Robust Bond Risk Premia*

Michael D. Bauer† and James D. Hamilton‡

April 16, 2015
Revised: December 31, 2016

Abstract

A consensus has recently emerged that variables beyond the level, slope, and curvature of the yield curve can help predict bond returns. This paper shows that the statistical tests underlying this evidence are subject to serious small-sample distortions. We propose more robust tests, including a novel bootstrap procedure specifically designed to test the spanning hypothesis. We revisit the evidence in six published studies, find most rejections of the spanning hypothesis to be spurious, and conclude that the current consensus is wrong. The evidence against the spanning hypothesis is much weaker than appears from the statistical evidence in these studies.

Keywords: yield curve, spanning, return predictability, robust inference, bootstrap
JEL Classifications: E43, E44, E47

*The views expressed in this paper are those of the authors and do not necessarily reflect those of others in the Federal Reserve System. We thank Anna Cieslak, John Cochrane, Greg Duffee, Graham Elliott, Robin Greenwood, Helmut Lütkepohl, Ulrich Müller, Hashem Pesaran and Glenn Rudebusch for useful suggestions, conference participants and discussants at the 7th Annual Volatility Institute Conference at the NYU Stern School of Business, the NBER Summer Institute 2015, the Federal Reserve System Macro Conference 2015 in Cleveland, the Federal Reserve Bank of San Francisco Fixed Income Research Conference 2015, the CESifo Conference on Macro, Money and International Finance 2016 in Munich, the Spring 2016 NBER Asset Pricing Workshop in Chicago, and the Western Finance Association Conference 2016 in Park City, as well as seminar participants at the Federal Reserve Bank of Boston, the Free University of Berlin, and the University of Hamburg for helpful comments, Javier Quintero and Simon Riddell for excellent research assistance, and Anh Le, Marcel Priebsch, Serena Ng and Anna Cieslak for the data used in their papers.

†Federal Reserve Bank of San Francisco, 101 Market St MS 1130, San Francisco, CA 94105, phone: 415-974-3299, e-mail: michael.bauer@sf.frb.org

‡University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0508, phone: 858-534-5986, e-mail: jhamilton@ucsd.edu

1 Introduction

Identifying the contribution of risk premia to long-term interest rates is crucial for monetary policy, investment strategy, and interpreting historical episodes such as the unprecedented low interest rates since 2008. Since the risk premium is just the difference between the current long rate and the expected average value of future short rates, the core question for estimating risk premia is how to construct short-rate expectations. Is it sufficient to consider the current yield curve, or should estimates incorporate additional information such as macroeconomic variables? This is the question we address in this paper.

A powerful argument can be made that the current yield curve itself should contain most (if not all) information useful for forecasting future interest rates and bond returns. Investors use information at time t—which we can summarize by a state vector z_t—to forecast future short-term interest rates and determine bond risk premia. Hence current yields are a function of z_t. Most modern macro-finance models have the implication that we would be able to back out the state vector z_t from the yield curve, in which case current yields themselves would be the only variables necessary to forecast interest rates and calculate risk premia.[1] Furthermore, it has long been recognized that the first three principal components (PCs) of yields, commonly labeled level, slope, and curvature, provide an excellent empirical summary of the entire yield curve (Litterman and Scheinkman, 1991), as they explain almost all of the cross-sectional variance of observed yields. This motivates what we term the "spanning hypothesis," a very practical and empirically focused interpretation of the question posed above: Do these three observed variables alone capture all the information that is useful for forecasting future yields and estimating bond risk premia? This question has been the focus of a number of influential studies,[2] and recent literature reviews by Gürkaynak and Wright (2012) and Duffee (2013a) identify it as a central issue in macro-finance. If the spanning hypothesis holds, this greatly simplifies estimation of monetary policy expectations and bond risk premia, as this estimation does not require any data or models involving macroeconomic series, other asset prices or quantities, volatilities, or survey expectations. Instead, all the necessary information is the shape of the current yield curve, summarized by its level, slope, and curvature.

Essentially all asset pricing models naturally imply some version of spanning of z_t by yields, but the spanning hypothesis as we define it here could still be violated for a number of reasons.

[1] For a detailed argument, see Duffee (2013b).

[2] As noted below, some of these studies have posed the question using a one- or two-variable summary of the information in the current yield curve rather than the first three principal components.

First, yields may of course depend on more than three state variables.[3] Second, even if three linear combinations of model-implied yields span z_t, this might be difficult to exploit in practice due to measurement error. In particular, Duffee (2011b) demonstrated that if the effects of some elements of z_t on yields nearly offset each other, those components will be very difficult to infer from current observed yields alone.[4] Third, the presence of non-linearities or structural breaks in the mapping from z_t into yields would naturally lead to a violation of our spanning hypothesis. Our paper does not address these theoretical possibilities, and instead focuses squarely on the empirical question whether additional variables like inflation are necessary to include for forecasting bond returns, or whether their implications for forecasting are already incorporated in the first three principal components of the yield curve.

It is also worth emphasizing under our spanning hypothesis macroeconomic variables can still be important determinants of interest rates and risk premia. Both theoretical models and empirical studies suggest important links between macroeconomic variables and the yield curve, and the literature has made much progress since the influential studies of Fama and Bliss (1987) and Campbell and Shiller (1991) that focused exclusively on the links between yields and risk premia.[5] Although macroeconomic variables are undoubtedly important drivers of yields and risk premia, our question here is what variables should be used for the *estimation* of these risk premia.

There is a growing consensus in the literature that the spanning hypothesis as we have defined it can be rejected by the observed data. This evidence comes from predictive regressions for bond returns on various predictors, controlling for information in the current yield curve. The variables that have been found to contain additional predictive power in such regressions include measures of economic growth and inflation (Joslin et al., 2014), factors inferred from a large set of macro variables (Ludvigson and Ng, 2009, 2010), long-term trends in inflation or inflation expectations (Cieslak and Povala, 2015), higher-order (fourth and fifth) PCs of bond yields (Cochrane and Piazzesi, 2005), the output gap (Cooper and Priestley, 2008), and measures of Treasury bond supply (Greenwood and Vayanos, 2014). These results suggest that there might be unspanned or hidden information that is not captured by the current yield curve but that is useful for forecasting.

But these predictive regressions have a number of problematic features. The true predictive variables under the null hypothesis are necessarily correlated with lagged forecast errors

[3]For example, in Bansal and Shaliastovich (2013) yields are functions of four state variables.

[4]Furthermore, Cieslak and Povala (2015) and Bauer and Rudebusch (2016) noted that in conventional affine yield-curve models, even small measurement errors can make it impossible to recover z_t from observed yields alone.

[5]Some important examples include Campbell and Cochrane (1999), Diebold et al. (2006), Bikbov and Chernov (2010), Rudebusch and Swanson (2012), and Bansal and Shaliastovich (2013).

because they summarize the information in the current yield curve. As a consequence they violate the condition of strict econometric exogeneity. In addition, the predictive variables are typically highly persistent. We show that this leads to substantial "standard error bias" in samples of the size commonly studied: estimated standard errors are too small, leading to spurious rejection of the spanning hypothesis even though it is true. This problem inherent in all tests of the spanning hypothesis has to our knowledge not previously been recognized. Mankiw and Shapiro (1986) and Stambaugh (1999) documented small-sample coefficient bias in predictive regressions with a persistent regressor that is not strictly exogenous.[6] By contrast, in our setting there is no coefficient bias pertaining to the additional predictors, and instead a downward bias of the estimated standard errors distorts the results of conventional inference. An additional problem is that the common predictive regressions are estimated in monthly data but with an annual excess bond return as the dependent variable, and the presence of overlapping observations introduces substantial serial correlation in the prediction errors. As a result, standard errors are even less reliable, and regression R^2 are harder to interpret. We demonstrate that the procedures commonly used for inference about the spanning hypothesis do not adequately address these issues and are subject to serious small-sample distortions.

We propose three procedures that researchers can use to obtain more robust inference in these predictive regressions. The first is a novel parametric bootstrap that generates data samples under the spanning hypothesis. We calculate the first three PCs of the observed set of yields and summarize their dynamics with a VAR fit to the observed PCs. Then we use a residual bootstrap to resample the PCs, and construct bootstrapped yields by multiplying the simulated PCs by the historical loadings of yields on the PCs and adding a small Gaussian measurement error. Thus by construction no variables other than the PCs are useful for predicting yields or returns in our generated data. We then fit a separate VAR to the proposed additional explanatory variables alone, and generate bootstrap samples for the predictors from this VAR. Using our novel bootstrap procedure, we can calculate the properties of any regression statistic under the spanning hypothesis.[7] This calculation demonstrates that the conventional tests reject the true null much too often. We show that the tests employed in published studies, which are intended to have a nominal size of five percent, have a true size between 8 and 61%. We then ask whether under the null it would be possible to observe similar

[6] Cavanagh et al. (1995) and Campbell and Yogo (2006) considered this problem using local-to-unity asymptotic theory.

[7] Our procedure notably differs from the bootstrap approach commonly employed in this literature, which generates artificial data under the expectations hypothesis, such as Bekaert et al. (1997), Cochrane and Piazzesi (2005), Ludvigson and Ng (2009, 2010), and Greenwood and Vayanos (2014).

patterns of predictability as researchers have found in the data. We find that in most of the studies this is indeed the case, meaning that much of the above-cited evidence against the spanning hypothesis might be spurious. These results provide a strong caution against using conventional tests for inference about bond risk premia, and we recommend that researchers instead use the bootstrap procedure proposed in this paper.

A second procedure that we propose for inference in this context is the Ibragimov and Müller (2010) approach to robust testing. This splits the sample into subsamples, estimates coefficients separately in each of these, and then performs a simple t-test on the coefficients across subsamples. We show that this approach has excellent size and power properties for tests of the spanning hypothesis. Applying it to the predictive regressions for excess bond returns studied in the literature, we find little to no evidence that variables other than the current yield curve are helpful for forecasting returns.

Finally, we take advantage of the data that have arrived since publication of these studies. This allows us to re-estimate the proposed models in new data, and to evaluate whether they improve true out-of-sample forecasts. We find that the proposed additional predictors are rarely helpful in the new data, reinforcing the case that the apparent strength of the in-sample evidence may be an artifact of the small-sample problems we highlight.

After revisiting the evidence in the six influential papers cited above we draw two main conclusions: First, conventional methods of inference are extremely unreliable in these predictive regressions, because they often suggest that variables are relevant for bond risk premia which in truth are irrelevant. New approaches for robust inference are needed, and we propose three in this paper. Second, when reconsidered with more robust methods for inference, the evidence against the spanning hypothesis appears weaker and much less robust than would appear from the published results.

Our paper is related to other studies that criticize return predictability in finance. Ferson et al. (2003) raised the possibility of finding spurious predictability if a persistent component of stock returns is unobserved. Welch and Goyal (2008) questioned the predictability of stock returns based on the observation that it largely disappears in out-of-sample analysis. Ang and Bekaert (2007) showed that the commonly employed Newey-West standard errors are not reliable for inference about stock return predictability at long horizons. Lewellen et al. (2010) showed that estimating factor models for equity risk premia can lead to spuriously high R^2 for truly irrelevant risk factors. Our paper parallels these studies by also documenting that published evidence on predictability and risk premia is fraught with serious econometric problems and appears to be partially spurious. But our work is distinct in that we describe a new, different econometric issue and focus on evidence on unspanned risks in bond returns

instead of predictability of stock returns. Bekaert et al. (1997) and Bekaert and Hodrick (2001) documented Stambaugh bias in predictive regressions for bond returns. Our paper shows that this bias matters not only for tests of the expectations hypothesis but more generally, and demonstrates exactly the effects on coefficient bias and standard error bias in the case of tests of the spanning hypothesis.

The paper is structured as follows: In Section 2 we describe the econometric problems of predictive regressions for bond returns, and propose practical solutions. In Sections 3 through 7 we revisit each of the prominent published studies that appear to show evidence against the spanning hypothesis. Section 8 concludes. An appendix includes theoretical derivations and additional empirical results.

2 Inference about the spanning hypothesis

Evidence against the spanning hypothesis typically comes from regressions of the form

$$y_{t+h} = \beta_1' x_{1t} + \beta_2' x_{2t} + u_{t+h}, \tag{1}$$

where the dependent variable y_{t+h} is the return or excess return on a long-term bond (or portfolio of bonds), x_{1t} and x_{2t} are vectors containing K_1 and K_2 predictors, respectively, and u_{t+h} is a forecast error. The predictors x_{1t} contain a constant and the information in the yield curve, typically captured by the first three PCs of observed yields, i.e., level, slope, and curvature. The null hypothesis of interest is

$$H_0: \quad \beta_2 = 0,$$

which says that the relevant predictive information is spanned by the information in the yield curve and that x_{2t} has no additional predictive power. A key feature of these regressions is that because the regressors in x_{1t} capture information in the current yield curve they are necessarily strongly correlated with the surprise returns, u_t, and hence not strictly exogenous. The predictors are also typically very persistent. We show in Sections 2.1-2.3 that this gives rise to a previously unrecognized problem, "standard error bias," that causes tests to reject the null hypothesis much too often. In addition, empirical work typically tries to predict returns over $h = 12$ months. Section 2.4 discusses how such use of overlapping returns, and the resulting serial correlation in u_{t+h}, leads to additional econometric problems. We provide solutions to these problems in Section 2.5.

2.1 The source of standard error bias

Here we explain the intuition for standard error bias in the case when $h = 1$ and u_{t+1} is white noise. According to the Frisch-Waugh Theorem, the OLS estimate of β_2 in (1) can always be viewed as having been obtained in two steps. First we regress x_{2t} on x_{1t} and calculate the residuals $\tilde{x}_{2t} = x_{2t} - \hat{A}_T x_{1t}$ for $\hat{A}_T = \left(\sum_{t=1}^T x_{2t}x_{1t}'\right)\left(\sum_{t=1}^T x_{1t}x_{1t}'\right)^{-1}$. Second we regress y_{t+1} on $\tilde{x}_{2t}$. The coefficient on $\tilde{x}_{2t}$ in this regression will be numerically identical to the coefficient on x_{2t} in the original regression (1).[8] The standard Wald statistic for a test about β_2 can be expressed as

$$W_T = \left(\sum\nolimits_{t=1}^T u_{t+1}\tilde{x}_{2t}'\right)\left(s^2\sum\nolimits_{t=1}^T \tilde{x}_{2t}\tilde{x}_{2t}'\right)^{-1}\left(\sum\nolimits_{t=1}^T \tilde{x}_{2t}u_{t+1}\right) \tag{2}$$

for $s^2 = (T - K_1 - K_2)^{-1}\sum_{t=1}^T (y_{t+1} - b_1'x_{1t} - b_2'x_{2t})^2$ and b_1 and b_2 the OLS estimates from (1). The validity of this test depends on whether W_T is approximately $\chi^2(K_2)$. If x_{1t} and x_{2t} are stationary and ergodic, the estimate $\hat{A}_T$ will converge to the true value $A = E(x_{2t}x_{1t}')\left[E(x_{1t}x_{1t}')\right]^{-1}$. In that case the sampling uncertainty from the first step is asymptotically irrelevant and W would have the same asymptotic distribution as if we replaced $\tilde{x}_{2t}$ with $x_{2t} - Ax_{1t}$, which gives rise to the standard result for stationary regressors that $W_T \xrightarrow{d} \chi^2(K_2)$.

If, however, the regressors are highly persistent, a regression of x_{2t} on x_{1t} behaves like a spurious regression. For example, if x_{1t} and x_{2t} are unit-root processes, the value of $\hat{A}_T$ is not tending to some constant but instead to a random variable $\tilde{A}$ that is different in every sample, even as the sample size T approaches infinity. If x_{1t} was strictly exogenous, this would not affect the asymptotic distribution of W_T. But in tests of the spanning hypothesis x_{1t} is necessarily correlated with u_t, and due to this lack of strict exogeneity $\sum_{t=1}^T \tilde{x}_{2t}u_{t+1}$ has a nonstandard limiting distribution with variance that is larger[9] than that of $\sum_{t=1}^T x_{2t}u_{t+1}$. By contrast, OLS hypothesis tests act as if the variance of $\sum_{t=1}^T \tilde{x}_{2t}u_{t+1}$ is *smaller* than than that of $\sum_{t=1}^T x_{2t}u_{t+1}$, since $\sum_{t=1}^T \tilde{x}_{2t}\tilde{x}_{2t}'$ is smaller by construction in every sample than $\sum_{t=1}^T x_{2t}x_{2t}'$. Therefore OLS standard errors are necessarily too small, W_T does not converge to a $\chi^2(K_2)$ distribution, and conventional t- or F-tests about the value of β_2 in (1) will reject more often than they should.[10]

[8] We provide a proof of this and other statements in this section in Appendix A.1.

[9] More formally, the difference between the two matrices is a positive definite matrix.

[10] In Appendix A.1 we go through this argument in more detail, and provide additional proofs. Note also that we have focused on conventional OLS standard errors that assume conditional homoskedasticity, but very similar reasoning applies when White's heteroskedasticity-robust standard errors are used.

2.2 A canonical example

In this section we explore the size of these effects in a canonical example, using first local-to-unity asymptotics and then small-sample simulations based on the model

$$y_{t+1} = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_{t+1} \tag{3}$$

where x_{1t} and x_{2t} are scalar AR(1) processes

$$x_{1,t+1} = \mu_1 + \rho_1 x_{1t} + \varepsilon_{1t} \tag{4}$$

$$x_{2,t+1} = \mu_2 + \rho_2 x_{2t} + \varepsilon_{2t} \tag{5}$$

with ε_{it} martingale-difference sequences and $x_{i0} = 0$. Our interest is in what happens when the persistence parameters ρ_i are close to unity. We first focus on the case without drift in these processes ($\mu_1 = \mu_2 = 0$). We assume that x_{1t} has correlation δ with the lagged value of u_{t+1}, whereas x_{2t} is uncorrelated with both x_{1t} and u_t:

$$E\begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ u_t \end{bmatrix} \begin{bmatrix} \varepsilon_{1s} & \varepsilon_{2s} & u_s \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \delta\sigma_1\sigma_u \\ 0 & \sigma_2^2 & 0 \\ \delta\sigma_1\sigma_u & 0 & \sigma_u^2 \end{bmatrix} \quad \text{if } t = s$$
$$= 0 \qquad \text{otherwise.}$$

Thus in this example when $\beta_2 = 0$, the variable x_{2t} has nothing to do with either x_{1s} or y_s for any t or s.

One device for seeing how the results in a finite sample of some particular size T differ from those predicted by conventional first-order asymptotics is to use a local-to-unity specification as in Phillips (1988) and Cavanagh et al. (1995):

$$x_{i,t+1} = (1 + c_i/T)x_{it} + \varepsilon_{i,t+1} \quad i = 1, 2. \tag{6}$$

For example, if our data come from a sample of size $T = 100$ when $\rho_i = 0.99$, the idea is to approximate the small-sample distribution of regression statistics by the asymptotic distribution obtained by taking $c_i = -1$ in (6) and letting $T \to \infty$.[11] The local-to-unity

[11] It is well known that approximations from such local-to-unity asymptotics are substantially better than those based on conventional first-order asymptotics which take $T \to \infty$ and treat $\rho_i = 0.99$ as a constant; see for example Chan (1988) and Nabeya and Sørensen (1994).

asymptotics turn out to be described by Ornstein-Uhlenbeck processes. For example

$$T^{-2}\sum_{t=1}^{T}(x_{it}-\bar{x}_i)^2 \Rightarrow \sigma_i^2\int_0^1 [J_{c_i}^{\mu}(\lambda)]^2 d\lambda$$

where $\Rightarrow$ denotes weak convergence as $T\to\infty$ and

$$J_{c_i}^{\mu}(\lambda)=J_{c_i}(\lambda)-\int_0^1 J_{c_i}(s)ds \qquad J_{c_i}(\lambda)=c_i\int_0^{\lambda} e^{c_i(\lambda-s)}W_i(s)ds+W_i(\lambda) \qquad i=1,2$$

with $W_1(\lambda)$ and $W_2(\lambda)$ denoting independent standard Brownian motion.[12]

We show in Appendix A.2 that under local-to-unity asymptotics the coefficient from a regression of x_{2t} on x_{1t} has the following limiting distribution:

$$A_T=\frac{\sum(x_{1t}-\bar{x}_1)(x_{2t}-\bar{x}_2)}{\sum(x_{1t}-\bar{x}_1)^2}\Rightarrow\frac{\sigma_2\int_0^1 J_{c_1}^{\mu}(\lambda)J_{c_2}^{\mu}(\lambda)d\lambda}{\sigma_1\int_0^1[J_{c_1}^{\mu}(\lambda)]^2d\lambda}\equiv(\sigma_2/\sigma_1)A, \tag{7}$$

where the last equality defines the random variable A. Under first-order asymptotics the influence of A_T would vanish as the sample size grows. But using local-to-unity asymptotics we see that A_T behaves similarly to the coefficient in a spurious regression and does not converge to zero—the true correlation between x_{1t} and x_{2t} in this setting—but to a random variable that differs across samples. The implication is that the t-statistic for b_2 can have a small-sample distribution that is very poorly approximated using first-order asymptotics. Appendix A.2 demonstrates that this t-statistic has a local-to-unity asymptotic distribution under the null hypothesis that is given by

$$\frac{b_2-\beta_2}{\{s^2/\sum\tilde{x}_{2t}^2\}^{1/2}}\Rightarrow\delta Z_1+\sqrt{1-\delta^2}Z_0 \tag{8}$$

$$Z_1=\frac{\int_0^1 K_{c_1,c_2}(\lambda)dW_1(\lambda)}{\left\{\int_0^1[K_{c_1,c_2}(\lambda)]^2d\lambda\right\}^{1/2}} \qquad Z_0=\frac{\int_0^1 K_{c_1,c_2}(\lambda)dW_0(\lambda)}{\left\{\int_0^1[K_{c_1,c_2}(\lambda)]^2d\lambda\right\}^{1/2}} \qquad K_{c_1,c_2}(\lambda)=J_{c_2}^{\mu}(\lambda)-AJ_{c_1}^{\mu}(\lambda)$$

for $s^2=(T-3)^{-1}\sum(y_{t+1}-b_0-b_1x_{1t}-b_2x_{2t})^2$ and $W_i(\lambda)$ independent standard Brownian processes for $i=0,1,2$. Conditional on the realizations of $W_1(.)$ and $W_2(.)$, the term Z_0 will be recognized as a standard Normal variable, and therefore Z_0 has an unconditional $N(0,1)$ distribution as well.[13] In other words, if x_{1t} is strictly exogenous ($\delta=0$) then the OLS t-test

[12] When $c_i=0$, (6) becomes a random walk and the local-to-unity asymptotics simplify to the standard unit-root asymptotics involving functionals of Brownian motion as a special case: $J_0(\lambda)=W(\lambda)$.

[13] The intuition is that for $v_{0,t+1}\sim$ i.i.d. $N(0,1)$ and $K=\{K_t\}_{t=1}^{T}$ any sequence of random variables

of $\beta_2 = 0$ will be valid in small samples even with highly persistent regressors. By contrast, if $\delta \neq 0$ the random variable Z_1 comes into play, which has a nonstandard distribution because the term $dW_1(\lambda)$ in the numerator is not independent of the denominator. In particular, Appendix A.2 establishes that $\text{Var}(Z_1) > 1$. Moreover Z_1 and Z_0 are uncorrelated with each other.[14] Therefore the t-statistic in (8) in general has a non-standard distribution with variance $\delta^2\text{Var}(Z_1) + (1-\delta^2)1 > 1$ which is monotonically increasing in $|\delta|$. This shows that whenever x_{1t} is correlated with u_t ($\delta \neq 0$) and x_{1t} and x_{2t} are highly persistent, in small samples the t-test of $\beta_2 = 0$ will reject too often when H_0 is true.[15]

We can quantify the magnitude of these effects using a simulation study. We generate values for x_{1t} and x_{2t} by drawing ε_{1t} and ε_{2t} as i.i.d. Gaussian random variables with $\sigma_1 = \sigma_2 = 1$, using $\mu_1 = \mu_2 = 0$ and different values of $\rho_1 = \rho_2 = \rho$, starting from $x_{10} = x_{20} = 0$. We generate $y_t = u_t = \delta\varepsilon_{1t} + \sqrt{1-\delta^2}v_t$ where v_t is a standard normal random variable. Hence, in our data-generating process (DGP) we have $\beta_0 = \beta_1 = \beta_2 = 0$, $\sigma_u = 1$, and $Corr(u_t, \varepsilon_{1t}) = \delta$. We simulate 1,000,000 samples, estimate regression (3) in each sample, and study the small-sample behavior of the t-statistic for the test of $H_0 : \beta_2 = 0$, using OLS standard errors and critical values from the Student-t distribution with 97 degrees of freedom.[16] In addition, we also draw from the local-to-unity asymptotic distribution of the t-statistic given in equation (8) using well-known Monte Carlo methods.[17]

The first panel of Table 1 shows the results of this exercise for different values of ρ and δ. If the regressors are either strictly exogneous ($\delta = 0$) or not serially correlated ($\rho = 0$), the true

that is independent of v_0, $\sum_{t=1}^{T} K_t v_{0,t+1}$ has a distribution conditional on K that is $N(0, \sum_{t=1}^{T} K_t^2)$ and $\sum_{t=1}^{T} K_t v_{0,t+1} / \sqrt{\sum_{t=1}^{T} K_t^2} \sim N(0,1)$. Multiplying by the density of K and integrating over K gives the identical unconditional distribution, namely $N(0,1)$. For a more formal discussion in the current setting, see Hamilton (1994, pp. 602-607).

[14] The easiest way to see this is to note that conditional on $W_1(.)$ and $W_2(.)$ the product has expectation zero, so the unconditional expected product is zero as well.

[15] Expression (8) can be viewed as a straightforward generalization of result (2.1) in Cavanagh et al. (1995) and expression (11) in Campbell and Yogo (2006). In their case the explanatory variable is $x_{1,t-1} - \bar{x}_1$ which behaves asymptotically like $J_{c_1}^{\mu}(\lambda)$. The component of u_t that is correlated with ε_{1t} leads to a contribution to the t-statistic given by the expression that Cavanagh et al. (1995) refer to as τ_{1c}, which is labeled as τ_c/κ_c by Campbell and Yogo (2006). This variable is a local-to-unity version of the Dickey-Fuller distribution with well-known negative bias. By contrast, in our case the explanatory variable is $\tilde{x}_{2.t-1} = x_{2,t-1} - A_T x_{1,t-1}$ which behaves asymptotically like $K_{c_1,c_2}(\lambda)$. Here the component of u_t that is correlated with ε_{1t} leads to a contribution to the t-statistic given by Z_1 in our expression (8). Unlike the Dickey-Fuller distribution, Z_1 has mean zero, so that there is no bias in b_2.

[16] Regular OLS standard errors are the correct choice in this simulation setup as the errors are not serially correlated ($h = 1$) and there is no heteroskedasticity.

[17] We simulate samples of size $\tilde{T}$ from near-integrated processes with $c_1 = c_2 = T(\rho - 1)$ and approximate the integrals in (8) using Rieman sums—see, for example, Chan (1988), Stock (1991), and Stock (1994). We use $\tilde{T} = 1000$, since even moderate sample sizes generally yield accurate approximations to the limiting distribution (Stock, 1991, uses $\tilde{T} = 500$).

size of the t-test of $\beta_2 = 0$ is equal to the nominal size of five percent. If, however, both $\rho \neq 0$ and $\delta \neq 0$, the true size exceeds the nominal size, and this size distortion increases in ρ and δ.[18] In the presence of high persistence the true size of this t-test can be quite substantially above the nominal size: For $\rho = 0.99$ and $\delta = 1$, the true size is around 15 percent, meaning that we would reject the true null hypothesis more than three times as often as we should. The size calculations are, not surprisingly, very similar for the small-sample simulations and the local-to-unity asymptotic approximations.

Figure 1 plots the size of the t-test for the case with $\delta = 1$ for sample sizes from $T = 50$ to 1000, based on the local-to-unity approximation.[19] When $\rho < 1$, the size distortion decreases with the sample size. For example for $\rho = 0.99$ the size decreases from 15 percent to about 9 percent. In contrast, when $\rho = 1$ the size distortions are not affected by the sample size, as indeed in this case the non-Normal distribution corresponding to (8) with $c_i = 0$ governs the distribution for arbitrarily large T.

The reason for the size distortions when testing $\beta_2 = 0$ is not coefficient bias. Table 1 shows that b_1 is downward biased but b_2 is unbiased. Instead, the reason that the t-test rejects too often is standard error bias, as the asymptotic standard errors underestimate the true sampling variability of the OLS estimates. As reported in the first panel of Table 1, the average OLS standard errors across simulations are up to about 30% too low relative to the true standard errors, calculated as the standard deviation of the coefficient estimates across simulations.

2.3 The role of trends

Up to now we have been considering the case when the true values of the constant terms μ_i in equations (4)-(5) are zero. As seen in the second and third panels of Table 1, the size distortions on tests about β_2 can nearly double when $\mu_i \neq 0$, and the bias in the estimate of β_1 increases as well.

We can understand what is going on most easily by considering the case when the roots ρ_i are exactly unity.[20] In that case, if μ_1 is zero and μ_2 is not, x_{2t} will exhibit a deterministic time trend and this ends up stochastically dominating the random walk component of x_{2t}. The regression (3) would then be asymptotically equivalent to a regression of y_{t+1} on $(1, x_{1t}, \mu_2 t)'$. When the correlation $\delta = 1$, the asymptotic distribution of a t-test of a true null hypothesis

[18] While in bond return regressions δ is typically negative (as we discuss below in Section 3), we can focus here on $0 \leq \delta \leq 1$, since only $|\delta|$ matters for the distribution of the t-statistic.

[19] The lines in Figure 1 are based on Monte Carlo simulations with 100,000 replications.

[20] The following results are proved formally in Appendix A.3.

about β_1 in regression (3) would be identical to that if we were to perform a Dickey-Fuller test of the true null hypothesis $\eta = 0$ in the regression

$$\Delta x_{1,t+1} = \mu_1 + \eta x_{1t} + \xi t + \varepsilon_{1,t+1}, \tag{9}$$

which is the well-known Dickey-Fuller Case 4 distribution described in Hamilton (1994, eq [17.4.55]). We know that the coefficient bias and size distortions are bigger when a time trend is included in regression (9) compared to the case when it is not (see Case 2 versus Case 4 in Hamilton (1994, Tables B.5 and B.6))). For the same reason we would find that the Stambaugh bias of b_1 in regression (3) becomes worse when a variable x_{2t} with a deterministic trend is added to the regression. The standard error bias for b_2 is also exacerbated when the true μ_2 is nonzero.

In the case when ρ_2 is close to but strictly less than unity, this problem would vanish asymptotically but is still a factor in small samples. An apparent trend shows up in a finite sample because when a stationary variable with mean $\mu_2/(1-\rho_2)$ is started from $x_{20} = 0$, it will tend to rise over a sample of size $T = 100$ toward its unconditional mean. As ρ_2 approaches unity, this trend within a finite sample becomes arbitrarily close to that seen in a true random walk with drift μ_2.

In most of the applications we study in this paper, the trend in explanatory variables like inflation is down over the sample rather than up. The distribution of b_1 is identical if x_2 begins at $x_{20} = 2\mu_2/(1-\rho_2)$ and then drifts down to its unconditional mean $\mu_2/(1-\rho_2)$ as when x_2 begins at $x_{20} = 0$ and drifts up to $\mu_2/(1-\rho_2)$, so the issue raised here applies in those settings in exactly the same way.

Note that the values we have used for simulation in Table 1 are representative of those that may be encountered in practice. For example, an AR(1) process fit to the trend inflation variable used by Cieslak and Povala (2015) over the sample 1985-2013 has $\rho_2 = 0.99$ and $\mu_2/\sigma_2 = 1.5$, an even stronger drift relative to innovation than the value $\mu_2/\sigma_2 = 1.0$ used in Table 1. And their variable has a value in 1985:1 that is 5 times the size of $\mu_2/(1-\rho_2)$, implying a downward drift over 1985-2013 that is 4 times as fast as in the Table 1 simulation.

It's interesting finally to consider the case when both x_{1t} and x_{2t} have trends (see panel 3 of Table 1). This turns out to be almost the same asymptotic distribution just discussed with a reinterpretation of the variables. Consider for example the case when both trends are the same ($\mu_1 = \mu_2$). Note that a regression of y_{t+1} on $(1, x_{1t}, x_{2t})'$ has the identical fitted values as a regression of y_{t+1} on $(1, x_{1t} - x_{2t}, x_{2t})'$, which again is asymptotically equivalent to a regression in which the second variable is a driftless unit-root process correlated with the lagged residual and the third variable is dominated by a deterministic time trend. Now the

Stambaugh bias will show up in the coefficient on $x_{1t} - x_{2t}$. Translating back in terms of the original regression of y_{t+1} on $(1, x_{1t}, x_{2t})'$ we would now find Stambaugh biases in both b_1 and b_2 that are mirror images of each other.

Note the implications of this example. When μ_1 and μ_2 are both nonzero, if we were to regress y_{t+1} on x_{1t} alone, there would be no Stambaugh bias and no problem with t-tests about β_1, because x_{1t} is dominated by the time trend. The same is true if we were to regress y_{t+1} on x_{2t} alone. But when both x_{1t} and x_{2t} are included in the regression, spurious conclusions about both coefficients would emerge.

The relevance of this result for tests of the spanning hypothesis is that there is a common perception that adding auxiliary trending variables x_{2t} may help to clean up the low frequency variation in x_{1t}. Certainly this is a possibility, and certainly we do find empirically that t-statistics on x_{1t} often increase tremendously when trending variables x_{2t} are added to the regression. But the above results suggest that we need to exercise great care in interpreting evidence of this form, since the presence of trends in the predictors can lead to more poorly sized tests and spurious rejections of the spanning hypothesis.

2.4 Overlapping returns

A separate econometric problem arises in predictive regressions for bond returns with holding periods that are longer than the sampling interval, i.e., $h > 1$. Most studies in this literature, and all those that we revisit in this paper, focus on predictive regressions for annual excess bond returns in monthly data, that is regression (1) with $h = 12$ and

$$y_{t+h} = p_{n-h,t+h} - p_{nt} - hi_{ht}, \tag{10}$$

for p_{nt} the log of the price of a pure discount n-period bond purchased at date t and $i_{nt} = -p_{nt}/n$ the corresponding zero-coupon yield. In that case, $E(u_t u_{t-v}) \neq 0$ for $v = 0, \ldots, h-1$, as the overlapping observations induce a MA$(h-1)$ structure for the error terms. This raises additional problems in the presence of persistent regressors that can be seen even using conventional first-order asymptotics, as we briefly note in this section.

If x_{1t} and x_{2t} are uncorrelated and the true value of $\beta_2 = 0$, we show in Appendix A.4 that under conventional first-order asymptotics

$$\sqrt{T} b_2 \xrightarrow{d} N(0, Q^{-1} S Q^{-1}), \tag{11}$$

$$Q = E(x_{2t} x_{2t}')$$

$$S = \sum_{v=-\infty}^{\infty} E(u_{t+h}u_{t+h-v}x_{2t}x'_{2,t-v}). \tag{12}$$

Note that even if x_{2t} is completely independent of u_t at all leads and lags, the product $u_{t+h}x_{2t}$ would be highly serially correlated when x_{2t} is persistent, since $E(u_{t+h}u_{t+h-v}x_{2t}x'_{2,t-v}) = E(u_tu_{t-v})E(x_{2t}x'_{2,t-v}) \neq 0$. Overlapping observations, in combination with persistent regressors, substantially increase the sampling variability of the OLS estimate b_2, because the long-run covariance matrix S will exceed the value $S_0 = E(u^2_{t+h}x_{2t}x'_{2t})$ that would be appropriate for serially uncorrelated residuals.

The standard approach is to use heteroskedasticity- and autocorrelation-consistent (HAC) standard errors to try to correct for this, for example, the estimators proposed by Newey and West (1987) or Andrews (1991). However, long-run variance estimation is notoriously difficult, particularly in small samples, and different HAC estimators of S can lead to substantially different empirical conclusions (Müller, 2014). That Newey-West standard errors are unreliable for inference with overlapping returns was demonstrated convincingly by Ang and Bekaert (2007). We emphasize that the higher the persistence of the predictors, the less reliable is HAC inference, since the effective sample size becomes very small. The reverse-regression approach of Hodrick (1992) and Wei and Wright (2013) can largely overcome the problem arising from overlapping returns. However, as we will see in the examples below, the small-sample problems highlighted in Sections 2.1-2.3 also plague reverse-regression inference.

There is another consequence of basing inference on overlapping observations that appears not to be widely recognized– it substantially reduces the reliability of R^2 as a measure of goodness of fit. Let R^2_1 denote the coefficient of determination in a regression that includes only x_{1t}, compared to R^2_2 for the regression that includes both x_{1t} and x_{2t}. We show in Appendix A.4 that again for the case when x_{1t} and x_{2t} are uncorrelated and $\beta_2 = 0$

$$T(R^2_2 - R^2_1) \xrightarrow{d} r'Q^{-1}r/\gamma \tag{13}$$

$$\gamma = E[y_t - E(y_t)]^2, \quad r \sim N(0, S).$$

The difference $R^2_2 - R^2_1$ converges in probability to zero, but in a given finite sample it is positive by construction. If $x_{2t}u_{t+h}$ is positively serially correlated, then S exceeds S_0 by a positive-definite matrix, and r exhibits more variability across samples. This means $R^2_2 - R^2_1$, being a quadratic form in a vector with a higher variance, would have both a higher expected value as well as a higher variance when $x_{2t}u_{t+h}$ is serially correlated compared to situations when it is not. This serial correlation in $x_{2t}u_{t+h}$ would contribute to larger values for $R^2_2 - R^2_1$ on average as well as to increased variability in $R^2_2 - R^2_1$ across samples. In other words, including x_{2t} could substantially increase the R^2 even if H_0 is true. We will use bootstrap approximations to

the small-sample distribution of $R_2^2 - R_1^2$, and demonstrate that the dramatic values sometimes reported in the literature are often entirely plausible even under the spanning hypothesis.

2.5 Three solutions for more robust inference

We now propose three approaches for more robust inference about the spanning hypothesis.

2.5.1 A bootstrap design to test the spanning hypothesis

Obviously the main question is whether the above considerations make a material difference for tests of the spanning hypothesis. We propose a parametric bootstrap that generates data under the spanning hypothesis to assess how serious these econometric problems are in practice.[21] With this bootstrap approach we can calculate the size of conventional tests to assess their robustness. In addition, we can use it to test the spanning hypothesis with better size and power than for conventional tests.[22]

Our bootstrap design is as follows: First, we calculate the first three PCs of observed yields which we denote

$$x_{1t} = (PC1_t, PC2_t, PC3_t)',$$

along with the weighting vector $\hat{w}_n$ for the bond yield with maturity n:

$$i_{nt} = \hat{w}_n' x_{1t} + \hat{v}_{nt}.$$

That is, $x_{1t} = \hat{W} i_t$, where $i_t = (i_{n_1 t}, \ldots, i_{n_J t})'$ is a J-vector with observed yields at t, and $\hat{W} = (\hat{w}_{n_1}, \ldots, \hat{w}_{n_J})'$ is the $3 \times J$ matrix with rows equal to the first three eigenvectors of the variance matrix of i_t. We use normalized eigenvectors so that $\hat{W}\hat{W}' = I_3$.[23] Fitted yields are obtained as $\hat{\imath}_t = \hat{W}' x_{1t}$. Three factors generally fit the cross section of yields very well, with fitting errors $\hat{v}_{nt}$ (pooled across maturities) that have a standard deviation of only a few basis

[21] An alternative approach would be a nonparametric bootstrap under the null hypothesis, using for example a moving-block bootstrap to re-sample x_{1t} and x_{2t}. However, Berkowitz and Kilian (2000) found that parametric bootstrap methods such as ours typically perform better than nonparametric methods.

[22] Cochrane and Piazzesi (2005) and Ludvigson and Ng (2009, 2010) also used the bootstrap to test $\beta_2 = 0$. They did so with bootstrap confidence intervals generated under the alternative hypothesis. But it is well known that bootstrapping under the null hypothesis generally leads to better numerical accuracy and more powerful tests (Hall and Wilson, 1991; Horowitz, 2001), and of course this is the only way to obtain bootstrap estimates of the size of conventional tests.

[23] We sign the eigenvectors so that the elements in the last column of $\hat{W}$ are positive, i.e., the loadings of the yield with the longest maturity are positive. Hence the signs of PC1 and PC2 correspond to the usual definition of level and slope of the yield curve.

points.[24] Then we estimate by OLS a VAR(1) for x_{1t}:

$$x_{1t} = \hat{\phi}_0 + \hat{\phi}_1 x_{1,t-1} + e_{1t} \qquad t = 1, \ldots, T. \tag{14}$$

This time-series specification for x_{1t} completes our simple factor model for the yield curve. Though this model does not impose absence of arbitrage, it captures both the dynamic evolution and the cross-sectional dependence of yields. A no-arbitrage model is a special case of this structure with additional restrictions on $\hat{W}$, but these restrictions typically do not improve forecasts of yields; see for example Duffee (2011a) and Hamilton and Wu (2014). Next we generate 5000 artificial yield data samples from this model, each with length T equal to the original sample length. We first iterate on

$$x^*_{1\tau} = \hat{\phi}_0 + \hat{\phi}_1 x^*_{1,\tau-1} + e^*_{1\tau}$$

where $e^*_{1\tau}$ denotes bootstrap residuals. We start every bootstrap sample at $x^*_{10} = x_{10}$, the starting value for the observed sample, to allow for a possible contribution of trends resulting from initial conditions as discussed in Section 2.3. Then we obtain the bootstrap yields using

$$i^*_{n\tau} = \hat{w}'_n x^*_{1\tau} + v^*_{n\tau} \tag{15}$$

for $v^*_{n\tau} \overset{iid}{\sim} N(0, \sigma_v^2)$. The standard deviation of the measurement errors, σ_v, is set to the sample standard deviation of the fitting errors $\hat{v}_{nt}$.[25] We thus have generated an artificial sample of yields $i^*_{n\tau}$ which by construction only the three factors in $x^*_{1\tau}$ have any power to predict, but whose covariance and dynamics are similar to those of the observed data i_{nt}.

We likewise fit a VAR(1) to the observed data for the proposed predictors x_{2t},

$$x_{2t} = \hat{\alpha}_0 + \hat{\alpha}_1 x_{2,t-1} + e_{2t}, \tag{16}$$

from which we then bootstrap 5000 artificial samples $x^*_{2\tau}$ in a similar fashion as for $x^*_{1\tau}$. The bootstrap residuals $(e'^*_{1\tau}, e'^*_{2\tau})$ are drawn from the joint empirical distribution of (e'_{1t}, e'_{2t}).

Using the bootstrapped samples of predictors and yields, we can then investigate the properties of any proposed test statistic involving $y^*_{\tau+h}$, $x^*_{1\tau}$, and $x^*_{2\tau}$ in a sample for which the dynamic serial correlation of yields and explanatory variables are similar to those in the

[24] For example, in the data of Joslin et al. (2014) this standard deviation is 6.5 basis points.

[25] Some evidence in the literature suggests that yield fitting errors are serially correlated (Adrian et al., 2013; Hamilton and Wu, 2014). We have also investigated a setting with serial correlation in $v^*_{n\tau}$ and found that this does not change any of our findings.

actual data but in which by construction the null hypothesis is true that $x^*_{2\tau}$ has no predictive power for future yields and bond returns.[26] Consider for example a t-test for significance of a parameter in β_2. Denote the t-statistic in the data by t and the corresponding t-statistic in bootstrap sample i as t^*_i. To obtain a bootstrap estimate of the size of this test we simply calculate the fraction of samples in which $|t^*_i|$ exceeds the usual asymptotic critical value. And to use the bootstrap to carry out the hypothesis test, we calculate the bootstrap p-value as the fraction of samples in which $|t^*_i| > |t|$, and reject the null if this is less than, say, five percent. Equivalently, we can calculate the bootstrap critical value as the 97.5th percentile of $|t^*_i|$ and reject the null if $|t|$ exceeds it.

Note that this bootstrap procedure does not generate a test with an exact size of 5%. First, under local-to-unity asymptotics the bootstrap is not a consistent test because the test statistics are not asymptotically pivotal—their distribution depends on the nuisance parameters c_1 and c_2, which cannot be consistently estimated.[27] Second, least squares typically underestimates the autocorrelation of highly persistent processes due to small-sample bias (Kendall, 1954; Pope, 1990), so that the VAR underlying our bootstrap would typically be less persistent than the true DGP. We can address the second issue by using bias-corrected VAR parameter estimates for generating bootstrap samples. We will use the bias correction proposed by Kilian (1998) and refer to this as the "bias-corrected bootstrap."[28] We have found that even the bias-corrected bootstrap tends to be slightly oversized. This means that if our bootstrap test fails to reject the spanning hypothesis, the reason is not that the test is too conservative, but that there simply is not sufficient evidence for rejecting the null.

In fact we can use the Monte Carlo simulations in Section 2.2 to calculate the size of our bootstrap test. In each sample i simulated from a known parametric model, we can: (i) calculate the t-statistic (denoted $\tilde{t}_i$) for testing the null hypothesis that $\beta_2 = 0$; (ii) estimate the autoregressive models for the predictors by using OLS on that sample; (iii) generate a

[26] For example, if y_{t+h} is an h-period excess return as in equation (10) then in our bootstrap

$$\begin{aligned} y^*_{\tau+h} &= ni^*_{n\tau} - hi^*_{h\tau} - (n-h)i^*_{n-h,\tau+h} \\ &= n(\hat{w}_n' x^*_{1\tau} + v^*_{n\tau}) - h(\hat{w}_h' x^*_{1\tau} + v^*_{h\tau}) - (n-h)(\hat{w}_{n-h}' x^*_{1,\tau+h} + v^*_{n-h,\tau+h}) \\ &= n(\hat{w}_n' x^*_{1\tau} + v^*_{n\tau}) - h(\hat{w}_h' x^*_{1\tau} + v^*_{h\tau}) - (n-h)[\hat{w}_{n-h}'(\hat{k}_h + e^*_{1,\tau+h} + \hat{\phi}_1 e^*_{1,\tau+h-1} + \cdots \\ &\quad + \hat{\phi}_1^{h-1} e^*_{1,\tau+1} + \hat{\phi}_1^h x^*_{1\tau}) + v^*_{n-h,\tau+h}] \end{aligned}$$

which replicates the date t predictable component and the MA$(h-1)$ serial correlation structure of the holding returns that is both seen in the data and predicted under the spanning hypothesis.

[27] This result goes back to Basawa et al. (1991). See also Hansen (1999) as well as Horowitz (2001) and the references therein.

[28] We have found in Monte Carlo experiments that the size of the bias-corrected bootstrap is closer to five percent than for the simple bootstrap.

single bootstrap sample using these estimated autoregressive coefficients; (iv) estimate the predictive regression on the bootstrap sample;[29] and (v) calculate the t-statistic in this regression, denoted t_i^*. We generate many samples from the maintained model, repeating steps (i)-(v), and then calculate the value c such that $|t_i^*| > c$ in 5% of the samples. Our bootstrap procedure amounts to the recommendation of rejecting H_0 if $|\tilde{t}_i| > c$, and we can calculate from the above simulation the fraction of samples in which this occurs. This number tells us the true size if we were to apply our bootstrap procedure to the chosen parametric model. This number is reported in the second-to-last column of Table 1. We find in these settings that our bootstrap has a size above but fairly close to five percent.

We will repeat the above procedure to estimate the size of our bootstrap test in each of our empirical applications, taking a model whose true coefficients are those of the VAR estimated in the sample as if it were the known parametric model, and estimating VAR's from data generated using those coefficients. To foreshadow those results, we will find that the size is typically quite close to or slightly above five percent. In addition, we will show that our bootstrap procedure has good power properties. The implication is that if our bootstrap procedure fails to reject the spanning hypothesis, we should conclude that the evidence against the spanning hypothesis in the original data is not persuasive.

2.5.2 An alternative robust test for predictability

HAC inference is concerned with accurately estimating the matrix S in (12), but does not address the issue of standard error bias. However, we have found one existing approach for robust inference that does address this issue, based on the Ibragimov and Müller (2010) method for testing a hypothesis about a scalar coefficient. The original dataset is divided into q subsamples and the statistic is estimated separately over each subsample. If these estimates across subsamples are approximately independent and Gaussian (which is not a bad approximation to (8)), then a standard t-test with q degrees of freedom can be carried out to test hypotheses about the parameter. Müller (2014) provided evidence that this test has excellent size and power properties in regression settings where standard HAC inference is seriously distorted. Our simulation results, to be discussed below, show that this test also performs very well when testing the spanning hypothesis. We will report results for the Ibragimov-Müller (IM) test with the number of subsamples q equal to either 8 and 16 (as in

[29] In this simple Monte Carlo setting, we bootstrap the dependent variable as $y_\tau^* = \hat{\phi}_1 x_{1,\tau-1}^* + u_\tau^*$ where u_τ^* is resampled from the residuals in a regression of y_t on $x_{1,t-1}$, and is jointly drawn with $\varepsilon_{1\tau}^*$ and $\varepsilon_{2\tau}^*$ to maintain the same correlation as in the data. By contrast, in our empirical analysis the bootstrapped dependent variable is calculated from the bootstrapped bond yields, obtained using (15), and the definition of y_{t+h} (for example, as an annual excess return).

Müller, 2014). A notable feature of the IM test is that by its nature it captures "robustness" of the empirical findings not only with respect to serial correlation but also with respect to parameter instability across subsamples, as we will see in several of the empirical applications.

We can use the same Monte Carlo simulation as before to estimate the size of the IM test in the simple setting with two scalar predictors. The results are shown in the last column of Table 1. When coefficient bias is absent in estimates of β_2, the size of the IM test is quite close to five percent. The reason is that the IM test estimates the sampling variability of the test statistic by using variation across subsamples. In this way, it solves the problem of standard error bias that conventional t-tests are faced with. Note, however, that the IM test is unreliable in the presence of coefficient bias, because it splits the sample into smaller subsamples, which magnifies small-sample coefficient bias. For this reason, it is likely unreliable for testing hypotheses about β_1. We will calculate the small-sample size and power of the IM test in each of our empirical applications below, and will show that for tests of the spanning hypothesis $H_0 : \beta_2 = 0$, the IM test generally has very good size and power.

2.5.3 New data: subsample stability and out-of-sample forecasting

Our third approach to assess claims of return predictability is to confront published results with new data. To circumvent econometric problems of predictability regressions a common practice is to perform pseudo out-of-sample (OOS) analysis, splitting the sample into an initial estimation and an OOS period. We are skeptical of this approach because the researcher has access to the full sample when formulating the model, and the sample-split is arbitrary. However, for each of the studies that we revisit a significant amount of new data have come in since the original research. This gives us an opportunity both to reestimate the models over a sample period that includes new data, and further to evaluate the true out-of-sample forecasting performance of each proposed model.

3 Economic growth and inflation

In this section we examine the evidence reported by Joslin et al. (2014) (henceforth JPS) that macro variables may help predict bond returns. We will follow JPS and focus on predictive regressions as in equation (1) where y_{t+h} is an excess bond return for a one-year holding period ($h = 12$), x_{1t} is a vector consisting of a constant and the first three PCs of yields, and x_{2t} consists of a measure of economic growth (the three-month moving average of the Chicago Fed National Activity Index, GRO) and of inflation (one-year CPI inflation expectations from the Blue Chip Financial Forecasts, INF). While JPS also presented model-based evidence in favor

of unspanned macro risks, all of those results stem from the substantial in-sample predictive power of x_{2t} in these excess return regressions. The sample contains monthly observations over the period 1985:1-2007:12.[30]

3.1 Predictive power according to $\bar{R}^2$

JPS found that for the ten-year bond, the (adjusted) $\bar{R}^2$ of regression (1) increased from 0.20 to 0.37 when x_{2t} is included. For the two-year bond, the change is even more striking, with $\bar{R}^2$ increasing from 0.14 to 0.48. JPS interpreted this as strong evidence that macroeconomic variables have predictive power for excess bond returns beyond the information in the yield curve, and concluded that "macroeconomic risks are unspanned by bond yields" (p. 1203). We report the $\bar{R}^2$ for an average excess-return on 2- to 10-year bonds in the first row of Table 2, where the first three entries are based on the same data set that was used by JPS.[31] The entry $\bar{R}^2_1$ gives the $\bar{R}^2$ for the regression with only x_{1t} as predictors, and $\bar{R}^2_2$ corresponds to the case when x_{2t} is added to the regression. For this specification, $\bar{R}^2$ also increases quite substantially, by 19 percentage points.

However, there are some warning flags for these predictive regressions. First, the predictors are very persistent; the first-order sample autocorrelations of PC1 and PC2 are 0.98 and 0.97, respectively, while that of INF is 0.99. Second, the sample is relatively small, with 276 observations. Third, the dependent variable is an annual overlapping return, i.e., $h = 12$. In the presence of these three warning flags even large increases in $\bar{R}^2$ may be plausible under the null hypothesis, as suggested by the arguments in Section 2.4.

The second row of Table 2 reports the mean $\bar{R}^2$ across 5000 replications of the bootstrap described in Section 2.5.1, that is, the average value we would expect to see for these statistics in a sample of the size used by JPS in which x_{2t} in fact has no true ability to predict y_{t+h} but whose serial correlation properties are similar to those of the observed data. The third row gives 95% bootstrap intervals, that is, the 2.5th and 97.5th percentiles of the bootstrap distributions which impose the null hypothesis. The variability of the $\bar{R}^2$ is very high. Values for $\bar{R}^2_2$ as high as 60% would not be uncommon, as indicated by the bootstrap intervals. Most notably, adding the regressors x_{2t} often substantially increases the $\bar{R}^2$—even increases of 20

[30] The last observation corresponds to the annual excess returns from December 2007 to December 2008.

[31] In Table 2 we have attempted to summarize results for R^2 or $\bar{R}^2$ across different studies on a comparable basis that is as close as possible to that in the original study. In the case of JPS, they reported results for only the 2-year and 10-year bonds and not an average. In Table B.1 in Appendix B.1 we present analogous results for each individual bond from two through ten years maturity. The increase in $\bar{R}^2$ when adding macro variables is particularly pronounced for short-term bonds, but most of our conclusions apply to these short maturities as well.

percentage points are not uncommon—although x_{2t} has no predictive power in population by construction. According to the bootstrap small-sample distribution of $\bar{R}^2$, the increase in the data of 19 percentage points is not inconsistent with the spanning hypothesis.

Since the persistence of x_{2t} is high, it may be important to adjust for small-sample bias in the VAR estimates. For this reason we also carried out the bias-corrected (BC) bootstrap. The expected values and 95% confidence intervals are reported in the bottom two rows of the top panel in Table 2. As expected, more serial correlation in the generated data (due to the bias correction) increases the mean and the variability of the $\bar{R}^2$ and of their difference, so that $\bar{R}^2_2 - \bar{R}^2_1$, our statistic of main interest, is even more comfortably within the bootstrap interval.

3.2 Testing the spanning hypothesis

Is the predictive power of macro variables statistically significant? JPS only reported $\bar{R}^2$ for their excess return regression, but one is naturally interested in formal tests of the spanning hypothesis in JPS' excess return regressions. We report coefficient estimates and test statistics in Table 3. The common approach to address the serial correlation in the residuals due to overlapping observations is to use the standard errors and test statistics proposed by Newey and West (1987), and in regressions for annual returns using monthly data researchers typically use 18 lags (see among many others Cochrane and Piazzesi, 2005; Ludvigson and Ng, 2009). In the second row of Table 3 we report the resulting t-statistic for each coefficient along with the Wald test of the hypothesis $\beta_2 = 0$, calculated using Newey-West standard errors with 18 lags. The third row reports the p-values for these statistics if they were interpreted using the conventional asymptotic approximation. According to this popular test, GRO and INF appear strongly significant, both individually and jointly. In particular, the Wald statistic has a p-value below 0.1%.

However, the small-sample problems described in Section 2 likely distort these test results. The canonical correlation between innovations in one-month excess returns and innovations in the three yield PCs (the generalization of the parameter δ in Section 2.2) is 0.99. Such a high correlation will be present by construction in all tests of the spanning hypothesis, because x_{1t} includes yield PCs that explain current yields very well, and so innovations to x_{1t} are necessarily highly correlated with surprise returns realized at t. Furthermore, we noted above that in this empirical application the autocorrelations of the predictors are high, and the sample size is relatively small. Our theory predicts that standard error bias will be severe in this application. In addition, the well-known small-sample problems of Newey-West standard errors are also likely to be particularly pronounced in this setting.

We therefore employ our bootstrap to carry out tests of the spanning hypothesis that account for these small-sample issues. Again, we use both simple (OLS) and BC bootstrap. For each, we report five-percent critical values for the t- and Wald statistics, calculated as the 95th percentiles of the bootstrap distribution, as well as bootstrap p-values, i.e., the frequency of bootstrap replications in which the bootstrapped test statistics are at least as large as in the data. Using either the simple or BC bootstrap, the coefficient on GRO is insignificant even at the 10% level, and the coefficient on INF is marginally significant at the 5% level. The bootstrap p-value for the Wald test of the spanning hypothesis is slightly below 5% for the simple bootstrap and slightly above 5% for the BC bootstrap. These tests result in much weaker evidence against the spanning hypothesis than one would have thought based on conventional asymptotic interpretation of the test statistics.

We also report in Table 3 the p-values for the IM test of the individual significance of the coefficients. The coefficients on GRO and INF are not significant at conventional levels based on this test.

Using the bootstrap we can easily calculate the true size of the conventional HAC, bootstrap, and IM tests, which all have a nominal size of five percent. These are reported in the *Size* section of the top panel of Table 3. For the conventional HAC tests, this is calculated as the frequency of bootstrap replications in which the t- and Wald statistics exceed the usual asymptotic critical values. The results reveal that the true size of these conventional tests is 20-37% instead of the presumed five percent. These substantial size distortions are also reflected in the bootstrap critical values, which far exceed the conventional critical values. The bootstrap and the IM tests, in contrast, have a size that is estimated to be very close to five percent, eliminating almost all of the size distortions of the more conventional tests.

We can also use our bootstrap to evaluate the power of our proposed tests. To do so, we simply add $\hat{\beta}_2 x^*_{2\tau}$ to the value generated by our bootstrap for $y^*_{\tau+h}$, where $\hat{\beta}_2$ is the coefficient on x_{2t} in the original data sample. We now have a generated sample in which x_{2t} in fact does predict y_{t+h}, and with a magnitude that is exactly that claimed in the original study. We repeat this to obtain 5000 samples in which the spanning hypothesis does not hold, and in each sample calculate all our tests. We find that in 92% of these samples, the HAC t-statistic exceeds 3.8, the value that our simple bootstrap suggests we need to see before rejecting the spanning hypothesis. In other words, our bootstrap test has high power, and should reject the hypothesis if it were indeed false. The IM test also has good power to reject the null hypothes. This suggests that the reason that neither the IM nor the bootstrap test reject the spanning hypothesis is not a lack of power, but the fact that empirical spanning is a reasonable description of the observed sample.

3.3 New data

What happens when we augment the sample with the eight years of new data that have arrived since the original analysis by JPS?[32] The last three columns of the top panel of Table 2 show that the in-sample improvement in $\bar{R}^2$ when x_{2t} is included in the regression is substantially smaller over the 1985-2015 data set than was found on the original JPS data set, and the improvement is far from statistically significant.[33] And as seen in the second panel of Table 3, the values of the conventional HAC t- and F-tests are substantially smaller on the longer data set than was found in the original data, and in fact the coefficient on GRO is now no longer statistically significant even if the t-statistic was interpreted in the usual way. The t-statistic on inflation would still appear to be significant if interpreted using the conventional asymptotic distribution, but based on the bootstrap small-sample distribution it is clearly insignificant.

Row 1 of Table 4 reports the pure OOS forecast comparison for y_{t+h} the average 12-month excess return across 2- to 10-year bonds.[34] Whereas in the original JPS in-sample regression, the addition of x_{2t} improved the mean squared prediction error by 24%, the addition of x_{2t} leads to a deterioration in the OOS prediction error by 140%. Moreover, the OOS performance of the predictive model that imposes the spanning hypothesis and leaves out x_{2t} is significantly better than the unrestricted model according to the Diebold and Mariano (1995) test.[35]

Adding new observations to the JPS data set substantially weakens the evidence against the spanning hypothesis. But if the null hypothesis were truly false, we would expect to find the evidence against it become stronger, not weaker, when we use a bigger data set. We conclude on the basis of the bootstrap, the IM test, and newly available data that the JPS evidence on unspanned macro risks is far from convincing.

4 Factors of large macro data sets

Ludvigson and Ng (2009, 2010) found that factors extracted from a large macroeconomic

[32] We update the yield data using unsmoothed Fama-Bliss yields provided to us by Anh Le.

[33] This also turns out to be the case for every individual bond, including the 2-year bond; see Table B.1 in Appendix B.1.

[34] We used a recursive scheme where we re-estimate the predictive regressions by extending the estimation window each month of the newly available data.

[35] In related work, Giacoletti et al. (2016) evaluate the real-time OOS forecasting performance of a model similar to that used in JPS. They find that including macro variables only helps for predicting very short-term yields and only over a specific subsample, but that overall "'macro rules' add little to the forecast accuracy of the basic yields-only rule" (p. 29). While this supports the spanning hypothesis, they find some incremental predictive power when including survey forecast disagreement.

data set are helpful in predicting excess bond returns, above and beyond the information contained in the yield curve. Here we revisit this evidence, focusing on the results in Ludvigson and Ng (2010) (henceforth LN). They started with a panel data set of 131 macro variables observed over 1964:1-2007:12 and extracted eight macro factors using the method of principal components. These factors, which we will denote by $F1$ through $F8$, were then related to future one-year excess returns on two- through five-year Treasury bonds. They also included the return-forecasting factor that was proposed by Cochrane and Piazzesi (2005), denoted as CP, which is the linear combination of forward rates that best predicts the average excess return across maturities. Based on comparisons of $\bar{R}^2$ of regressions with and without macro factors, as well as HAC inference using Newey-West standard errors, LN concluded that macro factors help predict excess returns, even when controlling for information in the yield curve using the CP factor.

We focus on regressions that are very similar to those estimated by LN, with two differences: First, we capture the information in the yield curve using the first three PCs of yields, while LN use the CP factor. Second, we do not carry out LN's preliminary specification search—they considered many different combinations of the factors along with squared and cubic terms—in order to focus squarely on hypothesis testing for a given regression specification.[36] Our regressions take the same form as (1), where now y_{t+h} is the average one-year excess bond return for maturities of two through five years, x_{1t} contains a constant and three yield PCs, and x_{2t} contains eight macro PCs. As before, our interest is in testing the hypothesis $H_0 : \beta_2 = 0$.

Table 2 shows that in LN's data set the $\bar{R}^2$ increases by 10 percentage points when the macro factors are included, consistent with LN's findings. The first three rows of Table 5 show the coefficient estimates, HAC t- and Wald statistics (using Newey-West standard errors with 18 lags as in LN), and p-values based on the conventional asymptotic distributions of these test statistics. There are five macro factors that appear to be statistically significant at the ten-percent level, among which three are significant at the five-percent level. The Wald statistic for H_0 far exceeds the critical values for conventional significant levels (the five-percent critical value for a $\chi^2(8)$ distribution is 15.5). Taken at face value, this evidence suggests that macro factors have strong predictive power, above and beyond the information contained in the yield curve.

How robust are these econometric results? We first check the warning flags. As usual, the first two yield PCs are very persistent, with autocorrelations of 0.98 and 0.94. The most

[36] We were able to closely replicate the results in LN's tables 4 through 7, and have also applied our techniques to those regressions, which led to qualitatively similar results.

persistent macro variables have first-order autocorrelations of around 0.75, so the persistence of x_{2t} is lower than in the data of JPS but still considerable. As always, the yield PCs strongly violate strict exogeneity by construction, for the reasons explained in the previous section. Based on these indicators, it appears that small-sample problems may well distort the results of conventional inference methods.

To address the potential small-sample problems in this context, we bootstrapped 5000 data sets of artificial yields and macro data in which H_0 is true in population. The samples each contain 516 observations, which corresponds to the length of the original data sample. We report results only for the simple bootstrap without bias correction, because the bias in the VAR for x_{2t} is estimated to be small. Note that LN also considered bootstrap inference, but their main bootstrap design imposed the expectations hypothesis, in order to test whether excess returns are predictable by macro factors and the CP factor. Using this setting, LN produced convincing evidence that excess returns are predictable, which is fully consistent with our results. Our null hypothesis of interest, however, is that excess returns are predictable only by current yields. While LN also reported results for a bootstrap under the alternative hypothesis, our bootstrap generates samples under the spanning hypothesis, and therefore allows us to provide a more accurate assessment of the spanning hypothesis, and to estimate the size of conventional tests under the null.

Table 2 shows that the observed increase in predictive power from adding macro factors to the regression, measured by $\bar{R}^2$, would not be implausible if the null hypothesis were true, as the increase in $\bar{R}^2$ is within the 95% bootstrap interval. And as seen in Table 5, our bootstrap finds that only three coefficients are significant at the ten-percent level (instead of five using conventional critical values), and one at the five-percent level (instead of three). While the Wald statistic is significant even compared to the critical value from the bootstrap distribution, the evidence is weaker than when using the asymptotic distribution.

Table 5 also reports p-values for the IM test using $q = 8$ and 16 subsamples. Only the coefficient on $F7$ is significant at the 5% level using this test, and then only for $q = 16$. The failure to reject the null based on the IM tests is a reflection of the fact that the parameter estimates are often unstable across subsamples. Duffee (2013b, Section 7) has also noted problems with the stability of the results in Cochrane and Piazzesi (2005) and Ludvigson and Ng (2010) across different sample periods.

We again use the bootstrap to estimate the size and power of the different tests with a nominal size of five percent. The results, reported in Table 5, reveal that the conventional t-tests have modest size distortions, with true size of 9-14% instead of the nominal five percent. But the Wald test is seriously distorted, with a true size of 32 percent. The Wald test

compounds the problems resulting from the non-standard small-sample distribution of each of the eight coefficient estimates for x_{2t}, and therefore ends up with a large size distortion. By contrast, our proposed bootstrap and IM tests have close to correct size, and also have good power.

Again there are several years of data that have arrived since the original LN analysis was conducted.[37] We repeated our analysis using the same 1985-2015 sample period that we used to reassess the results of JPS. There it was a strictly larger sample than the original, but here, in the case of LN, our second sample adds data at the end but leaves some out at the beginning. Reasons for interest in this sample period include the significant break in monetary policy in the early 1980s, the advantages of having a uniform sample period for comparison across all the different studies considered in our paper, and investigating robustness of the original claims in describing data since the papers were originally published. The results, shown in the right panel of Table 2 and the bottom panel of Table 5, show that over the later sample period, the evidence for the predictive power of macro factors is quite weak. The increases in $\bar{R}^2$ in Table 2 are not statistically significant, being squarely within the bootstrap intervals under the spanning hypothesis. The Wald test rejects H_0 when using asymptotic critical values, but is very far from significant when using bootstrap critical values. And the IM tests find no evidence of predictive power of the macro factors.

We also repeated the pure OOS exercise and report the results in the second row of Table 4, using relations estimated over data from 1964 through T to predict the value of $T+1$ for all the dates T since LN's original analysis. In contrast to the results for JPS (in the first row), we find that the unrestricted model which includes macro variables does better both in-sample and OOS than the model that only includes yield PCs. Adding the eight macro factors reduces the MSE for predicted returns over the 2009-2015 period by 25%. However, this improvement is not large enough to be statistically significant based on the DM test.

Overall, these results again show that conventional measures of fit and hypothesis tests are not reliable for assessing the spanning hypothesis. Furthermore, the evidence that macro factors have predictive power beyond the information already contained in yields is weaker than the results in LN would initially have suggested. Both small-sample econometric problems as well as subsample stability raise concerns about the robustness of the results.[38]

[37] To construct the macro factors for the 1985-2015 sample period, we used the macro data set of McCracken and Ng (2014) and transformed the data and extracted the PCs in the same way as LN did. Using the data constructed in this way, we also obtained results similar to LN's over their original sample period.

[38] Appendix B.2 reports additional results for predictive regressions with return-forecasting factors, using an empirical approach that was also advocated by LN. These results reinforce our conclusions.

5 Trend inflation

Cieslak and Povala (2015) (henceforth CPO) presented evidence that measures of the trend in inflation can help to estimate risk premia in bond returns. They established this result using a variety of measures of trend inflation, and found that these measures generally contained substantial predictive power for annual excess bond returns beyond the predictive information contained in yields. Their strongest results (and the specification we investigate here) calculates the trend in inflation using a very slowly adjusting weighted average of observed inflation rates,

$$\tau_t = (1-\nu)\sum_{i=0}^{t-i}\nu^i \pi_{t-i}, \tag{17}$$

for π_t the month t year-over-year inflation rate as measured by the CPI and $\nu = 0.987$. This measure is plotted in Figure 2 along with the yield on a 10-year bond. Both τ_t and nominal interest rates exhibited an upward trend until the early 1980s and a distinct downward trend since then. The variable τ_t is also extremely persistent, with an autocorrelation of 0.9985. To reproduce CPO's key results in a similar structure to those used in discussing the previous two studies, let y_{t+h} denote a weighted average[39] of the annual excess returns on 2- to 10-year bonds, x_{1t} a constant and the first three PCs of yields, and $x_{2t} = \tau_t$. Table 2 confirms CPO's conclusion that adding the inflation trend results in an enormous increase in the $\bar{R}^2$, in this case from 0.12 to 0.46.

To address small-sample problems we again generate bootstrap samples using the same setup as before, with a VAR(1) for yield PCs and an AR(1) for the inflation trend.[40] We use bias-corrected coefficient estimates due to the high persistence of the yield PCs and the inflation trend. Table 2 shows that although our bootstrap suggests that a large increase in $\bar{R}^2$ would be expected in this setting, the observed increase is substantially larger than could be explained under the spanning hypothesis.

Table 6 reports coefficient estimates and hypothesis tests for these regressions. The first three rows report estimates of the predictive regression using only x_{1t}, reproducing CPO's finding that information in yields has only moderate predictive power. The next three rows show that once the inflation trend is added, not only does the trend appear to be highly significant, but the predictive power of PC1 and PC2 also increases substantially. CPO calculated standard errors using the Wei and Wright (2013) reverse regression (RR) approach as a way to

[39] We use the same type of weighted average of excess returns as CPO, where returns are divided by the bond's duration before being averaged.

[40] While more sophisticated bootstrap designs for inflation and the inflation trend are possible—e.g., calculating the bootstrapped inflation trend as a moving average of inflation simulated from an ARIMA model—we have found that our key results remain essentially unaffected by this choice.

mitigate the problems identified in Section 2.4. The RR approach uses the insight of Hodrick (1992)—that it is beneficial to base inference in predictions for overlapping returns on estimates of predictions for one-period, non-overlapping returns which use cumulated predictors—and extends Hodrick's approach to perform inference about other hypotheses than the absence of predictability. But this does not eliminate the small-sample problems we raise in this paper, as seen in the *Size* section of Table 6. Our bootstrap estimates suggest that in this setting the RR t-test would reject a true null hypothesis 45% of the time instead of the intended 5%.[41] As indicated by the bootstrap critical value, we would need to see a t-statistic above 3.6 to reject the null at the 5% level. The actual observed t-statistic, however, is 6.2, providing strong evidence against the null even taking into account small-sample inference problems.

The bootstrap allows us to drill down further and try to understand the role of each of the econometric problems described in Sections 2.2, 2.3 and 2.4 for these large size distortions. We found in this setting that an HAC test using Newey-West standard errors with the usual 18 lags has an even larger size, 56%. That is, the RR approach partially alleviates the small-sample problem arising from overlapping observations. To assess how much of the problem still remains even with the RR test, we can compare the results to a setting with $h = 1$, where y_{t+1} contains monthly excess returns[42] and conventional hypothesis tests are carried out using White's heteroskedasticity robust standard errors (an example of this common approach is Duffee, 2013b, Section 7). In that case the t-test has a true size of 34%. The fact that this is noticeably below 45% suggests that RR does not completely solve the problem of overlapping observations. The last question is about the role of trends. Our theory in Section 2.3 suggests that conventional tests should be severely distorted in CPO's setting because both x_{1t} and x_{2t} are trending—despite the increase in the 1970s and early 1980s, both the level of yields and τ_t trend down substantially over the whole sample. We can verify this theoretical prediction by bootstrapping the test statistics in a setting where trends are absent by construction. We do so by initializing the bootstrap simulations at the population mean, and not at the first observation in the sample as we normally do, with the result that there is no drift in the predictors. The size of the RR t-test in that case is 16%, compared to the 45% we reported in Table 6 for the case when the bootstrap captures the trends in the predictors.[43] This confirms that the presence of trends substantially magnifies the size distortions that arise from standard

[41] It is well-known that reverse-regression standard errors, just like Hodrick's standard errors, do not eliminate the problem of Stambaugh bias; note for example the size distortions in Table 1 of Wei and Wright (2013). Therefore it is unsurprising that this approach does not eliminate standard error bias.

[42] We calculate monthly excess returns using the approximation $y_{t+1}^{n-1} \approx y_{t+1}^{n}$ and the one-month Treasury yield from Anh Le's unsmoothed Fama-Bliss yield data.

[43] In the setting with monthly excess returns we found that in the absence of trends the t-test has a size of 14%, compared to the size of 34% in the presence of trends which we reported in the text.

error bias, i.e., from the correlation of x_{1t} and realized excess returns.[44]

In contrast to the size-adjusted (bootstrap) RR tests, the IM tests do not reject the spanning hypothesis. These tests seem to be somewhat oversized, with the true size estimated around 14% instead of the nominal 5%, meaning that the failure to reject using this test is not the result of having relied on an undersized test. Figure 3 provides some insight into why IM fails to reject. The figure plots the coefficients on each predictor across the $q = 8$ subsamples. The coefficients are standardized by dividing them by the sample standard deviation across the eight estimated coefficients for each predictor. Thus, the IM t-statistics, which are reported in the legend of Figure 3, are equal to the means of the standardized coefficients across subsamples, multiplied by $\sqrt{8}$. The figure shows that $PC1$ and $PC2$ have more consistent predictive power across subsamples than does the trend, whose coefficient appears to be strongly positive in the second subsample but negative in the first, seventh, and eighth subsamples.

Again we reestimate the predictive regressions over the 1985-2015 period that we have used as a common comparison with the other studies. The increase in $\bar{R}^2$ is smaller and no longer statistically significant on this dataset, as seen in the last three columns of Table 2. Compared to the bootstrap small-sample distribution, the reverse-regression t-statistic is only marginally statistically significant at the 5% level, as seen in the second panel of Table 6. And the IM test again fails to reject H_0 over this sample period. Note that in the post-1985 sample the downward drift in the level of yields and the inflation trend is even more pronounced, adding to the econometric problems caused by trends in explanatory variables.

We also evaluated the true OOS usefulness of τ_t using new data after the end of CPO's sample period, which we report in line 3 of Table 4. Whereas within CPO's original sample the trend reduces the MSE by 40%, for the data that have come in since 2011 including the inflation trend actually increases the MSE by a factor of 12. However, due to the short OOS period, even this dramatic deterioration in forecast accuracy is not statistically significant, based on the DM statistic.

Our results in this case study again lead to the conclusion that conventional tests are very unreliable for inference about the spanning hypothesis. The small-sample issues are most severe in CPO's case because of the extreme persistence of the predictor $x_{2t} = \tau_t$ and the presence of pronounced trends. It is certainly noteworthy that CPO's key result survives even accounting for the very serious small-sample problems, at least in their original data set. On

[44]By contrast, in the JPS data we found that the biggest single source of the size distortions is the use of overlapping returns and Newey-West standard errors. And in the LN data it is a combination of the overlapping returns and the presence of a relatively large number of predictors in x_{2t}, which magnifies the size distortions.

the other hand, the IM test and the addition of new data suggest that the result may not be quite as robust as it initially seemed.

6 Higher-order PCs of yields

Cochrane and Piazzesi (2005) (henceforth CP) documented several striking new facts about excess bond returns. They showed that a tent-shaped combination of forward rates predicts annual excess returns on different long-term bonds with an R^2 of up to 37% (and even up to 44% when lags are included). Importantly for our context, CP found that the first three PCs of yields—level, slope, and curvature—did not fully capture this predictability, but that the fourth and fifth PC were also very helpful. As usual, the first three PCs explain a large share of the cross-section variation in yields (99.97% in their data), but CP found that the other two PCs, which explain only 0.03% of the cross-section variation in yields, are statistically important for predicting excess bond returns. In particular, the fourth PC appeared "very important for explaining expected returns" (p. 147). Here we assess the robustness of this finding, by revisiting the null hypothesis that only the first three PCs, but not higher-order PCs, predict excess returns.

The last panel of Table 2 shows (unadjusted) R^2 for predictive regressions for the average excess bond return using three and five PCs as predictors, and the first entries replicate the results of CP. In Table 7 we report the results of HAC inference for the regressions with 5 PCs using Newey-West standard errors with 18 lags, and the Wald statistic is identical to that reported by CP in their Table 4. The p-values indicate that $PC4$ is very strongly statistically significant, and that the spanning hypothesis would be rejected.

We then use our bootstrap procedure to obtain robust inference about the relevance of the predictors PC4 and PC5. We find that the CP results cannot be accounted for by small-sample size distortions. One reason is that the persistence of higher-order PCs is quite low, so that the size distortions of conventional tests are small. The other reason is that the Newey-West t-statistic on $PC4$ is far too large to be accounted for by the kinds of factors identified in Section 2. Likewise the increase in R^2 reported by CP would be quite implausible under the null hypothesis, as it falls far outside the 95% bootstrap interval under the null.

Interestingly, however, Table 7 shows that the IM tests fail to reject the null hypothesis that $\beta_2 = 0$. These tests conclude that the coefficients on $PC4$ and $PC5$ are not statistically significant, and find only the level and slope to be robust predictors of excess bond returns. Using the bootstrap we find that in this case the IM tests have the correct size, close to

five percent. The power of the IM tests is very high for finding significance of PC4.[45] The excellent size and power of the IM tests for PC4 should give us pause in concluding that this variable really helps to predict bond returns. Figure 4 again conveys why the IM test for $q = 8$ fails to reject. The figure shows that $PC1$ and $PC2$ have consistent predictive power across subsamples, whereas the coefficient on $PC4$ switches signs several times. The strong association between $PC4$ and excess returns is mostly driven by the fifth subsample, which starts in September 1983 and ends in July 1988.[46]

It is worth emphasizing the similarities and differences between the tests of interest to CP and in our own paper. Their central claim, with which we concur, is that the factor they have identified is a useful and stable predictor of bond returns. However, this factor is a function of all 5 PC's, and the first 3 of these account for 76% of the variation of the CP factor. Our claim is that it is the role of $PC1$-$PC3$ in the CP factor, and not the addition of $PC4$ and $PC5$, that makes this factor a robust predictor of bond returns. Our test for structural stability differs from those performed in CP and their accompanying online appendix. CP conducted tests of the usefulness of their return-forecasting factor for predicting returns across different subsamples, a result that we have been able to reproduce and confirm. Our tests, by contrast, look at stability of the role of each individual PC. While we agree with CP that the first three PC's indeed have a stable predictive relation, the predictive power of the 4th and 5th PC is much more tenuous, and is insignificant in most of the subsample periods that CP considered.[47]

In the last three columns of Table 2 and the bottom panel of Table 7 we report results for the 1985–2015 sample period. In this case, the increase in R^2 due to inclusion of higher-order PCs is comfortably inside the 95% bootstrap intervals, and the coefficients on $PC4$ and $PC5$ are not significant for any method of inference.

CP's sample period ended more than ten years prior to the time of this writing, giving us the longest true OOS period among the studies considered. The last row of Table 4 shows that in contrast to the in-sample estimates, where including $PC4$ and $PC5$ reduces the MSE by 11%, OOS predictive power deteriorates by 20% when the null hypothesis is not imposed. While the DM test does not reject the hypothesis that both models have equal predictive accuracy in population, restricting the predictive model to use only the level, slope and curvature leads to

[45] The power is low for PC5. The reason is that our alternative hypothesis uses the coefficient estimate of β_2 from the actual data, where PC5 is a very weak predictor.

[46] Consistent with this finding, an influence analysis of the predictive power of $PC4$ indicates that the observations with the largest leverage and influence are almost all clustered in the early and mid 1980s.

[47] Duffee (2013b, Section 7) also documented that extending CP's sample period to 1952–2010 alters some of their key results, and we have found that over Duffee's sample period the predictive power of higher-order PCs disappears.

more stable and more accurate return predictions in this particular sample. We conclude from both our in-sample and OOS results that the evidence for predictive power of higher-order factors is tenuous and sample-dependent, and that there is no compelling evidence that the first three PCs of yields are insufficient to estimate bond risk premia.[48]

7 Other studies

Several other studies have also reported evidence that might appear to be inconsistent with the spanning hypothesis. Cooper and Priestley (2008) concluded that the output gap contains useful information for forecasting interest rates, while Greenwood and Vayanos (2014) found the same for measures of Treasury bond supply. We have repeated our analysis using the datasets in these studies and found that evidence against the spanning hypothesis in these two cases is even weaker than for any of the studies discussed in Sections 3 to 6. Details of our investigations are reported in Appendices B.3 and B.4.

8 Conclusion

Conventional tests of whether variables other than the level, slope and curvature can help predict bond returns have significant size distortions, and the R^2 of the regression can increase dramatically when other variables are added to the regression even if they have no true explanatory power.

We proposed three strategies for dealing with this problem: First, a simple bootstrap based on PCs; second, a robust t-test based on subsample estimates proposed by Ibragimov and Müller (2010); and third, examining the proposed variables' usefulness in new data, preferably in a true out-of-sample forecasting exercise. We used these methods to revisit six different widely cited studies, and found in each case that the evidence that variables other than the current level, slope and curvature predict excess bond returns is substantially less convincing than the original research would have led us to believe.

We emphasize that these results do not mean that fundamentals such as inflation, output, and bond supplies do not matter for interest rates. Instead, our conclusion is that any effects of these variables can be summarized in terms of the level, slope, and curvature. Once these three factors are included in predictive regressions, no other variables appear to have robust

[48] Cattaneo and Crump (2014) also investigated the robustness of the results of Cochrane and Piazzesi (2005) and obtained even more negative results: Using a new HAC test proposed by Müller (2014) they did not reject the null hypothesis that the CP factor had no predictive power in a variety of in-sample and OOS specifications.

forecasting power for future yields or returns. Our results cast doubt on the claims for the existence of unspanned macro risks and suggest that it may not be necessary to look beyond the information in the yield curve to estimate risk premia in bond markets.

References

Adrian, Tobias, Richard K. Crump, and Emanuel Moench (2013) "Pricing the Term Structure with Linear Regressions," *Journal of Financial Economics*, Vol. 110, pp. 110–138.

Andrews, Donald W. K. (1991) "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, Vol. 59, pp. 817–858.

Ang, Andrew and Geert Bekaert (2007) "Stock return predictability: Is it there?" *Review of Financial studies*, Vol. 20, pp. 651–707.

Bansal, Ravi and Ivan Shaliastovich (2013) "A Long-Run Risks Explanation of Predictability Puzzles in Bond and Currency Markets," *Review of Financial Studies*, Vol. 26, pp. 1–33.

Basawa, Ishwar V, Asok K Mallik, William P McCormick, Jaxk H Reeves, and Robert L Taylor (1991) "Bootstrapping unstable first-order autoregressive processes," *Annals of Statistics*, pp. 1098–1101.

Bauer, Michael D. and Glenn D. Rudebusch (2016) "Resolving the Spanning Puzzle in Macro-Finance Term Structure Models," *Review of Finance*, forthcoming.

Bekaert, G., R.J. Hodrick, and D.A. Marshall (1997) "On biases in tests of the expectations hypothesis of the term structure of interest rates," *Journal of Financial Economics*, Vol. 44, pp. 309–348.

Bekaert, Geert and Robert J Hodrick (2001) "Expectations hypotheses tests," *The journal of finance*, Vol. 56, pp. 1357–1394.

Berkowitz, Jeremy and Lutz Kilian (2000) "Recent developments in bootstrapping time series," *Econometric Reviews*, Vol. 19, pp. 1–48.

Bikbov, Ruslan and Mikhail Chernov (2010) "No-Arbitrage Macroeconomic Determinants of the Yield Curve," *Journal of Econometrics*, Vol. 159, pp. 166–182.

Campbell, John Y. and John H. Cochrane (1999) "By force of habit: A consumption-based explanation of aggregate stock market behavior," *Journal of Political Economy*, Vol. 107, pp. 205–251.

Campbell, John Y. and Robert J. Shiller (1991) "Yield Spreads and Interest Rate Movements: A Bird's Eye View," *Review of Economic Studies*, Vol. 58, pp. 495–514.

Campbell, John Y and Motohiro Yogo (2006) "Efficient tests of stock return predictability," *Journal of financial economics*, Vol. 81, pp. 27–60.

Cattaneo, Matias D. and Richard K. Crump (2014) "Comment," *Journal of Business & Economic Statistics*, Vol. 32, pp. 324–329.

Cavanagh, Christopher L, Graham Elliott, and James H Stock (1995) "Inference in Models with Nearly Integrated Regressors," *Econometric theory*, Vol. 11, pp. 1131–1147.

Chan, Ngai Hang (1988) "The parameter inference for nearly nonstationary time series," *Journal of the American Statistical Association*, Vol. 83, pp. 857–862.

Cieslak, Anna and Pavol Povala (2015) "Expected Returns in Treasury Bonds," *Review of Financial Studies*, Vol. 28, pp. 2859–2901.

Cochrane, John H. and Monika Piazzesi (2005) "Bond Risk Premia," *American Economic Review*, Vol. 95, pp. 138–160.

Cooper, Ilan and Richard Priestley (2008) "Time-Varying Risk Premiums and the Output Gap," *Review of Financial Studies*, Vol. 22, pp. 2801–2833.

D'Amico, Stefania and Thomas B. King (2013) "Flow and stock effects of large-scale treasury purchases: Evidence on the importance of local supply," *Journal of Financial Economics*, Vol. 108, pp. 425–448.

Diebold, Francis X. and Robert S. Mariano (1995) "Comparing Predictive Accuracy," *Journal of Business & economic statistics*, Vol. 13, pp. 253–263.

Diebold, Francis X., Glenn D. Rudebusch, and S. Boragan Aruoba (2006) "The Macroeconomy and the Yield Curve: A Dynamic Latent Factor Approach," *Journal of Econometrics*, Vol. 131, pp. 309–338.

Duffee, Gregory R. (2011a) "Forecasting with the Term Structure: the Role of No-Arbitrage," Working Paper January, Johns Hopkins University.

——— (2011b) "Information In (and Not In) the Term Structure," *Review of Financial Studies*, Vol. 24, pp. 2895–2934.

——— (2013a) "Bond Pricing and the Macroeconomy," in Milton Harris George M. Constantinides and Rene M. Stulz eds. *Handbook of the Economics of Finance*, Vol. 2, Part B: Elsevier, pp. 907–967.

——— (2013b) "Forecasting Interest Rates," in Graham Elliott and Allan Timmermann eds. *Handbook of Economic Forecasting*, Vol. 2, Part A: Elsevier, pp. 385–426.

Fama, Eugene F. and Robert R. Bliss (1987) "The Information in Long-Maturity Forward Rates," *The American Economic Review*, Vol. 77, pp. 680–692.

Ferson, Wayne E, Sergei Sarkissian, and Timothy T Simin (2003) "Spurious Regressions in Financial Economics?" *Journal of Finance*, Vol. 58, pp. 1393–1414.

Giacoletti, Marco, Kristoffer T. Laursen, and Kenneth J. Singleton (2016) "Learning, Dispersion of Beliefs, and Risk Premiums in an Arbitrage-free Term Structure Model," unpublished manuscript.

Greenwood, Robin and Dimitri Vayanos (2014) "Bond Supply and Excess Bond Returns," *Review of Financial Studies*, Vol. 27, pp. 663–713.

Gürkaynak, Refet S. and Jonathan H. Wright (2012) "Macroeconomics and the Term Structure," *Journal of Economic Literature*, Vol. 50, pp. 331–367.

Hall, Peter and Susan R. Wilson (1991) "Two Guidelines for Bootstrap Hypothesis Testing," *Biometrics*, Vol. 47, pp. 757–762.

Hamilton, James D. (1994) *Time Series Analysis*: Princeton University Press.

Hamilton, James D. and Jing Cynthia Wu (2012) "Identification and estimation of Gaussian affine term structure models," *Journal of Econometrics*, Vol. 168, pp. 315–331.

——— (2014) "Testable Implications of Affine Term Structure Models," *Journal of Econometrics*, Vol. 178, pp. 231–242.

Hansen, Bruce E (1999) "The grid bootstrap and the autoregressive model," *Review of Economics and Statistics*, Vol. 81, pp. 594–607.

Hodrick, Robert J (1992) "Dividend yields and expected stock returns: Alternative procedures for inference and measurement," *Review of Financial studies*, Vol. 5, pp. 357–386.

Horowitz, Joel L. (2001) "The Bootstrap," in J.J. Heckman and E.E. Leamer eds. *Handbook of Econometrics*, Vol. 5: Elsevier, Chap. 52, pp. 3159–3228.

Ibragimov, Rustam and Ulrich K. Müller (2010) "t-Statistic Based Correlation and Heterogeneity Robust Inference," *Journal of Business and Economic Statistics*, Vol. 28, pp. 453–468.

Joslin, Scott, Marcel Priebsch, and Kenneth J. Singleton (2014) "Risk Premiums in Dynamic Term Structure Models with Unspanned Macro Risks," *Journal of Finance*, Vol. 69, pp. 1197–1233.

Kendall, M. G. (1954) "A note on bias in the estimation of autocorrelation," *Biometrika*, Vol. 41, pp. 403–404.

Kilian, Lutz (1998) "Small-sample confidence intervals for impulse response functions," *Review of Economics and Statistics*, Vol. 80, pp. 218–230.

King, Thomas B. (2013) "A Portfolio-Balance Approach to the Nominal Term Structure," Working Paper 2013-18, Federal Reserve Bank of Chicago.

Lewellen, Jonathan, Stefan Nagel, and Jay Shanken (2010) "A skeptical appraisal of asset pricing tests," *Journal of Financial Economics*, Vol. 96, pp. 175–194.

Litterman, Robert and J. Scheinkman (1991) "Common Factors Affecting Bond Returns," *Journal of Fixed Income*, Vol. 1, pp. 54–61.

Ludvigson, Sydney C. and Serena Ng (2009) "Macro Factors in Bond Risk Premia," *Review of Financial Studies*, Vol. 22, pp. 5027–5067.

Ludvigson, Sydney C and Serena Ng (2010) "A Factor Analysis of Bond Risk Premia," *Handbook of Empirical Economics and Finance*, p. 313.

Mankiw, N. Gregory and Matthew D. Shapiro (1986) "Do we reject too often? Small sample properties of tests of rational expectations models," *Economics Letters*, Vol. 20, pp. 139–145.

McCracken, Michael W. and Serena Ng (2014) "FRED-MD: A Monthly Database for Macroeconomic Research," working paper, Federal Reserve Bank of St. Louis.

Modigliani, Franco and Richard Sutch (1966) "Innovations in interest rate policy," *The American Economic Review*, pp. 178–197.

Müller, Ulrich K. (2014) "HAC Corrections for Strongly Autocorrelated Time Series," *Journal of Business and Economic Statistics*, Vol. 32.

Nabeya, Seiji and Bent E Sørensen (1994) "Asymptotic distributions of the least-squares estimators and test statistics in the near unit root model with non-zero initial value and local drift and trend," *Econometric Theory*, Vol. 10, pp. 937–966.

Newey, Whitney K and Kenneth D West (1987) "A Simple, Positive Semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, Vol. 55, pp. 703–08.

Phillips, Peter CB (1988) "Regression theory for near-integrated time series," *Econometrica: Journal of the Econometric Society*, pp. 1021–1043.

Pope, Alun L. (1990) "Biases of Estimators in Multivariate Non-Gaussian Autoregressions," *Journal of Time Series Analysis*, Vol. 11, pp. 249–258.

Rudebusch, Glenn D. and Eric T. Swanson (2012) "The bond premium in a DSGE Model with Long-Run Real and Nominal Risks," *American Economic Journal: Macroeconomics*, Vol. 4, pp. 105–143.

Stambaugh, Robert F. (1999) "Predictive regressions," *Journal of Financial Economics*, Vol. 54, pp. 375–421.

Stock, James H (1991) "Confidence intervals for the largest autoregressive root in US macroeconomic time series," *Journal of Monetary Economics*, Vol. 28, pp. 435–459.

Stock, James H. (1994) "Unit roots, structural breaks and trends," in Robert F. Engle and Daniel L. McFadden eds. *Handbook of Econometrics*, Vol. 4: Elsevier, Chap. 46, pp. 2739–2841.

Tobin, James (1969) "A general equilibrium approach to monetary theory," *Journal of money, credit and banking*, Vol. 1, pp. 15–29.

Vayanos, Dimitri and Jean-Luc Vila (2009) "A Preferred-Habitat Model of the Term Structure of Interest Rates," NBER Working Paper 15487, National Bureau of Economic Research.

Wei, Min and Jonathan H Wright (2013) "Reverse Regressions And Long-Horizon Forecasting," *Journal of Applied Econometrics*, Vol. 28, pp. 353–371.

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

Table 1: Simulation study of standard error bias

ρ	δ	Coefficient bias		SE bias	Size			
		β_1	β_2	(%)	simulated	asymptotic	bootstrap	IM
$\mu_1 = \mu_2 = 0$								
0.99	0.0	0.000	0.000	-4.7	0.050	0.047	0.048	0.044
0.00	1.0	-0.010	0.000	-0.6	0.050	0.051	0.050	0.048
0.90	1.0	-0.052	0.000	-15.4	0.085	0.086	0.057	0.050
0.99	0.8	-0.055	0.000	-23.2	0.113	0.112	0.072	0.046
0.99	1.0	-0.068	0.000	-29.8	0.151	0.151	0.082	0.048
$\mu_1 = 0, \mu_2 = 1$								
0.99	0.0	0.000	0.000	-5.1	0.050		0.049	0.039
0.00	1.0	-0.010	0.000	-0.5	0.050		0.050	0.048
0.90	1.0	-0.053	0.000	-17.1	0.089		0.057	0.048
0.99	0.8	-0.071	0.000	-42.4	0.183		0.077	0.031
0.99	1.0	-0.088	0.000	-50.8	0.268		0.085	0.029
$\mu_1 = 1, \mu_2 = 1$								
0.99	0.0	0.000	0.000	-4.0	0.050		0.047	0.043
0.00	1.0	-0.010	0.000	-0.5	0.050		0.050	0.047
0.90	1.0	-0.037	0.017	-12.0	0.081		0.054	0.051
0.99	0.8	-0.036	0.035	-12.1	0.168		0.056	0.345
0.99	1.0	-0.045	0.044	-16.0	0.241		0.058	0.488

Coefficient bias, standard error bias, and size distortions in simulation study with sample size $T = 100$ and DGP with $\beta_0 = \beta_1 = \beta_2 = 0$, $\sigma_1 = \sigma_2 = \sigma_u = 1$ for different values of $\rho_1 = \rho_2 = \rho$ and δ. The coefficient bias is reported as $E(\hat{\beta}_i) - \beta_i$. The standard error bias is reported as $E[(\hat{\sigma}_{\hat{\beta}_2}) - \sigma_{\hat{\beta}_2}]/\sigma_{\hat{\beta}_2}$. The last four columns report the size (i.e., frequency of rejections) of tests of $H_0 : \beta_2 = 0$ with a nominal size of five percent, for a conventional t-test—according to both regressions in simulated small samples and the local-to-unity asymptotic distribution—for the bootstrap test, and the Ibragimov-Mueller (IM) test. For details on the simulation study refer to main text.

Table 2: In-sample predictive power in excess-return regressions

	$\bar{R}^2_1$	$\bar{R}^2_2$	$\bar{R}^2_2 - \bar{R}^2_1$	$\bar{R}^2_1$	$\bar{R}^2_2$	$\bar{R}^2_2 - \bar{R}^2_1$
JPS	Original sample: 1985–2008			Later sample: 1985–2015		
Data	0.19	0.38	0.19	0.17	0.23	0.06
Bootstrap	0.32	0.38	0.06	0.28	0.33	0.05
	(0.11, 0.54)	(0.16, 0.60)	(-0.00, 0.20)	(0.08, 0.49)	(0.11, 0.54)	(-0.00, 0.17)
BC bootstrap	0.36	0.42	0.06	0.29	0.34	0.06
	(0.09, 0.61)	(0.15, 0.66)	(-0.00, 0.22)	(0.07, 0.53)	(0.11, 0.57)	(-0.00, 0.21)
Ludvigson-Ng	Original sample: 1964–2007			Later sample: 1985–2015		
Data	0.25	0.35	0.10	0.14	0.24	0.10
Bootstrap	0.21	0.24	0.03	0.29	0.35	0.06
	(0.05, 0.38)	(0.08, 0.42)	(-0.00, 0.11)	(0.08, 0.51)	(0.13, 0.56)	(-0.00, 0.19)
Cieslak-Povala	Original sample: 1971–2011			Later sample: 1985–2015		
Data	0.12	0.46	0.34	0.17	0.38	0.22
BC bootstrap	0.15	0.22	0.07	0.28	0.34	0.07
	(0.02, 0.34)	(0.06, 0.40)	(-0.00, 0.21)	(0.04, 0.53)	(0.11, 0.58)	(-0.00, 0.23)
Cochrane-Piazzesi	Original sample: 1964–2003			Later sample: 1985–2015		
Data	0.26	0.35	0.09	0.15	0.18	0.03
Bootstrap	0.21	0.22	0.01	0.29	0.31	0.01
	(0.05, 0.41)	(0.06, 0.41)	(0.00, 0.02)	(0.09, 0.52)	(0.10, 0.53)	(0.00, 0.05)

Adjusted $\bar{R}^2$ for regressions of annual excess bond returns on three PCs of the yield curve ($\bar{R}^2_1$) and on three yield PCs together with the additional proposed predictors x_{2t}, well as the difference in adjusted $\bar{R}^2$. The additional predictors, which are described in more detail in the text, are: for JPS, measures of growth and inflation; for Ludvigson-Ng, eight PCs of a large set of macro variables; for Cieslak-Povala, a moving-average estimate of the inflation trend; and for Cochrane-Piazzesi, the fourth and fifth PC of yields. The results in the left half of the table are for the original sample period in each paper; the right half of the table is for the 1985–20015 sample period. The excess bond return is an average across bond maturities: for JPS, from two to ten years; for Ludvigson-Ng, from two to five years; for Cieslak-Povala, from two to ten years (a weighted average); and for Cochrane-Piazzesi, from two to five years. The first row of each panel reports the values of the statistics in the original data. The following rows report bootstrap mean and 95%-confidence intervals (in parentheses). The bootstrap, which is described in the text, imposes the null hypothesis that x_{2t} has no incremental predictive power. For Cochrane-Piazzesi, the results are for the unadjusted R^2.

Table 3: Joslin-Priebsch-Singleton: statistical inference in excess-return regressions

	$PC1$	$PC2$	$PC3$	GRO	INF	Wald
Original sample: 1985–2008						
Coefficient	1.038	1.922	3.145	-0.022	-6.187	
HAC statistic	5.423	4.466	0.797	-2.537	-4.121	24.844
HAC p-value	**0.000**	**0.000**	0.426	**0.012**	**0.000**	**0.000**
Bootstrap 5% c.v.				3.157	3.848	22.634
Bootstrap p-value				0.108	**0.038**	**0.040**
BC bootstrap 5% c.v.				3.282	4.068	25.486
BC bootstrap p-value				0.116	**0.047**	0.052
IM $q = 8$	**0.004**	0.064	**0.011**	0.600	0.647	
IM $q = 16$	**0.001**	**0.009**	0.061	0.066	0.785	
Size						
HAC				0.204	0.283	0.369
Bootstrap				0.064	0.066	0.062
IM $q = 8$				0.052	0.051	
IM $q = 16$				0.046	0.050	
Power						
Bootstrap				0.198	0.917	0.895
IM $q = 8$				0.404	0.732	
IM $q = 16$				0.643	0.877	
Later sample: 1985–2015						
Coefficient	0.432	1.820	3.141	-0.002	-2.979	
HAC statistic	2.532	3.559	1.101	-0.269	-2.107	4.441
HAC p-value	**0.012**	**0.000**	0.272	0.788	**0.036**	0.109
Bootstrap 5% c.v.				3.100	3.545	18.894
Bootstrap p-value				0.847	0.231	0.432
BC bootstrap 5% c.v.				3.173	3.825	21.812
BC bootstrap p-value				0.858	0.255	0.473
IM $q = 8$	**0.001**	0.175	0.054	0.784	0.843	
IM $q = 16$	0.957	**0.010**	0.574	0.363	0.281	

Predictive regressions for annual excess bond returns, averaged over two- through ten-year bond maturities, using yield PCs and two macro variables that are described in the text. Results in the top panel are for the same sample period used in Joslin et al. (2014); the data used for the bottom panel is extended to December 2015. HAC statistics and p-values are calculated using Newey-West standard errors with 18 lags. The column "Wald" reports results for the χ^2 test that GRO and INF have no predictive power; the other columns report results for individual t-tests. We obtain bootstrap distributions of the test statistics under the null hypothesis that GRO and INF have no predictive power—the main text describes the design of the simple and bias-corrected (BC) bootstraps. Critical values (c.v.'s) are the 95th percentile of the bootstrap distribution of the test statistics, and p-values are the frequency of bootstrap replications in which the test statistics are at least as large as in the data. We also report p-values for t-tests using the methodology of Ibragimov and Müller (2010) (IM), splitting the sample into either 8 or 16 blocks. Under *Size* we report estimates of the size of the tests, based on simulations from the simple bootstrap under the null hypothesis. Under *Power* we report power estimates using a bootstrap under the alternative hypothesis, as described in the text. p-values below 5% are emphasized with bold face.

Table 4: In-sample vs. out-of-sample predictive power

	In-sample			Out-of-sample			
	R_1^2	R_2^2	MSE-ratio	Start	N	MSE-ratio	DM p-value
Joslin-Priebsch-Singleton	0.191	0.380	0.760	2009:1	72	2.392	**0.045**
Ludvigson-Ng	0.259	0.360	0.850	2008:1	84	0.751	0.365
Cieslak-Povala	0.149	0.489	0.599	2012:1	36	12.073	0.253
Cochrane-Piazzesi	0.267	0.344	0.891	2004:1	132	1.202	0.127

In-sample vs. out-of-sample (OOS) predictive power for excess bond returns (averaged across maturities) of a restricted model with three PCs and an unrestricted model with additional predictors as suggested in each of four published studies. The in-sample period is the original sample period used in each study. The OOS period starts 13 months after the end of the in-sample period and ends in December 2015. N indicates the number of OOS observations. The columns also show in-sample R^2 for the restricted and unrestricted model, the in-sample ratio of mean-squared-errors (MSE) for the unrestricted relative to the restricted model, and the OOS MSE ratio, as well as the p-value of the Diebold-Mariano (DM) test for equal forecast accuracy.

Table 5: Ludvigson-Ng: statistical inference in excess-return regressions

	$PC1$	$PC2$	$PC3$	$F1$	$F2$	$F3$	$F4$	$F5$	$F6$	$F7$	$F8$	Wald
Original sample: 1964–2007												
Coefficient	0.136	2.052	-5.014	0.742	0.146	-0.072	-0.528	-0.321	-0.576	-0.401	0.551	
HAC statistic	1.552	2.595	-2.724	1.855	0.379	-0.608	-1.912	-1.307	-2.220	-2.361	3.036	42.073
HAC p-value	0.121	**0.010**	**0.007**	0.064	0.705	0.543	0.056	0.192	**0.027**	**0.019**	**0.003**	**0.000**
Bootstrap 5% c.v.				2.551	2.495	2.230	2.515	2.435	2.610	2.455	2.326	29.477
Bootstrap p-value				0.147	0.753	0.580	0.131	0.284	0.096	0.062	**0.012**	**0.010**
IM $q = 8$	**0.001**	**0.001**	0.447	0.119	0.755	0.566	0.187	0.656	0.513	0.111	0.485	
IM $q = 16$	**0.000**	**0.014**	0.279	0.270	0.062	0.606	0.488	0.342	0.094	**0.021**	0.693	
Size												
HAC				0.128	0.112	0.082	0.125	0.109	0.135	0.121	0.095	0.323
Bootstrap				0.056	0.056	0.049	0.049	0.045	0.055	0.061	0.055	0.053
IM q=8				0.047	0.047	0.049	0.050	0.044	0.052	0.048	0.043	
IM q=16				0.052	0.053	0.050	0.050	0.050	0.048	0.054	0.046	
Power												
Bootstrap				0.462	0.067	0.083	0.407	0.194	0.360	0.493	0.785	0.951
IM q=8				0.424	0.066	0.091	0.433	0.214	0.414	0.550	0.892	
IM q=16				0.642	0.082	0.109	0.695	0.331	0.676	0.774	0.988	
Later sample: 1985–2015												
Coefficient	0.226	1.227	-0.977	-0.895	0.848	0.025	0.227	0.011	-0.099	0.074	-0.278	
HAC statistic	2.801	1.258	-0.297	-2.505	1.920	0.223	1.009	0.044	-0.445	0.442	-1.070	29.170
HAC p-value	**0.005**	0.209	0.767	**0.013**	0.056	0.824	0.314	0.965	0.657	0.659	0.286	**0.000**
Bootstrap 5% c.v.				3.003	3.338	2.705	3.133	3.034	2.862	2.525	2.593	45.119
Bootstrap p-value				0.095	0.238	0.876	0.501	0.977	0.744	0.726	0.417	0.179
IM $q = 8$	**0.003**	0.085	0.501	0.203	0.426	0.696	0.250	0.759	0.340	0.219	0.347	
IM $q = 16$	**0.004**	0.171	0.469	0.579	0.997	0.629	0.676	0.742	0.731	0.329	0.891	
Size												
HAC				0.186	0.229	0.150	0.200	0.169	0.168	0.119	0.135	0.550

Predictive regressions for annual excess bond returns, averaged over two- through five-year bond maturities, using yield PCs and factors from a large data set of macro variables, as in Ludvigson and Ng (2010). The top panel shows the results for the original data set used by Ludvigson and Ng (2010); the bottom panel uses a data sample that starts in 1985 and ends in 2015. The bootstrap is a simple bootstrap without bias correction. For a description of the statistics in each row, see the notes to Table 3. p-values below 5% are emphasized with bold face.

Table 6: Cieslak-Povala: statistical inference in excess-return regressions

	$PC1$	$PC2$	$PC3$	τ
Original sample: 1971–2011				
Only yield PCs				
Coefficient	0.003	0.293	-0.028	
RR t-statistic	0.505	2.266	-0.338	
RR p-value	0.614	**0.024**	0.735	
Yield PCs plus inflation trend				
Coefficient	0.191	0.604	0.341	-1.019
RR t-statistic	5.429	5.367	1.240	-6.173
RR p-value	**0.000**	**0.000**	0.215	**0.000**
BC bootstrap RR 5% c.v.				3.572
BC bootstrap RR p-value				**0.001**
IM $q = 8$	**0.001**	**0.026**	**0.046**	0.240
IM $q = 16$	**0.000**	**0.033**	0.523	0.823
Size				
RR				0.445
IM $q = 8$				0.138
IM $q = 16$				0.136
Power				
IM $q = 8$				0.406
IM $q = 16$				0.478
Later sample: 1985–2015				
Only yield PCs				
Coefficient	0.023	0.221	0.321	
RR t-statistic	1.985	1.194	0.814	
RR p-value	**0.048**	0.233	0.416	
Yield PCs plus inflation trend				
Coefficient	0.136	0.430	0.645	-0.693
RR t-statistic	4.520	3.449	2.299	-3.694
RR p-value	**0.000**	**0.001**	**0.022**	**0.000**
BC bootstrap RR 5% c.v.				3.619
BC bootstrap RR p-value				**0.043**
IM $q = 8$	**0.003**	0.896	0.156	0.649
IM $q = 16$	**0.001**	0.103	0.964	0.584

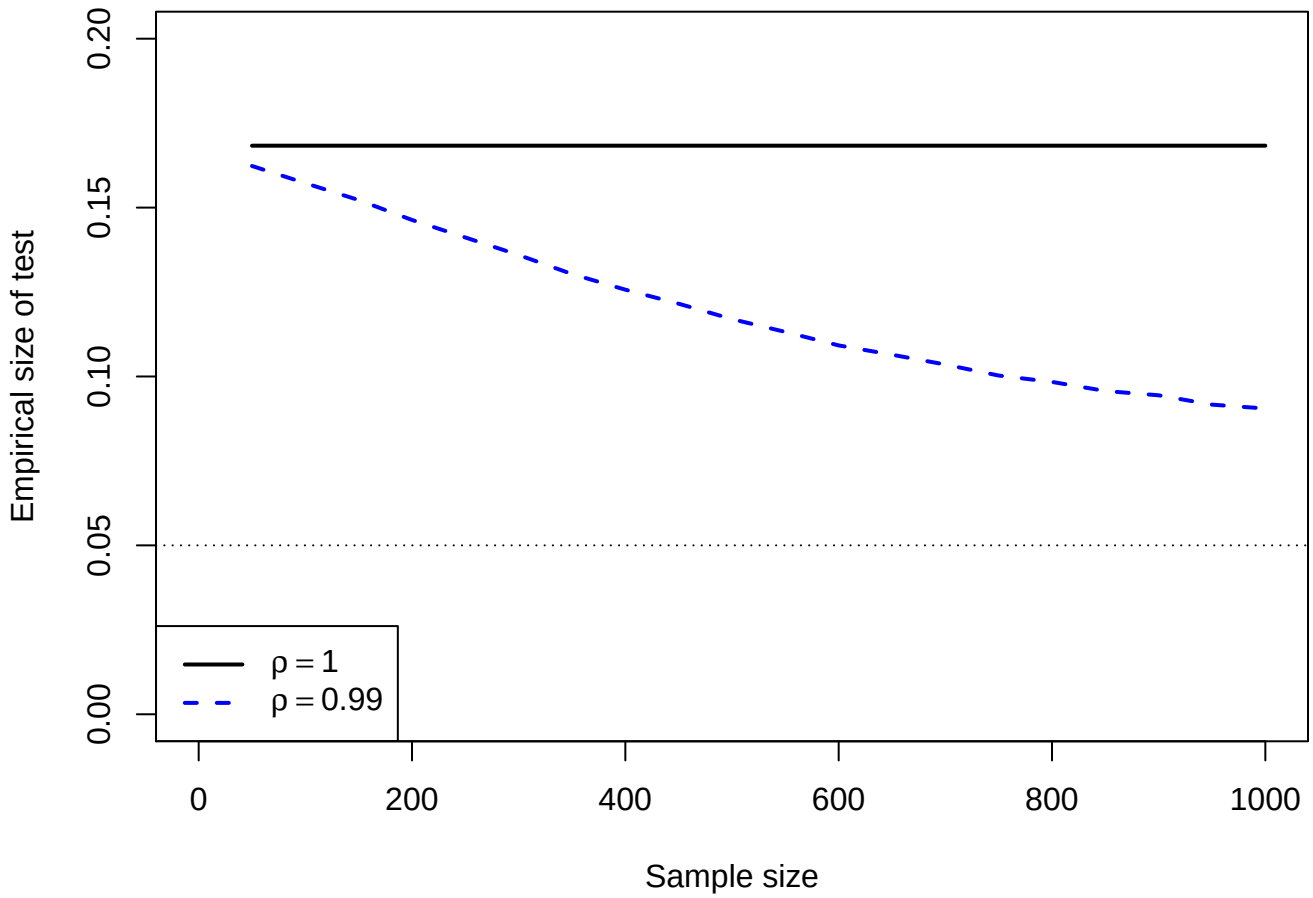
Predictive regressions for annual excess bond returns (weighted average over two- through ten-year bond maturities) using yield PCs and a moving-average estimate of inflation trend. The data used for the top panel covers the same sample period as in Cieslak and Povala (2015); the data used for the bottom panel starts in 1985 and ends in 2015. Reverse regression (RR) statistics and p-values are calculated using the reverse regression delta method of Wei and Wright (2013). We obtain bootstrap distributions of the test statistics under the null hypothesis that only PCs have predictive power, in order to calculate bootstrap critical values and p-values, and to estimate the size of tests. We also report p-values for t-tests using the methodology of Ibragimov and Müller (2010) (IM), splitting the sample into either 8 or 16 blocks. The last two rows of the top panel report the power of the IM test using a bootstrap under the alternative hypothesis. See the text for a description of the bootstrap designs. p-values below 5% are emphasized with bold face.

Table 7: Cochrane-Piazzesi: statistical inference in excess-return regressions

	$PC1$	$PC2$	$PC3$	$PC4$	$PC5$	Wald
Original sample: 1964–2003						
Data	0.127	2.740	-6.307	-16.128	-2.038	
HAC statistic	1.724	5.205	2.950	5.626	0.748	31.919
HAC p-value	0.085	**0.000**	**0.003**	**0.000**	0.455	**0.000**
Bootstrap 5% c.v.				2.263	2.198	8.183
Bootstrap p-value				**0.000**	0.491	**0.000**
IM $q = 8$	**0.003**	**0.013**	0.988	0.063	0.161	
IM $q = 16$	**0.000**	**0.020**	0.605	0.876	0.126	
Size						
HAC				0.085	0.075	0.106
Bootstrap				0.051	0.052	0.055
IM $q = 8$				0.048	0.045	
IM $q = 16$				0.051	0.041	
Power						
Bootstrap				0.996	0.151	0.992
IM $q = 8$				0.992	0.109	
IM $q = 16$				0.995	0.123	
Later sample: 1985–2015						
Data	0.105	1.595	3.512	-9.049	-9.537	
HAC statistic	1.862	2.248	0.990	-1.328	-1.291	4.020
HAC p-value	0.063	**0.025**	0.323	0.185	0.198	0.134
Bootstrap 5% c.v.				2.435	2.396	9.639
Bootstrap p-value				0.276	0.282	0.264
IM $q = 8$	**0.001**	0.245	0.116	0.675	0.201	
IM $q = 16$	**0.001**	0.077	0.301	0.150	0.865	

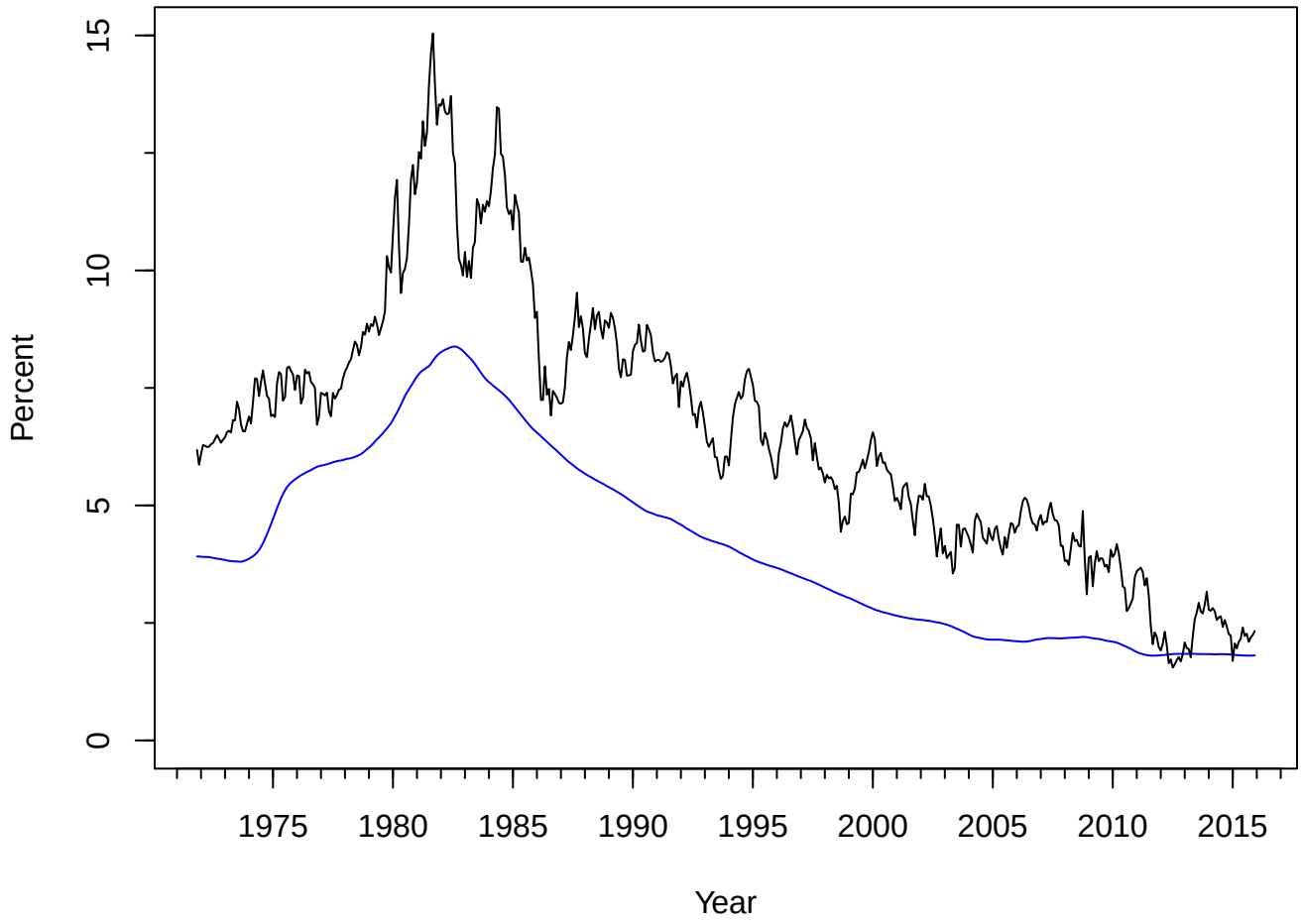
Predicting annual excess bond returns, averaged over two- through five-year bonds, using principal components (PCs) of yields. The null hypothesis is that the first three PCs contain all the relevant predictive information. The data used in the top panel is the same as in Cochrane and Piazzesi (2005)—see in particular their table 4. HAC statistics and p-values are calculated using Newey-West standard errors with 18 lags. Bootstrap distributions are obtained under the null hypothesis, using the bootstrap procedure described in the main text. We also report p-values for t-tests using the methodology of Ibragimov and Müller (2010) (IM), splitting the sample into either 8 or 16 blocks. Under *Size* we report estimates of the size of the tests based on the bootstrap samples. Under *Power* we report power estimates using a bootstrap under the alternative hypothesis, as described in the text. p-values below 5% are emphasized with bold face.

Figure 1: Size distortions and sample size in simulation study



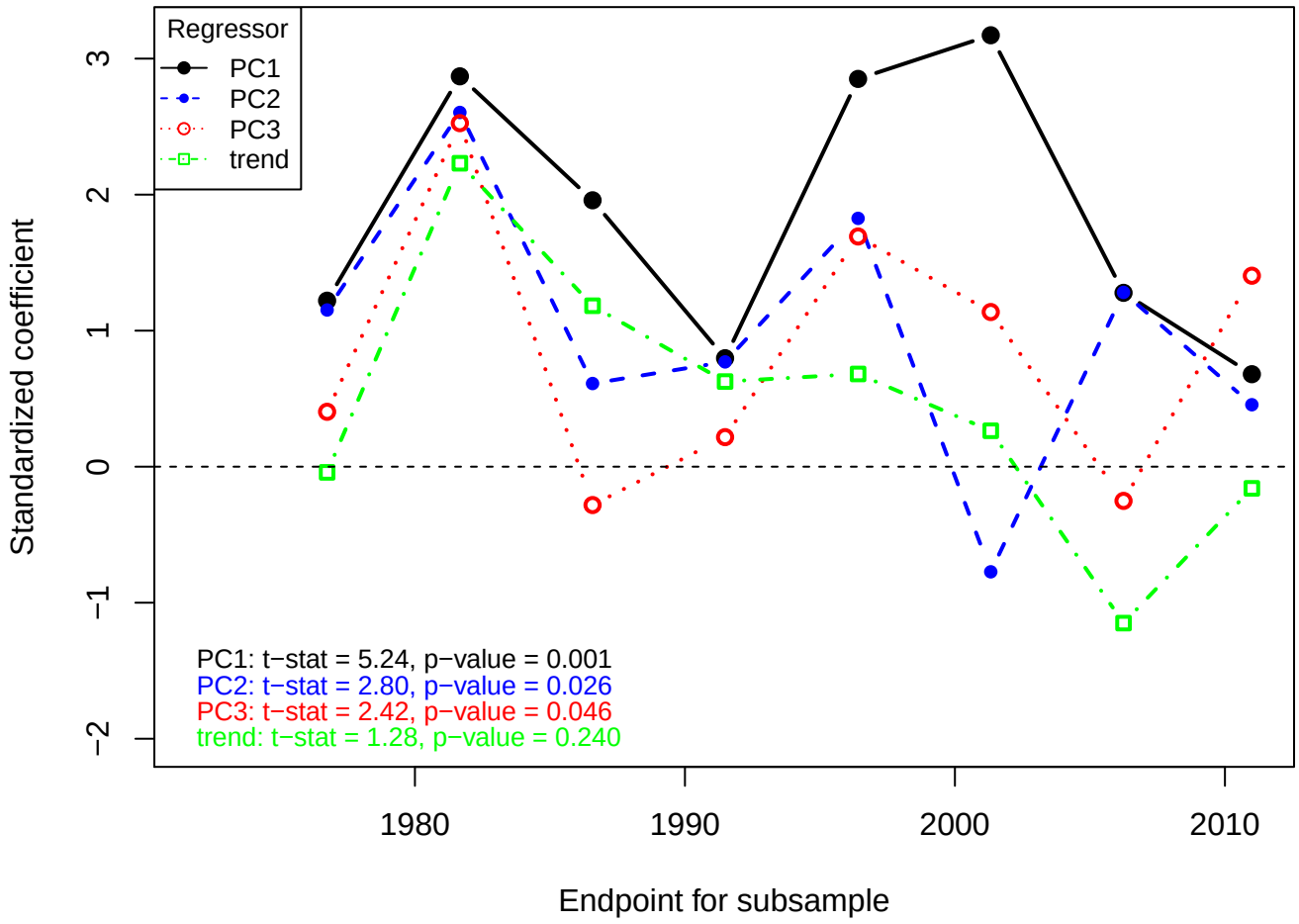
True size of conventional t-test of $H_0 : \beta_2 = 0$ with nominal size of 5% according to local-to-unity asymptotic distribution, for different sample sizes. DGP parameters are $\delta = 1$, $c_1 = c_2 = 0$, $\beta_0 = \beta_1 = \beta_2 = 0$, $\sigma_1 = \sigma_2 = \sigma_u = 1$, and $\rho_1 = \rho_2 = \rho$ either equal to one or 0.99. For details on the simulation study refer to main text.

Figure 2: Cieslak-Povala: ten-year yield and inflation trend



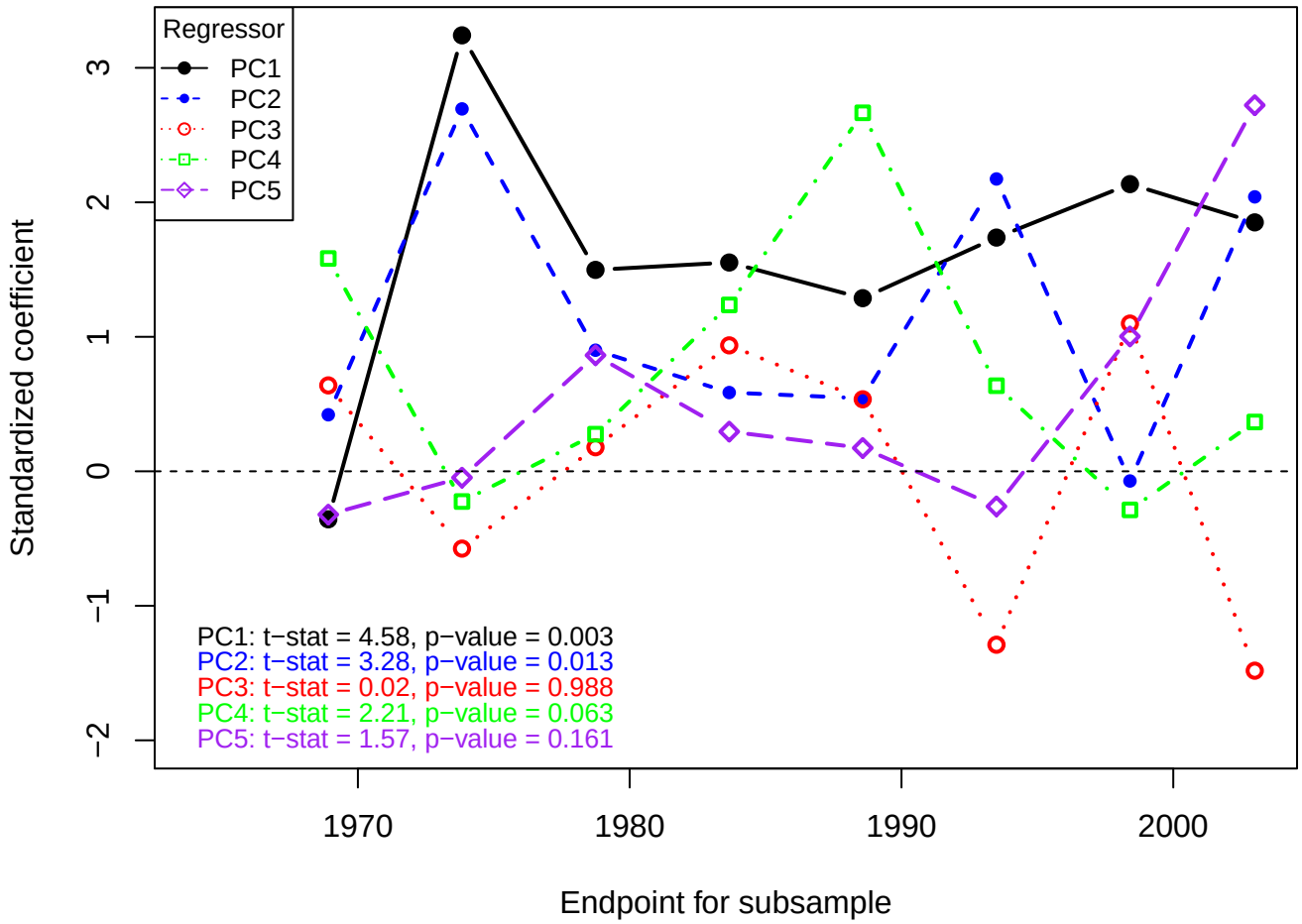
Bold line: yield on 10-year bond. Thin line: trend inflation as estimated by equation 17.

Figure 3: Cieslak-Povala: predictive power across subsamples



Standardized coefficients on principal components (PCs) and inflation trend across eight different subsamples, ending at the indicated point in time. Standardized coefficients are calculated by dividing through the sample standard deviation of the coefficient across the eight samples. Text labels indicate t-statistics and p-values of the Ibragimov-Mueller test with $q = 8$. Note that the t-statistics are equal to means of the standardized coefficients multiplied by $\sqrt{8}$. The sample period is the same as in Cieslak and Povala (2015).

Figure 4: Cochrane-Piazzesi: predictive power across subsamples



Standardized coefficients on principal components (PCs) across eight different subsamples, ending at the indicated point in time. Standardized coefficients are calculated by dividing through the sample standard deviation of the coefficient across the eight samples. Text labels indicate t-statistics and p-values of the Ibragimov-Mueller test with $q = 8$. Note that the t-statistics are equal to means of the standardized coefficients multiplied by $\sqrt{8}$. The data and sample period is the same as in Cochrane and Piazzesi (2005).

Appendix

A Derivations of theoretical results

A.1 Derivations for Section 2.1

Let $y = (y_{1+h}, y_{2+h}, ..., y_{T+h})'$ and stack x'_{1t} and x'_{2t} into $(T \times K_1)$ and $(T \times K_2)$ matrices denoted X_1 and X_2. Note that the OLS estimates of equation (1) satisfy

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix}.$$

Premultiply the first row by $X_2'X_1(X_1'X_1)^{-1}$ and subtract the result from the second,

$$(X_2'M_1X_2)b_2 = X_2'M_1y,$$

for $M_1 = I_T - X_1(X_1'X_1)^{-1}X_1'$. Using the fact that M_1 is symmetric and idempotent we have

$$X_2'M_1X_2 = (M_1X_2)'M_1X_2 = \sum \tilde{x}_{2t}\tilde{x}'_{2t} \tag{18}$$

$$b_2 = \left(\sum \tilde{x}_{2t}\tilde{x}'_{2t}\right)^{-1}\left(\sum \tilde{x}_{2t}y_{t+h}\right). \tag{19}$$

Substituting equation (1) into (19) and using the facts that $\sum \tilde{x}_{2t}x'_{1t} = 0$ (by the orthogonality property of residuals) and that $\sum \tilde{x}_{2t}x'_{2t} = \sum \tilde{x}_{2t}\tilde{x}'_{2t}$ (again by idempotence of M_1) gives

$$b_2 = \beta_2 + \left(\sum \tilde{x}_{2t}\tilde{x}'_{2t}\right)^{-1}\left(\sum \tilde{x}_{2t}u_{t+h}\right) \tag{20}$$

from which the Wald test is

$$\begin{aligned}&(b_2 - \beta_2)'s^{-2}\textstyle\sum_{t=1}^{T}\tilde{x}_{2t}\tilde{x}'_{2t}(b_2 - \beta_2) \\ &= \left(\textstyle\sum_{t=1}^{T} u_{t+1}\tilde{x}'_{2t}\right)\left(s^2\textstyle\sum_{t=1}^{T}\tilde{x}_{2t}\tilde{x}'_{2t}\right)^{-1}\left(\textstyle\sum_{t=1}^{T}\tilde{x}_{2t}u_{t+1}\right)\end{aligned}$$

as claimed in (2)

Note that if $u|X_1, X_2 \sim N(0, \sigma_u^2 I_T)$, then K_2^{-1} times expression (2) would have an exact $F(K_2, T - K_1 - K_2)$ distribution for every sample size T and any stationary or nonstationary process for x_{2t}. Under the weaker assumption that $E(u_{t+1}|x_t, x_{t-1}, ..., x_1) = 0$ but $E(u_t|x_t, x_{t-1}, ..., x_1) \neq 0$, the Wald statistic (2) will still be asymptotically $\chi^2(K_2)$ under standard first-order stationary asymptotics, as can be seen from equation (35) below for the special case $h = 1$ and $S = \sigma_u^2 Q$. The problems arise when x_{1t} is correlated with u_t and furthermore x_t is highly persistent. In the case of unit-root processes these problems give (2) an asymptotic distribution that is not $\chi^2(K_2)$, and for near-unit-root processes they cause the small-sample distribution to be quite different from a $\chi^2(K_2)$.

The unit-root derivations this next paragraph assume a functional central limit theorem $T^{-1/2}x_{i,[T\lambda]} \Rightarrow B_i(\lambda)$ for $i = 1, 2$ with $0 \leq \lambda \leq 1$, $[T\lambda]$ the largest integer less than or equal to $T\lambda$, $B_i(\lambda)$ K_i-dimensional Brownian motion, and $\Rightarrow$ denoting weak convergence in probability measure. From the FCLT and the Continuous Mapping Theorem,

$$\begin{aligned}\hat{A}_T &= \left[T^{-1}\int_0^1 x_{2,[T\lambda]}x'_{1,[T\lambda]}d\lambda\right]\left[T^{-1}\int_0^1 x_{1,[T\lambda]}x'_{1,[T\lambda]}d\lambda\right]^{-1} \\ &\Rightarrow \left[\int_0^1 B_2(\lambda)B_1(\lambda)'d\lambda\right]\left[\int_0^1 B_1(\lambda)B_1(\lambda)'d\lambda\right]^{-1} \\ &\equiv \tilde{A}.\end{aligned}$$

Notice that

$$\begin{aligned}\sum_{t=1}^T \tilde{x}_{2t}u_{t+1} &= \sum_{t=1}^T x_{2t}u_{t+1} - \hat{A}_T\sum_{t=1}^T x_{1t}u_{t+1} \\ &= \sum_{t=1}^T x_{2t}u_{t+1} - \sum_{t=1}^T x_{2t}x'_{1t}Z_T \qquad (21)\end{aligned}$$

for $Z_T = \left(\sum_{t=1}^T x_{1t}x'_{1t}\right)^{-1}\left(\sum_{t=1}^T x_{1t}u_{t+1}\right)$ If x_{1t} is a unit-root process that is correlated with the lag of u_{t+1}, Z_T will have a nonstandard distribution. For example, if x_{1t} is a scalar random walk with $x_{1,t+1} = x_{1t} + u_{t+1}$, then Z_T has the same distribution as $\hat{\rho}_T - 1$ where $\hat{\rho}_T$ is the OLS coefficient from a regression of $x_{1,t+1}$ on x_{1t}, a distribution with a negative bias that is well-known from unit root regressions.[49] If x_{2t} is uncorrelated with x_{1t}, then unlike the Dicky-Fuller distribution, the second term in (21) is symmetric around zero and is uncorrelated with the first term, so that the variance of $\sum_{t=1}^T \tilde{x}_{2t}u_{t+1}$ is strictly greater than that of $\sum_{t=1}^T x_{2t}u_{t+1}$.

A.2 Derivations for Section 2.2

For our local-to-unity results we assume as in Stock (1994, eq (2.17)) that $T^{-1/2}x_{i,[T\lambda]} \Rightarrow \sigma_i J_{c_i}(\lambda)$. We first note from Phillips (1988, Lemma 3.1(d)) that

$$T^{-2}\sum(x_{1t} - \bar{x}_1)^2 \Rightarrow \sigma_1^2\left\{\int_0^1 [J_{c_1}(\lambda)]^2 d\lambda - \left[\int_0^1 [J_{c_1}(\lambda)]d\lambda\right]^2\right\} = \sigma_1^2\int [J_{c_1}^{\mu}]^2$$

where in the sequel our notation suppresses the dependence on λ and lets $\int$ denote integration over λ from 0 to 1. The analogous operation applied to the numerator of (7) yields

$$A_T = \frac{T^{-2}\sum(x_{1t} - \bar{x}_1)(x_{2t} - \bar{x}_2)}{T^{-2}\sum(x_{1t} - \bar{x}_1)^2} \Rightarrow \frac{\sigma_1\sigma_2\int J_{c_1}^{\mu}J_{c_2}^{\mu}}{\sigma_1^2\int [J_{c_1}^{\mu}]^2}$$

[49] See for example Hamilton (1994, eq [17.4.7])

as claimed in (7). Also

$$T^{-1/2}\bar{x}_2 = T^{-3/2}\sum x_{2t} = \int_0^1 T^{-1/2} x_{2,[T\lambda]} d\lambda \Rightarrow \sigma_2 \int_0^1 J_{c_2}(\lambda) d\lambda.$$

Since $\tilde{x}_{2t} = x_{2t} - \bar{x}_2 - A_T(x_{1t} - \bar{x}_1)$,

$$\begin{aligned} T^{-1/2}\tilde{x}_{2,[T\lambda]} &\Rightarrow \sigma_2 \left\{ J_{c_2}(\lambda) - \int_0^1 J_{c_2}(s)ds - A\left[J_{c_1}(\lambda) - \int_0^1 J_{c_1}(s)ds \right] \right\} \\ &= \sigma_2 \left\{ J^{\mu}_{c_2}(\lambda) - A J^{\mu}_{c_1}(\lambda) \right\} = \sigma_2 K_{c_1,c_2}(\lambda) \end{aligned}$$

$$T^{-2}\sum \tilde{x}^2_{2t} = \int_0^1 \{T^{-1/2}\tilde{x}_{2,[T\lambda]}\}^2 d\lambda \Rightarrow \sigma_2^2 \int_0^1 \{K_{c_1,c_2}(\lambda)\}^2 d\lambda. \tag{22}$$

Note we can write

$$\begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ u_t \end{bmatrix} = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ \delta\sigma_u & 0 & \sqrt{1-\delta^2}\sigma_u \end{bmatrix} \begin{bmatrix} v_{1t} \\ v_{2t} \\ v_{0t} \end{bmatrix}$$

where $(v_{1t}, v_{2t}, v_{0t})'$ is a martingale-difference sequence with unit variance matrix. From Lemma 3.1(e) in Phillips (1988) we see

$$\begin{aligned} T^{-1}\sum \tilde{x}_{2t}u_{t+1} &= T^{-1}\sum [x_{2t} - \bar{x}_2 - A_T(x_{1t} - \bar{x}_1)](\delta\sigma_u v_{1,t+1} + \sqrt{1-\delta^2}\sigma_u v_{0,t+1}) \\ &\Rightarrow \delta\sigma_2\sigma_u \int K_{c_1,c_2} dW_1 + \sqrt{1-\delta^2}\sigma_2\sigma_u \int K_{c_1,c_2} dW_0. \end{aligned} \tag{23}$$

Recalling (2), the t-test of a true null hypothesis about β_2 can be written as

$$\tau = \frac{\sum \tilde{x}_{2t}u_{t+1}}{\{s^2 \sum \tilde{x}^2_{2t}\}^{1/2}} = \frac{T^{-1}\sum \tilde{x}_{2t}u_{t+1}}{\{s^2 T^{-2}\sum \tilde{x}^2_{2t}\}^{1/2}} \tag{24}$$

where

$$s^2 \xrightarrow{p} \sigma_u^2. \tag{25}$$

Substituting (25), (23), and (22) into (24) produces

$$\tau \Rightarrow \frac{\sigma_2\sigma_u \left\{ \delta \int K_{c_1,c_2} dW_1 + \sqrt{1-\delta^2} \int K_{c_1,c_2} dW_0 \right\}}{\left\{ \sigma_u^2 \sigma_2^2 \int (K_{c_1,c_2})^2 \right\}^{1/2}}$$

as claimed in (8).

Last we demonstrate that the variance of Z_1 exceeds unity. We can write

$$Z_1 = \frac{\int_0^1 J^{\mu}_{c_2}(\lambda) dW_1(\lambda)}{\left\{ \int_0^1 [K_{c_1,c_2}(\lambda)]^2 d\lambda \right\}^{1/2}} - \frac{A \int_0^1 J^{\mu}_{c_1}(\lambda) dW_1(\lambda)}{\left\{ \int_0^1 [K_{c_1,c_2}(\lambda)]^2 d\lambda \right\}^{1/2}} \tag{26}$$

Consider the denominator in these expressions, and note that

$$\begin{aligned}\int_0^1 [J_{c_2}^{\mu}(\lambda)]^2 d\lambda &= \int_0^1 [J_{c_2}^{\mu}(\lambda) - AJ_{c_1}^{\mu}(\lambda) + AJ_{c_1}^{\mu}(\lambda)]^2 d\lambda \\ &= \int_0^1 [K_{c_1,c_2}(\lambda)]^2 d\lambda + \int_0^1 [AJ_{c_1}^{\mu}(\lambda)]^2 d\lambda \\ &> \int_0^1 [K_{c_1,c_2}(\lambda)]^2 d\lambda\end{aligned}$$

where the cross-product term dropped out in the second equation by the definition of A in (7). This means that the following inequality holds for all realizations:

$$\left| \frac{\int_0^1 J_{c_2}^{\mu}(\lambda) dW_1(\lambda)}{\left\{ \int_0^1 [K_{c_1,c_2}(\lambda)]^2 d\lambda \right\}^{1/2}} \right| > \left| \frac{\int_0^1 J_{c_2}^{\mu}(\lambda) dW_1(\lambda)}{\left\{ \int_0^1 [J_{c_2}^{\mu}(\lambda)]^2 d\lambda \right\}^{1/2}} \right|. \tag{27}$$

Adapting the argument made in footnote 13, the magnitude inside the absolute-value operator on the right side of (27) can be seen to have a $N(0,1)$ distribution. Inequality (27) thus establishes that the first term in (26) has a variance that is greater than unity. The second term in (26) turns out to be uncorrelated with the first, and hence contributes additional variance to Z_1, although we have found that the first term appears to be the most important factor.[50] In sum, these arguments show that $\text{Var}(Z_1) > 1$.

A.3 Derivations for Section 2.3

First consider the case when $\rho_1 = \rho_2 = 1$, $\mu_1 = 0$, $\mu_2 \neq 0$, and $Corr(\varepsilon_{1t}, u_t) = 1$. Then $T^{-1/2} x_{1,[T\lambda]} \Rightarrow \sigma_1 W_1(\lambda)$ for $W_1(\lambda)$ standard Brownian motion, $T^{-1/2} \sum_{t=1}^{T} u_{t+1} \Rightarrow \sigma_1 W_1(1)$, while $x_{2t} = \mu_2 t + \sum_{s=1}^{t} \varepsilon_{2s}$ gives $T^{-1} x_{2,[T\lambda]} \Rightarrow \mu_2 \lambda$ as in Hamilton (1994, pp. 495-497). Let $x_t = (1, x_{1t}, x_{2t})'$ so $b = \beta + \left(\sum_{t=1}^{T} x_t x_t' \right)^{-1} \sum_{t=1}^{T} x_t u_{t+1}$. Define

$$\Upsilon_T = \begin{bmatrix} T^{1/2} & 0 & 0 \\ 0 & T & 0 \\ 0 & 0 & T^{3/2} \end{bmatrix}.$$

[50] These claims are based on moments of the respective functionals as estimated from discrete approximations to the Ornstein-Uhlenbeck processes.

Then very similar algebra to that in Hamilton (1994, pp. 498-500) gives

$$\begin{aligned}\Upsilon_T(b-\beta) &= \left[\Upsilon_T^{-1}\textstyle\sum x_t x_t'\Upsilon_T^{-1}\right]^{-1}\left[\Upsilon_T^{-1}\textstyle\sum x_t u_{t+1}\right] \\ &\Rightarrow \begin{bmatrix} 1 & \sigma_1\int W_1(\lambda) & \mu_2/2 \\ \sigma_1\int W_1(\lambda) & \sigma_1^2\int[W_1(\lambda)]^2 & \mu_2\sigma_1\int \lambda W_1(\lambda) \\ \mu_2/2 & \mu_2\sigma_1\int \lambda W_1(\lambda) & \mu_2^2/3 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_1 W_1(1) \\ (1/2)\sigma_1^2[W^2(1)-1] \\ \mu_2\sigma_1[W_1(1)-\int W_1(\lambda)] \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \sigma_1/\mu_2 \end{bmatrix} \begin{bmatrix} 1 & \int W_1(\lambda) & 1/2 \\ \int W_1(\lambda) & \int[W_1(\lambda)]^2 & \int \lambda W_1(\lambda) \\ 1/2 & \int \lambda W_1(\lambda) & 1/3 \end{bmatrix}^{-1} \begin{bmatrix} W_1(1) \\ (1/2)[W^2(1)-1] \\ W_1(1)-\int W_1(\lambda) \end{bmatrix}.\end{aligned}$$

Observe that the middle element, $T(b_1-\beta_1)$ is the identical distribution as that of $T(\hat\rho-1)$ in the Case 4 unit root distribution in Hamilton (1994, p. 499), and the t-statistic $(b_1-\beta_1)/\hat\sigma_{b_1}$ is identical to the Case 4 Dickey-Fuller t statistic (Hamilton (1994, eq [17.4.55])).

Consider next the case when $\rho_1 = \rho_2 = 1$, $\mu_1 \neq 0$, $\mu_2 \neq 0$, $Corr(\varepsilon_{1t}, u_t) = 1$, and $Corr(\varepsilon_{1t}, \varepsilon_{2s}) = 0$ for all s. Let's evaluate first the characteristics of a transformed regression of y_{t+1} on $\tilde x_t = Hx_t$ for

$$H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -\mu_1/\mu_2 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\tilde b = (\textstyle\sum \tilde x_t \tilde x_t')^{-1}\sum \tilde x_t y_{t+1} = (H')^{-1}b$$
$$\tilde\beta = (H')^{-1}\beta.$$

Then

$$\begin{aligned}\tilde x_{1t} &= x_{1t} - (\mu_1/\mu_2)x_{2t} \\ &= \mu_1 t + \textstyle\sum_{s=1}^t \varepsilon_{1s} - (\mu_1/\mu_2)\left(\mu_2 t + \sum_{s=1}^t \varepsilon_{2s}\right) \\ &= \textstyle\sum_{s=1}^t \varepsilon_{1s} - (\mu_1/\mu_2)\sum_{s=1}^t \varepsilon_{2s}\end{aligned}$$

and

$$\begin{aligned}T^{-1/2}\tilde x_{1,[T\lambda]} &\Rightarrow \sigma_1 W_1(\lambda) - (\mu_1/\mu_2)\sigma_2 W_2(\lambda) \\ &\equiv \kappa(\lambda)\end{aligned}$$

$$\Upsilon_T(\tilde b - \tilde\beta) \Rightarrow \begin{bmatrix} 1 & \int \kappa(\lambda) & \mu_2/2 \\ \int \kappa(\lambda) & \int[\kappa(\lambda)]^2 & \mu_2\int \lambda\kappa(\lambda) \\ \mu_2/2 & \mu_2\int \lambda\kappa(\lambda) & \mu_2^2/3 \end{bmatrix}^{-1} \begin{bmatrix} \sigma_1 W_1(1) \\ \sigma_1\int \kappa(\lambda) dW_1 \\ \mu_2\sigma_1[W_1(\lambda) - \int W_1(\lambda)] \end{bmatrix}.$$

The middle element, $T(\tilde b_1 - \beta_1)$, has a distribution that approaches the Dickey-Fuller Case 4 as $\sigma_2 \to 0$ and is a related unit-root distribution for general $\sigma_2 > 0$.

Translating back in terms of the original regression, we have $b = H'\tilde b$, $b_1 = \tilde b_1$,

$$b_2 = \tilde b_2 - (\mu_1/\mu_2)\tilde b_1 = \tilde b_2 - (\mu_1/\mu_2)b_1$$

$$T(b_2 - \beta_2) = T(\tilde{b}_2 - \tilde{\beta}_2) - (\mu_1/\mu_2)T(b_1 - \beta_1)$$
$$\Rightarrow 0 - (\mu_1/\mu_2)T(b_1 - \beta_1)$$

since $T^{3/2}(\tilde{b}_2 - \tilde{\beta}_2) \sim O_p(1)$. Thus $b_2 - \beta_2$ has the same asymptotic distribution as $-(\mu_1/\mu_2)(b_1 - \beta_1)$, with t-tests on either b_1 or b_2 having a distribution related to the Dickey-Fuller Case 4. When x_{1t} and x_{2t} share the same trend ($\mu_1 = \mu_2$), the distribution of b_2 will simply be the negative of that of b_1.

By contrast, if we were to regress $y_{t+1} = \beta_0 + \beta_1 x_{1t} + u_{t+1}$ on x_{1t} alone, or $y_{t+1} = \beta_0 + \beta_2 x_{2t} + u_{t+1}$ on x_{2t} alone, t-tests on β_1 or β_2 would be asymptotically $N(0,1)$, from the same algebra as in Hamilton (1994, pp. 495-497). Thus for example if the true $\beta_1 \neq 0$ and $\beta_2 = 0$, when we do the regression on x_{1t} alone we would have perfectly appropriate tests about β_1, but if we add x_{2t} to the regression, tests about both β_1 and β_2 become distorted and x_{2t} could spuriously appear to be helpful in improving the estimate of β_1.

A.4 Derivations for Section 2.4

Note from (18) that

$$\sum \tilde{x}_{2t}\tilde{x}'_{2t} = \sum x_{2t}x'_{2t} - \left(\sum x_{2t}x'_{1t}\right)\left(\sum x_{1t}x'_{1t}\right)^{-1}\left(\sum x_{1t}x'_{2t}\right).$$

If x_t is covariance-stationary and ergodic for second moments,

$$\begin{aligned} T^{-1}\sum \tilde{x}_{2t}\tilde{x}'_{2t} &= T^{-1}\sum x_{2t}x'_{2t} - \left(T^{-1}\sum x_{2t}x'_{1t}\right)\left(T^{-1}\sum x_{1t}x'_{1t}\right)^{-1}\left(T^{-1}\sum x_{1t}x'_{2t}\right) \\ &\xrightarrow{p} E(x_{2t}x'_{2t}) - E(x_{2t}x'_{1t})\left[E(x_{1t}x'_{1t})\right]^{-1}E(x_{1t}x'_{2t}) \\ &= E(x_{2t}x'_{2t}) \equiv Q \end{aligned} \tag{28}$$

with the last line following from the assumption that x_{1t} and x_{2t} are uncorrelated. From (20) we also know

$$T^{1/2}(b_2 - \beta_2) = \left(T^{-1}\sum \tilde{x}_{2t}\tilde{x}'_{2t}\right)^{-1}\left(T^{-1/2}\sum \tilde{x}_{2t}u_{t+h}\right) \tag{29}$$

where

$$T^{-1/2}\sum \tilde{x}_{2t}u_{t+h} = T^{-1/2}\sum x_{2t}u_{t+h} - A_T T^{-1/2}\left(\sum x_{1t}u_{t+h}\right).$$

But if $E(x_{2t}x'_{1t}) = 0$, then $\text{plim}(A_T) = 0$, meaning

$$T^{-1/2}\sum \tilde{x}_{2t}u_{t+h} \xrightarrow{d} T^{-1/2}\sum x_{2t}u_{t+h}.$$

This will be recognized as $\sqrt{T}$ times the sample mean of a random vector with population mean zero, for which the Central Limit Theorem would take the form

$$T^{-1/2}\sum \tilde{x}_{2t}u_{t+h} \xrightarrow{d} r \sim N(0, S) \tag{30}$$

for S given in (12). Combining results (28), (29) and (30) gives (11).

To derive (13), let $b = (b'_1, b'_2)'$ denote the OLS coefficients when the regression includes both x_{1t} and x_{2t} and b_1^* the coefficients from an OLS regression that includes only x_{1t}. The

sum of squared residuals from the latter regression can be written

$$\begin{aligned} SSR_1 &= \sum (y_{t+h} - x_{1t}' b_1^*)^2 \\ &= \sum (y_{t+h} - x_t' b + x_t' b - x_{1t}' b_1^*)^2 \\ &= \sum (y_{t+h} - x_t' b)^2 + \sum (x_t' b - x_{1t}' b_1^*)^2 \end{aligned}$$

where all summations are over $t = 1, ..., T$ and the last equality follows from the orthogonality property of OLS. Thus the difference in SSR between the two regressions is

$$SSR_1 - SSR_2 = \sum (x_t' b - x_{1t}' b_1^*)^2. \tag{31}$$

It's also not hard to show that the fitted values for the full regression could be calculated as[51]

$$x_t' b = x_{1t}' b_1^* + \tilde{x}_{2t}' b_2. \tag{32}$$

Thus from (31) and (32),

$$SSR_1 - SSR_2 = \sum (\tilde{x}_{2t}' b_2)^2.$$

If the true value of β_2 is zero, then (20) becomes

$$b_2 = \left(\sum \tilde{x}_{2t} \tilde{x}_{2t}'\right)^{-1} \left(\sum \tilde{x}_{2t} u_{t+h}\right) \tag{33}$$

$$\begin{aligned} SSR_1 - SSR_2 &= b_2' \left(\sum \tilde{x}_{2t} \tilde{x}_{2t}'\right) b_2 \\ &= \left(T^{-1/2} \sum u_{t+h} \tilde{x}_{2t}'\right) \left(T^{-1} \sum \tilde{x}_{2t} \tilde{x}_{2t}'\right)^{-1} \left(T^{-1/2} \sum \tilde{x}_{2t} u_{t+h}\right). \end{aligned} \tag{34}$$

Results (28) and (30) then establish

$$SSR_1 - SSR_2 \xrightarrow{d} r' Q^{-1} r. \tag{35}$$

Recall that R^2 is defined as

$$R^2 = 1 - \frac{SSR}{\sum_{t=1}^{T} (y_{t+h} - \bar{y}_h)^2}$$

so the difference in R^2 is

$$R_2^2 - R_1^2 = \frac{(SSR_1 - SSR_2)}{\sum_{t=1}^{T} (y_{t+h} - \bar{y}_h)^2}.$$

[51] The easiest way to confirm the claim is to show that the residuals implied by (32) satisfy the orthogonality conditions required of the original full regression, namely, that they are orthogonal to x_{1t} and x_{2t}. That the residual $y_{t+h} - x_{1t}' b_1^* - \tilde{x}_{2t}' b_2$ is orthogonal to x_{1t} follows from the fact that $y_{t+h} - x_{1t}' b_1^*$ is orthogonal to x_{1t} by the definition of b_1^* while $\tilde{x}_{2t}$ is orthogonal to x_{1t} by the construction of $\tilde{x}_{2t}$. Likewise $y_{t+h} - \tilde{x}_{2t}' b_2$ is orthogonal to $\tilde{x}_{2t}$ by (19), and since x_{1t} is again orthogonal to $\tilde{x}_{2t}$ by the construction of $\tilde{x}_{2t}$, it follows that $y_{t+h} - x_{1t}' b_1^* - \tilde{x}_{2t} b_2$ is orthgonal to $\tilde{x}_{2t}$. Since $y_{t+h} - x_{1t}' b_1^* - \tilde{x}_{2t}' b_2$ is orthogonal to both x_{1t} and $\tilde{x}_{2t}$, it is also orthogonal to $x_{2t} = \tilde{x}_{2t} + A_T x_{1t}$.

Thus from (34),

$$T(R_2^2 - R_1^2) = \frac{(SSR_1 - SSR_2)}{\sum(y_{t+h} - \bar{y}_h)^2/T} \xrightarrow{d} \frac{r'Q^{-1}r}{\gamma}$$

as claimed in (13).

B Additional empirical results

B.1 Additional results for Joslin-Priebsch-Singleton

In Table B.1 we show additional results for the $\bar{R}^2$ in predictive regressions with three yield PCs and the macro variables GRO and INF proposed by Joslin et al. (2014). The dependent variables are the annual excess returns for bonds with maturity from two to ten years. That is, Table B.1 reports the same results for each individual bond which Table 2 reports in its top panel for the average excess return across bond maturities. To economize on space we only show the bootstrap results for the bias-corrected (BC) bootstrap.

The results in Table B.1 show that for all bond maturities, the increase in $\bar{R}^2$ when macro variables are added is often large although the spanning hypothesis is true in population. While for the two- to four-year bonds, the increase in $\bar{R}^2$ in the data is larger than the upper bound of the 95%-bootstrap interval, for the remaining bonds this statistic is within this interval, meaning that there is no statistical evidence against the spanning hypothesis.

B.2 Additional results for Ludvigson-Ng

LN also constructed a single return-forecasting factor using a similar approach as Cochrane and Piazzesi (2005). They regressed the excess bond returns, averaged across the two- through five-year maturities, on the macro factors plus a cubed term of $F1$ which they found to be important. The fitted values of this regression produced their return-forecasting factor, denoted by $H8$. Adding $H8$ to a predictive regression that includes the Cochrane-Piazessi factor CP substantially increases the $\bar{R}^2$, and leads to a highly significant coefficient on $H8$. LN emphasized this result and interpreted it as further evidence that macro variables have predictive power beyond the information in the yield curve.

Tables B.2 and B.3 replicate LN's results for these regressions on the macro- ($H8$) and yield-based (CP) return-forecasting factors.[52] Table B.2 shows coefficient estimates and statistical significance, while Table B.3 reports $\bar{R}^2$. In LN's data, both CP and $H8$ are strongly significant with HAC p-values below 0.1%. Adding $H8$ to the regression increases the $\bar{R}^2$ by 9-11 percentage points.

One advantage of our bootstrap approach is that we can calculate the small-sample properties under the null hypothesis of complicated transformations of the original data such as these. To this end, we simply add an additional step in the construction of our artificial data by calculating CP and $H8$ in each bootstrap data set as the fitted values from preliminary regressions in the exact same way that LN did in the actual data.

[52] These results correspond to those in column 9 in tables 4-7 in LN.

Table B.2 shows that the observed increases in $\bar{R}^2$ when adding $H8$ to the regression are generally within the 95% bootstrap confidence interval. That is, although LN find large increases in $\bar{R}^2$ using these same regression specifications, this is not convincing evidence against the spanning hypothesis, as such increases in goodness-of-fit are perfectly plausible under the null hypothesis.

According to the bootstrap p-values for the coefficients on $H8$ in Table B.3, the macro return-forecasting factor is no longer significant at the 1% level. Furthermore, the size distortions for conventional t-tests are very substantial: a test with nominal size of five percent based on asymptotic HAC p-values has a true size of 58-61 percent. This evidence suggests that conventional HAC inference can be particularly problematic when the predictors are return-forecasting factors. Table B.3 also shows that the bootstrap test has good size and power.

We also examined the same regressions over the 1985–2015 sample period with results shown in the right half of Table B.2 and in the bottom panel of Table B.3. The observed increases in $\bar{R}^2$ are squarely in line with what we would expect under the spanning hypothesis, as indicated by the confidence intervals in Table B.2. The return-forecasting factors would again appear to be highly significant based on HAC p-values, but the size distortions of these tests are again very substantial and the coefficients on $H8$ are in fact not statistically significant when using the bootstrap p-values.

This evidence suggests that conventional HAC inference can be particularly problematic when the predictors are return-forecasting factors. One reason for the substantially distorted inference is their high persistence; $H8$ and CP have autocorrelations that are around 0.8, and decline only slowly with the lag length. Another reason is that the return-forecasting factors are constructed in a preliminary estimation step, which introduces additional estimation uncertainty not accounted for by conventional inference. In such a setting other econometric methods—preferably a bootstrap exercise designed to assess the relevant null hypothesis—are needed to accurately carry out inference. For the case at hand, we conclude that a return-forecasting factor based on macro factors exhibits only very tenuous predictive power, much weaker than indicated by LN's original analysis and which disappears completely over a different sample period.

B.3 Bond supply: Greenwood-Vayanos

In addition to macro-finance linkages, a separate literature studies the effects of the supply of bonds on prices and yields. The theoretical literature on the so-called portfolio balance approach to interest rate determination includes classic contributions going back to Tobin (1969) and Modigliani and Sutch (1966), as well as more recent work by Vayanos and Vila (2009) and King (2013). A number of empirical studies document the relation between bond supply and interest rates during both normal times and over the recent period of near-zero interest and central bank asset purchases, including Hamilton and Wu (2012), D'Amico and King (2013), and Greenwood and Vayanos (2014). Both theoretical and empirical work has convincingly demonstrated that bond supply is related to bond yields and returns.

However, our question here is whether measures of Treasury bond supply contain information that is not already reflected in the yield curve and that is useful for predicting future bond

yields and returns. Is there evidence against the spanning hypothesis that involves measures of time variation in bond supply? At first glance, the answer seems to be yes. Greenwood and Vayanos (2014) (henceforth GV) found that their measure of bond supply, a maturity-weighted debt-to-GDP ratio, predicts yields and bond returns, and that this holds true even controlling for yield curve information such as the term spread. Here we investigate whether this result holds up to closer scrutiny. The sample period used in Greenwood and Vayanos (2014) is 1952 to 2008.[53]

To estimate the effects of bond supply on interest rates, GV estimate a broad variety of different regression specifications with yields and returns of various maturities as dependent variables. Here we are most interested in those regressions that control for the information in the yield curve. In the top panel of Table B.4 we reproduce their baseline specification in which the one-year return on a long-term bond is predicted using the one-year yield and bond supply measure alone. The second panel includes the spread between the long-term and one-year yield as an additional explanatory variable.[54] Like GV we use Newey-West standard errors with 36 lags.[55]

If we interpreted the HAC t-test using the conventional asymptotic critical values, the coefficient on bond supply is significant in the baseline regression in the top panel but is no longer significant at the conventional significance level of five percent when the yield spread is included in the regression, as seen in the second panel. But once again there are some warning flags that raise doubts about the validity of HAC inference. The bond supply variable is extremely persistent—the first-order autocorrelation is 0.998—and the one-year yield and yield spread are of course highly persistent as well. This leads us to suspect that the true p-value likely exceeds the purported 5.8%.

The bond return that GV used as the dependent variable in these regressions is for a hypothetical long-term bond with a 20-year maturity. We do not apply our bootstrap procedure here because this bond return is not constructed from the observed yield curve.[56] Instead we rely on IM tests to carry out robust inference. Neither of the IM tests finds the coefficient on bond supply to be statistically significant. In contrast, the coefficient on the term spread is strongly significant for the HAC test and both IM tests.

We consider two additional regression specifications that are relevant in this context. The first controls for information in the yield curve by including, instead of a single term spread, the first three PCs of observed yields.[57] It also subtracts the one-year yield from the bond return in order to yield an excess return. Both of these changes make this specification more closely comparable to those in the literature. The results are reported in the third panel of Table B.4. Again, the coefficient on bond supply is only marginally significant for the HAC t-test, and insignificant for the IM tests. In contrast, the coefficients on both PC1 and PC2 are strongly significant for the IM tests.

[53] As in JPS, the authors report a sample end date of 2007 but use yields up to 2008 to calculate one-year bond returns up to the end of 2007.

[54] These estimates are in GV's table 5, rows 1 and 6. Their baseline results are also in their table 2.

[55] There are small differences in our and their t-statistics that we cannot reconcile but which are unimportant for the results.

[56] GV obtained this series from Ibbotson Associates.

[57] These PCs are calculated from the observed Fama-Bliss yields with one- through five-year maturities.

Finally, we consider the most common specification where y_{t+h} is the one-year excess return, averaged across two- though five-year maturities. The last panel of Table B.4 shows that in this case, the coefficient on bond supply is insignificant according to the conventional Newey-West t-test. Using our bootstrap procedure we find that there is a significant size distortion for this hypothesis test, and the bootstrap p-value is substantially higher than the conventional p-value. The IM test suggests that PC1 and PC2 have significant predictive power for bond returns but fails to reject the spanning hypothesis.

Overall, we find that the results in GV do not constitute evidence against the spanning hypothesis. While bond supply exhibits a strong empirical link with interest rates, its predictive power for future yields and returns seems to be fully captured by the current yield curve.

B.4 Output gap: Cooper-Priestley

Another widely cited study that appears to provide evidence of predictive power of macro variables for asset prices is Cooper and Priestley (2008) (henceforth CPR). This paper focuses on one particular macro variable as a predictor of stock and bond returns, namely the output gap, which is a key indicator of the economic business cycle. The authors concluded that "the output gap can predict next year's excess returns on U.S. government bonds" (p. 2803). Furthermore, they also claimed that some of this predictive power is independent of the information in the yield curve, and implicitly rejected the spanning hypothesis (p. 2828).

We investigate the predictive regressions for excess bond returns y_{t+h} using the output gap at date $t-1$ (gap_{t-1}), measured as the deviation of the Fed's Industrial Production series from a quadratic time trend.[58] CPR lagged their measure by one month to account for the publication lag of the Fed's Industrial Production data. Table B.5 shows our results for predictions of the excess return on the five-year bond; the results for other maturities closely parallel these. The top two panels correspond to the regression specifications that CPR estimated.[59] In the first specification, the only predictor is gap_{t-1}. The second specification also includes $\tilde{CP}_t$, which is the Cochrane-Piazzesi factor CP_t after it is orthogonalized with respect to gap_t.[60] We obtain coefficients and $\bar{R}^2$ that are close to those published in CPR. We calculate both OLS and HAC t-statistics, where in the latter case we use Newey-West with 22 lags as described by CPR. Our OLS t-statistics are very close to the published numbers, and according to these the coefficient on gap_{t-1} is highly significant. However, the HAC t-statistics are only about a third of the OLS t-statistics, and indicate that the coefficient on gap is far from significant, with p-values above 20%.[61]

Importantly, neither of the specifications in CPR can be used to test the spanning hypothesis, because the CP factor is first orthogonalized with respect to the output gap. This defeats the purpose of controlling for yield-curve information, since any predictive power that is shared by the CP factor and gap will be exclusively attributed to the latter.[62] One way to

[58] We thank Richard Priestley for sending us this real-time measure of the output gap.

[59] The relevant results in CPR are in the top panel of their table 9.

[60] Note that the predictors $\tilde{CP}_t$ and gap_{t-1} are therefore not completely orthogonal.

[61] This indicates that CPR may have mistakenly reported the OLS instead of the Newey-West t-statistics

[62] In particular, finding a significant coefficient on gap in a regression with $\tilde{CP}$ cannot justify the conclusion that "gap is capturing risk that is independent of the financial market-based variable CP" (p. 2828).

test the spanning hypothesis is to include CP instead of $\tilde{CP}$, for which we report the results in the third panel of Table B.5. In this case, the coefficient on *gap* switches to a positive sign, and its Newey-West t-statistic remains insignificant. In contrast, both $\tilde{CP}$ and CP are strongly significant in these regressions.

Our preferred specification includes the first three PCs of the yield curve—see the last panel of Table B.5. The predictor *gap* is highly persistent, with a first-order autocorrelation coefficient of 0.975, so there are likely small-sample inference problems. Hence we also include results for robust inference using the bootstrap and IM tests. The *gap* variable has a positive coefficient with a HAC p-value of 19%, which rises to 36% when using our bootstrap procedure. The conventional HAC t-test is substantially oversized, as evident by the bootstrap critical value that substantially exceeds the conventional critical value. The IM tests do not reject the null. Overall, we do not find any evidence that the output gap predicts excess bond returns.

Table B.1: Joslin-Priebsch-Singleton: $\bar{R}^2$ for excess-return regressions

		Original sample: 1985–2008			Later sample: 1985–2015		
		$\bar{R}^2_1$	$\bar{R}^2_2$	$\bar{R}^2_2 - \bar{R}^2_1$	$\bar{R}^2_1$	$\bar{R}^2_2$	$\bar{R}^2_2 - \bar{R}^2_1$
Two-year	Data	0.14	0.48	0.34	0.13	0.26	0.13
bond	BC bootstrap	0.46	0.52	0.05	0.37	0.42	0.05
		(0.12, 0.78)	(0.17, 0.81)	(-0.00, 0.19)	(0.09, 0.65)	(0.14, 0.68)	(-0.00, 0.18)
Three-year	Data	0.12	0.41	0.29	0.10	0.22	0.12
bond	BC bootstrap	0.41	0.47	0.06	0.32	0.37	0.05
		(0.11, 0.72)	(0.17, 0.75)	(-0.00, 0.22)	(0.07, 0.60)	(0.11, 0.64)	(-0.00, 0.19)
Four-year	Data	0.15	0.40	0.26	0.13	0.21	0.08
bond	BC bootstrap	0.40	0.46	0.06	0.30	0.36	0.06
		(0.10, 0.69)	(0.16, 0.72)	(-0.00, 0.21)	(0.06, 0.57)	(0.11, 0.62)	(-0.00, 0.20)
Five-year	Data	0.16	0.38	0.22	0.14	0.21	0.07
bond	BC bootstrap	0.37	0.43	0.06	0.28	0.34	0.06
		(0.10, 0.64)	(0.15, 0.69)	(-0.00, 0.22)	(0.06, 0.55)	(0.10, 0.59)	(-0.00, 0.20)
Six-year	Data	0.18	0.39	0.20	0.16	0.22	0.06
bond	BC bootstrap	0.37	0.43	0.06	0.29	0.35	0.06
		(0.10, 0.65)	(0.16, 0.68)	(-0.00, 0.22)	(0.06, 0.55)	(0.11, 0.59)	(-0.00, 0.20)
Seven-year	Data	0.18	0.37	0.18	0.17	0.23	0.06
bond	BC bootstrap	0.34	0.40	0.06	0.27	0.33	0.06
		(0.09, 0.60)	(0.15, 0.65)	(-0.00, 0.22)	(0.06, 0.51)	(0.10, 0.56)	(-0.00, 0.21)
Eight-year	Data	0.21	0.38	0.17	0.19	0.24	0.05
bond	BC bootstrap	0.34	0.40	0.06	0.27	0.33	0.06
		(0.09, 0.58)	(0.15, 0.63)	(-0.00, 0.22)	(0.06, 0.50)	(0.11, 0.55)	(-0.00, 0.19)
Nine-year	Data	0.23	0.39	0.16	0.20	0.25	0.05
bond	BC bootstrap	0.34	0.40	0.06	0.28	0.33	0.05
		(0.09, 0.58)	(0.15, 0.63)	(-0.00, 0.22)	(0.07, 0.50)	(0.12, 0.55)	(-0.00, 0.20)
Ten-year	Data	0.20	0.36	0.16	0.20	0.26	0.06
bond	BC bootstrap	0.32	0.38	0.06	0.28	0.33	0.05
		(0.08, 0.56)	(0.13, 0.62)	(-0.00, 0.24)	(0.07, 0.51)	(0.11, 0.56)	(-0.00, 0.20)

Adjusted $\bar{R}^2$ for regressions of annual excess bond returns on three PCs of the yield curve ($\bar{R}^2_1$) and on three yield PCs together with the macro variables GRO and INF ($\bar{R}^2_2$), as well as the difference in adjusted $\bar{R}^2$. The macro data is described in the text. The results in the left half of the table are for the original sample period of Joslin et al. (2014); the data used in the right half is extended to December 2015. Each panel reports first the statistics in the data, and then the mean and the 95%-confidence intervals (in parentheses) of the bootstrap small-sample distribution. The bootstrap, which is explained in the text, imposes the null hypothesis that the macro variables have no predictive power.

Table B.2: Ludvigson-Ng: $\bar{R}^2$ for regressions with return-forecasting factors

	Original sample: 1964–2007			Preferred sample: 1985–2015		
	$\bar{R}^2_1$	$\bar{R}^2_2$	$\bar{R}^2_2 - \bar{R}^2_1$	$\bar{R}^2_1$	$\bar{R}^2_2$	$\bar{R}^2_2 - \bar{R}^2_1$
Two-year bond						
Data	0.31	0.42	0.11	0.16	0.23	0.07
Bootstrap	0.21	0.24	0.03	0.30	0.34	0.04
	(0.06, 0.39)	(0.09, 0.41)	(0.00, 0.11)	(0.08, 0.54)	(0.14, 0.56)	(-0.00, 0.14)
Three-year bond						
Data	0.33	0.43	0.10	0.15	0.22	0.07
Bootstrap	0.20	0.24	0.04	0.29	0.33	0.05
	(0.06, 0.38)	(0.09, 0.41)	(0.00, 0.11)	(0.08, 0.51)	(0.13, 0.54)	(-0.00, 0.15)
Four-year bond						
Data	0.36	0.45	0.09	0.19	0.26	0.06
Bootstrap	0.21	0.25	0.04	0.30	0.34	0.04
	(0.07, 0.39)	(0.10, 0.42)	(0.00, 0.11)	(0.09, 0.52)	(0.15, 0.54)	(-0.00, 0.15)
Five-year bond						
Data	0.33	0.42	0.09	0.18	0.23	0.06
Bootstrap	0.21	0.24	0.04	0.28	0.32	0.04
	(0.06, 0.39)	(0.10, 0.41)	(0.00, 0.11)	(0.09, 0.50)	(0.14, 0.53)	(-0.00, 0.15)

$\bar{R}^2$ for regressions of annual excess bond returns on yield and macro factors, as in Ludvigson and Ng (2010). $\bar{R}^2_1$ is for regressions with only the return-forecasting factor based on yield-curve information (CP), $\bar{R}^2_2$ is for regressions that also include the return-forecasting factor based on macro information ($H8$). The left side of the table shows results for the original data set used by Ludvigson and Ng (2010), and the right side shows results for a data sample that starts in 1985 and ends in 2015. We report the values of the statistics in the data, and the means and 95%-confidence intervals (in parentheses) for the bootstrap small-sample distributions, obtained under the null hypothesis that the macro variables have no predictive power. The bootstrap procedure is described in the main text.

Table B.3: Ludvigson-Ng: statistical inference in regressions with return-forecasting factors

	Two-year bond		Three-year bond		Four-year bond		Five-year bond	
	CP	$H8$	CP	$H8$	CP	$H8$	CP	$H8$
Original sample: 1964–2007								
Coefficient	0.335	0.331	0.645	0.588	0.955	0.776	1.115	0.937
HAC t-statistic	4.429	4.331	4.666	4.491	4.765	4.472	4.371	4.541
HAC p-value	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Bootstrap 5% c.v.		3.857		3.968		3.965		3.998
Bootstrap p-value		**0.019**		**0.021**		**0.023**		**0.019**
Size								
HAC		0.579		0.612		0.610		0.594
Bootstrap		0.049		0.059		0.054		0.049
Power								
Bootstrap		0.621		0.573		0.555		0.521
Later sample: 1985–2015								
Coefficient	0.343	0.334	0.645	0.650	1.066	0.900	1.280	1.073
HAC statistic	2.566	2.698	2.403	2.983	2.805	3.218	2.734	3.256
HAC p-value	**0.011**	**0.007**	**0.017**	**0.003**	**0.005**	**0.001**	**0.007**	**0.001**
Bootstrap 5% c.v.		4.226		4.282		4.337		4.212
Bootstrap p-value		0.315		0.248		0.180		0.172

Predictive regressions for annual excess bond returns, using return-forecasting factors based on yield-curve information (CP) and macro information ($H8$), as in Ludvigson and Ng (2010). The first panel shows the results for their original data and sample period; the second panel uses a data sample that starts in 1985 and ends in 2015. HAC t-statistics and p-values are calculated using Newey-West standard errors with 18 lags. We obtain bootstrap small-sample distributions of the t-statistics under the null hypothesis that macro factors and hence $H8$ have no predictive power, and report the bootstrap critical values (c.v.'s) and p-values, as well as estimates of the true size of conventional HAC t-tests and the bootstrap tests with 5% nominal coverage (see notes to Table 3). We also report estimates of the power of the bootstrap tests. The bootstrap procedures are described in the main text. p-values below 5% are emphasized with bold face.

Table B.4: Greenwood-Vayanos: predictive power of Treasury bond supply

	One-year yield	Term spread	$PC1$	$PC2$	$PC3$	Bond supply
Dependent variable: return on long-term bond						
Coefficient	1.212					0.026
HAC t-statistic	2.853					3.104
HAC p-value	**0.004**					**0.002**
IM $q = 8$	**0.030**					0.795
IM $q = 16$	**0.001**					0.925
Dependent variable: return on long-term bond						
Coefficient	1.800	2.872				0.014
HAC t-statistic	5.208	4.596				1.898
HAC p-value	**0.000**	**0.000**				0.058
IM $q = 8$	**0.006**	**0.013**				0.972
IM $q = 16$	**0.000**	**0.000**				0.557
Dependent variable: excess return on long-term bond						
Coefficient			0.168	5.842	-6.089	0.013
HAC t-statistic			1.457	4.853	1.303	1.862
HAC p-value			0.146	**0.000**	0.193	0.063
IM $q = 8$			**0.000**	**0.003**	**0.045**	0.968
IM $q = 16$			**0.000**	**0.000**	**0.023**	0.854
Dependent variable: avg. excess return for 2-5 year bonds						
Coefficient			0.085	1.669	-4.632	0.004
HAC statistic			1.270	3.156	2.067	1.154
HAC p-value			0.204	**0.002**	**0.039**	0.249
Bootstrap 5% c.v.						3.105
Bootstrap p-value						0.448
IM q = 8			**0.005**	0.134	0.714	0.494
IM q = 16			**0.008**	**0.011**	0.611	0.980

Predictive regressions for annual bond returns using Treasury bond supply, as in Greenwood and Vayanos (2014) (GV). The coefficients on bond supply in the first two panels are identical to those reported in rows (1) and (6) of Table 5 in GV. HAC t-statistics and p-values are constructed using Newey-West standard errors with 36 lags, as in GV. The last panel includes bootstrap critical values and p-values using small-sample distributions generated under the null hypothesis that bond supply does not contain additional predictive power—the bootstrap procedure is described in the text. The last two rows in each panel report p-values for t-tests using the methodology of Ibragimov and Müller (2010), splitting the sample into either 8 or 16 blocks. The sample period is 1952 to 2008. p-values below 5% are emphasized with bold face.

Table B.5: Cooper-Priestley: predictive power of the output gap

	gap	$\tilde{CP}$	CP	$PC1$	$PC2$	$PC3$
Coefficient	-0.126					
OLS t-statistic	3.224					
HAC t-statistic	1.077					
HAC p-value	0.282					
Coefficient	-0.120	1.588				
OLS t-statistic	3.479	13.541				
HAC t-statistic	1.244	4.925				
HAC p-value	0.214	**0.000**				
Coefficient	0.113		1.612			
OLS t-statistic	2.940		13.831			
HAC t-statistic	1.099		5.059			
HAC p-value	0.272		**0.000**			
Coefficient	0.147			0.001	0.043	-0.067
OLS t-statistic	3.524			4.359	11.506	3.690
HAC t-statistic	1.306			1.354	4.362	2.507
HAC p-value	0.192			0.176	**0.000**	0.012
Bootstrap 5% c.v.	2.933					
Bootstrap p-value	0.356					
IM $q = 8$	0.612			**0.002**	**0.011**	0.234
IM $q = 16$	0.243			**0.000**	**0.001**	0.064

Predictive regressions for the one-year excess return on a five-year bond using the output gap, as in Cooper and Priestley (2008) (CPR). $\tilde{CP}$ is the Cochrane-Piazzesi factor after orthogonalizing it with respect to *gap*, whereas CP is the usual Cochrane-Piazzesi factor. For the predictive regression, *gap* is lagged one month, as in CPR. HAC standard errors are based on the Newey-West estimator with 22 lags. The bootstrap procedure, which does not include bias correction, is described in the main text. The sample period is 1952 to 2003. p-values below 5% are emphasized with bold face.