# Can Multistep Nonparametric Regressions Beat Historical Average in Predicting Excess Stock Returns?

Najrin Khanom[1]

[1] Department of Accounting and Finance, School of Business and Entrepreneurship, North Central College, USA

Correspondence: Najrin Khanom, Assistant Professor of Finance, Department of Accounting and Finance, School of Business and Entrepreneurship, North Central College, USA.

Received: June 4, 2021 Accepted: July 12, 2021 Online Published: July 28, 2021

doi:10.5430/ijfr.v12n5p71 URL: https://doi.org/10.5430/ijfr.v12n5p71

**Abstract**

Several economic and financial variables are said to have predictive power over excess stock returns. Empirically there is little consensus among academics, whether these variables have predictive power or not. Results are often sensitive to the econometric model of choice. The econometric models can produce biased results due to the high degree of persistence in predictive variables. Apart from high persistence, the relationship between stock return and the predictive variable may also be misspecified in the model. In order to address possible non-linearities and endogeneity between the residuals and persistent independent variables in predictive regressions, multi-step non-parametric and semiparametric regressions are explored in this paper. In these regressions, the conditional mean and the residuals are estimated separately and then added to obtain the predicted excess stock returns. Goyal and Welch's (2008) predictive variables are used to predict excess S&P 500 returns. The predictive performance of both in-sample and out-of-sample of the two proposed models are compared with the historical average, Ordinary Least Squares (OLS) and non-parametric regressions. The performance of the models is evaluated using Root Mean Squared Errors (RMSEs). The explored models, particularly the two-step nonparametric model, outperform the compared models in-sample. Out-of-sample several variables are found to have predictive ability.

**Keywords:** predictive regressions, autocorrelations, semiparametric, predictive bias, equity premium

## 1. Introduction

This paper explores two multi-step non-parametric and semi-parametric methods, which estimate the conditional mean and the residuals separately. Preliminary work done in this area involved using OLS regression of returns on lagged instrument variables that have predictive power over stock returns. While this is not the first attempt to apply non-parametric to predict excess stock returns, see Jin et. al, (2013), Lee et. al, (2014), and Chen & Hong (2016), the models explored in this paper have not been applied before. Prior to the late twentieth century, the consensus in the finance literature was that excess stock returns were entirely unpredictable (Fama, 1970), attributing to the efficient market hypothesis. However, towards the end of the century, numerous studies came out that believed otherwise; several variables were found to have predictive power over excess stock return. Fama and French (1988a) and Poterba and Summers (1988) find that the statistical significance of their univariate model using only past returns improves greatly when predictive variables are added to the model. Among many economic variables that are found to have predictive powers, the most notable are short term interest rates (Fama & Schwert, 1977), yield spreads (Campbell J. Y., 1987), stock market volatility (Goyal & Santa-Clara, 2003; Yin, 2019), book-to-market ratios (Ponti and Schall, 1998), price-earnings ratios (Campbell and Shiller 1988), and dividend-price ratio (Campbell and Shiller, 1988; Fama and French, 1988b; Lettau and Van Nieuwerburgh, 2008). Li and Tsiakas (2017) find excess return to be predictable out-of-sample when many of these economic variables are used in a kitchen sink regression with shrinkage.

Given the noisy nature of stock returns a sizable portion of the series tends to remain unpredictable, however, based on in-sample tests there now seems to be a consensus among the financial economists that the series does contain a significant predictable component (Campbell, 2000). Using bivariate predictive regression Goyal & Welch (2008) show that these predicting variables perform poorly, in comparison with historical average excess stock return in out-of-sample forecasts. Campbell & Thompson (2008) on the other hand, using a priori knowledge about the regression parameters, impose sign restrictions on the regression parameters; and show that many predictive variables have better out-of-sample performance than historical average return. Baltas and Karyampas (2018) attribute the sensitivity in the predictive ability to stages in the business cycle, and Tsiakas, Li and Zhang (2020) find certain variables to have predictive power during expansions and some during recessions.

Controversy surrounding the out-of-sample performance of the predictive variables cast doubt over the predictive ability of these variables. Whether the contradicting results are due to model misspecification pose even serious concern. The non-robust results of return predictability may stem from the statistical tests performed (Lamoureux & Zhou, 1996). Using a linear model when the true data generation process is non-linear may seriously undermine forecasts. Chen & Hong (2016) point out that linear model might not be appropriate to capture the movements in stock return and suggest using non-parametric regressions, which can capture the linearities and non-linearities in the data without imposing parametric restrictions. According to Chen and Hong (2016) the restrictions imposed by Campbell and Thomspon are ways of introducing non-linearity into the model, they too like the latter find predictive variables to outperform historical average in a non-parametric setting. Parametric and non-parametric forecast combination models also reach a similar conclusion (Elliott et. al, 2013; Jin et. al, 2013).

Another plausible reason for contradicting results on the out-of-sample predictive ability of variables noted as predictive variables in the literature is due to the non-stationarities in the explanatory variables. Roll (2002) argues that in the presence of rational expectation if the innovations are identically and independently distributed then the expectation about a future quantity must follow a random walk. Stock prices are based on expectations about a future quantity, and explanatory variables like dividend yield and book to market ratio are in turn functions of stock prices. Thus, these explanatory variables must also follow a random walk. Unbalanced predictive regression of stationary stock return and non-stationary dividend yield may lead one to conclude that dividend yield has no predictive power. Structural breaks might also be present in the data, for instance, Fama and French (2001) have pointed out a dramatic fall in the proportion of firms paying dividends in the late 1970s. If not careful these structural breaks might be incorrectly categorized as non-stationarity. Apart from the term spread prior to 1952 and dividend yield in the period 1926 to 1994, they find the presence of unit root in all popular predictive variables. Using international data Torous, Valkanov and Yan (2004) show that when dividend to price ratio is stationary it has predictive power and not when it is non-stationary. Torous, Valkanov and Yan (2004) find the presence of unit root in almost all commonly used predictive variables, within a 95% confidence interval. In pre-1926 and post-1994 data Torous, Valkanov, & Yan's (2004) tests indicate the presence of unit root in dividend yield and when dividend yield from those sub-periods are used to predict stock excess return, the predictive power is lost. Thus, the presence of unit root in predictive variables might explain why in certain cases they are found to have predictive power and not in other cases.

Due to the possibility of a nonlinear relationship between excess stock return and predictive variables, and nonstationarities in the predictive variables this paper explores two multi-step non-parametric and semi-parametric methods, which estimate the conditional mean and the residuals separately. The motivation is to evaluate whether such augmented non-parametric regressions can predict excess stock return in-sample and out-of-sample. The empirical performances of the proposed models in this paper are compared with the historical mean model, simple OLS model, local constant and local linear non-parametric models, on the basis of the root mean squared (forecast) errors. Analysis is performed using Goyal and Welch's (2008) original data till 2005 and using the extended data till 2019. The results should be relevant to practitioners and academics attempting similar models to predict excess stock returns and help inform their decisions to proceed.

Several methods have been explored to correct this bias. Stambaugh (1999) for instance uses the analytical expression of the bias in univariate linear, popularly known as Stambaugh's bias, and corrects the biased estimates accordingly. The analytical expression of bias derived by Stambaugh (1999) holds only when the dependent variable is stationary and under normality. Both stationarity of predictive variables and normality in error terms are strong assumptions in models of excess return (Roll, 2002). Amihud and Hurvich (2004) propose using a two-step augmented regression where the conditional mean and residuals are estimated separately using linear regression. The work proposed in this paper follows Amihud and Hurvich's (2004) two-step augmented regression, where the parametric models are replaced with non-parametric and semiparametric counterparts.

The paper is organized as follows, section 2 presents the estimation of the two multi-step nonparametric and semiparametric regressions explored, along with the other models used for comparison, section 3 shares the empirical results, and section 4 concludes.

## 2. Estimation

### *2.1 OLS*

Preliminary studies use linear regression to predict excess return using other financial variables and their lags, that tend to move with excess return, such a model is shown by (1), where $r_t$ is the excess return and $x_{t-1}$ is a lagged explanatory variable. The parameters of the simple OLS regression are estimated by (2), where the t^th row of matrix **X** and vector R are (1, $x_{t-1}$) and ($r_t$), respectively, and the predicted return, $\hat{r_t}$, OLS is given by (3).

$$r_t = \alpha + \beta x_{t-1} + u_t \quad (1)$$

$$\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix} = (X'X)^{-1}X'R \quad (2)$$

$$\hat{r}_{t,\ OLS} = \hat{\alpha} + \hat{\beta}x_{t\ -\ 1} \quad (3)$$

OLS estimates are unbiased if all the information in $x_{t-1}$ has been used to predict $r_t$. As most financial variables are highly persistent, there is information about the lags in $x_{t-1}$ that is not independent of $u_t$. For instance, if the predicting variable, $x_{t-1}$, follows an AR (1) process like (4), then E($x_{t-1}|u_t$)≠0. If $x_{t-1}$ is persistent the error terms in (1) and (4) are not independent of each other and can be expressed using (5), where $\xi \neq 0$ and $\varepsilon_t$ are *i.i.d.* errors that are independent of $v_t$ and its lags. Thus, a simple OLS with autoregressive predicting variables will result in biased estimates.

$$x_t = \varphi + \rho x_{t-1} + v_t \quad (4)$$

$$u_t = \xi v_t + \varepsilon_t \quad (5)$$

### *2.2 Historical Average (HA)*

Goyal and Welch (2008) compare the simple OLS predicted returns with the Historical Average (HA) returns shown in (6), the predicted returns are the average of the past realized returns.

$$\hat{r}_{t,\ HA} = \frac{1}{t-1}\sum_{i=1}^{t-1} r_i \quad (6)$$

### *2.3 Nonparametric (NP)*

Instead of assuming the data generation process, to be a linear model, as shown in (1), the functional form can be expressed as $m(x_{t-1})$ using a local constant non-parametric model as shown in (7).

$$r_t = m(x_{t-1}) + u_t \quad (7)$$

For a discrete random $x_{t-1}$ there are n* observations in its neighborhood, let them be $x$, m($x_{t-1}$) is the average of the $r_t$'s corresponding to the x's (Pagan & Ullah, 1999). $h$ is the window width that determines the size of the neighborhood of $x_{t-1}$ that will be used to find m($x_{t-1}$), as shown in (8).

$$\hat{m} = \left(\frac{\sum_{t=1}^{T} I(-.5 < \psi t - 1 < .5)rt}{\sum_{t=1}^{T} I(-.5 < \psi t - 1 < .5)}\right) \quad (8)$$

where $\psi_{t-1} = (x - x_{t-1})/h$. A kernel function K can be used for smoothing as illustrated in (9).

$$\hat{m} = \frac{\sum_{t=1}^{T} K(\psi t - 1)rt}{\sum_{t=1}^{T} K(\psi t - 1)} \quad (9)$$

While local constant minimizes $\sum_{t=1}^{T}[r_t - m]^2 K(\psi_{t-1})$ with respect to $m$; local linear minimizes

$$\sum_{t=1}^{T}[r_t - m - (x_{t-1} - x)\beta]^2 K(\psi_{t-1})$$

Although the nonparametric regression addresses the specification bias stemming from selecting the functional form between $r_t$ and $x_{t-1}$, it does not take into account the predictive bias stemming from highly autoregressive $x_{t-1}$. This paper explores two new multistep nonparametric and semiparametric models to address that predictive regression bias.

### *2.4 Model 1: Multistep Semiparametric Model (Multistep SP)*

In the multistep semi-parametric model, excess stock returns are predicted using a combination of linear and non-linear models. Any linear relationship between the excess stock return and the predictive variable is first captured using an OLS regression as (1). The linear prediction is then re-scaled for additional nonlinearities. Any remaining non-linearities and the endogeneity between $x_{t-1}$ and $u_t$ are then addressed by nonparametrically estimating the residuals of (1), $u_t$, using the residuals of the AR(1) process of $x_{t-1}$, $v_t$. After running the OLS regressions (1) and (4) the residuals are saved and used in a nonparametric regression as shown in (10). The estimated values of $\hat{u}_{t,SP} = m(\hat{v}_t)$ are then used to update equation (1) as illustrated in (11). The predicted excess stock returns, $\hat{r}_{t,SP}$, is a sum of the predicted excess return from the OLS model in (1) and the predicted residual from (10).

$$u_t = m(v_t) + \varepsilon_t \tag{10}$$

$$\hat{r}_{t,\,SP} = \hat{\alpha}_{OLS} + \hat{\beta}_{OLS} x_{t-1} + \hat{u}_{t,\,SP} \tag{11}$$

*2.5 Model 2: Multistep Nonparametric Model (Multistep NP)*

The multistep nonparametric model is similar to the previous model discussed, except the linear regressions (1) and (4) are replaced with nonparametric regressions. Step 1: Excess stock returns are regressed on the predictive variables using nonparametric regressions as in (12) and the residuals, $\hat{u}_{t,NP}$ are saved. Step 2: Residuals of a nonparametric AR (1) process of $x_{t-1}$, $\hat{v}_{t,NP}$, described in (13) are saved. Step 3: $\hat{u}_{t,NP}$ is regressed on $\hat{v}_{t-1,NP}$, nonparametrically as in (14). Step 4: Excess stock returns are predicted as the sum of the predicted values of (12) and (14). An across-the-board non-parametric model addresses not only any nonlinear relationship between excess stock return and the predictive variable but also any nonlinear relationship the predictive variable may have with its own past.

$$r_{t,NP} = m(x_{t-1}) + u_{t,NP} \tag{12}$$

$$\hat{u}_{t,NP} = r_t - \hat{m}(x_{t-1})$$

$$x_{t,\,NP} = m_1(x_{t-1}) + v_{t,NP} \tag{13}$$

$$\hat{v}_{t,\,NP} = x_t - \hat{m}_1(x_{t-1})$$

$$\hat{u}_{t,\,NP} = m_2(\hat{v}_{t-1,NP}) + \epsilon_{t,NP} \tag{14}$$

$$\hat{r}_{t,\,NPP} = \hat{m}(x_{t-1}) + \hat{m}_2(\hat{v}_{t-1,NP}) \tag{15}$$

$$\hat{r}_{t,NPP} = \hat{r}_{t,NP} + \hat{u}_{t,NPP}$$

In the next section, the predictive performance in-sample and out-of-sample of the two proposed models are compared with the historical average, OLS and nonparametric regressions, for the predictive variables used in Goyal and Welch (2008) and Campbell and Thompson (2008).

## 3. Empirical Results

Annual S&P 500 Index return with dividends in excess of the risk-free return are predicted using the historical average in (6), OLS regression model in (1), nonparametric regression (NP) as in (7), proposed multistep semi-parametric (Multistep SP) and nonparametric models (Multistep NP). Data is collected from Amit Goyal's website.

Table 1 presents the in-sample Root Mean Squared Error (RMSE) for the five models in predicting the S&P 500 excess return for the years 1872-2005, both Local Constant (LC) and Local Linear (LL) regressions are used for the nonparametric and semiparametric models. The start date for the samples may differ due to the availability of data of the predictive variables. Bold typeface in each row indicates the model with the lowest RMSE. Column 2 reports the start year of the sample, the end year for all samples is 2005. The one-lag autocorrelation of the independent variable, $\rho$, is presented in column 3. In all cases, the multistep semiparametric and nonparametric models perform just as well if not better than, historic average, OLS and non-parametric regression.

As can be seen from Table 1 the multistep nonparametric model has the greatest number of cases with the lowest RMSE. The historical average has higher RMSE than the nonparametric methods in all cases. It should also be noted that local linear regressions outperform their local constant counterparts in estimating excess stock returns.

Table 1. In Sample RMSE in predicting the S&P 500 excess return for the years 1872-2005

| | | | Historic | OLS | Multistep SP | | NP | | Multistep NP | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Start | ρ | | | LC | LL | LC | LL | LC | LL |
| Default Yield Spread | 1920 | 0.796 | 0.186 | 0.186 | 0.186 | 0.176 | 0.186 | 0.171 | 0.186 | **0.169** |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Inflation | 1920 | 0.578 | 0.186 | 0.186 | 0.186 | 0.186 | 0.186 | 0.186 | 0.186 | **0.186** |
| Stock Variance | 1886 | 0.687 | 0.181 | 0.181 | 0.171 | **0.165** | 0.177 | 0.176 | 0.177 | 0.172 |
| Dividend Payout | 1873 | 0.692 | 0.178 | 0.178 | 0.178 | 0.178 | 0.178 | 0.178 | 0.178 | **0.166** |
| Long Term Yield | 1920 | 0.960 | 0.186 | 0.185 | 0.184 | 0.183 | 0.186 | 0.185 | **0.167** | 0.183 |
| Term Spread | 1921 | 0.598 | 0.187 | 0.186 | 0.186 | **0.184** | 0.187 | 0.185 | 0.187 | **0.184** |
| Treasury-bill rate | 1921 | 0.891 | 0.187 | 0.185 | 0.185 | 0.185 | 0.187 | 0.185 | 0.186 | **0.185** |
| Default return Spread | 1927 | 0.339 | 0.190 | 0.188 | 0.188 | 0.188 | 0.190 | 0.188 | 0.189 | **0.188** |
| Dividend/Price | 1873 | 0.859 | 0.178 | 0.176 | 0.171 | **0.171** | 0.173 | 0.174 | 0.173 | 0.171 |
| Dividend Yield | 1873 | 0.924 | 0.178 | 0.176 | 0.175 | 0.174 | 0.174 | 0.173 | 0.172 | **0.171** |
| Long term return | 1927 | 0.080 | 0.190 | 0.188 | 0.188 | **0.183** | 0.188 | 0.188 | 0.188 | 0.183 |
| Earning price ratio | 1873 | 0.725 | 0.178 | 0.176 | 0.176 | **0.175** | 0.177 | 0.176 | 0.177 | 0.175 |
| Book to market ratio | 1922 | 0.829 | 0.187 | 0.183 | 0.162 | 0.175 | 0.185 | 0.183 | **0.160** | 0.174 |
| Investment/capital | 1948 | 0.719 | 0.159 | 0.152 | **0.152** | **0.152** | 0.154 | 0.152 | 0.154 | 0.152 |
| Net equity expansion | 1928 | 0.459 | 0.189 | 0.177 | 0.177 | **0.149** | 0.171 | 0.168 | 0.171 | 0.164 |
| Percent equity issuing | 1928 | 0.490 | 0.189 | 0.178 | 0.178 | 0.178 | 0.170 | 0.169 | 0.170 | **0.169** |
| Consumption | 1946 | 0.566 | 0.156 | 0.143 | 0.143 | 0.126 | 0.114 | 0.120 | **0.104** | 0.117 |
| Dividend yield | 1928 | 0.929 | 0.189 | 0.186 | 0.186 | 0.184 | 0.179 | 0.179 | 0.179 | **0.178** |
| Earning price ratio | 1928 | 0.783 | 0.189 | 0.184 | 0.184 | 0.174 | 0.185 | 0.176 | 0.185 | **0.166** |
| Book to market ratio | 1928 | 0.828 | 0.189 | 0.183 | **0.159** | 0.183 | 0.185 | 0.183 | 0.160 | 0.183 |

Bold typeface in each row indicates the model with the lowest RMSE when compared till 4 decimal places. Start reports the start year of the sample. ρ is the one-lag autocorrelation of the independent variable. The dependent variable is risk premium with dividends.

Table 2. In Sample RMSE in predicting the S&P 500 excess return for the years 1872-2019

| | | | Historic | OLS | Multistep SP | | NP | | Multistep NP | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Start | ρ | | | LC | LL | LC | LL | LC | LL |
| Default Yield Spread | 1920 | 0.719 | 0.184 | 0.183 | 0.183 | 0.183 | 0.184 | 0.170 | 0.184 | **0.170** |
| Inflation | 1920 | 0.566 | 0.184 | 0.184 | 0.184 | **0.147** | 0.184 | 0.178 | 0.184 | 0.147 |
| Stock Variance | 1886 | 0.644 | 0.180 | 0.180 | 0.170 | **0.158** | 0.180 | 0.180 | 0.177 | 0.179 |
| Dividend Payout | 1873 | 0.620 | 0.177 | 0.177 | 0.171 | 0.174 | 0.177 | 0.176 | 0.176 | **0.166** |
| Long Term Yield | 1920 | 0.955 | 0.184 | 0.183 | 0.183 | 0.182 | 0.180 | 0.182 | 0.180 | **0.180** |
| Term Spread | 1921 | 0.614 | 0.185 | 0.183 | 0.183 | 0.183 | 0.184 | 0.183 | 0.184 | **0.183** |
| Treasury-bill rate | 1921 | 0.896 | 0.185 | 0.183 | 0.170 | 0.170 | 0.184 | 0.183 | 0.184 | **0.170** |
| Default return Spread | 1927 | -0.290 | 0.187 | 0.187 | 0.187 | **0.184** | 0.187 | 0.187 | 0.187 | 0.184 |
| Dividend/Price | 1873 | 0.886 | 0.177 | 0.177 | 0.172 | 0.171 | 0.177 | 0.175 | 0.177 | **0.170** |
| Dividend Yield | 1873 | 0.942 | 0.177 | 0.177 | 0.175 | 0.174 | 0.174 | 0.173 | 0.172 | **0.171** |
| Long term return | 1927 | -0.153 | 0.187 | 0.186 | 0.186 | **0.180** | 0.187 | 0.186 | 0.187 | 0.180 |
| Earning price ratio | 1873 | 0.716 | 0.177 | 0.176 | 0.176 | 0.176 | 0.177 | 0.176 | 0.177 | **0.174** |
| Book to market ratio | 1922 | 0.854 | 0.185 | 0.182 | **0.162** | 0.174 | 0.184 | 0.182 | 0.165 | 0.173 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Investment/capital | 1948 | 0.725 | 0.161 | 0.153 | 0.153 | **0.152** | 0.152 | 0.153 | **0.152** | 0.152 |
| Net equity expansion | 1928 | 0.630 | 0.186 | 0.178 | 0.178 | 0.175 | 0.169 | 0.155 | 0.169 | **0.154** |
| Percent equity issuing | 1928 | 0.539 | 0.186 | 0.179 | 0.179 | 0.179 | 0.172 | 0.171 | 0.172 | **0.171** |
| Consumption | 1946 | 0.803 | 0.159 | 0.157 | 0.157 | 0.157 | 0.147 | 0.149 | 0.147 | **0.146** |
| Dividend yield | 1928 | 0.943 | 0.186 | 0.185 | 0.185 | 0.184 | 0.179 | 0.179 | **0.166** | 0.178 |
| Earning price ratio | 1928 | 0.749 | 0.186 | 0.184 | 0.184 | 0.184 | 0.185 | 0.184 | 0.185 | **0.178** |
| Book to market ratio | 1928 | 0.856 | 0.186 | 0.182 | **0.159** | 0.174 | 0.184 | 0.182 | 0.168 | 0.173 |

Bold typeface in each row indicates the model with the lowest RMSE when compared till 4 decimal places. Start reports the start year of the sample. ρ is the one-lag autocorrelation of the independent variable. The dependent variable is risk premium with dividends.

The out-of-sample Root Mean Squared Forecast Error (RMSFE) for the original data till 2005 of the aforementioned models is presented in Table 3. Rolling expanding window is used for estimation, with the first sample using 20 years of data. The estimated model is used to forecast the one year ahead excess S&P 500 return. The bold typeface indicates the model with the lowest RMSFE for respective predictive variables. The historic model outperforms the other models in the out-of-sample analysis in half of the cases. In the other half of the variables studied the predictive models were able to out predict the historical average in terms of lower forecast errors. In out-of-sample local constant regressions tend to produce lower forecast errors than corresponding local linear models. The nonparametric and semiparametric models that outperform the historical average in-sample but not in out-of-sample analysis likely suffer from overfitting. Although no model consistently outperforms the others studied, it does indicate which model is better suited based on the variable in question. It is not unusual to expect that each of these variables have unique relationships or possibly influences on stock returns, and one particular model may not be suitable for all. The last three rows present results for dividend yield, earnings price ratio and book to market ratio, for samples starting in year 1928. It can be seen that the results are also sensitive to the starting year. Earning to price ratio does not appear to have predictive ability based on the models tested when the sample starts from 1873. However, changing the start year to 1928 changed the predictive performance of the models, and all the studied models are able to outperform the historic average. Measures such as RMSFE can be swayed by extremely large forecast errors, even if they are rare.

Table 3. Out of Sample RMSFE in predicting the S&P 500 excess return for the years 1872-2005

| | | Historic | OLS | Multistep SP | | NP | | Multistep NP | |
|---|---|---|---|---|---|---|---|---|---|
| | Start | | | LC | LL | LC | LL | LC | LL |
| Default Yield Spread | 1920 | 0.1589 | 0.1597 | 0.1619 | 0.2474 | **0.1585** | 0.1636 | 0.1594 | 0.1915 |
| Inflation | 1920 | 0.1589 | 0.1603 | **0.1587** | 0.1648 | 0.1823 | 0.2325 | 0.1822 | 0.2276 |
| Stock Variance | 1886 | **0.1928** | 0.2162 | 0.2252 | 3.7626 | 0.2055 | 0.2117 | 0.2075 | 0.2196 |
| Dividend Payout | 1873 | 0.1858 | 0.1885 | 0.1896 | **0.1850** | 0.1862 | 0.1905 | 0.1891 | 0.1912 |
| Long Term Yield | 1920 | **0.1589** | 0.1637 | 0.1663 | 0.1726 | 0.1671 | 0.2109 | 0.1672 | 0.2586 |
| Term Spread | 1921 | **0.1582** | 0.1587 | 0.1587 | 0.1609 | 0.1610 | 0.1597 | 0.1613 | 0.1793 |
| Treasury-bill rate | 1921 | **0.1582** | 0.1599 | 0.1663 | 0.1696 | 0.1599 | 0.1713 | 0.1670 | 0.2428 |
| Default return Spread | 1927 | 0.1592 | **0.1588** | **0.1588** | 0.1663 | 0.1602 | 0.1681 | 0.1604 | 0.1650 |
| Dividend/Price | 1873 | **0.1858** | 0.1862 | **0.1858** | 0.1875 | 0.1871 | 0.1899 | 0.1888 | 0.1948 |
| Dividend Yield | 1873 | **0.1858** | 0.1861 | 0.1859 | 0.1896 | 0.1872 | 0.1946 | 0.1877 | 0.1941 |
| Long term return | 1927 | **0.1592** | 0.1639 | 0.1643 | 0.1700 | 0.1612 | 0.1683 | 0.1617 | 0.1700 |
| Earning price ratio | 1873 | **0.1858** | 0.1864 | 0.1909 | 0.2279 | 0.1920 | 0.1998 | 0.1913 | 0.2401 |
| Book to market ratio | 1922 | 0.1587 | 0.1593 | 0.1611 | 0.1631 | 0.1585 | 0.1595 | **0.1563** | 0.1802 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Investment capital ratio | 1948 | 0.1658 | **0.1616** | **0.1616** | 0.1625 | 0.1652 | **0.1616** | 0.1699 | 0.1672 |
| Net equity expansion | 1928 | 0.1611 | 0.1648 | 0.1648 | 2.0524 | **0.1609** | 0.3773 | 0.1610 | 0.2567 |
| Percent equity issuing | 1928 | 0.1611 | 0.1584 | 0.1584 | 0.1591 | 0.1580 | 0.1581 | 0.1587 | **0.1576** |
| Consumption | 1946 | 0.1608 | 0.1454 | 0.1496 | 0.1585 | 0.1516 | **0.1360** | 0.1544 | 0.1366 |
| Dividend yield | 1928 | **0.1611** | 0.1705 | 0.1701 | 0.1795 | 0.1648 | 0.1776 | 0.1647 | 0.1875 |
| Earning price ratio | 1928 | 0.1611 | 0.1578 | 0.1584 | **0.1564** | 0.1604 | 0.1588 | 0.1600 | 0.1593 |
| Book to market ratio | 1928 | **0.1611** | 0.1743 | 0.1765 | 0.1781 | 0.1637 | 0.1911 | 0.1631 | 0.1821 |

Bold typeface in each row indicates the model with the lowest RMSFE. Start reports the start year of the sample when compared till 5 decimal places. Expanding window is used for estimation, with 20 years bands. The dependent variable is risk premium with dividends

Out-of-sample analysis extended till 2019 are presented in Table 4. In the extended data, the gains from non-parametric and semiparametric are reduced and historical average tends to dominate in most variables. However, dividend yield spread, book to market ratio, investment capital and percent of equity issuing continue to show predictive powers in the extended data. Local linear models tend to do better in-sample compared to local constant, whereas out of sample local constant produces lower forecast errors.

Table 4. Out of Sample RMSFE in predicting the S&P 500 excess return for the years 1872-2019

| | | Historic | OLS | Multistep SP | | NP | | Multistep NP | |
|---|---|---|---|---|---|---|---|---|---|
| | Start | | | LC | LL | LC | LL | LC | LL |
| Default Yield Spread | 1920 | 0.1611 | 0.1615 | 0.1630 | 0.2318 | **0.1608** | 0.1650 | 0.1615 | 0.1881 |
| Inflation | 1920 | **0.1611** | 0.1622 | 0.1615 | 0.1628 | 0.1704 | 0.1794 | 0.1703 | 0.1865 |
| Stock Variance | 1886 | **0.1905** | 0.2103 | 0.2202 | 0.2537 | 0.2094 | 0.2791 | 0.2125 | 0.3137 |
| Dividend Payout | 1873 | **0.1845** | 0.1870 | 0.1878 | 0.1858 | 0.1848 | 0.1890 | 0.1873 | 0.1884 |
| Long Term Yield | 1920 | **0.1611** | 0.1648 | 0.1671 | 0.1885 | 0.1666 | 0.2080 | 0.1671 | 0.2506 |
| Term Spread | 1921 | 0.1605 | **0.1603** | **0.1603** | 0.1627 | 0.1629 | 0.1612 | 0.1629 | 0.1816 |
| Treasury-bill rate | 1921 | **0.1605** | 0.1611 | 0.1656 | 0.1697 | 0.1617 | 0.1673 | 0.1681 | 0.3756 |
| Default return spread | 1927 | **0.1616** | 0.1643 | 0.1643 | 0.1728 | 0.1628 | 0.1743 | 0.1635 | 0.1763 |
| Dividend/Price | 1873 | **0.1845** | 0.1854 | 0.1856 | 0.1871 | 0.1864 | 0.1897 | 0.1880 | 0.1941 |
| Dividend Yield | 1873 | **0.1845** | 0.1858 | 0.1895 | 0.1889 | 0.1866 | 0.1935 | 0.1868 | 0.1925 |
| Long term return | 1927 | **0.1616** | 0.1665 | 0.1665 | 0.1730 | 0.1639 | 0.1701 | 0.1642 | 0.1737 |
| Earning price ratio | 1873 | **0.1845** | 0.1860 | 0.1888 | 0.2231 | 0.1905 | 0.1979 | 0.1898 | 0.2340 |
| Book to market ratio | 1922 | 0.1610 | 0.1627 | 0.1672 | 0.1647 | 0.1611 | 0.1629 | **0.1595** | 0.1797 |
| Investment capital ratio | 1948 | 0.1680 | 0.1610 | 0.1610 | 0.1640 | 0.1644 | **0.1606** | 0.1674 | 0.1721 |
| Net equity expansion | 1928 | **0.1632** | 0.1715 | 0.1769 | 0.4457 | 0.1687 | 0.1852 | 0.1690 | 0.1844 |
| Percent equity issuing | 1928 | 0.1632 | 0.1639 | 0.1639 | 0.1647 | **0.1624** | 0.1637 | 0.1632 | 0.1655 |
| Consumption | 1946 | **0.1643** | 0.1651 | 0.1653 | 0.1650 | 0.1676 | 0.1650 | 0.1675 | 0.1662 |
| Dividend yield | 1928 | **0.1632** | 0.1727 | 0.1728 | 0.1803 | 0.1684 | 0.1785 | 0.1689 | 0.1843 |
| Earning price ratio | 1928 | **0.1632** | 0.1638 | 0.1644 | 0.1875 | 0.1645 | 0.1852 | 0.1650 | 0.1924 |
| Book to market ratio | 1928 | **0.1632** | 0.1758 | 0.1758 | 0.1791 | 0.1662 | 0.1873 | 0.1653 | 0.1791 |

Bold typeface in each row indicates the model with the lowest RMSFE when compared till 5 decimal places. Start reports the start year of the sample. Expanding window is used for estimation, with 20 years bands. The dependent variable is risk premium with dividends.

## 4. Conclusion

Predictability of stock return is an elusive subject, and whether certain variables have predictive power over stock return has yet to cease the interest of many academics and practitioners. The presence of high autocorrelation in the predictive variables and possible non-linearities in their relationship with stock return further complicates the matter. In order to address the possible non-linearity and endogeneity between the residuals due to the persistent independent variables in the predictive regression, multistep semiparametric and non-parametric methods are explored, where the conditional mean and the residuals are estimated separately and added to obtain the predicted excess stock return. Using Goyal and Welch's (2008) predictive variables, the proposed models particularly the multistep nonparametric model produce better estimates of the excess S&P 500 return in-sample than the historical average and OLS regression. Out-of-sample the results are mixed, while in many variables the historical average dominates in terms of producing lower forecast errors, there are several variables, that are able to better predict the stock excess returns than the historical average. Future research in this area can focus on studying individual variables and their relationship with excess stock returns to find the most suitable forecasting model. Different estimation and forecast windows may also provide forecasting opportunities. In order to reduce overfitting often encountered in non-parametric regression, possible regularization parameters can be explored.

## References

Amihud, Y., & Hurvich, C. M. (2004). Predictive regressions: A reduced-bias estimation method. *Journal of Financial and Quantitative Analysis*, *39*(4), 813-841. https://doi.org/10.1017/S0022109000003227

Baltas, N., & Karyampas, D. (2018). Forecasting the equity risk premium: The importance of regime-dependent evaluation. *Journal of Financial Markets*, *38*, 83-102. https://doi.org/10.1016/j.finmar.2017.11.002

Campbell, J. Y. (1987). Stock returns and the term structure. *Journal of Financial Economics*, *18*(2), 373-399. https://doi.org/10.1016/0304-405X(87)90045-6

Campbell, J. Y. (2000). Asset pricing at the millennium. *Journal of Finance*, *55*(4), 1515-1567. https://doi.org/10.1111/0022-1082.00260

Campbell, J. Y., & Shiller, R. J. (1988). The dividend-price ratio and expectations of future dividends. *The Review of Financial Studies*, *1*(3), 195-228. https://doi.org/10.1093/rfs/1.3.195

Campbell, J. Y., & Thompson, S. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average?. *Review of Financial Studies*, *21*(4), 1509-1531. https://doi.org/10.1093/rfs/hhm055

Chen, Q., & Hong, Y. (2016). *Predictability of equity returns over different time horizons: A nonparametric approach*. Available at SSRN 3390982. https://doi.org/10.2139/ssrn.3390982

Elliott, G., Gargano, A., & Timmermann, A. (2013). Complete subset regressions. *Journal of Econometrics*, *177*(2), 357-373. https://doi.org/10.1016/j.jeconom.2013.04.017

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, *25*(2), 383-417. https://doi.org/10.1111/j.1540-6261.1970.tb00518.x

Fama, E. F., & French, K. R. (1988a). Permanent and temporary components of stock prices. *Journal of Political Economy*, *96*(2), 246-273. https://doi.org/10.1086/261535

Fama, E. F., & French, K. R. (1988b). Dividend yields and expected stock returns. *Journal of Financial Economics*, *22*(1), 3-25. https://doi.org/10.1016/0304-405X(88)90020-7

Fama, E. F., & French, K. R. (2001). Disappearing dividends: Changing firm characteristics or lower propensity to pay?. *Journal of Financial Economics*, *60*(1), 3-43. https://doi.org/10.1016/S0304-405X(01)00038-1

Fama, E. F., & Schwert, G. W. (1977). Asset returns and inflation. *Journal of Financial Economics*, *5*, 115-146. https://doi.org/10.1016/0304-405X(77)90014-9

Goyal, A., & Santa-Clara, P. (2003). Idiosyncratic risk matters!. *The Journal of Finance*, *58*(3), 975-1007. https://doi.org/10.1111/1540-6261.00555

Goyal, A., & Welch, I. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, *21*(4), 1455-1508. https://doi.org/10.1093/rfs/hhm014

Jin, S., Su, L., & Ullah, A. (2013). Robustify financial time series forecasting with bagging. *Econometric Reviews, 33*(5-6), 575-605. https://doi.org/10.1080/07474938.2013.825142

Lamoureux, C., & Zhou, G. (1996). Temporary components of stock returns: what do the data tell us?. *Review of Financial Studies, 9*(4), 1033-1059. https://doi.org/10.1093/rfs/9.4.1033

Lee, T., Tu, Y., & Ullah, A. (2014). Nonparametric and semiparametric regressions subject to monotonicity constraints: Estimation and forecasting. *Journal of Econometrics, 182*(1), 196-210. https://doi.org/10.1016/j.jeconom.2014.04.018

Lettau, M., & Van Nieuwerburgh, S. (2008). Reconciling the return predictability evidence. *The Review of Financial Studies*, *21*(4), 1607-1652. https://doi.org/10.1093/rfs/hhm074

Li, J., & Tsiakas, I. (2017). Equity premium prediction: The role of economic and statistical constraints. *Journal of Financial Markets, 36*, 56-75. https://doi.org/10.1016/j.finmar.2016.09.001

Pagan, A., & Ullah, A. (1999). *Non-parametric econometrics*. Cambridge: Cambridge University Press.

Ponti, J., & Schall, L. D. (1998). Book-to-market ratios as predictors of market returns. *Journal of Financial Economics, 49*(2), 141-160. https://doi.org/10.1016/S0304-405X(98)00020-8

Poterba, J. M., & Summers, L. H. (1988). Mean reversion in stock returns: Evidence and implications. *Journal of Financial Economics*, *22*(1), 27-59. https://doi.org/10.1016/0304-405X(88)90021-9

Roll, R. (2002). Rational infinitely-lived asset prices must be non-stationary. *Journal of Banking and Finance*, *26*(6), 1093-1097. https://doi.org/10.1016/S0378-4266(02)00207-8

Stambaugh, R. (1999). Predictive regressions. *Journal of Financial Economics*, *54*(3), 375-421. https://doi.org/10.1016/S0304-405X(99)00041-0

Stone, C. J. (1977). Consistent non-parametric regression. *Annuals of Statistics*, *5*(4), 595-645. https://doi.org/10.1214/aos/1176343886

Torous, W., Valkanov, R., & Yan, S. (2004). On predicting stock returns with nearly integrated explanatory variables. *The Journal of Business*, *77*(4), 937-966. https://doi.org/10.1086/422634

Tsiakas, I., Li, J., & Zhang, H. (2020). Equity premium prediction and the state of the economy. *Journal of Empirical Finance*, *58*, 75-95. https://doi.org/10.1016/j.jempfin.2020.05.004

Yin, A. (2019). Out-of-sample equity premium prediction in the presence of structural breaks. *International Review of Financial Analysis*, *65*, 101385. https://doi.org/10.1016/j.irfa.2019.101385

## Copyrights