therefore, the inference for predictability would have been the same had we imposed an AR(1) assumption for these predictor variables, as in the empirical work by Lewellen (2004) and Stambaugh (1999). For the series for which BIC estimated $p > 1$, we repeated our estimates imposing an AR(1) model to see if that changes inference. We find that the confidence intervals for β are essentially the same except for the S&P 500 dividend–price ratio, for which the estimated lag length is $p = 3$. Imposing $p = 1$, the 95% confidence interval based on the Bonferroni Q-test is $[-0.002, 0.190]$ in the full sample and $[0.037, 0.317]$ in the subsample through 1994. These confidence intervals under the AR(1) model show considerably more predictability than those under the AR(3) model, reported in Table 5. This can be explained by the fact that the predictor variable appears more stationary under the AR(1) model; the 95% confidence interval for ρ is now $[0.784, 0.973]$ for the full sample and $[0.563, 0.856]$ for the subsample through 1994. These findings suggest that one does not want to automatically impose an AR(1) assumption on the predictor variable.

To summarize the empirical results, we find reliable evidence for predictability with the earnings–price ratio, the T-bill rate, and the yield spread. The evidence for predictability with the dividend–price ratio is weaker, and we do not find unambiguous evidence for predictability using our Bonferroni Q-test. The Bonferroni Q-test gives tighter confidence intervals than the Bonferroni t-test due to better power. The power gain is empirically important in the full sample through 2002.

4.4. Connection to previous empirical findings

The empirical literature on the predictability of returns is rather large, and in this section, we attempt to interpret the main findings in light of our analysis in the last section.

4.4.1. The conventional t-test

The earliest and most intuitive approach to testing predictability is to run the predictive regression and to compute the t-statistic. One would then reject the null hypothesis $\beta = 0$ against the alternative $\beta > 0$ at the 5% level if the t-statistic is greater than 1.645. In the third column of Table 5, we report the t-statistics from the predictive regressions. Using the conventional critical value, the t-statistics are mostly "significant," often greater than two and sometimes greater than three. Comparing the full sample through 2002 (Panel A) and

the subsample through 1994 (Panel B), the evidence for predictability appears to have weakened in the last eight years. In the late 1990s, stock returns were high when the valuation ratios were at historical lows. Hence, the evidence for predictability "went in the wrong direction."

However, one may worry about statistical inference that is so sensitive to the addition of eight observations to a sample of 115 (for the S&P 500 data) or 32 observations to a sample of 273 (for the quarterly CRSP data). In fact, this sensitivity is evidence of the failure of first-order asymptotics. Intuitively, when a predictor variable is persistent, its sample moments can change dramatically with the addition of a few data points. Since the t-statistic measures the covariance of excess returns with the lagged predictor variable, its value is sensitive to persistent deviations in the predictor variable from the mean. This is what happened in the late 1990s when valuation ratios reached historical lows. Tests that are derived from local-to-unity asymptotics take this persistence into account and hence lead to valid inference.

Using the Bonferroni Q-test, which is robust to the persistence problem, we find that the earnings–price ratio predicts returns in both the full sample and the subsample through 1994. There appears to be some empirical content in the claim that the evidence for predictability has weakened, with the Bonferroni confidence interval based on the Q-test shifting toward zero. Using the Bonferroni confidence interval based on the t-test, we reject the null of no predictability in the subsample through 1994, but not in the full sample. The "weakened" evidence for predictability in the recent years puts a premium on the efficiency of test procedures.

As additional evidence of the failure of first-order asymptotics, we report the OLS point estimates of β in the fourth column of Table 5. As Eqs. (15)–(16) show, the point estimate $\widehat{\beta}$ does not necessarily lie in the center of the robust confidence interval for β. Indeed, $\widehat{\beta}$ for the valuation ratios is usually closer to the upper bound of the Bonferroni confidence interval based on the Q-test, and in a few cases (dividend–price ratio in Panel C), $\widehat{\beta}$ falls strictly above the confidence interval. This is a consequence of the upward finite-sample bias of the OLS estimator arising from the persistence of these predictor variables (see Lewellen, 2004; Stambaugh, 1999).

4.4.2. *Long-horizon tests*

Some authors, notably Campbell and Shiller (1988) and Fama and French (1988), have explored the behavior of stock returns at lower frequencies by regressing long-horizon returns onto financial variables. In annual data, the dividend–price ratio has a smaller autoregressive root than it does in monthly data and is less persistent in that sense. Over several years, the ratio has an even smaller autoregressive root. Unfortunately, this does not eliminate the statistical problem caused by persistence because the effective sample size shrinks as one increases the horizon of the regression.

Recently, a number of authors have pointed out that the finite-sample distribution of the long-horizon regression coefficient and its associated t-statistic can be quite different from the asymptotic distribution due to persistence in the regressor and overlap in the returns data; see Ang and Bekaert (2001), Hodrick (1992), and Nelson and Kim (1993) for computational results, and Torous et al. (2004) and Valkanov (2003) for analytical results. Accounting for these problems, Torous et al. (2004) find no evidence for predictability at long horizons using many of the popular predictor variables. In fact, they find no evidence

for predictability at any horizon or time period, except at quarterly and annual frequencies in the period 1952–1994.

Long-horizon regressions can also be understood as a way to reduce the noise in stock returns, because under the alternative hypothesis that returns are predictable, the variance of the return increases less than proportionally with the investment horizon (see Campbell, 2001; Campbell et al., 1997, Chapter 7). The procedures developed in this paper and in Lewellen (2004) have the important advantage that they reduce noise not only under the alternative but also under the null. Thus, they increase power against local alternatives, while long-horizon regression tests do not.

4.4.3. More recent tests

In this section, we discuss more recent papers that have taken the issue of persistence seriously to develop tests with the correct size even if the predictor variable is highly persistent or I(1).

Lewellen (2004) proposes to test the predictability of returns by computing the Q-statistic evaluated at $\beta_0 = 0$ and $\rho = 1$ (i.e., $Q(0,1)$). His test procedure rejects $\beta = 0$ against the one-sided alternative $\beta > 0$ at the α-level if $Q(0,1) > z_\alpha$. Since the null distribution of $Q(0,1)$ is standard normal under local-to-unity asymptotics, Lewellen's test procedure has correct size as long as $\rho = 1$. If $\rho \neq 1$, this procedure does not in general have the correct size. However, Lewellen's procedure is a valid (although conservative) one-sided test as long as $\delta \leqslant 0$ and $\rho \leqslant 1$. As we have shown in Panel C of Table 5, the 5% one-sided test using the monthly dividend–price ratio rejects when $\rho = 1$, confirming Lewellen's empirical findings.

Based on finance theory, it is reasonable to assume that the dividend–price ratio is mean reverting, at least in the very long run. However, we might not necessarily want to impose Lewellen's parametric assumption that the dividend–price ratio is an AR(1) with $\rho \leqslant 1$. In the absence of knowledge of the true data-generating process, the purpose of a parametric model is to provide a flexible framework to approximate the dynamics of the predictor variable in finite samples, such as in Eqs. (21)–(22) in Appendix A. Allowing for the possibility that $\rho > 1$ can be an important part of that flexibility, especially in light of the recent behavior of the dividend–price ratio. In addition, we allow for possible short-run dynamics in the predictor variable by considering an AR(p), which Lewellen rules out by imposing a strict AR(1).

Another issue that arises with Lewellen's test is that of power. As discussed in Lewellen (2004) and illustrated in Fig. 3, the test can have poor power when the predictor variable is stationary (i.e., $\rho < 1$). For instance, the annual earnings–price ratio for the S&P 500 index has a 95% confidence interval $[0.768, 0.965]$ for ρ. As reported in Panel A of Table 5, the lower bound of the confidence interval for β using the Bonferroni Q-test is 0.043, rejecting the null of no predictability. However, the Q-test at $\rho = 1$ results in a lower bound of -0.023, failing to reject the null. Therefore, the poor power of Lewellen's test understates the strength of the evidence that the earnings–price ratio predicts returns at annual frequency. Similarly, Lewellen's procedure always leads to wider confidence intervals than the Bonferroni Q-test in the subsample through 1994, when the valuation ratios are less persistent.

Torous et al. (2004) develop a test of predictability that is conceptually similar to ours, constructing Bonferroni confidence intervals for β. One difference from our approach is that they construct the confidence interval for ρ using the ADF test, rather than the more

powerful DF-GLS test of Elliott et al. (1996). The second difference is that they use the long-horizon t-test, instead of the Q-test, for constructing the confidence interval of β given ρ. Their choice of the long-horizon t-test is motivated by their objective of highlighting the pitfalls of long-horizon regressions.

A key insight in Torous et al. (2004) is that the evidence for the predictability of returns depends critically on the unknown degree of persistence of the predictor variable. Because we cannot estimate the degree of persistence consistently, the evidence for predictability can be ambiguous. This point is illustrated in Figs. 4–5, where we find that the dividend–price ratio predicts returns if its autoregressive root ρ is sufficiently small. In this paper, we have confirmed their finding that the evidence for predictability by the dividend–price ratio is weaker once its persistence has been properly accounted for.

A different approach to dealing with the problem of persistence is to ignore the data on predictor variables and to base inference solely on the returns data. Under the null that returns are not predictable by a persistent predictor variable, returns should behave like a stationary process. Under the alternative of predictability, returns should have a unit or a near-unit root. Using this approach, Lanne (2002) fails to reject the null of no predictability. However, his test is conservative in the sense that it has poor power when the predictor variable is persistent but not close enough to being integrated.[10] Lanne's empirical finding agrees with ours and those of Torous et al. (2004). From Figs. 4–5, we see that the valuation ratios predict returns provided that their degree of persistence is sufficiently small. In addition, we find evidence for predictability with the yield spread, which has a relatively low degree of persistence compared to the valuation ratios. Lanne's test would fail to detect predictability by less persistent variables like the yield spread.

As revealed by Fig. 3, all the feasible tests considered in this paper are biased. That is, the power of the test can be less than the size, for alternatives sufficiently close to zero. Jansson and Moreira (2003) have made recent progress in the development of unbiased tests for predictive regressions. They characterize the most powerful test in the class of unbiased tests, conditional on ancillary statistics. In principle, their test is useful for testing the predictability of returns, but in practice, implementation requires advanced computational methods; see Polk et al. (2003) for details. Until these tests become easier to implement and are shown to be more powerful in Monte Carlo experiments, we see our procedure as a practical alternative.

5. Conclusion

The hypothesis that stock returns are predictable at long horizons has been called a "new fact in finance" (Cochrane, 1999). That the predictability of stock returns is now widely accepted by financial economists is remarkable given the long tradition of the "random walk" model of stock prices. In this paper, we have shown that there is indeed evidence for predictability, but it is more challenging to detect than previous studies have suggested. The most popular and economically sensible candidates for predictor variables

[10]In fact, Campbell et al. (1997, Chapter 7) construct an example in which returns are univariate white noise but are predictable using a stationary variable with an arbitrary autoregressive coefficient.

(such as the dividend–price ratio, earnings–price ratio, or measures of the interest rate) are highly persistent. When the predictor variable is persistent, the distribution of the t-statistic can be nonstandard, which can lead to over-rejection of the null hypothesis using conventional critical values.

In this paper, we have developed a pretest to determine when the conventional t-test leads to misleading inferences. Using the pretest, we find that the t-test leads to valid inference for the short-term interest rate and the long-short yield spread. Persistence is not a problem for these interest rate variables because their innovations have sufficiently low correlation with the innovations to stock returns. Using the t-test with conventional critical values, we find that these interest rate variables predict returns in the post-1952 sample.

For the dividend–price ratio and the smoothed earnings–price ratio, persistence is an issue since their innovations are highly correlated with the innovations to stock returns. Using our pretest, we find that the conventional t-test can lead to misleading inferences for these valuation ratios. In this paper, we have developed an efficient test of predictability that leads to valid inference regardless of the degree of persistence of the predictor variable. Over the full sample, our test reveals that the earnings–price ratio reliably predicts returns at various frequencies (annual to monthly), while the dividend–price ratio predicts returns only at annual frequency. In the post-1952 sample, the evidence for predictability is weaker, but the dividend–price ratio predicts returns if we can rule out explosive autoregressive roots.

Taken together, these results suggest that there is a predictable component in stock returns, but one that is difficult to detect without careful use of efficient statistical tests.

Appendix A. Generalizing the model and the distributional assumptions

The AR(1) model for the predictor variable (2) is restrictive since it does not allow for short-run dynamics. Moreover, the assumption of normality (i.e., Assumption 1) is unlikely to hold in practice. This appendix therefore generalizes the asymptotic results in Section 3 to a more realistic case when the dynamics of the predictor variable are captured by an AR(p), and the innovations satisfy more general distributional assumptions.

Let L be the lag operator, so that $L^i x_t = x_{t-i}$. We generalize model (2) as

$$x_t = \gamma + \rho x_{t-1} + v_t, \tag{21}$$

$$b(L)v_t = e_t, \tag{22}$$

where $b(L) = \sum_{i=0}^{p-1} b_i L^i$ with $b_0 = 1$ and $b(1) \neq 0$. All the roots of $b(L)$ are assumed to be fixed and less than one in absolute value. Eqs. (21) and (22) together imply that

$$\Delta x_t = \tau + \theta x_{t-1} + \sum_{i=1}^{p-1} \psi_i \Delta x_{t-i} + e_t, \tag{23}$$

where $\theta = (\rho - 1)b(1)$, $\psi_i = -\sum_{j=i}^{p-1} a_j$, and $a(L) = L^{-1}[1 - (1 - \rho L)b(L)]$. In other words, the dynamics of the predictor variable are captured by an AR(p), which is written here in the augmented Dickey–Fuller form.

We assume that the sequence of innovations satisfies the following fairly weak distributional assumptions.

Assumption A.1 (*Martingale difference sequence*). *Let* $\mathscr{F}_t = \{w_s | s \leqslant t\}$ *be the filtration generated by the process* $w_t = (u_t, e_t)'$. *Then*

1. $\mathrm{E}[w_t|\mathscr{F}_{t-1}] = 0$,
2. $\mathrm{E}[w_t w_t'] = \Sigma$,
3. $\sup_t \mathrm{E}[u_t^4] < \infty$, $\sup_t \mathrm{E}[e_t^4] < \infty$, and $\mathrm{E}[x_0^2] < \infty$.

In other words, w_t is a martingale difference sequence with finite fourth moments. The assumption allows the sequence of innovations to be conditionally heteroskedastic as long as it is covariance stationary (i.e., unconditionally homoskedastic). Assumption 1 is a special case when the innovations are i.i.d. normal and the covariance matrix Σ is known.

We collect known asymptotic results from Phillips (1987, Lemma 1) and Cavanagh et al. (1995) and state them as a lemma for reference.

Lemma A.1 (*Weak convergence*). *Suppose* $\rho = 1 + c/T$ *and Assumption* A.1 *holds. The following limits hold jointly.*

1. $T^{-3/2}\sum_{t=1}^{T} x_t^{\mu} \Rightarrow \omega \int J_c^{\mu}(s)\,\mathrm{d}s$,
2. $T^{-2}\sum_{t=1}^{T} x_{t-1}^{\mu 2} \Rightarrow \omega^2 \int J_c^{\mu}(s)^2\,\mathrm{d}s$,
3. $T^{-1}\sum_{t=1}^{T} x_{t-1}^{\mu} v_t \Rightarrow \omega^2 \int J_c^{\mu}(s)\,\mathrm{d}W_e(s) + \frac{1}{2}(\omega^2 - \sigma_v^2)$,
4. $T^{-1}\sum_{t=1}^{T} x_{t-1}^{\mu} u_t \Rightarrow \sigma_u \omega \int J_c^{\mu}(s)\,\mathrm{d}W_u(s)$,

where $\omega = \sigma_e / b(1)$ *and* $\sigma_v^2 = \mathrm{E}[v_t^2]$.

When the predictor variable is an AR(1), the Q-statistic (9) has a standard normal asymptotic distribution under the null. Under the more general model (21)–(22) which allows for higher-order autocorrelation, the statistic (9) is not asymptotically pivotal. However, a suitably modified statistic

$$Q(\beta_0, \rho) = \frac{\sum_{t=1}^{T} x_{t-1}^{\mu}[r_t - \beta_0 x_{t-1} - \sigma_{ue}/(\sigma_e \omega)(x_t - \rho x_{t-1})] + \frac{T}{2}\sigma_{ue}/(\sigma_e \omega)(\omega^2 - \sigma_v^2)}{\sigma_u (1-\delta^2)^{1/2} (\sum_{t=1}^{T} x_{t-1}^{\mu 2})^{1/2}} \tag{24}$$

has a standard normal asymptotic distribution by Lemma A.1. Eq. (14) in the confidence interval for the Bonferroni Q-test becomes

$$\beta(\rho) = \frac{\sum_{t=1}^{T} x_{t-1}^{\mu}[r_t - \sigma_{ue}/(\sigma_e \omega)(x_t - \rho x_{t-1})] + \frac{T}{2}\sigma_{ue}/(\sigma_e \omega)(\omega^2 - \sigma_v^2)}{\sum_{t=1}^{T} x_{t-1}^{\mu 2}}. \tag{25}$$

In the absence of short-run dynamics (i.e., $b(1) = 1$ so that $\omega^2 = \sigma_v^2 = \sigma_e^2$), the Q-statistic reduces to (9). The correction term involving $(\omega^2 - \sigma_v^2)$ is analogous to the correction of the Dickey and Fuller (1981) test by Phillips and Perron (1988).

Appendix B. Asymptotic power of the *t*-test and the *Q*-test

This appendix derives the asymptotic distribution of the t-statistic and the Q-statistic under the local alternative $\beta = \beta_0 + b/T$. These asymptotic representations are used to

compute the power functions of the various test procedures in Section 3.5. The underlying model and distributional assumptions are the same as in Appendix A.

The t-statistic can be written as

$$t(\beta_0) = \frac{b(T^{-2}\sum_{t=1}^{T} x_{t-1}^{\mu 2})^{1/2}}{\sigma_u} + \frac{T^{-1}\sum_{t=1}^{T} x_{t-1}^{\mu} u_t}{\sigma_u (T^{-2}\sum_{t=1}^{T} x_{t-1}^{\mu 2})^{1/2}}.$$

By Lemma A.1 (see also Cavanagh et al., 1995),

$$t(\beta_0) \Rightarrow \frac{b\omega\kappa_c}{\sigma_u} + \delta\frac{\tau_c}{\kappa_c} + (1-\delta^2)^{1/2} Z, \tag{26}$$

where Z is a standard normal random variable independent of $(W_e(s), J_c(s))$. Note that the asymptotic distribution of the t-statistic is not affected by heteroskedasticity in the innovations. Intuitively, the near nonstationarity of the predictor variable dominates any stationary dynamics in the variables.

The three types of Q-test considered in Section 3.5 correspond to $Q(\beta_0, \widetilde{\rho})$ (see Eq. (24)) for particular choices of $\widetilde{\rho}$:

1. Infeasible Q-test: $\widetilde{\rho} = 1 + c/T$, where c is the true value assumed to be known.
2. Bonferroni Q-test: $\widetilde{\rho} = 1 + \overline{c}/T$, where $\overline{c}$ depends on the DF-GLS statistic and δ.
3. Sup-bound Q-test: $\widetilde{\rho} = 1$.

Under the local alternative, the Q-statistic is

$$Q(\beta_0, \widetilde{\rho}) = \frac{b(T^{-2}\sum_{t=1}^{T} x_{t-1}^{\mu 2})^{1/2}}{\sigma_u(1-\delta^2)^{1/2}} + \frac{\delta(\widetilde{c}-c)(T^{-2}\sum_{t=1}^{T} x_{t-1}^{\mu 2})^{1/2}}{\omega(1-\delta^2)^{1/2}} + \frac{T^{-1}\sum_{t=1}^{T} x_{t-1}^{\mu}(u_t - \sigma_{ue}/(\sigma_e\omega)v_t) + \frac{1}{2}\sigma_{ue}/(\sigma_e\omega)(\omega^2 - \sigma_v^2)}{\sigma_u(1-\delta^2)^{1/2}(T^{-2}\sum_{t=1}^{T} x_{t-1}^{\mu 2})^{1/2}},$$

where $\widetilde{c} = T(\widetilde{\rho} - 1)$. By Lemma A.1,

$$Q(\beta_0, \widetilde{\rho}) \Rightarrow \frac{b\omega\kappa_c}{\sigma_u(1-\delta^2)^{1/2}} + \frac{\delta(\widetilde{c}-c)\kappa_c}{(1-\delta^2)^{1/2}} + Z, \tag{27}$$

where $\widetilde{c}$ is understood to be the joint asymptotic limit (with slight abuse of notation). Let $\Phi(z)$ denote one minus the standard normal cumulative distribution function, and let z_α denote the $1-\alpha$ quantile of the standard normal. The power function for the right-tailed test (i.e., $b>0$) is therefore given by

$$\pi_Q(b) = \mathrm{E}\left[\Phi\left(z_\alpha - \frac{b\omega\kappa_c}{\sigma_u(1-\delta^2)^{1/2}} - \frac{\delta(\widetilde{c}-c)\kappa_c}{(1-\delta^2)^{1/2}}\right)\right], \tag{28}$$

where the expectation is taken over the distribution of $(W_e(s), J_c(s))$.

Following Stock (1991, Appendix B), the limiting distributions (26) and (27) are approximated by Monte Carlo simulation. We generate 20,000 realizations of the Gaussian AR(1) (i.e., model (2) under Assumption 1, $\gamma = 0$, and $\sigma_e^2 = 1$) with $T = 500$ and $\rho = 1 + c/T$. The distribution of κ_c is approximated by $(T^{-2}\sum_{t=1}^{T} x_{t-1}^{\mu 2})^{1/2}$, and τ_c is approximated by $T^{-1}\sum_{t=1}^{T} x_{t-1}^{\mu} e_t$.

References

Ang, A., Bekaert, G., 2001. Stock return predictability: is it there? Unpublished working paper. Columbia University.

Billingsley, P., 1999. Convergence of Probability Measures, second ed. Wiley Series in Probability and Statistics. Wiley, New York.

Campbell, J.Y., 1987. Stock returns and the term structure. Journal of Financial Economics 18, 373–399.

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

Campbell, B., Dufour, J.-M., 1991. Over-rejections in rational expectations models: a non-parametric approach to the Mankiw–Shapiro problem. Economics Letters 35, 285–290.

Campbell, B., Dufour, J.-M., 1995. Exact nonparametric orthogonality and random walk tests. Review of Economics and Statistics 77, 1–16.

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

Campbell, J.Y., Yogo, M., 2002. Efficient tests of stock return predictability. Working Paper 1972. Harvard Institute of Economic Research.

Campbell, J.Y., Yogo, M., 2005. Implementing the econometric methods in "Efficient tests of stock return predictability". Unpublished working paper. University of Pennsylvania.

Campbell, J.Y., Lo, A.W., MacKinlay, A.C., 1997. The Econometrics of Financial Markets. Princeton University Press, Princeton, NJ.

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

Chan, N.H., 1988. The parameter inference for nearly nonstationary time series. Journal of the American Statistical Association 83, 857–862.

Cochrane, J.H., 1999. New facts in finance. Working Paper 7169. National Bureau of Economic Research.

Cox, D.R., Hinkley, D.V., 1974. Theoretical Statistics. Chapman & Hall, London.

Dickey, D.A., Fuller, W.A., 1981. Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica 49, 1057–1072.

Elliott, G., Stock, J.H., 1994. Inference in time series regression when the order of integration of a regressor is unknown. Econometric Theory 10, 672–700.

Elliott, G., Stock, J.H., 2001. Confidence intervals for autoregressive coefficients near one. Journal of Econometrics 103, 155–181.

Elliott, G., Rothenberg, T.J., Stock, J.H., 1996. Efficient tests for an autoregressive unit root. Econometrica 64, 813–836.

Evans, G.B.A., Savin, N.E., 1981. Testing for unit roots: 1. Econometrica 49, 753–779.

Evans, G.B.A., Savin, N.E., 1984. Testing for unit roots: 2. Econometrica 52, 1241–1270.

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

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

Fama, E.F., Schwert, G.W., 1977. Asset returns and inflation. Journal of Financial Economics 5, 115–146.

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

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

Jansson, M., Moreira, M.J., 2003. Optimal inference in regression models with nearly integrated regressors. Unpublished working paper. Harvard University.

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

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

Lanne, M., 2002. Testing the predictability of stock returns. Review of Economics and Statistics 84, 407–415.

Lehmann, E.L., 1986. Testing Statistical Hypotheses, second ed. Springer Texts in Statistics. Springer, New York.

Lehmann, E.L., 1999. Elements of Large Sample Theory, second ed. Springer Texts in Statistics. Springer, New York.

Lewellen, J., 2004. Predicting returns with financial ratios. Journal of Financial Economics 74, 209–235.
Macaulay, F.R., 1938. Some Theoretical Problems Suggested by the Movements of Interest Rates, Bond Yields, and Stock Prices in the United States Since 1856. National Bureau of Economic Research, New York.
Mankiw, N.G., Shapiro, M.D., 1986. Do we reject too often? Small sample properties of tests of rational expectations models. Economics Letters 20, 139–145.
Nelson, C.R., Kim, M.J., 1993. Predictable stock returns: the role of small sample bias. Journal of Finance 48, 641–661.
Phillips, P.C.B., 1987. Towards a unified asymptotic theory for autoregression. Biometrika 74, 535–547.
Phillips, P.C.B., Perron, P., 1988. Testing for a unit root in time series regression. Biometrika 75, 335–346.
Polk, C., Thompson, S., Vuolteenaho, T., 2003. New forecasts of the equity premium. Unpublished working paper. Harvard University.
Richardson, M., Stock, J.H., 1989. Drawing inferences from statistics based on multiyear asset returns. Journal of Financial Economics 25, 323–348.
Shiller, R.J., 2000. Irrational Exuberance. Princeton University Press, Princeton, NJ.
Stambaugh, R.F., 1999. Predictive regressions. Journal of Financial Economics 54, 375–421.
Stock, J.H., 1991. Confidence intervals for the largest autoregressive root in US macroeconomic time series. Journal of Monetary Economics 28, 435–459.
Stock, J.H., 1994. Unit roots, structural breaks and trends. In: Engle, R.F., McFadden, D.L. (Eds.), Handbook of Econometrics, vol. 4. Elsevier Science, New York, pp. 2739–2841.
Torous, W., Valkanov, R., Yan, S., 2004. On predicting stock returns with nearly integrated explanatory variables. Journal of Business 77, 937–966.
Valkanov, R., 2003. Long-horizon regressions: theoretical results and applications. Journal of Financial Economics 68, 201–232.