Robust Bond Risk Premia[*]

Michael D. Bauer[†]and James D. Hamilton[‡]

April 16, 2015

Abstract

Recent studies appear to have found evidence that information not reflected in the yield curve helps predict interest rates and excess bond returns. These studies reject the Markov property of the yield curve and conclude that there is unspanned or hidden information that should be used in forecasting. We revisit the evidence of these papers using novel econometric techniques that address the difficult problems surrounding inference about predictability of highly persistent series. We reach the opposite conclusion: only the level and the slope of the yield curve are robust predictors of excess bond returns, and there is no robust and convincing evidence for unspanned macro risk. In other words, the Markov property of the yield curve seems alive and well.

Keywords: yield curve, spanning, bond returns, small-sample bias, robust inference
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 Ulrich Müller and Glenn Rudebusch for helpful comments, and Javier Quintero and Simon Riddell for excellent research assistance.

[†] Federal Reserve Bank of San Francisco, michael.bauer@sf.frb.org

[‡] University of California at San Diego, jhamilton@ucsd.edu

1 Introduction

The nominal yield on a 10-year U.S. Treasury bond has been below 2% much of the time since 2011, a level never seen previously. To what extent does this represent unprecedently low expected interest rates extending through the next decade, and to what extent does it reflect an unusually low risk premium resulting from a flight to safety and large-scale asset purchases by central banks that depressed the long-term yield? Finding the answer is a critical input for monetary policy, investment strategy, and understanding the lasting consequences of the financial and economic disruptions of 2008.

In principle one can measure the risk premium by the difference between the current long rate and the expected value of future short rates. But what information should go into constructing that expectation of future short rates? 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 necessarily a function of z_t, reflecting the general fact that current asset prices incorporate all current information. This suggests that we may be able to back out the state vector z_t from the observed yield curve.[1] The "invertibility" or "spanning" hypothesis states that the current yield curve contains all the information that is useful for predicting future interest rates or determining risk premia. Notably, under this hypothesis, the yield curve is first-order Markov.

It has long been recognized that three factors can provide an excellent summary of the information in yields (Litterman and Scheinkman, 1991), and that these factors correspond to the level, slope, and curvature of the yield curve. Hence it would seem that these three factors should be everything one needs to forecast future yields and to estimate bond risk premia. This hypothesis—that all the relevant information is spanned by level, slope and curvature of the yield curve—is an important benchmark case that should be carefully investigated. Importantly, if it holds then finding answers to questions about short-rate expectations and bond risk premia does not require any data or models involving macroeconomic series, other asset prices or quantities, volatilities, measures of monetary policy, or survey expectations. Instead, the only piece of information needed to answer these questions is the shape of the current yield curve. The bar for rejecting this hypothesis, which we will refer to as the spanning hypothesis, certainly should be high.

However, a number of recent studies have produced evidence that variables other than

[1] Specifically, this invertibility requires that (a) we observe at least as many yields as there are state variables in z_t, and (b) there are no knife-edge cancellations or pronounced nonlinearities; see for example Duffee (2013).

the first three principal components of current yields seem to be useful for predicting future interest rates. Joslin et al. (2014) found that measures of economic growth and inflation contain substantial predictive power for excess bond returns beyond the information in the yield curve. Ludvigson and Ng (2009, 2010) documented that factors inferred from a large set of macro variables help predict bond returns. Cooper and Priestley (2008) found that the output gap helps predict excess bond returns. Cochrane and Piazzesi (2005) reported evidence that information in the fourth and fifth principal component of yields has predictive power. Greenwood and Vayanos (2014) found that measures of Treasury bond supply appear to help forecast yields and returns. Each of these findings suggests that the Markov property of the yield curve should be rejected and that there is unspanned or hidden information that is not captured by the current yield curve but that is useful for forecasting.

The key evidence in all these studies comes from regressions of yields or excess returns on a vector x_t of predictive variables that are highly serially correlated and that include variables that are strongly correlated with lagged values of the dependent variable. Although these regressions have a fundamentally different structure from that considered by Mankiw and Shapiro (1986) and Stambaugh (1999), small-sample problems that are related to those identified by these researchers turn out to be potentially important for investigation of the spanning hypothesis. We demonstrate in this paper that the procedures researchers have been using to deal with problems raised by serial correlation of the regressors and regression residuals are subject to significant small-sample distortions. We show for example that the tests employed by Ludvigson and Ng (2009), which are intended to have a nominal size of 5%, can have a true size of up to 56%. We further demonstrate that the predictive relations found by all of these researchers exhibit much weaker performance over subsequent data than they had over the samples originally analyzed by the researchers.

We propose two procedures that researchers could use that would give substantially more robust small-sample inference. The first is a bootstrap procedure that is designed to test the null hypothesis of interest. We calculate the first three principal components of the observed set of yields and summarize their dynamics with a VAR fit to the observed principal components. We generate a time series for the yield for a bond of maturity n by multiplying the simulated principal components by the historical weighting vector for that yield on the principal components and adding a small Gaussian measurement error. Thus by construction no variables other than the principal components are useful for predicting yields in our generated data. We then fit a separate VAR to the proposed additional explanatory variables, and generate a realization of these that is completely independent of the generated yields. We can then calculate the properties of any statistic under the null hypothesis that the additional explanatory variables have no predictive power. We find using this bootstrap procedure that much of the evidence of predictability reported by earlier researchers in fact fails to pass the usual standards for statistical significance. Notably, while other studies have employed the bootstrap to carry out inference, they have almost invariably done so for testing the expectations hypothesis, which is much stronger and much less plausible than the spanning hypothesis. In contrast, our study is the first to use a bootstrap design that is tailored to test the relevant null hypothesis, namely that nothing else but three yield-curve factors contains information relevant for predicting yields and returns.

A second procedure that we propose for inference in this context is the approach for robust testing recently suggested by Ibragimov and Müller (2010). We have found this approach to have excellent size and power properties in settings similar to the ones encountered by researchers testing for predictive power for interest rates and bond returns. The suggestion of Ibragimov and Müller (2010) is to split the sample into subsamples, estimate coefficients separately in each of these, and to perform a simple t-test on the coefficients across subsamples. Applying this type of test to the predictive regressions for yields and bond returns studied in the literature, we find that the only robust predictors are the level and the slope of the yield curve, while the evidence on all other predictors lacks robustness.

We carefully revisit the evidence in five very influential papers cited above, all of which appear to provide evidence against the null hypothesis of invertibility/spanning. We draw two conclusions from our investigation. First, the claims going back to Fama and Bliss (1987) and Campbell and Shiller (1991) that excess returns can be predicted from the level and slope of the yield curve remain quite robust. We emphasize that this conclusion is fully consistent with the Markov property of the yield curve. Second, the newer evidence on the predictive power of macro variables, higher-order principal components of the yield curve, or other variables, is subject to more serious econometric problems and overall appears weaker and much less robust. Overall, we do not find convincing evidence to reject the baseline hypothesis that the current yield curve, and in particular three factors summarizing this yield curve, contains all the information necessary to infer interest rate forecasts and bond risk premia. In other words, the Markov property of the yield curve seems alive and well.

2 Inference about the spanning hypothesis

The evidence against the spanning hypothesis in all of the studies referred to in the introduction 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 a yield, a yield curve factor (such as the level of the yield curve), or a bond return that we wish to predict, x_{1t} and x_{2t} are vectors containing K_1 and K_2 predictors, respectively, and u_{t+h} is an orthogonal forecast/projection error. The predictors x_{1t} contain a constant and the information in the yield curve, typically captured by the first three principal components (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.

The evidence produced in these studies comes in two forms, the first based on simple descriptive statistics such as how much the R^2 of the regression increases when the variables x_{2t} are added and the second from formal statistical tests of the hypothesis that $\beta_2 = 0$. In this section we show how two features of the specification—serial correlation in the error term u_t and the fact that x_{1t} is not strictly exogenous, for example because it includes lagged dependent variables—can matter significantly for both forms of evidence.

2.1 Consequences of serially correlated errors

Our first observation is that in regressions in which x_{1t} and x_{2t} are strongly serially correlated and the dependent variable is an excess holding yield for $h > 1$, we should not be surprised to see substantial increases in R^2 when x_{2t} is added to the regression even if the true coefficient is zero. It is well known that in small samples serial correlation in the residuals can increase both the bias as well as the variance of a regression R^2 (see for example Koerts and Abrahamse (1969) and Carrodus and Giles (1992)). To see how much difference this could make in the current setting, consider the unadjusted R^2 defined as

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

where SSR denotes the regression sum of squared residuals. The increase in R^2 when x_{2t} is added to the regression is thus given by

$$R_2^2 - R_1^2 = \frac{(SSR_1 - SSR_2)}{\sum_{t=1}^{T}(y_{t+h} - \bar{y}_h)^2}. \tag{3}$$

We show in Appendix A that when x_{1t}, x_{2t}, and u_{t+h} are stationary and satisfy standard regularity conditions, if the null hypothesis is true ($\beta_2 = 0$) and the extraneous regressors are uncorrelated with the valid predictors ($E(x_{2t}x_{1t}') = 0$), then

$$T(R_2^2 - R_1^2) \xrightarrow{d} r'Q^{-1}r/\gamma \tag{4}$$

$$\gamma = E[y_t - E(y_t)]^2$$

$$r \sim N(0, S), \tag{5}$$

$$Q = E(x_{2t}x_{2t}') \tag{6}$$

$$S = \textstyle\sum_{v=-\infty}^{\infty} E(u_{t+h}u_{t+h-v}x_{2t}x_{2,t-v}'). \tag{7}$$

Result (4) implies that the difference $R_2^2 - R_1^2$ itself converges in probability to zero under the null hypothesis that x_{2t} does not belong in the regression, meaning that the two regressions asymptotically should have the same R^2.

In a given finite sample, however, R_2^2 is larger than R_1^2 by construction, and the above results give us an indication of how much larger it would be in a given finite sample. If $x_{2t}u_{t+h}$ is serially uncorrelated, then (7) simplifies to $S_0 = E(u_{t+h}^2 x_{2t}x_{2t}')$. On the other hand, 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 $R_2^2 - R_1^2$, 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.

When the dependent variable y_{t+h} is something like a one-year holding return, $E(u_t u_{t-v}) \neq 0$ for $v = 0, \ldots, 11$, due to the overlapping observations. The explanatory variables x_{2t} often are highly serially correlated, so $E(x_{2t}x_{2,t-v}') \neq 0$. Thus even if x_{2t} is completely independent of u_t at all leads and lags, the product will be highly serially correlated,

$$E(u_{t+h}u_{t+h-v}x_{2t}x_{2,t-v}') = E(u_t u_{t-v})E(x_{2t}x_{2,t-v}') \neq 0.$$

This serial correlation in $x_{2t}u_{t+h}$ would contribute to larger values for $R_2^2 - R_1^2$ on average as well as to increased variability in $R_2^2 - R_1^2$ across samples. In other words, including x_{2t} could substantially increase the R^2 even if H_0 is true.

These results on the asymptotic distribution of $R_2^2 - R_1^2$ could be used to design a test of H_0. However, we show in the next section that in small samples the bias and variability of $R_2^2 - R_1^2$ can be even greater than predicted by (4). For this reason, in this paper we will rely on the exact small-sample distribution of the statistic $R_2^2 - R_1^2$, and demonstrate that the dramatic values sometimes reported in the literature are not implausible under the spanning hypothesis.[2]

Serial correlation of the residuals also affects the sampling distribution of the OLS estimate of β_2. In Appendix A we verify using standard algebra that under the null hypothesis $\beta_2 = 0$ the OLS estimate b_2 can be written as

$$b_2 = \left(\sum_{t=1}^{T} \tilde{x}_{2t}\tilde{x}_{2t}'\right)^{-1} \left(\sum_{t=1}^{T} \tilde{x}_{2t}u_{t+h}\right) \tag{8}$$

where $\tilde{x}_{2t}$ denotes the sample residuals from OLS regressions of x_{2t} on x_{1t}:

$$\tilde{x}_{2t} = x_{2t} - A_T x_{1t} \tag{9}$$

$$A_T = \left(\sum_{t=1}^{T} x_{2t}x_{1t}'\right) \left(\sum_{t=1}^{T} x_{1t}x_{1t}'\right)^{-1}. \tag{10}$$

If x_{2t} and x_{1t} are stationary and uncorrelated with each other, as the sample size grows, $A_T \xrightarrow{p} 0$ and b_2 has the same asymptotic distribution as

$$b_2^* = \left(\sum_{t=1}^{T} x_{2t}x_{2t}'\right)^{-1} \left(\sum_{t=1}^{T} x_{2t}u_{t+h}\right), \tag{11}$$

namely

$$\sqrt{T} b_2 \xrightarrow{d} N(0, Q^{-1}SQ^{-1}). \tag{12}$$

with Q and S the matrices defined in (6) and (7). Again we see that positive serial correlation causes S to exceed the value S_0 that would be appropriate for serially uncorrelated residuals.

[2] The same conclusions necessarily also hold for the adjusted $\bar{R}^2$ defined as

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

for k_i the number of coefficients estimated in model i, from which we see that

$$T(\bar{R}_2^2 - \bar{R}_1^2) = \frac{[T/(T-k_1)]SSR_1 - [T/(T-k_2)]SSR_2}{\sum_{t=1}^{T}(y_{t+h} - \bar{y}_h)^2/(T-1)}$$

which has the same asymptotic distribution as (4). In our small-sample investigations below, we will analyze either R^2 or $\bar{R}^2$ as was used in the original study that we revisit.

In other words, serial correlation in the error term increases the sampling variability of the OLS estimate b_2.

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, even if a consistent estimator of S is available, it may perform poorly in small samples. In practice, different HAC estimators can lead to substantially different empirical conclusions (Müller, 2014), and the small-sample variability can be significantly greater than predicted by (12). We will demonstrate empirically in the subsequent sections that this is a serious problem when carrying out inference about bond return predictability.

2.2 Consequences of weak exogeneity

A second feature of the studies examined in this paper is that the valid explanatory variables x_{1t} are correlated with lagged values of the error term. That is, these regressors are only weakly exogenous. This turns out to matter a great deal when x_{1t} and x_{2t} are highly serially correlated. Ours is a different setting from that considered by Mankiw and Shapiro (1986), Stambaugh (1999) and Campbell and Yogo (2006), who studied tests of the hypothesis $\beta_1 = 0$ in a setting of the form

$$y_{t+1} = \beta_1' x_{1t} + u_{t+1} \tag{13}$$

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

with x_{1t} a scalar and $E(u_t \varepsilon_{1t}) \neq 0$. Because the regressors x_{1t} are not strictly exogenous, Stambaugh (1999) showed that the OLS estimate of β_1 in (13) will be biased in small samples and this can significantly affect the small-sample inference when x_{1t} is highly serially correlated (ρ_1 large). By contrast, in our study the question is whether the vector $\beta_2 = 0$ in (1) is zero. The problem we identify can arise even when x_{2t} is strictly exogenous, that is, uncorrelated with u_t at all leads and lags. However, as in the case of Stambaugh bias, the small-sample problem in our setting arises from the fact that the other regressors x_{1t} are not strictly exogenous, and the problem is most dramatic when x_{1t} and x_{2t} are both highly serially correlated. We now explain why this is the case.

Consider the following simple example. Suppose that x_{1t} and x_{2t} are scalars that follow independent highly persistent processes,

$$x_{it} = \rho_i x_{i,t-1} + \sigma_i v_{it} \tag{14}$$

where ρ_i is close to one, v_{it} is a martingale difference sequence with unit variance and finite fourth moments, and v_{1t} is independent of v_{2t} at all leads and lags. We consider the consequences of OLS estimation of (1) in the special case where $h = 1$, x_{1t} and x_{2t} scalars, and the first regressor is the lag of the dependent variable: $x_{1t} = y_t$:

$$y_{t+1} = \beta_1 x_{1t} + \beta_2 x_{2t} + u_{t+1} \tag{15}$$

Because x_{1t} is a lagged dependent variable, it is correlated with lags of the error term, and strict exogeneity is violated. This is a simple example to illustrate the problems that can arise when x_{1t} includes variables that are correlated with lags of the dependent variable. Note for this example $x_{2t}u_{t+1} = \sigma_1 x_{2t} v_{1,t+1}$ is serially uncorrelated and the standard OLS t-test of $\beta_2 = 0$ asymptotically has a $N(0, 1)$ distribution.

One device for seeing how the results in a finite sample of some particular size T might differ from those predicted by the asymptotic distribution is to use a local-to-unity specification as in Phillips (1988):

$$x_{it} = (1 + c_i/T)x_{i,t-1} + \sigma_i v_{it} \qquad i = 1, 2. \tag{16}$$

For example, if our data come from a sample of size $T = 100$ when $\rho_i = 0.95$, the idea is to represent this with a value of $c_i = -5$ in (16). The claim is that analyzing the properties as $T \to \infty$ of a model characterized by (16) with $c_i = -5$ gives a better approximation to the actual distribution of a regression in a sample of size $T = 100$ and $\rho_i = 0.95$ than is provided by other methods; see for example Chan (1988) and Nabeya and Sørensen (1994).

The local-to-unity asymptotics turn out to be described by Ornstein-Uhlenbeck processes

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

where $W_1(\lambda)$ and $W_2(\lambda)$ denote independent standard Brownian motion. When $c_i = 0$, (16) becomes a random walk and the local-to-unity asymptotics simplify to the standard unit-root asymptotics involving integrals of Brownian motion as a special case: $J_0(\lambda) = W(\lambda)$.

When the null hypothesis is true, the t-test of $H_0 : \beta_2 = 0$ can be written as

$$\frac{b_2}{\hat{\sigma}_{b_2}} = \frac{\sum \tilde{x}_{2t} u_{t+1}}{\{s^2 \sum \tilde{x}_{2t}^2\}^{1/2}} \tag{17}$$

for $\tilde{x}_{2t}$ given by (9) and $s^2 = (T-2)^{-1} \sum (y_{t+1} - b_1 x_{1t} - b_2 x_{2t})^2$. We show in Appendix B that if x_{1t}, x_{2t}, and y_t are characterized by (15)-(16) with $x_{1t} = y_t$, as $T \to \infty$ the t-statistic has an asymptotic distribution given by

$$\frac{b_2}{\hat{\sigma}_{b_2}} \xrightarrow{d} \frac{\int J_{c_2}(\lambda)dW_1(\lambda) - \left[\int J_{c_2}(\lambda)J_{c_1}(\lambda)d\lambda\right]\left[\int [J_{c_1}(\lambda)]^2 d\lambda\right]^{-1}\left[\int J_{c_1}(\lambda)dW_1(\lambda)\right]}{\left\{\int [J_{c_2}(\lambda)]^2 d\lambda - \left[\int J_{c_2}(\lambda)J_{c_1}(\lambda)d\lambda\right]^2\left[\int [J_{c_1}(\lambda)]^2 d\lambda\right]^{-1}\right\}^{1/2}} \tag{18}$$

where $\int$ denotes integration over λ from 0 to 1. This distribution has significantly fatter tails than a $N(0,1)$, which explains why in small samples the OLS t-test is more likely to reject a true null hypothesis than is predicted by the standard asymptotic theory. The feature that makes this distribution nonstandard is the fact that because of lagged dependent variables, the processes $W_1(\lambda)$ and $J_{c_1}(\lambda)$ in the numerator are the same as appearing in the denominator.

One can get further insight into what accounts for these results by considering an alternative regression which according to standard asymptotics should have the same asymptotic distribution as the OLS t test but which under local-to-unity asymptotics is seen to behave quite differently in small samples. Under the null hypothesis $\beta_2 = 0$, the estimate b_2 from OLS estimation of (15) could be calculated by regressing the true residual u_{t+1} on $\tilde{x}_{2t}$ where $\tilde{x}_{2t}$ denotes the residual from a regression of x_{2t} on x_{1t}. We noted above that this would be predicted under the usual asymptotics to have the same asymptotic distribution as b_2^* in equation (11), which is obtained from regressing u_{t+1} directly on x_{2t}. It's interesting to consider the t-statistic associated with the latter regression,

$$\frac{b_2^*}{\hat{\sigma}_{b_2^*}} = \frac{\sum x_{2t}u_{t+1}}{\{s^{*2}\sum x_{2t}^2\}^{1/2}}$$

for $s^{*2} = (T-1)^{-1}\sum(u_{t+1} - b_2^* x_{2t})^2$. We show in Appendix B that under the local-to-unity assumptions in (16),

$$\frac{b_2^*}{\hat{\sigma}_{b_2^*}} \xrightarrow{d} \frac{\int J_{c_2}(\lambda)dW_1(\lambda)}{\left\{\int [J_{c_2}(\lambda)]^2 d\lambda\right\}^{1/2}}. \tag{19}$$

Although expression (19) might appear nonstandard, in fact it turns out simply to be a $N(0,1)$ distribution.[3] In other words, if we were actually able to perform a regression of u_{t+1} on x_{2t}, then even in small samples we would not have any problems. The problem is that, while the usual asymptotics that were used in Section 2.1 would predict that b_2^* should have the same distribution as b_2, in small samples the distributions can be quite different.[4]

[3] One can see this by noting as in Hamilton (1994, pp. 602-607) that conditional on $W_2(.)$ the expression in (19) has $N(0,1)$ distribution for all realizations of $W_2(.)$, and therefore has an unconditional $N(0,1)$ distribution when integrated with respect to the distribution of $W_2(.)$.

[4] Yet another calculation that helps shed light is if the regression is run as specified in (15) but the regressors

We can understand this difference by directly comparing the distributions in (18) and (19). Note that the denominator of (18) is strictly less than that in (19) for all realizations, a factor that would tend to make (18) bigger than (19). In other words, treating the t-statistic as $N(0,1)$ will result in rejecting a true null hypothesis too often. Based on our simulations, the difference in the denominators of (18) and (19) appears to be the main source of the small-sample problem.[5]

To summarize, the OLS estimate of β_2 in (1) can be obtained in three steps: (1) regress x_{2t} on x_{1t}, (2) regress y_{t+h} on x_{1t}, and (3) regress the residuals from (2) on the residuals of (1). The small-sample properties of the first regression are very different when x_{1t} and x_{2t} are highly persistent, and when x_{1t} is correlated with lags of y, this can end up mattering a great deal for the small-sample distribution of the final result.

We examined these implications in detail by generating values for x_{1t} and x_{2t} from (14) with v_{1t} and v_{2t} independent standard Normal variables. Under our data-generating process (DGP), the true values for the magnitudes in (15) are given by $\beta_1 = \rho_1$, $\beta_2 = 0$, and $u_{t+1} = x_{1,t+1} - \rho_1 x_{1t} = \sigma_1 v_{1t}$. We generated 5,000 artificial data samples.[6] We set $\sigma_1 = \sigma_2 = 1$, since all test statistics are invariant to these parameters. We vary ρ_1, ρ_2, and T, in order to study the effects of persistence and sample size on inference. In each simulated sample, we run a regression of y_{t+1} on x_{1t} and x_{2t}, including an intercept. Our interest is in the inference about β_2, and with this simulation design we can study the small-sample behavior of the relevant sample statistics.[7]

Table 1 reports the performance of a standard t-test with a nominal size of 5 percent using OLS standard errors. It shows the frequency of rejections of the null hypothesis, and numbers higher than 5 percent indicate a small-sample size distortion. For values of ρ_1 and ρ_2 at or

x_{1t} and x_{2t} are both strictly exogenous, that is, $u_{t+1} = \sigma_3 v_{3,t+1}$ with v_{3t} completely independent of v_{1t} and v_{2t}. In this case (18) would be replaced by

$$\frac{\int J_{c_2}(\lambda)dW_3(\lambda) - \left[\int J_{c_2}(\lambda)J_{c_1}(\lambda)d\lambda\right]\left[\int [J_{c_1}(\lambda)]^2 d\lambda\right]^{-1}\left[\int J_{c_1}(\lambda)dW_3(\lambda)\right]}{\left\{\int [J_{c_2}(\lambda)]^2 d\lambda - \left[\int J_{c_2}(\lambda)J_{c_1}(\lambda)d\lambda\right]^2 \left[\int [J_{c_1}(\lambda)]^2 d\lambda\right]^{-1}\right\}^{1/2}}$$

which, conditioning on $W_1(.)$ and $W_2(.)$ will be recognized as the $N(0,1)$ distribution (using the same reasoning as in footnote 3). The small-sample problems thus arise from the interaction between high serial correlation of x_1 and x_2 and the fact that x_1 includes lagged dependent variables.

[5] The numerators also differ, but our simulations reveal that the difference in denominators is the most important factor.

[6] We start the simulations at $x_{1,0} = x_{2,0} = 0$, following standard practice of making all inference conditional on date 0 magnitudes.

[7] Our simulation setup differs from that in Mankiw and Shapiro (1986) in two respects. First, in our case the innovations to x_{1t} and the error term in the regression u_t are perfectly correlated, since they are identical. Second, we include an additional (spurious) persistent independent variable in our regression.

above 0.95, the size distortions are substantial, and the t-test rejects the null two or three times as often as it should. The size distortions are generally smaller for longer samples, except for values of ρ_1 and ρ_2 near unity. To see more clearly the dependence of the size distortions on the sample size, Figure 1 plots the empirical size of the t-test for β_2 for different sample sizes from $T = 50$ to $T = 1000$. For the case $\rho_1 = \rho_2 = 0.99$ the empirical size declines from about 15 percent to about 9 percent. In contrast, when $\rho_1 = \rho_2 = 1$ the size distortions remain substantial even for larger sample sizes, as indeed in this case the non-Normal distribution corresponding to (18) with $c_i = 0$ governs the distribution for arbitrarily large T.

Why do conventional t-tests go so wrong in this setting, despite the fact that OLS coefficient estimates are consistent and the conventional OLS standard errors are asymptotically valid? To investigate the source of the size distortion we study the coefficient bias and the accuracy of the OLS standard errors in our simulations. Table 2 shows statistics for four different simulation settings which reveal the source of the problem. The top rows in each panel show the mean of the coefficient estimates and the corresponding bias, which indicate that estimates of β_1 are strongly downward biased, due to the lack of strict exogeneity, with the size of the bias decreasing in the sample size. In contrast, estimates of β_2 are unbiased—clearly the problem with hypothesis tests of $\beta_2 = 0$ do not arise from Stambaugh bias as traditionally understood. The reason for the size distortions is not coefficient bias, but the fact that the standard errors substantially underestimate the sampling variability. This is evident from comparing the standard deviation of the coefficient estimates across simulations—which is an estimate of the true small-sample standard error—and the average OLS standard errors. The difference, which we term “standard error bias,” is substantial: in our simulation study the standard errors are about 30% too low. The last row shows the size of a test that uses the “true” standard errors. The fact that this is close to the nominal size of 0.05 demonstrates that standard error bias accounts for the size distortions of the test for β_2. In the case of β_1 we illustrate the role of bias by considering t-tests of the hypothesis that β_1 is equal to its true value. Here, using the correct standard error does not eliminate the size distortion, because it is caused by a combination of coefficient bias and standard error bias.

Our simulations also establish that lack of strict exogeneity aggravates the potential problems with R^2 that Section 2.1 demonstrated could arise from serial correlation alone even if there were no lagged dependent variables. To investigate this issue, we compare simulation results for our original DGP to those for a second DGP, in which $y_{t+1} = \rho_1 x_{1t} + u_{t+1}$ and u_t is normally distributed with unit variance and independent of v_{1t} and v_{2t} at all leads and lags. That is, for this DGP both x_{1t} and x_{2t} are strictly exogenous, so that there is no bias and t-tests have exactly the correct size in any sample size. Since we want to investigate the distribution of R^2 we now take $y_{t+1} - x_{1t}$ to be the dependent variable in the regressions, so that $\beta_1 = \rho_1 - 1$ and the population R^2 is $(1-\rho_1)/2$—this is akin to predicting yield changes or bond returns.[8] Table 3 shows summary statistics for the distributions of R_1^2, R_2^2, and $R_2^2 - R_1^2$, for the case with $T = 50$ and $\rho_1 = \rho_2 = 0.99$. Under strict exogeneity, $R_2^2 - R_1^2$ is on average 2.1%, with a standard deviation of 2.9%. Lack of strict exogeneity raises both the mean (to 3.7%) and the variability (to 4.4%) of the distribution of $R_2^2 - R_1^2$. In Figure 2 we show the dependence on the sample size by plotting the average R_1^2 and R_2^2 for both DGPs from $T = 50$ to $T = 1000$, as well as the true population R^2 (0.5%). The results illustrate that lack of strict exogeneity raises the R^2 as well as $R_2^2 - R_1^2$ for all sample sizes, but particularly so for small samples.

To summarize, the lack of strict exogeneity of a subset of the regressors can have significant consequences for the small-sample inference, even if interest lies in the predictors that themselves are strictly exogenous. In all of the empirical studies that we consider below, the predictors in x_{1t} are correlated with past error terms. The reason is that they correspond to information in current yields, and the dependent variable is either a future bond return or the future level of the yield curve. Hence, lack of strict exogeneity is a serious concern in all tests of the spanning hypothesis. This is a separate issue from serial correlation in the residuals, and one that to the best of our knowledge has not been recognized in the predictability literature. Both issues become particularly serious when the predictors are persistent and when the sample sizes are small. Unfortunately, this is exactly the type of situation that researchers are faced when carrying out inference about predictability of interest rates, since the relevant time series are highly persistent and the sample periods typically studied are relatively short.[9] In Table 4 we report the estimated autocorrelation coefficients for the predictors used in the published studies that we investigate in the following sections. Clearly, many of the predictors considered in the literature are highly persistent, and we need to be particularly concerned about the aforementioned small-sample issues. Adding extraneous regressors that in reality contribute nothing to prediction may lead to an artificially large increase in the R^2 and inflated values for t- and F-statistics. In the following sections we quantify just how important this may be for a number of influential studies.

[8] In a regression of y_{t+1} on x_{1t} the population R^2 is ρ_1^2 which is too close to one to be useful or interesting.

[9] Reliable interest rate data is only available since about the 1960s, which leads to situations with about 40-50 years of monthly data. Going to higher frequencies—such as weekly or daily—does not increase the effective sample sizes, since it typically increases the persistence of the series and at introduces additional noise.

2.3 A bootstrap design for investigating the spanning hypothesis

The above analysis suggests that it is of paramount importance to base inference on the correct small-sample distributions of the relevant test statistics. While some studies use the bootstrap for this purpose, they typically do so by generating samples under the absence of predictability (Cochrane and Piazzesi, 2005; Ludvigson and Ng, 2009; Greenwood and Vayanos, 2014). By contrast, in our paper we propose a bootstrap to specifically test the spanning hypothesis $H_0 : \beta_2 = 0$.

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

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

along with the weighting vector $\hat{h}_n$ for bond n:

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

That is, $x_{1t} = \hat{H} y_t$, where $y_t = (y_{n_1 t}, \ldots, y_{n_J t})'$ is a J-vector with observed yields at t, and $\hat{H} = (\hat{h}_{n_1}, \ldots, \hat{h}_{n_J})'$ is the $3 \times J$ matrix with rows equal to the first three eigenvectors of the covariance matrix of x_t. We use normalized eigenvectors so that the matrix $\hat{H}$ is orthonormal. Fitted yields can be obtained using $\hat{y}_t = \hat{H}' x_{1t}$. 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 points.[10]

Then we estimate by OLS a VAR(12) for x_{1t}:

$$x_{1t} = \hat{\mu} + \hat{\phi}_1 x_{1,t-1} + \hat{\phi}_2 x_{1,t-2} + \cdots + \hat{\phi}_{12} x_{1,t-12} + e_{1t} \qquad t = 1, ..., T.$$

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.

Next we generate 1000 artificial yield data samples from this model, each with length T equal to the original sample length. We first iterate[11] on

$$x^*_{1\tau} = \hat{\mu} + \hat{\phi}_1 x^*_{1,\tau-1} + \hat{\phi}_2 x^*_{1,\tau-2} + \cdots + \hat{\phi}_{12} x^*_{1,\tau-12} + \varepsilon^*_{1\tau}$$

[10] For example, in the case study of Joslin et al. (2014) in Section 3, the standard deviation is 6.5 basis points.

[11] We start the recursion from $\tilde{x}_{1,1} = \ldots = \tilde{x}_{1,12} = 0$ and drop the first 500 realizations, so that we effectively start with a draw from the unconditional distribution of x_{1t}.

with $\varepsilon^*_{1\tau}$ *i.i.d.* draws from the empirical distribution of e_{1t}.[12]. Then we obtain the artificial yields using

$$y^*_{n\tau} = \hat{h}'_n x^*_{1\tau} + v^*_{n\tau}$$

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

We likewise fit a VAR(p) to the observed data[13] for the proposed predictors x_{2t},

$$x_{2t} = \hat{\alpha}_0 + \hat{\alpha}_1 x_{2,t-1} + \hat{\alpha}_2 x_{2,t-2} + \cdots + \hat{\alpha}_p x_{2,t-p} + e_{2t},$$

from which we then generated 1000 artificial samples $x^*_{2\tau}$. We can then investigate the properties of any proposed test statistic involving $y^*_{n\tau}$, $x^*_{1\tau}$, and $x^*_{2\tau}$ in a sample for which the dynamic serial correlation of yields and explanatory variables are similar to those in the actual data but in which $x^*_{2\tau}$ is independent by construction at all leads and lags from $y^*_{n\tau}$. In other words, in our bootstrap samples, the null hypothesis is true that macroeconomic variables have no predictive power for yields, i.e., there are no unspanned macro risks.

There is one important way in which this bootstrap likely understates the true magnitude of the problem, because the eigenvalues of the estimated values for $\left|\sum_{j=1}^{12} \hat{\phi}_j\right|$ and $\left|\sum_{j=1}^{p} \hat{\alpha}_j\right|$ are likely to be smaller than the corresponding population magnitudes—least squares estimates typically underestimate the autocorrelation of highly persistent processes (Kendall, 1954; Pope, 1990). Therefore, we also consider a variant of the bootstrap design described above, in which the generated samples use not the OLS estimates $\hat{\phi}_j$ and $\hat{\alpha}_j$ but instead use bias-corrected VAR estimates, using a simple bootstrap adopted by Kilian (1998).

2.4 An alternative robust test for predictability

There is of course a very large literature addressing the problem of HAC inference. This literature is concerned with accurately estimating the matrix S in (7) but does not address what we have identified as the key issue, which is the small-sample difference between the statistics in (8) and (11). We have looked at a number of alternative approaches in terms of how well they perform in our bootstrap experiments. We found that the most reliable existing

[12] We also experimented with a Monte Carlo design in which $\varepsilon^*_{1\tau}$ was drawn from a Student-t dynamic conditional correlation GARCH model (Engle, 2002) fit to the residuals e_{1t} with similar results to those obtained using resampled residuals as described in the text.

[13] We choose the lag length p according to the Schwarz-Bayes Information Criterion (SBIC). For example, in the case study on Joslin et al. (2014) in Section 3, the SBIC prescribes four lags.

test appears to be the one suggested by Ibragimov and Müller (2010), who proposed a novel 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, then a standard t-test with q degrees of freedom can be carried out to test hypotheses about the 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 (not reported) show that this test also performs very well in the specific settings that we consider in this paper, namely inference about predictive power of certain variables for future interest rates and excess bond returns. Throughout this paper, we report two sets of results for the Ibragimov-Müller (IM) test, setting the number of subsamples q equal to either 8 and 16 (as in Müller, 2014). A notable feature of the IM test is that it allows us to carry out inference that is robust not only against serial correlation in regressors and error terms, but also robust with respect to parameter instability across subsamples, as we will discuss below.

3 Predicting yields using economic growth and inflation

In this section we examine some of the evidence reported by Joslin et al. (2014) (henceforth JPS) that macro variables may help predict bond yields.

3.1 Excess bond returns

We begin with some of the most dramatic findings reported by JPS, which come from predictive regressions as in equation (1) where y_{t+h} is an excess bond return for a one-year holding period, x_{1t} is a vector consisting of a constant and the first three principal components of yields, and x_{2t} a vector consisting measures of economic growth and inflation. JPS found that for y_{t+h} the excess return on a ten-year bond over the risk-free one-year yield, the adjusted $\bar{R}^2$ of regression (1) when x_{2t} is excluded is only 0.20 when the regression was estimated over the period 1985:1-2007:12. But when they added x_{2t}, consisting of economic growth measured by a three-month moving average of the Chicago Fed National Activity Index (GRO) and inflation measured by one-year CPI inflation expectations from the Blue Chip Financial Forecasts (INF), the $\bar{R}^2$ increased to 0.37. For y_{t+h} the excess return on a two-year bond, the change is even more striking, with $\bar{R}^2$ increasing from 0.14 without the macro variables to 0.48 when they are included. JPS interpreted these adjusted $\bar{R}^2$ as strong evidence that macroeconomic variables have predictive power for excess bond returns beyond the information in the yield curve itself, and concluded from this evidence that "macroeconomic risks are unspanned by bond yields" (p. 1203).

However, the predictors in x_{2t} are very persistent. As shown in Table 4, the first-order sample autocorrelations for GRO and INF are 0.91 and 0.99, respectively. The results in Sections 2.1 and 2.2 thus suggest the change in $\bar{R}^2$ should be interpreted with some caution.

The first row of Table 5 reports the actual $\bar{R}^2$ for the 2-year and 10-year excess return regressions as in (1), and essentially replicates the results in JPS.[14] The entry $\bar{R}_1^2$ gives the adjusted $\bar{R}^2$ 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. The second row reports the mean $\bar{R}^2$ across 1000 replications of the bootstrap described in Section 2.3, 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 but whose serial correlation properties are similar to those of the observed data. The third row gives 95% confidence intervals for the estimated $\bar{R}^2$, constructed from the appropriate quantiles of the bootstrap distribution of the test statistics.

For all predictive regressions, the variability of the adjusted $\bar{R}^2$ is very high. Values for $\bar{R}_2^2$ up to 55% would not be uncommon. Most notably, adding the regressors x_{2t} often substantially increases the adjusted R^2, by up to 25 percentage points or more, although x_{2t} has no predictive power in population by construction. For the ten-year bond, JPS report an increase of 17 percentage points when adding macro variables, but our results show that this increase is in fact not statistically significant at conventional significance levels. Only for the two-year bond is $\bar{R}_2^2 - \bar{R}_1^2$ slightly outside our bootstrap confidence interval.

Since the persistence of x_{2t} is high, it may be important to adjust for small-sample bias in the VAR estimates. Hence we also carried out the bias-corrected (BC) bootstrap. The expected values and 95% confidence intervals for $\bar{R}^2$ are reported in rows 4 and 5 of Table 5. As expected, with more serial correlation in the generated data, the variability of the adjusted $\bar{R}^2$, as well as their difference, increases. Consequently, the statistical evidence for predictive power of GRO and INF would be regarded as even weaker.

The theory described in Section 2.1 predicts that this problem should go away as the sample size grows. The second panel of Table 5 updates the analysis to include an additional 7 years of data. As expected, the value of $\bar{R}_2^2$ that is observed in the data falls significantly when new data are added. And although the bootstrap 95% confidence intervals are tighter with

[14] The yield data set of JPS includes the six-month and the one- through ten-year Treasury yields. After calculating annual returns for the two- to ten-year bonds, JPS discard the six, eight, and nine-year yields before fitting PCs and their term structure models. Here, we need the fitted nine-year yield to construct the return on the ten-year bond, so we keep all 11 yield maturities. While our PCs are therefore slightly different than those in JPS, the only noticeable difference is that our adjusted $\bar{R}^2$ in the regressions for the two-year bond with yield PCs and macro variables is 0.49 instead of their 0.48.

the longer data set, the conclusion that there is no statistically significant evidence of added predictability provided by x_{2t} is even more compelling. Both for the two-year and ten-year bond, the increases in adjusted $\bar{R}^2$ from adding macro variables as predictors lie comfortably inside the bootstrap confidence intervals.

3.2 Predicting the level of the yield curve

JPS went on to estimate yield-curve models in which it is assumed that the macro factors x_{2t} directly help predict the principal components x_{1t}. The first block of their proposed vector autoregression takes the form

$$x_{1,t+1} = c + \phi_1 x_{1t} + \Gamma_1 x_{2t} + \varepsilon_{1,t+1}. \tag{20}$$

The estimates reported in Table 3 of their paper result from a yield-curve model with overidentifying restrictions that are implied by the no-arbitrage assumption and tight restrictions on risk pricing. Here we analyze properties of simple direct estimation of (20), whose estimates turn out to be close to the structural estimates reported in JPS. We will focus on the first row of (20), which is a regression of the first principal component of the yields in period $t+1$ (approximately equal to an average of the yields) on the first three principal components, economic growth and inflation at t.[15] This corresponds to regression (1) with x_{1t} and x_{2t} the same as before, but with $h = 1$ and y_{t+1} equal to $PC1_{t+1}$. This regression is the crucial forecasting equation, since forecasts of *any* yield are dominated by the forecast for the level of the yield curve. The estimated coefficients from this regression are reported in the first row of Table 6. These are comparable to the estimates reported in the first row of JPS Table 3.

The standard errors in JPS original Table 6 incorporate the restrictions implied by the structural model but make no allowance for possible serial correlation of the product $x_t u_{t+1}$. One popular approach to guard against this possibility is to use the HAC standard errors and test statistics proposed by Newey and West (1987). In the second row of our Table we report the resulting t-statistic for each coefficient (using 18 lags for the Newey-West correction) along with the Wald test of the hypothesis $\beta_2 = 0$. The third row reports p-values assuming that the usual asymptotic interpretations (Normal or χ^2_2,respectively) of these HAC-calculated statistics are accurate.

We then used our bootstrap to calculate the properties of the HAC tests for data with serial correlation properties similar to those observed in the sample. Surprisingly, we find

[15] To make our estimates comparable to those of JPS, we rescale our PCs in the same way that they do (see footnote 19 of JPS).

that the true size of these tests is 12-22% instead of the presumed 5%. When we take the added step of bias-correcting the bootstrap, none of the tests is statistically significant at the 5% level, though the Wald test would come very close to rejecting ($p = 0.052$).

We again find that the statistical evidence of predictability declines significantly when more data are added to the sample, as seen in the second panel of Table 6. When the data set is extended through 2013, the HAC t-statistics would no longer be statistically significant at 5% even if interpreted assuming the usual asymptotics, and are far from significant when we take into account the serial correlation of the actual data.

The bottom two rows for each panel in Table 6 report the p-values for the IM test of the individual significance of the coefficients. In both samples, the level of the yield curve ($PC1$) is a strongly significant predictor, with p-values below two percent for both IM tests. This will turn out to be a consistent finding in all the data sets that we will look at– the level or slope of the yield curve appear to be robust predictors of bond risk premia, consistent with an old literature going back to Fama and Bliss (1987) and Campbell and Shiller (1991). The low p-values are also consistent with the conclusion from our unreported Monte Carlo investigation that IM has good power to reject a false null hypothesis.

By contrast, in both samples the coefficients on GRO and INF are not statistically significant at conventional significance levels based on the IM test, consistent with the conclusion drawn from our bootstrap calculations.

We conclude that the evidence in JPS on the predictive power of macro variables for yields and bond returns is not robust. Notwithstanding, JPS noted that theirs is only one of several papers claiming to have found such evidence. We turn in the next section to another influential study.

4 Predicting yields using factors of large macro data sets

Ludvigson and Ng (2009, 2010) found that factors extracted from a large macroeconomic data set are helpful in predicting excess bond returns, above and beyond the information contained in the yield curve, adding further evidence for the claim of unspanned macro risks and against the hypothesis of invertibility. Here we revisit this evidence, focusing on the results in Ludvigson and Ng (2010) (henceforth LN).

LN 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. The authors carried out an extensive specification search in which they considered many different combinations of the factors along with squared and cubic terms. They also included in their specification search the bond-pricing factor proposed by Cochrane and Piazzesi (2005), which is the linear combination of forward rates that best predicts the average excess return across maturities, and which we denote here by CP. LN's conclusion was that macro factors appear to help predict excess returns, even when controlling for the CP factor. This conclusion is mostly based on comparisons of adjusted $\bar{R}^2$ in regressions with and without the macro factors and on HAC inference using Newey-West standard errors.

4.1 Robust inference about coefficients on macro factors

One feature of LN's design obscures the evidence relevant for the null hypothesis that is the focus of our paper. Their null hypothesis is that the CP factor alone provides all the information necessary to predict bond yields, whereas our null hypothesis of interest is that the 3 variables $(PC1, PC2, PC3)$ contain all the necessary information. Their regressions in which CP alone is used to summarize the information in the yield curve could not be used as a basis to reject our null hypothesis. For this reason, we begin by examining similar predictive regressions to those in LN in which excess bond returns are regressed on three PCs of the yields and all eight of the LN macro factors. We further leave aside the specification search of LN in order to focus squarely on hypothesis testing for a given regression specification.[16] These regressions take the same form as (1), where now $y_{t+h} = rx^{(n)}_{t,t+12}$ is the one-year return on an n-year bond in excess of the one-year yield, 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 7 reports regression results for the excess return on the two-year and the five-year bond. We first focus on the results obtained in LN's original sample, reported in the top panel. The first three rows for each set of results show the coefficient estimates, HAC t- and Wald statistics (using Newey-West with 18 lags as in LN), and p-values based on the asymptotic distributions of these test statistics. For the two-year bond, there are five macro factors that appear to be statistically significant at the ten-percent level, among which two are significant at the one-percent level. The same is true for the five-year bond.[17] In both cases, the Wald statistic for H_0 far exceeds the critical values for conventional significant levels (the 5%-critical value for a χ^2_8-distribution is 15.5). Table 8 also reports adjusted $\bar{R}^2$ for the

[16] 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.

[17] The p-value for $F4$ is rounded from 0.0997 to 0.100.

restricted ($\bar{R}_1^2$) and unrestricted ($\bar{R}_2^2$) regressions, and shows that this measure of fit increases by 14 percentage points for the two-year bond and 10 percentage points for the five-year bond when the macro factors are included. Taken at face value, this evidence suggests that macro factors have strong predictive power, above and beyond the information contained in the yield curve, consistent with the overall conclusions of LN.

How robust are these econometric results? We again use the bootstrap to test H_0, as described in 2.3. The yield factors x_{1t} are again the first three PCs of observed yields, and in this data, the (pooled) fitting errors have a standard deviation of 4.3 basis points. The predictor x_{2t} is now an (8×1) vector of macro factors, for which we estimate a VAR with two lags.[18] As before, we simulate 1000 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—the bias in the VAR for x_{2t} is estimated to be small.

Before turning to the results, it is worth noting the differences between our bootstrap exercise and the bootstrap carried out by LN. Their bootstrap is designed to test the null hypothesis that excess returns are not predictable against the alternative that they 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 all the results in our paper as well. Our null hypothesis of interest, however, is that excess returns are predictable only by current yields. Our bootstrap, in contrast to the bootstrap of LN, is designed to test this hypothesis.

Our bootstrap reveals that the tests using asymptotic p-values have serious size distortions. The true size of the t-tests is 8-16 percent, instead of the nominal five percent. For the Wald test, the size distortion is particularly high, with a true size of about 33 percent. Due to these size distortions, the bootstrapped p-values are larger than the asymptotic p-values, and several coefficients are now less significant or not significant at all. The Wald statistics are nevertheless still significant using bootstrap p-values. However, Table 8 shows that the observed increase in predictive power from adding macro factors to the regression, measured by the adjusted $\bar{R}^2$, would not be implausible if the null hypothesis were true. For the two-year bond, this increase is only barely outside the 95% bootstrap confidence interval, and for the five-year bond, this increase is within the confidence interval.

Table 7 also reports p-values for the two IM tests, using $q = 8$ and 16 subsamples. The interpretation of the results is complicated by the fact that some coefficients are significant for $q = 8$ but not for $q = 16$, or the other way around. The overall picture is, however, quite

[18] The lag length of two is based on the SBIC.

clear: The only predictors that are robustly significant for both the two-year and five-year bond are the level and the slope of the yield curve. There are no macro factors for which the IM tests show similarly strong evidence of a predictive relation.

These results imply that the evidence that macro factors have predictive power *beyond the information already contained in yields* is somewhat weaker than the results in LN would initially have suggested. For some of the coefficients, some of the tests remain statistically significant at the 5% level. But many of the tests that initially appeared to be significant fail to reject the null hypothesis at conventional levels when interpreted correctly. Our overall conclusion is that once small-sample concerns are taken into account, any evidence against the null hypothesis of no unspanned factors is much weaker than would have originally appeared to be the case.

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 (2013, 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. To explore this further we repeated our analysis using the same 1985-2013 sample period that was used in the second panel of Tables 5 and 6. Note that whereas in the case of JPS this was a strictly larger sample than the original, in the case of LN our second sample adds data at the end but leaves some out at the beginning. Reasons for interest in the this sample period include the significant break in monetary policy after 1984, 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.[19]

We used the macro data set of McCracken and Ng (2014), to extract macro factors in the same way as LN over the more recent data.[20] The bottom panels of Tables 7 and 8 display the results. Over this sample period, the evidence for the predictive power of macro factors is considerably weaker. Notably, the Wald tests reject H_0 for both bond maturities (at the ten-percent level for the five-year bond) when using asymptotic critical values, but do not reject when using bootstrap critical values. The increases in adjusted $\bar{R}^2$ in Table 8 are not statistically significant, and the IM tests find essentially no evidence of predictive power of the macro factors.

[19] We also analyzed the full 1964-2013 sample and obtained similar results as over the 1964-2007 sample.

[20] Using this macro data set and the same sample period as LN we obtained results that were very similar to those in the original paper, which gives us confidence in the consistency of the macro data set.

4.2 Robust inference about return-forecasting factors

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 $H8$. The CP factor of Cochrane and Piazzesi (2005) is similarly constructed using a regression on five forward rates. Adding $H8$ to a predictive regression with CP substantially increases the adjusted $\bar{R}^2$, 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.

Table 9 replicates LN's results for these regressions for the two- and five-year bond maturity.[21] In their data, both CP and $H8$ are strongly significant with HAC p-values below 0.1%. Adding $H8$ to the regression increases the adjusted $\bar{R}^2$ by 11 and 9 percentage points, respectively, for the two-year and five-year bond. How plausible would it have been to obtain these results if macro factors have in fact no predictive power? In order to answer this question, we adjust our bootstrap design to handle regressions with return-forecasting factors CP and $H8$. 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. The results in Table 9 show that the size distortions for tests of the significance of the macro return-forecasting factor are enormous: a test with nominal size of 5% that uses asymptotic HAC p-values has a true size of 54-56%. The bootstrap p-values increase substantially, and $H8$ is no longer significant at the 1% level (though it would still be significant at the 5% level). The observed increases in adjusted $\bar{R}^2$ when adding $H8$ to the regression fall inside the 95% bootstrap confidence intervals. One reason for the substantially distorted inference using conventional statistics is the high persistence of the return-forecasting factors. Table 4 shows that both $H8$ and CP have autocorrelations that are near 0.8 at first order, and decline only slowly with the lag length.

We also examined the same regressions over the 1985–2013 sample period with results shown in the bottom panel of Table 9. In this sample, the return-forecasting factors would again appear to be highly significant based on HAC p-values, but the coefficients on $H8$ are not statistically significant when using the correct bootstrap p-values. The size distortions are even larger in this sample, up to 63%, due to the smaller sample size. The observed increases in adjusted $\bar{R}^2$ are near the bootstrap mean of this statistic, i.e., they are squarely in line with

[21] These results correspond to those in column 9 in tables 4 and 7 in LN.

what we would expect under the null.

This evidence suggests that conventional HAC inference can be even more problematic if return-forecasting factors are constructed in a preliminary step, and that 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 does not exhibit nearly as strong and robust predictive power for excess bond returns as may have appeared to be the case in LN's original analysis.

5 Predicting yields using higher-order PCs of yields

In a seminal paper, Cochrane and Piazzesi (2005) (henceforth CP) documented several striking new facts about excess bond returns. Focusing on returns with a one-year holding period, they showed that the same linear combination of forward rates predicts excess returns on different long-term bonds, that the coefficients of this linear combination have a tent shape, and that the predictive regressions using this one variable delivers 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 significant predictors of future bond returns (see CP's Table 4 on p. 147, row 3). In particular, the fourth PC, while explaining only a tiny fraction of the cross-sectional variation in yields, appeared "very important for explaining expected returns" (p. 147).

The null hypothesis of interest for us is that only the first three PCs predict yields and excess returns, and that higher-order PCs do not contain additional predictive power. This null is more restrictive than the invertibility assumption/Markov property of the yield curve: under invertibility, it could well be the case that higher-order PCs are informative about the state variables relevant for predicting yields and returns. Our more restrictive null hypothesis is motivated by the long-standing evidence that three factors are sufficient to fully capture the shape and evolution of the yield curve, which goes back at least to Litterman and Scheinkman (1991). In the CP data, the first three PCs explain 99.97% (!) of the variation in the five Fama-Bliss yields (see page 147 of CP). It is very surprising, and indeed hard to believe, that the remaining 0.03% of the variation in yields contain any substantial information relevant for predicting yields and returns. In other words, information in the yield curve that we cannot see with our bare eyes is hard to use for forecasting. Hence, invertibility/spanning should also hold if the information set is restricted to the first three PCs of the yield curve. This is why we are interested in the robustness of the finding that these PCs are not sufficient to capture bond risk premia.

First, we replicate the relevant results of CP using their original data. We estimate the predictive regression for the average excess bond return using five PCs as predictors, and carry out HAC inference in this model. The results are in the top panel of Table 10. The Wald statistic and R_1^2 and R_2^2 are identical to those reported by CP. The p-values indicate that $PC4$ is very strongly statistically significant, and that our null hypothesis would be rejected.

In contrast to the results found for JPS in Section 3 and LN in Section 4, our bootstrap finds that the CP results cannot be accounted for by serial correlation alone. The reason is that the predictors $PC4$ and $PC5$ are less persistent. As shown in Table 4, their first-order autocorrelation coefficients are only 0.43 and 0.23, respectively. There are some size distortions for the Newey-West HAC statistics– the true size for the t-tests is 8-10 percent, and for the Wald test it is 12 percent– but these are not big enough to overturn CP's conclusion. Furthermore, the increase in R^2 reported by CP would be quite implausible to observe under the null hypothesis, given that it is far outside the 95% bootstrap interval under the null.

To address the econometric issues of return-forecasting regressions with overlapping returns, CP use three different types of HAC standard errors. Table 10 reported results only for Newey-West, but we have also investigated the cases with Hansen-Hodrick (HH) and the "Simplified Hansen-Hodrick" (SHH) standard errors used by CP. We found that while HH leads to similar size distortions as NW, the tests using SHH standard errors are correctly sized in our simple bootstrap setting, because the assumptions about autocorrelation and conditional homoskedasticity are correct by construction. However, in an alternative bootstrap setting where the errors are conditionally heteroskedastic (see footnote 12), this method of inference suffers from some size distortions as well.[22] Overall, we conclude that while results from standard HAC inference can lead to some overconfidence in the results, the broad conclusions of CP cannot be attributed to size distortions of these tests.

It is nevertheless of interest that the IM t-tests would fail to reject the null hypothesis that $\beta_2 = 0$ even when the inference is based on CP's original sample. These indicate 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. Figure 3 provides some intuition about why the IM tests fail to reject. It shows the coefficients on each predictor across the $q = 8$

[22] CP also carry out three different bootstrap exercises, which are in principle well suited to deal with small-sample issues. However, these are generally designed to test the null hypothesis that excess returns are unpredictable (the expectations hypothesis). Although their "small T" inference is intended to give the correct small-sample distribution of the relevant Wald statistic, none of their bootstrap simulations generate artificial data under the null hypothesis that we are interested in.

subsamples used in the IM test. The coefficients are standardized by dividing them by the sample standard deviation across the eight estimated coefficients for each predictor. Thus, t-statistics, which are also reported in Figure 3, are equal to the means of the standardized coefficients across subsamples, multiplied by $\sqrt{8}$. The figure shows that $PC1$ and $PC2$ had much more consistent predictive power across subsamples than $PC4$, whose coefficient 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.[23] This illustrates that the IM test, which is designed to produce inference that is robust to serial correlation, at the same time delivers results that are robust to sub-sample instability. Only the level and slope have predictive power for excess bond returns in the CP data that is truly robust in both meanings of the word.

The evidence of CP indeed seems to be quite sensitive to sample choice. Duffee (2013, Section 7) found that extending CP's sample period to 1952–2010 alters some of their key results. Similarly, we have found that using Duffee's sample period the predictive power of higher-order PCs disappears. Here we focus on our preferred sample period, from 1985 to 2013, for which we report results in the bottom panel of Table 10. In this case, the coefficients on $PC4$ and $PC5$ are not significant for any method of inference, and the increase in R^2 due to inclusion of higher-order PCs are comfortably in the 95% bootstrap intervals. At the same time, the predictive power of the level and slope of the yield curve is quite strong also in this sample. Although the standard HAC t-test fails to reject that the coefficient on the level is zero, the same test finds the coefficient on the slope to be significant, and the IM tests imply that both coefficients are significant.

Since CP used a sample period that ended more than ten years prior to the time of this writing, we can carry out a true out-of-sample test of our hypothesis of interest. We estimate the same predictive regressions as in CP, for excess returns on two- to five-year bonds as well as for the average excess return across bond maturities. The first two columns of Table 11 report the in-sample R^2 for the restricted models (using only $PC1$ to $PC3$) and unrestricted models (using all PCs). Then we construct expected future excess returns from these models using yield PCs[24] from 2003:1 through 2012:12, and compare these to realized excess returns for holding periods ending in 2004:1 through 2013:12. Table 11 shows the resulting root-mean-squared forecast errors (RMSEs). For all bond maturities, the model that leaves out $PC4$ and $PC5$ performs substantially better, with reductions of RMSEs around 20 percent.

[23] Consistent with this finding, an influential 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.

[24] Principal components are calculated throughout using the loadings estimated over the original CP sample period.

The test for equal forecast accuracy of Diebold and Mariano (2002) rejects the null, indicating that the performance gains of the restricted model are statistically significant. Figure 4 shows the forecast performance graphically, plotting the realized and predicted excess bond returns. Clearly, both models did not predict future bond returns very well, expecting mostly negative excess returns over a period when these turned out to be positive. In fact, the unconditional mean, estimated over the CP sample period, was a better predictor of future returns. This is evident both from Figure 4, which shows this mean as a horizontal line, and from the RMSEs in the last column of Table 11. Nevertheless, the unrestricted model implied expected excess returns that were more volatile and significantly further off than those of the restricted model from the future realizations. Restricting the predictive model to use only the level, slope and curvature leads to more stable and more accurate return predictions.

We conclude from both our in-sample and out-of-sample results that the evidence for predictive power of higher-order factors is tenuous and sample-dependent. To estimate bond risk premia in a robust way, we recommend using only those predictors that consistently show a strong associations with excess bond returns, namely the level and the slope of the yield curve.

6 Predicting yields using measures of bond supply

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.[25]

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 where they control for the information in the yield curve, namely their results for regressions of future bond returns on the current one-year yield, the term spread, and their preferred measure of bond supply. In the top panel of Table 12 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.[26] Like GV we use Newey-West standard errors with 36 lags.[27]

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 5% when the yield spread is included in the regression, as seen in the second panel. But once again the predictors in these regressions are extremely persistent, leading us to suspect that the true p-value likely exceeds the purported 0.058 —the first-order autocorrelations of the yield spread and the bond supply variable are 0.960 and 0.998, respectively, as reported in Table 4.

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.[28] 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.[29] 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 12. Again, the coefficient on bond supply is only marginally significant for the HAC

[25] 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.

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

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

[28] GV obtained this series from Ibbotson Associates.

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

t-test, and insignificant for the IM tests. In contrast, the coefficients on both PC1 and PC2 are strongly significant for the IM tests.

Finally, we consider a different, more common excess bond return. Instead of the return on a hypothetical 20-year bond, we report results for one-year excess returns calculated from the commonly used Fama-Bliss yields. In particular we use, like CP and Section 5, the average excess return for bonds with two- though five-year maturities. The last panel of Table 12 shows that in this case, the coefficient on bond supply is insignificant. As usual, there is robust evidence that PC1 and PC2 have predictive power for bond returns. In this case we can apply our bootstrap procedure, and the bootstrap p-value is even higher, but we omit the bootstrap results for the sake of brevity. Even without accounting for small-sample problems there is no evidence against the spanning hypothesis based on GV's bond supply variable.

Overall, the results in the Greenwood-Vayanos data lead to the conclusions that the level and slope of the yield curve are strong predictors of excess bond returns, whereas the predictive power of bond supply measures is very tenuous and not robust.

References

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

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.

Carrodus, Mark L and David EA Giles (1992) "The exact distribution of R 2 when the regression disturbances are autocorrelated," *Economics Letters*, Vol. 38, pp. 375–380.

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

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 (2002) "Comparing predictive accuracy," *Journal of Business & economic statistics*, Vol. 20.

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

Engle, Robert (2002) "Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models," *Journal of Business & Economic Statistics*, Vol. 20, pp. 339–350.

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

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

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.

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, p. 11971233.

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.

Koerts, Johannes and Adriaan Pieter Johannes Abrahamse (1969) *On the theory and application of the general linear model*: Rotterdam University Press Rotterdam.

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.

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

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.

Appendix

A Conventional asymptotic results

Here we provide details of the claims made in Section 2.1. 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 SSR from the latter regression can be written

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

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

$$SSR_1 - SSR_2 = \textstyle\sum(x_t'b - x_{1t}'b_1^*)^2. \tag{21}$$

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

$$x_t'b = x_{1t}'b_1^* + \tilde{x}_{2t}'b_2 \tag{22}$$

where $\tilde{x}_{2t}$ denotes the residuals from regressions of the elements of x_{2t} on x_{1t} and b_2 can be obtained from an OLS regression of $y_{t+h} - x_{1t}'b_1^*$ on $\tilde{x}_{2t}$.[30] Thus from (21) and (22),

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

If the true value of β_2 is zero, then by plugging (1) into the definition of b_2 and using the fact that $\sum\tilde{x}_{2t}x_{1t}'\beta_1 = 0$ (which follows from the orthogonality of $\tilde{x}_{2t}$ with x_{1t}) we see that

$$b_2 = \left(\textstyle\sum\tilde{x}_{2t}\tilde{x}_{2t}'\right)^{-1}\left(\textstyle\sum\tilde{x}_{2t}u_{t+h}\right) \tag{23}$$

$$\begin{aligned} SSR_1 - SSR_2 &= b_2'\left(\textstyle\sum\tilde{x}_{2t}\tilde{x}_{2t}'\right)b_2 \\ &= \left(T^{-1/2}\textstyle\sum u_{t+h}\tilde{x}_{2t}'\right)\left(T^{-1}\textstyle\sum\tilde{x}_{2t}\tilde{x}_{2t}'\right)^{-1}\left(T^{-1/2}\textstyle\sum\tilde{x}_{2t}u_{t+h}\right). \end{aligned} \tag{24}$$

[30] That is, $b_2 = \left(\sum\tilde{x}_{2t}\tilde{x}_{2t}'\right)^{-1}\left(\sum\tilde{x}_{2t}(y_{t+h} - x_{1t}b_1^*\right)$ for $\tilde{x}_{2t}$ defined in (9) and (10). The easiest way to confirm the claim is to show that the residuals implied by (22) 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 orthogonality of $y_{t+h} - x_{1t}'b_1^* - \tilde{x}_{2t}'b_2$ to $\tilde{x}_{2t}$ follows directly from the definition of b_2. 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 $x_{2t} = \tilde{x}_{2t} + A_T x_{1t}$.

If x_t is stationary and ergodic, then it follows from the Law of Large Numbers that

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

which equals Q in (6) in the special case when $E(x_{2t}x_{1t}') = 0$. For the last term in (24) we see from (9) and (10) that

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

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

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

This will be recognized as $\sqrt{T}$ times the sample mean of a random vector with population mean zero, so from the Central Limit Theorem

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

implying from (24) that

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

Thus from (3),

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

as claimed in (4).

Expression (23) also implies that

$$\sqrt{T}b_2 = \left(T^{-1}\textstyle\sum \tilde{x}_{2t}\tilde{x}_{2t}'\right)^{-1}\left(T^{-1/2}\textstyle\sum \tilde{x}_{2t}u_{t+h}\right) \xrightarrow{d} Q^{-1}r$$

from which (12) follows immediately.

B Local-to-unity asymptotic results

Here we provide details behind the claims made in Section 2.2. Note that

$$\begin{aligned} \textstyle\sum \tilde{x}_{2t}\tilde{x}_{2t}' &= \textstyle\sum (x_{2t} - A_T x_{1t})(x_{2t} - A_T x_{1t})' \\ &= \textstyle\sum x_{2t}x_{2t}' - \left(\textstyle\sum x_{2t}x_{1t}'\right)\left(\textstyle\sum x_{1t}x_{1t}'\right)^{-1}\left(\textstyle\sum x_{1t}x_{2t}'\right) \end{aligned}$$

and

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

Taking the scalar case for illustration, we have as in Phillips (1988, p. 1026) that $T^{-2}\sum x_{2t}^2 \Rightarrow \sigma_2^2 \int [J_{c_2}(\lambda)]^2 d\lambda$, $T^{-2}\sum x_{2t}x_{1t} \Rightarrow \sigma_2\sigma_1 \int J_{c_2}(\lambda)J_{c_1}(\lambda)d\lambda$, and $T^{-2}\sum x_{1t}^2 \Rightarrow \sigma_1^2 \int [J_{c_1}(\lambda)]^2 d\lambda$ where

$\Rightarrow$ denotes weak convergence. Under the null hypothesis, the t-test of $\beta_2 = 0$ can be written

$$\frac{\sum \tilde{x}_{2t} u_{t+1}}{\left\{s^2 \sum \tilde{x}_{2t}^2\right\}^{1/2}} = \frac{T^{-1} \sum \tilde{x}_{2t} \varepsilon_{1,t+1}}{s \left\{T^{-2} \sum \tilde{x}_{2t}^2\right\}^{1/2}}$$

for $s^2 = (T-2)^{-1} \sum (y_{t+1} - b_1 x_{1t} - b_2 x_{2t})^2$. But $s^2 \xrightarrow{p} \sigma_1^2$ and

$$\begin{aligned} T^{-1} \sum \tilde{x}_{2t} \varepsilon_{1,t+1} &= T^{-1} \sum x_{2t} \varepsilon_{1,t+1} - \left(T^{-2} \sum x_{2t} x_{1t}\right) \left(T^{-2} \sum x_{1t}^2\right)^{-1} \left(T^{-1} \sum x_{1t} \varepsilon_{1,t+1}\right) \\ &\Rightarrow \sigma_2 \sigma_2 \int J_{c_2}(\lambda) dW_1(\lambda) - \frac{\left[\sigma_1 \sigma_2 \int J_{c_2}(\lambda) J_{c_1}(\lambda) d\lambda\right] \left[\sigma_1^2 \int J_{c_1}(\lambda) dW_1(\lambda)\right]}{\left[\sigma_1^2 \int [J_{c_1}(\lambda)]^2 d\lambda\right]}. \end{aligned}$$

Combining these results produces (18) as claimed.

Similar analysis identifies (19) as the local-to-unity asymptotic distribution of

$$\frac{\sum x_{2t} u_{t+1}}{\left\{s^2 \sum x_{2t}^2\right\}^{1/2}}.$$

Table 1: Size distortions in simulation study

	ρ_2						
ρ_1	1	0.999	0.99	0.95	0.9	0.5	0
$T = 50$							
1	16.0	15.3	14.3	11.7	10.5	6.1	5.1
0.999	15.6	14.8	15.0	12.7	11.0	5.7	4.9
0.99	16.5	15.3	14.3	12.9	11.9	6.3	5.1
0.95	15.1	14.3	14.2	12.2	11.6	6.1	4.9
0.9	13.0	12.5	12.7	11.4	10.8	6.4	5.3
0.5	7.1	6.4	6.8	6.7	6.9	4.9	4.5
0	5.4	5.4	5.4	5.3	5.6	4.8	4.7
$T = 200$							
1	16.3	16.3	13.3	9.9	8.0	5.6	5.2
0.999	16.6	16.9	14.0	10.2	8.3	5.7	5.3
0.99	17.2	16.6	14.1	10.4	8.9	5.0	5.2
0.95	11.0	11.2	10.7	8.5	7.6	5.1	4.9
0.9	8.6	8.8	8.4	8.1	6.8	5.3	5.2
0.5	5.4	5.4	6.0	5.0	5.2	5.8	5.0
0	4.9	5.3	5.1	4.5	4.6	4.8	4.7

Frequency of rejections (in percent) using a standard t-test with 5% nominal size. For details about the simulation design, please refer to main text.

Table 2: Bias and size distortions in simulation study

	$\rho_1 = \rho_2 = 1$		$\rho_1 = \rho_2 = 0.99$	
	β_1	β_2	β_1	β_2
True coefficient	1.000	0.000	0.990	0.000
$T = 50$				
True coefficient	1.000	0.000	0.990	0.000
Mean coeff. estimate	0.864	0.001	0.853	-0.002
Coefficient bias	-0.136	0.001	-0.137	-0.002
Sample SD of coeff. estimates	0.100	0.101	0.103	0.105
Mean standard error	0.070	0.071	0.073	0.073
Standard error bias	-0.029	-0.030	-0.030	-0.031
Empirical size	0.403	0.150	0.366	0.150
Empirical size using sample SD	0.220	0.049	0.215	0.049
$T = 200$				
True coefficient	1.000	0.000	0.990	0.000
Mean coeff. estimate	0.964	0.000	0.956	0.000
Coefficient bias	-0.036	0.000	-0.034	0.000
Sample SD of coeff. estimates	0.027	0.027	0.028	0.028
Mean standard error	0.018	0.018	0.020	0.020
Standard error bias	-0.009	-0.009	-0.008	-0.008
Empirical size	0.432	0.171	0.287	0.132
Empirical size using sample SD	0.215	0.054	0.193	0.051

Analysis of bias in estimated coefficients and standard errors. For details about the simulation design, please refer to main text.

Table 3: R^2 in simulation study

Strict exog.	Statistic	Mean	95% CI	Std. dev.
Yes	R_1^2	2.2	[0.0, 9.8]	2.9
	R_2^2	4.4	[0.1, 15.6]	4.1
	$R_2^2 - R_1^2$	2.1	[0.0, 10.6]	2.9
No	R_1^2	6.2	[0.0, 17.8]	4.9
	R_2^2	9.9	[0.9, 23.7]	6.1
	$R_2^2 - R_1^2$	3.7	[0.0, 15.4]	4.4

Small-sample distributions of R^2 (in percent) in simulations with strictly vs. weakly endogenous regressors x_{1t}. The sample size is $T = 50$, and the regressors have autocorrelation coefficients of $\rho_1 = \rho_2 = 0.99$. The true R^2 is $\frac{1-\rho_1}{2} = 0.5\%$. For details about the simulation design, please refer to main text.

Table 4: Persistence of predictors in published studies

Study	Predictor	Original sample			Later sample		
		1	6	12	1	6	12
JPS	PC1	0.974	0.840	0.696	0.983	0.890	0.784
	PC2	0.973	0.774	0.467	0.968	0.753	0.444
	PC3	0.849	0.380	0.216	0.833	0.395	0.272
	GRO	0.910	0.507	0.260	0.947	0.589	0.250
	INF	0.986	0.897	0.815	0.985	0.892	0.822
LN	PC1	0.984	0.904	0.821	0.984	0.891	0.785
	PC2	0.944	0.734	0.537	0.959	0.718	0.422
	PC3	0.601	0.254	0.113	0.749	0.339	0.192
	F1	0.766	0.381	0.088	0.700	0.463	0.139
	F2	0.748	0.454	0.188	0.499	0.386	0.128
	F3	-0.233	0.035	-0.085	-0.123	-0.066	-0.151
	F4	0.455	0.207	0.151	0.486	0.215	0.031
	F5	0.361	0.207	0.171	0.136	0.186	-0.020
	F6	0.422	0.476	0.272	0.033	0.031	-0.014
	F7	-0.111	0.134	0.054	-0.032	-0.059	-0.072
	F8	0.225	0.087	0.093	-0.328	0.099	0.005
	H8	0.777	0.627	0.331	0.580	0.463	0.313
	CP	0.773	0.531	0.377	0.886	0.615	0.379
CP	PC1	0.980	0.880	0.767	0.984	0.891	0.785
	PC2	0.940	0.721	0.539	0.959	0.718	0.422
	PC3	0.592	0.237	0.110	0.749	0.339	0.192
	PC4	0.425	0.137	0.062	0.649	0.232	0.068
	PC5	0.227	0.157	-0.135	0.543	0.167	-0.103
	CP	0.767	0.522	0.361	0.889	0.634	0.399
GV	yield	0.984	0.905	0.827			
	spread	0.960	0.762	0.580			
	PC1	0.988	0.925	0.860			
	PC2	0.942	0.722	0.521			
	PC3	0.582	0.233	0.094			
	supply	0.998	0.990	0.974			

Persistence, measured by autocorrelations with lags of one, six, and twelve months, of predictors used in published predictability studies: JPS stands for Joslin et al. (2014), LN stands for Ludvigson and Ng (2010), CP stands for Cochrane and Piazzesi (2005), and GV stands for Greenwood and Vayanos (2014). The predictors are described in the corresponding sections in the main text. The original sample is the one used in the published study, whereas the later sample is from 1985 to 2013.

Table 5: Joslin-Priebsch-Singleton: predicting excess bond returns

	Two-year bond			Ten-year bond		
	$\bar{R}_1^2$	$\bar{R}_2^2$	$\bar{R}_2^2 - \bar{R}_1^2$	$\bar{R}_1^2$	$\bar{R}_2^2$	$\bar{R}_2^2 - \bar{R}_1^2$
Original sample: 1985–2008						
Data	0.14	0.49	0.35	0.20	0.37	0.17
Simple bootstrap	0.18	0.25	0.08	0.26	0.33	0.06
	(0.02, 0.42)	(0.06, 0.51)	(0.00, 0.27)	(0.07, 0.48)	(0.12, 0.55)	(0.00, 0.25)
BC bootstrap	0.18	0.28	0.10	0.24	0.32	0.08
	(0.01, 0.45)	(0.05, 0.56)	(0.00, 0.32)	(0.04, 0.50)	(0.09, 0.57)	(0.00, 0.29)
Later sample 1985–2013						
Data	0.12	0.28	0.16	0.20	0.28	0.08
Simple bootstrap	0.16	0.23	0.07	0.22	0.28	0.06
	(0.02, 0.37)	(0.05, 0.45)	(0.00, 0.24)	(0.03, 0.46)	(0.08, 0.51)	(0.00, 0.20)
BC bootstrap	0.15	0.23	0.08	0.24	0.30	0.06
	(0.01, 0.40)	(0.04, 0.48)	(0.00, 0.25)	(0.03, 0.50)	(0.07, 0.54)	(0.00, 0.21)

Adjusted R^2 for regressions of annual excess bond returns on three PCs of the yield curve ($\bar{R}_1^2$) and on three yield PCs together with the macro variables GRO and INF ($\bar{R}_2^2$), as well as the difference in adjusted R^2 (rows may not add up due to rounding). GRO is the three-month moving average of the Chicago Fed National Activity Index, and INF is one-year expected inflation measured by Blue Chip inflation forecasts. The first panel shows the results for the original data set used by Joslin et al. (2014); the second panel uses a data sample that is extended to December 2013. For each data sample and bond maturity, we report $\bar{R}^2$ for the restricted and unrestricted regressions in the data, as well as the mean and 95%-confidence intervals (in parentheses) for the bootstrap distribution of $\bar{R}^2$ for these regressions obtained under the null hypothesis that the macro variables have no predictive power. See the text for a description of the experimental design for the simple bootstrap and the bias-corrected (BC) bootstrap.

Table 6: Joslin-Priebsch-Singleton: predicting the level of the yield curve

	$PC1$	$PC2$	$PC3$	GRO	INF	Wald
Original sample: 1985–2008						
Coefficient	0.928	-0.013	-0.097	0.092	0.118	
HAC statistic	41.205	1.312	0.508	2.214	2.400	17.075
HAC p-value	0.000	0.191	0.612	0.028	0.017	0.000
Simple bootstrap 5% c.v.'s				2.608	2.829	11.821
Simple bootstrap p-values				0.090	0.099	0.021
Simple bootstrap true size				0.120	0.171	0.219
BC bootstrap 5% c.v.'s				2.926	3.337	17.270
BC bootstrap p-values				0.127	0.145	0.052
BC bootstrap true size				0.159	0.224	0.289
IM $q = 8$	0.000	0.864	0.436	0.339	0.456	
IM $q = 16$	0.000	0.709	0.752	0.153	0.554	
Later sample: 1985–2013						
Coefficient	0.958	-0.013	-0.209	0.024	0.087	
HAC statistic	54.015	1.371	1.326	0.786	1.874	5.999
HAC p-value	0.000	0.171	0.186	0.432	0.062	0.050
Simple bootstrap 5% c.v.'s				2.899	2.913	13.821
Simple bootstrap p-values				0.508	0.195	0.229
Simple bootstrap true size				0.151	0.172	0.229
BC bootstrap 5% c.v.'s				2.631	3.112	15.095
BC bootstrap p-values				0.564	0.240	0.265
BC bootstrap true size				0.140	0.224	0.265
IM $q = 8$	0.000	0.725	0.815	0.302	0.310	
IM $q = 16$	0.020	0.381	0.805	0.157	0.719	

Inference about predictive power of yield PCs and macro variables (described in the notes to Table 5) for one-month-ahead level of the yield curve (the first PC): HAC statistics and p-values are calculated using Newey-West standard errors with 18 lags. The column "Wald" reports χ^2-statistics for the null hypothesis that GRO and INF have no predictive power; the other column report results for individual t-tests. We obtain bootstrap distributions of the regression coefficients under the null hypothesis; 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 statistic is at least as large as in the data. We also report the bootstrap true size of a test with 5% nominal coverage—values higher than 0.05 indicate the presence of a small-sample size distortion. See the text for a description of the experimental design for the simple bootstrap and the bias-corrected (BC) bootstrap. 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.

Table 7: Ludvigson-Ng: yield and macro factors

	*PC*1	*PC*2	*PC*3	*F*1	*F*2	*F*3	*F*4	*F*5	*F*6	*F*7	*F*8	Wald
A. Original sample: 1964–2007												
Two-year bond												
Coefficient	-0.071	-0.973	2.825	0.471	-0.008	-0.085	-0.346	-0.083	-0.209	-0.133	0.254	
HAC statistic	1.797	2.640	3.515	2.350	0.043	1.442	2.652	0.673	1.698	1.675	2.888	54.514
HAC *p*-value	0.073	0.009	0.000	0.019	0.966	0.150	0.008	0.501	0.090	0.095	0.004	0.000
Bootstrap 5% c.v.'s				2.703	2.632	2.256	2.736	2.790	2.859	2.494	2.430	30.409
Bootstrap *p*-values				0.092	0.974	0.218	0.060	0.607	0.221	0.194	0.026	0.002
Bootstrap true size				0.153	0.121	0.088	0.150	0.145	0.157	0.116	0.112	0.337
IM $q = 8$	0.002	0.007	0.356	0.052	0.404	0.217	0.007	0.526	0.545	0.177	0.241	
IM $q = 16$	0.000	0.229	0.021	0.016	0.290	0.793	0.137	0.629	0.248	0.034	0.426	
Five-year bond												
Coefficient	-0.198	-3.106	6.496	0.883	0.269	-0.061	-0.663	-0.584	-0.916	-0.655	0.800	
HAC statistic	1.525	2.684	2.332	1.581	0.479	0.354	1.649	1.680	2.423	2.643	3.101	38.341
HAC *p*-value	0.128	0.008	0.020	0.115	0.632	0.723	0.100	0.094	0.016	0.008	0.002	0.000
Bootstrap 5% c.v.'s				2.644	2.639	2.233	2.637	2.668	2.714	2.416	2.419	30.072
Bootstrap *p*-values				0.254	0.689	0.766	0.200	0.190	0.074	0.036	0.016	0.012
Bootstrap true size				0.154	0.142	0.084	0.137	0.134	0.140	0.111	0.109	0.330
IM $q = 8$	0.001	0.001	0.060	0.155	0.689	0.800	0.205	0.778	0.474	0.057	0.383	
IM $q = 16$	0.000	0.022	0.326	0.497	0.335	0.556	0.473	0.257	0.201	0.031	0.598	
B. Later sample: 1985–2013												
Two-year bond												
Coefficient	0.088	-0.447	-0.077	0.529	-0.043	0.057	-0.143	0.189	0.193	0.037	0.002	
HAC statistic	2.019	0.967	0.051	3.123	0.096	0.649	0.743	0.801	2.432	0.401	0.032	20.464
HAC *p*-value	0.044	0.334	0.959	0.002	0.924	0.517	0.458	0.424	0.016	0.688	0.974	0.009
Bootstrap 5% c.v.'s				2.986	3.008	2.581	2.706	3.044	2.430	2.425	2.443	39.092
Bootstrap *p*-values				0.041	0.942	0.641	0.535	0.569	0.050	0.756	0.980	0.252
Bootstrap true size				0.171	0.196	0.134	0.153	0.172	0.114	0.107	0.109	0.433
IM $q = 8$	0.001	0.054	0.672	0.037	0.477	0.465	0.966	0.765	0.104	0.802	0.571	
IM $q = 16$	0.002	0.929	0.579	0.336	0.708	0.891	0.191	0.865	0.912	0.859	0.493	
Five-year bond												
Coefficient	0.202	-1.906	-6.465	0.520	-0.527	0.238	-0.861	-0.307	0.420	0.031	-0.190	
HAC statistic	1.260	1.207	1.021	0.881	0.342	0.733	1.374	0.354	1.597	0.090	0.706	14.626
HAC *p*-value	0.209	0.228	0.308	0.379	0.733	0.464	0.170	0.723	0.111	0.929	0.481	0.067
Bootstrap 5% c.v.'s				3.008	2.977	2.508	2.635	2.905	2.458	2.401	2.244	35.949
Bootstrap *p*-values				0.529	0.823	0.578	0.293	0.786	0.185	0.934	0.557	0.482
Bootstrap true size				0.162	0.167	0.124	0.136	0.179	0.113	0.092	0.085	0.429
IM $q = 8$	0.098	0.003	0.022	0.280	0.553	0.619	0.715	0.588	0.192	0.875	0.335	
IM $q = 16$	0.169	0.061	0.526	0.481	0.604	0.858	0.258	0.373	0.785	0.752	0.181	

Inference about predictive power of yield PCs and factors from a large data set of macro variables for annual excess returns. For a description, see the notes to Table 6.

Table 8: Ludvigson-Ng: $\bar{R}^2$ for yield and macro factors

	Two-year bond			Five-year bond		
	$\bar{R}_1^2$	$\bar{R}_2^2$	$\bar{R}_2^2 - \bar{R}_1^2$	$\bar{R}_1^2$	$\bar{R}_2^2$	$\bar{R}_2^2 - \bar{R}_1^2$
Original sample: 1964–2007						
Data	0.26	0.38	0.12	0.26	0.34	0.09
Bootstrap	0.27	0.30	0.03	0.27	0.30	0.03
	(0.11, 0.45)	(0.13, 0.48)	(0.00, 0.11)	(0.10, 0.46)	(0.13, 0.49)	(0.00, 0.11)
Later sample: 1985–2013						
Data	0.13	0.25	0.12	0.15	0.17	0.02
Bootstrap	0.16	0.20	0.05	0.18	0.22	0.04
	(0.02, 0.37)	(0.04, 0.43)	(-0.01, 0.18)	(0.02, 0.42)	(0.04, 0.46)	(-0.01, 0.15)

Adjusted $\bar{R}^2$ for regressions of annual excess bond returns on three PCs of the yield curve ($\bar{R}_1^2$) and on three yield PCs together with eight macro factors ($\bar{R}_2^2$), as well as the difference in $\bar{R}^2$ (rows may not add up due to rounding). The first panel shows the results for the original data set used by Ludvigson and Ng (2010); the second panel uses a data sample that starts in 1985 and ends in 2013. For each data sample and bond maturity, we report the adjusted R^2 for the restricted and unrestricted regressions in the data, as well as the mean and 95%-confidence intervals (in parentheses) for the bootstrap distribution of $\bar{R}^2$ for these regressions obtained under the null hypothesis that the macro factors have no predictive power. See the text for a description of the experimental design for the bootstrap.

Table 9: Ludvigson-Ng: return-forecasting factors

	CP	$H8$	$\bar{R}_1^2$	$\bar{R}_2^2$	$\bar{R}_2^2 - \bar{R}_1^2$
A. Original sample: 1964–2007					
Two-year bond					
Data	0.335	0.331	0.31	0.42	0.11
HAC t-statistic	4.429	4.331			
HAC p-value	0.000	0.000			
Bootstrap 5% c.v./mean $\bar{R}^2$		4.044	0.27	0.31	0.03
Bootstrap p-value/95% CIs		0.029	(0.11, 0.45)	(0.14, 0.48)	(0.00, 0.11)
Bootstrap true size		0.542			
Five-year bond					
Data	1.115	0.937	0.33	0.42	0.09
HAC t-statistic	4.371	4.541			
HAC p-value	0.000	0.000			
Bootstrap 5% c.v./mean $\bar{R}^2$		4.031	0.27	0.31	0.03
Bootstrap p-value/95% CIs		0.018	(0.11, 0.47)	(0.15, 0.49)	(0.00, 0.11)
Bootstrap true size		0.564			
B. Later sample: 1985–2013					
Two-year bond					
Data	0.349	0.371	0.15	0.23	0.07
HAC t-statistic	2.644	3.348			
HAC p-value	0.009	0.001			
Bootstrap 5% c.v./mean $\bar{R}^2$		4.112	0.14	0.20	0.05
Bootstrap p-value/95% CIs		0.146	(0.01, 0.35)	(0.05, 0.39)	(0.00, 0.17)
Bootstrap true size		0.596			
Five-year bond					
Data	1.320	1.021	0.17	0.21	0.05
HAC t-statistic	2.946	3.270			
HAC p-value	0.003	0.001			
Bootstrap 5% c.v./mean $\bar{R}^2$		4.149	0.18	0.23	0.05
Bootstrap p-value/95% CIs		0.171	(0.02, 0.40)	(0.07, 0.43)	(0.00, 0.15)
Bootstrap true size		0.632			

Inference about predictive power of return-forecasting factors CP and H8 used by Ludvigson and Ng (2010) for annual excess returns. For a description of HAC inference see the notes to Table 6. For a description of the bootstrap design, see text.

Table 10: Cochrane-Piazzesi: in-sample evidence

	$PC1$	$PC2$	$PC3$	$PC4$	$PC5$	Wald	R_1^2	R_2^2	$R_2^2 - R_1^2$
Original sample: 1964–2003									
Data	0.127	-2.740	6.307	16.128	-2.038		0.26	0.35	0.09
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./mean $\bar{R}^2$				2.441	2.190	8.571	0.30	0.31	0.01
Bootstrap p-value/95% CIs				0.000	0.494	0.000	(0.13, 0.50)	(0.13, 0.50)	(0.00, 0.03)
Bootstrap true size				0.097	0.078	0.116			
IM $q = 8$	0.002	0.030	0.873	0.237	0.233				
IM $q = 16$	0.000	0.004	0.148	0.953	0.283				
Later sample: 1985–2013									
Data	0.104	-1.586	-3.962	-9.196	9.983		0.14	0.17	0.03
HAC statistic	1.619	2.215	1.073	1.275	1.351	4.174			
HAC p-value	0.106	0.027	0.284	0.203	0.178	0.124			
Bootstrap 5% c.v./mean $\bar{R}^2$				2.656	2.367	11.321	0.18	0.21	0.02
Bootstrap p-value/95% CIs				0.317	0.283	0.289	(0.03, 0.41)	(0.05, 0.42)	(0.00, 0.09)
Bootstrap true size				0.140	0.113	0.175			
IM $q = 8$	0.011	0.079	0.044	0.803	0.435				
IM $q = 16$	0.001	0.031	0.215	0.190	0.949				

Inference about predictive power of principal components (PCs) of yields for excess bond returns and the null hypothesis that the first three PCs contain all the relevant predictive information in the yield curve. The dependent variable is the average one-year excess return for two- through five-year bonds. The data used in the top panel is the same as in Cochrane and Piazzesi (2005)—see in particular their table 4. For a description of HAC inference see the notes to Table 6. For a description of the bootstrap design, see text.

Table 11: Cochrane-Piazzesi: out-of-sample forecast accuracy

n	R_2^2	R_1^2	$RMSE_2$	$RMSE_1$	DM	p-value	$RMSE_{mean}$
2	0.321	0.260	2.120	1.769	2.149	0.034	1.067
3	0.341	0.242	4.102	3.232	2.167	0.032	1.946
4	0.371	0.266	5.848	4.684	2.091	0.039	2.989
5	0.346	0.270	7.374	6.075	2.121	0.036	3.987
average	0.351	0.264	4.845	3.917	2.133	0.035	2.385

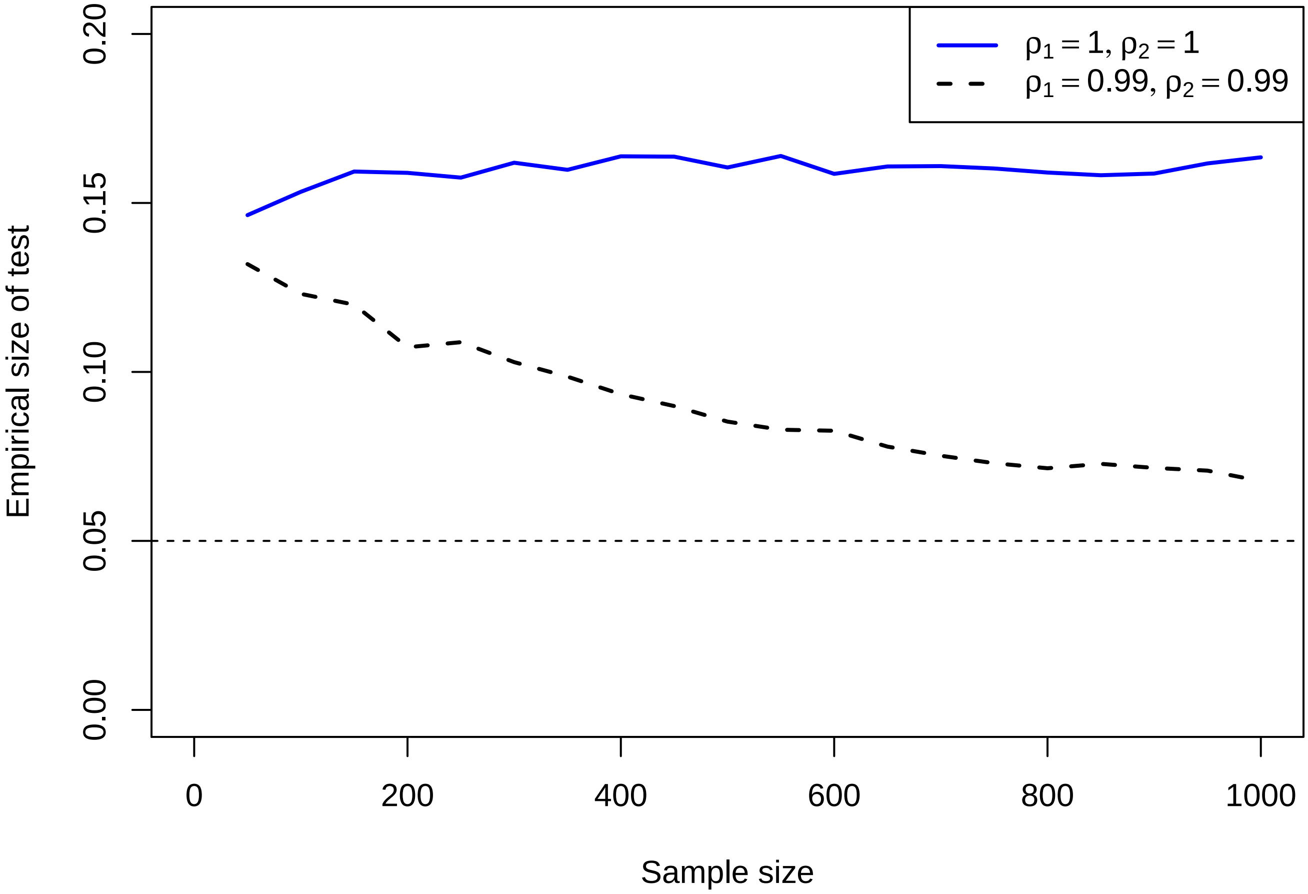
In-sample vs. out-of-sample predictive power for excess bond returns (averaged across maturities) of restricted model (1) with three PCs and unrestricted model (2) with five PCs. The in-sample period is from 1964 to 2002 (the last observation used by Cochrane-Piazzesi), and the out-of-sample period is from 2003 to 2013. The second and third column show in-sample R^2. The fourth and fifth column show root-mean-squared forecast errors (RMSEs) of the two models. The column labeled "DM" reports the z-statistic of the Diebold-Mariano test for equal forecast accuracy, and the following column the corresponding p-value. The last column shows the RMSE when forecasts are the in-sample mean excess return.

Table 12: Greenwood-Vayanos

	One-year	Term				Bond
	yield	spread	$PC1$	$PC2$	$PC3$	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						
Coefficients			-0.168	-5.842	6.089	0.013
HAC t-stat.			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 t-statistic			1.270	3.156	2.067	1.154
HAC p-value			0.204	0.002	0.039	0.249
IM $q = 8$			0.005	0.134	0.714	0.494
IM $q = 16$			0.008	0.011	0.611	0.980

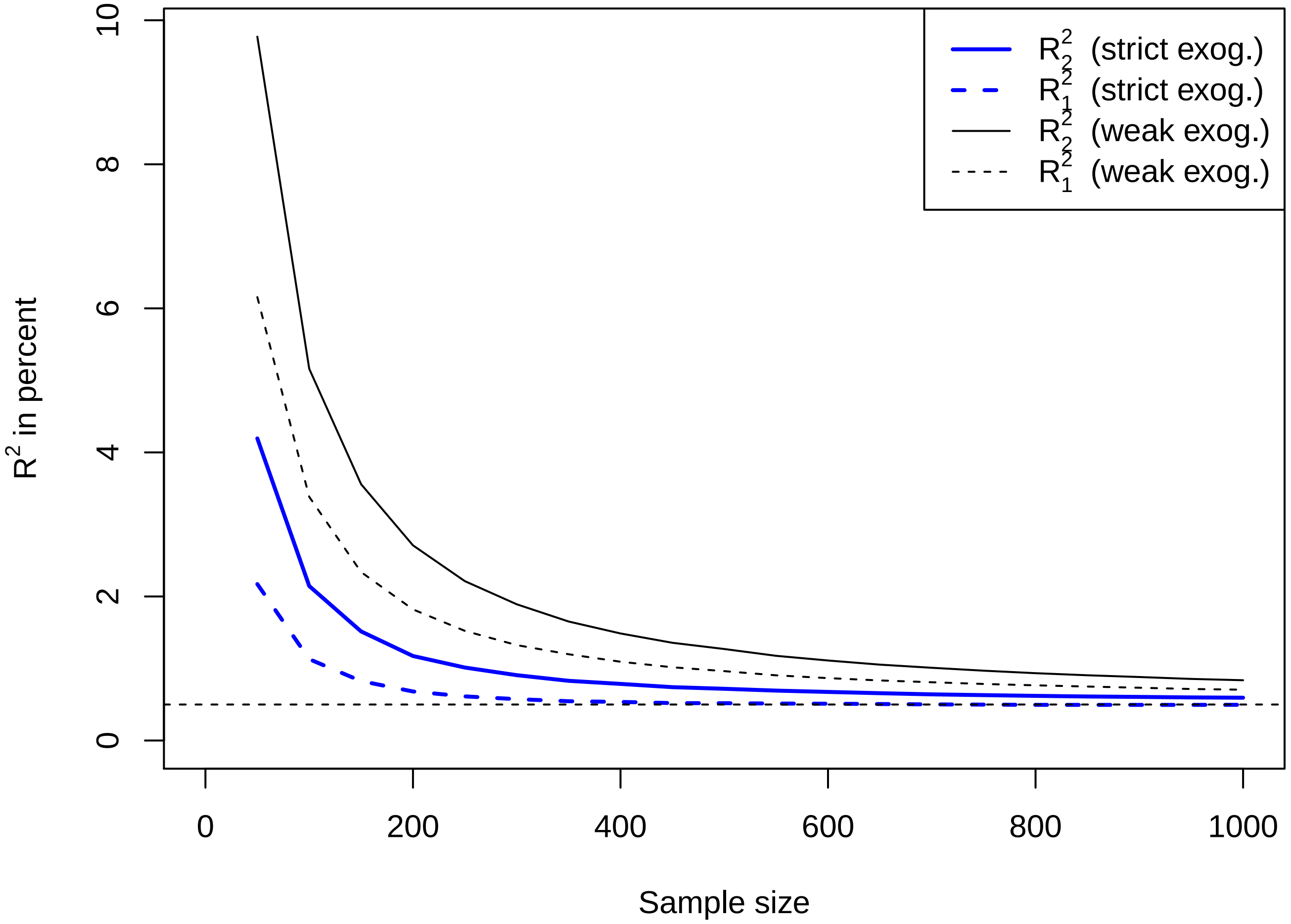
Predictive regressions for one-year bond returns using Treasury bond supply, as in Greenwood and Vayanos (2014) (GVG). The coefficients on bond supply in the first two panels are identical to those reported in row (1) and (6) of table 5 in GV. HAC t-statistics and p-values are constructed using Newey-West standard errors with 36 lags, as in GV. The sample period is 1952 to 2008.

Figure 1: Size distortions in simulation study



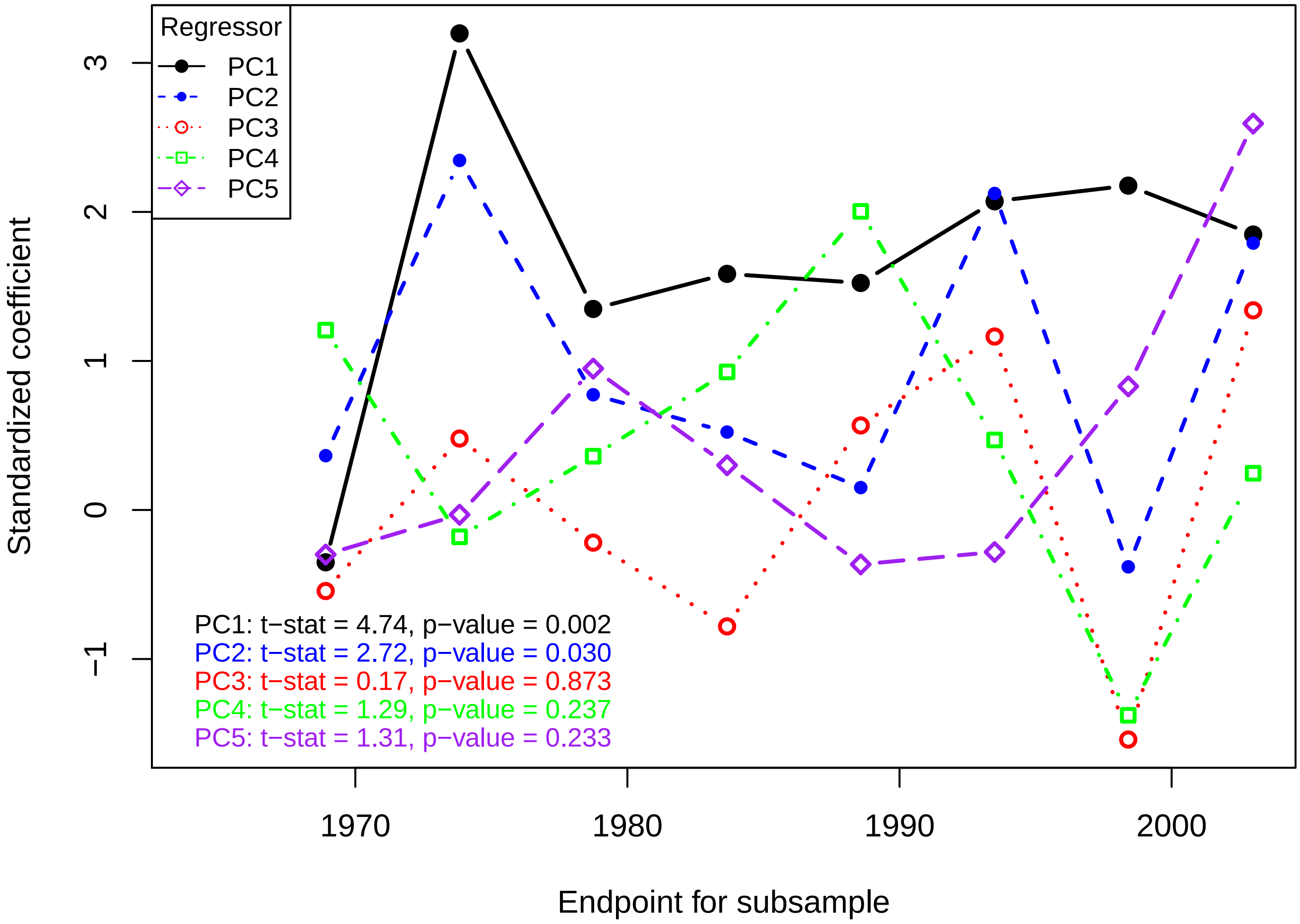
Frequency of rejections using a standard *t*-test with 5% nominal size, for different sample sizes. For details about the simulation design, please refer to main text.

Figure 2: Size distortions in simulation study



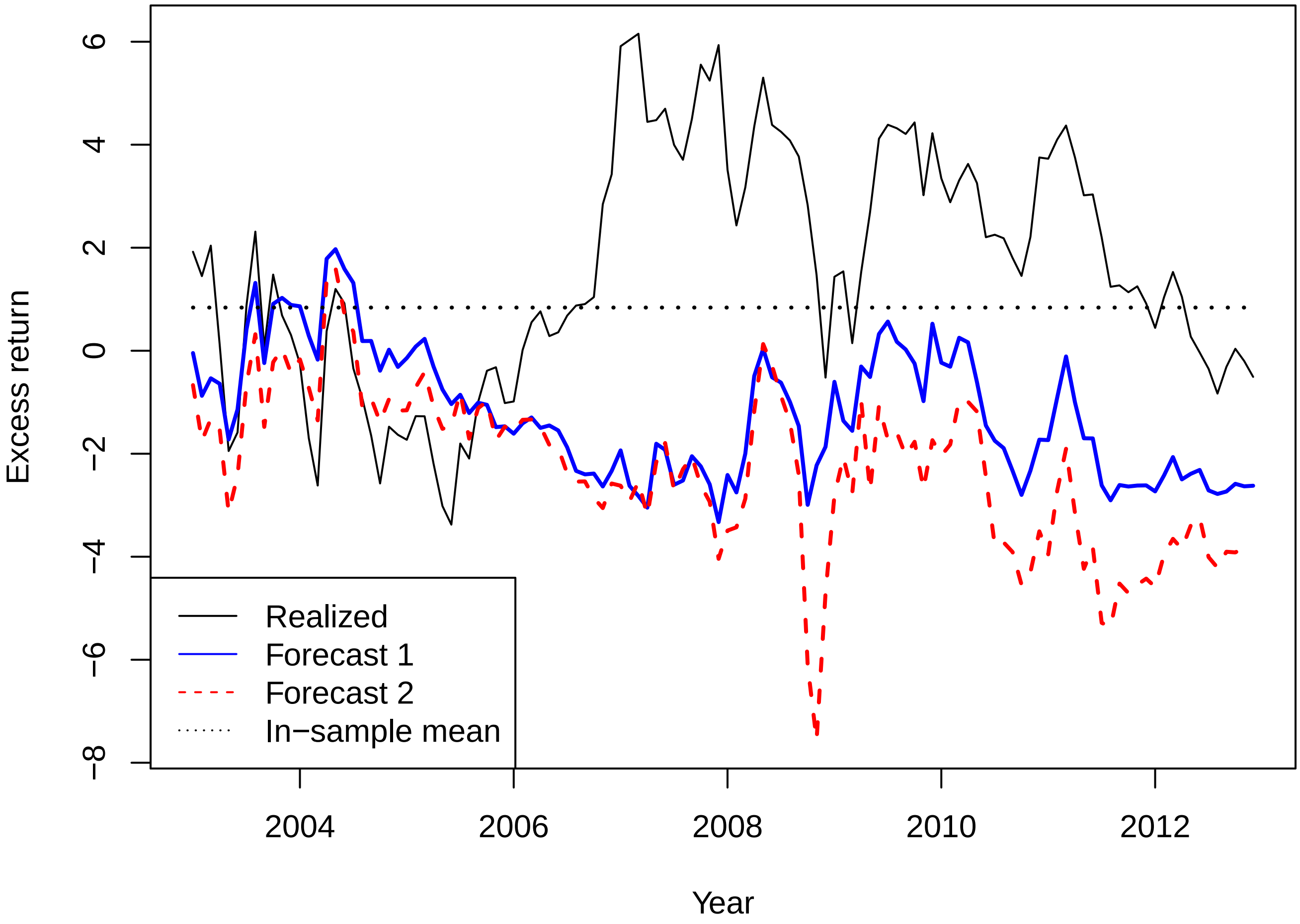
R^2 (in percent) for restricted and unrestricted predictive regressions, for simulations with strictly vs. weakly endogenous regressors x_{1t}. For details about the simulation design, please refer to main text.

Figure 3: Cochrane-Piazzesi: predictive power of PCs 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).

Figure 4: Cochrane-Piazzesi: out-of-sample forecasts



Realizations vs. out-of-sample forecasts of excess bond returns (averaged across maturities) from restricted model (1) with three PCs and unrestricted model (2) with five PCs. The in-sample period is from 1964 to 2002 (the last observation used by Cochrane-Piazzesi), and the out-of-sample period is from 2003 to 2013. The figure also shows the in-sample mean excess return.