FEB22009-16: Financial Case Study

Predicting Aggregate Stock Returns with Short Interest and Forecast Combinations

Xiao Xiao and **Francine Gresnigt**

Department of Econometrics, Erasmus University Rotterdam

Email: xiao@ese.eur.nl and gresnigt@ese.eur.nl

1 Introduction

Stock return predictability remains an issue of intense debate. On the one hand, studies such as Welch and Goyal (2008) argue that over long time spans no single predictor variable can outperform the historical mean return in terms of forecast accuracy. On the other hand, several papers report forecast improvements upon the historical mean return. For instance, Rapach et al. (2010) show that despite the inferior performance of individual predictive regression model, combinations of individual models are able to deliver forecasts that are both statistically as economically superior relative to the historical average. Moreover, in a recent study by Rapach et al. (2016) it is shown that even a single predictor is able to generate positive out-of-sample R^2, in contrast to the results in Campbell and Thompson (2008) and Welch and Goyal (2008). They show that short interest, aggregated across securities, is arguably the strongest predictor of aggregate stock returns. Short interest ratio for a public company is the ratio of tradable shares being shorted to the total shares in the market, which can be viewed as a measure of market pessimism. It outperforms a host of popular return predictors both in sample and out of sample, such as dividend-price ratio, earnings-price ratio, book-to-market

ratio, default yield spread, etc. Their evidence indicates that short sellers are informed traders who anticipate future aggregate cash flows and associated market returns.

In this project, we first compare the information content of the short interest in predicting aggregate stock returns with other predictors. Hereafter, we explore whether the forecast performance can be further improved using forecast combinations that include both short interest and other variables. Recent studies including Aiolfi and Favero (2005), Elliott and Timmermann (2008), and Rapach et al. (2010) demonstrate that superior return forecasts may be obtained by combining forecasts from several different models. In particular, Rapach et al. (2010) find that a simple equally-weighted average of forecasts based on a single predictor outperforms the historical mean as well as each of the individual forecasts. While the equally-weighted average of forecasts provides a useful benchmark, we may also consider more advanced forecast combination schemes. Furthermore, in addition to examining the statistical performance of various forecast combination schemes, an important component of this project is to assess their economic value, when these forecasts are used in actual investment strategies.

2 Predictive regression models

We consider one-period ahead forecasts of excess stock returns based on linear regression models. The predictive regression model takes the form:

$$r_t = \beta_0 + \beta x_{t-1} + \epsilon_{t-1},$$

where r_t is the S&P500 log excess return in period t and x_t is the predictor variable. We are interested in testing the significance of β and the in-sample R^2 in the regression.

To examine the robustness of the in-sample results, we are also interested in the results for out-of-sample tests of return predictability. Such tests are important in light of Welch and

Goyal (2008), who show that the in-sample predictive ability of a variety of plausible return predictors generally does not hold up in out-of-sample tests. Corresponding to each predictor, we compute a predictive regression forecast as

$$\hat{r}_t = \hat{\beta}_0 + \hat{\beta} x_{t-1}. \tag{1}$$

where $\hat{\beta}_0$ and $\hat{\beta}$ are the OLS estimates of β_0 and β. The estimates of β_0 and β_1 should be based on data up to and including time $t-1$, so that it actually is an 'honest' out-of-sample forecast.

The out-of-sample R^2 is calculated using the definition in Campbell and Thompson (2008):

$$R^2_{OS} = 1 - \frac{\sum_{t=1}^{T}(r_t - \hat{r}_t)^2}{\sum_{t=1}^{T}(r_t - \bar{r}_t)^2}$$

where $\hat{r}_t$ is the fitted value from a predictive regression estimated through period t-1, and $\bar{r}_t$ is the historical average return estimated through period t-1.

After evaluating the predictive performance of short interest and other variables, we may employ model averaging or forecast combinations to construct a weighted average of forecasts from a range of models of the form in Equation (1) with different variables in x_t and z_t, that is, we may consider a forecast of the form:

$$\hat{r}_t^{(c)} = \omega_{1,t}\hat{r}_t^{(1)} + \omega_{2,t}\hat{r}_t^{(2)} + \ldots + \omega_{n,t}\hat{r}_t^{(n)} \tag{2}$$

where $\hat{r}_t^{(c)}$ denotes the combined forecast, $\hat{r}_t^{(i)}$ denotes the forecast obtained from model i and n is the number of models considered.

Two key issues in the implementation of a forecast combination approach are (i) the choice of individual forecasts $\hat{r}_t^{(i)}$ and (ii) the weighting scheme $\omega_{i,t}$. Concerning the first issue, we may use the individual forecasts or the forecasts of models that contain more variables, see Elliott et al. (2013). Concerning the second issue, simple forecast combination schemes (such as the equally-weighted average) often work quite well, but sometimes further improvement is possible by employing a more advanced forecast combination scheme, such as weighting schemes

based on past forecasting performance, in terms of mean squared prediction error (MSPE) or some alternative measure of forecast accuracy. The details of how to combine forecast estimates are discussed in Rapach et al. (2010).

3 Data

We aim to predict monthly excess returns on the Standard & Poor's (S&P) 500 index. The sample period runs from January 1973 until December 2014.

In the Excel table, we have the data the variable of our main interest, the short interest, as used by Rapach et al. (2016). They construct an aggregate short interest series using firm-level short interest data from Compustat, covering a variety of asset classes, including common equities, ADRs, ETFs, and REITs. The raw short interest numbers from Compustat are reported as the number of shares that are held short in a given firm. They normalize these numbers by dividing the level of short interest by each firm's shares outstanding from CRSP. The aggregate short interest is the equal-weighted mean of all asset-level short interest data (EWSI). The value-weighted short interest data (VWSI) is also provided.

The Excel table contains the relevant variables which can be used to calculate the commonly accepted monthly predictor variables:

1. Log dividend-price ratio (DP): log of a twelve-month moving sum of dividends paid on the S&P 500 index minus the log of stock prices (S&P 500 index). $DP = \log(D12) - \log(Index)$

2. Log dividend yield (DY): log of a twelve-month moving sum of dividends minus the log of lagged stock prices $DY = \log(D12) - \log(Index_{t-1})$.

3. Log earnings-price ratio (EP): log of a twelve-month moving sum of earnings on the S&P

500 index minus the log of stock prices. $EP = \log(E12) - \log(Index)$

4. Log dividend-payout ratio (DE): log of a twelve-month moving sum of dividends minus the log of a twelve-month moving sum of earnings. $DE = log(D12) - log(E12)$

5. Excess stock return volatility (RVOL): computed using a twelve-month moving standard deviation estimator, as in Mele (2007). $RVOL = \sqrt{svar}$

6. Book-to-market ratio (BM): book-to-market value ratio for the DJIA.$BM = b/m$

7. Net equity expansion (NTIS): ratio of a twelve-month moving sum of net equity issues by NYSE-listed stocks to the total end-of-year market capitalization of NYSE stocks. $NTIS = ntis$

8. Treasury bill rate (TBL): interest rate on a three-month Treasury bill (secondary market). $TBL = tbl$

9. Long-term yield (LTY): long-term government bond yield. $LTY = lty$

10. Long-term return (LTR): return on long-term government bonds. $LTR = ltr$

11. Term spread (TMS): long-term yield minus the Treasury bill rate. $TMS = lty - Rfree$

12. Default yield spread (DFY): difference between Moody's BAA- and AAA-rated corporate bond yields. $DFY = AAA - BAA$

13. Default return spread (DFR): long-term corporate bond return minus the long-term government bond return.$DFR = corpr - ltr$

14. Inflation (INFL): calculated from the CPI for all urban consumers. $INFL = infl$

Finally, we measure the market excess return as the log return on the S&P 500 index minus the log return on a one-month Treasury bill.

4 Research question and related issues of interest

The key research questions to be addressed in this project are:

1) Compared to other predictor variables, how is the forecasting performance of short interest in predicting aggregate stock returns?

2) Can forecasting combinations using short interest and other variables further improve the forecasting performance, both statistically and economically?

Several other issues/questions also deserve attention:

1) Make sure that you avoid 'look-ahead' bias – that is, for predicting the excess return in period t+1, make sure that only historically available information up to and including period t is used for estimating the coefficients in the predictive regressions and for determining the forecast combination weights.

2) Is it better to use an expanding window or a moving window for estimation of the predictive regression models? This question is related to the issue of structural breaks in the relation between excess returns and predictor variables. Intuitively, in case these relations are stable, it is best to use all available information for estimation; however in case of structural breaks a moving window approach may be better. If a moving window approach is adopted, an important choice that has to be made is the length of the moving window.

5 Econometrics Software

Matlab is preferred in this project. The reason is that David Rapach provides complete dataset and Matlab programs for the paper Rapach et al. (2016) on his website, which is also

available on Blackboard. It is straightforward to replicate their tables using their program and convenient to build extensions based on their program. The additional functions which are used in Rapach's program can be downloaded from the following website: https://nl.mathworks.com/matlabcentral/fileexchange/45093-time-frequency-generalized-phase-synchrony-for-eeg-signal-analysis?focused=3805362&tab=function Other software are also acceptable for this project.

6 Road map for this project

1) Use summary statistics to get an overview of the dataset.

2) Replicate Table 3 (in-sample predictive regression) and Table 5 (out-of-sample test) in Rapach et al. (2016). If you use the Matlab program provided by David Rapach, make sure you understand the complete code that generates these two tables. Results for 1 month are mandatory.

To make sure that you do not simply copy the tables in the paper, we would like you to provide additional results with some variations. You can choose **at least two variations** from the following options: **(a)** Report a different horizon (e.g. 2 months) **(b)** Split the sample to two parts **(c)** Use a different in-sample and out-of-sample period. **(d)** Provide multi-variate regression results both in-sample and out-of-sample.

3) Apply forecast combination methods and analyze whether the predictive ability can be further improved. Make the implementation choices (such as the length of moving window, etc) with proper justification.

4) Evaluate the economic value of the predictability from an asset allocation perspective in terms of utility gain or Sharpe ratio. Details can be found in Campbell and Thompson (2008)

and Rapach et al. (2010). Compare the economic value of predictability for single predictors and combined predictors.

5) Bonus is possible if other relevant econometric techniques are implemented, comprehensive analysis are included or more economic insights are provided in the report.

7 Research Proposal

The research proposal should be handed in before 17:00 on Wednesday, June 7th. You should include the econometric techniques that you intend to use and how you plan to solve the research question in the proposal with maximum 2 pages.

The meeting for discussing the proposal is scheduled from **13:00 to 17:00 on Thursday, June 8th**. The table in the next page shows the schedule. Each group member has to be present. On June 13th, we organize the office hour from **9:00-11:00 in H10-31**. Not every group member has to be present.

8 Final Report

The results of this assignment have to be reported in a scientific report with similar requirements as in the previous two cases. The report should be written in a logic structure. The content must be self-contained with enough model and implementation details to understand your results. Make sure you make discussions on your results and include a general conclusion. You have to hand in your final report via Blackboard before June 16th before 17:00. Late reports will not be graded.

Schedule for the proposal meeting on June 8th

Time/Room	H11-8	H6-1	H11-32	H8-28
13:00	Team 1	Team 11	Team X1	Team 21
13:15	Team 2	Team 12	Team X2	Team 22
13:30	Team 3	Team 13	Team X4	Team 23
13:45	Team 4	Team 14	Team X5	Team 24
14:00	Team 5	Team 15	Team X6	Team 25
14:15	Team 6	Team 16	Team X7	Team 26
14:30	Team 7	Team 17	Team X8	Team 27
14:45			Team X9	Team 28
15:00	Team 8	Team 18	Team X10	Team 29
15:15	Team 9	Team 19	Team X11	Team 30
15:30	Team 10	Team 20	Team X12	Team 31
15:45	Team S9	Team S12	Team X13	Team 32
16:00	Team S10	Team S13	Team X14	Team X3
16:15	Team S11	Team S14	Team X15	Team S15
16:30			Team X16	Team S16
16:45			Team X17	Team S17

References

Marco Aiolfi and Carlo A Favero. Model uncertainty, thick modelling and the predictability of stock returns. *Journal of Forecasting*, 24(4):233–254, 2005.

John Y Campbell and Samuel B Thompson. Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies*, 21(4):1509–1531, 2008.

Graham Elliott and Allan Timmermann. Economic forecasting. *Journal of Economic Literature*, 46(1):3–56, 2008.

Graham Elliott, Antonio Gargano, and Allan Timmermann. Complete subset regressions. *Journal of Econometrics*, 177(2):357–373, 2013.

David E Rapach, Jack K Strauss, and Guofu Zhou. Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *Review of Financial Studies*, 23(2):821–862, 2010.

David E Rapach, Matthew C Ringgenberg, and Guofu Zhou. Short interest and aggregate stock returns. *Journal of Financial Economics*, 121(1):46–65, 2016.

Ivo Welch and Amit Goyal. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4):1455–1508, 2008.