Problem 5.2
Revised Problem Statement and Data Description

Revised Problem Statement

In this exercise we study empirically whether the out-of-sample stock market return predictability of well-known valuation ratios can be improved by imposing simple theoretical restrictions on the predictive regressions. The data for this question can be found in an Excel spreadsheet on the textbook website[1] together with an accompanying explanatory document offering more details on the suggested implementation of the predictive regressions.

Consider the regression

$$R^e_{t+1} \equiv R_{t+1} - R_{f,t+1} = \alpha + \beta x_t + u_{t+1}, \tag{5.94}$$

where R_{t+1} denotes the one-quarter-ahead real return to the S&P 500 index and x_t is a predictor variable. Motivated by the claim of Welch and Goyal (2008) that the historical average excess stock return forecasts future excess stock returns out of sample better than regressions of excess returns on predictor variables, we evaluate the out-of-sample performance of forecasts based on predictor variable x_t using the out-of-sample R^2 statistic computed as

$$R^2_{OS} = 1 - \frac{\sum_{t=0}^{T-1}\left(R^e_{t+1} - \widehat{R}^e_{t+1}\right)^2}{\sum_{t=0}^{T-1}\left(R^e_{t+1} - \overline{R}^e_{t+1}\right)^2}, \tag{5.95}$$

where $\widehat{R}^e_{t+1}$ is the fitted value from regression (5.94) from the start date $-T_{IE}$ of the estimation sample through date t and $\overline{R}^e_{t+1}$ is the historical arithmetic average excess return estimated from $-T_{IE}$ through t. Here T_{IE} is the size of the initial estimation period, and T is the size of the out-of-sample forecast evaluation period. A positive value for R^2_{OS} means that the predictive regression has lower average mean-squared prediction error than the historical average excess return.

1. Calculate the in-sample R^2 statistics for the dividend yield, $x_t = D_t/P_t$, and the smoothed earnings yield, $x_t = X_t/P_t$, when regression (5.94) is estimated by standard ordinary least squares (OLS) over the full sample from 1872 to 2016.

2. Calculate the out-of-sample R^2 statistics for the two valuation ratios when regression (5.94) is estimated by standard OLS, with 1872-1926 as the initial estimation period and 1927-2016 as the out-of-sample forecast evaluation period.

   Compare the values you obtained for the in-sample and out-of-sample R^2 statistics. Are your results consistent with Welch and Goyal's (2008) claim?

[1] http://press.princeton.edu/titles/11177.html.

3. Repeat the calculations of the previous part for the out-of-sample R^2 statistics but now impose the (rather weak) theoretical restrictions that the slope β in the predictive regression and the forecast for the excess return are both nonnegative. That is, calculate the return forecast as

$$\widehat{R}^e_{t+1} = \max\left\{0, \widehat{\alpha}_{t+1} + \max\{0, \widehat{\beta}_{t+1}\}x_t\right\}, \tag{5.96}$$

where $\widehat{\alpha}_{t+1}$ and $\widehat{\beta}_{t+1}$ denote the intercept and slope estimates from the standard OLS regression, all estimated through period t.

Is there a significant improvement in the out-of-sample explanatory power of the two valuation ratios?

In the remaining parts of the exercise, we examine whether the forecasting performance of the dividend yield improves once we impose the theoretical restrictions of the drifting steady-state valuation model of section 5.5.2. Following equation (5.87), we use a version of the dividend yield adjusted for dividend growth and the real interest rate as our predictor variable:

$$x_t = \frac{D_t}{P_t}(1 + G_t) + \exp(\mathrm{E}_t[g_{t+1}]) + \frac{1}{2}\mathrm{Var}_t(r_{t+1}) - \mathrm{E}_t[1 + R_{f,t+1}], \tag{5.97}$$

where $\mathrm{E}_t[g_{t+1}]$ and $\mathrm{Var}_t(r_{t+1})$ denote market participants' conditional expectation of future log dividend growth and the conditional variance of log returns and $\mathrm{E}_t[1 + R_{f,t+1}]$ is the conditional expectation of the *real* riskfree rate.

4. Construct an estimate of (5.97) using the historical sample means of dividend growth and the real riskfree rate, and the historical sample variance of log stock returns up to date t. Even though the model assumes market participants know the value of D_{t+1} at date t, to avoid any look-ahead bias construct a real-time estimate of x_t assuming D_{t+1} is not in the econometrician's information set at date t.

   Discuss alternative procedures that you could use to construct real-time estimates of $\mathrm{E}_t[g_{t+1}]$, $\mathrm{Var}_t(r_{t+1})$ and $\mathrm{E}_t[1 + R_{f,t+1}]$.

5. Repeat the calculations of parts (b) and (c) for x_t given by (5.97). Compare the forecasting performance of this adjusted version of the dividend yield with that of the (unadjusted) dividend yield.

6. Finally, fully impose the theoretical restriction of equation (5.87) by calculating the predicted return as[2]

$$\widehat{R}^e_{t+1} = x_t, \tag{5.98}$$

   where x_t is given by (5.97). What is the out-of-sample R^2 statistic now? Discuss your conclusions from this exercise.

Note: This problem is based on Campbell and Thompson (2008).

[2]Equivalently, impose a zero intercept and a unit slope in predictive regression (5.94).

Data description

The file ‘‘Problem5.2_data.xlsx’’ (also provided in csv format) contains the data for Problem 5.2. The dataset runs from the last quarter of 1871 to the last quarter of 2016 and consists of quarterly series for the real return to the S&P 500 index (column 2), the real riskfree rate (column 3), and the aggregate real dividend (column 4), dividend yield (column 5), and smoothed earnings yield (column 6) for the S&P 500 index. The choice of initial estimation period (1872-1926) and out-of-sample forecast evaluation period (1926-2016) in the problem is guided by the fact that historical data before the CRSP era starting in 1926 are of significantly lower quality, as discussed below.

The S&P 500 real return series (Rm, column 2) from 1926Q1-2016Q4 is constructed from the monthly CRSP Index Portfolio total return for the S&P500 universe, after compounding to quarterly frequency and adjusting for inflation. The return for a quarter $t+1$ is the net simple realized value-weighted return to stocks in the index from the end of quarter t to the end of quarter $t+1$, adjusted for growth in the Consumer Price Index (CPI) over the quarter. The CPI series is taken from Shiller's monthly historical dataset on the S&P500 index. The return series before 1926Q1 is constructed using data from Shiller's dataset. In particular, the return for quarter $t+1$ is estimated using the standard return formula $(P_{t+1}+D_t/4)/P_t$ where P_t is the real index price in the last month of quarter t (an average of the daily closing prices during the month), and D_t (column 4 in this problem's dataset) is the 12-month trailing sum of real dividends paid by S&P 500 stocks.

The real riskfree rate series (Rf, column 3) from 1926Q1-2016Q4 is constructed from the CRSP risk-free rate file that is part of the CRSP US treasury database. The database contains a monthly series of the yields to maturity in the secondary market of a Treasury bill that matures approximately 90 days in the future. The nominal riskfree rate estimate (one-quarter-ahead riskfree return) for quarter $t+1$ is then constructed as the yield to maturity quoted at the end of the previous quarter t. Because there was no short-term riskfree debt before the 1920s, for the 1871-1925 period we proxy for the riskfree rate using the New York City commercial paper rates series from the NBER Macrohistory database, which consists of monthly averages of yields to 60-90 day prime commercial paper bills, after applying a credit spread discount of 88 basis points per annum (the average credit spread during 1926, the first year for which we have CRSP Treasury data). We adjust the series for inflation using realized CPI growth over the quarter to obtain an estimate of the real riskfree rate.

The real dividend series (D, column 4) is the 12-month trailing sum of aggregate dividends paid by S&P 500 stocks up to a given quarter, adjusted for inflation, and is taken directly from Shiller's dataset on the S&P500 index. The series is constructed using annual estimates on dividends paid prior to 1926, and using quarterly updates from Standard and Poor's thereafter.

The dividend yield (dp, column 5) for a given quarter is the ratio of the 12-month trailing sum of real dividends (column 4) divided by the real S&P 500 index price for the end-of-quarter month. The dividend and price series are from Shiller's dataset.

The smoothed earnings yield (xp, column 6) for a given quarter is the inverse of the cyclically-adjusted price-earnings ratio (PE10 or CAPE) for the end-of-quarter month from Shiller's dataset. The latter is computed as the real S&P price index divided by an average of real aggregate earnings over the previous 10 years. Because the CAPE series starts in 1881, the simple earnings yield (12-month trailing sum of earnings divided by the price index) is used for the 1871-1881 period, again from Shiller's dataset.