Financial Models and their Assumptions

Spring 2020

Joseph A. Knapp

University of Maryland, Baltimore County

1 Project Motivation and Problem

The idea of this project is to analyze different financial metrics and models and their underlying assumptions. It is known that these assumptions generally do not hold and are overly simplified, but some of the metrics/models are still widely used. This project will evaluate the assumptions of the metrics/models, how well they hold, and if they are useful.

2 Existing Approaches

Many of the metrics/models assume the efficient markets hypothesis (EMH) be true. EMH explains that "competition among investors works to eliminate all positive NPV (net present value) trading opportunities" [2, Pg 295]. Due to competition in the market, easily obtained public information such as media reports, company financial statements, press releases, etc. are readily available and under the EMH many expect "competition between investors to be fierce and the stock price to react nearly instantaneously to such news." [2, Pg 296]. There is also information which is difficult and expensive to obtain. Larger firms have access to all public data as well as the expertise to gather better insight into data which the general public may not be able to. "While the fundamental information may be public, the interpretation of how that information will affect the firm's future cash flows is itself private information. When private information is relegated to the hands of a relatively small number of investors, these investors may be able to profit by trading on their information. In this case, the efficient markets hypothesis will not hold in the strict sense. However, as these informed traders begin to trade, they will tend to move prices, so over time prices will begin to reflect their information as well." [2, Pg 297]. If profitability is great enough, others will attempt to do the same, competition takes over, and EMH should hold.

Studies in the past have tested weather EMH for public data and private information holds or not. The following summarizes studies which support the EMH. In 1953 M. G. Kendall performed analysis to find cycles in stock prices using time-series analysis. Instead Kendal discovered price movements more or less follow a random walk. "An analysis of stock-exchange movements revealed little serial correlation within series and little lag correlation between series. Unless individual stocks behave differently from the average of similar stocks, there is no hope of being able to predict movements on the exchange" [3] which implies public information (past prices) has little bearing on future prices and thus EMH holds. A 2017 study which followed several stocks for over a decade. Scatter plots of stocks prices at time t against time $t-1$ were created and showed the "concentration of points being around the origin, and no bias towards any quadrants" [4]. Because lagged stock prices have no correlation, EMH also holds in that the public information on stock prices is not of use. Studies on EMH in the past have also shown that 'informed traders' and large firms with expert analysis are confined by the EMH. A 2017 study analyzed the performance of actively managed funds compared to the market and found that such 'expert analysists' which have all public and private data where only able to beat the Wilkshire5000 index 40% of the time [4].

Other studies have shown the EMH to be flawed mostly due to anomalies. One such anomaly is referred to as the January Effect, the "perceived seasonal increase in stock prices during the month of January" [5]. The effect is generally attributed to the selling off of stocks in December that experienced significant losses the year before. The December sell-off causes stock prices to drop, which can trigger a buying frenzy in January causing January to have 'superior gains'. Several studies have shown this to true to a certain extent, but it is becoming less of an anomaly [4][5].

Another anomaly is called the Momentum Effect and the Reversal Effect. The Momentum Effect implies a positive serial correlation in stock prices (stock prices continue in the same direction) due to an inefficient market where investors underreact to new information and the market does not adjust instantaneously (as it should if EMH held) [4][6]. The Reversal Effect shows the opposite of the Momentum Effect. It suggests a "negative correlation in stock prices and that a mean reversion in stock prices is in effect as investors overreact to new information" [4]. Again this is due to an inefficient market where investors but investors overreact to new information and the market overreacts.

Using the EMH and Mean-Variance Portfolio Theory, several metrics have been created to analyze returns different securities. These include the Sharpe ratio, Beta, CAPM, and Alpha.

3 EDA and Proposed Solution

The historical data was collected on stocks in the S&P400 [7] and S&P500 [8] between January 1, 2000 and January 1, 2020 using the Yahoo-Finance API [9]. Data was collected for each stock on each date, and included the stocks daily adjusted closing price, stock splits, and dividends paid as well as their Sector and Subindustry information.

My main dataset was a collection of stock data, and stock metrics calculated based on its historical pricing data.

Several of the metrics/columns were created to evaluate the stocks

- `Symbol`: stock price on date of calculation

- `Date`: date in the data frame the calculations were based on
- `StockPrice`: stock price on date of calculation
- `ExpectedMarketReturn`: the historical average return of the S&P500 Index

$$E[R_{Mkt}] = \frac{1}{T}\sum R_{Mkt,t} \quad s.t. \quad R_{Mkt,t} = \frac{\mathrm{Market_{t+1}} - \mathrm{Markete_t}}{\mathrm{Market_t}}$$

- `ExpectedReturn`: the historical average return of the stock

$$E[R_i] = \frac{1}{T}\sum R_{i,t} \quad s.t. \quad R_{i,t} = \frac{\mathrm{Price_{t+1}} - \mathrm{Price_t}}{\mathrm{Price_t}}$$

- `Volatility`: the historical standard deviation of stock the stock’s returns

$$\sigma_i = \sqrt{Var[R_i]} = \sqrt{\frac{\sum\left(R_{i,t} - E[R_i]\right)^2}{T-1}}$$

- `Sharpe`: the Sharpe ratio of the stock

$$= \frac{E[R_i] - r_f}{\sigma_i}$$

  - Where r_f is the risk-free rate: the annual return on a ten-year treasury note
- `Beta`: the Beta of a stock with respect to the S&P500 Index

$$\beta_i = \frac{Cov[R_i, R_{Mkt}]}{\sigma_i^2}$$

- `CAPM`: estimates a stocks return based on its systematic risk (correlation to market risk premium)

$$r_{CAPM} = r_f + \beta_i\left(E[R_{Mkt}] - r_f\right)$$

- `Alpha`: The difference between a stocks `Expected Return` and its `CAPM` return estimate

$$\alpha_i = E[R_i] - r_{CAPM}$$

- `TVarLow`: the expected return of a stock given it’s returns are negative

$$= E[R_i | R_i < 0]$$

- `TVarHigh`: the expected return of a stock given it's returns are positive

$$= E[R_i | R_i > 0]$$

- `Sector`: business sector of company, includes: 'Real Estate', 'Communication Services', 'Materials', 'Industrials', 'Consumer Staples', 'Health Care', 'Consumer Discretionary', 'Information Technology', 'Utilities', 'Financials', 'Energy']
- `AnnualReturn`: a dependent metric, the return of an asset one year past the date of calculation
- `AnnualRiskPremium`: a dependent metric, the excess return over the risk-free rate one year past the date of calculation
- `AnnualMarketPremium`: a dependent metric, the excess return over the S&P500 return over the same year past the date of calculation

3a EDA – Scatterplots of Metrics vs. Annual Returns

A scatterplot was used to visually inspect any of the metrics/models for any correlation with the annual return, see Figure 3a. Some variables of interest may be Volatility, Sharpe, Beta, TVarLow and TVarHigh. While none of them stand out, some correlation may exist based on their plots

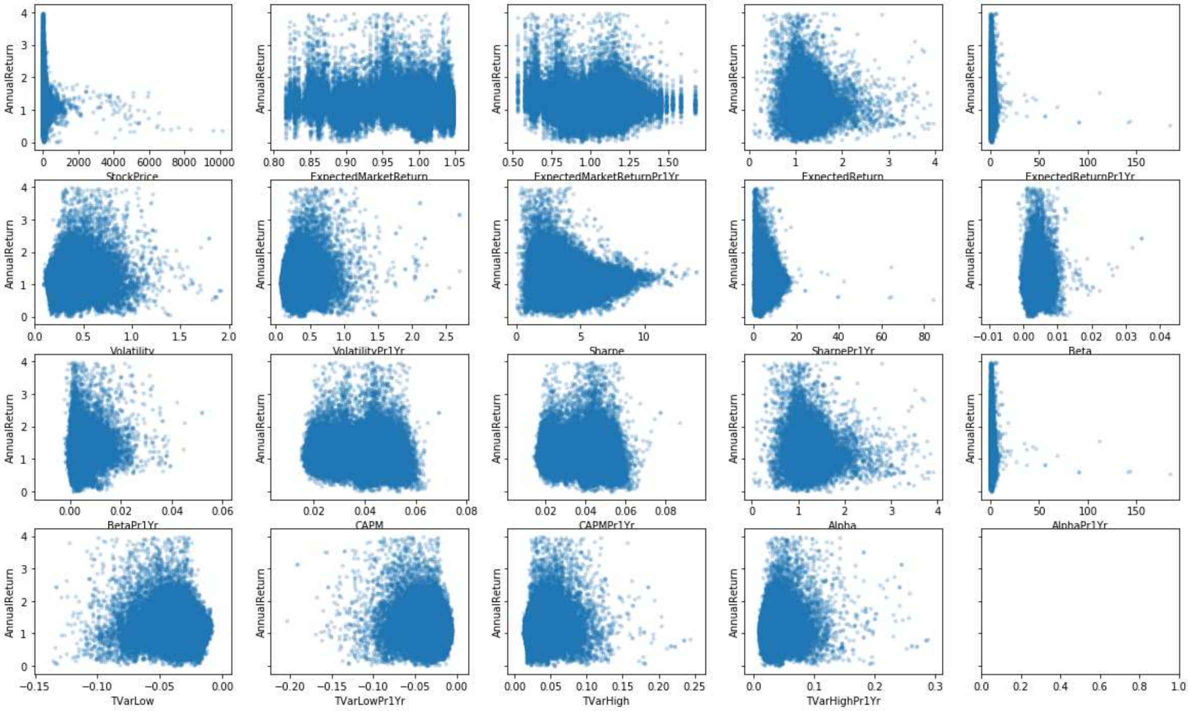

Figure 3a: *Shows a scatterplot of each of the metrics against the Annual Return of the stock*

3b EDA – Heatmap of Correlations among Parameters

A correlation heatmap was used to visually inspect any correlation between metrics/models, see Figure 3b. The correlation between many of the variables may be an issue and is likely due to the expected return calculation being used in most of the metrics calculated.

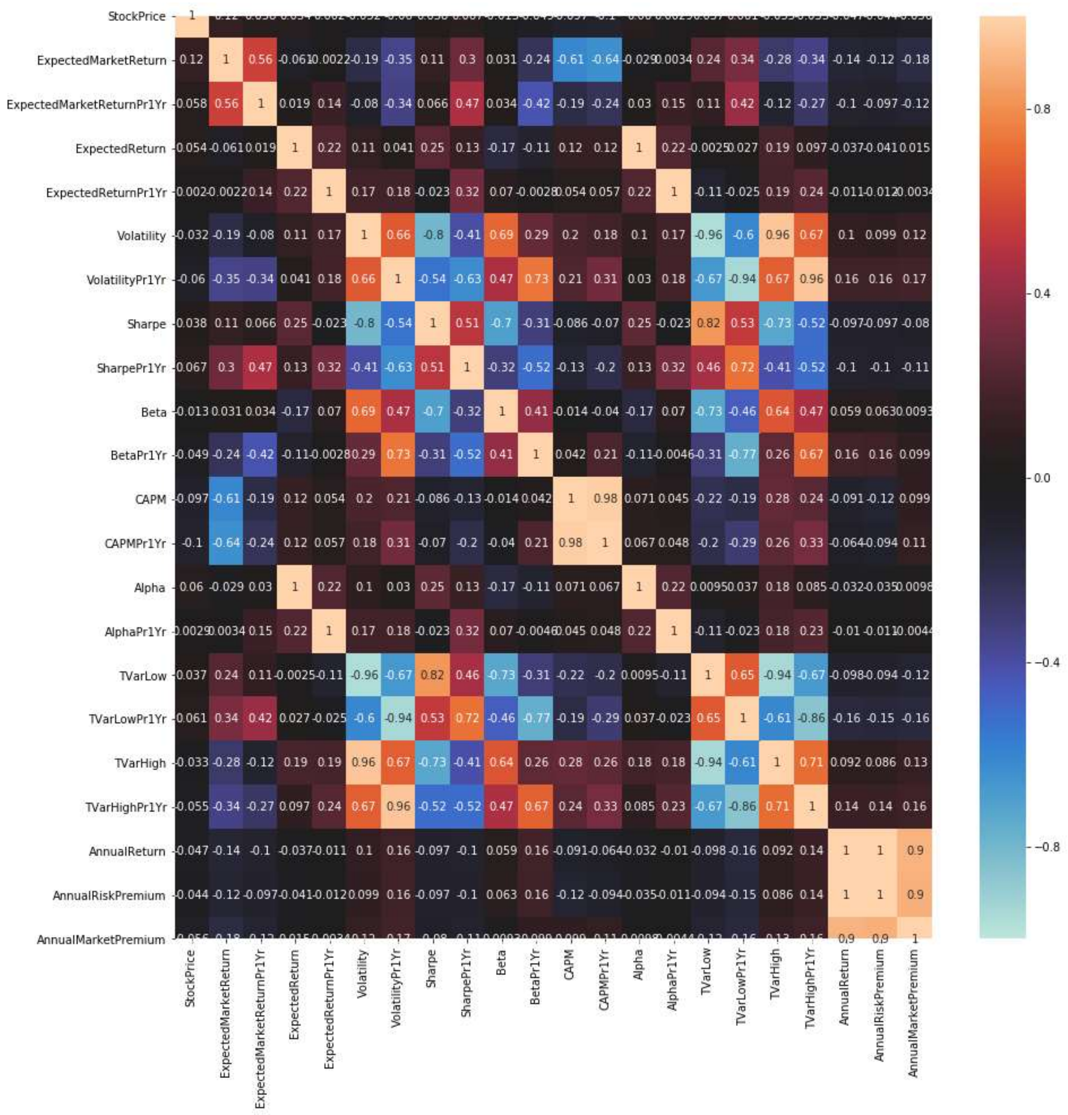

Figure 3b: *Shows a heatmap of each of the metrics/models against each other based on their correlation coefficients.*

3c EDA – Cumulative Distribution Function of Annual Returns among Parameters

To evaluate the distribution of annual returns based on values of each metric/model a cumulative distribution was used, see Figure 3c. This plot shows the cumulative probability of annual returns each metric/model. About all the models show between a 20%-50% chance of loosing value, but also that between 50%-80% change of getting a positive return. Models which are shifted to right have greater expected annual returns for the same probability. As the probabilities get above 50%, there is greater dispersion in annual return for a given probability.

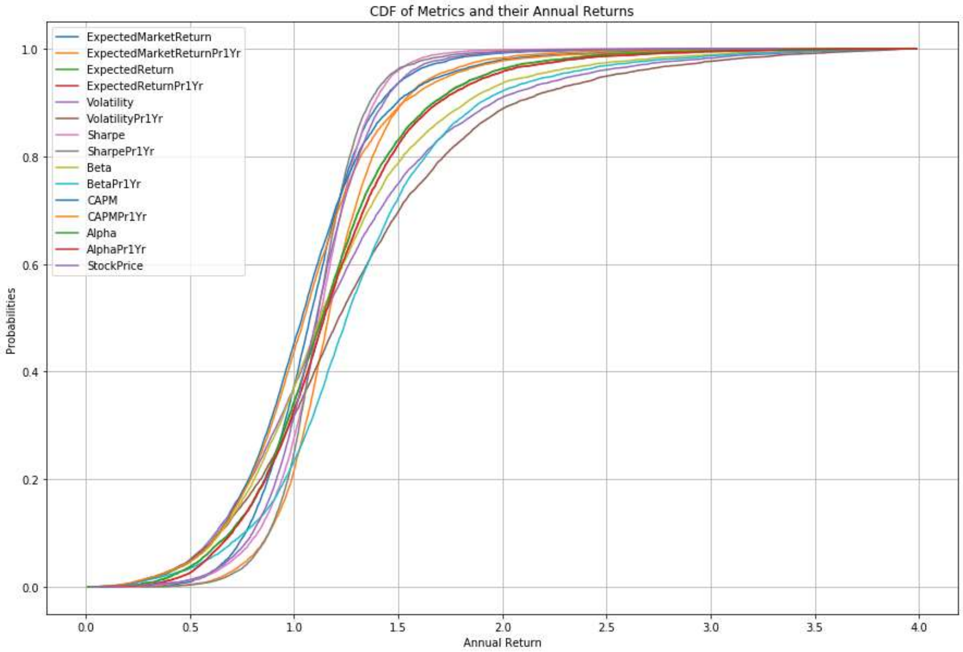

Figure 3c: *Shows cumulative distribution of each metric/model relative to its Annual Return*

4a - Implementation

The main idea for implementation is to test and train different models on the dataset. Assuming the models perform relatively well they will be put into an ensemble of different models and their collection of results will be used to analyze and predict different stocks.

4b Models and Results

The following models will be trained and tested for their accuracy and bias variance tradeoffs.

4b.1 – OLS Multiple Linear Regression

With OLS linear regression the first step was to regress all the parameters onto the dependent variable Annual Return and calculate the F-statistic. The F-statistic is used in hypothesis testing where the null hypothesis is that all the regression coefficients are equal to zero, and the alternative hypothesis is that at least one of the parameters is related (not equal to zero) to Annual Return. Running all the parameters against the dependent variable Annual Return resulted in a F-statistic much greater than one, and a F-statistic p-value of 0.00, indicating the null hypothesis can be rejected and there likely is a parameter that Annual Return is related to.

The first linear regression model was tuned and trained using forward stepwise selection and 5-fold cross validation. The model was fitted by averaging the residual sum of squares of the five folds then selecting the set of parameters which had the lowest average residual sum of squares. Figure 4b.1a depicts how the R-squared score and residual sum of squares react to the addition of parameters to the model.

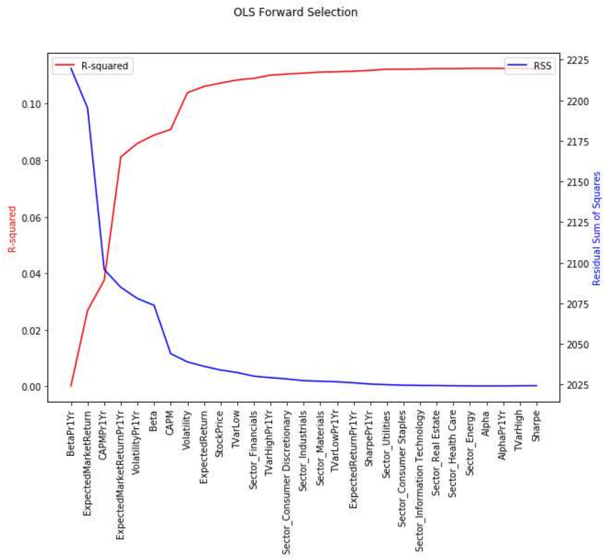

Figure 4b.1a: *The x-axis depicts the parameters and the order they were added to the model. For instance, 'BetaPr1Yr' was the first parameter added as it is the most left. Each x-tick/parameter*

shown implies the model was fitted using that parameter and all those to the left of it on the x-axis. For instance, ‘CAPMPr1Yr’ x-tick depicts the results of the model using ‘BetaPr1Yr, ExpectedMarketReturn, CAPMPr1Yr’.

None of the models perform particularly well, all have a low R-squared scores and the residual sum of squares cannot be reduced by more than 9% even with all the parameters. Below is the printout of the results of each model. It depicts the Parameters used in the model, the predicted average RSS followed by its 5-fold RSS scores, and the predicted R-squared score followed by its 5-fold R-squared scores.

```
--------
Parameters: ['BetaPr1Yr']
RSS: 2219.974 : [2246.21, 2224.2, 2148.63, 2233.63, 2247.2]
R^2: 0.0002 : [0.0288, 0.0171, 0.0226, 0.0352, 0.0294]
--------
Parameters: ['BetaPr1Yr', 'ExpectedMarketReturn']
RSS: 2195.586 : [2217.85, 2200.95, 2121.14, 2210.72, 2227.27]
R^2: 0.0266 : [0.0411, 0.0274, 0.0351, 0.0451, 0.038]
--------
Parameters: ['BetaPr1Yr', 'ExpectedMarketReturn', 'CAPMPr1Yr']
RSS: 2096.08 : [2111.93, 2100.69, 2020.05, 2116.23, 2131.5]
R^2: 0.0374 : [0.0869, 0.0717, 0.0811, 0.0859, 0.0794]
--------
Parameters: ['BetaPr1Yr', 'ExpectedMarketReturn', 'CAPMPr1Yr',
'ExpectedMarketReturnPr1Yr']
RSS: 2085.076 : [2099.42, 2093.3, 2009.98, 2102.94, 2119.74]
R^2: 0.0811 : [0.0923, 0.0749, 0.0857, 0.0917, 0.0844]
--------
```

```
Parameters:         ['BetaPr1Yr',         'ExpectedMarketReturn',         'CAPMPr1Yr',
'ExpectedMarketReturnPr1Yr', 'VolatilityPr1Yr']
RSS: 2078.112 : [2088.91, 2089.5, 2001.36, 2097.7, 2113.09]
R^2: 0.0859 : [0.0968, 0.0766, 0.0896, 0.0939, 0.0873]
--------
Parameters:         ['BetaPr1Yr',         'ExpectedMarketReturn',         'CAPMPr1Yr',
'ExpectedMarketReturnPr1Yr', 'VolatilityPr1Yr', 'Beta']
RSS: 2073.93 : [2086.55, 2081.45, 1999.83, 2094.14, 2107.68]
R^2: 0.0888 : [0.0978, 0.0802, 0.0903, 0.0955, 0.0896]
--------
Parameters:         ['BetaPr1Yr',         'ExpectedMarketReturn',         'CAPMPr1Yr',
'ExpectedMarketReturnPr1Yr', 'VolatilityPr1Yr', 'Beta', 'CAPM']
RSS: 2044.02 : [2062.76, 2053.72, 1976.4, 2058.79, 2068.43]
R^2: 0.0907 : [0.1081, 0.0924, 0.101, 0.1107, 0.1066]
--------
```

The biggest increase in R-squared occurs when ‘ExpectedMarketReturnPr1Yr’ is added to the model and so my final trained model will include ['BetaPr1Yr', 'ExpectedMarketReturn', 'CAPMPr1Yr', 'ExpectedMarketReturnPr1Yr', 'VolatilityPr1Yr', 'Beta', 'CAPM'] parameters.

Next the second linear regression model was tuned and trained using backward stepwise selection and 5-fold cross validation. The model was fitted by averaging the residual sum of squares of the five folds then selecting the set of parameters which had the lowest average residual sum of squares. Figure 4b.1b depicts how the R-squared score and residual sum of squares react to the removal of parameters.

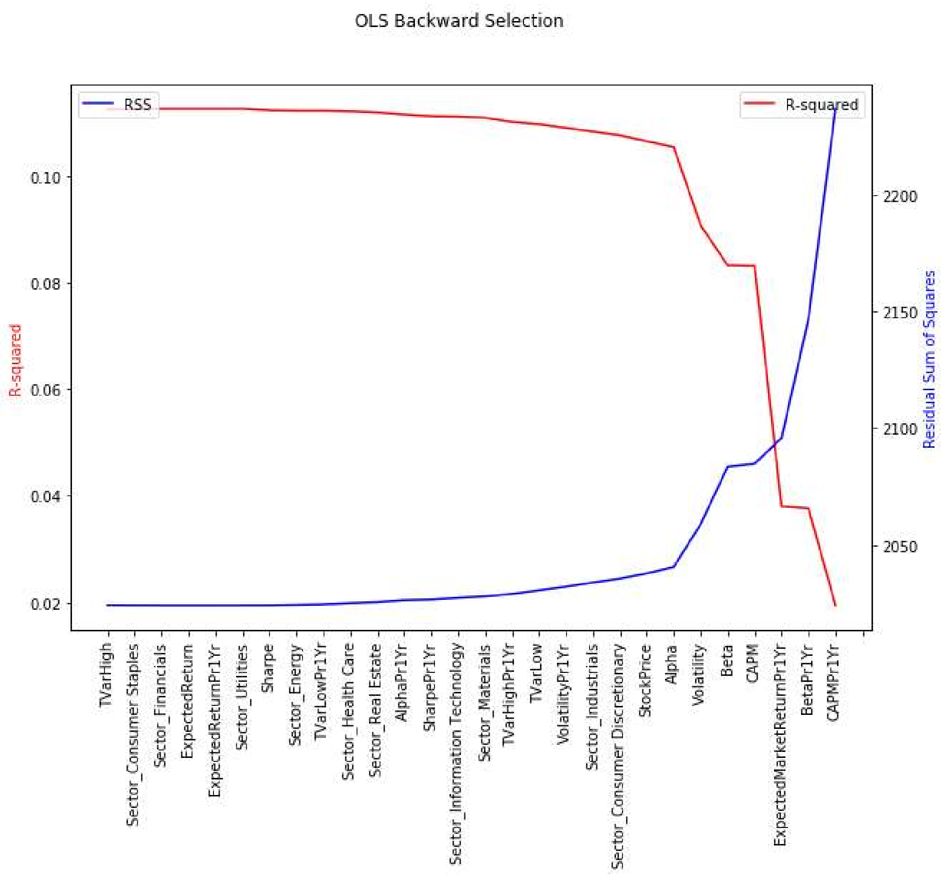

Figure 4b.1b: *The x-axis depicts the parameters and the order they were removed from the model. For instance, 'TVarHigh' was the first removed as it is the most left. Each x-tick/parameter shown implies the model was fitted using that parameter and all those to the right of it on the x-axis. For instance, 'CAPM' x-tick depicts the results of the model using 'ExpectedMarketReturnPr1Yr, BetaPr1Yr, CAPMPr1Yr, ExpectedMarketReturn,'.*

The biggest decrease in R-squared associated with a big increase in RSS occurs when 'Alpha' is removed from the model and so my final trained model will include [ExpectedMarketReturn, CAPMPr1Yr, ExpectedMarketReturnPr1Yr, CAPM, Beta, Volatility] parameters.

```
--------
RemovedParameters: Alpha
```

```
RSS: 2040.6 : [2046.37, 2090.12, 1991.52, 2074.4, 2000.59]
R^2: 0.1054 : [0.1088, 0.1041, 0.1075, 0.1002, 0.1063]
--------
RemovedParameters: Volatility
RSS: 2058.938 : [2066.99, 2114.44, 2007.87, 2092.08, 2013.31]
R^2: 0.0907 : [0.0998, 0.0936, 0.1002, 0.0925, 0.1006]
--------
RemovedParameters: Beta
RSS: 2083.482 : [2091.45, 2139.08, 2035.87, 2110.84, 2040.17]
R^2: 0.0832 : [0.0892, 0.0831, 0.0876, 0.0844, 0.0886]
--------
RemovedParameters: CAPM
RSS: 2084.74 : [2093.93, 2140.47, 2037.09, 2111.62, 2040.59]
R^2: 0.0831 : [0.0881, 0.0825, 0.0871, 0.084, 0.0885]
--------
RemovedParameters: ExpectedMarketReturnPr1Yr
RSS: 2095.778 : [2106.31, 2148.58, 2048.94, 2122.85, 2052.21]
R^2: 0.038 : [0.0827, 0.079, 0.0818, 0.0792, 0.0833]
--------
RemovedParameters: BetaPr1Yr
RSS: 2146.81 : [2163.02, 2189.52, 2099.63, 2174.0, 2107.88]
R^2: 0.0376 : [0.058, 0.0614, 0.0591, 0.057, 0.0584]
--------
RemovedParameters: CAPMPr1Yr
RSS: 2236.626 : [2247.39, 2290.59, 2188.84, 2261.23, 2195.08]
R^2: 0.0194 : [0.0213, 0.0181, 0.0191, 0.0191, 0.0194]
--------
```

Using the forward stepwise and backward stepwise selection methods yielded similar parameters. While the R-squared scored remain low, given the volatility and complexity of the stock market and that both selection methods chose similar models leads me to believe while the OLS multiple linear regression model is not great, given the parameters at hand it is simple and has a good bias variance trade-off.

4b.2 – Multiple Linear Ridge Regression

Next, a ridge regression was model was trained and tested. Ridge regression uses L2 regularization which helps to avoid overfitting. This is done using a tuning parameter known as *alpha* which minimize the values of the coefficients of the model. As alpha is increased, the coefficients are forced towards zero.

The ridge regression model was tuned and trained by adjusting the tuning parameter alpha and using 5-fold cross validation. The model was fitted by averaging the residual sum of squares of the five folds for each change in alpha. Figure 4b.2 depicts how the R-squared score and residual sum of squares react to changes in alpha.

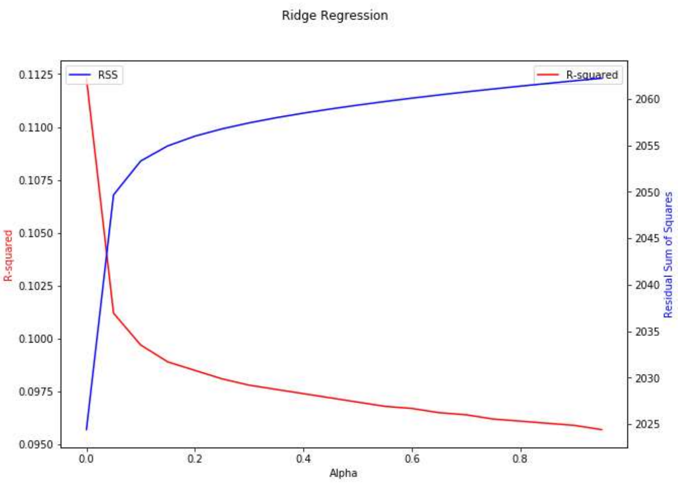

Figure 4b.2: *The x-axis depicts the tuning parameter alpha for Ridge Regression and its effect on RSS and R-squared*

For my data ridge regression performed best when alpha was closest to zero; resulting in the largest R-squared score and smallest RSS. This implies that using all the parameters with little to no tuning of alpha produces the same results as OLS linear regression with all the parameters and therefore ridge regression is inferior to OLS linear regression for this dataset. Below are the results of ridge regression for the smaller values of alpha.

```
--------
Alpha: 0.0
RSS: 2024.9763
R^2: 0.1122
--------
Alpha: 0.05
RSS: 2049.9599
R^2: 0.1012
--------
Alpha: 0.1
RSS: 2053.5621
R^2: 0.0996
--------
Alpha: 0.15
RSS: 2055.1746
R^2: 0.0989
--------
--------
Alpha: 0.2
RSS: 2056.2095
```

```
R^2: 0.0985

--------

Alpha: 0.25

RSS: 2056.9896

R^2: 0.0981

--------

Alpha: 0.3

RSS: 2057.6297

R^2: 0.0978

--------
```

4b.3 – Multiple Linear Lasso Regression

The last linear model to be trained and tested is the lasso regression model. Lasso regression uses L1 regularization which also helps to avoid overfitting. This is also done using a tuning parameter known as alpha which also minimizes parameter coefficients and even forces some to zero. Lasso regression performs its own parameter selection in this way. As alpha is increased, the coefficients are reduced, and some may equal zero causing that parameter to be ignored by the model

The lasso regression model was tuned and trained by adjusting the tuning parameter alpha and using 5-fold cross validation. The model was fitted by averaging the residual sum of squares of the five folds for each alpha. Figure 4b.3 depicts how the R-squared score and residual sum of squares react as alpha increases.

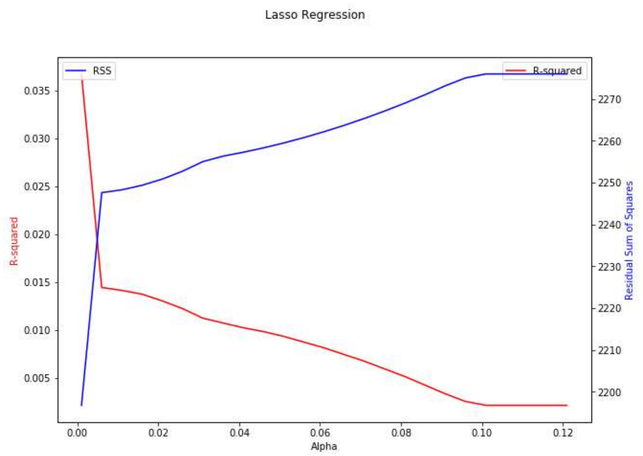

Figure 4b.3: *The x-axis depicts the tuning parameter alpha for lasso regression and its effect on RSS and R-squared*

For my data lasso regression performed best when alpha was closest to zero; resulting in the largest R-squared score and smallest RSS. However the largest R-squared score was around 4% and the RSS hardly changed. Therefore lasso regression has performed the worst when it comes to linear regression

```
--------
Alpha: 0.001
#Coef: 13
RSS:
2196.7366
R^2: 0.0367
--------
Alpha: 0.006
#Coef: 3
RSS:
2247.6154
R^2: 0.0144
--------
Alpha: 0.011
#Coef: 3
RSS:
2248.2991
R^2: 0.0141
--------
Alpha: 0.016
#Coef: 3
RSS:
2249.3828
R^2: 0.0137
--------
Alpha: 0.021
#Coef: 3
RSS:
2250.8667
R^2: 0.013
--------
Alpha: 0.026
#Coef: 3
RSS:
2252.7508
R^2: 0.0122
--------
Alpha: 0.031
#Coef: 3
```

```
RSS:
2255.035
R^2: 0.0112
--------
Alpha: 0.036
#Coef: 2
RSS:
2256.3446
R^2: 0.0107
--------
Alpha: 0.041
#Coef: 2
RSS:
2257.2739
R^2: 0.0102
--------
Alpha: 0.046
#Coef: 2
RSS:
2258.3237
R^2: 0.0098
--------
Alpha: 0.051
#Coef: 2
RSS:
2259.4939
R^2: 0.0093
--------
Alpha: 0.056
#Coef: 2
RSS:
2260.7844
R^2: 0.0087
--------
Alpha: 0.061
#Coef: 2
RSS:
2262.1954
R^2: 0.0081
--------
Alpha: 0.066
#Coef: 2
RSS:
2263.7268
R^2: 0.0074
--------
Alpha: 0.071
#Coef: 2
RSS:
2265.3786
R^2: 0.0067
--------
Alpha: 0.076
#Coef: 2
RSS:
2267.1508
R^2: 0.0059
--------
Alpha: 0.081
#Coef: 2
RSS:
2269.0433
R^2: 0.0051
--------
Alpha: 0.086
#Coef: 2
RSS:
2271.0563
R^2: 0.0042
--------
Alpha: 0.091
#Coef: 2
RSS:
2273.1897
R^2: 0.0033
--------
Alpha: 0.096
#Coef: 1
RSS:
2275.0781
R^2: 0.0025
--------
Alpha: 0.101
#Coef: 1
RSS:
2276.0121
R^2: 0.0021
--------
```

4c Conclusions

In conclusion I believe there are many improvements to be made. I believe I have a collinearity problem in my metrics and that is something to analyze more in depth.

After training and testing many models and not getting any concrete results I was unable to find any evidence that the efficient market hypothesis holds, and consistent superior profits are not possible. The results of the models also implies that the mean-variance portfolio theory is not effective, nor does it enable you to build a portfolio which beats the market simply on past pricing data.

5 References

1. Derivatives Markets – McDonald
2. Corporate Finance – Berk DeMarzo
3. Analysis of Economic Time-Series – (Kendal 1952) https://www.jstor.org/stable/2980947?read-now=1&seq=1#page_scan_tab_contents
4. Coaching Actuaries – https://coachingactuaries.com/
5. Investopedia – January Effect https://www.investopedia.com/terms/j/januaryeffect.asp
6. Investopedia – Reversal Effect https://www.investopedia.com/terms/r/reversal.asp
7. S&P400 Wikipedia - https://en.wikipedia.org/wiki/List_of_S%26P_400_companies
8. S&P500 Stocks - https://en.wikipedia.org/wiki/List_of_S%26P_500_companies
9. Yahoo-Finance API - https://pypi.org/project/yfinance/