technische universität
dortmund

Case Studies

Winter 2023

Project I
Forecasting the equity premium: Approach using ordinary least squares

Ly Le Thi
November 17, 2023

in collaboration with Alexander Langnau, Sudhir Pratap Singh Rathore, and Jonathan Albert Karras

Lecturers: Prof. Matei Demetrescu
Dr. Paul Navas Alban

Contents

1 Introduction

Excess returns are defined as the return obtained from the return on stock (or portfolio of stocks) above the risk-free rate, which is usually calculated based on the yield of the latest short-term government treasury bill. The prediction of excess stock returns has been a central theme in financial economics. Accurate forecasts of excess stock returns are important for investors, policymakers, and financial analysts as they impact investment decisions and portfolio management strategies. In this paper, we go deeper into stock excess return forecasting by looking at different forecasting models, and we aim to evaluate the performance of these models relative to historical average excess stock returns.

The motivation for this research stems from Goyal and Welch's project (Welch and Goyal, 2008), where they contend that the historical average excess stock return outperforms predictive regressions that use a multitude of predictor variables. Their findings lead to the reexamining of the effectiveness of complex predictive models in forecasting excess stock returns and highlight the potential of simple historical average-based forecasts. In this study, we explore the predictive power of various models relative to the historical average return.

To achieve this goal, the Ordinary least squares (OLS) regression is employed to produce forecasts of excess stock returns. These forecasts will be generated using lagged predictor variables and investigated for different time frequencies, including monthly, quarterly, and yearly series within the dataset. For historical time series, we will fit autoregressive models with varying lags p (AR(p)) and select the optimal lag value using Akaike Information Criteria (AIC). Additionally, we establish linear predictive models, first for each predictor individually and then using multiple predictors in the model. Best subset selection and step-forward selection are also applied to select the best subset of predictors with the support of AIC. The quality of these forecasts is assessed using the mean squared forecast error (MSFE) as an evaluation metric, providing a comprehensive measure of forecast accuracy. The results show that AR(1) for monthly, AR(4) for quarterly, and AR(2) for annual data is an appropriate choice in which AR(1) works better than other OLS regression models for monthly data, while OLS model with the combination of all predictors works better than historical models and other combinations of the predictors in term of MSFEs.

In addition to this introduction, the report comprises four additional sections. The second section of this report provides a description of the dataset. The following section describes the different statistical methods and software used in this report, such as the ordinary least squares, autoregressive processes, mean squared forecast error (MSFE), and AIC. The fourth section of this report states all the results that are obtained using the statistical methods mentioned in the third section of the report and discusses the limitations of the findings. Finally, the fifth section summarizes all the results.

2 Data

2.1 All predictors

The dataset used in this project is taken from Welch Goyal's project which is provided on the website $https : //sites.google.com/view/agoyal145$. The original multivariate dataset contains a total of twenty-two time series variables which have been calculated in different frequencies (monthly, quarterly, and annually). For the purpose of this research, we drop three variables from the original dataset. Firstly, the two series $CRSP - SPvw$ and $CRSP - SPvwx$ since they are just stock returns series computed by Welch Goyal's project and will not be used as predictors for our target variable, excess returns time series. We also remove the series Cross-Sectional Premium since more than half the data of this series is the "Not a Number" value and we only have the data from 1947 to 2003, not untill the end period of this research 2022. The final dataset which is used for this research then contains nineteen variables and all variables are plotted in Figure 1

The monthly dataset contains fifteen series variables with 1824 data points starting from 1871-01 and ending in 2022-12. For the quarterly dataset, four extra variables were added compared to monthly data with a total of 608 data points from the first quarter of 1871 to the last quarter of 2022. Analogously, in the annual dataset, three extra variables were added compared to monthly data with a total of 152 data points from 1871-2022. We can see that the time series S&P 500 index, Dividend 12, Earning 12, Dividends-3-Year period, and Earnings-3-Year period have an increasing trend since they are the cumulative sum of the dividends (or earnings) over the measures period. The time series Treasure-Bills, Corporate Bond Returns (including AAA and BAA), Long-Term Yield, and Risk-free rate all have increasing trends from 1950 and reach the peak in the early 1980s then decrease until the end of the research period (2022). The other variables have large volatility in the research period.

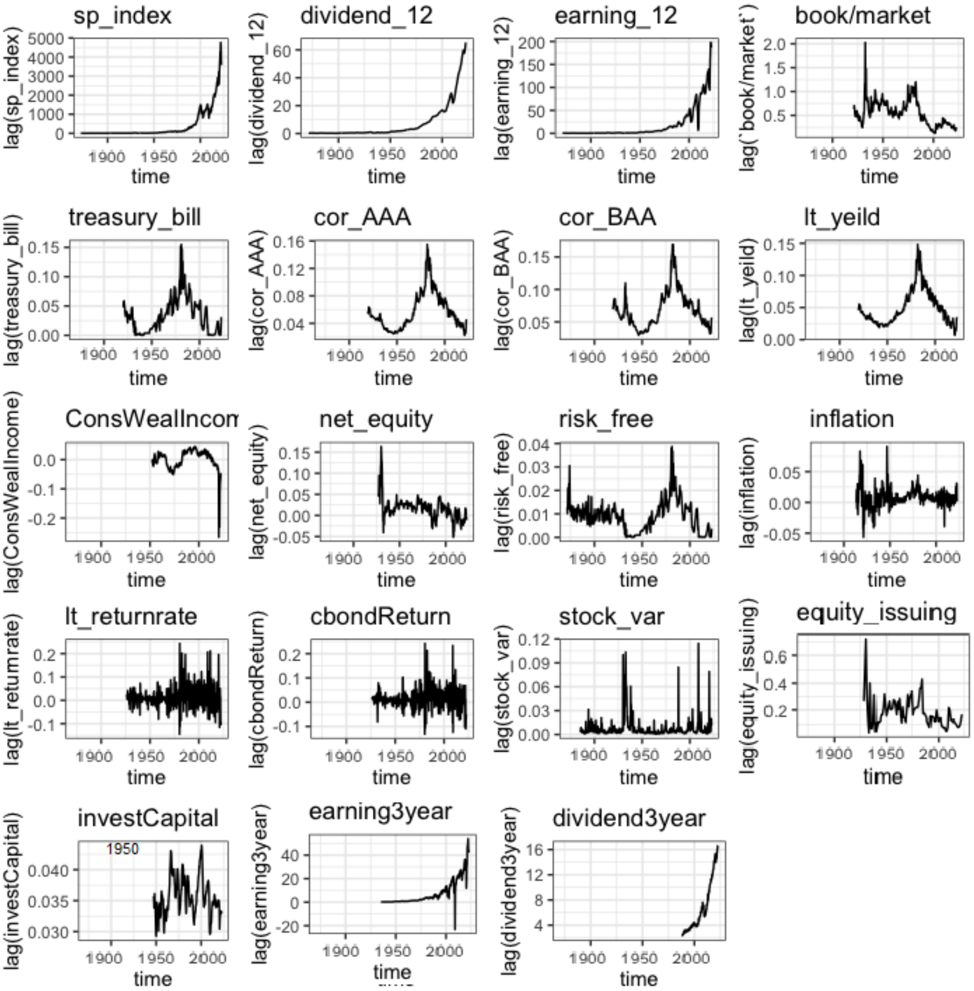


Figure 1: Plot of all nineteen variables over time.

2.2 Excess returns

Excess return is the surplus return on the investment above the risk-free rate, providing insights into the additional return earned for taking on additional risk. By subtracting the risk-free rate from the asset's return, it's possible to measure the additional return generated by the asset in comparison to a virtually risk-free investment.

Based on the definition of stock excess returns, the time series excess returns are generated for all three frequencies - monthly, quarterly, and annual- by first calculating the percentage change between consecutive values in the S&P 500 index data, effectively determining the daily returns of the index with the help of *Delt* function in package *quantmod* in R. The intuitively of the *Delt* package is taking the percentage change of stock price to have the stock returns $R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$ with P_t is the S&P 500 stock price at time t.

Following this, the function subtracts the risk-free rate from the calculated returns to derive the excess returns $ER_t = R_t - RF_t$ with ER_t as the excess return at time t, R_t as the stock return which has been calculated above and RF_t is the risk-free rate at time t. This is a common approach in finance known as the ”excess return formula.” This calculation is fundamental in finance as it helps in assessing the performance of an investment against a low-risk alternative, providing insights into the excess returns generated, the additional compensation for taking on additional risk and aiding in investment decision-making. The time series excess returns are plotted in Figure 2 for monthly data, and for quarterly data and annual data, plots will be placed in Appendix

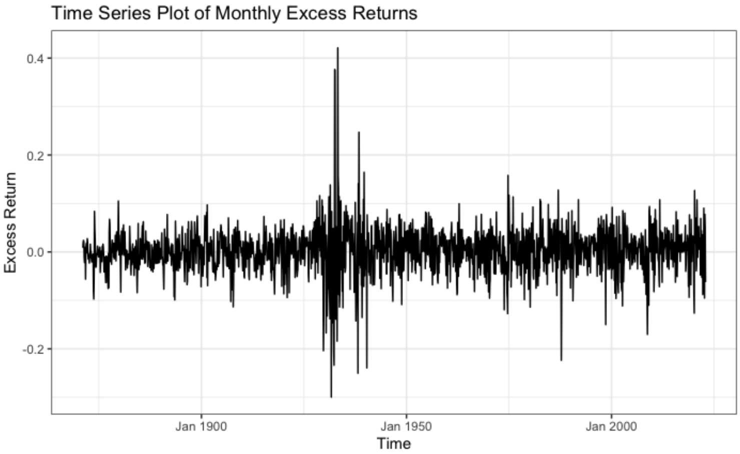


Figure 2: Monthly excess returns time series.

All three-time series excess returns for three frequencies show a constant or stationary trend. The time series excess returns vary around the mean that close to zero and do not show any upward or downward trend along the research times. This can be because the excess returns are generated by taking first differences in short-term future prices that will remove the trend of the S&P 500 and then subtract the risk-free rate. Taking first differences can make the variables stationary[1]. In table 1 the minimum observation, the first quantile, the median, the mean, variance and the third quantile, and the maximum observation for the excess returns time series for different frequencies (monthly, quarterly, and annually time series data).

Table 1: Summary statistics of excess returns series.

Excess return	Min	1st Quantile	Median	Mean	Variance	3rd Quantile	Max
Monthly	-0.30	-0.02	0.003	0.002	0.002	0.03	0.42
Quarterly	-0.40	-0.04	0.01	0.006	0.009	0.06	0.86
Annual	-0.48	-0.09	0.02	0.03	0.035	0.15	0.46

[1] A time series is said to be stationary if its statistical properties, such as the mean, variance, and autocorrelation, remain constant over time. In other words, stationarity implies that the underlying data-generating process does not change with time.

Analyzing the monthly excess returns series, the data reveals a balanced trend with a mean value equal to -0.3, and close to zero variance (0.002) indicating overall stability. The range extends from a minimum of -0.02 to a maximum of 0.42, suggesting potential fluctuations in first differences of future S&P 500 prices. The same fluctuation happens with quarterly and annual data around the mean -0.4, -0.48, and variance 0.009, and 0.035 respectively. The results also show that the excess returns data less fluctuate with a higher frequency and that monthly data has the smallest variance, then quarterly and annual data have higher fluctuation.

The autocorrelations function (ACF) and the partial-autocorrelations (PACF) of the three-time series excess returns at monthly, quarterly, and annually are plotted in Figure 3. Autocorrelation, also known as serial correlation, refers to the correlation observed between a variable in a time series and its previous values. While, partial autocorrelation assesses the direct connection between a current observation and its historical values, conditioned on all the information in the past value in between (Hamilton, 1994). From ACF and PACF plots, an idea of which past observations are related to their own past value can be obtained. They also will be considered as the benchmark for fitting AR(p) models in this project.

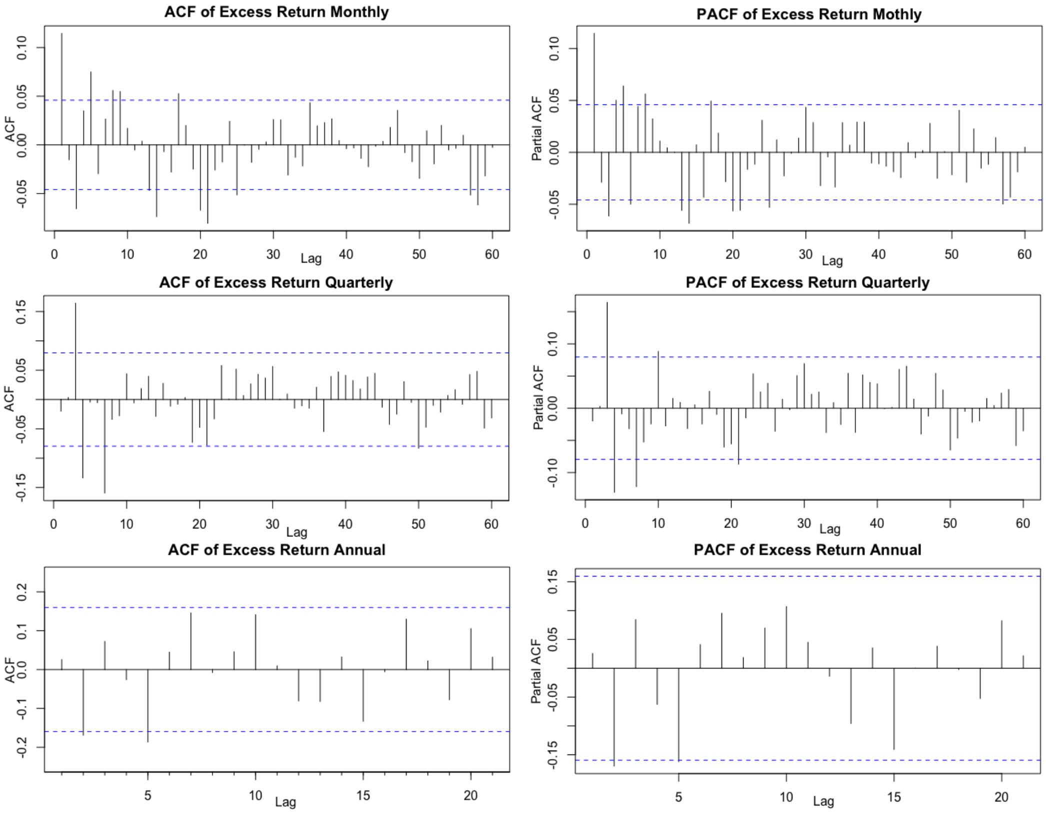


Figure 3: Plot of the autocorrelations and the partial-autocorrelations of excess return series for three frequencies.

In each plot, the blue lines depict the 95% confidence interval and are indicators of the significance threshold. The lags which are beyond the blue lines are considered important. For all three frequencies, the ACF plots exhibit a gradually decaying pattern as we look at lags further into the past, especially after a lag of 21 (for monthly data), 7 (for quarterly data), and 5 (for annual data), there is no correlation exceed the confidence interval line. Looking at PACF plots for three models, we can see the significant cut-off after lag 1 for monthly data, after lag 1 the correlations are very close to the confidence interval line. For quarterly excess return data, after lagging 3 and 4, the correlation seems not so significant, and for annual data the correlation is quite weak, and only at lag 2 does the correlation level exceed the confidence interval line and lie on the confidence interval line at lag 5. Accordingly, the AR(p) model will be fitted with the range of lag p based on observations on ACF and PACF plots.

3 Methods

In this section, we provide an overview and description of the various methods used in this project, as well as the software utilized for their implementation.

3.1 Autoregressive Process

An autoregressive (AR) process is a representation of a type of random process where the current variable is modeled as a linear combination of its past values and a stochastic component. An autoregressive process of order p (AR(p)) satisfies the equation

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t, \tag{1}$$

where Y_t is the value of the time series variable at time t, $\phi_1, \phi_2, \ldots, \phi_p$ are the autoregressive coefficients and $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$ are the lag (or past) values of the time series variable

An AR(p) process is called stationary if all the roots of its characteristic polynomial which is given by

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0,$$

lie outside the unit circle. In this project, we assume that the time series variables follow a Gaussian AR(p) process of the form in equation (1) where $\epsilon_t \sim i.i.d\ \mathcal{N}(0, \sigma^2)$. The vector of parameters that needs to be estimated is $\boldsymbol{\theta} = (c, \phi_1, \phi_2, \ldots, \phi_p, \sigma^2)$. The first p observations in the sample $(y_1, y_2, \ldots, y_p)$ are stored as a $(p \times 1)$ vector $\mathbf{y}_p$. $\mathbf{y}_p$ is the realization of a p-dimensional Gaussian variable. The mean of $\mathbf{y}_p$ is $\boldsymbol{\mu}_p$ which is also a $(p \times 1)$ vector with each element given by $\mu = \frac{c}{1-\phi_1-\phi_2-\cdots-\phi_p}$. Let $\sigma^2\mathbf{V}_p$ be a $(p \times p)$ variance-covariance matrix of $(Y_1, Y_2, \ldots, Y_p)$:

$$\sigma^2\mathbf{V}_p = \begin{bmatrix} E(Y_1-\mu)^2 & E(Y_1-\mu)(Y_2-\mu) & \ldots & E(Y_1-\mu)(Y_p-\mu) \\ E(Y_2-\mu)(Y_1-\mu) & E(Y_2-\mu)^2 & \ldots & E(Y_2-\mu)(Y_p-\mu) \\ \vdots & \vdots & \ldots & \vdots \\ E(Y_p-\mu)(Y_1-\mu) & E(Y_p-\mu)(Y_2-\mu) & \ldots & E(Y_p-\mu)^2 \end{bmatrix}.$$

The density of the first p observations follow a $\mathcal{N}(\boldsymbol{\mu}_p, \sigma^2\mathbf{V}_p)$ distribution, which is

$$f_{Y_p,Y_{p-1},\ldots,Y_1}(y_p, y_{p-1}, \ldots, y_1; \boldsymbol{\theta})$$

$$= (2\pi)^{-p/2}(\sigma^{-2})^{p/2}|\mathbf{V}_p^{-1}|^{1/2}\exp[-\frac{1}{2\sigma^2}(\mathbf{y}_p - \boldsymbol{\mu}_p)'\mathbf{V}_p^{-1}(\mathbf{y}_p - \boldsymbol{\mu}_p)].$$

The log-likelihood for the whole sample is

$$\mathcal{L}(\boldsymbol{\theta}) = \log f_{Y_T,Y_{T-1},\ldots,Y_1}(y_T, y_{T-1}, \ldots, y_1; \boldsymbol{\theta})$$

$$= -\frac{T}{2}log(2\pi) - \frac{T}{2}log(\sigma^2) + \frac{1}{2}log|\mathbf{V}_p^{-1}| - \frac{1}{2\sigma^2}(\mathbf{y}_p - \boldsymbol{\mu}_p)'\mathbf{V}_p^{-1}(\mathbf{y}_p - \boldsymbol{\mu}_p)$$
$$- \sum_{t=p+1}^{T} \frac{(y_t - c - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p})^2}{2\sigma^2}. \qquad (2)$$

The log-likelihood is maximized to get the estimates of the parameters

$$\boldsymbol{\theta} = (c, \phi_1, \ldots, \phi_p, \sigma^2),$$

as $\hat{\boldsymbol{\theta}} = (\hat{c}, \hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\sigma}^2)$ (Hamilton, 1994).

3.2 Ordinary Least Square

Ordinary Least Squares regression, referred to as OLS, is one of the most common techniques for estimating the parameters in a linear regression model, which describes the relationship between one or more independent quantitative variables and a dependent variable (simple or multiple linear regression). This model is described mathematically as equation (3) below.

$$y_t = {x_t}^\intercal \beta + \epsilon_t = \beta_0 + \beta_1 x_{t,1} + \beta_2 x_{t,2} + \cdots + \beta_k x_{t,k} + \epsilon_t, \tag{3}$$

with $t = 1, 2, \ldots, T$, T is the last time point or the total number of observations of regression model. The target of regression model is the N time points of vector $Y = (y_1, y_2, \ldots, y_N)^\intercal$. The part of the right-hand side is called the regression or the regression function which involves the k-dimensional (column) vector $(x_{t1}, x_{t2}, \ldots, x_{tK})^\intercal$, k-dimensional vector of the random component $(\epsilon_1, \epsilon_2, \ldots, \epsilon_T)^\intercal$ and the coefficients $(\beta_0, \beta_1, \beta_2, \ldots, \beta_k)^\intercal$ are called the regression coefficients, which represents the change in the dependent variable when the k^{th} regressor increases by one unit while other regressors are held constant. In the language of calculus, this can be expressed as $\frac{\partial y_t}{\partial x_{tk}} = \beta_k$

For the matrix notation, define the $T \times k$ matrix of regressors X

$$\underset{T\times k}{\mathbf{X}} = \begin{bmatrix} {x_1}^\intercal \\ {x_2}^\intercal \\ \vdots \\ {x_n}^\intercal \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & \ldots & x_{k,1} \\ 1 & x_{1,2} & \ldots & x_{k,2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,T} & \ldots & x_{k,T} \end{bmatrix}$$

we can rewrite the linear regression equation as $\underset{T\times 1}{\mathbf{Y}} = \underset{T\times k}{X} \underset{k\times 1}{\beta} + \underset{T\times 1}{\varepsilon}$

In the context of linear regression, OLS aims to determine the set coefficients $\beta_0, \beta_1, \ldots, \beta_k$ that minimize the sum of the squared differences between the observed and predicted values, which is known as the sum of square errors (SSR) with $SSR = \sum_{t=1}^{T} (y_t - \hat{y}_t)^2$. In that way, the vector β of the coefficients can be estimated by the following formula:

$$\beta = (X^T X)^{-1} X^T Y$$

In order to derive the vector β of the coefficients, OLS model must fulfill four main assumptions:

- Linear relationship between the dependent variable and the regressors

- There must be no relationship between the X's and the error term. For time series models, exogeneity condition is not satisfied since the current time of regressors will be correlated to its past and its future.

- The rank of the $T \times k$ data matrix, X, is k with probability 1.

- Homoscedasticity and independence of error terms: the errors (residuals) have constant variance and are uncorrelated with each other.

3.3 Akaike Information Criterion

To detect the optimal number of orders (p) in an AR(p) model, and messure the best subset selections, the Akaike information criterion (AIC) has been used. AIC is a mathematical method for evaluating how well a model fits the data it was generated from. In this project, AIC is used to compare different possible models and determine which one is the best fit for the data. AIC has the general form:

$$AIC(m) = 2K - 2\log(\hat{\mathcal{L}}) \tag{4}$$

Here m denotes the model, K denotes the number of parameters in the model and $\hat{\mathcal{L}}$ is the associated likelihood of the fitted model given parameters K. The optimum lag order is chosen such that AIC is minimum over all possible lag orders (Aho et al., 2014).

3.4 Mean Square Forecast Error

To evaluate the accuracy of forecasts in this project, we employed Mean Square Forecast Error (MSFE). MSFE is used to measure the accuracy of forecasts in time series analysis. It's computed by taking the average of the squared differences between forecasted values and actual values over a specified period as formula

$$MSFE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2} \tag{5}$$

.

MSFE is the key metric for evaluating the performance of forecasting methods or models and helps in comparing different forecasting techniques. A high RMSE implies that the fitted values are further away from the actual data, the lower MSFE indicates the better forecasting accuracy.

3.5 Best subset selection

To have the best subset selection, we undertake the fitting of individual OLS models for each conceivable combination of the K predictors. This involves fitting all K models containing precisely one predictor, all $\binom{K}{2} = (K(K-1)/2)$ models with exactly two predictors, and so forth. Subsequently, a comprehensive evaluation is conducted across all resulting models with the objective of pinpointing the optimal one. Typically, this task is divided into two stages as Procedure (1):

Let M_0 represent the null model, devoid of any predictors, predicting the sample mean for each observation. For each $k = 1, 2, \ldots, K$:

(1) Fit all $\binom{K}{k}$ models comprising exactly k predictors.

(2) Identify the best model among the $\binom{K}{k}$ models, denoting it as M_k. The criteria for 'best' involve minimizing the Sum of Squares Residual (SSR) or, equivalently, maximizing R^2.

(3) Employ a selection criterion, such as the AIC for this project, to choose a singular best model from the set $M_0, M_1, ..., M_K$.

Instead of the general process of the best subset selection, which involves two stages and allows for potential manual selection, we also explored the forward stepwise model selection method. This approach entails iteratively incorporating predictors based on their individual contributions to face with computational challenges for the best subset selection, especially as the number of predictors (K) increases. The computational efficiency of forward stepwise selection makes it a viable alternative to best subset selection. Unlike best subset selection, which considers all 2^K possible models, forward stepwise selection explores a much smaller set of models with $K - k$ models. It also starts with a null model M_0 as general method in Procedure (1), but different in the range of $k = 1, 2, \ldots, K - 1$. The procedures also used AIC as criteria for selection (Hamilton, 2013).

3.6 Statistical Software

The programming language R (R Core Team, 2021) is used for analyzing and processing the dataset along with the package dplyr (Wickham et al., 2022). The *ggplot*2 package (Wickham, 2016) is also used in this project for data visualization. For calculation of the excess return series we employed the *Delt* function in package *quantmod*, for subset selection package *lmSubsets* (Hofmann, 2022) has been used.

4 Results

This section is dedicated to analyzing the data using the methods defined in the preceding section. It is divided into the following subsections:

4.1 AR(p) Processes

This section represents the results after fitting the AR(p) process of the form (1) to the excess return series with 1823 observations (for monthly data), 607 observations (for quarterly data), and 151 observations (for annual data) by maximum likelihood estimation given by equation (2). Corporate with the plot for ACF and PACF of the excess returns time series, the lag orders are allowed to vary from 1 to 5, and the corresponding AIC is noted. The optimum orders are then determined such that the corresponding model has the lowest AIC. The formula for AIC used is the same as in equation (3). The particular combination of lag orders that have the lowest AIC is noted in Table 2.

Table 2: Optimum models with the lowest AIC for excess return series.

Excess returns	AR(p) Process	AIC
Monthly	AR(1)	-5958.87
Quarterly	AR(4)	-1136.68
Annual	AR(2)	-77.09

Let X_t, Y_t, and Z_t denote the monthly, quarterly, and annual short-term future returns of S&P 500 index at time point t respectively, $t = 1, 2, \ldots, T$ with T up to 1823 for monthly data model, 607 for quarterly model and 151 for annual data model. The regression equation obtained after estimating the parameters for monthly excess returns by maximum likelihood estimation is given as:

$$\hat{X_t} = 0.0019 + 0.1145X_{t-1} + \epsilon_t,$$

The regression equation obtained for quarterly excess returns series is:

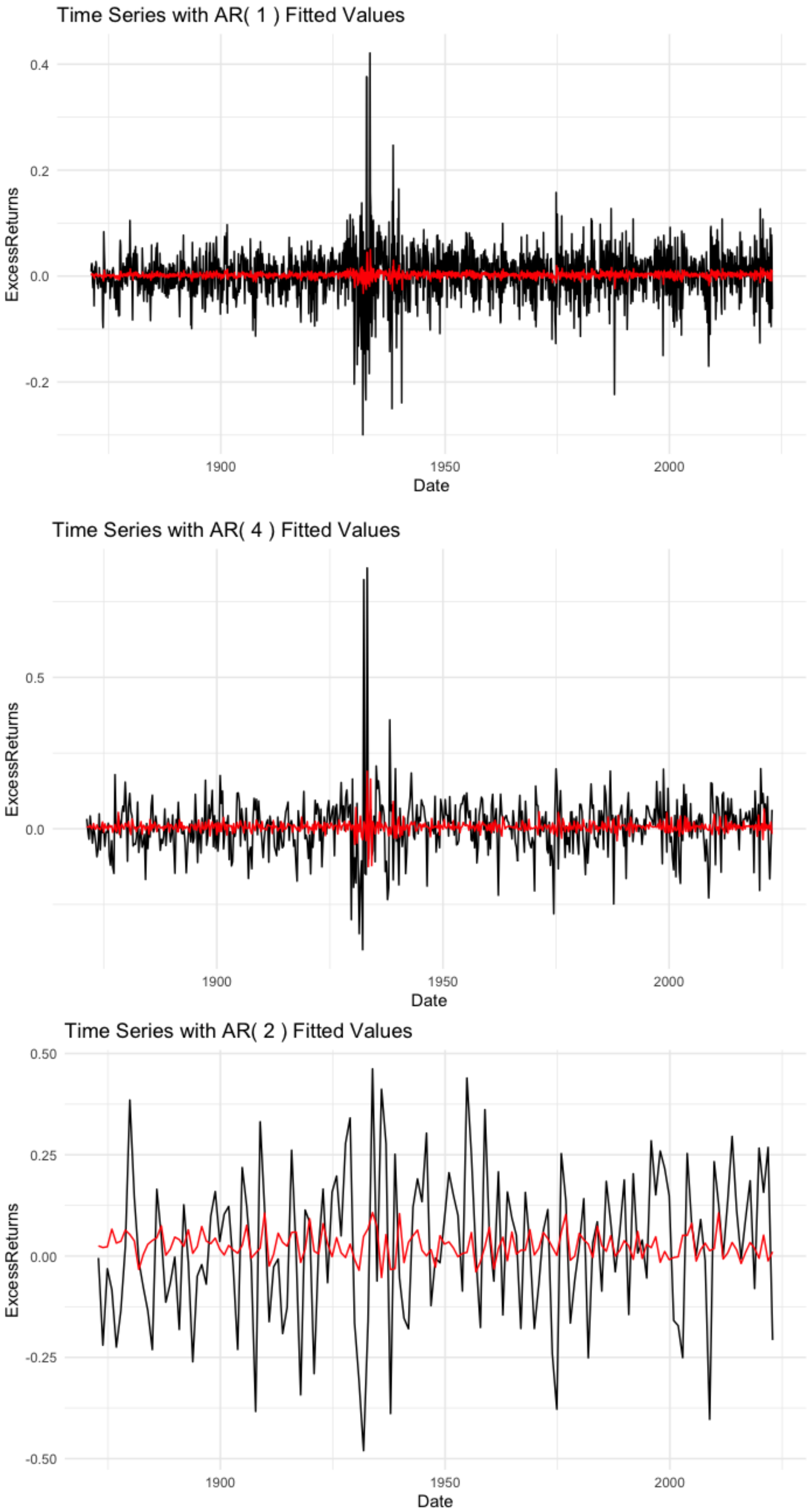
$$\hat{Y_t} = 0.0065 + 0.0011Y_{t-1} + 0.0074Y_{t-2} + 0.1617Y_{t-3} - 0.1302Y_{t-4} + \epsilon_t,$$

The regression equation obtained for annual excess returns time series is:

$$\hat{Z_t} = 0.0256 + 0.0318Z_{t-1} - 0.1735Z_{t-2} + \epsilon_t,$$

0.02145, 0.0677 and 0.0224 are the mean of the process for monthly, quarterly, and annual excess returns time series process respectively, created by Autoregressive models. All estimated coefficients give the relationship between the current value of the process and the past values.

Based on the full-sample fit AR(1) model for monthly, AR(4) for quarterly, and AR(2) for annual excess returns time series data, we generate the forecast at all times in the sample. To create the forecasts, the models are refitted each time. We assume that the forecasts follow a normal distribution. The results are plotted in Figure 4 for monthly, quarterly and annual fit



Time Series with AR(1) Fitted Values

Time Series with AR(4) Fitted Values

Time Series with AR(2) Fitted Values

Figure 4: Plot of the forecasted and the actual value fitted by AR(1) model for monthly excess returns data.

The value of MSFEs is also calculated for each AR(p) process with MSFEs are $0.0471, 0.0939, 0.1825$ for monthly, quarterly, and annual predictive models respectively. It's a very good prediction for AR(1) process of monthly data, moderate predictive performance for AR(4) of quarterly and AR(2) of annual.

While MSFE is a valuable metric for comparing forecasting models in controlled settings or historical data analysis such as autoregressive processes, its practical applicability might be limited in certain real-life forecasting scenarios due to the challenge of obtaining true future values since MSFE requires a time series of forecasts and corresponding actual values to calculate the differences. Also its assumptions of stationary forecast errors since MSFE assumes that the forecast errors are stationary over time.

4.2 Fitted model using OLS

To comprehensively assess the impact of external factors on time series excess returns, our methodology involved constructing predictive models through the estimation of individual OLS models for each predictor in isolation. Subsequently, we aggregated all predictors into a unified model to evaluate both the individual contributions of each regressor and the collective impact of all predictors on the excess return predictive models. It was crucial to lag each predictor one period before setting up the model to account for temporal dependencies. This process allowed us to obtain individual autoregressive models for each predictor.

MSFEs are then computed for each model, and the results are presented in the Appendix. Notably, for monthly data, the model featuring only a 12-month Dividends predictor outperformed the combined predictor model in terms of MSFE, yielding a value of 0.0474. However, for quarterly and annual data, the combination of all predictors demonstrated superior predictive performance compared to the models employing individual predictors alone.

Analyzing the MSFEs for models with each predictor in isolation revealed that the Dividends 12-month period (for monthly data), Investment Capital Ratio (for quarterly data), and Stock Variance (for annual data) exhibited the smallest MSFEs, with values of 0.0474, 0.0762, and 0.1727, respectively. Conversely, Net Equity Expansion (for monthly and annual data) and Long-Term Rate of Return (for quarterly data) displayed relatively poor predictive performance with the highest MSFEs among the variables, with values of 0.0540, 0.1806, and 0.1088, respectively. These findings contribute to the understanding of the predictive power of individual predictors and emphasize the importance of considering their combined influence for accurate time series forecasting.

4.3 Best subset selection

The outcomes obtained from the best subset selection process are documented for excess returns prediction models. The model incorporating the best subsets, determined through both the general approach and the forward stepwise selection method, will be detailed in the table in the Appendix. The predictive performance of each model is assessed using MSFEs and is noted in 4. Upon comparison between the general method and the forward stepwise selection, it becomes evident that, in most instances, the forward stepwise selection method outperforms the general selection procedure. It can be understandable since the algorithm selects the variable in forward stepwise selection (as discussed in the Method part), uses a smaller set of models, and provides the greatest improvement in model performance. Its step-by-step approach allows forward stepwise selection to adapt to the most influential predictors, resulting in a more efficient and refined model.

4.4 Predictive performance of all models

MSFE is employed in this research to evaluate the predictive performances of each forecasting model (AR(p) and OLS for each combination of the data and OLS with best subsets selection method). MSFE is calculated using equation (5) for each model, each frequency and noted in Table 4

Table 3: MSFE for all forecasting model

Variables	MSFE monthly	MSFE quarterly	MSFE annual
AR(p)	0.0471	0.0939	0.1825
OLS model each predictor (min)	0.0474	0.0762	0.1806
OLS model each predictor (max)	0.0540	0.1088	0.1940
OLS model - all predictors	0.0536	0.0684	0.1404
OLS model - best subset	0.0539	0.0695	0.1464
OLS model - forward stepwise	0.0537	0.0695	0.1445

Depending on the results of MSFE from Table 4 we can see that all models are providing accurate predictions for monthly and quarterly data (MSFE ranging from 0.04 to 0.1), especially very good performance for monthly data (just around 0.05 for all models) but moderate performance for annual data (greater than 0.1). Across each type of model, for monthly data, AR(1) has the best predictive performance in terms of MSFE when it has the smallest MSFE (0.0471) compared to other models, and follow up closely is the model with only Dividend 12 months as predictors with MSFE is 0.0474, while the model with only Net Equity as predictor has highest MSFE among all models (0.054). For quarterly and annual data, the model with the combination of all predictors has the best predictive performance in terms of MSFEs with MSFEs are 0.0684 and 0.1404 respectively.

5 Conclusion

In this project, we constructed models for forecasting excess returns series using a multivariate time series that includes external variables, as well as the excess returns time series generated from the $S\&P500$ index and the risk-free rate for all frequencies, monthly, quarterly, and annual. The procedure began with fitting AR(p) models for the excess returns time series for 1823 time points for monthly data model, 607 time points for quarterly model and 151 time points for annual data model, employing visualization techniques such as ACF and PACF to control lag of AR(p) models, and utilizing AIC for selecting the optimal lag p. Subsequently, we developed linear models using OLS regression, exploring three different approaches for constructing predictive models: utilizing each predictor individually, combining all predictors into a single model, and selecting the best subset of predictors through both general and forward stepwise selection procedures.

For the full-sample fit, AR(1) models were employed for monthly data, AR(4) for quarterly data, and AR(2) for annual data in the excess returns time series. The generated forecasts demonstrated excellent predictive performance for the AR(1) model in monthly data, moderate performance for AR(4) in quarterly data, and AR(2) in annual data, as evaluated by the mean squared forecast error (MSFE). Notably, the AR(1) model exhibited the best predictive performance, boasting the smallest MSFE compared to other models.

In terms of combining predictors, the model with the inclusion of all predictors demonstrated superior predictive performance for quarterly and annual data, achieving the lowest MSFE. A comparative analysis between the general method and forward stepwise selection revealed that, in most cases, forward stepwise selection method outperformed the general selection procedure. This can be attributed to the forward stepwise algorithm's adaptive nature, selecting variables incrementally to provide the greatest improvement in model performance. The step-by-step approach leads to a more efficient and refined model.

Bibliography

Ken Aho, DeWayne Derryberry, and Teri Peterson. *Model selection for ecologists: the worldviews of AIC and BIC*. Ecological Society of America, 2014.

James D Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, New Jersey, 1994.

James D Hamilton. *An introduction to statistical learning with applications in R*. Springer, 2013.

Marc Hofmann. *ggplot2: Elegant Graphics for Data Analysis*. CRAN, 2022. URL `https://github.com/marc-hofmann/lmSubsets.R`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

Ivo Welch and Amit Goyal. *A Comprehensive Look at The Empirical Performance of Equity Premium Prediction*, 2008.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2022. URL `https://CRAN.R-project.org/package=dplyr`. R package version 1.0.9.

Appendix

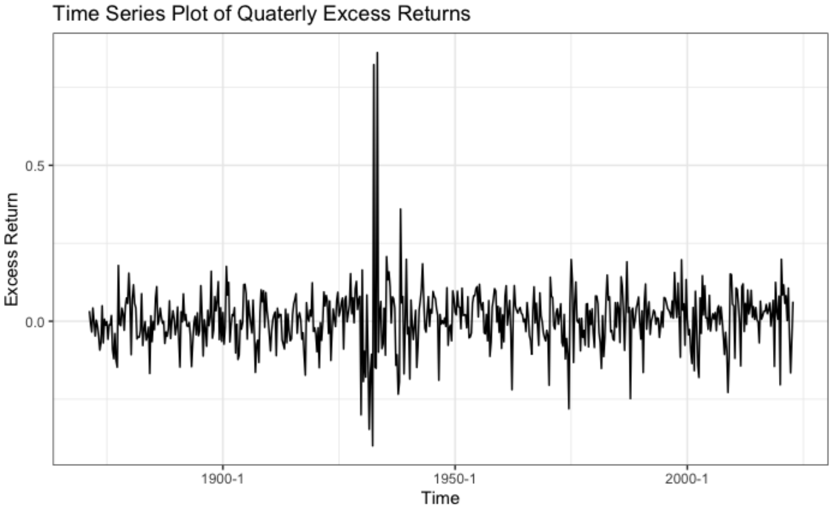


Figure 5: Plot of time series excess returns

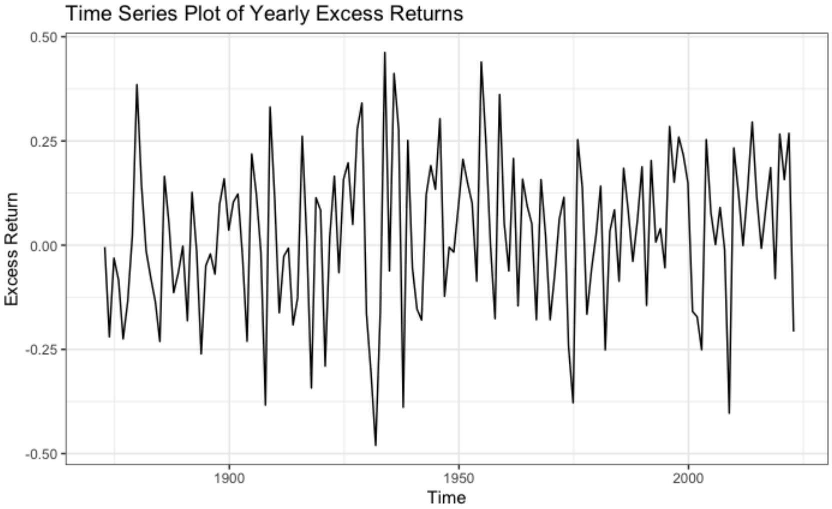


Figure 6: Plot of time series excess returns

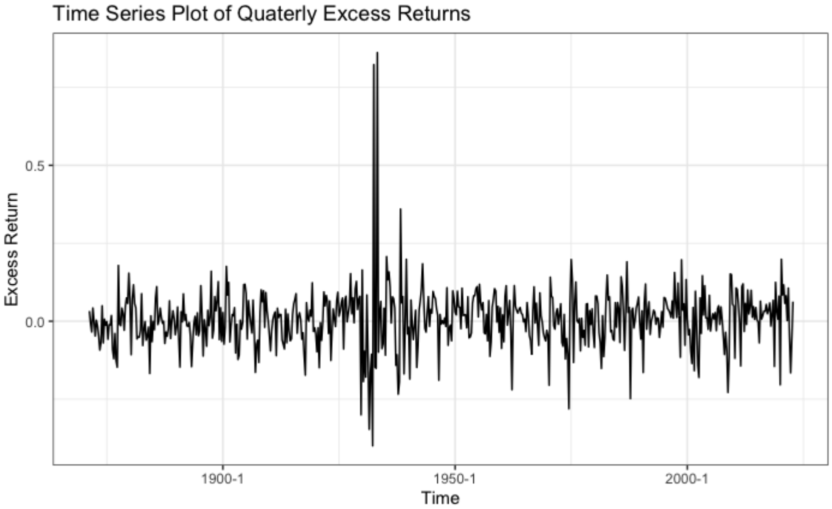
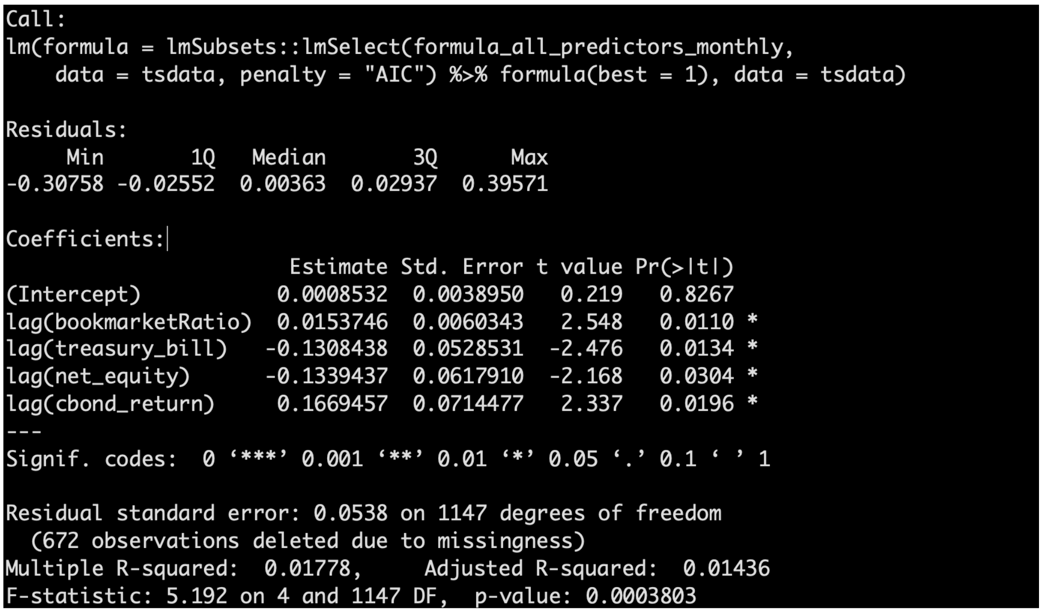


Figure 7: Plot of time series excess returns

Table 4: MSFE for forecasting models for each and al predictors using OLS

Model monthly	MSFE monthly	MSFE quarterly	MSFE annual
dividend12	0.0474	0.09595	0.1842
earning12	0.04741	0.09596	0.18337
bookmarketRatio	0.05292	0.10582	0.18121
treasurybill	0.0529	0.1062	0.1899
corAAA	0.0528	0.1060	0.1882
corBAA	0.0528	0.1057	0.1842
longtermyeild	0.0528	0.1060	0.1892
netequity	0.0540	0.1087	0.1933
consumWealIncome	NA	0.079	0.1698
inflation	0.0520	0.1046	0.1913
longreturnrate	0.0538	0.1088	0.1932
cbondreturn	0.0529	0.1070	0.1904
investmentCapital	0.0761	0.0965	0.1647
dividend3year	NA	0.0781	NA
earning3year	NA	0.0835	NA
All predictors	**0.0535**	**0.0684**	**0.1403**



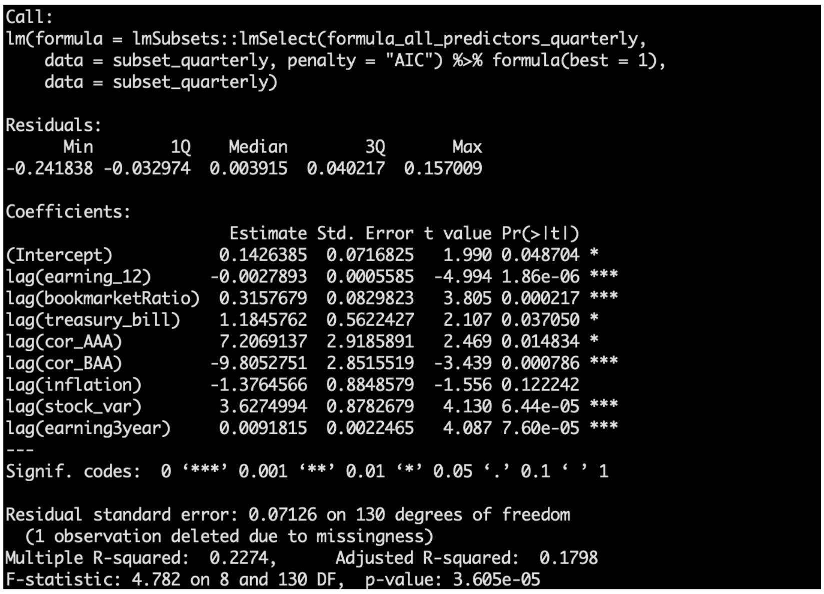
```
Call:
lm(formula = lmSubsets::lmSelect(formula_all_predictors_monthly,
    data = tsdata, penalty = "AIC") %>% formula(best = 1), data = tsdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.30758 -0.02552  0.00363  0.02937  0.39571 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)  
(Intercept)          0.0008532  0.0038950   0.219   0.8267  
lag(bookmarketRatio)  0.0153746  0.0060343   2.548   0.0110 *
lag(treasury_bill)   -0.1308438  0.0528531  -2.476   0.0134 *
lag(net_equity)      -0.1339437  0.0617910  -2.168   0.0304 *
lag(cbond_return)     0.1669457  0.0714477   2.337   0.0196 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0538 on 1147 degrees of freedom
  (672 observations deleted due to missingness)
Multiple R-squared:  0.01778,	Adjusted R-squared:  0.01436 
F-statistic: 5.192 on 4 and 1147 DF,  p-value: 0.0003803
```

Figure 8: Results for best subset monthly



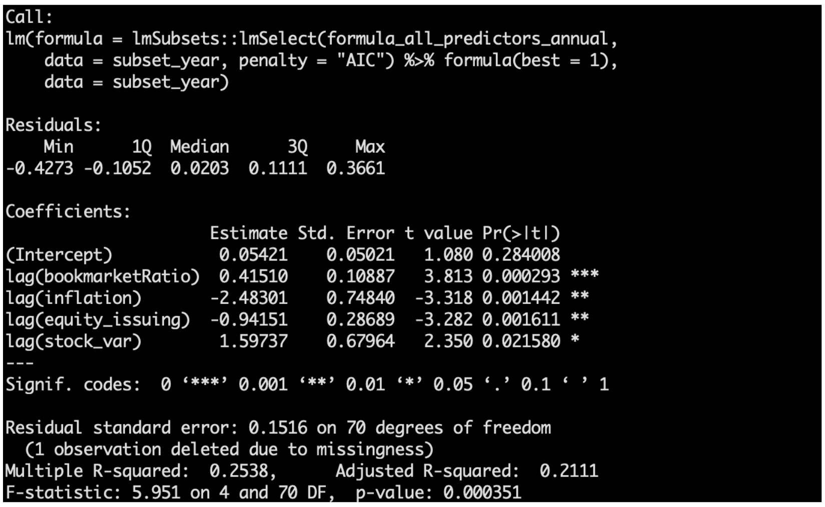
```
Call:
lm(formula = lmSubsets::lmSelect(formula_all_predictors_quarterly,
    data = subset_quarterly, penalty = "AIC") %>% formula(best = 1),
    data = subset_quarterly)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.241838 -0.032974  0.003915  0.040217  0.157009 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          0.1426385  0.0716825   1.990 0.048704 *  
lag(earning_12)     -0.0027893  0.0005585  -4.994 1.86e-06 ***
lag(bookmarketRatio)  0.3157679  0.0829823   3.805 0.000217 ***
lag(treasury_bill)    1.1845762  0.5622427   2.107 0.037050 *  
lag(cor_AAA)          7.2069137  2.9185891   2.469 0.014834 *  
lag(cor_BAA)         -9.8052751  2.8515519  -3.439 0.000786 ***
lag(inflation)       -1.3764566  0.8848579  -1.556 0.122242    
lag(stock_var)        3.6274994  0.8782679   4.130 6.44e-05 ***
lag(earning3year)     0.0091815  0.0022465   4.087 7.60e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07126 on 130 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.2274,	Adjusted R-squared:  0.1798 
F-statistic: 4.782 on 8 and 130 DF,  p-value: 3.605e-05
```

Figure 9: Results for best subset quarterly



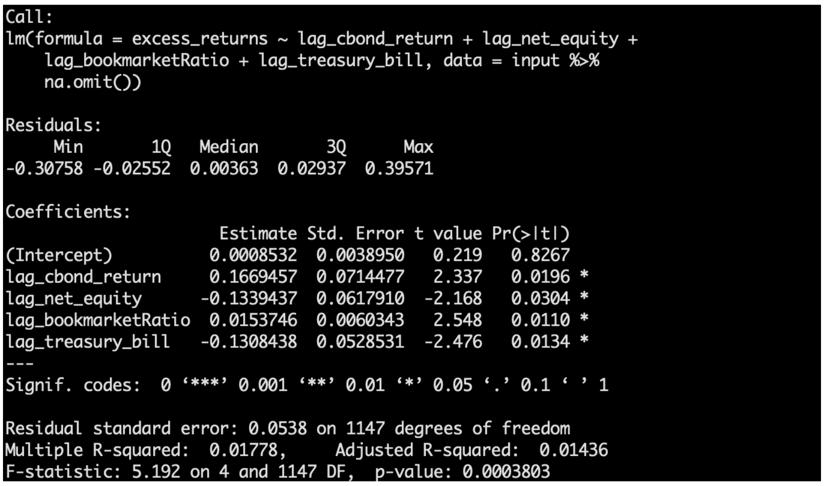
```
Call:
lm(formula = lmSubsets::lmSelect(formula_all_predictors_annual,
    data = subset_year, penalty = "AIC") %>% formula(best = 1),
    data = subset_year)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4273 -0.1052  0.0203  0.1111  0.3661

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.05421    0.05021   1.080 0.284008
lag(bookmarketRatio)  0.41510    0.10887   3.813 0.000293 ***
lag(inflation)       -2.48301    0.74840  -3.318 0.001442 **
lag(equity_issuing)  -0.94151    0.28689  -3.282 0.001611 **
lag(stock_var)        1.59737    0.67964   2.350 0.021580 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1516 on 70 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.2538,	Adjusted R-squared:  0.2111
F-statistic: 5.951 on 4 and 70 DF,  p-value: 0.000351
```

Figure 10: Results for best subset annual



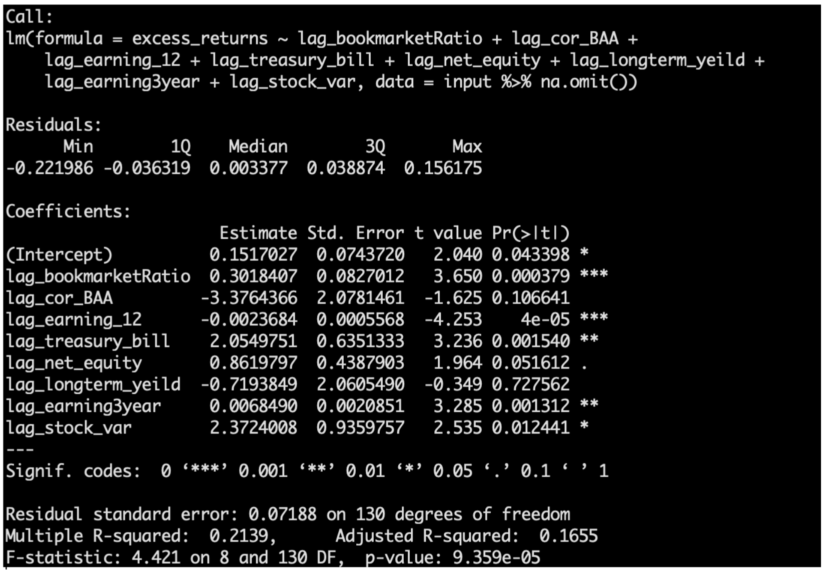
```
Call:
lm(formula = excess_returns ~ lag_cbond_return + lag_net_equity +
    lag_bookmarketRatio + lag_treasury_bill, data = input %>%
    na.omit())

Residuals:
     Min       1Q   Median       3Q      Max
-0.30758 -0.02552  0.00363  0.02937  0.39571

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.0008532  0.0038950   0.219   0.8267
lag_cbond_return     0.1669457  0.0714477   2.337   0.0196 *
lag_net_equity      -0.1339437  0.0617910  -2.168   0.0304 *
lag_bookmarketRatio  0.0153746  0.0060343   2.548   0.0110 *
lag_treasury_bill   -0.1308438  0.0528531  -2.476   0.0134 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0538 on 1147 degrees of freedom
Multiple R-squared:  0.01778,	Adjusted R-squared:  0.01436
F-statistic: 5.192 on 4 and 1147 DF,  p-value: 0.0003803
```

Figure 11: Results for forward stepwise selection monthly



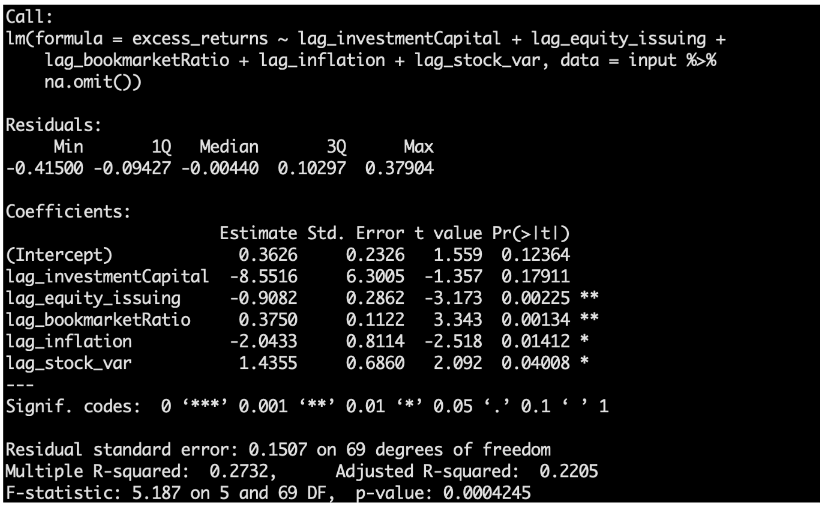
```
Call:
lm(formula = excess_returns ~ lag_bookmarketRatio + lag_cor_BAA +
    lag_earning_12 + lag_treasury_bill + lag_net_equity + lag_longterm_yeild +
    lag_earning3year + lag_stock_var, data = input %>% na.omit())

Residuals:
      Min        1Q    Median        3Q       Max
-0.221986 -0.036319  0.003377  0.038874  0.156175

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.1517027  0.0743720   2.040 0.043398 *
lag_bookmarketRatio  0.3018407  0.0827012   3.650 0.000379 ***
lag_cor_BAA         -3.3764366  2.0781461  -1.625 0.106641
lag_earning_12      -0.0023684  0.0005568  -4.253    4e-05 ***
lag_treasury_bill    2.0549751  0.6351333   3.236 0.001540 **
lag_net_equity       0.8619797  0.4387903   1.964 0.051612 .
lag_longterm_yeild  -0.7193849  2.0605490  -0.349 0.727562
lag_earning3year     0.0068490  0.0020851   3.285 0.001312 **
lag_stock_var        2.3724008  0.9359757   2.535 0.012441 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07188 on 130 degrees of freedom
Multiple R-squared:  0.2139,	Adjusted R-squared:  0.1655
F-statistic: 4.421 on 8 and 130 DF,  p-value: 9.359e-05
```

Figure 12: Results for forward stepwise selection quarterly



```
Call:
lm(formula = excess_returns ~ lag_investmentCapital + lag_equity_issuing +
    lag_bookmarketRatio + lag_inflation + lag_stock_var, data = input %>%
    na.omit())

Residuals:
     Min       1Q   Median       3Q      Max
-0.41500 -0.09427 -0.00440  0.10297  0.37904

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             0.3626     0.2326   1.559  0.12364
lag_investmentCapital  -8.5516     6.3005  -1.357  0.17911
lag_equity_issuing     -0.9082     0.2862  -3.173  0.00225 **
lag_bookmarketRatio     0.3750     0.1122   3.343  0.00134 **
lag_inflation          -2.0433     0.8114  -2.518  0.01412 *
lag_stock_var           1.4355     0.6860   2.092  0.04008 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1507 on 69 degrees of freedom
Multiple R-squared:  0.2732,	Adjusted R-squared:  0.2205
F-statistic: 5.187 on 5 and 69 DF,  p-value: 0.0004245
```

Figure 13: Results for forward stepwise selection annual