Predictive Regressions with Persistent Regressors: Monte Carlo evidence for the Bonferroni Q-test

Jan Philipp Wöltjen
Seminar in Statistics CAU Kiel

September 12, 2019

Abstract

Campbell and Yogo (2006) propose a Bonferroni-type procedure for valid inference of predictive regressions with almost nonstationary regressors. In this study, the rationale behind their method is explained. Empirical findings are replicated and extended with out-of-sample data that emerged since the method's publication. Furthermore, the method's behavior is investigate via Monte Carlo simulation, highlighting practically relevant scenarios where it over-rejects the null hypothesis of no predictability.

1 Introduction

The question whether asset returns can be predicted with the help of observable variables is of great interest to financial institutions and the public as a whole. One of the most economically intuitive relationships exists between the price of a company's stock and the discounted earnings it will generate in the future. Since future earnings are not observable, current earnings or dividends or moving averages of them are often used as a proxy. To statistically test whether these variables are indeed predictive, one may test the significance of coefficients in a regression of the supposedly predictive variable onto equity market returns. The most straightforward way of doing this is via the t-test. Elliott and Stock (1994) find that this test has a nonstandard distribution under the null if the largest autoregressive root of the regressor is close to unity and the innovations of regressor and regressand are correlated. Given the presence of these nuisances, the t-test will lead to invalid inference. In this context, Campbell and Yogo (2006) develop a test that exploits information about the autoregressive root of the regressor and thus tries to overcome this issue. Section 2 formalizes the regression setup and outlines the problem. Section 3 explains the derivation of the Campbell and Yogo's (2006) test. This is done by first considering optimal tests, albeit under unrealistic assumptions, in section 3.1. This section also establishes the conditions for a nonstandard distribution of the t-statistic. A pretest based on these conditions is described in section 3.2. Section 3.3 details the use of Bonferroni's inequality in making the test feasible. The test is further generalized to more realistic assumptions in section 3.4 and its mathematical implementation is described. Section 3.5 summarizes the power advantages of the test found by Campbell and Yogo (2006). Section 4 proceeds with evaluating finite-sample rejection rates of the feasible Q-test under various innovation distributions and violations of assumptions made in section 3.4. First, Monte Carlo evidence of Campbell and Yogo (2006) is scrutinized and replicated in section 4.1. Section 4.2 demonstrates over-rejection in

finite-samples when the largest autoregressive root of the regressor is nonlocal-to-unity. An attempt of modeling innovations by a multivariate GARCH process is then made in section 4.3. This is further refined in section 4.4 where it is hypothesized that innovation variances of both the regressor and the regressand share a common stochastic trend that is modeled by a GARCH process. Section 4.5 concludes the Monte Carlo evidence by simulation of the common stochastic trend by GJR-GARCH with conditional leptokurtosis. Empirical findings are replicated and evaluated on out-of-sample data in section 5. Finally, section 6 concludes.

2 The regression setup and its problem

With log-returns r_t observable at time t and lagged supposedly predictive variable x_{t-1} observable at $t-1$, Campbell and Yogo (2006) consider the regression system

$$r_t = \alpha + \beta x_{t-1} + u_t \tag{1}$$
$$x_t = \gamma + \rho x_{t-1} + e_t. \tag{2}$$

They further assume normality, i.e.,

$$w_t = (u_t, e_t)' \overset{iid}{\sim} \mathrm{N}(\mathbf{0}, \mathbf{\Sigma}) \tag{3}$$

where $\mathbf{\Sigma} = \begin{bmatrix} \sigma_u^2 & \sigma_{ue} \\ \sigma_{ue} & \sigma_e^2 \end{bmatrix}$ is known.

Whether x_{t-1} is indeed predictive may be ascertained by evaluating the statistically significance of the β coefficient obtained from an ordinary least squares [OLS] regression, i.e., testing the null hypothesis $H_0 : \beta = 0$. A common approach is to perform a t-test of β. Elliott and Stock (1994), however, show that the t-test leads to invalid inference if x_t is persistent, i.e., if ρ is close to unity, and the noise terms u_t and e_t are highly correlated. Since the regressor is a valuation ratio, which is a function of the stock price, and the regressand is the stock's return, which is also a function of the current price, there is strong reason to believe that u_t and e_t will indeed be highly (negatively) correlated. Sample correlations, $\hat{\delta}$, in Table 9 on page 15 will confirm this belief. It will also be seen that the regressor is highly persistent, manifested by the 90% confidence interval for the largest autoregressive root containing unity. Hence, worry about the validity of the t-test is justified. A pretest, testing the null of the actual size of the t-test being greater than some deemed-acceptable nominal size α based on the joint presence of both aforementioned conditions, is explained in more detail in section 3.2.

3 Derivation of the tests

3.1 Infeasible tests

In developing their test, Campbell and Yogo (2006) start by considering the t-statistic

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

where $x_{t-1}^{\mu} = x_{t-1} - T^{-1} \sum_{t=1}^{T} x_{t-1}$ denotes the demeaned regressor and $\widehat{\beta}$ is the OLS estimate of β. They argue that this test ignores information contained in the system (1, 2) and can

thus not be optimal. To see this, consider the joint log likelihood function of the system (1, 2)

$$L(\beta,\rho,\alpha,\gamma) = -\frac{1}{1-\delta^2}\sum_{t=1}^{T}\left[\frac{(r_t-\alpha-\beta x_{t-1})^2}{\sigma_u^2} - 2\delta\frac{(r_t-\alpha-\beta x_{t-1})(x_t-\gamma-\rho x_{t-1})}{\sigma_u\sigma_e} + \frac{(x_t-\gamma-\rho x_{t-1})^2}{\sigma_e^2}\right] \tag{5}$$

and observe that the t-test squared is equal to the likelihood ratio test statistic

$$\max_{\beta,\rho,\alpha,\gamma} L(\beta,\rho,\alpha,\gamma) - \max_{\rho,\alpha,\gamma} L(\beta_0,\rho,\alpha,\gamma) = t(\beta_0)^2. \tag{6}$$

But this test statistic turns out to be the same if only the marginal log likelihood

$$L(\beta,\alpha) = -\sum_{t=1}^{T}(r_t-\alpha-\beta x_{t-1})^2 \tag{7}$$

is used in its computation. Thus the t-test ignores information about ρ.

To reason about how to incorporate information about ρ into the test Campbell and Yogo (2006) assume ρ to be known at first. If, further, the assumption $\alpha=\gamma=0$ is made, the only unknown variable that remains is β. Now, the likelihood function can be denoted as $L(\beta)$. If one restricts oneself to consider only the simple alternative of the form $\beta=\beta_1$ one can reason by the Neyman–Pearson Lemma that the most powerful test is of the form

$$\begin{aligned}\sigma_u^2\left(1-\delta^2\right)\left(L(\beta_1)-L(\beta_0)\right) =& 2(\beta_1-\beta_0)\sum_{t=1}^{T}x_{t-1}\left[r_t-\beta_{ue}(x_t-\rho x_{t-1})\right] \\ &-\left(\beta_1^2-\beta_0^2\right)\sum_{t=1}^{T}x_{t-1}^2 > C,\end{aligned} \tag{8}$$

with $\beta_{ue}=\sigma_{ue}/\sigma_e^2$ and C being some constant. This optimal test is not uniformly most powerful [UMP], however, since it is a weighted sum of minimal sufficient statistics and the weights depend on β_1. For that reason, Campbell and Yogo (2006) propose to condition the test on the ancillary statistic $\sum_{t=1}^{T}x_{t-1}^2$. As a result the test can be simplified to

$$\sum_{t=1}^{T}x_{t-1}\left[r_t-\beta_{ue}(x_t-\rho x_{t-1})\right] > C, \tag{9}$$

which is UMP for alternatives of the form $\beta_1>\beta_0$ when ρ is known.

For the test-statistic to have a standard normal distribution under the null, Campbell and Yogo (2006) propose to recenter and rescale it, which results in

$$\frac{\sum_{t=1}^{T}x_{t-1}\left[r_t-\beta_0 x_{t-1}-\beta_{ue}(x_t-\rho x_{t-1})\right]}{\sigma_u\left(1-\delta^2\right)^{1/2}\left(\sum_{t=1}^{T}x_{t-1}^2\right)^{1/2}} > C. \tag{10}$$

Revoking the previous assumption of $\alpha=\gamma=0$ and generalizing instead to any unknown α and γ by replacing x_{t-1} by its demeaned value x_{t-1}^{μ}, the test

$$Q(\beta_0,\rho) = \frac{\sum_{t=1}^{T}x_{t-1}^{\mu}\left[r_t-\beta_0 x_{t-1}-\beta_{ue}(x_t-\rho x_{t-1})\right]}{\sigma_u\left(1-\delta^2\right)^{1/2}\left(\sum_{t=1}^{T}x_{t-1}^{\mu 2}\right)^{1/2}} > C \tag{11}$$

is UMP conditional on $\sum_{t=1}^{T} x_{t-1}^{\mu 2}$. This Q-statistic, as Campbell and Yogo (2006) call it, has an intuitive interpretations when $\beta_0 = 0$. It is the t-statistic of the $\beta^\star$ coefficient in the regression

$$r_t - \beta_{ue}\left(x_t - \rho x_{t-1}\right) = \alpha^\star + \beta^\star x_{t-1} + v_t. \tag{12}$$

Regression (12) can be interpreted as regressing the de-noised returns onto the regressor x_{t-1}, where the information contained in ρ and the correlation of the shocks is exploited. When $\beta_{ue} = 0$, i.e., the correlation of the shocks is zero, the Q-statistic simplifies to the t-statistic which converges to a standard-normal distribution in this case. When $\beta_{ue} \neq 0$, the t-statistic does not converge to a standard normal distribution. Instead, as shown by Elliott and Stock (1994), under local-to-unity asymptotic theory it has the null distribution

$$t\left(\beta_0\right) \Rightarrow \delta\frac{\tau_c}{\kappa_c} + \left(1-\delta^2\right)^{1/2} Z, \tag{13}$$

where $\left(W_u(s), W_e(s)\right)'$ is a two-dimensional Wiener process with correlation δ, $J_c(s)$ is defined by $\mathrm{d}J_c(s) = cJ_c(s)\mathrm{d}s + \mathrm{d}W_e(s)$ with $J_c(0) = 0$, Z is a standard normal random variable independent of $\left(W_e(s), J_c(s)\right)$, $\kappa_c = \left(\int J_c^\mu(s)^2\mathrm{d}s\right)^{1/2}$, and $\tau_c = \int J_c^\mu(s)\mathrm{d}W_e(s)$.

Within the local-to-unity asymptotic theory the persistence of the process x_t is modeled as $\rho = 1 + c/T$ where c is a constant and T is the sample size. Hence, when $c < 0$, x_t is $I(0)$ but highly persistent and its sample moments do not converge in probability to constants but to functionals of a diffusion process instead. The t-test is not feasible since it depends on the unknown parameter ρ through τ_c/κ_c, which also makes it non-standard.

3.2 A pretest

From equation (13) it's easy to see that if $\delta = 0$, the nuisance term $\delta\frac{\tau_c}{\kappa_c}$ vanishes and the t-statistic collapses to the standard normal random variable Z. It therefore makes sense to test whether δ is different from zero. If it isn't, one can infer that the t-statistic is approximately standard normal and inference based on it is valid. This, however, is not the only condition in which there is no size distortion. Phillips (1987) shows that if x_t is not persistent, τ_c/κ_c converges to a different standard normal random variable $Z^\star$. Hence the t-statistic, as a sum of two independent standard normal random variables, is likewise standard normal.

Campbell and Yogo (2006) use these facts as bases for a pretest. This test supposes that an actual size $\alpha^\star \geq \alpha$ may be acceptable if its not much greater than the nominal size α. It then tests whether $\alpha^\star$ is greater than a prespecified acceptable size. It would be tempting to use a unit root test and infer no size distortion if its null of nonstationarity can be rejected. Elliott and Stock (1994) show, however, that this does not guarantee valid inference via the t-statistic. Instead, Campbell and Yogo (2006) use a unit root test statistic to construct a confidence interval for c by inverting its alternative distribution in the spirit of Stock (1991). δ, on the other hand, can be consistently estimated from the residuals of regressions (1, 2). The size of the t-statistic forms a two-dimensional surface in the c-δ-parameter space. If the confidence interval for c lies strictly outside the region where the size is above the acceptable threshold for $\hat{\delta}$, the null of unacceptable size distortion can be rejected. This test is implemented with tables provided by Campbell and Yogo (2006), who use the DF-GLS test statistic proposed by Elliott et al. (1996) to construct the confidence interval for c. The application accompanying uses the closest tabulated values to the estimated values without interpolating.

3.3 Making the Q-test feasible

When the pretest cannot reject the null of unacceptable size distortion inference cannot be based the t-test without adjustment. While the Q-test is UMP when ρ and δ are known, in practice this is not the case. Furthermore, ρ cannot be estimated consistently since its OLS estimator converges at rate T. Instead, analogously to the description in the previous section, Campbell and Yogo (2006) construct a $100\left(1-\alpha_1\right)\%$ confidence interval, $C_\rho\left(\alpha_1\right)$ for ρ from a unit root test statistic. They then construct $C_{\beta|\rho}\left(\alpha_2\right)$, a $100\left(1-\alpha_2\right)\%$ conditional on ρ confidence interval for β. To marginalize ρ and thus getting an unconditional confidence interval for β they take the union over $\rho \in C_\rho\left(\alpha_1\right)$

$$C_\beta(\alpha) = \bigcup_{\rho \in C_\rho(\alpha_1)} C_{\beta|\rho}\left(\alpha_2\right). \tag{14}$$

$C_\beta(\alpha)$ has coverage of at least $100(1-\alpha)\%$ with $\alpha = \alpha_1 + \alpha_2$. This follows from Bonferroni's inequality which gives this method its name. When the regressor is a valuation ratio, a negative value for δ should be expected. Further, note that $\beta(\rho)$ is given by

$$\beta(\rho) = \frac{\sum_{t=1}^{T} x_{t-1}^{\mu}\left[r_t - \beta_{ue}\left(x_t - \rho x_{t-1}\right)\right]}{\sum_{t=1}^{T} x_{t-1}^{\mu 2}}. \tag{15}$$

Hence, the estimate of β declines linearly in ρ and its confidence interval is

$$C_\beta(\alpha) = \left[\underline{\beta}\left(\overline{\rho}\left(\overline{\alpha}_1\right), \alpha_2\right), \overline{\beta}\left(\underline{\rho}\left(\underline{\alpha}_1\right), \alpha_2\right)\right], \tag{16}$$

with $C_\rho\left(\alpha_1\right) = \left[\underline{\rho}\left(\underline{\alpha}_1\right), \overline{\rho}\left(\overline{\alpha}_1\right)\right]$ being the confidence interval for ρ, $\underline{\alpha}_1 = \Pr\left(\rho < \underline{\rho}\left(\underline{\alpha}_1\right)\right)$, $\overline{\alpha}_1 = \Pr\left(\rho > \overline{\rho}\left(\overline{\alpha}_1\right)\right)$, and $\alpha_1 = \underline{\alpha}_1 + \overline{\alpha}_1$. We also know that the Q-statistic is normally distributed. Let $z_{\alpha_2/2}$ denote the $1-\alpha_2/2$ quantile of the standard normal distribution. Then the lower bound is given by

$$\underline{\beta}\left(\overline{\rho}\left(\overline{\alpha}_1\right), \alpha_2\right) = \beta\left(\overline{\rho}\left(\overline{\alpha}_1\right)\right) - z_{\alpha_2/2}\sigma_u\left(\frac{1-\delta^2}{\sum_{t=1}^{T} x_{t-1}^{\mu}}\right)^{1/2} \tag{17}$$

and the upper bound is given by

$$\overline{\beta}\left(\underline{\rho}\left(\underline{\alpha}_1\right), \alpha_2\right) = \beta\left(\underline{\rho}\left(\underline{\alpha}_1\right)\right) + z_{\alpha_2/2}\sigma_u\left(\frac{1-\delta^2}{\sum_{t=1}^{T} x_{t-1}^{\mu 2}}\right)^{1/2}. \tag{18}$$

If this confidence interval does not contain zero, the regressor is inferred to have predictive power. While this test is valid, it is conservative, however, since Bonferroni's inequality is likely to be strict, i.e., the coverage of $C_\beta(\alpha)$ is likely greater than $100(1-\alpha)\%$. For that reason, Campbell and Yogo (2006) refine the confidence interval based on a numerical method proposed by Cavanagh et al. (1995). To obtain a tighter confidence interval with significance level $\widetilde{\alpha}$, according to this method, α_2 is fixed first. $\Pr\left(\underline{\beta}\left(\overline{\rho}\left(\overline{\alpha}_1\right), \alpha_2\right) > \beta\right)$ is then evaluated for each δ, and $\overline{\alpha}_1$ is selected such that

$$\Pr\left(\underline{\beta}\left(\overline{\rho}\left(\overline{\alpha}_1\right), \alpha_2\right) > \beta\right) \leq \widetilde{\alpha}/2 \tag{19}$$

holds for all c and, importantly,

$$\Pr\left(\underline{\beta}\left(\overline{\rho}\left(\overline{\alpha}_1\right), \alpha_2\right) > \beta\right) = \widetilde{\alpha}/2 \tag{20}$$

holds for some c. Similarly, $\underline{\alpha}_1$ is selected such that

$$\Pr\left(\overline{\beta}\left(\underline{\rho}\left(\underline{\alpha}_1\right), \alpha_2\right) < \beta\right) \leq \widetilde{\alpha}/2 \tag{21}$$

holds again for all c with equality at some c. A one-sided Bonferroni test based a confidence interval thus obtained has size $\widetilde{\alpha}/2$ for some permissible c while a two-sided test has at most size $\widetilde{\alpha}$.

3.4 Implementation of the feasible Q-test

The feasible Q-test, generalized to less restrictive assumptions, is implemented in the R programming language of R Core Team (2013) following the steps proposed by Campbell and Yogo (2005). Suppose the regressor follows an $AR(p)$ process. Further, the innovations are not necessarily normally distributed but are a martingale difference sequence, i.e.,

$$\mathrm{E}\left[w_t | \mathscr{F}_{t-1}\right] = 0 \tag{22}$$

where $\mathscr{F}_t = \{w_s | s \leq t\}$ denotes the filtration generated by $w_t = (u_t, e_t)'$. Its fourth moments are assumed to be finite and the unconditional variance is fixed, i.e.,

$$\begin{array}{l} \mathrm{E}\left[w_t w_t'\right] = \Sigma \\ \sup_t \mathrm{E}\left[u_t^4\right] < \infty, \sup_t \mathrm{E}\left[t_t^4\right] < \infty, \text{ and } \mathrm{E}\left[x_0^2\right] < \infty. \end{array} \tag{23}$$

The dynamics of x_t are written as

$$x_t = \gamma + \rho x_{t-1} + v_t, \tag{24}$$

where ρ is the largest autoregressive root and $b(L)v_t = e_t$ with $b(L) = \sum_{i=0}^{p-1} b_i L^i$, $b_0 = 1$, and $b(1) \neq 0$. In this generalized case, the Q-statistic is not asymptotically standard normal distributed under the null. This is corrected by modifying it to

$$Q\left(\beta_0, \rho\right) = \frac{\sum_{t=1}^{T} x_{t-1}^{\mu}\left[r_t - \beta_0 x_{t-1} - \sigma_{ue}/\left(\sigma_e \omega\right)\left(x_t - \rho x_{t-1}\right)\right] + \frac{T}{2}\sigma_{ue}/\left(\sigma_e\omega\right)\left(\omega^2 - \sigma_v^2\right)}{\sigma_u\left(1-\delta^2\right)^{1/2}\left(\sum_{t=1}^{T} x_{t-1}^{\mu 2}\right)^{1/2}}, \tag{25}$$

where $\omega = \sigma_e / b(1)$.

Implementing the feasible Q-test, first $\hat{\beta}$ and its standard error $\mathrm{SE}(\widehat{\beta})$ are obtained by OLS estimation of equation (1). x_t is written in its augmented Dickey-Fuller form

$$\Delta x_t = \tau + \theta x_{t-1} + \sum_{i=1}^{p-1} \psi_i \Delta x_{t-i} + e_t, \tag{26}$$

where $\psi_i = -\sum_{j=i}^{p-1} a_j$, $a(L) = L^{-1}[1 - (1-\rho L)b(L)]$, and $\theta = (\rho - 1)b(1)$. Estimation of $\widehat{\psi}_i (i = 1, \ldots, p-1)$ is done via OLS where p is either explicitly specified or estimated via the Bayesian information criterion [BIC]. With $\widehat{u}_t$ and $\widehat{e}_t$ being the residuals of regression (1) and

(2), respectively,

$$\widehat{\sigma}_u^2 = \frac{1}{T-2}\sum_{t=1}^{T}\widehat{u}_t^2 \tag{27}$$

$$\widehat{\sigma}_e^2 = \frac{1}{T-2}\sum_{t=1}^{T}\widehat{e}_t^2 \tag{28}$$

$$\widehat{\sigma}_{ue} = \frac{1}{T-2}\sum_{t=1}^{T}\widehat{u}_t\widehat{e}_t \tag{29}$$

$$\widehat{\delta} = \frac{\widehat{\sigma}_{ue}}{\widehat{\sigma}_u\widehat{\sigma}_e} \tag{30}$$

$$\widehat{\omega}^2 = \widehat{\sigma}_e^2 / \left(1 - \sum_{i=1}^{p-1}\widehat{\psi}_i\right)^2 \tag{31}$$

are then computed. OLS estimation of equation (24) yields $\widehat{\rho}$ and its standard error $\mathrm{SE}(\widehat{\rho})$. The residual $\widehat{v}_t$ is used to compute

$$\widehat{\sigma}_v^2 = (T-2)^{-1}\sum_{t=1}^{T}\widehat{v}_t^2. \tag{32}$$

Now, to compute the DF-GLS statistic, $(x_0, x_1 - \rho_{GLS}x_0, \ldots, x_T - \rho_{GLS}x_{T-1})'$ is regressed onto $(1, 1-\rho_{GLS}, \ldots, 1-\rho_{GLS})'$ with $\rho_{GLS} = 1 - 7/T$. The coefficient μ_{GLS} is used to compute $\overline{x}_t = x_t - \mu_{GLS}$ which, in turn, is needed for the regression

$$\Delta\overline{x}_t = \theta\overline{x}_{t-1} + \sum_{i=1}^{p-1}\psi_i\Delta\overline{x}_{t-i} + e_t, \tag{33}$$

which is estimated with no intercept. Finally, the t-statistic for θ is computed. This is the DF-GLS statistic. To confirm that no error is conducted in this estimation procedure, the DF-GLS statistic thus obtained is compared to the implementation in the urca package of Pfaff (2008).

The confidence interval $[\underline{\rho}, \overline{\rho}] = [1 + \underline{c}/T, 1 + \overline{c}/T]$ is obtained by first finding $[\underline{c}, \overline{c}]$ via tables 2–11 of Campbell and Yogo (2005), where the closest values to the estimated DF-GLS statistic and $\hat{\delta}$ are used. Let a temporary variable $r^\star$ be defined by

$$r_t^* = r_t - \widehat{\sigma}_{ue}\widehat{\sigma}_e^{-2}\left(x_t - \rho x_{t-1}\right). \tag{34}$$

Now the regression

$$r_t^* = \alpha + \beta x_{t-1} + u_t \tag{35}$$

is run for each $\rho = \{\underline{\rho}, \overline{\rho}\}$ to get $\widehat{\beta}(\rho)$. $[\underline{\beta}(\rho), \overline{\beta}(\rho)]$ is then obtained by computing

$$\underline{\beta}(\rho) = \widehat{\beta}(\rho) + \frac{T-2}{2}\frac{\widehat{\sigma_{ue}}}{\widehat{\sigma}_e\widehat{\omega}}\left(\frac{\widehat{\omega}^2}{\widehat{\sigma}_v^2} - 1\right)\mathrm{SE}(\widehat{\rho})^2 - 1.645\left(1 - \widehat{\delta}^2\right)^{1/2}\mathrm{SE}(\widehat{\beta}) \tag{36}$$

and

$$\overline{\beta}(\rho) = \widehat{\beta}(\rho) + \frac{T-2}{2}\frac{\widehat{\sigma}_{ue}}{\widehat{\sigma}_e\widehat{\omega}}\left(\frac{\widehat{\omega}^2}{\widehat{\sigma}_v^2} - 1\right)\mathrm{SE}(\widehat{\rho})^2 + 1.645\left(1 - \widehat{\delta}^2\right)^{1/2}\mathrm{SE}(\widehat{\beta}) \tag{37}$$

A 5% one-sided or 10% two-sided test of $H_0 : \beta = 0$ is based on the 90% Bonferroni confidence interval

$$[\underline{\beta}(\overline{\rho}), \overline{\beta}(\underline{\rho})]$$

3.5 Power comparison

Campbell and Yogo (2006) evaluate the power of the feasible Q-test under local-to-unity asymptotics. In this framework, the OLS estimators $\hat{\beta}$ and $\hat{\rho}$ are consistent at rate T. Thus, to have a meaningful comparison, local alternatives of the form $\beta = \beta_0 + b/T$ are considered where b is a constant. The feasible Q-test is found to dominate the feasible t-test in all cases considered with increasing relative power gain in increasing $|\delta|$ and decreasing $|c|$. Their analysis shows that in the case of a highly persistent regressor, using the Q-test over the t-test provides a comparatively important gain in power. Their numerical refinement of the Bonferroni method provides a substantial further gain in power for the Q-test since this test exploits information about ρ. This makes its confidence interval for β given ρ more sensitive to ρ. Hence, without the refinement the Bonferroni Q-test is too conservative.

4 Monte Carlo simulation

This section, next to the out-of-sample results in section 5, accommodates one of the main contributions of this study. First, Monte Carlo evidence of Campbell and Yogo (2006) is replicated providing reassurance that both parties implemented the method correctly. Next, the Bonferroni Q-test is tested by violating key assumptions about the regressor's persistence. Furthermore, innovations are modeled by increasingly realistic distributions, for which justification in the form of real-world parameter estimates is given.

4.1 Simulation with Gaussian innovations

To confirm the correct implementation of the procedure, Table 1 on the next page replicates Table 3 of Campbell and Yogo (2006) with the same input parameters. For each parameter combination, 10,000 Monte Carlo samples are drawn from a multivariate standard normal distribution with correlation δ. These variates represent the innovations, which are used to simulate sample paths according to system (1, 2). The null hypothesis $H_0 : \beta = 0$ is tested against the alternative $H_1 : \beta > 0$ at the 5% significance level. If the Bonferroni confidence interval lies strictly above zero, the one-sided test is rejected.

Next, slightly different input parameters are chosen. If the results differ qualitatively, the previous findings might be the result of selection bias. As demonstrated by Table 2 on page 10, no such suspicion is warranted. As long as the sample size is at least 100 and ρ is local-to-unity, the Bonferroni Q-test has acceptable finite sample rejection rates. The infeasible Q-test, unsurprisingly, always has rejection rates close to α. The t-test, on the other hand, significantly over-rejects when the innovations are highly correlated and ρ is local-to-unity. It should be noted, however, that the two-sided Bonferroni Q-test tends to under-reject for the chosen parameters.

4.2 Simulation with weakly persistent regressors and Gaussian innovations

As argued by Phillips (2014), the feasible Q-test as derived in section 3 is asymptotically invalid when the largest autoregressive root of the regressor is not local-to-unity. In the following, finite-sample behavior is demonstrated as ρ diverges from unity. Again, 10,000 Monte Carlo simulations are run with bivariate Gaussian innovations. A relatively small sample size is chosen to keep the DF-GLS statistic small in absolute value and thus stay in

the vicinity of the implementation of Campbell and Yogo (2005). Table 3 on page 11 shows that as ρ distances itself from the unity neighborhood the Bonferroni Q-test significantly over-rejects the true null of no predictability. Over-rejection is largest for highly correlated innovations and large $|c|$ where it surpasses 200%.

4.3 Simulation with CCC-GARCH innovations

Campbell and Yogo (2006) evaluate the robustness of the Q-test under a fat-tailed distribution. In particular, they find finite-sample rejection probabilities to be unchanged from the Gaussian case when the innovations follow a Student t-distribution with five degrees of freedom. As shown by Cavaliere (2005), unit root tests may be invalid under conditional heteroskedasticity. Assumptions in the form of equation (23) imply validity of the Q-test under conditionally heteroskedastic innovations only if they are covariance stationary. It is therefore interesting to observe the behavior of the Bonferroni Q-test if innovations are modeled by a generalized autoregressive conditional heteroskedasticity [GARCH] process that approaches covariance nonstationarity. If the regressor is a valuation ratio, it seems sensible that there is a more or less stable correlation between the regressor and stock returns. A multivariate GARCH model that allows such a restriction is the Constant Conditional Correlations GARCH [CCC-GARCH] model of Bollerslev et al. (1990). It is defined as follows. Let $\boldsymbol{z_t}$ be a bivariate standard normal random variable, i.e., $\boldsymbol{z}_t \overset{iid}{\sim} \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_2)$. Then the innovations are modeled as

$$\boldsymbol{w}_t = \boldsymbol{H}_t^{1/2} \boldsymbol{z}_t, \tag{38}$$

Table 1: Finite-sample rejection rates

	Obs	c	ρ	δ	t-test	Bonf. Q-test	Q-test
1	50	0	1.000	-0.95	0.4160	0.0826	0.0483
2	50	0	1.000	-0.75	0.2916	0.0837	0.0515
3	50	-2	0.961	-0.95	0.2714	0.0868	0.0482
4	50	-2	0.961	-0.75	0.2079	0.0881	0.0532
5	50	-20	0.608	-0.95	0.0977	0.1206	0.0515
6	50	-20	0.608	-0.75	0.0840	0.1078	0.0484
7	100	0	1.000	-0.95	0.4217	0.0616	0.0480
8	100	0	1.000	-0.75	0.2930	0.0616	0.0497
9	100	-2	0.980	-0.95	0.2698	0.0587	0.0505
10	100	-2	0.980	-0.75	0.2104	0.0588	0.0489
11	100	-20	0.802	-0.95	0.1063	0.0622	0.0471
12	100	-20	0.802	-0.75	0.0874	0.0514	0.0500
13	250	0	1.000	-0.95	0.4259	0.0476	0.0483
14	250	0	1.000	-0.75	0.2970	0.0506	0.0536
15	250	-2	0.992	-0.95	0.2866	0.0507	0.0481
16	250	-2	0.992	-0.75	0.2092	0.0466	0.0492
17	250	-20	0.920	-0.95	0.1080	0.0406	0.0517
18	250	-20	0.920	-0.75	0.0944	0.0369	0.0501

This Table displays the finite-sample rejection rates of the true null for right-tailed and two-tailed Bonferroni Q-tests of predictability at α=0.05 and α=0.1, respectively. 10,000 Monte Carlo simulations are performed.

Table 2: Finite-sample rejection rates

	Obs	c	ρ	δ	t-test	Bonf. Q-test	Q-test	Bonf. Q-test (two-sided)
1	50	0	1.000	-0.99	0.4442	0.0863	0.0544	0.0864
2	50	0	1.000	-0.90	0.3786	0.0943	0.0506	0.0953
3	50	-1	0.980	-0.99	0.3502	0.0857	0.0486	0.0898
4	50	-1	0.980	-0.90	0.3107	0.0949	0.0489	0.0964
5	50	-10	0.804	-0.99	0.1440	0.0821	0.0512	0.1004
6	50	-10	0.804	-0.90	0.1226	0.0840	0.0472	0.0929
7	50	-15	0.706	-0.99	0.1114	0.0895	0.0485	0.1112
8	50	-15	0.706	-0.90	0.1035	0.0981	0.0468	0.1112
9	150	0	1.000	-0.99	0.4467	0.0508	0.0508	0.0509
10	150	0	1.000	-0.90	0.3804	0.0524	0.0494	0.0554
11	150	-1	0.993	-0.99	0.3584	0.0475	0.0493	0.0611
12	150	-1	0.993	-0.90	0.3137	0.0572	0.0532	0.0625
13	150	-10	0.934	-0.99	0.1510	0.0468	0.0480	0.0720
14	150	-10	0.934	-0.90	0.1309	0.0534	0.0484	0.0672
15	150	-15	0.901	-0.99	0.1273	0.0446	0.0520	0.0819
16	150	-15	0.901	-0.90	0.1133	0.0484	0.0489	0.0695
17	200	0	1.000	-0.99	0.4477	0.0435	0.0478	0.0438
18	200	0	1.000	-0.90	0.3786	0.0501	0.0520	0.0546
19	200	-1	0.995	-0.99	0.3582	0.0449	0.0483	0.0607
20	200	-1	0.995	-0.90	0.3127	0.0542	0.0468	0.0615
21	200	-10	0.950	-0.99	0.1490	0.0401	0.0497	0.0666
22	200	-10	0.950	-0.90	0.1352	0.0486	0.0502	0.0634
23	200	-15	0.925	-0.99	0.1276	0.0369	0.0487	0.0786
24	200	-15	0.925	-0.90	0.1234	0.0482	0.0491	0.0670

This Table displays the finite-sample rejection rates of the true null for right-tailed and two-tailed Bonferroni Q-tests of predictability at α=0.05 and α=0.1, respectively. 10,000 Monte Carlo simulations are performed.

where $\boldsymbol{H_t}$ denotes the covariance matrix at time t. $\boldsymbol{H}_t$ is a positive definite matrix whose Cholesky decomposition factors into

$$\boldsymbol{H}_t^{1/2} = \boldsymbol{D}_t \boldsymbol{R}_t^{1/2}, \tag{39}$$

with $\boldsymbol{D}_t$ being a diagonal matrix with the conditional standard deviations

$$\boldsymbol{D}_t = \begin{bmatrix} \sqrt{h_{1t}^2} & 0 \\ 0 & \sqrt{h_{2t}^2} \end{bmatrix} \tag{40}$$

and $\boldsymbol{R}_t$ being the positive definite conditional correlation matrix

$$\boldsymbol{R}_t = \begin{bmatrix} 1 & \delta_t \\ \delta_t & 1 \end{bmatrix}. \tag{41}$$

Since correlations are assumed to be constant through time, $\boldsymbol{R}_t = \boldsymbol{R}$. The conditional variances

$$h_{it}^2, \quad t = 1, \ldots, T, \quad i = 1, 2$$

Table 3: Finite-sample rejection rates (non local to unity autoregressive root)

	Obs	c	ρ	δ	t-test	Bonf. Q-test	Q-test	Bonf. Q-test (two-sided)
1	50	0	1.000	-0.95	0.4147	0.0946	0.0480	0.0957
2	50	0	1.000	-0.75	0.2810	0.0838	0.0475	0.0870
3	50	0	1.000	-0.50	0.1758	0.0846	0.0536	0.0940
4	50	-5	0.902	-0.95	0.1887	0.0896	0.0499	0.0980
5	50	-5	0.902	-0.75	0.1487	0.0830	0.0522	0.0891
6	50	-5	0.902	-0.50	0.1081	0.0701	0.0512	0.0811
7	50	-10	0.804	-0.95	0.1325	0.0894	0.0543	0.1018
8	50	-10	0.804	-0.75	0.1126	0.0789	0.0462	0.0867
9	50	-10	0.804	-0.50	0.0887	0.0689	0.0510	0.0820
10	50	-20	0.608	-0.95	0.0987	0.1266	0.0506	0.1469
11	50	-20	0.608	-0.75	0.0906	0.1087	0.0506	0.1216
12	50	-20	0.608	-0.50	0.0682	0.0800	0.0443	0.0948
13	50	-30	0.412	-0.95	0.0811	0.1666	0.0455	0.1961
14	50	-30	0.412	-0.75	0.0787	0.1272	0.0500	0.1460
15	50	-30	0.412	-0.50	0.0719	0.1003	0.0523	0.1195

This Table displays the finite-sample rejection rates of the true null for right-tailed and two-tailed Bonferroni Q-tests of predictability at α=0.05 and α=0.1, respectively, with non local to unity autoregressive root. 10,000 Monte Carlo simulations are performed.

in $\boldsymbol{D}_t$ are modeled by independent univariate GARCH processes, i.e.,

$$\begin{aligned} &\epsilon_t = h_t \eta_t, \quad \eta_t \overset{iid}{\sim} N(0,1) \\ &h_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i h_{t-i}^2 \\ &\omega > 0, \quad \alpha_i \geq 0, \quad i = 1, \ldots, q, \quad \beta_i \geq 0, \quad i = 1, \ldots, p. \end{aligned} \tag{42}$$

Simulation results are shown in Table 4.

Table 4: Finite-sample rejection rates (CCC-GARCH(1, 1))

	Obs	c	ρ	δ	$\hat{\delta}$	α_{GARCH}	β_{GARCH}	t-test	BQ	BQ_{2sided}
1	100	-2	0.98	-0.95	-0.6002	0.0999	0.90	0.1819	0.0635	0.0697
2	100	-2	0.98	-0.95	-0.6053	0.0999	0.80	0.1685	0.0633	0.0723
3	100	-2	0.98	-0.95	-0.6080	0.0999	0.50	0.1677	0.0645	0.0744
4	100	-2	0.98	-0.95	-0.6081	0.0999	0.40	0.1663	0.0596	0.0678
5	100	-2	0.98	-0.95	-0.6077	0.0999	0.30	0.1616	0.0600	0.0696
6	100	-2	0.98	-0.95	-0.6077	0.0999	0.20	0.1675	0.0610	0.0680
7	100	-2	0.98	-0.95	-0.6086	0.0999	0.10	0.1741	0.0667	0.0742

This Table displays the finite-sample rejection rates of the true null for right-tailed [BQ] and two-tailed [BQ_{2sided}] Bonferroni Q-tests of predictability at α=0.05 and α=0.1, respectively, with innovations following a CCC-GARCH(1, 1) process. 10,000 Monte Carlo simulations are performed.

4.4 Simulation with common GARCH innovations

In the CCC-GARCH model it has been assumed that there are no volatility transmissions between the returns and the regressor. This assumption seems unrealistic and it reduces the correlation between the residuals as demonstrated by $\hat{\delta}$. There are various multivariate GARCH models suited for incorporating spill-over effects such as the BEKK model of Engle and Kroner (1995). It seems plausible, though, that the variances of both the regressor and the returns share a common stochastic trend which may be imagined as the overall economic uncertainty. This seems reasonable since it is hard to imagine having high variance in returns without having high variance in valuation ratios and vice versa. Parameter estimates of the univariate GARCH model based on monthly returns from 1926 to 2003 are displayed in Table 5. Since $\hat{\alpha}+\hat{\beta} = 0.983$ is close to 1, at which point the covariance stationarity condition is violated, concern of conditional heteroskedasticity of real returns leading to invalid inference using the feasible Q-test is justified.

Table 5: Parameter Estimates of the GARCH(1, 1)

	Estimate	Rob. Std. Error	Rob. t value	Pr(>|t|)
ω	0.0001	0.00002	2.553	0.011
α	0.107	0.020	5.331	0.000
β	0.876	0.020	44.365	0.000

Data is taken from Motohiro Yogo's website. Returns are the CRSP value weighted index log-returns from 1926 to 2003 minus the risk free rate.

Simulation is performed similar as in the CCC-GARCH model but now $h_{1,t} = h_{2,t}$ in equation (40). Results in Table 6 on the following page compared with Table 1 on page 9 indicate that the one-sided Bonferroni Q-test tends to over-reject the true null if innovations are modeled as described above.

4.5 Simulation with common GJR-GARCH innovations

Glosten, Jagannathan and Runkle (1993) [GJR] observe that in financial markets negative shocks contribute disproportionately more to future return volatility than positive shocks. To model the dynamics of real markets more adequately, innovation variances are modeled by the square of a common univariate GJR-GARCH process. This process models conditional variance as

$$h_t^2 = \omega + (\alpha + \gamma S_{t-1})\,\epsilon_{t-1}^2 + \beta h_{t-1}^2, \tag{43}$$

where

$$S_{t-1} = \begin{cases} 1 & \text{if } \epsilon_{t-1} < 0 \\ 0 & \text{if } \epsilon_{t-1} \geq 0. \end{cases}$$

When $\gamma > 0$ negative shocks increase volatility more than positive shocks. The GJR-GARCH process is covariance stationary if $\alpha + \gamma/2 + \beta < 1$. Further, to allow for conditional leptokurtosis η_t in equation (42) now follows a generalized error distribution with parameter θ, i.e.,

$$\eta_t \overset{iid}{\sim} GED(\theta).$$

Table 6: Finite-sample rejection rates (common stochastic trend)

	Obs	c	ρ	δ	α_{GARCH}	β_{GARCH}	t-test	BQ	BQ_{2sided}
1	50	-2	0.961	-0.9500	0.0999	0.90	0.2815	0.0952	0.0990
2	50	-2	0.961	-0.9500	0.0999	0.80	0.2762	0.1009	0.1048
3	50	-2	0.961	-0.9500	0.0999	0.50	0.2706	0.0942	0.0977
4	50	-2	0.961	-0.9500	0.0999	0.40	0.2676	0.0961	0.1004
5	50	-2	0.961	-0.9500	0.0999	0.30	0.2628	0.0932	0.0972
6	50	-2	0.961	-0.9500	0.0999	0.20	0.2704	0.0908	0.0943
7	50	-2	0.961	-0.9500	0.0999	0.10	0.2677	0.0935	0.0984
8	100	-2	0.980	-0.9500	0.0999	0.90	0.3135	0.0733	0.0771
9	100	-2	0.980	-0.9500	0.0999	0.80	0.2812	0.0703	0.0771
10	100	-2	0.980	-0.9500	0.0999	0.50	0.2780	0.0683	0.0747
11	100	-2	0.980	-0.9500	0.0999	0.40	0.2797	0.0676	0.0742
12	100	-2	0.980	-0.9500	0.0999	0.30	0.2758	0.0664	0.0734
13	100	-2	0.980	-0.9500	0.0999	0.20	0.2746	0.0629	0.0703
14	100	-2	0.980	-0.9500	0.0999	0.10	0.2749	0.0656	0.0715

This Table displays the finite-sample rejection rates of the true null for right-tailed [BQ] and two-tailed [BQ_{2sided}] Bonferroni Q-tests of predictability at α=0.05 and α=0.1, respectively, with the variances of innovations sharing a common stochastic trend which is modeled as a univariate GARCH(1, 1) process. 10,000 Monte Carlo simulations are performed.

Parameter estimates for monthly returns of the CRSP value weighted index from 1926 to 2003 are displayed in Table 7. Since $\hat{\alpha} + \hat{\gamma}/2 + \hat{\beta} = 0.958$ real return volatilities are close to nonstationary. Monte Carlo paths are simulated with $\omega = 0.000118$, $\gamma = 0.0999$, $\theta = 1.5$,

Table 7: Parameter Estimates of the GJR-GARCH(1, 1)

	Estimate	Rob. Std. Error	Rob. t value	Pr(>|t|)
ω	0.0001	0.00004	2.982	0.003
α	0.040	0.025	1.599	0.110
β	0.861	0.028	30.881	0.000
γ	0.114	0.045	2.562	0.010
θ	1.515	0.089	17.058	0.000

Data is taken from Motohiro Yogo's website. Returns are the CRSP value weighted index log-returns from 1926 to 2003 minus the risk free rate.

$\alpha = 0$ and letting β increase to 0.95 such that the process approaches nonstationarity. The simulation outcome is displayed in Table 8 on the next page. It is observed that under realistic volatility modeling, the right-tailed Bonferroni Q-test tends to over-reject the true null of no predictability. Even for the larger sample size of 100, the amount of over-rejection of the right-tailed Q-test at 5% significance level surpasses 50%. The null for the smaller sample size is over-rejected more than 100%.

Table 8: Finite-sample rejection rates (common stochastic trend GJR-GARCH)

	Obs	c	ρ	δ	α_{GARCH}	β_{GARCH}	t-test	BQ	BQ_{2sided}
1	50	-2	0.961	-0.9500	0.0000	0.95	0.2836	0.1001	0.1033
2	50	-2	0.961	-0.9500	0.0000	0.90	0.2736	0.0946	0.0986
3	50	-2	0.961	-0.9500	0.0000	0.80	0.2635	0.0952	0.0993
4	100	-2	0.980	-0.9500	0.0000	0.95	0.3054	0.0757	0.0802
5	100	-2	0.980	-0.9500	0.0000	0.90	0.2811	0.0662	0.0727
6	100	-2	0.980	-0.9500	0.0000	0.80	0.2801	0.0685	0.0734

This Table displays the finite-sample rejection rates of the true null for right-tailed [BQ] and two-tailed [BQ_{2sided}] Bonferroni Q-tests of predictability at α=0.05 and α=0.1, respectively, with the variances of innovations sharing a common stochastic trend which is modeled as a univariate GJR-GARCH(1, 1) process. 10,000 Monte Carlo simulations are performed.

5 Empirical results

5.1 Replication of empirical results from Campbell and Yogo (2006)

Table 9 on the following page replicates the key results of Tables 4 and 5 of Campbell and Yogo (2006). The Bonferroni Q-test is implemented as described in section 3.4. The regressors under consideration are the dividend-to-price ratio [ldp] and the earnings-to-price ratio [lep] where earnings, following Shiller (2000), are averaged over a moving 10-year period. Three different sampling frequencies, annual, quarterly, and monthly, are considered. Following Campbell and Yogo (2006), the confidence interval for β is scaled by $\widehat{\sigma}_e/\widehat{\sigma}_u$ such that interpretation is made easier as the coefficient in equation 1 with unit variance innovations. Right-tailed tests with null hypothesis $H_0 : \beta = 0$ against alternative $H_1 : \beta > 0$ are considered where the null is rejected at 5% significance level if the lower bound of β's 90% confidence interval is greater than zero. Apart from rounding errors, the results are nearly the same except for the following discrepancies. First, the DF-GLS statistic for $p > 1$ is higher than reported by Campbell and Yogo (2006). As previously mentioned, the DF-GLS implementation of section 3.4 was verified with a different open-source implementation. Second, based on the BIC, a lower p of 1 was selected for monthly ldp. Investigations showed that the difference in BIC when $p = 1$ compared with $p = 2$ is very small. Thus the different selection is likely due to rounding errors. For all datasets, both regressors are highly negatively correlated with returns. The confidence interval for ρ is very near to unity in all cases, and containing it in some. Hence, the pretest of section 3.2 indicates non-standard null distribution of the t-test for every sampling frequency and regressor, invalidating inference based on t-statistics, which otherwise would be statistically significant. Based on the Bonferroni Q-test, the null hypothesis of no predictability can be rejected at the 5% significance level for ldp and lep at the annual sampling frequency and for all datasets, respectively.

5.2 Out-of-sample results

Based on an extended dataset taken from Amid Goyal's website, only lep on a quarterly frequency can be inferred to be predictive of excess log-returns. This discrepancy may arise from two causes. First, returns in the extended dataset are of the S&P500 instead of the CRSP value-weighted index. Also, due to the lack of availability of the 1-month T-bill rate, all risk free-rate computations are based on the 3-month T-bill in the extended dataset. This introduces minor differences in datasets since Campbell and Yogo (2006) use the 1-month

Table 9: Replicated estimates of model parameters from Campbell and Yogo (2006)

dataset	x	obs	$\hat{\delta}$	CI_ρ	DF-GLS	p	t-stat	pt	$\hat{\beta}$	CI_β
SP_A	lep	123	-0.96	[0.786,0.931]	-2.888	1	2.76	0	0.127	[0.043,0.225]
SP_A	ldp	123	-0.85	[0.940,1.006]	-1.247	3	1.95	0	0.083	[-0.024,0.136]
CRSP_A	lep	77	-0.96	[0.778,0.960]	-2.229	1	2.77	0	0.162	[0.040,0.273]
CRSP_A	ldp	77	-0.72	[0.926,1.010]	-1.033	1	2.53	0	0.158	[0.013,0.186]
CRSP_Q	lep	305	-0.99	[0.944,0.992]	-2.191	1	2.91	0	0.047	[0.011,0.066]
CRSP_Q	ldp	305	-0.94	[0.962,0.999]	-1.696	1	2.06	0	0.034	[-0.009,0.044]
CRSP_M	lep	913	-0.99	[0.985,1.000]	-1.859	1	2.66	0	0.013	[0.001,0.018]
CRSP_M	ldp	913	-0.95	[0.990,1.001]	-1.433	1	1.70	0	0.008	[-0.005,0.010]

Data is taken from Motohiro Yogo's website at https://drive.google.com/file/d/0BzR-ojpYuaFMcnZteHFyWUVIUFU/view. The regressors are the 10-year moving average earnings to current price ratio [lep] and the dividend to price ratio [ldp] in logs. Observations are recorded on a monthly [CRSP M], quarterly [CRSP Q], and annual basis [CRSP A] for CRSP (1926– 2002). For S&P (1880– 2002) data is only available on an annual basis [SP A]. Stock returns are the SP 500 value weighted index log-returns from 1880 to 2003 minus the risk free rate and the CRSP value weighted index log-returns from 1926 to 2003 minus the risk free rate. p denotes the optimal number of lags selected by BIC. pt is a boolean denoting whether the pretest of section 3.2 rejects the null of actual size greater than 0.075. CI_ρ and CI_β denote the confidence interval for ρ and β, respectively.

T-bill as a risk-free proxy for their monthly regressions. Second, as will be seen in Table 11 on the next page, predictability seems to have been weakening in recent years. Hence, a time-extended sample will show less evidence of it.

Table 10: Replicated estimates of model parameters from Campbell and Yogo (2006) with an extended dataset

dataset	x	obs	$\hat{\delta}$	CI_ρ	DF-GLS	p	t-stat	pt	$\hat{\beta}$	CI_β
CRSP_A	lep	92	-0.97	[0.827,0.979]	-2.092	1	2.12	0	0.114	[-0.01,0.180]
CRSP_A	ldp	92	-0.86	[0.875,0.986]	-1.731	1	0.97	0	0.042	[-0.069,0.107]
CRSP_Q	lep	365	-0.98	[0.973,1.002]	-1.515	1	3.16	0	0.048	[0.001,0.039]
CRSP_Q	ldp	365	-0.95	[0.976,1.002]	-1.386	1	1.82	0	0.023	[-0.012,0.026]
CRSP_M	lep	1104	-0.99	[0.993,1.002]	-1.343	1	2.40	0	0.010	[-0.002,0.009]
CRSP_M	ldp	1104	-0.98	[0.993,1.002]	-1.221	1	1.20	0	0.004	[-0.005,0.006]

Data is taken from Amit Goyal's Website. Stock returns are the SP 500 index log-returns from 1926 to 2017 from the Center for Research in Security Press (CRSP) minus the rolled over 3-month T-bill rate. lep is the log 10 year moving average earnings/price ratio (1926 to 2017). ldp is the log dividend/price ratio (1926 to 2017). p denotes the optimal number of lags selected by BIC. pt is a boolean denoting whether the pretest of section 3.2 rejects the null of actual size greater than 0.075. CI_ρ and CI_β denote the confidence interval for ρ and β, respectively.

Since the original publication of Campbell and Yogo (2006) a sufficiently large amount of data has been generated at the quarterly and monthly sampling frequencies that merits a standalone out-of-sample analysis. Table 11 on the following page displays model parameter estimates since 2004 until the last available data point in 2017. This relatively small dataset allows no rejection of the null hypothesis at the 5% significance level for any of the regressors and sampling frequencies considered. Both regressors are still highly persistent at all sampling frequencies, demonstrated by the confidence interval for ρ containing unity in all cases.

Table 11: Out-of-sample estimates of model parameters

dataset	x	obs	$\hat{\delta}$	CI_ρ	DF-GLS	p	t-stat	pt	$\hat{\beta}$	CI_β
CRSP_Q	lep	53	-0.99	[0.792,1.023]	-1.607	1	1.48	0	0.101	[-0.059,0.202]
CRSP_Q	ldp	53	-0.97	[0.778,1.001]	-1.708	1	0.86	0	0.060	[-0.134,0.142]
CRSP_M	lep	168	-0.99	[0.955,1.014]	-1.211	1	1.02	0	0.020	[-0.025,0.041]
CRSP_M	ldp	168	-0.98	[0.961,1.012]	-1.091	1	0.59	0	0.012	[-0.052,0.012]

Data is taken from Amit Goyal's Website. Stock returns are the SP 500 index log-returns from 2004 to 2017 from the Center for Research in Security Press (CRSP) minus the rolled over 3-month T-bill rate. lep is the log 10 year moving average earnings/price ratio (2004 to 2017). ldp is the log dividend/price ratio (2004 to 2017). p denotes the optimal number of lags selected by BIC. pt is a boolean denoting whether the pretest of section 3.2 rejects the null of actual size greater than 0.075. CI_ρ and CI_β denote the confidence interval for ρ and β, respectively.

6 Conclusion

The Bonferroni Q-test is an asymptotically valid test of predictability if the regressor's largest autoregressive root is local-to-unity and the innovations of regressor and regressand are bivariate Gaussian. It has been shown in the original work by Campbell and Yogo (2006) that the Bonferroni Q-test has important power advantages over the Bonferroni t-test. This study confirms acceptable finite-sample rejection rates when innovations are Gaussian and the assumptions of the Bonferroni Q-test are fulfilled. When the regressor is nonlocal-to-unity, however, the Q-test can over-reject the null in finite samples. Further, if innovations are modeled by a covariance nonstationarity approaching GARCH process, the right-tailed test again tends to over-reject the null of no predictability. The practical importance of this effect is demonstrated by GARCH parameter estimates based on S&P 500 returns. Previous evidence of predictability of the earnings-to-price ratio and dividend-to-price ratio has been replicated. Some doubt, however, is casted by the Monte Carlo evidence that finds over-rejection of the null for typical variance dynamics of real markets. Out-of-sample evidence, based on data emerged since the first publication of the method, is eroding. In particular, both regressors are not found statistically significant at the 5% level at the monthly as well as the quarterly sampling frequency. In all cases, the pretest cannot reject the null of unacceptable size distortion of the t-test. Hence, inference based on the t-test is invalid. Campbell and Yogo (2005)

References

Bollerslev, T. et al. (1990). Modelling the coherence in short-run nominal exchange rates: a multivariate generalized arch model, *Review of Economics and statistics* **72**(3): 498–505.

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

Campbell, J. Y. and Yogo, M. (2006). Efficient tests of stock return predictability, *Journal of financial economics* **81**(1): 27–60.

Cavaliere, G. (2005). Unit root tests under time-varying variances, *Econometric Reviews* **23**(3): 259–292.

Elliott, G., Rothenberg, T. J. and Stock, J. H. (1996). Efficient tests for an autoregressive unit root, *Econometrica* **64**(4): 813–836.

Elliott, G. and Stock, J. H. (1994). Inference in time series regression when the order of integration of a regressor is unknown, *Econometric theory* **10**(3-4): 672–700.

Engle, R. F. and Kroner, K. F. (1995). Multivariate simultaneous generalized arch, *Econometric theory* **11**(1): 122–150.

Pfaff, B. (2008). *Analysis of Integrated and Cointegrated Time Series with R*, second edn, Springer, New York. ISBN 0-387-27960-1.
URL: *http://www.pfaffikus.de*

Phillips, P. C. (1987). Towards a unified asymptotic theory for autoregression, *Biometrika* **74**(3): 535–547.

Phillips, P. C. (2014). On confidence intervals for autoregressive roots and predictive regression, *Econometrica* **82**(3): 1177–1195.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Shiller, R. C. (2000). Irrational exuberance, *Philosophy and Public Policy Quarterly* **20**(1): 18–23.

Stock, J. H. (1991). Confidence intervals for the largest autoregressive root in us macroeconomic time series, *Journal of monetary economics* **28**(3): 435–459.