Noname manuscript No.
(will be inserted by the editor)

Forecasting benchmarks of long-term stock returns via machine learning

Ioannis Kyriakou · Parastoo Mousavi · Jens Perch Nielsen · Michael Scholz

Abstract Recent advances in pension product development seem to favour alternatives to the risk free asset often used in the financial theory as a performance standard for measuring the value generated by an investment or a reference point for determining the value of a financial instrument. To this end, in this paper, we apply the simplest machine learning technique, namely, a fully nonparametric smoother with the covariates and the smoothing parameter chosen by cross-validation to forecast stock returns in excess of different benchmarks, including the short-term interest rate, long-term interest rate, earnings-by-price ratio, and the inflation. We find that, net-of-inflation, the combined earnings-by-price and long-short rate spread form our best-performing two-dimensional set of predictors for future annual stock returns. This is a crucial conclusion for actuarial applications that aim to provide real-income forecasts for pensioners.

Keywords Benchmark · Cross-validation · Prediction · Stock returns

Ioannis Kyriakou
Faculty of Actuarial Science and Insurance, Cass Business School, City, University of London, London, United Kingdom
Tel.: +44 20 7040 8738
E-mail: ioannis.kyriakou@city.ac.uk

Parastoo Mousavi
Faculty of Actuarial Science and Insurance, Cass Business School, City, University of London, London, United Kingdom
E-mail: Parastoo.Mousavi.2@cass.city.ac.uk

Jens Perch Nielsen
Faculty of Actuarial Science and Insurance, Cass Business School, City, University of London, London, United Kingdom
Tel.: +44 20 7040 0990
E-mail: jens.nielsen.1@city.ac.uk

Michael Scholz
Department of Economics, University of Graz, Graz, Austria
Tel.: +43 316 380 7112
E-mail: michael.scholz@uni-graz.at

1 Introduction

One of the key messages of Merton (2014) is that pension forecasts must be in real terms. Perhaps, the simplest way of accommodating this challenge is to change the benchmark, i.e., the monetary unit everything is measured in terms of, to inflation rather than the risk free interest rate often used. In three recent pension product development papers from the sponsored project by the Institute and Faculty of Actuaries (IFoA) "Minimizing longevity and investment risk while optimizing future pension plans", Donnelly et al. (2018) and Gerrard et al. (2018a, 2018b) suggest changing the classical benchmark, i.e., the risk free asset, to inflation.

The previous contributions did not consider the econometric challenges of using different benchmarks. The purpose of the current research is to make the first few investigations on suitable benchmark selection from an econometric perspective. We achieve this by machine learning based on the cross-validated time series approach of Nielsen and Sperlich (2003), Scholz et al. (2015) and Scholz et al. (2016) to optimize the fully nonparametric statistical estimation and forecasting of the risky asset returns in excess of four different benchmarks: the risk free rate, the long-term interest rate, the earnings-by-price ratio, and the inflation. Our method lets the data speak in themselves via training and learning, while being intuitively informative so that we can identify the covariates driving the system.

The paper benefits by both a theoretical contribution, that is, a study of the convergence properties of the local-linear smoother we use to solve the regression problem, but also an important empirical contribution that follows from the application. In particular, we assess the performance of the different benchmarks in terms of forecasting next year's excess returns given prominent covariates from the literature, such as dividend-by-price ratio, earnings-by-price ratio, short interest rate, long interest rate, the term spread, the inflation, as well as this year's lagged excess stock return. We apply single benchmarking, where only the stock returns are adjusted according to the benchmark, or full benchmarking with additionally transformed covariates using the same benchmark. In summary, our investigations show that the latter approach uncovers the predictability of earnings which, when combined with the long-short spread, in real terms result in optimal forecasts with a predictive power of at least 18%. This is important for long-term saving strategies, where one is interest in real value, corroborating the change of the classical risk free asset benchmark to inflation, as suggested in the abovementioned researches.

The remaining of this paper is organized as follows. In Section 2, we provide our definition of machine learning and adapt to our context of long-term stock return prediction. In Section 3, we present our underlying financial model, the adopted local-linear smoother and its theoretical properties. In Section 4, we present our validation criterion for the model selection. We then provide in Section 5 a description of our dataset and exhibit our empirical findings from different validated scenarios: we study in Section 5.2 a single benchmarking approach with the dependent variable measured on the original nominal scale

and extend in Section 5.3 to the case of both the independent and dependent variables adjusted according to the benchmark (full benchmarking approach). Section 5.4 summarizes the key points of our analysis and Section 6 concludes the paper.

2 Machine learning and prediction of long-term stock returns

We define machine learning as a way of working that involves the following key processes. First, the problem, the audience and the potential client must be articulated. Second, machine learners must have domain knowledge. This is what distinguishes them from applied statisticians. Machine learners not only know the data very well, but, in fact, more than that. For example, if it is a business question that needs to be solved, then the machine learner knows the business from experience; if it is a medical question, then the machine learner is someone who has worked closely with the relevant medical doctors and other health personnel for long before conducting the analysis. Therefore, the machine learner is in a position to ask for extra data or even, perhaps, manipulate the data having a deep insight knowledge of the business, medical, or any other question. Third, new techniques must be qualified against earlier ones via validation, which should normally be the final selection criterion. Finally, prior knowledge has to be channeled to the statistical model used for validation, which then has to be conducted consistently with correct underlying statistical principles.

Our study follows the aforementioned key processes closely. More specifically, our audience is the entire community of pension savers wishing for more meaningful and better communicated pension products. The IFoA, that is, the biggest actuarial organization with 33,000 members globally, is our client and sponsor of the work. Our domain knowledge on how to conduct machine learning on yearly stock data derives from more than 20 years of practical and academic work in the area of pensions. Also, we adhere to the principle of outcome selection via validation; this is our only criterion, besides common sense, when selecting our preferred models for forecasting stock returns under different benchmarks. Our domain knowledge of the pension industry and the theory of pension research have given us the insight to consider different benchmarks than just the short interest rate benchmark. In fact, the inflation benchmark might fit better in what our audience and client look for as the goal of investing is to increase wealth, or purchasing power. In addition, investors aim to anticipate the factors that impact portfolio performance and make decisions based on their expectations; inflation is one of those factors that affects a portfolio. However, inflation's varying impact on stocks confuses the decision to trade positions already held or to take new positions and, thus, taking out inflation could give a clearer picture.

In this paper, we apply the simplest machine learning technique, namely, a fully nonparametric smoother with the covariates and the smoothing parameter chosen by cross-validation. The result is a full-blown machine learning

algorithm which lets the data speak in themselves via training and learning, while being intuitively informative so that we can identify the covariates driving the system.

3 The underlying financial model

In this section, we focus on nonlinear relationships between stock returns in excess of a reference rate or benchmark, Y, and a set of explanatory variables, X. We aim to investigate different benchmark models and their predictability.

We consider a battery of benchmarks including the short-term interest rate, the long-term interest rate, the earnings-by-price ratio, and the inflation. More specifically, we investigate stock returns $S_t = (P_t + D_t)/P_{t-1}$, where D_t denotes the (nominal) dividends paid during year t and P_t the (nominal) stock price at the end of year t, in excess (log-scale) of a given benchmark $B_{t-1}^{(A)}$:

$$Y_t^{(A)} = \ln \frac{S_t}{B_{t-1}^{(A)}},$$

where $A \in \{R, L, E, C\}$ with, respectively,

$$B_t^{(R)} = 1 + \frac{R_t}{100}, \quad B_t^{(L)} = 1 + \frac{L_t}{100}, \quad B_t^{(E)} = 1 + \frac{E_t}{P_t}, \quad B_t^{(C)} = \frac{CPI_t}{CPI_{t-1}},$$

R_t is the short-term interest rate, L_t the long-term interest rate, E_t the earnings accruing to the index in year t, and CPI_t the consumer price index for year t. The predictive nonparametric regression model is

$$Y_t^{(A)} = m(X_{t-1}) + \xi_t, \tag{1}$$

where

$$m(x) = \mathbb{E}(Y^{(A)}|X = x),\ x \in \mathbb{R}^q, \tag{2}$$

is an unknown smooth function and ξ_t is a martingale difference process, i.e., serially uncorrelated zero-mean random error terms, given the past, of an unknown conditionally heteroscedastic form $\sigma(x)$.

Our aim is to forecast the excess stock returns $Y_t^{(A)}$ using popular lagged predictive variables X_{t-1} including the: i) dividend-by-price ratio $d_{t-1} = D_{t-1}/P_{t-1}$; ii) earnings-by-price ratio $e_{t-1} = E_{t-1}/P_{t-1}$; iii) short-term interest rate $r_{t-1} = R_{t-1}/100$; iv) long-term interest rate $l_{t-1} = L_{t-1}/100$; v) inflation $\pi_{t-1} = (CPI_{t-1} - CPI_{t-2})/CPI_{t-2}$; vi) term spread $s_{t-1} = l_{t-1} - r_{t-1}$; and vii) excess stock return $Y_{t-1}^{(A)}$. Other popular explanatory variables could be the consumption, wealth, income ratio (cay), or the book-to-market ratio (b/m), which have been used in predictive regressions, as, for example, in Welch and Goyal (2008). Currently, we consider only the aforementioned variables due to data restrictions (see Section 5.1).

In the next section, we address the regression problem of estimating the conditional mean function (2). We present consistency results and asymptotic normality for the local-linear (LL) smoother which we then implement in Section 5.

3.1 The local-linear smoother

Consider a sample of real random variables $\{(X_t, Y_t), t = 1, \ldots, n\}$ which are strictly stationary and weakly dependent. To measure the strength of dependence in the time series, we limit ourselves to the *strong-* or α-*mixing*[1] defined, for example, in Doukhan (1994), where

$$\alpha_\tau = \sup_{t \in \mathbb{N}} \sup_{A \in \mathcal{F}_{t+\tau}^{\infty}, B \in \mathcal{F}_{-\infty}^{t}} |P(A \cap B) - P(A)P(B)|,$$

$\mathcal{F}_i^j$ denotes the σ-algebra generated by $\{X_k, i \leq k \leq j\}$. In addition, α_τ approaches zero as $\tau \to \infty$. Note that weakly dependent data rule out, for example, processes with long-range dependence and nonstationary processes with unit-roots. We further assume that the sequence $\{(X_t, Y_t), t = 1, \ldots, n\}$ is *algebraic* α-*mixing*, i.e., $\alpha = O(\tau^{-(1+\epsilon)})$ for some $\epsilon > 0$.

Consider now the prediction problem (1)–(2). A common estimator for $m(x)$ is the Nadaraya–Watson (NW) estimator (local-constant kernel method) given by

$$\hat{m}_{NW}(x) = \frac{\hat{p}(x)}{\hat{f}(x)}, \tag{3}$$

where the probability density function of X_t, $f(x)$, is estimated for a given fixed value of $x = (x_1, \ldots, x_q)' \in \mathbb{R}^q$ by

$$\hat{f}(x) = \frac{1}{n} \sum_{t=1}^{n} K_h(X_t - x)$$

and

$$\hat{p}(x) = \frac{1}{n} \sum_{t=1}^{n} Y_t K_h(X_t - x).$$

K_h denotes some kernel function, for example, the product kernel

$$K_h(X_t - x) = \prod_{s=1}^{q} \frac{1}{h_s} k\left(\frac{X_{ts} - x_s}{h_s}\right),$$

which depends on a set of bandwidths $(h_1, \ldots, h_q)$ and higher-order kernels k (the order $\nu > 0$ of the kernel is defined as the order of the first nonzero moment), i.e., univariate symmetric functions satisfying $\int k(u)du = 1$, $\int u^l k(u)du = 0$ $(l = 1, \ldots, \nu - 1)$, and $\int u^\nu k(u)du =: \kappa_\nu > 0$. X_{ts} denotes the sth component of X_t $(s = 1, \ldots, q)$.

Under the standard assumptions of serial dependence with a required rate α as stated above, bounded density $f(x)$, controlled tail behaviour of conditional expectations, $h_s \to 0$ $(s = 1, \ldots, q)$ and $nH_q = nh_1 \ldots h_q \to \infty$ as $n \to \infty$, Li and Racine (2007), for example, show the following result of pointwise convergence.

[1] First proofs of consistency and asymptotic normality for nonparametric regression were introduced in the statistical literature in a seminal paper on kernel estimation with strong mixing data by Robinson (1983).

Theorem 1 *Under the given assumptions,*

$$|\hat{m}_{NW}(x) - m(x)| = O_p\left(\sum_{s=1}^{q} h_s^2 + \frac{1}{\sqrt{nH_q}}\right).$$

Several generalizations of Theorem 1 have been proposed in the literature. For example, Hansen (2006) proves the uniform and almost sure convergence of the NW estimator, while Scholz et al. (2016) show the quasi-complete convergence of the estimator in the case of generated regressors and weakly dependent data. Li and Racine (2007) further show the asymptotic normality of the estimator by calculating the bias term $B_s(x) = \frac{\kappa_2}{2}\left(f(x)m_{ss}(x) + 2f_s(x)m_s(x)\right)/f(x)$, where subscripts s and ss denote, respectively, first and second order derivatives, and $\kappa_2 = \int u^2 k(u)du$.

Theorem 2 *Under the given assumptions,*

$$\sqrt{nH_q}\left(\hat{m}_{NW}(x) - m(x) - \sum_{s=1}^{q} h_s^2 B_s(x)\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\kappa^q \sigma^2(x)}{f(x)}\right),$$

where $\kappa = \int k^2(u)du$.

The extension to the LL estimator $\hat{m}_{LL}(x)$ is almost straightforward. For notational convenience, we focus on the case $q = 1$. Then, upon defining

$$s_j(x) = \sum_{t=1}^{n} K_h(X_t - x)(X_t - x)^j,$$
$$t_j(x) = \sum_{t=1}^{n} Y_t K_h(X_t - x)(X_t - x)^j$$

for $j = 0, 1, 2$, we get

$$\hat{m}_{LL}(x) := \frac{t_0(x)s_2(x) - t_1(x)s_1(x)}{s_0(x)s_2(x) - s_1^2(x)} = \frac{\sum_{t=1}^{n} Y_t C_h(X_t - x)}{\sum_{t=1}^{n} C_h(X_t - x)} \tag{4}$$

with the kernel function

$$C_h(X_t - x) = \frac{1}{nh}\sum_{s \neq t} K_h(X_t - x)(X_s - X_t)K_h(X_s - x)(X_s - x)$$

representing a discretized version of $C(u) := \int K(u)(v-u)K(v)vdv$. Note that (4) is of the same form as (3) and that the kernel C has similar properties to K. Applying Theorem 1 yields the pointwise convergence result for the LL estimator.

Theorem 3 *Under the given assumptions,*

$$|\hat{m}_{LL}(x) - m(x)| = O_p\left(h^2 + \frac{1}{\sqrt{nh}}\right).$$

For a multivariate extension and asymptotic normality, refer, for example, to Masry (1996).

As many time series exhibit a nonstationary behaviour, the focus of the statistical literature is broadened in recent years to the so-called locally stationary processes (Dahlhaus, 1997). Processes are locally stationary whenever it is possible to approximate the behaviour of the process over short periods of time in a stationary way. For example, Vogt (2012) studies nonparametric models allowing for locally stationary regressors and a regression function that changes over time. He develops asymptotic theory for the NW estimator which has rescaled time as one covariate. Vogt (2012) states that his convergence result is not valid in a forecasting context. However, Cheng et al. (2018) provide predictive models and estimation theory for the local-constant case and locally stationary regressors. They apply their methods to monthly stock market data and find improved predictability of their models compared to traditional linear predictive regression models. We do not apply a similar strategy to our annual data because i) using an additional regressor for rescaled time increases the dimensionality of our problem and it is not clear whether this is beneficial in our scarce data environment (curse of dimensionality); (ii) most of our regressors do not seem to be highly persistent on an annual basis.

4 The principle of validation: model selection and the choice of smoothing parameter

As we use a nonparametric technique, we require an adequate measure of predictive power. Classical in-sample measures, such as the R^2 or adjusted R^2, are not appropriate. For example, R^2 favours the most complex model and is often inconsistent (see Valkanov, 2003), whereas standard penalization for complexity via a degree-of-freedom adjustment becomes meaningless in nonparametrics as it is unclear what the degrees of freedom are in this setting. Moreover, in prediction, we are not interested in how well a model explains the variation inside the sample but, instead, in its out-of-sample performance; hence, we aim to estimate the prediction error directly.

For the purpose of model as well as bandwidth selection, we use a generalized version of the validated R^2, the R^2_V, introduced by Nielsen and Sperlich (2003) based on a leave-k-out cross-validation. This method of finding the smoothing parameter has shown to be suitable also in a time series context. Our validation criterion is defined as

$$R^2_V = 1 - \frac{\sum_t (Y_t^{(A)} - \hat{m}_{-t})^2}{\sum_t (Y_t^{(A)} - \bar{Y}_{-t}^{(A)})^2}, \tag{5}$$

where leave-k-out estimators are used: $\hat{m}_{-t}$ for the nonparametric function m and $\bar{Y}_{-t}^{(A)}$ for the unconditional mean of $Y_t^{(A)}$. Both are computed by removing k observations around the tth time point. Here, we use $k = 1$, that is, the

classical leave-one-out estimator. Nevertheless, it is well-known that cross-validation often requires omitting more than one data point and, possibly, additional correction when the omitted fraction of data is considerable (see, for example, Burman et al., 1994).

R_V^2 measures the predictive power of a given model compared to the cross-validated historical mean; positive R_V^2 implies that the predictor-based regression model (1) outperforms the historical average excess stock return. Moreover, cross-validation not only punishes instances of overfitting, but also allows finding the optimal (predictive) bandwidth for non- and semi-parametric estimators (see Györfi et al., 1990); more recently, Bandi et al. (2016) have also studied optimality of the cross-validated bandwidth under stationary or nonstationary behaviour. Hence, in general, R_V^2 can be used for both model selection and optimal bandwidth choice.

5 Predicting excess returns based on different benchmarks

5.1 Data

In this paper, we take the long-term actuarial view and base our predictions on annual US data provided by Robert Shiller. This dataset, which is made available from `http://www.econ.yale.edu/~shiller/data.htm`, includes, among other variables, long-term changes of the Standard and Poor's (S&P) Composite Stock Price Index, bond price changes, consumer price index changes, and interest rate data from 1872 to 2015. This is an updated and revised version of Shiller (1989, Chapter 26), which provides a detailed description of the data. Various long-term studies use the same dataset, such as Chen et al. (2012), Elliott et al. (2013) and Favero et al. (2011).

Including structural changes in the modelling process is important, hence the length of this period allows for this. For example, Harvey et al. (2018) investigate the stability of predictive regression models and develop a real-time monitoring procedure for the emergence of predictive regimes. Rapach and Wohar (2004) find significant evidence of structural breaks in seven out of eight predictive regressions of S&P 500 returns, and three out of eight in CRSP (Center for Research in Security Prices) equal-weighted returns. Pesaran and Timmermann (2002) find that a linear predictive model that incorporates structural breaks improves the out-of-sample statistical forecasting power. In addition, from a statistical perspective, as we apply a fully non-parametric method, more observations are required than a linear regression to produce consistent estimates but also to have more information that will help us understand better the underlying data generating process.

Clearly, there are not many historical years in our records and data sparsity is an important issue in our approach. It could be argued that using monthly, weekly, or even daily data to the extent these are available would be preferred. However, it cannot be overlooked that prediction can be very different for yearly, monthly, weekly and daily data and that a good model for monthly

data might not be for yearly data, and vice versa. We take the long-term view using yearly data and predict at a one-year horizon as we are interested in actuarial models of long-term savings and potential econometric improvements to such models (see, for example, Guillén et al., 2013a, 2013b, Owadally et al., 2013, Bikker et al., 2012, Guillén et al., 2014, or Gerrard et al., 2014). For this, the methodology we adopt for validating our sparse long-term yearly data originates from the actuarial literature (see Nielsen and Sperlich, 2003).

5.2 Single benchmarking approach

In this section, we consider a single benchmarking approach where only the dependent variable is adjusted according to some benchmark, as shown in (1), while the independent variable(s) is (are) measured on the original nominal scale. The model (1) is estimated with a local-linear kernel smoother using the quartic kernel and the optimal bandwidth chosen by cross-validation, i.e., by maximizing the R_V^2 given by (5). Moreover, it should be kept in mind that the nonparametric method can estimate linear functions without any bias, given that we apply a local-linear smoother. Thus, the linear model is automatically embedded in our approach.

We study the empirical findings of R_V^2 values based on different validated scenarios shown in Table 1. Overall, we find that the term spread s is in itself the most important predictor under the different benchmarks. This remains quite the case also when combined with additional information. Inflation is another predictor that performs well when used concurrently as a benchmark. Hence, these are aspects where we especially focus our spotlight on in our discussion.

More specifically, if we constrain prediction to using only one-dimensional covariates, then the term spread s is the best predictor under the short interest benchmark $B^{(R)}$ with $R_V^2 = 13.2\%$, but also does quite well under the inflation benchmark $B^{(C)}$ with $R_V^2 = 9.9\%$. Imposing an additional covariate to s has, in general, a decreasing effect on R_V^2 however not substantial. Under $B^{(R)}$, R_V^2 remains in the majority of the combinations within the range 12.0–13.2%. Under other benchmarks, for example, $B^{(L)}$, s yields R_V^2 in the range 7.6–8.8%; under $B^{(E)}$, R_V^2 lies in the range 6.4–8.7%. The two-dimensional covariates $(Y^{(R)}, s)$, $(Y^{(L)}, s)$ and $(Y^{(E)}, s)$ result in lower predictive power than the previous ranges under $B^{(R)}$, $B^{(L)}$ and $B^{(E)}$ with, respectively, R_V^2 values 9%, 4.4% and 5.1%; while still below the best-performing ranges, (π, s) is found slightly better under $B^{(R)}$, $B^{(L)}$ and $B^{(E)}$ with, respectively, R_V^2 values 10.3%, 5.8% and 5.7%. It is worth noting that the occasional reduction in predictive power in the two-dimensional compared to the one-dimensional case is not particularly surprising as we use a fully nonparametric smoother that requires more observations than a linear regression to produce consistent estimates when fitting higher-dimensional models, so that our cross-validated R_V^2 might rank one-dimensional better than two-dimensional models. (Note that this is

Table 1 Predictive power for dependent variable $Y_t^{(A)}$: the single benchmarking approach. The prediction problem is defined in (1). The predictive power (%) is measured by R_V^2 as defined in (5). The benchmarks $B^{(A)}$ considered are based on the short-term interest rate ($A \equiv R$), long-term interest rate ($A \equiv L$), earnings-by-price ratio ($A \equiv E$), and consumer price index ($A \equiv C$). The predictive variables used are X_{t-1}, given by the dividend-by-price ratio d_{t-1}, earnings-by-price ratio e_{t-1}, short-term interest rate r_{t-1}, long-term interest rate l_{t-1}, inflation π_{t-1}, term spread s_{t-1}, excess stock return $Y_{t-1}^{(A)}$, or the possible different pairwise combinations as indicated.

Benchmark $B^{(A)}$	Explanatory variable(s) X_{t-1}						
	$Y^{(A)}$	d	e	r	l	π	s
Short-term rate	-1.5	-1.0	-0.3	4.0	-0.1	-1.4	13.2
Long-term rate	-1.8	-0.7	0.0	2.1	-0.1	-1.4	8.8
Earnings-by-price	-1.7	-1.4	-1.5	-0.1	-0.9	-1.2	8.7
Inflation	-1.3	-0.2	-1.5	1.1	-0.8	10.5	9.9
	$(Y^{(A)}, d)$	$(Y^{(A)}, e)$	$(Y^{(A)}, r)$	$(Y^{(A)}, l)$	$(Y^{(A)}, \pi)$	$(Y^{(A)}, s)$	
Short-term rate	-2.3	-2.0	1.9	-2.3	-2.8	9.0	
Long-term rate	-2.2	-1.8	0.0	-2.5	-3.0	4.4	
Earnings-by-price	-3.5	-3.7	-2.0	-2.8	-2.8	5.1	
Inflation	-1.2	-3.2	-0.6	-2.5	10.2	6.9	
	(d, e)	(d, r)	(d, l)	(d, π)	(d, s)		
Short-term rate	-2.7	3.0	-1.7	-2.4	12.0		
Long-term rate	-2.3	1.4	-1.4	-2.2	7.6		
Earnings-by-price	-3.8	-1.6	-2.3	-2.6	6.6		
Inflation	-2.0	0.9	-1.2	9.7	9.7		
	(e, r)	(e, l)	(e, π)	(e, s)			
Short-term rate	5.4	-0.8	-1.1	13.0			
Long-term rate	3.7	-0.3	-0.7	8.6			
Earnings-by-price	-1.5	-2.4	-2.7	6.4			
Inflation	0.0	-2.5	11.5	8.1			
	(r, l)	(r, π)	(r, s)				
Short-term rate	10.4	2.5	12.2				
Long-term rate	5.9	0.5	7.8				
Earnings-by-price	6.6	-1.6	8.2				
Inflation	6.6	9.7	8.8				
	(l, π)	(l, s)					
Short-term rate	-2.0	12.5					
Long-term rate	-2.1	8.1					
Earnings-by-price	-1.9	8.1					
Inflation	10.1	9.3					
	(π, s)						
Short-term rate	10.3						
Long-term rate	5.8						
Earnings-by-price	5.7						
Inflation	16.1						

not the case for a linear model estimated with OLS based on the usual R^2 measure which would always choose the most complex model.)

Remarkable is the case of predictor π, either in itself or combined with covariates $Y^{(C)}, d, e, r, l$, under the inflation benchmark $B^{(C)}$ leading to R_V^2 in the range 9.7–11.5%. In addition, when put together with the term spread, the resulting combination (π, s) under $B^{(C)}$ is the clear winner reaching an astonishing $R_V^2 = 16.1\%$. Given that the inflation benchmark might be the most important one for many pension product applications, this high predic-

tive power is very promising. Under the $B^{(R)}$, $B^{(L)}$ and $B^{(E)}$ benchmarks, the performance of covariate π deteriorates and, in fact, the historical average excess return in these cases surpasses the predictor-based regression model, as implied by negative R^2_V, unless it is combined with covariate r or s.

Finally, other covariates in themselves or combined also lead to negative R^2_V values; such predictor examples include Y, d, e, l and their pairwise combinations under any benchmark. On the contrary, the short-term rate r individually or combined with other covariates boosts, with a few exceptions, the predictive power of our nonparametric regression model.

5.3 Full benchmarking approach

The second step now is to analyze whether an adequate transformation of the explanatory variables can further improve predictions. Recall that fully nonparametric models suffer in several aspects by the curse of dimensionality, in particular, as in our framework, where we confront sparsely distributed annual observations in higher dimensions. In statistics, it is well-known that importing more structure in the estimation process can help reduce or circumvent such problems. For example, Nielsen and Sperlich (2003) investigate an additive functional structure in the context of predictability of excess stock returns (as proposed in the statistical literature by Stone, 1985). Their results indicate a more complex structure than additivity, as the fully nonparametric models always do better in terms of validated R^2 than the additive counterparts. Scholz et al. (2015) propose a semiparametric bias reduction method for the purpose of importing more structure based on a multiplicative correction with a parametric pilot estimate. Alternatively, Scholz et al. (2016) make use of economic theory saying that the price of a stock is driven by fundamentals and investors should focus on forward earnings and profitability. They include information on the same years' – instead of last years' – explanatories and improve predictions.

Here, we propose an extension of the study in Section 5.2 using economic structure in the sense that we consider adjusting both the independent and dependent variables according to the same benchmark. For example, in the full benchmarking approach with an inflation benchmark, both excess returns and covariates are expressed in terms of inflation. In pension research, it is sensible to employ such a model with all returns and covariates net-of-inflation. This, in turn, provides a simple scaling when working on long-term forecasts in real terms.

In general, in our full benchmarking approach, the prediction problem is reformulated as

$$Y_t^{(A)} = m(X_{t-1}^{(A)}) + \xi_t, \tag{6}$$

where we use transformed predictive variables

$$X_{t-1}^{(A)} = \begin{cases} \frac{1+X_{t-1}}{B_{t-1}^{(A)}}, & X \in \{d, e, r, l, \pi, s\} \\ Y_{t-1}^{(A)} \end{cases}, \quad A \in \{R, L, E, C\}. \tag{7}$$

Table 2 Predictive power for dependent variable $Y_t^{(A)}$: the full benchmarking approach. The prediction problem is defined in (6). The predictive power (%) is measured by R_V^2 as defined in (5). The benchmarks $B^{(A)}$ considered are based on the short-term interest rate ($A \equiv R$), long-term interest rate ($A \equiv L$), earnings-by-price ratio ($A \equiv E$), and consumer price index ($A \equiv C$). The predictive variables used are transformed $X_{t-1}^{(A)}$ using the indicated benchmark $B_{t-1}^{(A)}$ as shown in (7). X_{t-1} are given by the dividend-by-price ratio d_{t-1}, earnings-by-price ratio e_{t-1}, short-term interest rate r_{t-1}, long-term interest rate l_{t-1}, inflation π_{t-1}, term spread s_{t-1}, or the possible different pairwise combinations as indicated. "–" are not applicable cases of matched covariate with benchmark.

Benchmark $B^{(A)}$	Explanatory variable(s) $X_{t-1}^{(A)}$						
	$Y^{(A)}$	$d^{(A)}$	$e^{(A)}$	$r^{(A)}$	$l^{(A)}$	$\pi^{(A)}$	$s^{(A)}$
Short-term rate	-1.5	4.2	7.0	–	13.2	-1.2	13.2
Long-term rate	-1.8	-0.2	1.0	8.9	–	-1.5	8.9
Earnings-by-price	-1.7	-2.3	–	0.3	-1.2	0.0	8.7
Inflation	-1.3	10.7	13.3	6.6	10.8	–	8.9
	$(Y^{(A)}, d^{(A)})$	$(Y^{(A)}, e^{(A)})$	$(Y^{(A)}, r^{(A)})$	$(Y^{(A)}, l^{(A)})$	$(Y^{(A)}, \pi^{(A)})$	$(Y^{(A)}, s^{(A)})$	
Short-term rate	2.7	5.0	–	8.7	-2.7	8.7	
Long-term rate	-1.9	-0.8	4.3	–	-3.1	4.3	
Earnings-by-price	-4.1	–	-1.7	-3.2	-2.5	5.0	
Inflation	11.2	12.3	5.8	10.1	–	6.0	
	$(d^{(A)}, e^{(A)})$	$(d^{(A)}, r^{(A)})$	$(d^{(A)}, l^{(A)})$	$(d^{(A)}, \pi^{(A)})$	$(d^{(A)}, s^{(A)})$		
Short-term rate	4.1	–	11.5	2.6	11.5		
Long-term rate	-1.2	7.0	–	-1.8	7.0		
Earnings-by-price	–	-2.8	-3.7	-3.2	4.9		
Inflation	11.2	9.8	10.3	–	16.2		
	$(e^{(A)}, r^{(A)})$	$(e^{(A)}, l^{(A)})$	$(e^{(A)}, \pi^{(A)})$	$(e^{(A)}, s^{(A)})$			
Short-term rate	–	12.9	5.2	12.9			
Long-term rate	8.4	–	-0.1	8.4			
Earnings-by-price	–	–	–	–			
Inflation	12.4	12.0	–	18.7			
	$(r^{(A)}, l^{(A)})$	$(r^{(A)}, \pi^{(A)})$	$(r^{(A)}, s^{(A)})$				
Short-term rate	–	–	–				
Long-term rate	–	5.7	–				
Earnings-by-price	4.9	-1.6	6.5				
Inflation	15.1	–	16.9				
	$(l^{(A)}, \pi^{(A)})$	$(l^{(A)}, s^{(A)})$					
Short-term rate	10.2	–					
Long-term rate	–	–					
Earnings-by-price	-2.4	6.3					
Inflation	–	16.7					
	$(\pi^{(A)}, s^{(A)})$						
Short-term rate	10.2						
Long-term rate	5.7						
Earnings-by-price	5.6						
Inflation	–						

This model can be interpreted as a way of reducing dimensionality of the estimation procedure as $X_{t-1}^{(A)}$ encompasses an additional predictive variable.

Results of this empirical study are presented in Table 2. We find that, in the majority of the cases, the full outruns the single benchmarking approach presented in Table 1 in terms of R_V^2. In addition, by full benchmarking, several cases of inability of the predictor-based regression model to beat the historical

average excess return are now surmounted with R^2_V becoming positive. The term spread s retains its superior predictability with perceptible improvement brought in two-dimensional settings when paired with the dividend-by-price ratio d, the earnings-by-price ratio e, the short rate r, or the long rate l under the $B^{(C)}$ benchmark, reaching up to a notable $R^2_V = 18.7\%$ when using specifically predictors $(e^{(C)}, s^{(C)})$. Otherwise, as in the single benchmarking approach, we experience similarly some decrease in predictability when adding covariates to s, with $(Y^{(R)}, s^{(R)})$, $(Y^{(L)}, s^{(L)})$ and $(Y^{(E)}, s^{(E)})$, although resulting in positive R^2_V, still performing worse under $B^{(R)}$, $B^{(L)}$ and $B^{(E)}$.

A few more interesting comments relating to benchmarks $B^{(R)}$, $B^{(L)}$ and $B^{(E)}$ are in order. We find that under $B^{(R)}$ and $B^{(L)}$, $s^{(R)}$ and $s^{(L)}$ have, respectively, the same predictive power as $l^{(R)}$ and $r^{(L)}$, either in themselves or when combined with a common covariate such as Y, d, e, and this is considerably improved from the single benchmarking approach: for example, changing from predictor l under $B^{(R)}$ to $l^{(R)}$ increases R^2_V from -0.1% to 13.2% or from predictor r under $B^{(L)}$ to $r^{(L)}$ increases R^2_V from 2.1% to 8.9%. Nevertheless, the highest predictability power for these benchmarks is close to that from the single approach originating from s under $B^{(R)}$ and $B^{(L)}$; similar remark applies in the $B^{(E)}$ benchmark where $s^{(E)}$ is the best predictor achieving at most the same level of predictability as s under $B^{(E)}$ in the single approach.

We turn now attention to $B^{(C)}$. Here, remarkable in the full approach is the contribution of the earnings-by-price ratio $e^{(C)}$ whose predictability, contrary to that of e under $B^{(C)}$ in the single approach, is now put on show. We find that $e^{(C)}$, $(Y^{(C)}, e^{(C)})$, $(e^{(C)}, s^{(C)})$ result in R^2_V of 13.3%, 12.3% and 18.7% against maximum R^2_V of 9.9%, 6.9% and 8.1% from individual s, $(Y^{(C)}, s)$, (e, s) under $B^{(C)}$ in Table 1. Notable contributions from other covariates are those of $(d^{(C)}, s^{(C)})$, $(r^{(C)}, s^{(C)})$ and $(l^{(C)}, s^{(C)})$ with R^2_V of 16.2%, 16.9% and 16.7% against maximum R^2_V of 9.7%, 8.8% and 9.3% from (d, s), (r, s), (l, s) under $B^{(C)}$ in Table 1. Overall, using earnings-by-price as an individual predictor and two-dimensional predictors that include the term spread capture the best predictive performances; the winning pair is earnings-by-price ratio and term spread.

5.4 Synopsis and further discussion

In summary, our study indicates that with single benchmarking, the spread in nominal terms has a significantly higher predictability than the earnings-by-price or even the earnings-by-price together with the spread. This is consistent with Kothari et al. (2006) who state that lagged earnings exhibit no predictive power for future annual returns. Besides, evidence, for example, from Resnick and Shoesmith (2002) suggests that the value of the yield spread holds important information about the probability of a bear stock market.

Our full benchmarking analysis complements the original conclusion of Kothari et al. (2006) by showing that, net-of-inflation, i.e., in real terms, the earnings-by-price beats the spread, with their combination performing best.

This is an important observation as benchmarking fits well in building models net-of-inflation, as discussed in Donnelly et al. (2018) and Gerrard et al. (2018a, 2018b), while at the same time expanding to double benchmarking endows earnings with predictive power. This is a crucial implication for pension research or other long-term saving strategies, where one should look at real value and implement double benchmarking including both spread and earnings.

In addition, as part of their comprehensive analysis of a long literature including articles based on different techniques, variables, and time periods, and sometimes contradicting results, Welch and Goyal (2008) find that predictive linear regressions using prominent variables, including our choices, do result in poor predictability, in-sample and out-of-sample. Nevertheless, following their recommendation, we explore the possibility of an alternative model approach, here, a simple nonparametric regression model and different benchmarks. Contrary to them, our method with double benchmarking leads, as highlighted earlier, to favourable predictive results implying that nonlinear and/or nonparametric models are necessary in order to represent the complicated relationship between earnings, prices, the yield curve and the stock returns.

6 Conclusion

In this communication, we define machine learning as a working framework comprising the following summarized key ingredients: articulation, domain knowledge, final selection by validation, conduct of validation consistently with underlying statistical principles and properly channeled prior knowledge. We then apply to forecasting stock returns in excess of different benchmarks, including the inflation, long interest rate and earnings-by-price ratio to supplement the short interest rate which is by far the most commonly used in finance. Indeed, this paper expresses an interest in going beyond this as different benchmarks might be important, for example, when modelling returns in real terms (inflation benchmark) or modelling returns in excess of long-term interest rate.

We use predictors such as the dividend-by-price ratio, earnings-by-price ratio, short interest rate, long interest rate, the term spread, the inflation, as well as the lagged excess stock return. We also investigate the option of full benchmarking, meaning that, not only returns are benchmarked, but also the covariates used to predict them. The full benchmarking approach can also be seen as an example of a dimension reduction technique, where more information is included in the nonparametric prediction without extra cost in the form of increasing problem dimensionality. The summarizing conclusion of our analysis is that, in real terms, the combination of earnings-by-price and long-short rate spread within our nonparametric model setting has the best predictive outcome, which is important for long-term saving strategies.

Acknowledgements The third author thanks the Institute and Faculty of Actuaries in the UK for financial support through the grant "Minimizing Longevity and Investment Risk while Optimizing Future Pension Plans".

References

Bandi, F. M., Corradi, V. and Wilhelm, D. (2016) Possibly nonstationary cross-validation. Cemmap Working Paper CWP11/16.

Bikker, J. A., Broeders, D. W. G. A., Hollanders, D. A. and Ponds, E. H. M. (2012) Pension funds' asset allocation and participant age: A test of the life-cycle model. *Journal of Risk and Insurance*, **79**, 595–618.

Burman, P., Chow, E. and Nolan, D. (1994) A cross-validatory method for dependent data. *Biometrika*, **81**, 351–358.

Chen, L., Da, Z. and Priestley, R. (2012) Dividend smoothing and predictability. *Management Science*, **58**, 1834–1853.

Cheng, T., Gao, J. and Linton, O. (2018) Multi-step non- and semi-parametric predictive regression for short and long horizon stock return prediction. Cemmap Working Paper CWP03/18.

Dahlhaus, R. (1997) Fitting time series models to nonstationary processes. *The Annals of Statistics*, **25**, 1–37.

Donnelly, C., Guillen, M., Nielsen, J. and Pérez-Marín, A. (2018) Implementing individual savings decisions for retirement with bounds on wealth. *ASTIN Bulletin*, **48**, 111–137.

Doukhan, P. (1994) *Mixing*. New York: Springer.

Elliott, G., Gargano, A. and Timmermann, A. (2013) Complete subset regressions. *Journal of Econometrics*, **177**, 357–373.

Favero, C. A., Gozluklu, A. E. and Tamoni, A. (2011) Demographic trends, the dividend-price ratio, and the predictability of long-run stock market returns. *Journal of Financial and Quantitative Analysis*, **46**, 1493–1520.

Gerrard, R., Guillén, M., Nielsen, J. P. and Pérez-Marín, A. M. (2014) Long-run savings and investment strategy optimization. *The Scientific World Journal*, **2014**.

Gerrard, R., Hiabu, M., Kyriakou, I. and Nielsen, J. P. (2018a) Communication and personal selection of pension saver's financial risk. *European Journal of Operational Research (in press)*.

— (2018b) Self-selection and risk sharing in a modern world of life-long annuities. *British Actuarial Journal (forthcoming)*.

Guillén, M., Jarner, S. F., Nielsen, J. P. and Pérez-Marín, A. M. (2014) Risk-adjusted impact of administrative costs on the distribution of terminal wealth for long-term investment. *The Scientific World Journal*, **2014**.

Guillén, M., Konicz, A. K., Nielsen, J. P. and Pérez-Marín, A. M. (2013a) Do not pay for a Danish interest guarantee. The law of the triple blow. *Annals of Actuarial Science*, **7**, 192–209.

Guillén, M., Nielsen, J. P., Pérez-Marín, A. M. and Petersen, K. S. (2013b) Performance measurement of pension strategies: a case study of Danish life-cycle products. *Scandinavian Actuarial Journal*, **2013**, 49–68.

Györfi, L., Härdle, W., Sarda, P. and Vieu, P. (1990) *Nonparametric Curve Estimation from Time Series.* Lecture Notes in Statistics. Heidelberg: Springer.

Hansen, B. (2006) Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, **24**, 726–748.

Harvey, D., Leybourne, S., Sollis, R. and Taylor, R. (2018) Detecting regimes of predictability in the U.S. equity premium. Essex Finance Center Working Paper No37: 10-2018.

Kothari, S. P., Lewellen, J. and Warner, J. B. (2006) Stock returns, aggregate earnings surprises, and behavioral finance. *Journal of Financial Economics*, **79**, 537–568.

Li, Q. and Racine, J. (2007) *Nonparametric Econometrics.* Princeton and Oxford: Princeton University Press.

Masry, E. (1996) Multivariate regression estimation local polynomial fitting for time series. *Stochastic Processes and their Applications*, **65**, 81–101.

Merton, R. (2014) The crisis in retirement planning. *Harvard Business Review*, **92**, 43–50.

Nielsen, J. and Sperlich, S. (2003) Prediction of stock returns: A new way to look at it. *ASTIN Bulletin*, **33**, 399–417.

Owadally, I., Haberman, S. and Gómez Hernández, D. (2013) A savings plan with targeted contributions. *Journal of Risk and Insurance*, **80**, 975–1000.

Pesaran, H. and Timmermann, A. (2002) Market timing and return predictability under model instability. *Journal of Empirical Finance*, **8**, 495–510.

Rapach, D. and Wohar, M. (2004) Structural change and the predictability of stock returns. St. Louis University and University of Nebraska at Omaha Working Paper.

Resnick, B. G. and Shoesmith, G. L. (2002) Using the yield curve to time the stock market. *Financial Analysts Journal*, **58**, 82–90.

Robinson, P. (1983) Nonparametric estimators for time series. *Journal of Time Series Analysis*, **4**, 185–207.

Scholz, M., Nielsen, J. and Sperlich, S. (2015) Nonparametric prediction of stock returns based on yearly data: The long-term view. *Insurance: Mathematics and Economics*, **65**, 143–155.

Scholz, M., Sperlich, S. and Nielsen, J. (2016) Nonparametric long term prediction of stock returns with generated bond yields. *Insurance: Mathematics and Economics*, **69**, 82–96.

Shiller, R. J. (1989) *Market Volatility.* Cambridge: MIT Press.

Stone, C. (1985) Additive regression and other nonparametric models. *Annals of Statistics*, **13**, 689–705.

Valkanov, R. (2003) Long-horizon regressions: theoretical results and applications. *Journal of Financial Economics*, **68**, 201–232.

Vogt, M. (2012) Nonparametric regression for locally stationary time series. *The Annals of Statistics*, **40**, 2601–2633.

Welch, I. and Goyal, A. (2008) A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, **21**,

1455–1508.