A Multivariate Nonparametric Test for Return and Volatility Timing[1]

Wessel Marquering
Department of Financial Management,
Erasmus University Rotterdam,
F4 - 26, P.O.Box 1738, 3000 DR Rotterdam,
The Netherlands.
Tel.: +31-10-4082786.
Fax: +31-10-4089017.
E-mail: W.Marquering@fbk.eur.nl.

Marno Verbeek
Department of Financial Management and Econometric Institute,
Erasmus University Rotterdam,
F4 - 36, P.O.Box 1738, 3000 DR Rotterdam,
The Netherlands.
Tel.: +31-10-4082702
Fax: +31-10-4089017.
E-mail: M.Verbeek@fbk.eur.nl.
Erasmus University Rotterdam

July 5, 2004

[1]The authors would like to thank Frank De Jong, Hans Dewachter, Joop Huij, Christopher Polk, Keunkwan Ryu, Allan Timmermann and an anonymous referee for helpful comments and suggestions. We also wish to thank Allan Timmermann for providing the core dataset used in this paper. Address all correspondence to: Wessel Marquering, Department of Financial Management, Erasmus University Rotterdam, F4 - 26, P.O.Box 1738, 3000 DR Rotterdam, The Netherlands. Tel.: +31-10-4082786. Fax: +31-10-4089017. E-mail: W.Marquering@fbk.eur.nl.

Abstract

This paper develops a novel approach to simultaneously test for market timing in stock index returns and volatility. The tests are based on the estimation of a system of regression equations with indicator variables and provide detailed information about the statistical significance of alternative market timing components.

JEL classification: C14; C22; C52; C53; G12; G14.

Keywords: Nonparametric; Market Timing; Predictability of Stock Returns and Volatility; Realized Volatility.

1 Introduction

In this note we propose a regression-based approach to test for market timing in returns and volatility. This approach provides detailed information about alternative sources of market timing and allows one to examine the relationships between the performances of the return and volatility forecasts. The tests also allow to distinguish positive and negative timing ability. Because a standard multivariate test may reject when forecasts and realized values are negatively related, our approach is much more insightful. The approach is illustrated using forecasts for the monthly return on the S&P 500 index and its volatility, over the period 1966-2001.

2 Nonparametric Tests for Market Timing

Denote the excess return on a stock market index by $r^e_{mt} = r_{mt} - r_{ft}$, where r_{ft} denotes the (time-varying) riskless return. The variance of the excess return in period t is denoted by VOL_t. We shall assume that volatility is observed ex post. In recent empirical work, it is common to employ a measure of realized volatility, estimated on the basis of high frequency data. Further, we consider a series of excess return forecasts, denoted $\hat{r}^e_{m,t}$ and a series of volatility forecasts, denoted $\hat{VOL}_t$. These forecasts, generated using genuinely available information only, can be constructed in alternative ways.

Table 1: **Return Contingency Table**

This 2×2 contingency table is a cross-tabulation of the signs of $r^e_{m,t}$ and $\hat{r}^e_{m,t}$, obtained from recursive out-of-sample forecasts.

	$r^e_{m,t} \leq 0$	$r^e_{m,t} > 0$	Total
$\hat{r}^e_{m,t} \leq 0$	m_{11}	m_{12}	m_{10}
$\hat{r}^e_{m,t} > 0$	m_{21}	m_{22}	m_{20}
Total	m_{01}	m_{02}	m

To compare predicted and actual excess returns, Henriksson and Merton (1981) (HM henceforth) propose a nonparametric test based upon a cross-tabulation of the signs of $\hat{r}^e_{m,t}$ and $r^e_{m,t}$, as in Table 1. The HM test statistic, which is asymptotically standard normally distributed under the null hypothesis, is given by

$$HM = \frac{m_{11} - \frac{m_{10}m_{01}}{m}}{\sqrt{\frac{m_{10}m_{01}m_{20}m_{02}}{m^2(m-1)}}}, \tag{1}$$

where m_{11} is the number of correct bear market forecasts, m_{01}, m_{10} are the numbers of bear markets and bear market forecasts, respectively, while m_{02}, m_{20} denote the number of bull markets and bull market forecasts, respectively. The total number of

evaluation periods is m. The HM test is asymptotically equivalent to a one-tailed test on the significance of the slope coefficient α_1 in

$$I_{\{\hat{r}^e_{m,t}>0\}} = \alpha_0 + \alpha_1 I_{\{r^e_{m,t}>0\}} + \nu_t, \tag{2}$$

where $I_{\{.\}}$ denotes the indicator function. The above tests can also be applied to volatility forecasts, provided we define a positive benchmark, ξ say, to distinguish between positive and negative forecasting errors.

To analyze the joint forecasting performance of returns and volatility, a generalization of the contingency table extended to four dimensions can be used. In the 4×4 contingency table presented in Table 2, the diagonal cells represent the correctly predicted pairs. A test statistic in this more-dimensional case is the generalized Henriksson and Merton test statistic for a 4×4 contingency table, given by

$$HM = \sum_{i,j=1}^{4} \frac{\left(n_{ij} - \frac{n_{i0}n_{0j}}{n}\right)^2}{\frac{n_{i0}n_{0j}}{n}}, \tag{3}$$

where n_{ij} is the number of observations in the category (i, j), and n_{i0} and n_{0j} are the ith row and the jth column totals. Under the null hypothesis, the test statistic is asymptotically Chi-squared distributed with 9 degrees of freedom. The null hypothesis that is actually tested here is independence between forecasts and realizations, so that a rejection does not necessarily imply *positive* timing ability.

Table 2: **Return and Variance Contingency Table**

This 4×4 contingency table is a cross-tabulation of the signs of the pairs ($r^e_{m,t}$, $VOL_t - \xi$) and ($\hat{r}^e_{m,t}$, $V\hat{O}L_t - \xi$), obtained from recursive out-of-sample forecasts.

	$r^e_{m,t} \leq 0$, $VOL_t > \xi$	$r^e_{m,t} \leq 0$, $VOL_t \leq \xi$	$r^e_{m,t} > 0$, $VOL_t > \xi$	$r^e_{m,t} > 0$, $VOL_t \leq \xi$	Total
$\hat{r}^e_{m,t} \leq 0$, $V\hat{O}L_t > \xi$	n_{11}	n_{12}	n_{13}	n_{14}	n_{10}
$\hat{r}^e_{m,t} \leq 0$, $V\hat{O}L_t \leq \xi$	n_{21}	n_{22}	n_{23}	n_{24}	n_{20}
$\hat{r}^e_{m,t} > 0$, $V\hat{O}L_t > \xi$	n_{31}	n_{32}	n_{33}	n_{34}	n_{30}
$\hat{r}^e_{m,t} > 0$, $V\hat{O}L_t \leq \xi$	n_{41}	n_{42}	n_{43}	n_{44}	n_{40}
Total	n_{01}	n_{02}	n_{03}	n_{04}	n

To focus upon combinations of predicted signs and actual signs which are particularly interesting, we propose several alternative tests based upon auxiliary regressions. We define two sets of dummy variables. The first set includes four dummies corresponding to the different outcomes in the columns of Table 2. That is,

$$\begin{aligned}
I_{A,t} &= 1 \quad \text{if} \quad r^e_{m,t} \leq 0 \text{ and } VOL_t > \xi, \quad 0 \text{ otherwise,} \\
I_{B,t} &= 1 \quad \text{if} \quad r^e_{m,t} \leq 0 \text{ and } VOL_t \leq \xi, \quad 0 \text{ otherwise,} \\
I_{C,t} &= 1 \quad \text{if} \quad r^e_{m,t} > 0 \text{ and } VOL_t > \xi, \quad 0 \text{ otherwise,} \\
I_{D,t} &= 1 \quad \text{if} \quad r^e_{m,t} > 0 \text{ and } VOL_t \leq \xi, \quad 0 \text{ otherwise.}
\end{aligned} \tag{4}$$

The second set includes four dummies corresponding to the different row outcomes, and are denoted as $I_{\hat{A},t}$, $I_{\hat{B},t}$, $I_{\hat{C},t}$ and $I_{\hat{D},t}$, respectively. They are defined in a similar way replacing actual outcomes by predicted outcomes. Now consider the following set of equations

$$\begin{aligned} I_{A,t} &= \alpha_{10} + \alpha_{11} I_{\hat{A},t} + \alpha_{12} I_{\hat{B},t} + \alpha_{13} I_{\hat{C},t} + v_{A,t}, \\ I_{B,t} &= \alpha_{20} + \alpha_{21} I_{\hat{A},t} + \alpha_{22} I_{\hat{B},t} + \alpha_{24} I_{\hat{D},t} + v_{B,t}, \\ I_{C,t} &= \alpha_{30} + \alpha_{31} I_{\hat{A},t} + \alpha_{33} I_{\hat{C},t} + \alpha_{34} I_{\hat{D},t} + v_{C,t}, \\ I_{D,t} &= \alpha_{40} + \alpha_{42} I_{\hat{B},t} + \alpha_{43} I_{\hat{C},t} + \alpha_{44} I_{\hat{D},t} + v_{D,t}. \end{aligned} \tag{5}$$

To prevent perfect multicollinearity, in each equation the dummy variable corresponding to the forecast with two incorrect signs is omitted. The coefficients α_{jj}, $j = 1, 2, 3, 4$, correspond to the increase in the probability of observing a given pair of signs if the signs are predicted correctly, relative to the case where both signs are predicted incorrectly. Similarly, the coefficients α_{ij}, $i \neq j$, $j = 1, 2, 3, 4$, measure the increase in probability due to forecasting only one given outcome correctly.

The null hypothesis of independence between forecasts and realizations states that all partial slope coefficients in (5) are equal to zero. The corresponding Wald test is asymptotically equivalent to the Henriksson-Merton test in (3). However, the partial slope coefficients provide more detailed information about the forecasting performance. For example, positive timing ability requires the four diagonal coefficients α_{11} to α_{44} to be positive. A further desirable feature is that, in each of the equations, the dummy for the best pair of forecasts has a coefficient that is larger than the other two dummies. This means that the probability of a given pair of outcomes of the equations is larger if both signs are correctly predicted than if only one sign is correctly predicted. For the first equation, this requires that α_{11} is larger than both α_{12} and α_{13}.

The system of equations in (5) is singular, because $v_{A,t} + v_{B,t} + v_{C,t} + v_{D,t} = 1$ for each t. This means that the coefficients in one equation can be expressed as linear functions of those in the other three equations.[1] This means that either equation can be dropped in estimation with equivalent results.

3 Return and volatility timing for S&P 500

To illustrate the testing procedure suggested above, we apply it to the S&P 500 index over the period January 1966-August 2001. Forecasts for the S&P 500 excess return and its volatility were generated using linear regression models, estimated recursively over

[1] It can be shown that the following restrictions hold: $\alpha_{40} = 1 - \alpha_{10} - \alpha_{20} - \alpha_{30} - \alpha_{11} - \alpha_{21} - \alpha_{31}$; $\alpha_{42} = \alpha_{11} + \alpha_{21} + \alpha_{31} - \alpha_{12} - \alpha_{22}$; $\alpha_{43} = \alpha_{11} + \alpha_{21} + \alpha_{31} - \alpha_{13} - \alpha_{33}$; and $\alpha_{44} = \alpha_{11} + \alpha_{21} + \alpha_{31} - \alpha_{24} - \alpha_{34}$.

all months up to the most recent. The excess returns is forecasted from the 3-month Treasury bill, the price-earnings ratio, dividend yield, inflation, industrial production, the 12-month Treasury bill, monetary growth, and the commercial paper-Treasury yield spread. To avoid look-ahead bias, we include the financial variables with a one-month lag, and the macroeconomic variables with a two-month lag; see Marquering and Verbeek (2004) for more details about the data and the forecasting model.

In a similar fashion we consider a linear model for the logarithm of realized volatility, which is explained from the same set of variables and its own lag. The employed measure for realized volatility is

$$\sum_{i=1}^{N_t}(r_{i,t}-\overline{r_t})^2\left[1+2N_t^{-1}\sum_{j=1}^{N_t-1}(N_t-j)\hat{\phi}_t^j\right], \tag{6}$$

where N_t is the number of trading days in month t, $r_{i,t}$ the return on day i in month t, and $\overline{r_t}$ denotes the average daily return in month t. Further, $\hat{\phi}_t$ is the first-order autocorrelation coefficient estimated using daily returns within month t.[2] This expression follows French, Schwert and Stambaugh (1987) and Akgiray (1989) and allows for serial correlation in daily stock returns, for example because of infrequent or non-synchronous trading. The functional form implies that predicted volatilities are nonnegative by construction. The first forecasts we consider are for January 1970, so that the first 48 months of our sample period are used to estimate the initial forecasting model.

A formal comparison of the return and volatility forecasts with their realized values is obtained by the market timing tests discussed in Section 2. First, Table 3 presents the contingency table corresponding to the HM test for market timing in returns. This table is simply a cross-tabulation of the signs of $\hat{r}^e_{m,t}$ and $r^e_{m,t}$. Over the entire sample period, the sign of $r^e_{m,t}$ is predicted correctly in 209 out of 380 months, such that the proportion of correctly predicted signs, or "hit ratio", is 55.0%. The idea behind the HM test is that there is an indication of market timing if the sum of the (estimated) conditional probabilities of a correct forecast exceeds one. From the figures in Table 3, we see that the estimated probability of a correct forecast conditional on a down market is 0.56, and the probability of a correct forecast conditional on an up market is 0.54. Consequently, the sum of the conditional probabilities of a correct forecast equals 1.10, which exceeds unity, thus providing an indication of market timing ability. This is confirmed by the HM test statistic, which exceeds the one-sided 5% critical value of 1.64.

To compare predicted and actual volatilities, we choose a benchmark of $\xi = 0.0012$, corresponding to a monthly standard deviation of 3.46%. This number closely cor-

[2] The average estimated daily autocorrelation coefficient equals 0.12, and varies between -0.50 and 0.61.

Table 3: **Nonparametric Market Timing Test**

This 2×2 contingency table is a cross-tabulation of the signs of $r^e_{m,t}$ and $\hat{r}^e_{m,t}$. The p-value for the HM test is one-sided and based on the asymptotic standard normal distribution. Figures correspond to the period 1970:1–2001:8 ($T = 380$).

	$r^e_{m,t} \leq 0$	$r^e_{m,t} > 0$	Total
$\hat{r}^e_{m,t} \leq 0$	93	98	191
$\hat{r}^e_{m,t} > 0$	73	116	189
Total	166	214	380
Proportion of correctly predicted signs: 55.0%			
Henriksson-Merton test: 1.9808 ($p = 0.023$)			

Table 4: **Nonparametric Volatility Timing Test**

This 2×2 contingency table is a cross-tabulation of the signs of $VOL_t - \xi$ and $\hat{VOL}_t - \xi$, with $\xi = 0.0012$. The p-value for the HM test is one-sided and based on the asymptotic standard normal distribution. Figures correspond to the period 1970:1–2001:8 ($T = 380$).

	$VOL_t \leq \xi$	$VOL_t > \xi$	Total
$\hat{VOL}_t \leq \xi$	103	55	158
$\hat{VOL}_t > \xi$	76	146	222
Total	179	201	380
Proportion of correctly predicted signs: 65.5%			
Henriksson-Merton test: 5.9503 ($p = 0.0000$)			

responds to the sample median of actual volatility. The 2×2 contingency table is presented in Table 4, and provides a highly significant HM test statistic of 5.95, while the percentage of correctly predicted signs is 65.5%. The results indicate the presence of statistically significant volatility timing at a monthly frequency. Using daily data, Fleming, Kirby and Ostdiek (2001) also find significant volatility timing. While volatility is much more persistent at daily than at monthly frequencies, some statistically significant predictability remains at the monthly frequency.

Next, we consider the joint forecasting power of the two models. Table 5 presents the 4×4 contingency table and the HM test statistic corresponding to market timing in both moments. In this table the diagonal cells represent the correctly predicted pairs. Both signs are predicted correctly in 141 out of 380 cases (37.1%). The realization of the generalized HM test statistic is 57.48 (with a 5% critical value of 16.92), which clearly indicates a significant relationship between the predicted and realized pairs.

To evaluate the question whether the rejection of the test is due to positive timing ability, we explore several alternative tests based upon auxiliary regressions, as introduced in Section 2. This allows us to focus the tests upon combinations of predicted signs and actual signs which are particularly interesting. The estimates for the system

Table 5: **Nonparametric Market Timing Test for both Moments**

This 4×4 contingency table is a cross-tabulation of the pairs $(r^e_{m,t},\, VOL_t - \xi)$ and $(\hat{r}^e_{m,t},\, \hat{VOL}_t - \xi)$, obtained from recursive out-of-sample forecasts, and where $\xi = 0.0012$. The nonparametric test for market timing in first and second moment is a χ^2 test with 9 degrees of freedom. Figures correspond to the period 1970:1–2001:8 ($T = 380$).

	$r^e_{m,t} \leq 0$, $VOL_t > \xi$	$r^e_{m,t} \leq 0$, $VOL_t \leq \xi$	$r^e_{m,t} > 0$, $VOL_t > \xi$	$r^e_{m,t} > 0$, $VOL_t \leq \xi$	Total
$\hat{r}^e_{m,t} \leq 0$, $\hat{VOL}_t > \xi$	51	15	32	16	114
$\hat{r}^e_{m,t} \leq 0$, $\hat{VOL}_t \leq \xi$	12	15	15	35	77
$\hat{r}^e_{m,t} > 0$, $\hat{VOL}_t > \xi$	23	18	40	27	108
$\hat{r}^e_{m,t} > 0$, $\hat{VOL}_t \leq \xi$	14	18	14	35	81
Total	100	66	101	113	380
Proportion of correctly predicted pairs: 37.1%					
χ^2-test: 57.4815 ($p = 0.0000$)					

in (5), for the period January 1970 to August 2001, are presented in Table 6. For ease of interpretation we present the results for the full system, even though the system is singular.

The null hypothesis of independence implies that all partial slope coefficients in (5) are equal to zero and is soundly rejected ($p = 0.0000$). Note that the value of the Wald statistic based on (5) is close to that of the non-parametric one of 57.48, given in Table 5, which is no surprise given that these tests are asymptotically equivalent.

Note that three out of four diagonal elements are significantly positive, indicating positive timing abilities. If we jointly test the set of restrictions $\alpha_{11} = \alpha_{22} = \alpha_{33} = \alpha_{44} = 0$, we obtain a test statistic of 43.08, which is highly significant. As mentioned above, a desirable feature is that, in each of the equations, the dummy for the best pair of forecasts has a coefficient that is larger than the other two dummies. This means that the probability of a given pair of outcomes is larger if both signs are correctly predicted than if only one sign is correctly predicted. The most notable violation of this implication is that the estimate for α_{24} exceeds that of α_{22}, indicating that the model is not very well capable to time whenever simultaneously excess returns are negative and volatility is low ($r^e_{m,t} \leq 0$ and $VOL_t \leq \xi$), which – in this sample period – occurs in 21% of the months. For the other combinations the simultaneous timing is reasonably well. Further, we tested whether the probability of a correct return forecast is independent of the probability of a correct volatility forecast. This resulted in a test statistic of 0.877, which is insignificant for a standard normal distribution. Apparently, the fact that in a given month the sign of the excess return is predicted correctly does not increase or decrease the probability that the sign of volatility (relative to 0.0012) is predicted correctly in the same month (or vice versa). Finally, Pesaran and Timmermann (1995) suggest that the predictability of excess returns is larger at times when volatility is

Table 6: **Regression-Based Tests for Market Timing in both Moments**

The tests for market timing are based upon (5); t-statistics are given in parentheses, heteroskedasticity corrected t-statistics in square brackets. The predictive failure test is a χ^2 test with 9 degrees of freedom. Estimation results over the period 1970:1–2001:8.

Explanatory Variable	$I_{A,t}$	$I_{B,t}$	$I_{C,t}$	$I_{D,t}$
Constant	0.1778	0.1818	0.1899	0.1607
	(4.0186)	(4.7111)	(3.9270)	(3.7899)
	[4.3880]	[4.6657]	[4.2803]	[4.6066]
$I_{\hat{A},t}$	0.2597	−0.0479	0.0780	–
	(4.3716)	(−0.9041)	(1.2314)	
	[4.1793]	[−0.9455]	[1.2756]	
$I_{\hat{B},t}$	−0.0132	0.0207	–	0.2823
	(−0.2043)	(0.3576)		(4.2817)
	[−0.2267]	[0.3460]		[4.2689]
$I_{\hat{C},t}$	0.0141	–	0.1738	0.1019
	(0.2314)		(2.6801)	(1.6462)
	[0.2490]		[2.6406]	[1.8032]
$I_{\hat{D},t}$	–	0.0404	−0.0121	0.2615
		(0.7224)	(−0.1826)	(4.1163)
		[0.6869]	[−0.2013]	[4.1574]
Wald test: 53.9998 ($p = 0.0000$)				

high. From the above results, we can easily test independence of the sign of the return forecast and volatility. This results in a value of 2.290, which implies a clear rejection based on a standard normal distribution. Apparently, periods with larger shocks more than average correspond to periods with a correctly predicted up- or down-market.

4 Diagnostics and small sample properties

In this section we discuss the small sample properties of the proposed tests. Asymptotically the test statistics are normally or χ^2-distributed. However, as the regression models in (2) and (5) use forecasts, either based on rolling windows or expanding windows, we expect the forecasts to be serially correlated. Consequently, the errors in the equations will not be i.i.d. and the properties of the test statistics must be further examined. Table 7 presents some diagnostics statistics of the residuals of equation (2) applied to returns and volatility. Not surprisingly, the Jarque-Bera test for normality strongly rejects the null hypothesis for both equations, which can be attributed to low kurtosis. The p-values corresponding to the White test for hetereskedasticity suggests no violation of homoskedastic errors at the usual significance levels. Based on the Ljung-Box statistics, serial correlation is highly significant at various lags. These tests confirm that the errors in the test equations are not i.i.d. and we will examine the consequences

of this below.

Table 7: **Diagnostics Residuals Model (2)**

	Return	Volatility
Skewness	-0.2005	-0.1947
Kurtosis	1.0484	1.3775
Jarque-Bera (p-value)	0.0000	0.0000
White (p-value)	0.7227	0.1843
$Q(1)$ (p-value)	0.0000	0.0000
$Q(5)$ (p-value)	0.0000	0.0000
$Q(10)$ (p-value)	0.0000	0.0000

Notes: This table gives descriptive statistics for the residuals of equation (2), $Q(r)$ denotes the Ljung-Box test statistic for rth order serial correlation in the residuals.

To investigate the small sample properties of the bivariate timing test we carried out a small Monte Carlo experiment. Since the serial correlation of the errors is of main concern, we examine whether the results are robust to dependence between subsequent forecasts. To do so, we generate 380 (the actual number of observations) monthly returns and sign forecasts. The returns are randomly drawn from a binomial distribution with $p = 0.5$, indicating positive or negative excess returns. Forecasts are generated conditional on the previous forecast to incorporate serial dependence. Denoting the probability that a forecast has the same sign as the previous forecast by pr, we employ values ranging from 0.5 (no dependence) to 0.95, with an interval of 0.05.[3] For each value of pr, we run 10,000 simulations and compute the HM test statistic. From the resulting sampling distribution we determine the probability that the test statistic exceeds the critical value of 1.64. Figure 1 shows the relation between the (Monte Carlo) p-values (based on 10,000 replications) and the degree of serial correlation in the forecasts (pr). We find that the p-values are close to 5 percent for all simulations, and that there is no clear relation between the p-value and the degree of serial dependence. Consequently, the actual size is very close to the nominal size (5%).

Table 8 presents the diagnostics statistics of the residuals of the regressions in (5). The p-values corresponding to the Ljung-Box statistics for serial dependence again show that serial correlation is highly significant at various lags, although somewhat less pro-

[3] More specifically, we generate sign predictions using the following first-order Markov process:

$$I_{\{\hat{r}^e_{m,t}\}} = int\left[\frac{I_{\{\hat{r}^e_{m,t-1}\}} - 0.5}{0.5} * (pr - 0.5) + x\right],$$

where int rounds to the nearest integer, and x is a random draw from the uniform (0,1) distribution.

Table 8: **Diagnostics Residuals Model (5)**

	A	B	C	D
Skewness	1.0206	1.6320	1.0763	0.7392
Kurtosis	2.4761	3.7146	2.3129	1.8539
Jarque-Bera (p-value)	0.0000	0.0000	0.0000	0.0000
White (p-value)	0.0832	0.3723	0.0202	0.1940
$Q(1)$ (p-value)	0.0252	0.5182	0.1367	0.1750
$Q(5)$ (p-value)	0.0364	0.0000	0.0006	0.0040
$Q(10)$ (p-value)	0.0032	0.0000	0.0003	0.0065

Notes: This table gives descriptive statistics for the residuals of equations (5), where A, B, C, and D refer to the first, second, third and fourth equation, respectively. $Q(r)$ denotes the Ljung-Box test statistic for rth order serial correlation.

nounced than in the bivariate case. Again, the errors will not be i.i.d. and we will examine the consequences of this using a small Monte Carlo experiment. For the multivariate model, we generated 380 monthly volatilities and volatility forecasts in a similar way as above, but imposing that actual volatility is serially correlated (with $pr = 0.69$, corresponding to the sample moment). We perform 10,000 simulations to estimate the probability that the test statistic (testing the null that all nine partial slope coefficients are zero) exceeds the critical value of 16.92. The results are displayed in Figure 2, where the serial dependence in the forecasts ranges from 0.5 (no dependence) to 0.95. From the figure we see that the multivariate tests are somewhat oversized: actual sizes are greater than the nominal size of 5%. Further, the Monte Carlo p-values increase with the degree of dependence in the forecasts. For very high levels of dependence, the

Figure 1: Monte Carlo p-values of univariate timing test

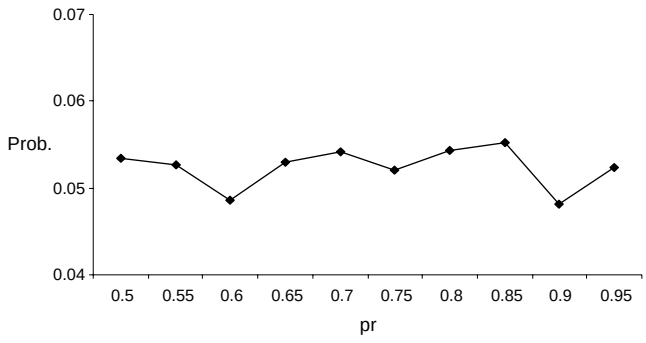
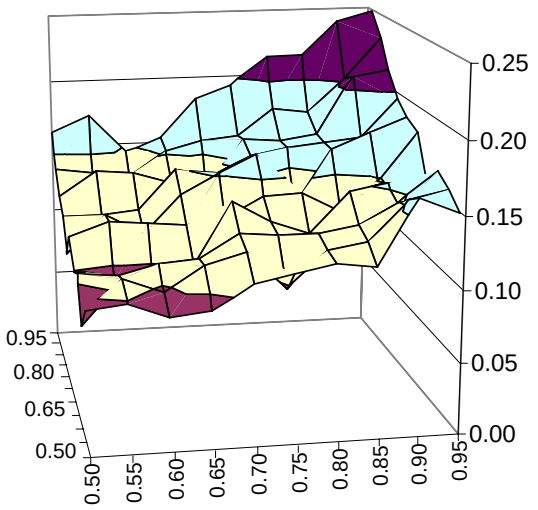


Figure 2: Monte Carlo p-values of multivariate timing test



Monte Carlo p-value is close to 0.25. Note that the values of pr for return and volatility forecasts in our sample are 0.6 and 0.7, respectively. While these Monte Carlo results suggest that the multivariate test may reject somewhat too often when based on serially correlated forecasts, the empirical findings in the previous section are still highly significant. Overall, the results of these Monte Carlo experiments suggest that the test statistics are reasonably robust to serial dependence in the forecasts.

5 Concluding Remarks

In this paper we proposed a simple framework for simultaneously evaluating return and volatility forecasts, using non-parametric and regression-based market timing tests. Employing a measure of realized volatility, the approach allows us to test for predictability in returns and volatility simultaneously, as well as for several alternative forms of dependence between returns and volatility (forecasts). The tests require no assumptions concerning "optimal" portfolio weights that can be derived from these forecasts, unlike the approaches in, e.g., Fleming, Kirby and Ostdiek (2001) and Marquering and Verbeek (2004).

We illustrated the testing framework by investigating out-of-sample forecasts for both returns and volatility of the S&P 500 index, based on recursive regression models. Over the period 1970 – 2001, all tests indicate statistically significant market timing in both returns and volatility. The joint test for the presence of market timing in both

moments indicates positive timing ability in the return and volatility pairs. Further the results indicate that there is no systematic relationship between the quality of the return and volatility forecasts. That is, it is not the case that a good return forecast typically corresponds to a bad volatility forecast or vice versa. However, we do find that the predictability of returns is larger in times when volatility is high. The results are robust to the relatively high dependence in return forecasts.

References

[1] Akgiray, V. (1989), "Conditional Heteroscedasticity in Time Series of Stock Returns: Evidence and Forecasts", Journal of Business, 62, 55-80.

[2] Fleming, J., C. Kirby and B. Ostdiek (2001), "The Economic Value of Volatility Timing", Journal of Finance, 56, 329-352.

[3] Fleming, J., C. Kirby and B. Ostdiek (2003), "The Economic Value of Volatility Timing Using "Realized" Volatility", Journal of Financial Economics, 67, 473-509.

[4] French, K.R., G.W. Schwert and R.F. Stambaugh (1987), "Expected Stock Returns and Volatility", Journal of Financial Economics, 19, 3-29.

[5] Henriksson, R.D. and R.C. Merton (1981), "On Market Timing and Investment Performance. II. Statistical Procedures for Evaluating Forecasting Skills", Journal of Business, 54, 513-533.

[6] Marquering, W. and M. Verbeek (2004), "The Economic Value of Predicting Stock Index Returns and Volatility", Journal of Financial and Quantitative Analysis, 39, 407-429.

[7] Pesaran, M.H. and A.G. Timmermann (1995), "Predictability of Stock Returns: Robustness and Economic Significance", Journal of Finance, 50, 1201-1228.

[8] Pesaran, M.H. and A.G. Timmermann (2000), "A Recursive Modelling Approach to Predicting UK Stock Returns", Economic Journal, 110, 159-191.