Predicting ETF Prices:
A Simple Linear Regression Project

Steven Rivera and Team, Spring 2021

Section 1: Introduction

In Osaka, Japan 1697, the Dojima Rice Exchange was born which laid the groundwork for what we know as the modern banking system, then in 1730 Dojima ushered in the first spot market and futures market which evolved into what is now known as the modern stock exchange. Since that time, we have gambled, won, lost and learned the art that is the stock market and investment. The market as we know it, is as much about sentiment and policy changes as it is about investment. This is most evident during the frenzy of a bull market, where the common investor is caught up in the furor and fervor over gains and stock prices that seems like it could not possibly go down and a bear market where each loss paradoxically leads to a big sell-off and further loss.

As the market has evolved, many have attempted to predict prices to hedge against losses and increase gains through the market, yet from its humble beginnings in Dojima to today, it is still just as difficult to predict where prices will go in the long term and just as difficult if not impossible to pinpoint where a bull or bear market will begin and end. Certainly, with technology and the flow of information and a better understanding of key fundamental factors we can generalize the direction of gains and losses, but, again, the nature of the market is dealing heavily with unknowns that are a function of the political landscape, monetary policy, individual company performance against predicted performance, etc.

The market has helped alleviate the risks of investment by creating a basket of products that tracks the entire market, each sector and other specialized collections of stocks based on type, goal, morality, etc. These are called Exchange Traded Funds or ETFs. When purchasing one share of any ETF, the buyer is really purchasing a small percentage of each company being tracked by that ETF providing an instant diversification or spread of the risk. If any single company being tracked by an ETF happens to severely underperform or goes bankrupt, the investor will lose only a fraction of the original investment based on the percentage of the ETF's holdings. This will allow for both the natural increase in value for an ETF over the long term, but also provide a less volatile rise in price with steadier highs and lows, making it an attractive way to invest for less risk averse retail investors. Considering the smoothing and normalization of price fluctuations against risk, we believe ETFs may also allow us to build a model that could better track and predict prices in the market.

Section 2: Data Preparation

Our dataset contains 1,680 exchange traded funds (ETFs) with categorical variables describing the form, function and goals of each fund and numerical variables detailing the breakdown of stocks, sectors and credit ratings of the fund. Additionally, each fund provides past performance and market benchmark comparisons. The data was acquired from www.kaggle.com. The title of the dataset from www.kaggle.com is “US Funds dataset from Yahoo Finance”. The 1,680 ETFs in the dataset were scraped from www.finance.yahoo.com.

To prepare the data for modeling, we sorted through the columns and determined which variables would be unnecessary or unhelpful. In the end, we removed most categorical variables that would have no bearing on our analysis. This includes things like fund family (iShares, XGlobal, SPDR, etc.), investment strategy, currency type (all ETF prices are listed in the US market and as such are already in USD), the size type (small, medium, large) and investment type (growth, value, blend).

Furthermore, we removed the sector variables (like tech, commodities, finance, healthcare, industrial, etc.) , which provide a percentage breakdown of an ETF's holdings by each sector and weight of agency ratings that range from AAA to C and lower and also include another category for 'Other' that is a rating provided when a company is held in the ETF, but is in default (usually denoted with a D-F)

After removing these variables we are left with many real-number data points that would also need to be reviewed, sorted and removed. We started with the question of what we want to evaluate and we are left with a baseline of ETF returns from 2013-2017 while using the 2018 returns as our comparative variable for prediction. That allowed us to remove a good portion of real-number variables. To start that section, we removed price/earnings (P/E) ratio, price/book (P/B) ratio, price/sales (P/S) ratio and price/cash flow (P/CF). Next, we removed the category returns over the short and long term. These variables provide a return on category for time values ranging from one (1) month to one (1) year, including year-to-date (YTD), five (5) year and ten (10) year.

Finally, we are left with ETF returns from 2013-2018. To complete our usable data set, we removed ETF rows without returns from all three (3) years 2014-2017. This leaves us with a dataset containing 896 ETFs and six (6) features that include ETF returns from 2013-2018.

Section 3: Data Analysis

After thoroughly analyzing the histograms for our ten numerical variables shown in **Figure 1** and **Figure 2**, there were three interesting observations that we made. The first observation that we made was that the *asset_stocks* and *asset_bonds* histograms were polar opposites of each other. The *asset_stocks* histogram shows that a majority of ETF portfolios are primarily made up of stocks, and a minority of the ETF portfolios have zero stocks in them. On the other hand, the *asset_bonds* histogram shows that a minority of ETF portfolios primarily consist of bonds, and a majority of the ETF portfolios have zero bonds in them. The next observation that we made was that the *price_earnings_ratio* histogram is skewed to the right and most of the ETFs have a low price-earnings ratio. The final observation that we made about our histograms was that the Fund Return histograms between the years 2013 and 2018 were very random when it comes to how the ETFs were performing. The two Fund Return histograms that seemed to show the highest median and mean were the histograms for *fund_return_2013* and *fund_return_2017*.

Once we completed analyzing the histograms for our data set, we decided to look at the five number summaries for our numerical variables. While there are a lot of different ways to use the five number summaries for our variables, we found that the most important things to compare were the interquartile range, maximum, and minimum of each fund return variable. When comparing the interquartile range of each variable with each other, we noticed that the interquartile range for *fund_return_2013* is significantly higher than the other fund return variables. This shows us that a majority of the values for *fund_return_2013* are more spread out compared to the values of the other fund return variables. After that, we compared the maximum and minimum of each numerical variable. The main variable that we decided to focus on for this was our dependent variable. The range for *fund_return_2018* was 71.5, which is less than half of two of our independent variables. This indicates that the outliers for our dependent variable are not significantly far away from the interquartile range compared to most of our independent variables.

Next, we looked over our scatter plot matrix shown in **Figure 3** to see if there were any interesting findings. The first observation that we made was that *asset_stocks* and *asset_bonds* have zero correlation with all of the other variables. The next observation that we made was that it was very hard to tell if there was any correlation between all of the fund return variables. For the most part, the different values of each variable were scattered all throughout the scatter plot.

Finally, we looked at the correlation matrix to see the correlation between each of our variables shown in **Figure 4**. The first observation that we made was that our dependent variable, *fund_return_2018*, has either a weak correlation or no correlation with every independent variable not including *fund_return_2014*. When it comes to the correlation between *fund_return_2018* and *fund_return_2014*, we see that both variables have a moderate negative correlation between each other. The next interesting observation that we made was that both *asset_stocks* and *asset_bonds* had little to no correlation with any of the other variables in our data set. While this was the case, both *asset_stocks* and *asset_bonds* had a strong negative correlation between each other.

Based on the data output so far we can not say definitively that previous year returns for ETFs can be a predictor of future returns. We understood prior to choosing our dataset that the chaotic nature of the stock market would not lend itself to a simple task of predicting future returns, however we believe, like very long term stock prices, there would be a pattern to follow that might allow us to build an accurate model and at this stage we think there is still a chance of doing that.

Our histogram output shows us the rate of return of the funds for the given year, we can see from above in the individual histograms there is a noticable difference to shape and trend for our independent variables vs our dependent variable (fund return 2018). However, since there appears to be a slightly similar shape of the histograms for 2014, 2015 and surprisingly for the fund return YTD as well, there could be some merit to featuring those variables, but we will need to dig deeper when we build our models.

Further, we look at the five number summary and we again can't find a close, direct correlation from the previous return years against the return year for 2018. The values of the mean are of particular interest in this case. Certainly, the max and min are useful, but focusing only on the mean, we have a negative value where all other return years except 2015 are positive, some like 2017 and 2013 have highly positive values which could skew our prediction values possibly making any model built around those values wrong and non-functional. We do find a somewhat similar spread and center for 2018 vs. fund return YTD, though the min and max values are going to distribute the data in a wider range than 2018 which could cause issues with a prediction model.

The scatterplot summary is very telling of issues we may run into as we create our models. As described above, the nil to very weak relationship for 2013, 2016, 2017 and fund return YTD might make it impossible to accurately determine future values with any degree of certainty. There is hope for a good model for 2014 and 2015, even with a positively linear but weak relationship. We can confirm this with the correlation matrix where return YTD, 2014 and to a lesser but hopefully still significant degree, 2013 and 2015 shows a positive linear relationship does exist.

Since we have chosen to work through a dataset and are trying to predict a subset of the stock market, we recognize some of the challenges we face. There is a highly variable rate of return for each of the years per the histograms for the corresponding years, making the chances of prediction suspect. Additionally, the variables for Asset Stocks and Asset bonds, while on the surface are helpful for understanding the makeup of an ETFs holdings, may, on reflection, not be a helpful or accurate predictor. We are also going to probably need to dismiss return YTD and PE ratio, since the values account for 2020 returns, though we believe there could still be some correlation for current year returns and the oversold/overbought positions for ETFs during a dip in the market like in 2018 vs 2020. There is a question about the performance of the market on a downturn due to the covid pandemic in 2020 and how it would also reflect the volatile and negative market from 2018 due to a collection of negative signals like the trade war with China and public sentiment regarding the Federal reserve raising interest rates too quickly. As stated, we do find an interesting weak linear relationship, however we aren't positive it will be a useful indicator for a prediction model.

Based on the distinctive graphs and insights we have discussed so far, now we can carry out some data analysis with R language. We will use multiple regression models taking fund returns in 2018 as dependent variables. The aim is to discover what factors and how they affect the current year returns. First of all, we should use some plots to see whether they have a strong relationship with fund return in 2018. This step is important because adding too many variables without selecting will greatly increase the burden. Some irrelevant variables will also cause bad effects on the model analysis.

We will include asset stocks, asset bonds, price earning ratio, fund_return_ytd, funds return of recent years as explanatory variables to model fund return in 2018. Among these independent variables, price earning ratio and fund_return_ytd (fund return year to date) are two important ones that have a strong linear relationship with the dependent variable. Price earning ratio stands for the value of a stock. The higher it is, the more earnings investors will get. Fund return year to date means the amount of fund return accumulated in the single year. It will also attract people to invest more money in the stock market. But 2020 is unique due to Covid-19 Pandemic, which also affected the stock market to undergo unprecedented changes. Therefore, the final model may be not suitable for prediction in fund return of 2020.

As we look through the data statistics, we find that the current year return may inflate hugely. That's because of the psychology of speculation. People are money chasers. When they see an attractive stock most of them would blindly follow suit and spend lots of money on it. As we can imagine, such people may not have a harvest year in the end. Therefore, we decide to include extra columns as variables like annual return or difference of two successive years of past several years or annual difference of and assign them specific weight values. In this way, we can include the influence of recent years.

We also need to take many other things into consideration like interaction terms and second order terms. For example, some data shows that a relatively lower return of last year and a lower price earning rate might work together to boost fund return in the next year. So we should take price_earning_ratio* fund_return as an interaction term. We will add and delete some independent variables in the model to check the F-test and T-test so that we can decide whether to include them or not.

Section 4: Model Building

This project will explore price prediction in the stock market through multiple linear regression analysis to try and build models that can accurately predict prices specifically in the ETF market. This will be implemented using a dataset that contains nearly all ETFs that are currently available for trade in the United States and are readily available to trade on at least one of the major stock exchanges. We will be focusing our efforts on differing aspects of the data in an attempt to use ETF prices from 2013-2017 to build a model that can predict 2018 ETF prices.

Stephanie:

For my individual milestones, I will be comparing the predictive performance of two regression models. One that is trained and tested on the Stock ETFs and the other that is trained and tested on the Bond ETFs. I will first show how I created a regression model to predict 2018 ETF returns using only the Stock ETFs and will then show how I built a regression model using only the Bond ETFs. I will then compare the two models and their predictive performances.

I started off by doing some preliminary steps before modelling. First, I plotted a histogram of all of the features to determine if any of them needed transformations. I found that all of the distributions were approximately normal, so I decided not to transform any of the variables. I then plotted the correlation matrix to check the pairwise correlation between all pairs of features. I found that none of the features had high enough correlation values to cause multicollinearity. I also found that all VIF values of the features were at acceptable levels.

Because our feature set if very small, I first tried to make a regression model using all of the explanatory variables (*fund_return_2013, fund_return_2014, fund_return_2015, fund_return_2016 and fund_return_2017*) to predict *fund_return_2018*. This resulted in a model with an adjusted-R2 value of 0.313. Because this value was so low, I decided to add interaction and second order terms.

I first added all 10 possible interaction terms to the original dataset. I then used both forward selection and backward elimination to determine which interaction terms to include in the model. I found that forward selection and backward elimination produced the same models. After this step, the regression model was made up of the following variables: *fund_return_2013, fund_return_2014, fund_return_2015, fund_return_2016, fund_return_2017, fr_2013*fr_2014, fr_2013*fr_2016, fr_2013*fr_2017, fr_2015*fr_2017, fr_2016*fr_2017* and reached an adjusted-R2 value of 0.468.

I used a similar procedure to determine which quadratic terms to include in the model. I first added all of the squared features to the dataset. I then used forward selection and backward elimination to determine which features to include in the model. Again, the resulting feature sets from forward selection and backward selection were the same. Using this technique, I found that the significant quadratic terms to include in the model were *fund_return_2013_SQ, fund_return_2015_SQ, fund_return_2016_SQ,* and *fund_return_2017_SQ.*

Next, I created a regression model with all of the original explanatory variables, the interaction terms and the quadratic terms determined above. Here, I found the F-test value was significant, indicating that at least one of the Betas in the model is not equal to 0. I then evaluated the individual t-test. I found that two of the terms were not significant and decided to drop them. This resulted in my final model with the following explanatory variables: *fund_return_2013, fund_return_2014, fund_return_2015, fund_return_2016, fund_return_2017, fr_13*fr_14, fr_13*fr_16, fr_13*fr_17, fr_15*fr_17, fund_return_2013_SQ, fund_return_2015_SQ, fund_return_2016_SQ* and *fund_return_2017_SQ*. This model reached an adjusted-R2 value of 0.533, meaning that 53.3% of the variance in fund_return_2018 is explained by the model.

Finally, I created an 80-20 train-test split and 10-fold cross validation to evaluate the prediction performance of the model. I first trained the model on 80% of the data and tested that model on the other 20% of the data. I created a correlation matrix of the value predicted by the model vs. the actual 2018_fund_return values in the test set to assess the performance of this model. They appeared to be slightly correlated with a correlation value of 0.619. I then performed 10-fold cross validation and found the average MSE to be 92.9 or 9.6 R-MSE.

I used a very similar procedure to create my second regression model that predicts 2018 ETF returns using only the Bond ETFs. I started off with some preprocessing steps. I first performed two tests to check for multicollinearity in the features. I checked for pairwise correlation between all pairs of features using a correlation matrix. The only very strong correlation that I noticed was between *fund_return_2017* and *fund_return_2018*, however because *fund_return_2018* is our response variable, I decided to leave both of the features in the model. Next, I computed the VIF values for the independent variables in the model. All had values less than 3, so I again decided not to drop any features.

I then plotted the histograms of all of the features to evaluate their distributions. I found that all of the independent variables looked approximately normal and decided not to transform them. I did notice that the response variable had a left skewed distribution. Because there are negative values in this variable, I was not able to perform a log transformation. I did try a cube root transformation on the response variable, but it did not make a significant improvement on the adjusted-R2 value and made the model more difficult to interpret, so I decided not to include this transformation.

Next, I created my first regression model. I used Forward Selection and Backward Elimination to determine which first order terms to include in the model. Both of these techniques resulted in a regression model that included all 5 of the independent variables. This base model had an adjusted-R2 value of **0.8419**. I then created possible interaction terms and used Forward Selection and Backward elimination to determine which interaction terms to include in the model. Based on their t-test p-values, I found that the following interaction terms were significant enough to include in the model: *fund_return_2013*fund_return_2015, fund_return_2013*fund_return_2016* and *fund_return_2014*fund_return_2016.* Adding these interaction terms increased the adjusted-R2 value to **0.8825**. I then added all 5 quadratic terms and found that the only significant quadratic term was *fund_return_2013_SQ.*

My final step before evaluating my model was performing Residual Analysis. The first assumption about the residuals that I check is whether the residual had a mean of zero. I used R's **mean()** function to calculate this and found that the sum of the residuals was indeed zero. Next, I plotted a histogram of the residuals and found that they were approximately normal. I performed the Durbin Watson test on the model to check for independence of errors. Here, I found that the test had a p-value of 0.6, meaning that we cannot reject the null hypothesis that the residuals are not autocorrelated. Finally, I plotted all of the independent variables against the residuals to check for homoscedasticity and found that all of the variances were approximately constant.

Lastly, I used cross validation with an 80-20 train-test split and 10-fold cross validation to evaluate the prediction performance of the model. After training the data on the 80% training set, I used this model to predict the values in the 20% test set. The predicted values and actual values had a high correlation value of 0.844. The plot of the predicted values against the actual values is shown below. Due to the very small size of the Bond ETF dataset, this metric alone is not good enough to fully evaluate the model. To get a better picture of the model performance, I then performed 10-fold cross validation. The average MSE across the 10 folds was 3.78 with values of 27.4, 0.26, 0.87, 3.3, 1.74, 0.41, 0.22, 4.84, 0.26 and 0.39.

Based on the Adjusted-R2 value, we can see that with a value of 0.8827, the Bond ETFs model is much better at explaining the variance in 2018 ETF returns than the Stock ETF regression model with an Adjusted-R2 value of 0.533. Additionally, based on the correlation values and average MSE values, it is clear that the Bond ETF regression model also performs much better at predicting 2018 ETF return values.

Steven:

Naturally, the entire market will sometimes underperform for the fiscal year and this will include the ETF market. While less volatile, we can still expect negative returns in a downturn year. This section will focus specifically on fund return years that ended in a negative average return as we believe it could indicate the return to the natural average from years where hyper-increases in prices drive all prices beyond fair market values. For that reason, I attempted to build a model with a dependent variable that had a negative average return (2018) and begin with a single independent variable that also had a negative average return.

To start building the model, I simply ran a summary on the cleaned dataset and looked for a negative average return in the independent variables to correspond to the negative average return of the dependent variable's mean return value of -9.807. I only needed to find a negative value, the number value is of no consequence to the overall model as long as the average is negative. After looking at all other mean returns from 2013 to 2017 I found the only other negative average return was in 2015 with a value of -4.74. This allowed me to build a model that would include one dependent variable (ETF returns for 2018) and one independent variable (ETF returns for 2015).

After determining that I would be building a simple linear model based on negative average returns, I was able to analyze the statistical relationship between the dependent and independent variables and found there is a positive linear relationship, though a moderate to weak one. The linear regression analysis passed the T-Test and F-Test that gave an indication of a good model and an R-Squared value of 18.0% which was a cause of concern since we would hope the variability of the data would be explained by the model in a more robust way, however this is not altogether unexpected due to the nature of the question we are trying to answer. Though it does mean further digging was required to fully understand the relationship of the variables.

This led to in-depth residual analysis that was more than a little disappointing as not much more was gleaned with respect to the simple linear model. The first test was checking if the mean value was zero (0). For my model comparing returns for 2018 against returns for 2015, the mean value of the residuals was: 1.604411e-13. Naturally, this is not zero (0), but I felt it was an acceptable value and moved forward. Next, I looked at the distribution of the residuals and found that while normal, it was skewed with many peaks and decided it would be valuable to check the distribution against normalized z-score values (bottom). Other than small variations, there was little to no change in the output.

I then performed a series of tests of the residuals that checked the soundness of the model and the viability of the variables that were used. First, I tested the residuals against the fitted values to find a large and varied scatter of the data as expected from the summary output that also had no values that weighted heavily enough that removing them changed the output in any significant way. I can say with certainty there is no sense of any pattern and it must be concluded that the data is homoscedastic, which suggests that there is no value in transforming either the independent or dependent variables and the final model would remain a simple linear regression model.

In the end, I find that this model is only slightly acceptable for prediction of prices in the ETF market. An already difficult and nearly impossible thing to do in predicting the stock market is not made easier with this dataset or model and throws into sharp relief the unknown unknowns that investors and predictive analysts face. I do not find this to be a 100% impossible task, but the randomness and variability in returns year to year seem to be the biggest hurdle.

Michael:
Stock prices are affected by a number of different factors, which include inflation, economic policies, and the condition of the company. Since these factors influence the prices of stocks, they can also be used to predict the prices of individual stocks. In order to simplify the problem of having many different features affecting the stock prices, I have decided to focus on the fund returns from previous years to see if it is possible to build a regression model using data from previous years.

I will first be building a model using the two oldest fund return variables as the explanatory variables. I will also consider adding interaction terms and second-order terms. This is being done to see how much predictive power a model using older fund returns has.

To begin with, I needed to decide which explanatory variables to attempt to include in this model. Since the fund return variables for 2016 and 2017 would be included in the next model, they will not be included in this model. Also, the fund return variable for 2015 will not be included in either model, since the fund return in 2015 cannot be considered an older or newer fund return. In addition, the variables for the fund symbol and the fund extended name will be excluded from this model. These two variables are categorical variables, which means that it would be better to not include them in the model.

The final selected explanatory variables for this model include *fund_return_2013*, *fund_return_2014*, and *asset_bonds*. I have also included two second-order terms and one interaction term to the model. The interaction is between *fund_return_2013* and *fund_return_2014*. The second-order terms are both of the fund return variables squared.

During the process of checking which of the explanatory variables should be removed from the model, the only variable that was removed was *asset_stocks*. This variable was removed due to there being multicollinearity in the model. This was because both *asset_stocks* and *asset_bonds* were highly correlated with each other. Therefore, one of these two variables needed to be removed from the model.

After the model had been built, it was important to evaluate the MSE, the R-squared, and the F-test of the model. The MSE of the model is 41.8729, the R-squared is 0.4505, and the F-test is significant. The F-test tells us that at least one of the betas of the independent variables is not equal to zero. The R-squared tells us that 45.05 % of the variability in the dependent variable can be explained by this model. The MSE is telling us that there are a lot of outliers in the dataset, but the model is not overfitting the training data.

After looking at the t-test for each independent variable, it was found that all of the independent variables were significant in this model. This means that the beta for each independent variable is not equal to zero.

When running the backwards and forwards selection processes, the results are similar to what I ended up with for my final model. The backwards selection process had the exact same results for the final model. On the other hand, the forwards selection process did not add the interaction term between the fund returns for 2013 and 2014.

Next, I will be building a model using the two most recent fund return variables as the explanatory variables. I will also consider adding interaction terms and second-order terms. This is being done to see how much explanatory power a model using newer fund returns has. Then I will transform some variables in both models to see if there is any improvement in the models. Lastly, I will compare the models to see which one has more explanatory power.

The first thing that I needed to do was decide which explanatory variables to attempt to include in this model. Since the fund return variables for 2013 and 2014 were included in the previous model, they will not be included in this model. Also, the fund return variable for 2015 will not be included in either model, since the fund return in 2015 cannot be considered an older or newer fund return. In addition, the variables for the fund symbol and the fund extended name will be excluded from this model. These two variables are categorical variables, which means that it would be better to not include them in the model.

The final selected explanatory variables for this model include *fund_return_2016, fund_return_2017*, and *asset_bonds*. I have also included one second-order term and one interaction term to the model. The interaction is between *fund_return_2016* and *fund_return_2017*. The second-order term is the fund return variable for 2017 squared. Unlike the previous model where there were two second-order terms for the fund return variables, this model only has one second-order term.

During the process of checking which of the explanatory variables should be removed from the model, the only variable that was removed was *asset_stocks*. This variable was removed due to there being multicollinearity in the model. This was because both *asset_stocks* and *asset_bonds* were highly correlated with each other. Therefore, one of these two variables needed to be removed from the model.

Next, I attempted to transform some of the independent variables and the dependent variable in both of the models. The main independent variables that were focused on were the fund return variables in both models. After I performed a transformation on the specific variables in each model, I decided that there was no benefit to transforming any of the variables in both models. For most of the transformations, the distribution of the values started to become abnormal.
After building both of the models using training and test data, it was important to evaluate the MSE, the R-squared, and the F-test of both of the models. Starting with the older model, it passes the F-test, the adjusted R-squared is 0.463, and the MSE is 43.1. Since the model passes the F-

test, we can reject the null hypothesis that says that all betas equal zero. The adjusted R-squared indicates that 46.3 percent of the variability in the dependent variable is explained by this model. For the newer model, it passes the F-test, the adjusted R-squared is 0.404, and the MSE is 48. Since the model passes the F-test, we can reject the null hypothesis that says that all betas equal zero. The adjusted R-squared indicates that 40.4 percent of the variability in the dependent variable is explained by this model.

After looking at the t-test for each independent variable, it was found that all of the independent variables were significant in both of the models. This means that the beta for each independent variable is not equal to zero.

Overall, the older model has a higher R-squared compared to the newer model. This means that the older model is able to explain more variability in the 2018 fund return variable. Also, the older model has a correlation of 0.609 with the actual data, and the newer model has a correlation of 0.534 with the actual data. This also shows that the older model has a better chance of predicting the values in the actual data compared to the newer model. All in all, it is safe to say that the older fund returns have more significance when it comes to explaining the variability of the 2018 fund return.

Lizhi:
Stock price is influenced by a number of factors including political and economic policies, market interest rates, inflation, and companies' conditions and some other features. Based on our dataset, to simplify the problem, one useful method to predict the stock price is to look at the prices of past few years and find out the change patterns. The change patterns may be related to the business nature of the company, and we could try to extract clusters from them.
In this report, I focus especially on predicting prices without the data of bond ETFs and discuss whether this model without bond ETFs will work better and provide more accurate predictions than models with bond ETFs. My individual effort is to discover whether removing variables of asset_bond will make a difference, so I removed funds data that are composed mainly of bonds. I applied a linear model in this research and used the method of stepAIC to find the most proper model.

In Individual Milestone 1, I discussed with my group members and understood the task I needed to do. Then I built a basic model and modified variables and parameters to find the most suitable one. We have already cleaned data before. Now the data has 12 total columns, which are fund_symbol, fund_extended_name, asset_stocks, asset_bonds, price_earnings_ratio, fund_return_ytd and fund_returns from 2013 to 2018. However, some variables are not very relevant to this study and need to be removed. Besides, some data are missing and we need to delete them to prevent any errors in later research. After going through the overall data, I find that asset_bond and asset_stock are percentage numbers that are either near 100% or 0%. So we can transform the float value into integer 1 or 0. The final selected explanatory variables are fund_return from 2013 to 2017 and second order terms of them and interaction terms of recent 3 years. The explanatory variable is fund_return in 2018.

I removed some columns. Because they are just labels of explanation and don't contain any useful values such as fund_symbol and fund_extended_name. To make our study easier to carry out, we also removed price_earning_rate. In this model, MSE is 6.286. Adj-R2 is 0.4802. F-value is 63.31. P-value < 2.2e-16. The analysis of t-tests are as follows. The final explanatory variables are fund return in 2017, 2015, 2014 and 2013. Second order terms included are 2017, 2016, 2015 and 2013. The interaction terms of 2017 and 2016 are also useful. Fund_return_2015, fund_return_2014, SQ17 and inter17_16 are extremely significant. Fund_return_2017, SQ16 and SQ15 are very significant. Fund_return_2013 and SQ13 are not very significant but they increase the adj-R2. Therefore, we also include them in our model. After trying backward and forward selection processes, I got some different values. But the value didn't change greatly. The model with highest adj-R2 has been mentioned in the last paragraph. We've used cor() to check if there is any multicollinearity.

In Individual Milestone 2, I learned about the theory of durbinWaston test and residual test and applied those methods in my research. I also compared results from other group members and combined some useful ideas. First, I drew a histogram of the residuals and checked whether the residual is normally distributed. Second, I ran the durbinWatson test, here are the results. P-value indicates that we should optimize the model next time. I also plotted(model1) to see the relationship between residuals and other factors. From the Q-Q plot, we can see this dataset is normal. In the last graph, some points are labeled out and indicate that we should remove them to make our model suit the dataset better. In sum, it appears that fund return is most related to its history record, especially in the last 3 years. Some second order terms and interaction terms also work.

Section 5: Discussion

We have chosen the difficult task to build linear regression models that can accurately predict prices in the ETF market. Our dataset provided the focus of the goal that previous return years in stock and bond ETFs could be valuable in predicting future prices. We determined from the outset that prices could be predicted based on positive or negative return years, when comparing to stock ETFs vs bond ETFs and if there is consistent returns when comparing the two most recent return years vs. the earliest two return years when viewed as a cycle. Here we present those models and findings with a comprehensive and thorough analysis of these questions based around returns in the ETF market for 2018.

Our analysis involved using different subsets of the dataset to model 2018 fund returns. First we created a regression model to predict 2018 fund returns using only the 2015 fund return feature. For our next model, we used the two most recent fund return years to model 2018 fund returns. Next, we divided that data based on whether the ETFs were composed of mainly stocks or bonds. For these two subsets, we created multiple regression models to predict 2018 fund returns using the fund return values for 2013-2017. Lastly, we created a multiple regression model that did not consider bond data. In developing these models, we used both forward selection and backward elimination. Additionally, we considered interaction terms, quadratic terms and residual analysis when developing these models.

Our results suggest several key findings. First, using data from the past three to five years shows better prediction performance than just using data from last year. Second, according to adjusted-R2, ETFs in the bond market can be predicted more stably and reliably than that in the stock market. Third, we find that data in 2013 and 2014 have a much clearer linear relationship with data in 2018, indicating that it's better to apply longer-period data in stock price prediction. Last but not least, to predict the ETFs of 2018, the interaction term of 2016 and 2017 is also helpful.

Steven built a model that attempted to track and predict prices in a downturn year where the entire ETF market ended the year with negative returns. The dependent variable, returns in 2018, was compared against the only other negative return year in our dataset, 2015. He found that his simple linear regression model produced weak results that are only slightly acceptable for prediction of prices in the ETF market. The model passed the T-Test for the independent variable as well as the F-Test that allows us to accept that return year 2015 is a valid factor when comparing to 2018, however the R-Squared output states that only 17.93% of the variability can be explained by the model. Further analysis of the residuals shows normal distribution, homoscedasticity and variance are within acceptable range without any noticeable pattern to suggest transformation of the variables and lacking any significant overleveraged values drastically skewing the results. This tells us that we are missing key factors and using negative return years will not be the strongest indicator of future negative returns. In short, while there is a statistically significant impact on 2018 prices given 2015 prices, it only has a slight impact on actual prediction power for this model.

In order to compare the explanatory power of the fund returns from 2013 and 2014 to the explanatory power of the fund returns from 2016 and 2017, Michael built two regression models containing the dependent variable, independent variables, second-order terms, and interaction terms. The dependent variable in both of the models was the 2018 fund return. The independent variables in the first model were the 2013 fund return, the 2014 fund return, and the asset bonds variable. The independent variables in the second model were the 2016 fund return, the 2017 fund return, and the asset bonds variable. The second-order terms in the first model were for both of the fund returns in the model. The one second-order term in the second model was for the 2017 fund return variable. The interaction term for the first model was between both of the fund return variables in the model. The interaction term for the second model was also between both of the fund return variables in the model. After building both of these models, Michael found that both of the models passed the F-test. This indicates that the null hypothesis is rejected, which means that at least one of the betas in each model does not equal zero. After this was done, the t-tests for all of the explanatory variables were checked, and all of the explanatory variables passed their t-tests. This indicates that the beta for each variable does not equal zero. Next, the adjusted R-squared of both models were compared against each other. The adjusted R-squared was 0.463 for the first model, and the adjusted R-squared was 0.404 for the second model. This indicates that the first model explains 46.3% of the variability in the dependent variable, and the second model explains 40.4% of the variability in the dependent variable. Lastly, the results of the testing data from using both models was compared with the actual data. It was found that the testing data from using the first model has a correlation of 0.609 with the actual data. Also, it was found that the testing data from using the second model has a correlation of 0.534 with the actual data. After fully comparing both of these models, Michael came to the conclusion that the 2013 and 2014 fund returns were able to predict the 2018 ETF fund returns better than the 2016 and 2017 fund returns.

Stephanie built two multiple regression models to compare the predictive performance of a regression model trained only on the majority stock ETFs to the performance of a regression model trained only on the majority bond ETFs. She found that the bond ETF model greatly outperformed the stock ETF model when trying to predict 2018 fund return values based on the 2013-2017 fund return values. The bond ETF regression model was able to explain 88.3% of the variability in the dependent variable 2018 fund return while the stock etf regression model was only able to explain 53.3% of the variability in 2018 fund return. Furthermore, the bond ETF model performed much better than the stock ETF model based on the 10-fold cross validation results. The bond ETF model had an average root MSE of 1.94 calculated from the 10 folds. This value means that 95% of the time, the predictions of the model will be correct within 3.88 of the true 2018 fund return value. The stock ETF model had a much greater root MSE value of 9.6, meaning that 95% of the time, the prediction will be within 19.2 of the true 2018 fund return value. This shows that the stock ETF model predictions are much less reliable than the bond ETF model predictions. Overall, the bond ETF return values fell in a much smaller range than the stock ETF return values. This is likely due to the much more unstable nature of stocks compared to bonds. Stock ETF returns tend to have much greater variance than bond ETFs, leading to highly negative and highly positive return values. This means that there is potential for much greater returns in the stock ETFs compared to the bond ETFs, but there is also potential for much greater loss.

Based on the beta weights in the regression models and correlation values between different pairs of variables, we were able to draw some conclusions about the relationship between the 2018 fund return value and the fund return values for the remaining years. First, we found that the relationship between 2018 fund return values and the fund return values of 2013, 2014, 2015, 2016, 2017 appear to be different based on whether the ETFs are composed of majority stocks versus majority bonds. When evaluating the pairwise correlation between 2018 fund return and the remaining years using only the stock ETFs, we found that 2018 fund return had a positive correlation with 2013 fund return, 2014 fund return, and 2015 fund return and a negative correlation with 2016 fund return and 2017 fund return. In the regression model using only the bond ETFs, 2018 fund return had a positive correlation with 2013 fund return and 2015 fund return and a negative correlation with 2014 fund return, 2016 fund return and 2017 fund return. We also found that the correlation values for the bond ETFs were much higher than the correlation values for the stock ETFs. More specifically, 2018 fund return had a very strong negative correlation with 2018 fund return, indicating that they may be good predictors for 2018 fund return when looking only at the bond ETFs. This was further confirmed in the final bond ETF regression model where 2016 fund return and 2017 fund return had the largest beta values. We also found that 2015 fund return appeared to have a fairly strong positive correlation with 2018 fund return in both the bond ETF and stock ETF regression models.

In conclusion, there are many different features that can be used to predict ETF prices, but we decided to focus on building regression models using previous ETF fund returns and the type of each ETF as our features. The main goals of our experiments were to compare stock ETFs vs bond ETFs, compare fund returns from the two most recent years vs fund returns from the two earliest years, and determine if ETF prices could be predicted from both positive and negative return years. In order to accomplish each of these goals, we needed to build different regression models using methods such as forward selection, backward elimination, and residual analysis. Our key findings from all of our experiments include the earlier fund returns having more predictive power than the recent funds returns, the one downturn year having little impact on the predictive power of the model, and that it is much easier to predict the prices of bond ETFs compared to the prices of stock ETFs. All in all, while it was difficult trying to build regression models to predict ETF prices using this specific dataset, there were enough major findings from our experiments to make using this dataset worthwhile.

Section 6: Conclusion

Overall, we found various interesting outputs for multiple linear regression modeling of ETF markets. We can say with certainty that negative average returns are a poor indicator of future downturn prices and perform significantly worse when testing against semi-random multi-year models. In those models we emphasized the cyclic nature of the stock market, albeit with a small scope of two return years and found a higher correlation and stronger prediction pattern than the negative average return years.

Moreover, we found when we disregard prices entirely and focus on the makeup of the ETFs by comparing stock vs bond ETFs against our dependent variable, we were able to build more robust models with stronger statistical significance. Surprisingly, the best model statistically is a comparison of 2018 pricing to ETFs that only hold bonds and we attribute the strength of the model to the bond market's propensity to remain fairly stable in positive and negative return years.

To conclude, we believe that past price performance is hardly an indicator of current or future price performance and we would be better served taking note of the composition of an ETF when trying to predict prices with multiple linear regression. However, we think that making use of traditional stock analysis methods in conjunction with time-series modeling would give us a more vigorous and complete view of price movements.

Appendix

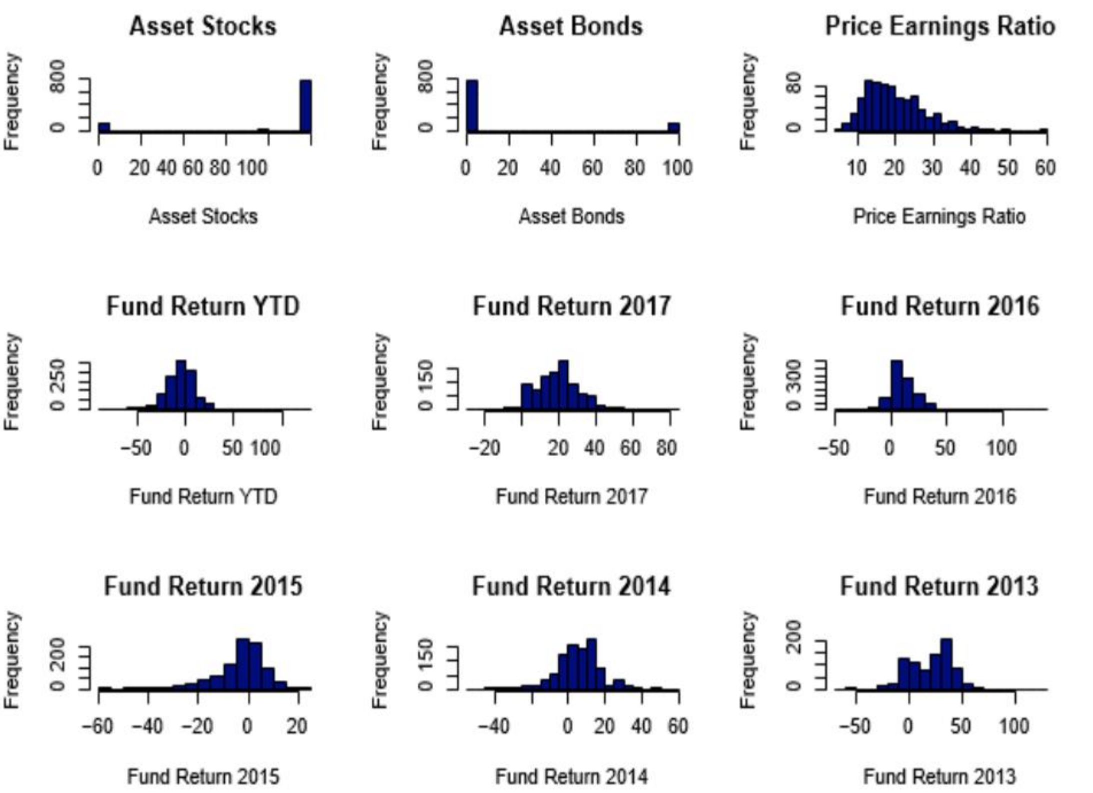


Fig. 1 Histograms of the numeric predictors.

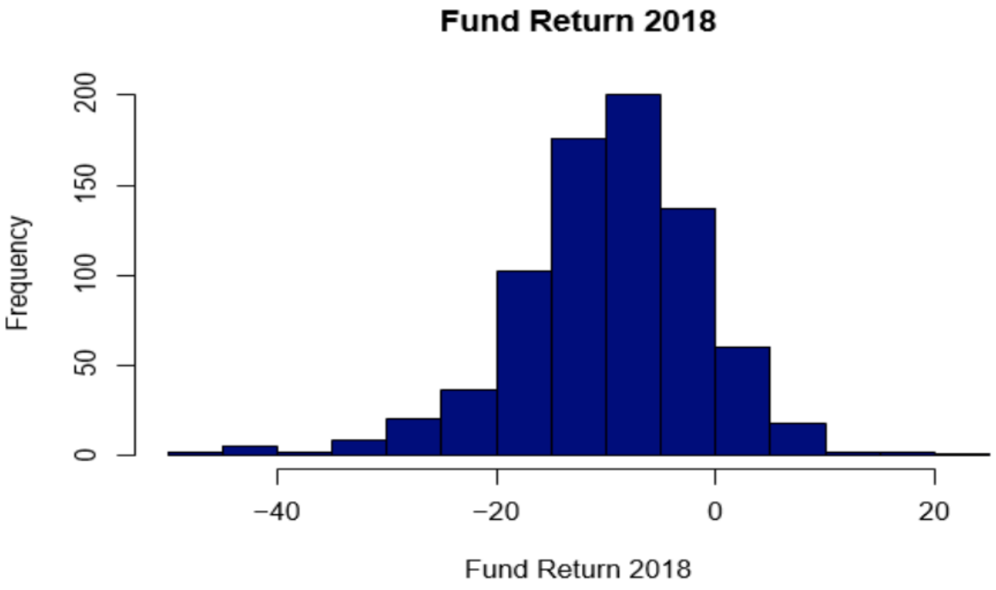


Fig 2. Histogram of the response variable.

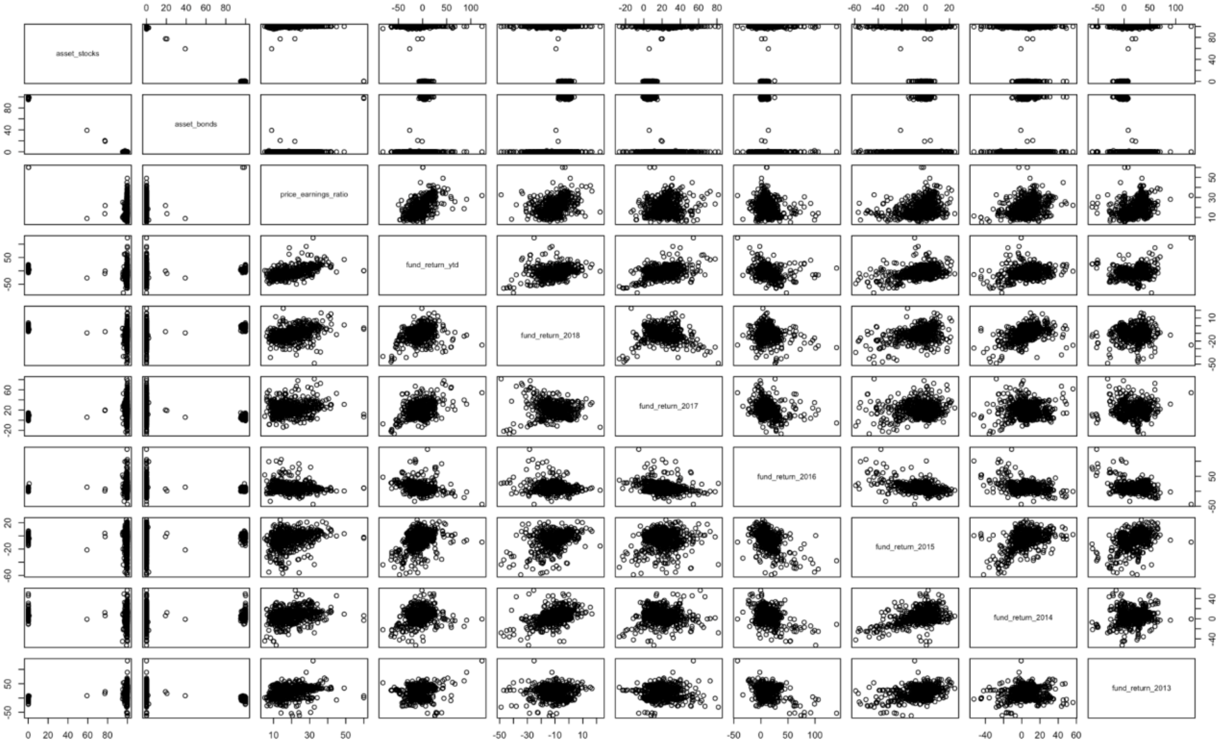
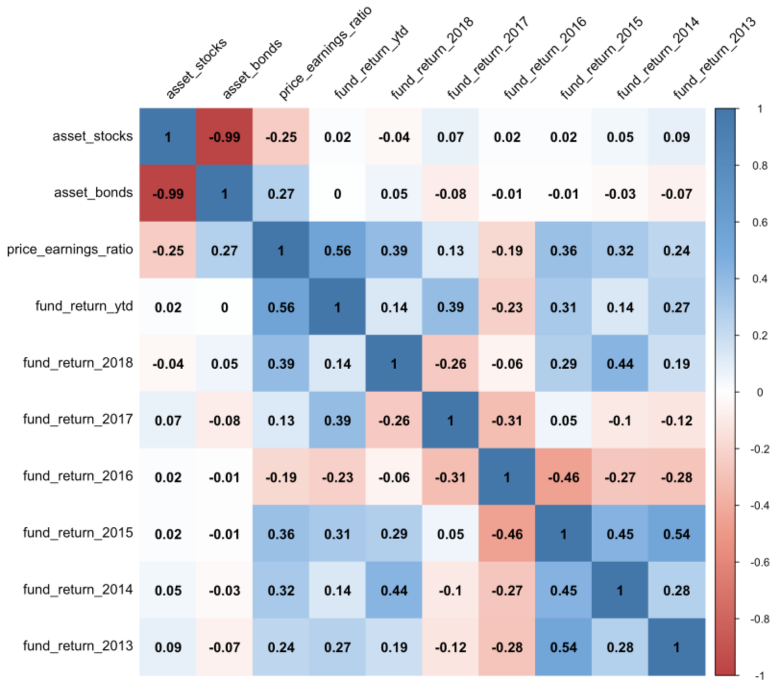


Fig 3. Pairwise scatter plots between all pairs of features in the dataset.



	asset_stocks	asset_bonds	price_earnings_ratio	fund_return_ytd	fund_return_2018	fund_return_2017	fund_return_2016	fund_return_2015	fund_return_2014	fund_return_2013
asset_stocks	1	-0.99	-0.25	0.02	-0.04	0.07	0.02	0.02	0.05	0.09
asset_bonds	-0.99	1	0.27	0	0.05	-0.08	-0.01	-0.01	-0.03	-0.07
price_earnings_ratio	-0.25	0.27	1	0.56	0.39	0.13	-0.19	0.36	0.32	0.24
fund_return_ytd	0.02	0	0.56	1	0.14	0.39	-0.23	0.31	0.14	0.27
fund_return_2018	-0.04	0.05	0.39	0.14	1	-0.26	-0.06	0.29	0.44	0.19
fund_return_2017	0.07	-0.08	0.13	0.39	-0.26	1	-0.31	0.05	-0.1	-0.12
fund_return_2016	0.02	-0.01	-0.19	-0.23	-0.06	-0.31	1	-0.46	-0.27	-0.28
fund_return_2015	0.02	-0.01	0.36	0.31	0.29	0.05	-0.46	1	0.45	0.54
fund_return_2014	0.05	-0.03	0.32	0.14	0.44	-0.1	-0.27	0.45	1	0.28
fund_return_2013	0.09	-0.07	0.24	0.27	0.19	-0.12	-0.28	0.54	0.28	1

Fig 4. Pairwise correlation values between all pairs of features in the dataset.

Appendix Data Dictionary:

Variable Name	Variable Type	Variable Description
fund_symbol	Categorical	Symbol used to identify each ETF.
fund_extended_name	Categorical	The full name for each ETF.
asset_stocks	Numerical	The percentage of assets in the ETF portfolio that are stocks.
asset_bonds	Numerical	The percentage of assets in the ETF portfolio that are bonds.
fund_return_2018	Numerical	The ETF total return percentage in 2018. This is also the dependent variable.
fund_return_2017	Numerical	The ETF total return percentage in 2017.
fund_return_2016	Numerical	The ETF total return percentage in 2016.
fund_return_2015	Numerical	The ETF total return percentage in 2015.
fund_return_2014	Numerical	The ETF total return percentage in 2014.
fund_return_2013	Numerical	The ETF total return percentage in 2013.

Table 1. Data Dictionary