Statistics and Econometrics Series 12
March 2001

Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

Prediction of stocks: A new way to look at it.

Jens Pech Nielsen and Stefan Sperlich[1]

Abstract

While the traditional R^2 value is useful to evaluate the quality of a fit, it does not work when it comes to evaluating the predictive power of estimated financial models in finite samples. In this paper we introduce a validated R_V^2 value that is Taylor made for prediction. Based on data from the Danish stock market, using this measure we find that the dividend-price ratio has good predictive power for time horizons between one year and five years. We explain how the R^2s for different time horizons could be compared, respectively, how they must not be interpreted. For our data we can conclude that the quality of prediction is almost the same for the five different time horizons. This is in contradiction to earlier studies based on the traditional R^2 value, where it has been argued that the predictive power increases with the time horizon up to a horizon of about five or six years. Furthermore, we find that while inflation and interest rate do not add to the predictive power of the dividend-price ratio then last years excess stock return does.

Keywords: Prediction, Stocks, Bonds, Cross Validation.

[1] Nielsen, Codanhus, 60 Gammel Kongevej, DK - 1790 Copenhagen V, Denmark. npj@codan.dk; Sperlich, Universidad Carlos III de Madrid, Depto. de Estadística y Econometría, c/ Madrid 126, E - 28903 Getafe - Madrid, Spain. stefan@est-econ.uc3m.es. The research was supported by the Danish Social Science Research fund and the Spanish DGES Nr.PB98-0025. Comments and suggestions from Thomas Engsted and Carsten Tanggård are gratefully acknowledged.

1 Introduction

Long term investors have the contradicting aims of minimizing their risk and maximizing return over the long run. Much financial literature investigates trading patterns and strategy among long term investors, for example, Barber and Terrance (2000) argue for a buy-and-hold type of strategy that does not eat up returns by trading costs and many professional advisers argue that stocks are better over the long run, see Siegel (1998) and Jagannathan and Kocherlakota (1996) for particular easily read accounts on this. Other professional financial advisers say that expected returns in financial markets vary over time and contain a significant predictable component. Consequently time periods exist where the long term investor might choose to sell stocks and buy bonds, because the return on stocks in these time periods do not match the risk involved. The dividend-price ratio and the earning-price ratio, in particular, has proven to have some predictive power for future stock returns, see Campbell, Lo, and MacKinlay (1997, Chapter 7) for an up-to-date account regarding the predictability of the dividend-price ratio, and see Shiller (2000, p.8) for a recent warning of an overvalued American stock exchange based on the earning-price ratio.

Campbell, Lo, and MacKinlay (1997) argued that the predictable component of stock yields is increasing with the time horizon, since the measure of fit, the R^2, increases rapidly with the time horizon. We modify this point of view in the present paper, firstly by showing that for increasing time horizon a significant increase of the R^2 value is indeed necessary to maintain the same quality of fit and secondly by introducing a validated R^2 (we will note it by R^2_V value) that is Taylor made for prediction purposes while the traditional R^2 values only measure the quality of in-sample fit. We investigate our new point of view of prediction by analyzing yearly Danish data from the period $1922-1996$. Our conclusions are surprising. First: based on our recalculated scale for the R^2 values we can conclude that if the traditional R^2 values are used for prediction, then prediction on a one year basis gives the best power of prediction. This is somehow in contrast to Campbell, Lo, and MacKinlay (1997) and many others (see e.g. Richardson and Stock, 1989) which give the impression that longer time horizons are preferable for prediction due to their higher R^2 values. However, traditional R^2 can not be used for prediction. Our adjusted measure of predictive power, the R^2_V value, tells us that time horizons between one and four years seem optimal for predictions and that the quality of predictions is big enough to point out when expected excess return on stocks, compared to bonds, is below zero.

Traditional parametric or nonparametric one-dimensional regression, with dividend-price as the dependent variable, does have good prediction power whereas knowledge of inflation and interest rates do not add to this predictive power. However, our study shows that the one year lagged returns do. The best predictive filter on a one year basis turns out to be a two-dimensional fully nonparametric estimator based on the dividend-price ratio and last years lagged excess return. Last years excess return enters with a tendency towards reversal, such that good years tend to follow bad years and vice versa. The dividend-price ratio is, however, still the most indicative parameter while estimating the excess returns of the coming years. Moreover, based on the current level on the dividend-price ratio in Denmark (around 1%) we can conclude that expected excess returns on stocks are indeed below zero, for all the considered time horizons with good prediction power, namely one, two, three, four and five years time horizons. If this finding is credible, then it can conclude that the current market and political situation in Denmark is out of balance, since all institutional investors heavily increase their percentages of stocks in their portfolios right now. On average, an increase from around 20% invested in stocks to around 40% invested in stocks have been seen for long term institutional investors in Denmark over the last five years. The model of this paper argue that this strategy increase the risk without increasing the average return. Unfortunately the currently used stochastic models of the pension industry, see Wilkie (1986) and Wilkie (1990) for the by far most popular actuarial prediction models, do not have any ability to warn the pension industry and its customers towards periods with high risks and low returns on stocks. We believe that the considerations of this paper can be helpful while developing a modern information system for the long term investor.

2 The basic relationship between stock returns and economic factors

One traditional equation for the value of a stock is

$$P_t = \sum_{j=1}^{\infty} (1+\gamma)^{-j}(1+g)^{j-1} D_t.$$

where most of the entering quantities on the right hand side are unknown, γ, discount rate, g, constant growth of dividend yields, i inflation and D_t real dividend yield paid out during the period t. This model was introduced to the financial theory by Williams (1938) and Gordon and Shapiro (1956). Campbell and Shiller (1988) referred to the model as the "dividend-ratio" in absence of uncertainty, see also Goetzman and Jorion (1993), Hodrick (1992), and Fama and French (1988). For simplicity the discount rate and the growth rate does not depend on time in this model although this is well known to be incorrect. The point of the above identity is however, that it gives a strong indication that the price of stocks depend on quantities such that dividend yield, interest rate and inflation. The two latter being highly correlated with almost any relevant discount rate. It is also clear from the above identity that a decrease in discount rate, which is highly correlated with an increase in bond yield, are related to an increase in the stock return and vice versa. The correlation of 0.5 of stock returns and long term bond yields is therefore not surprising.

Now let us look at another fundamental equation characterizing the stock market, namely the following formula for the log dividend-price ratio, which can be derived from a first-order Taylor approximation of the identity relating the one-period log stock return to log stock prices and log dividends, see Campbell and Shiller (1988):

$$\delta_t = E_t \sum_{j=1}^{\infty} \rho^{j-1}(r_{t+j} + S_{t+j} - \Delta d_{t+j}) + k \tag{1}$$

where $\delta_t \equiv \log(D/P)_t \equiv d_t - p_t$. D_t and P_t are real dividends paid during period t and real stock prices at the end of period t, respectively. r_{t+j} is the one-period log real interest rate from period $t+j-1$ to $t+j$, and S_{t+j} is the log excess stock return from period $t+j-1$ to $t+j$, i.e. the log stock return in excess of a short-term interest rate. ρ is equal to $(1+\exp(\overline{\delta}))^{-1}$, where $\overline{\delta}$ is the mean log dividend-price ratio over the sample. E_t and Δ are the conditional expectations operator and the first-difference operator, respectively, and k is a constant arising from the linearization. Another way of writing the above formula was given in Campbell

(1991) and derives the following basic expression for the unexpected log excess stock return from period t to $t+1$

$$S_{t+1} - E_t S_{t+1} = (E_{t+1} - E_t) \left\{ \sum_{j=0}^{\infty} \rho^j \Delta d_{t+1+j} - \sum_{j=0}^{\infty} \rho^j r_{t+1+j} - \sum_{j=1}^{\infty} \rho^j S_{t+1+j} \right\} \tag{2}$$

(2) is a dynamic accounting identity stating that positive unexpected excess stock returns are associated either with higher expected future long-term dividend growth, and/or lower expected future long-term real stock returns, where the latter can be decomposed into real interest rates and excess stock returns. Thus, unexpected excess stock returns are the result of either news about future dividends, news about future real interest rates, or news about future excess stock returns (or a combination of the three). This type of analyzes was replicated in Engsted and Tanggård (2000) for the Danish data. Based on the above we find it reasonable to consider a regression of the future excess stock returns on the actual ones, dividend by price yields, short-term interest rate and inflation. The data and particular model will be specified in detail in the proceeding sections.

3 The data and our definition of prediction

In this paper we use the annual Danish stock market data from Lund and Engsted (1996), respectively the extended sample period $1922-1996$ from Engsted and Tanggård (2000). Considered are the time series (S_t, D_t, I_t, r_t), where S_t is stock return, D_t is dividend yield, I_t is inflation and r_t is the short-term interest rate. The stock index is based on a value weighted portfolio of individual stocks chosen to obtain maximum coverage of the marked index of the Copenhagen Stock Exchange. In constructing the data corrections were made for stock splits and new equity issues below market prices. P_t is the (nominal) stock price at the end of year t, while D_t denotes (nominal) dividends paid during year t divided by the stock price at the end of year t (the appendix in Lund and Engsted (1996) contains a detailed description of the data). As a measure of the short-term interest rate we use the Danish Central Bank's discount rate up to 1975, spliced together with a short-term zero-coupon yield for the period thereafter. In computing real values, we deflate nominal values by the consumption deflator. The real excess stock return is defined as

$$S_t = \log\left\{(P_t + D_t)/P_{t-1}\right\} - r_{t-1}.$$

The resulting average of these excess stock returns are 2.1% for the period $1922-1996$ and 3.2% for the after war period $1947-1996$. The problem of prediction is considered as follows: Let S_t be the excess stock return at time t and let W_t be some one-dimensional or multidimensional stochastic process that we wish to use for prediction. We establish prediction through the following model

$$S_t = g(W_{t-1}) + \epsilon_t, \;\; g \in G \tag{3}$$

where the error variables ϵ_t are independent mean zero stochastic variables and G is some set of the possible functional relationships between the independent and the dependent variable. We will see below that G can be chosen as a parametric family of functions or as a nonparametric set of functions where smoothness is the only restriction or as a nonparametric set of functions with restriction. In this last case we consider the restriction that g is additive when W_{t-1} is multidimensional. Let us for convenience assume that a $\widehat{g}$ and an average $\widehat{\mu}$ is estimated from a data set that is independent of our data set, then the optimal estimator within the set G could be based on the quality of the prediction evaluated as

$$R_A^2 = 1 - E\left\{S_t - \widehat{g}(W_{t-1})\right\}^2 / E\left(S_t - \widehat{\mu}\right)^2 . \tag{4}$$

R_A^2 is therefore one minus the standardized average prediction error of the chosen prediction strategy. If we knew R_A^2 for a number of different model and estimation strategies then we would simply choose that combination of modeling and estimation that gives the smallest average prediction error. We do not, however, know these $R_A^{2\prime}s$ and moreover the $\widehat{g}$ and the average $\widehat{\mu}$ has to be estimated from our data set, i.e. are not independent of them. Therefore we have to consider approximations to the average prediction error based on the data. In our case we the biggest set of predicting variables we consider is

$$W_t = \left(S_{t-1}, D_t, I_t, R_t\right). \tag{5}$$

We also consider all possible three-dimensional, two-dimensional and one-dimensional subsets of the stochastic process W_t. For example, we consider

$$\overline{W}_t = (S_t, D_t) \tag{6}$$

and

$$\overline{\overline{\overline{W}}}_t = D_t. \tag{7}$$

In the results below it turns out that $\overline{W}_t$ and $\overline{\overline{\overline{W}}}_t$ are better to use for prediction than the big four-dimensional vector W_t that seems to introduce too much noise due to the many variables.

4 What is a good prediction?

It can be difficult to say what a good prediction is. Campbell, Lo and Mackinlay (1997, p.269) suggest the traditional R^2 measure as a way of understanding the predictive power of an estimated model and they show that R^2 is increasing with the horizon considered. This last fact has been observed by a number of authors including Goetzmann and Jorion (1993) who also noted that the increased R^2 for the longer horizons is followed by an increased variance of the estimated slope leaving results of tests almost indifferent of the time horizon. It therefore seems that R^2 values are not directly comparable for different time horizons. Based on a direct comparison of R^2 values then one could get the impression that long horizon excess stock returns are more predictable than short horizon returns. Campbell, Lo and Mackinlay (1997, p.271) point this out theoretically and notice that when the forecasting variable is highly persistent, then the R^2 statistic can continue to rise out to very long time horizons.

In Table 2 below we calculate for the time horizons $T = 1$ to $T = 6$ those R_A^2 values, see Section 3, that represent an improved prediction of the conditional mean μ, $\Delta\mu$, of respectively 0.01 to 0.05, where the excess return over T years is defined as

$$S_{t+1} + \ldots + S_{t+T}$$

with S_t being the 1-year log excess stock return from $t-1$. In Table 1 is given the estimated variance of the observed returns for the time horizons $T = 1$ to $T = 6$. Below we recalculate how much an improvement in variation compared to a given total variation means when calculated as an improvement on the estimated mean. For example, the value of the needed improvement of R^2 to correspond to an improvement of the estimation accuracy of the mean of yearly 2% is calculated for $T = 4$ as

$$\frac{(1.02^4 - 1)^2}{0.082914} = 8.2\%.$$

T	$Mean$	$Variance\ (s_T^2)$
1	0.020949	0.029557
2	0.035825	0.050473
3	0.050328	0.064505
4	0.070561	0.082914
5	0.089415	0.087686
6	0.11041	0.091136

Table 1: Mean and estimated variance of overlapping excess stock returns for different time horizons T.

This is off course just a rule of thumb based on a most simple mode and it is therefore not the final recalculation of the effect on the mean of an improvement of a calculated R^2 value when considering a complicated dynamic time series structures. We consider, however, this approximation to be good enough for our purposes and use it both, for the classical R^2 value and the validated R_V^2 value that we define in Section 5. Note also that 2% on the mean is measured in basis points. When we say that an average one year prediction of returns is 2% better on the mean than another prediction, then it is comparable to the simple case, where a true mean of 5% is estimated by 4% by the good predictor and 2% by the bad predictor.

Based on the empirical variation estimated in Table 1, we are able to construct Table 2 below that give the sought for correspondence between an improvement measured on the total variation and an improvement measure on the estimated predicted mean. We consider returns over the time horizons $T = 1$ to $T = 6$. Table 2 is crucial for the interpretation of the results in the rest of the paper, since it is a lot more interesting to relate to an improvement on the estimated mean than to improvements on the (total) variation.

Since the mean of the excess stock returns are 2.1% for the period $1922 - 1996$ and 3.2% for the period $1947 - 1996$, then R^2 values for $T = 1$ in the neighborhood of 3% gives an improved prediction of the mean that is higher than the total mean itself - in other words a quite powerful level of prediction. It is also seen from the table that the R^2 values needed, for a fixed improvement of the mean, almost can be approximated by a linear relationship for $T = 1$ to $T = 5$ from $1923 - 1996$. This is interesting, since it can be observed in the table of Campbell, Lo and Mackinlay (1997, p.269) based on American data, that the improvement of the R^2 value over time is indeed almost linear over time like our observations based on the Danish data contains the same type of observation, see Section 5 below. Hence the improvement of R^2 values for longer time horizons are indeed of an order of magnitude corresponding to more or less the same level of prediction.

Above we gave an empirical explanation that a linear improvement of R^2 values over time indeed corresponds to the same quality of prediction. For a better understanding we now give additionally a theoretical explanation that such a linear improvement of R^2 values over time

horizon, T	1922-1996 s_T	$\Delta\mu = 0.01$ R^2	$\Delta\mu = 0.02$ R^2	$\Delta\mu = 0.03$ R^2	$\Delta\mu = 0.04$ R^2	$\Delta\mu = 0.05$ R^2
1	17,2%	0,3%	1,4%	3,0%	5,4%	8,5%
2	22,5%	0,8%	3,2%	7,3%	13,2%	20,8%
3	25,4%	1,4%	5,8%	13,3%	24,2%	38,5%
4	28,8%	2,0%	8,2%	19,0%	34,8%	56,0%
5	29,6%	3,0%	12,4%	28,9%	53,5%	87,1%
6	30,2%	4,2%	17,5%	41,3%	77,2%	-

horizon, T	1947-1996 s_T	$\Delta\mu = 0.01$ R^2	$\Delta\mu = 0.02$ R^2	$\Delta\mu = 0.03$ R^2	$\Delta\mu = 0.04$ R^2	$\Delta\mu = 0.05$ R^2
1	19,4%	0,3%	1,1%	2,4%	4,2%	6,6%
2	24,8%	0,7%	2,7%	6,0%	10,8%	17,1%
3	27,4%	1,2%	5,0%	11,5%	20,7%	33,1%
4	30,8%	1,7%	7,2%	16,6%	30,4%	48,0%
5	31,2%	2,7%	11,1%	26,0%	48,1%	79,3%
6	31,8%	3,7%	15,7%	37,2%	69,5%	-

Table 2: Corresponding R^2s for improvement, $\Delta\mu$, of the conditional mean prediction T-year excess stock returns for different time horizons T.

are to be expected, at least for a horizon below some T_{max} value. For example Campbell, Lo and Mackinlay (1997, p.269) has an empirical T_{max} value equal to six years.

Consider the following definition for the (classic) R^2 value:

$$R^2 = 1 - \frac{not\ explained\ Variation}{total\ Variation}, \tag{8}$$

where total variation is defined above and

$$\begin{aligned} total\ Variation &= \sum_{i=1}^{n} (y_i - \bar{y})^2, \text{ where } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \\ not\ explained\ Variation &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2, \end{aligned}$$

and $\widehat{y}_i$ is the estimator of y_i based on the estimated model.

Let us consider a simple linear model with centered variables, i.e.

$$S_t = \beta X_t + \varepsilon_t, \ t = 1, \ldots, n$$

where the ε are iid error terms, independent of X. Imagine now that most of the *total Variation* of $S, V(S)$, is caused by the variation of ε such that the averaged *explained Variation* $V(\beta X) = \beta^2 V(X)$ is rather small compared to $V(S)$. This means that X compared to Y does not move much and we therefore might expect the explanatory power of X_t on S_t to hold almost unchanged for the following observations S_{t+1},. . .,S_{t+T-1}, at least up to some T_{max}. Then the following equations would hold (approximately):

$$S^*_{t,T} = \sum_{j=0}^{T-1} S_{t+j} = \beta_T X_t + \varepsilon^*_{t,T},$$

with

$$\varepsilon^*_{t,T} \approx \sum_{j=0}^{T-1} \varepsilon_{t,j}$$

and $\beta_T \approx T\beta$ as long as T is smaller than T_{max}. Thus the *explained Variation* follows the approximation

$$V(\beta_T X_1) \approx \beta^2 T^2 V(X_1).$$

Since

$$V(\varepsilon^*_{t,T}) \approx V(\sum_{j=0}^{T-1} \varepsilon_{t,j}) = TV(\varepsilon_1)$$

for $T \leq T_{max}$ and assuming that the correlation among the $S_t's$ is small as they are dominated by the iid ε_t, then

$$V(S^*_{t,T}) \approx T \cdot V(Y_t) = T \cdot V(Y_1).$$

We finally get that the improvement in the classical R^2 value

$$R^2 \approx \frac{\beta^2 T^2 V(X_1)}{T \cdot V(S_1)} = \frac{\beta^2 T V(X_1)}{V(Y_1)}$$

approximately linear for "small time horizons $T \leq T_{max}$. And note that this linear increase in R^2 value has nothing to do with an improved level of predictive power. It is so to speak a simple consequence of the law of large numbers on the error term. We therefore have explained empirically and theoretically that something not too far from a linear increase in the R^2 value as a function of the time horizon indeed is to be expected when the same level of accuracy is present in the estimation of the considered time horizons. This is important, since it gives a clear interpretation of results like the ones found in Campbell, Lo and Mackinlay (1997, p.269) or the ones we find on our Danish data in the next section. When analyzing the Danish data in the next section, we will see that with $X_t = \log(D/P)_{t-1}$ and S_t being our excess stock return, then we arrive at phenomena very close to the ones anticipated in this section. We will also see that the traditional R^2 value has to be replaced by another measure of prediction power that is based on the principle of validation and not, as the R^2 value, on the principle of goodness of fit.

5 Estimating and evaluating the power of prediction

In this section we enter the methodological question of finding a good estimator of prediction power, first we follow Campbell, Lo and Mackinlay (1997, p.269) and calculate R^2 for different prediction horizons. So, we consider the regression

$$S_{t+1} + ... + S_{t+T} = \alpha + \beta\delta_t + \epsilon_{t+T}, \tag{9}$$

where S_t is the 1-year log excess stock return from $t-1$ to t, and δ_t the log dividend-price ratio. The results are given in Table 3 together with the corresponding R^2 values.

If these reported R^2 values are good estimates of the R^2_A described in Section 5 above, then we can conclude that for the period $1922 - 1996$, predicting the time horizon T equal to 1 year gives the best power of prediction, namely corresponding to an improved prediction of the mean just above 3%. Measure in this way the improvement in prediction falls with increasing T. For T between 2 years and 4 years the power of prediction corresponds to between 2% and

	1922-1996			1947-1996		
horizon, T	β	$s(\beta)$	R^2	β	$s(\beta)$	R^2
1	0.080	6.3%	3.2%	0.116	7.9%	5.9%
2	0.157	8.2%	6.6%	0.217	10.9%	11.5%
3	0.233	9.6%	10.5%	0.308	12.0%	17.1%
4	0.331	11.3%	14.2%	0.423	12.6%	21.0%
5	0.364	12.7%	15.7%	0.438	13.6%	20.6%
6	0.382	14.8%	15.5%	0.502	17.8%	23.5%
7	0.343	14.2%	12.6%	0.465	17.2%	20.7%
8	0.273	14.1%	7.9%	0.406	15.4%	15.3%

Table 3: Predictability of T-year excess stock returns with model (9) on $\delta = \log$ D/P .

3% of the conditional mean. For longer time horizons the prediction power on the mean falls below 2%. For the period $1947 - 1996$ we have a similar picture, though prediction seems to me more accurate here. The predicting power corresponding to the time horizon T equal to 1 year gives the best power of prediction namely around 5% and the prediction power is decreasing with T from around 4% on the mean for T equal to 2 years down to between 2% and 3% for T equal to 5 years. Apparently we therefore should be able to make the convenient conclusion, that the optimal time horizon for prediction is 1 year. However, it is a well known fact from recent theory of mathematical statistics that these reported R^2 values are not good estimates for the corresponding $R_A^{2\prime}s$. They have some astonishing bad characteristics if they are used to select appropriate models for prediction. For example: the R^2 values are always increasing with complexity of the model. As a matter of fact it takes a quite clever model with a selective choice of the most important explanatory variables to beat even the simple mean in practical prediction, and complexity is one of the worst enemies of a good prediction.

To get some information on this last point consider a comparison of the regression fits for the parametric model and the more complicated nonparametric regression based on excess stock returns on inflation, interest rate, D/P and excess stock return, all one year lacked. For the nonparametric regression we used the local linear smoother with bandwidth equal to $2.0\sigma_X$, where σ_X is the vector of standard deviation of the regressors. For exact technical definitions, see Appendix 1. The quality of fit of these two models are given in Figure 1. Looking on these graphs, then a methodology considering R^2 values and the accuracy of fit would clearly select the more complex nonparametric method.

It is a known fact that adding new parameters to a parametric model gives almost always a bigger R^2. This lead to the introduction of adjusted R^2 and some other modifications, correcting for the number of estimated parameters. So some try to e.g. minimize the Akaike criteria or the prediction criteria of Amemiya (1985) to mention only some of them. But all these offered only reasonable corrections for parametric (linear) models and do further not solve the problem of “in–sample validation. Therefore we consider the more general out of sample criteria based on Cross Validation. This criteria will show that the prediction based on the four-dimensional nonparametric fit above is rather misleading and that it actually predicts outside the sample worse than a simple constant, while the linear predictor based on the dividend yield does have quite good predictive power. The problem using the complicated four-dimensional nonparametric model is overfitting and corresponds to the problem of overparameterization in parametric regression.

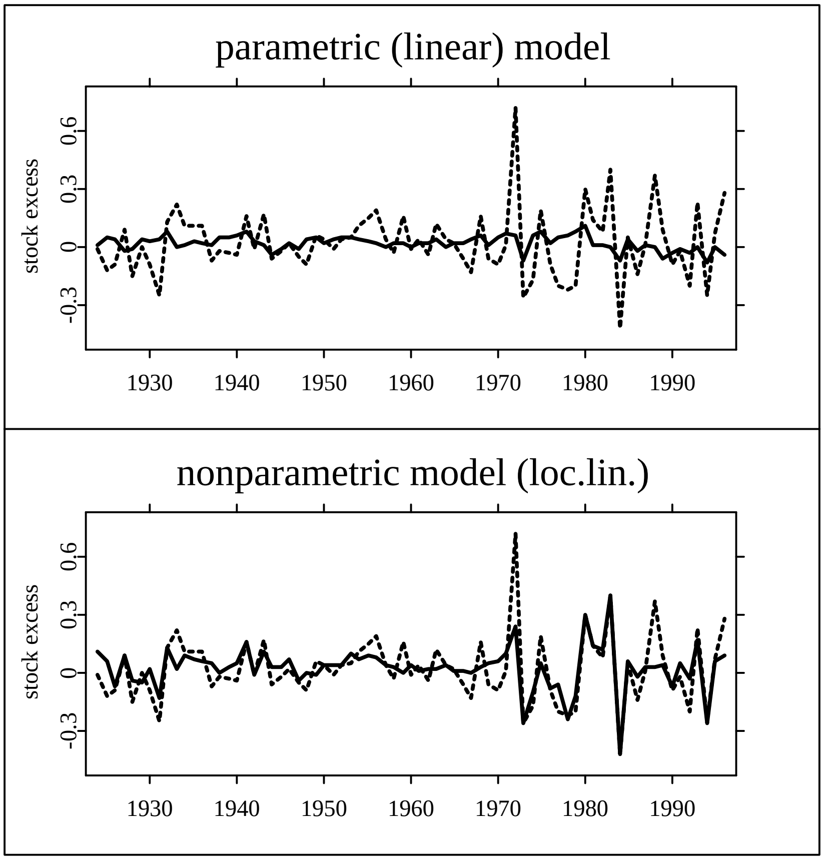

Figure 1: Parametric (upper) and nonparametric (lower) regression fit (solid lines) of stock returns on inflation, interest rate, D/P and excess stock return, all one year lacked. Dotted line is the real, observed data process.

The conclusion is that the R^2 values can be considered rather useless as evidence of prediction power. We define an appropriate replacement for the traditional R^2 value. The well known statistical method of Cross Validation is able to give us such a measure, see among many others Campbell, Lo and MacKinlay (1997, p.233). Below we introduce the validated measure of prediction, R_V^2, that is a reasonable estimator of the prediction error R_A^2 defined in Section 4 above.

For convenience of interpretation, the validated R_V^2 value is defined similarly to the traditional R^2. Recall first the expressions in (8). We do, however, replace the key components of the R^2 formula by its Cross Validation analogs, i.e.

$$\begin{aligned} CV - total\ Variation &= \sum_{i=1}^{n}(y_i - \bar{y}_{-i})^2, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad (10) \\ CV - not\ explained\ Variation &= \sum_{i=1}^{n}(\widehat{y}_{-i,i} - y_i)^2, \qquad (11) \end{aligned}$$

see Appendix 3. $\widehat{y}_{-i,i}$ is the prediction that we would construct for y_i based on our estimation methodology using all data except the i'th dependent and independent variables. This gives the validated measure of prediction:

$$R_V^2 = 1 - \frac{CV - not\ explained\ Variation}{CV - total\ Variation} \tag{12}$$

that is a data based estimator of the average prediction error R_A^2 described in Section 4. This estimator is a reliable measure for prediction power and it does not have the inherent weaknesses of the R^2 measure described above. We will for example often have that

$$\sum_{i=1}^{n}(y_i - \widehat{y}_{-i}(x_i))^2 > \sum_{i=1}^{n}(y_i - \bar{y}_{-i})^2$$

and consequently $R_V^2 < 0$. This just means that our model does even worse than taking the so far observed average when predicting Y. So the R_V^2 gives the percentage of how much better or worse the model does compared to taking the (so far observed) average, and we have $R_V^2 \in (-\infty, 1]$.

In Table 4 below we see the power of prediction of the the linear model based on the logarithm to the dividend-price ratio.

$log(D/P)$	1922-1996	1947-1996
horizon, T	R_V^2	R_V^2
1	-1.1%	-0.3%
2	2.2%	3.0%
3	4.6%	7.7%
4	7.4%	9.4%
5	6.5%	0.5%
6	5.2%	-19.5%

Table 4: Predictability of T-year excess stock returns on log D/P, model (9), evaluated with the R_V^2.

We see that the power of prediction is a lot less than the traditional R^2 values might suggest. For the time period $1922 - 1996$: While the linear model does worse than just estimating a constant mean for the time horizon $T = 1$, then Table 2 suggests that the power of prediction is between 1% and 2% measured on the mean for the time horizons T between 2 years and 6 years. The optimal time horizon for prediction seems to be 3 years or 4 years, where the power of prediction is closest to 2%, see Table 2. For the time period $1947 - 1996$: Also here the linear model does worse than just estimating a constant mean for the time horizon $T = 1$, and Table 2 suggests that the power of prediction is between 2% and 3% measured on the mean for the time horizons T between 2 years and 4 years. The model does not seem to predict well for the time horizons above 5 years.

In Table 4 we consider the same numbers based on the raw dividend price ratio without taking the logarithm. The R^2 and R_V^2 values based on this raw data are very similar but slightly better than the results presented in Table 3 above. For example, Table 3 suggests that the linear model does as a matter of fact predict more than 2% on the mean for the time horizon $T = 1$ for the period $1947 - 1996$.

Finally we consider the power of prediction by choosing the functional relationship between the dividend-price ratio and the return by a nonparametric kernel estimator. Since this functional

D/P	1922-1996			1947-1996		
horizon, T	β	R^2	R_V^2	β	R^2	R_V^2
1	2.384	3.8%	-0.2%	0.116	7.3%	1.3%
2	4.824	8.8%	4.7%	0.217	14.9%	8.2%
3	6.750	13.0%	7.8%	0.308	21.1%	14.2%
4	9.176	17.5%	10.3%	0.423	25.8%	16.0%
5	9.871	18.7%	10.3%	0.438	24.2%	9.5%
6	9.612	16.4%	6.9%	0.502	25.4%	-4.6%

Table 5: Predictability of T-year excess stock returns with model (9) on δ = D/P comparing classic R^2 with validated R_V^2.

relationship can be arbitrary, the above discussion on using the raw dividend price ratio or taking the logarithm is irrelevant. We get the following results:

D/P	1922-1996	1947-1996
horizon, T	R_V^2	R_V^2
1	-0.1%	3.3%
2	4.8%	10.8%
3	8.0%	16.7%
4	12.2%	20.5%
5	13.5%	22.6%
6	6.9%	17.8%

Table 6: Predictability of T-year excess stock returns using nonparametric models and evaluated with R_V^2. Explanatory variable was D/P.

When considering the period $1947-1996$, then data from the entire period, $1922-1996$, is used to fit the nonparametric functional relationship. The evaluation of the quality of the fit is, however, exclusively based on the data in the period $1947-1996$. While the nonparametric power of prediction for the period $1922-1996$ is already slightly better than the strictly linear power of prediction, we see a clear improvement of prediction power for the nonparametric method when considering the period $1947-1996$. Since the linear model over the entire period is enclosed as a special case of our nonparametric method, namely the special case corresponding to infinite bandwidth, the nonparametric selection method can point out the linear model as giving a better prediction than any other functional relationship. This does in fact happen for the time horizons 1 year, 2 years, 3 years and 6 years. We can therefore conclude that the greatest part of the improvement is due to the fact that we now have used data for the entire period for predicting the years $1947-1996$ instead of using only the data of that very period. For $1947-1996$ we get the astonishing prediction power corresponding to 4% on the mean for the horizon T equal to 2 years, see Table 2. For horizons $T=1, T=3, T=4$ the prediction power on the mean is above 3%. We therefore conclude that prediction of excess stock returns indeed seem possible and that the dividend-price ratio does have a significant role to play. We can also conclude that it does not seem to be impossible for the long term investor to decide whether the excess yield on stocks is positive or not. We also conclude that this prediction power can be obtained using dividend-price ratio information alone. The basis of this conclusions is that the prediction power described above is bigger than 2.1% on the mean for the period $1922 - 1996$ and it is bigger than 3.2% on the mean for the period $1947 - 1996$. This is indeed the case for all predictions with time horizon less than 4 years. In the next section we show that we can actually improve this prediction even more by including further information in our prediction.

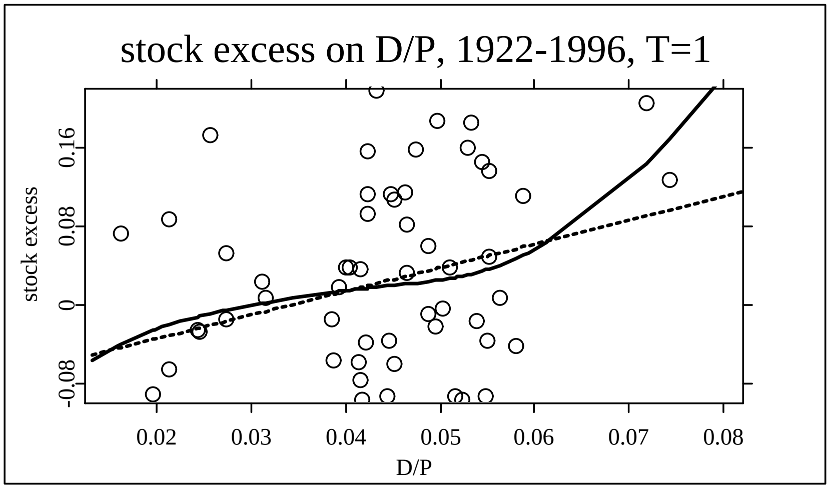

Figure 2: Parametric (dashed) and nonparametric (solid) regression fit of stock returns on D/P and real data points. Bandwidth= $2.4\sigma_X$

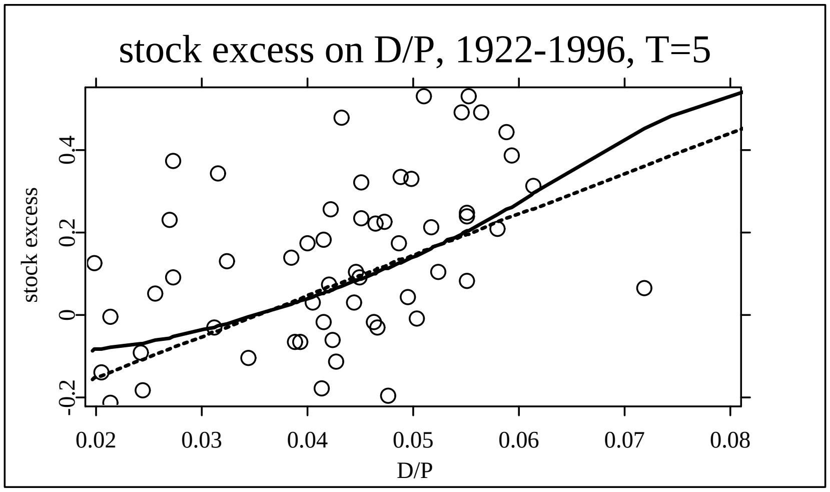

Figure 3: Parametric (dashed) and nonparametric (solid) regression fit of stock returns on D/P and real data points. Bandwidth= $3.4\sigma_X$

For a graphical visualization of the impact of the dividend-price ratio at excess stock returns, see Figure 2 and Figure 3 for respectively the one-year horizon and the five year horizon versions of the prediction of excess stock returns based on the dividend-price ratio. Both the parametric and nonparametric versions are shown. The graphs clearly indicate the impact of the dividend yield on future returns and we also see, that our current Danish level of the dividend-price ratio on around 1% is so low, that we indeed must conclude that according our predictive filter it is a dangerous time to invest in stocks and we should not expect the average excess return on stocks to match this danger. As of matter of fact our model predicts excess returns in the near future to have an average value below zero. So, it seems that the extra risk inherent in investments in stocks are not followed by a corresponding extra return on stocks in a situation with a general low level of dividend yields. As a consequence our advice to Danish long term investors is not to increase their percentage of stocks in their portfolio right now.

6 Looking for the right model

In this section we investigate the potential advantages that we can obtain by including other variables than just dividend divided by price in our prediction. Due to the complexity of the study of the section, we have chosen to restrict our investigation to a time horizon of one year. Based on the considerations given in Section 2, we have chosen to consider a time series regression problem of the following form:

$$S_t = g(S_{t-1}, D_{t-1}, I_{t-1}, r_{t-1}) + \epsilon_t \tag{13}$$

using the data described in Section 3. The full four-dimensional model corresponds to estimate the function g without any parametric assumptions nor assumptions of structure such as additivity or multiplicativity. This model is most often too complex for both to visualize and/or to predict well. The lack of prediction is due to the error of estimation rather than that the model is insufficient. Therefore we suggest some structure on g to predict well. We have chosen to consider additive models such that

$$g(s, d, i, r) = c + g_1(s) + g_2(d) + g_3(i) + g_4(r), \tag{14}$$

compare also Appendix 2, especially for estimation.

Furthermore we consider both the situation where the entering $g_i\prime s$ are nonparametric and the situation where all the entering $g_i\prime s$ are parametric and follow a linear model. In our study we consider three types of models with all combinations of subsets of (S_{t-1}, D_{t-1}, I_{t-1}, r_{t-1}). Namely (see above)

- ◊ Linear models
- ◊ Nonparametric additive models
- ◊ Fully nonparametric models.

The more complex the model is, the bigger the estimation error will be and the smaller the modeling error will be. To be able to choose among the entering models, we use the validated R_V^2 defined in Section 5. All in all, we have 41 models to consider, namely 15 linear models, 15 full models and 11 nonparametric additive models (leaving out the one-dimensional models that we counted among the full ones). As mentioned and explained in the appendices we always looked for the optimal bandwidths in the nonparametric procedures using Cross Validation.

Some first findings of the estimation respective model structure are the following:

Though the multidimensional nonparametric additive model reaches a positive R_V^2 for some of the considered models, the corresponding full model did always better. This is a clear indicator for having here a more complex structure than additivity. This is not surprising when we consider the complicated relationship between these variables as described in Section 2. From our calculated R_V^2 values we also concluded that the only linear model that does better than the simple constant is the linear model based on the dividend divided by price for the period $1947-1996$ as described in the sections before. However, best among all estimators is the fully nonparametric two-dimensional model based on dividend divided by price and lagged excess stock return. This two-dimensional model has a R_V^2 value of 1.16% for the period $1922-1996$ and 4.62% for the period $1947-1996$. For the time period this is much better than the negative values of the R_V^2 obtained in Section 5. For the time period $1947-1996$ we get a significant improvement from the 3.3% we obtained in Table 6. While 4.62% in R_V^2 value corresponds to a prediction accuracy of more than 4% on the mean, the 3.3% in R_V^2 value obtained in Table 6 corresponds to a prediction accuracy of about 3.5%, compare Table 1.

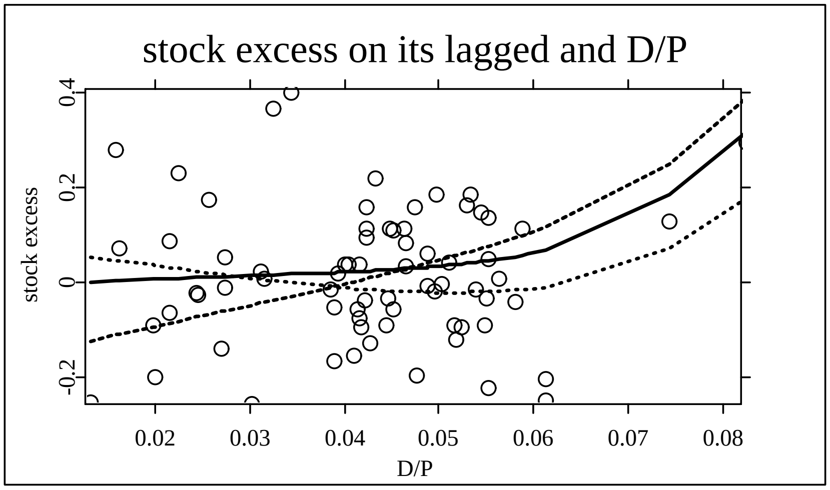

Figure 4: Nonparametric regression fit of excess returns on D/P and excess returns lagged at excess returns equal to -25% (dotted, starting above zero), 1% (solid), and 30% (dashed) in $1922-1996$. Bandwidth$= 3.8\sigma_X$

Once again have a look on the relation excess returns to dividend by price. In Figure 4 we see three slices from the two-dimensional predictive filter based on the dividend-price ratio and the lagged excess return of stocks. We plot the dependency on the dividend yield for three fixed values of excess returns: -25%, 1% and 30% corresponding to the lower 5% fractile, the median and the upper 95% fractile. We see a clear tendency of the excess stock return to be increasing with the dividend-price ratio and decreasing with last years excess return. Again, a clear indication based on this graph is that Danish investors should keep away for new investments in stocks, since they are just about to finish a magnificant year with a general Danish excess return on stocks above 30% resulting in a historical low dividend-price ratio of around 1%.

7 Appendix

Appendix 1. Local linear regression

In this appendix we give a brief insight into the algorithms of nonparametric flexible function regression. In particular we explain the local linear smoothing. The basic idea is to construct an estimator that lays a smooth surface (or hyperplane), e.g. in the one dimensional case a smooth line, into the point cloud that presents its functional form. The smoothness of that surface can be (pre-) determined by choosing a respectively large *smoothing parameter* (h), called bandwidth. Actually, often this parameter can also be data driven, see Appendix 3.

First, it is important to understand that this estimator works locally, e.g. we estimate the wanted function, the hyperplane, at each point we are interested in separately. This is, using the notation $E[Y|X=x]=m(x)$, $x \in I\!R^d$ having $(X_i, Y_i)_{i=1}^n$ observed and being interested in $m(x_0)$ for some point $x_0 \in I\!R^d$, we calculate $\widehat{m}(x_0)$. This is done by minimizing

$$\sum_{i=1}^{n} \left\{ Y_i - a_0 - a_1^T (X_i - x_0) \right\}^2 K_h(X_i - x_0) \qquad (15)$$

over $a_0 \in I\!R$, $a_1 \in I\!R^d$ and setting $\widehat{m}(x_0) = \hat{a}_0$. In equation (15) $K_h(v) = \prod_{j=1}^d \frac{1}{h} K(\frac{v_j}{h})$ is a $I\!R^d \to I\!R$ weight function. In our calculations we chose the so called quartic kernel, i.e. $K(u) = \frac{15}{16}(1-u^2)^2 1\!\!1\{|u| \le 1\}$. So we just use a weighted least squares estimator for linear regression that becomes a local estimator due to the weights K_h giving a lot of weight to points (X_i, Y_i) where X_i is close to x_0 but no weight to points far from x_0.

Here, in the weighting function comes the smoothing parameter h in: the larger h and consequently the environment with positive weighting, the smoother gets the resulting hyperplane whereas $h=0$ would be equivalent to interpolation of the Y_i's. Consistency, asymptotic theory and properties are well known and studied for the multivariate case in Ruppert and Wand (1994), for a general introduction see Fan and Gijbels (1996).

Remarks:

1. An often discussed question is how to choose bandwidth h in practice. As we are concerned about prediction, we take that bandwidth that is minimizing the "out of sample" prediction error using the Cross Validation measure, see Appendix 3. This is equivalent to maximizing our R_V^2. For more discussion of data driven bandwidth choice by Cross Validation in time series context, see e.g. Gyöfri, Härdle, Sarda, Vieu (1990).

2. The resulting vector $\hat{a}_1$ when minimizing equation (15) is a consistent estimate for the gradient $dm(x)/dx$. This can easily be understood when interpreting the expression $a_0 + a_1^T(X_i - x_0)$ as being the first terms of the Taylor approximation of $m(\cdot)$ around x_0 . Again, for more discussion see Fan and Gijbels (1996).

Appendix 2. Local linear additive regression

We speak of an additive model if the model $E[Y|X=x]=m(x)$, with $x \in I\!R^d$ is of the form

$$m(x) = c + \sum_{j=1}^{d} m_j(x_j), \quad \text{with } c = E[Y] \ \text{ and } E[m_j(X_j)] = 0 \qquad (16)$$

for identification. These models are quite popular thanks to its straight forward consequences in economic theory, interpretability (as only one dimensional functionals have to be considered), and some statistical properties as getting rid of problems in multidimensional smoothing ("curse of dimensionality", compare Stone, 1985). For the nonparmetric case, i.e. letting the additive components m_j arbitrary smooth functions, several procedures are known in the literature (see

Sperlich, 1998). In this article we focus only on the backfitting by Hastie, Tibshirani (1990). If $m(x)$ is really of additive form, this is a consistent and efficient procedure; if not, it still gives at least the projection on that additive model that fits the data best, for both see Mammen, Linton, Nielsen (1999). Actually, the backfitting tries to minimize $E[\{Y - m(X)\}^2]$ over all $m(\cdot)$ of additive form as in equation (16). This can be done by iteration; start with some initials $\widehat{m}_j^{[0]}(\cdot)$, $\hat{c} = \frac{1}{n}\sum_{i=1}^n Y_i$ and regress $Y - \hat{c} - \sum_{j \neq k}^d \widehat{m}_j^{[r-1]}(X_j)$ against X_k to get $\widehat{m}_k^{[r]}$ until the estimates do not differ from those yield in the last iteration.

For the regression the (one dimensional) local linear estimator, presented in Appendix 1 can be applied. This is exactly the procedure we did in our data analysis when modeling additively.

Remarks:

1. Certainly, there exist a growing amount of articles how to test additivity. But, a comparison of the Cross Validation values yield for the multidimensional local linear and the backfitting smoothing gives already an idea how far the *true model* is from additivity.

2. Bandwidths can again be chosen using Cross Validation, compare Appendices 3 and 1.

7.1 Appendix 3. Cross Validation

A typical question of interest, not only in prediction problems, is how to evaluate the different models. This concerns the model or variable selection as well as the bandwidth choice. In general, a natural way to evaluate an estimator is to look on the mean squared error or the expected squared difference between estimate and observation Y $E[\{Y - \widehat{m}(X)\}^2]$ which certainly itself can only be estimated. Additionally, as we speak about prediction, we would like to know how well the estimate works outside the considered sample. Both aspects are taken into account in the so called Cross Validation (CV) values, defined as

$$\text{CV - value} = \frac{1}{m}\sum_{l=1}^{m}\{y_l - \widehat{m}_{-l}(x_l)\}^2 \tag{17}$$

where $(X_l, Y_l)_{l=1}^m$ is the evaluation sample, e.g. can be the whole sample $(X_i, Y_i)_{i=1}^n$ itself, and $\widehat{m}_{-l}(x_l)$ the considered estimator evaluated at point x_l but determined without observation (x_l, y_l). This CV value is an approximation for the mean squared error (also for prediction) and a quite common used validation measure in nonparametric regression. For time series context and more references see e.g. Gyöfri, Härdle, Sarda, Vieu (1990).

Remark: It is important to eliminate always all information that is aimed to predict from the estimation of m. So, if we predict the increase of assets over a period of 4 years, the estimator $\widehat{m}_{-l}$ is calculated not only without the l^{th} observation but also without the three years before and after year l.

How can it be used for bandwidth or model selection? We give an example for bandwidth selection. we write $\widehat{m}$ as a function of the bandwidth ($\widehat{m}_h$) and look for that h that minimizes

$$\text{CV}(h) = \frac{1}{m}\sum_{l=1}^{m}\{y_l - \widehat{m}_{\mathbf{h},-l}(x_l)\}^2$$

This has been shown to give the optimal bandwidth in nonparametric regression; we refer again at Gyöfri et al.(1990). So the idea is always just to minimize the CV criteria and to take that model as the best that is minimizing equation (17).

References

Amemiya, T. (1985). *Advanced Econometrics.* Havard University Press, Cambridge.

Barber, B.M and O. Terrance (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *Journal of Finance* **55**, 773-806.

Campbell, J.Y. (1991). A variance decomposition for stock returns. *Economic Journal* **101**, 157-179.

Campbell, J.Y. and R.J. Shiller (1988). The dividend–price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* **1**, 195-228.

Campbell, J.Y., A.W. Lo and A.C. MacKinlay (1997). *The Econometrics of financial markets.* Princeton University Press, Princeton, New Jersey.

Engsted, T. and C. Tanggård (2000). The Danish stock and bond markets: Comovement, return predictability and variance decomposition. *forthcoming in Journal of Empirical Finance*

Fama, E. and K. French (1988). Dividend yields and expected stock returns. *Journal of Financial Economics* **22**, 3-25.

Fan, J. and I. Gijbels (1996). *Local polynomial regression.* Chapman and Hall, London.

Goetzmann, W.N. and P. Jorion (1993). Testing the predictive power of dividend yields. *Journal of Finance* **48**, 663-679.

Gordon, M. and P. Shapiro (1956). Capital Equilibrium analysis: The required rate of profit. *Management Science* **3**, 102-110.

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

Hastie, T.J. and R.J. Tibshirani (1990). *Generalized Additive Models.* Chapman and Hall.

Hodrick, R.J. (1992). Dividend yields and expected stock returns: Alternative procedures for inference and measurement. *Review of Financial Studies* **5**, 357-386.

Jagannathan, R. and N.R. Kocherlakota (1996). Why should older people invest less in stocks than younger people. *Federal Reserve Bank of Minneapolis. Quaterly Review, Summer 1996.* 11-20.

Lund, J. and T. Engsted (1996). GMM and present value tests of the C-CAPM: Evidence from the Danish, German, Swedish, and UK stock markets. *Journal of International Money and Finance* **15**, 497-521.

Mammen, E., O.B. Linton and J.P. Nielsen (1999). The existence and asymptotic properties of a backfitting projection algorithm under

Nielsen, J.P. and O.B. Linton (1998). An optimisation interpretation of integration and backfitting estimators for separable nonparametric models. *J. Roy. Statist. Soc., Ser. B.* **60**, 217-222.

Richardson, M. and J.H. Stock (1989). Drawing inferences from statistics based on multi-year asset returns. *Journal of Financial Economics* **25**, 323-348.

Ruppert, D. and M.P. Wand (1994). Multivariate locally weighted least squares regression. *Annals of Statistics* **22**, 1346-1370.

Siegel, J.J. (1998). *Stocks for the long run*, 2nd ed. New York: McGraw-Hill.

Shiller, R.J. (2000). *Irrational exuberrance.* Princeton University Press, Princeton, New Jersey.

Sperlich, S. (1998). *Additive Modelling and testing Model Specification.* Shaker Verlag, Aachen.

Stone, C.J. (1985). Additive regression and other nonparametric models. *Annals of Statistics* **13**, 685-705.

Wand, M.P. and M.C. Jones (1996). *Kernel Smoothing.* Chapman and Hall, London.

Wilkie, A.D. (1986). A stochastic investment model for actuarial use. *Transactions of the Faculty of Actuaries,* **39**, 341-403.

Wilkie, A.D. (1990). Modern portfolio theory - some actuarial problems. *Proceedings of the Afir Colloquium* **1**, 199-215.

Williams, J.B. (1938). *The Theory of Investment Value.* Harvard University Press, Cambridge.