Economics Letters 121 (2013) 546–549

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Testing the predictive power of the term structure without data snooping bias

Yi-Cheng Kao [a,*], Chung-Ming Kuan [b], Shikuan Chen [c]

[a] Department of Business Administration, Chung Yuan Christian University, Taiwan, ROC
[b] Department of Finance, National Taiwan University, Taiwan, ROC
[c] Department of International Business, National Taiwan University, Taiwan, ROC

HIGHLIGHTS

- We identify superior models of interest rates in predicting economic fluctuations.
- Models with 3-month Treasury bill interest rate have better performance.
- The best one consists of a short-term rate and a term spread.

ARTICLE INFO

Article history:
Received 1 August 2013
Received in revised form
13 October 2013
Accepted 18 October 2013
Available online 28 October 2013

JEL classification:
E43
E27
C12
C52

Keywords:
Data snooping
GDP growth
Stepwise SPA test
Term spread

ABSTRACT

It is well documented that the term structure of interest rates has predictive power for real economic growth. Applying the stepwise superior predictive ability test, we find that superior models contain both a short-term rate and a term spread.

1. Introduction

It is well documented in the literature that the term structure of interest rates contains useful information about future states of the economy. In particular, empirical studies show that the term spreads between short-term and long-term government bond rates have significant predictive power for the real GDP growth in the US. However, Ang et al. (2006) argue that using only one particular term spread may be inefficient since there is more information across the whole yield curve. On the other hand, Bordo and Haubrich (2008a,b), based on the historical evidence, suggest that the model using both the short-term rate and the term spreads is better than those using only the short-term rate on predicting real GNP growth.

Although there exist some theories regarding the predictive power of the term structure on real economic growth, economists still have no consensus on which theory indeed explains the relationship. Therefore, researchers construct various empirical models consisting of different short-term rates, long-term rates, and term spreads to evaluate the predictive power. Data snooping bias may arise when we reuse the same data set to test many models for the predictive power of the term structure. It is natural to question whether the predictive power found in the literature is real or due to chance. If the predictive power is real, then a relevant question would be: what are the best predictive models?

In this paper, we apply the Step-SPA test (Hsu et al., 2010) to identify superior models without data snooping bias. We construct

* Correspondence to: Department of Business Administration, Chung Yuan Christian University, No. 200, Chung Pei Road, Chung-Li, 32023, Taiwan, ROC. Tel.: +886 921 351 634; fax: +886 2 23626544.
E-mail addresses: kao@cycu.edu.tw, alpha_kao@yahoo.com.tw (Y.-C. Kao).

0165-1765/$ – see front matter
http://dx.doi.org/10.1016/j.econlet.2013.10.020

a set of 900 predictive models that includes various combinations of interest rates and term spreads. The null hypothesis is that none of those alternative models is superior to a benchmark. The result indicates that, for predicting the annual growth rates of real GDP, the best model consists of the short-term rate of 3-month Treasury bill and the term spread between this short rate and the yield of 5-year Treasury note.

2. The benchmark and alternative models

The following empirical model is commonly used to examine the relationship between the quarterly data of real GDP, Y_t, and the term spreads. It is the linear regression model of annualized real GDP growth rate over the next k quarters:

$$g_{k,t} = c + \gamma(LR_t - SR_t) + \varepsilon_{k,t}, \tag{1}$$

where $g_{k,t} = (400/k)(\ln Y_{t+k} - \ln Y_t)$, and LR_t and SR_t are respectively the long-term and short-term interest rates. Stock and Watson (2003) suggest that additional lagged variables should be included in (1) to improve the model performance. We thus extend (1) to construct a set of alternative models that contain various combinations of interest rates and term spreads:

$$g_{k,t} = c + \alpha(\mathbf{B})SR_t + \varepsilon_{k,t}, \tag{2}$$

$$g_{k,t} = c + \beta(\mathbf{B})LR_t + \varepsilon_{k,t}, \tag{3}$$

$$g_{k,t} = c + \gamma(\mathbf{B})(LR_t - SR_t) + \varepsilon_{k,t}, \tag{4}$$

$$g_{k,t} = c + \alpha(\mathbf{B})SR_t + \gamma(\mathbf{B})(LR_t - SR_t) + \varepsilon_{k,t}, \tag{5}$$

where $\alpha(\mathbf{B})$, $\beta(\mathbf{B})$, and $\gamma(\mathbf{B})$ are lag polynomials with j lags, so that

$$h(\mathbf{B})Z_t = h_0Z_t + h_1Z_{t-1} + h_2Z_{t-2} + \cdots + h_jZ_{t-j}.$$

This class of models includes the models with short-term rates, those with long-term rates, those with term spreads, and those with short-term rates and term spreads. To capture the variation of interest rates used in the literature, SR_t may be the effective federal funds rate (FF), the secondary market rate of 3-month Treasury bill (TB), and the yield rate of 1-year Treasury note (GR_1). For the same reason, LR_t may be the yield rates of 3, 5, and 10 year Treasury notes, denoted by GR_3, GR_5, and GR_{10}, respectively.

Since the data studied here are quarterly and the sample size is limited, we only consider models with lag polynomials of order $j = 0, 1, 2, 3, 4$. That is, as in Stock and Watson (2003), up to four lags of regressors are included in the models. Following Stock and Watson (2003) and Bordo and Haubrich (2008a,b), we also use the squared error as the performance measure in the test. Therefore, we need to specify the window length of rolling regressions, denoted by l. In Bordo and Haubrich (2008a), for example, they report the results based on the window length of 24 quarters. Here, we incorporate various models with $l = 24, 28, 32, 36, 40$ (corresponding to 6–10 years) into the alternative set. As a result, we have 4 groups of models, 5 different lengths of polynomials, 3×3 combinations of short rates and long rates, and 5 window lengths of rolling regressions. This yields a set containing 900 ($= 4 \times 5 \times 3 \times 3 \times 5$) alternative models.

As for the benchmark model in the Step-SPA test, a candidate is the AR model used in Stock and Watson (2003) and Bordo and Haubrich (2008a,b):

$$g_{k,t} = c + \delta(\mathbf{B})g_{k,t-1} + \varepsilon_{k,t}, \tag{6}$$

where $\delta(\mathbf{B})$ is a lag polynomial with $j = 4$ lags. The order of $j = 4$ is selected by the Akaike information criterion among $j = 1, 2, 3, 4$. Moreover, we need to specify the window length of rolling regressions, denoted by l^*, for the benchmark model (6). Note that each l^* characterizes a particular benchmark model. In this study, we consider 5 window lengths of $l^* = 24, 28, 32, 36, 40$ for (6). Also note that (6) is not nested in the alternative models, so that the asymptotic normality for the SPA and Step-SPA tests is valid (see Hansen, 2005, p. 367).

3. Testing method

We formally state the question as follows. Let m be the model index for the alternative models, so that we have $m = 1, 2, \ldots, 900$. In the mth alternative model with the window length of l, let $L_{m,t}(k)$ be the sequence of squared errors from rolling regressions such that

$$L_{m,t}(k) = (g_{k,t} - \hat{g}_{k,t})^2,$$

for $t = l + 1, \ldots, N - k$, where $\hat{g}_{k,t}$ is the predicted value from the model, and N and k are respectively the data length and the forecasting horizon as defined in (1). For the maximizing window length $l = 40$ and $k = 4$, the actual length of $L_{m,t}(k)$ is $N - k - l$, simply denoted by n. We thus relabel the sequence $L_{m,t}(k)$ as $t = 1, 2, 3, \ldots, n$. Similarly, let $L_{0,t}(k)$ be the sequence of squared errors from rolling regressions in the benchmark model. We further define $d_{m,t} \equiv L_{0,t}(k) - L_{m,t}(k)$, $\mathbf{d}_t \equiv (d_{1,t}, d_{2,t}, \ldots, d_{900,t})'$, and $\mu \equiv E(\mathbf{d}_t)$. Then, the null hypothesis of interest is

$$H_0 : \mu \leq 0. \tag{7}$$

That is, none of the alternative models is superior to the benchmark.

To implement the Step-SPA test, we first compute the studentized test statistic T as follows:

$$T \equiv \max\left[\max_{m=1,\ldots,900} \frac{n^{1/2}\bar{d}_m}{\hat{\omega}_m}, 0\right], \tag{8}$$

where $\bar{d}_m = n^{-1}\sum_{t=1}^{n} d_{m,t}$ and $\hat{\omega}_m^2$ is a consistent estimator of $\mathrm{var}(n^{1/2}\bar{d}_m)$. Following Hansen (2005, p. 372), we use the $\hat{\omega}_m^2$ based on Politis and Romano (1994).

Hansen (2005) suggests that the null distribution should be approximated by $N(\hat{\mu}, \hat{\Omega})$ in which the estimator for μ_m is

$$\hat{\mu}_m = \bar{d}_m \mathbf{1}\left[\frac{n^{1/2}\bar{d}_m}{\hat{\omega}_m} \leq -\sqrt{2\log\log n}\right],$$

for $m = 1, 2, \ldots, 900$, where $\mathbf{1}[A]$ is the indicator function of the event A. That is, a poor model with $\bar{d}_m$ smaller than a threshold will be re-centered in constructing the null distribution. Therefore, we have $\hat{\mu}_m = 0$ almost surely if $\mu_m = 0$, and $\hat{\mu}_m \to \mu_m$ in probability if $\mu_m < 0$. Including this re-centering process would improve the testing power; see Hansen (2005, p. 371).

We now describe how to bootstrap the null distribution and determine the p value and the critical value in the Step-SPA test. Given the parameter $q \in (0, 1]$ of the geometric distribution, the stationary bootstrap of Politis and Romano (1994) enables us to reproduce pseudo time-series sets of $\mathbf{d}_t$, denoted by $(d^*_{1,b,t}, d^*_{2,b,t}, \ldots, d^*_{900,b,t})'$ for $b = 1, 2, \ldots, B$, where B is the number of bootstraps. Defining $h^*_{m,b,t} = d^*_{m,b,t} - \hat{\mu}_m$ and $\bar{h}^*_{m,b} = n^{-1}\sum_{t=1}^{n} h^*_{m,b,t}$,