# Bayesian Portfolio Construction and Error Minimization

David Debreceni

Steven's Institute of Technology

Advisor: William Rubens, PhD

December 14, 2019

# Abstract

The focus of this paper is to compare the process of security selection between frequentist linear regression and Bayesian linear regression. Utilizing random security selection from the Russell 1000 constituent list and a well-documented portfolio optimization process, Markowitz Mean-Variance optimization, portfolios with a 10-year performance record will be built. The portfolios built with each of the processes will be compared at a holistic level as well as the individual security selection. Measuring the residual error at each point to identify where any performance gains can be had. The goal is to demonstrate that the residual error in the Bayesian process is smaller thus generating a better-quality portfolio.

# Contents

# Introduction

The quest for generating portfolios that consistently have higher returns over their designated benchmark is the main reason for many of the modern portfolio analysis techniques used today. These techniques have their roots in some basic but solid fundamental principals of analysis. Starting in 1934 with Benjamin Graham and David Dodd writing Security Analysis highlighting the various ways an investor can analyze a company to provide Intrinsic Value. The idea of intrinsic value of a business was determined by its earning power. The phrase of earning power must imply a confident expectation of certain future results. (Graham & Dodd, 2009) It is this expectation of results that has led to other forms of analysis focused mostly on returns. Methods such as the Mean Variance optimization developed by Harry Markowitz. This process takes a universe of securities and through mathematical processing, finds the optimal portfolio. However, for this process to work in an efficient manner there needs to be some assumptions made about the returns of the securities, or rather the expected returns of the securities.

One of the first formulas designed to attempt to predict security returns systematically is the Capital Asset Pricing Model (CAPM). CAPM for an individual security will take the expected return for a security minus the risk-free market rate to determine the true return. When combined with the Mean-Variance optimization, this process became the most efficient way to select an optimized portfolio at that time. Over time this process has been added to due to additional studies showing deficiencies in the CAPM process only. The most common extensions of this is the Fama French model.

Calculating portfolios by hand can be both error prone and time consuming. As processing power of computers has improved the ability to better analyze data has as well. The main

improvements that have been seen have been the ability to process more historical data for more complete return analysis and portfolio optimization. In this analysis I will utilize Mean Variance Optimization in two different forms to minimize the residual error in the resulting portfolios.

## Literature Review

The process of portfolio optimization has been researched by many different sources. In this paper I am focusing on the Bayesian process that was highlighted in "Bayesian Portfolio Analysis" (Avramov and Zhou, 2009). This paper demonstrates that with quality priors, a Bayesian process can be utilized for better security selection. The paper focuses on how using factors as a random variable in a Bayesian process could possibly predict values more reliably. Combining this process with the popular Mean Variance optimization (Michaud, R. O., 2014) and the process of trying to identify a way of minimizing the error in a portfolio. As described in the above paper, I attempt to extend and prove the thought that a linear regression maximizes error rather than minimizes it.

Portfolio optimization can be done via many ways, factor analysis is a very common process of attempting to pick securities. As noted in "Robust Optimisation for Factor Portfolios" (Oxford, 2016), Factor analysis can be an excellent proxy for security return analysis. However, the different approaches have their own hurdles. "The quantitative approach suffers from non-stationary expected returns, Markovian prices path and efficient markets amidst others. Robust optimisation can then be used as a tool to deal with uncertainty in the expected asset return. In fact, in mean-variance optimisation, errors in expected returns should be the primary focus." This paper will attempt to focus on the expected errors mentioned.

## Problem of residual errors

Classical Mean Variance optimization assumes that the investor prefers a portfolio of securities that offers maximum expected return for some given level of risk. (Michaud, R. O., 2014) These returns can be estimated using various methods such as factor based or historical return analysis. Once the various expected returns are generated, they are often combined using an optimizer to generate what is expected to be the most efficient portfolio. These portfolios are often measured as successful based upon some form of utility equation, often something like the Sharpe ratio or Mean Variance equations.

Sharpe Ratio

$$S = \frac{R - R_f}{\sigma}$$

Where
$R_f$ = Risk Free return asset
$R$ = Return of Asset or portfolio
$\sigma$ = Standard of Deviation of the asset

Mean Variance

$$MV = E[R] - \lambda\sigma^2$$

Where
$E[R]$ = Expected Return of asset
$\lambda$ = Risk tuning parameter
$\sigma$ = Standard Deviation (Risk)

The issue with these processes is they are often "Error Maximizers" where the process itself selects securities based upon the best possible expected return, overweighting securities that outperform while underweighting those that underperform. During the process of estimating returns there is always a residual error generated (ε ). This error is assumed to have a normal distribution with mean of zero and standard of deviation of 1. Since this is expected to be true, the residual error is thus ignored during the optimization process and not accounted for in the utility equations above.

The primary issue is this residual error is unknown and cannot be known for the future as the

actual future return is unknown. Since this is unknown and as mentioned cannot be accounted for properly in the portfolio optimization process, the error is amplified. In many current optimization processes this is addressed by adding additional constraints to the optimizer. However, I believe adding these constraints ultimately adds a bias forcing security returns to be estimated in the direction we believe they should go. In this analysis there are only two constraints that I am adding to the Linear Regression and Bayes Regression, weights must sum to one for a long only portfolio and the beta of the portfolio must be equal to one.

# Data selection for analysis

## *Base universe selection*

To adequately analyze the residual error in the portfolio construction process it is important to pick securities that have a long history of returns and high liquidity. The base universe I selected was the Russell 1000 Index. This index is comprised of approx. 1000 companies containing almost 90% of the current market capitalization in the United States. These companies will typically have high liquidity and stable returns. For this analysis I am consuming 20 years of data, 10 will be used as a training set, the balance will be used for testing.

## *Data Gathering process*

The security returns for the Russell 1000 constituents are obtained using Factset codes for Total Return. Total Return is used to ensure that dividend data is accounted for in the individual security returns. First, I gather a unique list of constituents over the 20 year history of the Russell 1000 index. Once this list is obtained, I gather all available security return history for each of these constituents during the 20-year period being analyzed. In addition, I download the

Total Return for the index itself over the 20 year period to use as a benchmark for the portfolios created.

## *Fama French Factors*

The Fama French factor analysis is an attempt to break down the equity risk premium of the securities market. There have been many studies that show that there is both an in sample and out of sample predictability with the Fama French 3 Factor model. In addition, there have been multiple papers demonstrating the unreliability of analyzing historical security returns alone. For this reason, I have chosen to use the factor model with security return's regressed over these values for this analysis. The model I will use includes these four factors: CAPM, SMB, HML and Momentum.

- CAPM - Rm-Rf, the excess return on the market, is the value-weighted return on all NYSE, AMEX, and NASDAQ stocks (from CRSP) minus the one-month Treasury bill rate (from Ibbotson Associates).
- SMB (Small Minus Big) – Size Factor - is the average return on three small portfolios minus the average return on three big portfolios,

  SMB =1/3 (Small Value + Small Neutral + Small Growth) - 1/3 (Big Value + Big Neutral + Big Growth)
- HML (High Minus Low) – Value Factor - is the average return on two value portfolios minus the average return on two growth portfolios

  HML = 1/2 (Small Value + Big Value) - 1/2 (Small Growth + Big Growth)
- Mom (Momentum) - Mom is the average return on the two high prior return portfolios minus the average return on the two low prior return portfolios,

  Mom =1/2 (Small High + Big High) - 1/2(Small Low + Big Low).

# Methodology

## *Common processes*

To ensure continuity between the two different analyses there are certain parts of this analysis that are shared. The core data used from Fama French and security returns will be identical throughout. This data will be narrowed down to a universe of 100 randomly selected securities. This is due to the need for computing power to analyze results greater than this. These securities are selected from the available constituents each month of the analysis. I store the constituents listed and use the same 100 securities during both the Linear Regression and the Bayes Regression.

Both processes will use the same optimizer and utility equation to determine the best portfolio of securities. The optimizer will use a Least Squares approach by utilizing the libraries available in SciPy's Minimize function. The goal is to maximize the following Mean Variance equation:

$$\max \ \omega \cdot R \ - 2 * \lambda * (\omega \cdot Q \cdot \omega^T)$$

$$s.t. \ \sum \omega = 1$$

$$\sum \omega * \beta_{mkt} = 1$$

Where: $\omega$ = weights

$\lambda$ = tuning parameter set to 0.5

$R$ = Vector of Security Returns (defined for each process independently)

$Q$ = Covariance Matrix (defined for each process independently)

## *Linear Regression*

In order to generate an Expected Return, it is important to choose an analysis that has been proven over time to have a level of predictive power. A Linear Regression is just this type of analysis, determining the relationship between a dependent variable and predictor variables. In this analysis I will be regressing the individual security returns over the Fama French factors for the same time frame. The goal is to get a Beta for the 4 factors that is a reliable predictor for the next month's return.

The overall process will start with a base of 10 years of monthly return data for securities and Fama French. As highlighted above the process will for each month starting at 10 years and continuing until the last available date regress each of the 100 constituents individually over the available Fama Factor returns. This process will use all known available data at the time of processing, i.e. data 0 : t. Once each security has been regressed for that month, all the Beta's, Intercepts and predicted Error will be used to generate the correct return vector and covariance matrix to be used in the optimizer.

Covariance Matrix - $Q = \beta \cdot \Sigma \cdot \beta^T + diag(\varepsilon)$

Expected Return Vector - $R = \alpha + \beta \cdot F + R_f$

Where :

$\beta$ = Betas from regression

$\Sigma$ = Fama 4 factor covariance for the current time frame

F = Last known historical Fama Return

$R_f$ = Last known risk free return

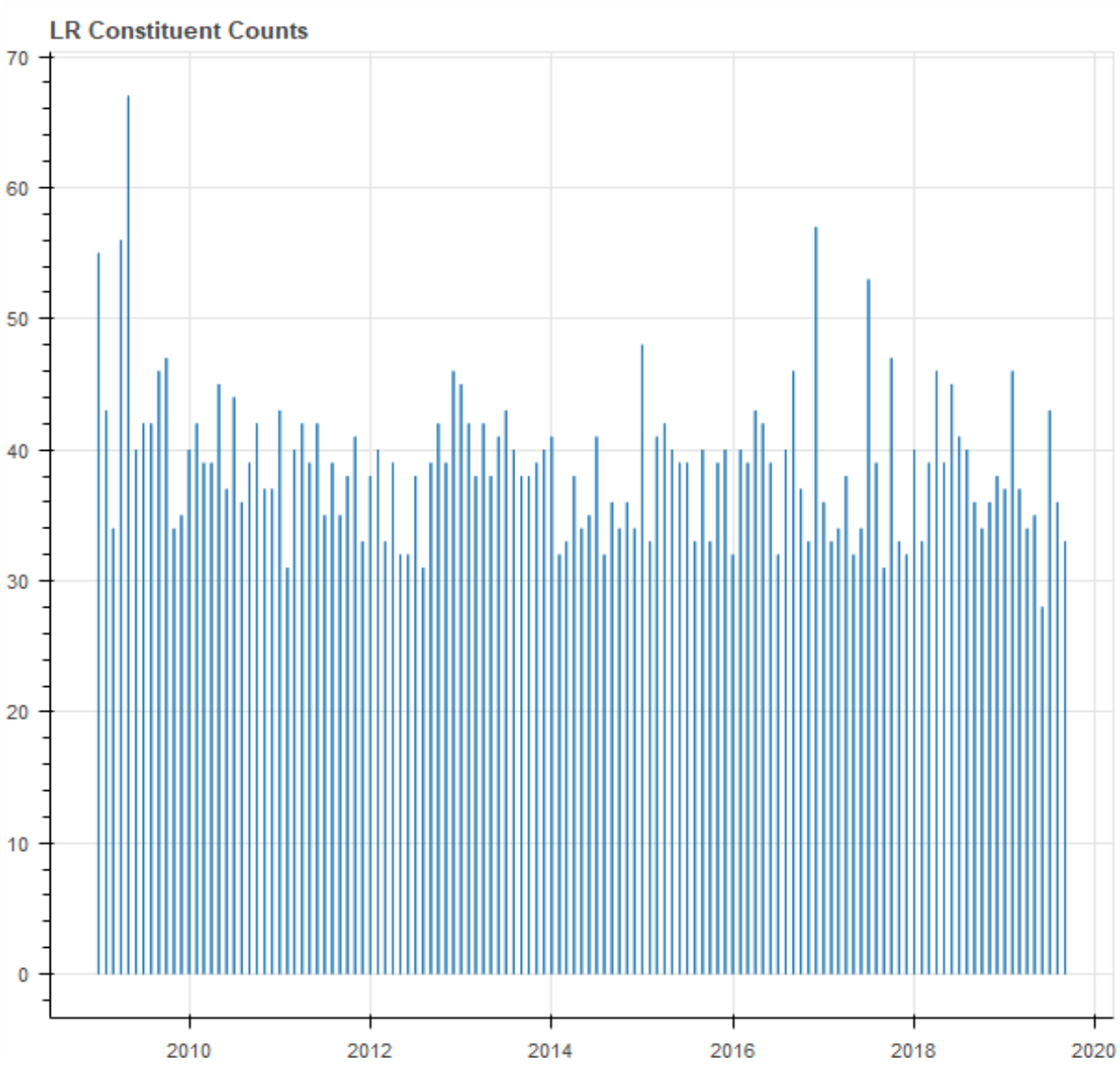
$\alpha$ = Intercepts from the regression

$\varepsilon$ = Variance from error estimates

All the above data is stored monthly for further analysis. Key items to focus on are the generated Expected Returns for each security based upon the equation above R and the final portfolio construction, weights for each security and portfolio return. Once this process has been run for all available months, I have approximately 10 years of monthly future predictions to check for optimal portfolio construction as well as the residual error.

First step in this process is to ensure that number of constituents that have been selected during this process was adequate. If too few securities were selected, then the portfolio would not be diversified enough to make the analysis of the final return meaningful. The Sharpe ratio (highlighted above) of the portfolio is a keyway to identify if this is truly a diversified portfolio. Below demonstrates that this process has indeed built a portfolio containing enough constituents of the 100 available to be diversified.

129 Months of Data




Constituents counts

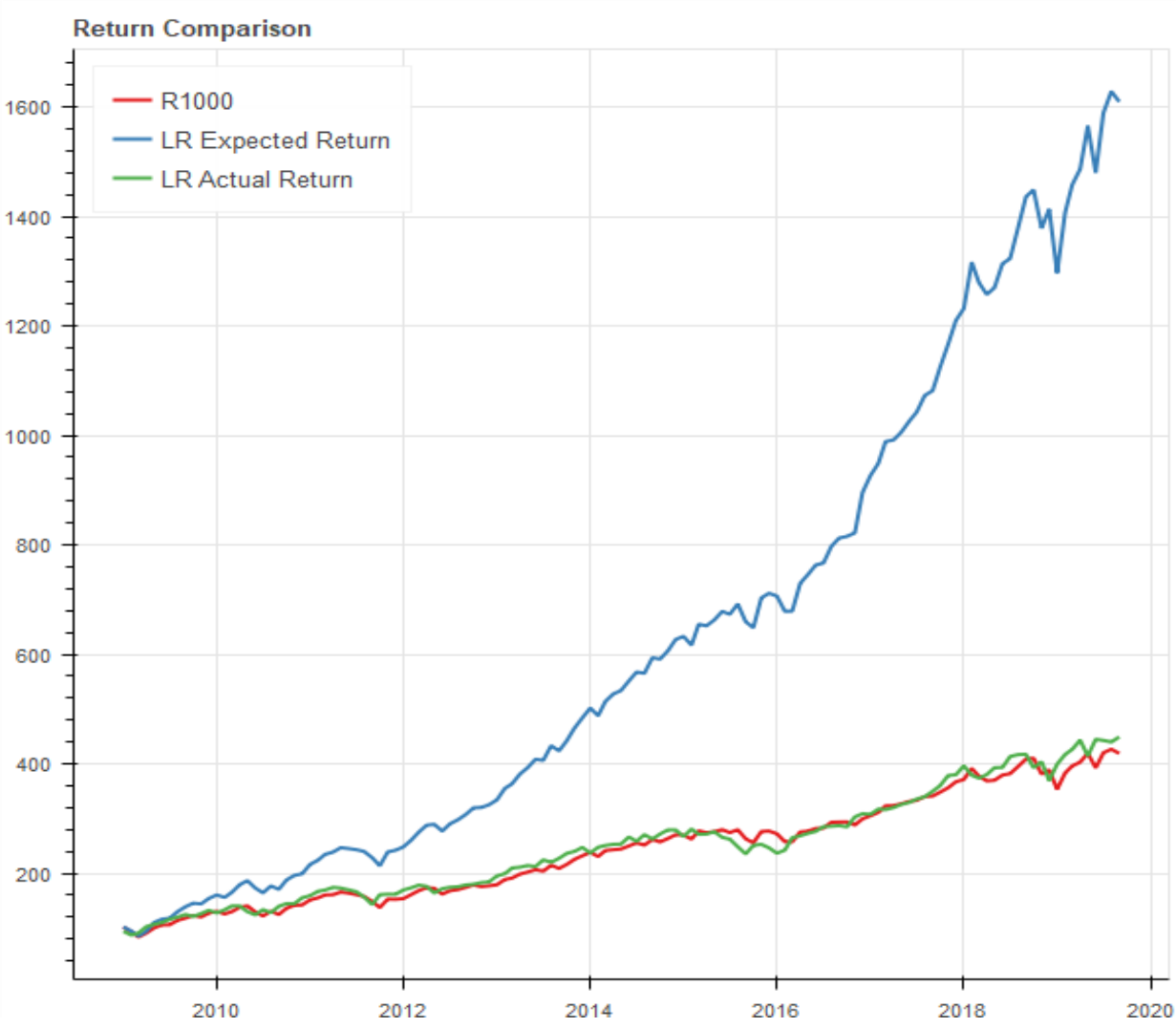
- Mean – 39 Constituents
- Min – 28 Constituents
- Max – 67 Constituents

Constituent weights

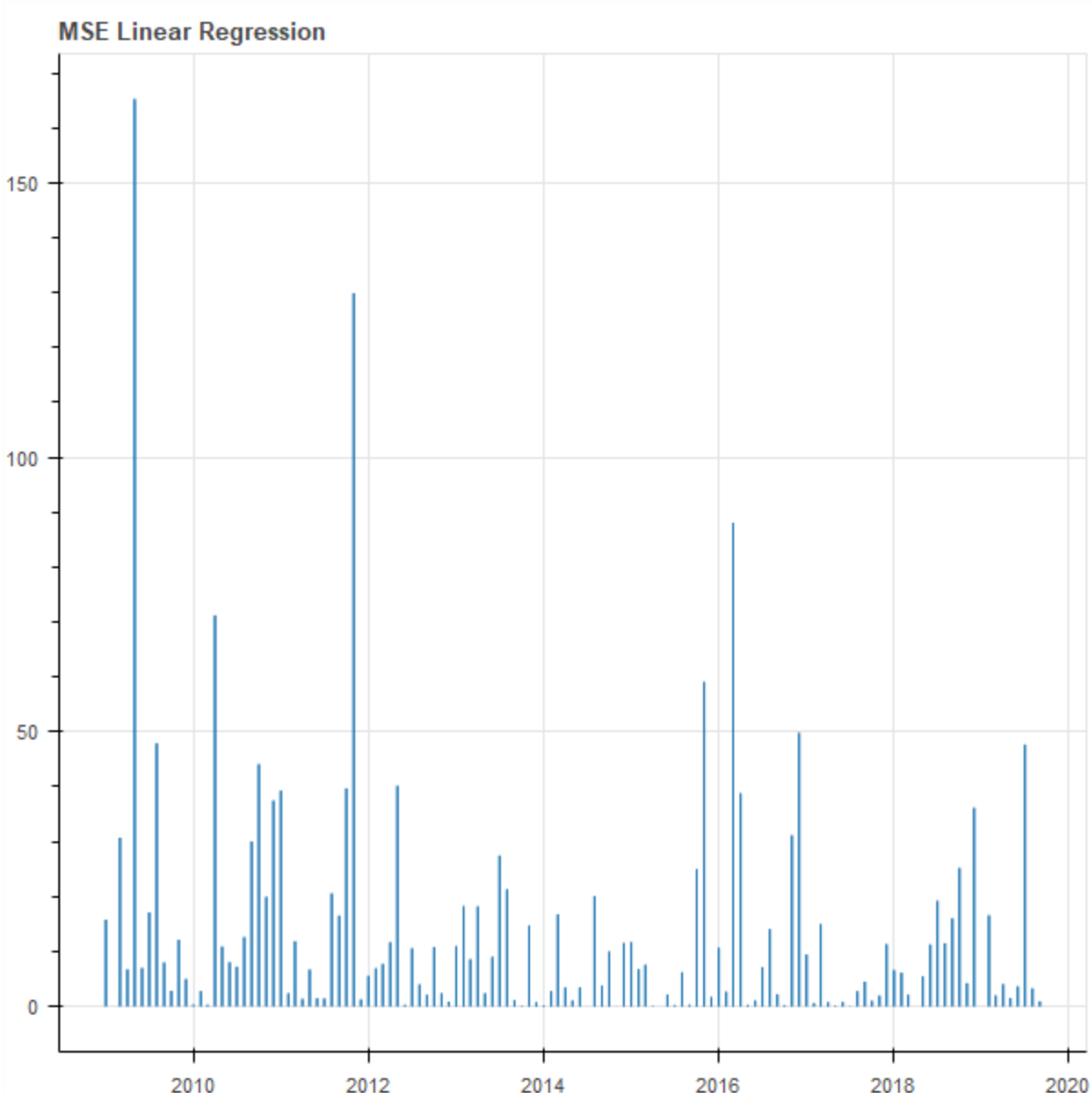
- Mean – 2.6%
- Min – 1%
- Max – 14%

Compounding return data is important to realize whether the predicted portfolio will be close in relation to the actual portfolio. The predicted portfolio is the one using the returns from the predicted monthly returns and the weights generated by the optimizer. The actual portfolio will use the same weights generated from the optimizer but use actual security returns for that month. In theory these two values should be close if the process worked correctly.

As you can see from the graph below, this is hardly the case. According to the chart below over the 10-year period I would have expected to see a portfolio returning ~1,604% vs what the actual portfolio would have returned from the suggested securities of ~449%.




Looking at the monthly Mean Squared Error for the Linear Regression process you can see that while there are certainly some months that have a very large MSE, there are quite a few that are very close to being in line. The average Mean Squared Error for the entire period is 35.8%




When analyzing the expected values vs, the actual values at a portfolio level in the Linear Regression optimization process that Expected values generate portfolios that are not in line with expected.

Expected Values
- Annualized Return – 29.5%
- Annualized Volatility – 1.28%
- Sharpe Ratio – 1.76

Actual Values
- Annualized Return – 15%
- Annualized Volatility – 1.24%
- Sharpe Ratio – 1.01

While there is considerable error in this process the goal is to see if this is better or worse than a Bayesian Regression.

## *Bayesian Regression*

A Bayesian model is one that uses probability to represent the uncertainty in the data. Rather than attempting to fit this process to a specific line, this process will use a Monte Carlo process to attempt to represent as many possible scenarios as possible. The Bayesian process operates differently as it treats the Expected Return as a random quantity with an unknown probability distribution. This distribution is continually updated based upon prior data; in this case I will start with the same 10 years of core data.

$$P(Return \mid factors) = \frac{P(factors \mid Return) \ * P(Return)}{P(factors)}$$

Utilizing the library PYMC3 I will process data in the same manner as was done for the Linear Regression. Each month's security return data will be regressed over the Fama French factors. To get the best results the PYMC3 library will take 2000 samples over 2 chains. This way there is a check to ensure there is a level of continuity in the results. To obtain the Beta and Intercepts from this process mean values from the traces generated will be used.

Unlike in the Linear Regression the Expected Return is calculated based upon a normal probability equation - $N(\alpha + \beta \cdot F, 1)$. This equation will be used when calculating the posterior sample. The PYMC3 library will have its input value changed to be the last known Fama French factor (same as the Linear Regression) then based upon the already established probability distributions from the model built with the priors will run 500 samples in a Monte Carlo scenario to predict the individual security returns from the factors. These predicted returns are then averaged to generate the Expected Return for that security. Once all securities have

been run, the expected returns and covariance matrix are passed to the optimizer, using the same Utility equation as before in the Linear Regression analysis.

Covariance - $Q = \beta \cdot \Sigma \cdot \beta^T + diag(\varepsilon)$

Expected Return - $R$

Where:

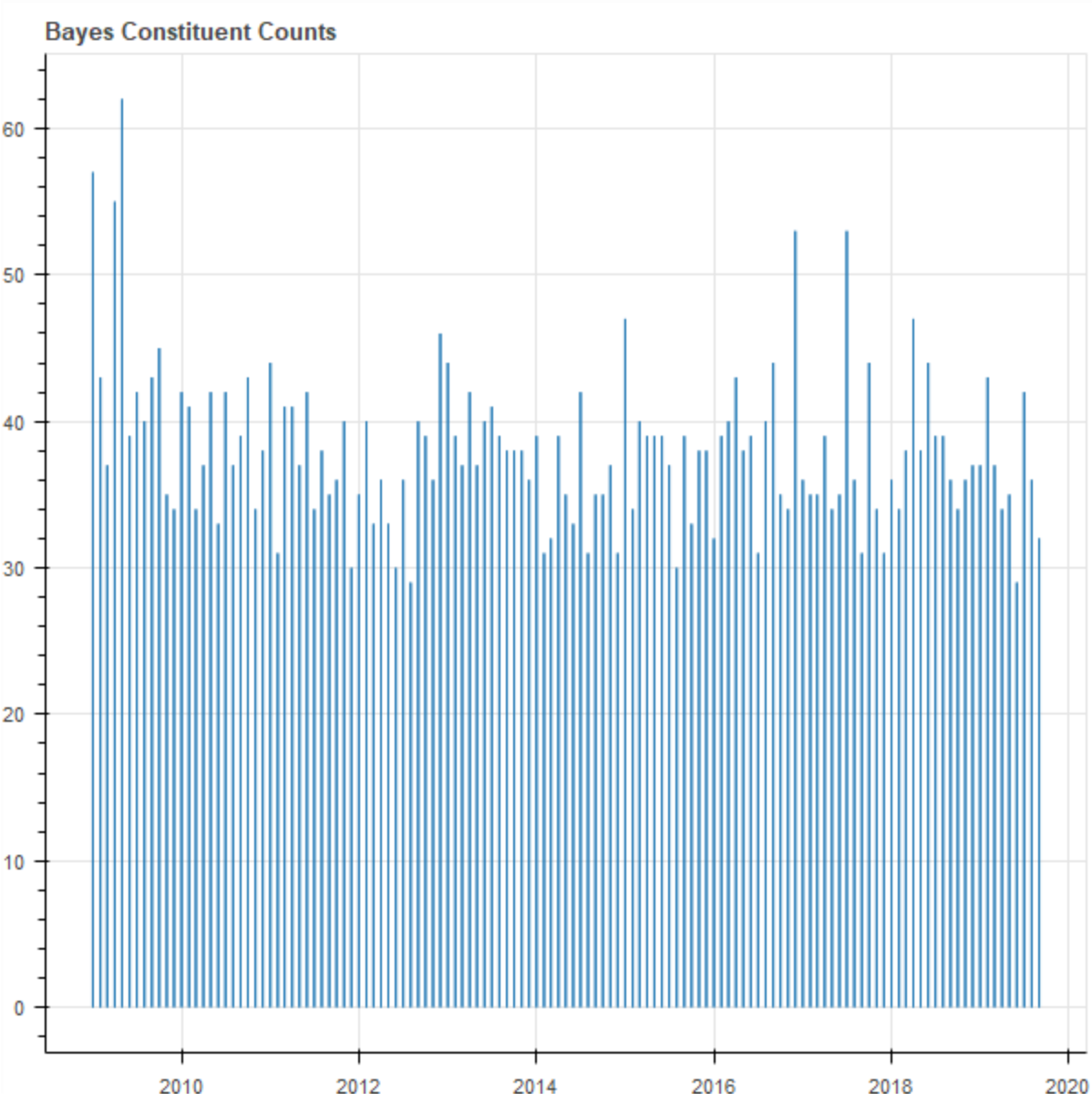
$\beta$ = Mean of Beta's from the 2000 sample Bayes process

$\Sigma$ = Fama 4 factor covariance for the current time frame

$R$= Mean of posterior predictions from Bayesian process

$\varepsilon$ = Mean of sigma values generated during the Bayesian process, squared

As above the first item to check is the portfolio constituent count. As before a solid sampling is expected to ensure that we are diversified.




Constituents counts

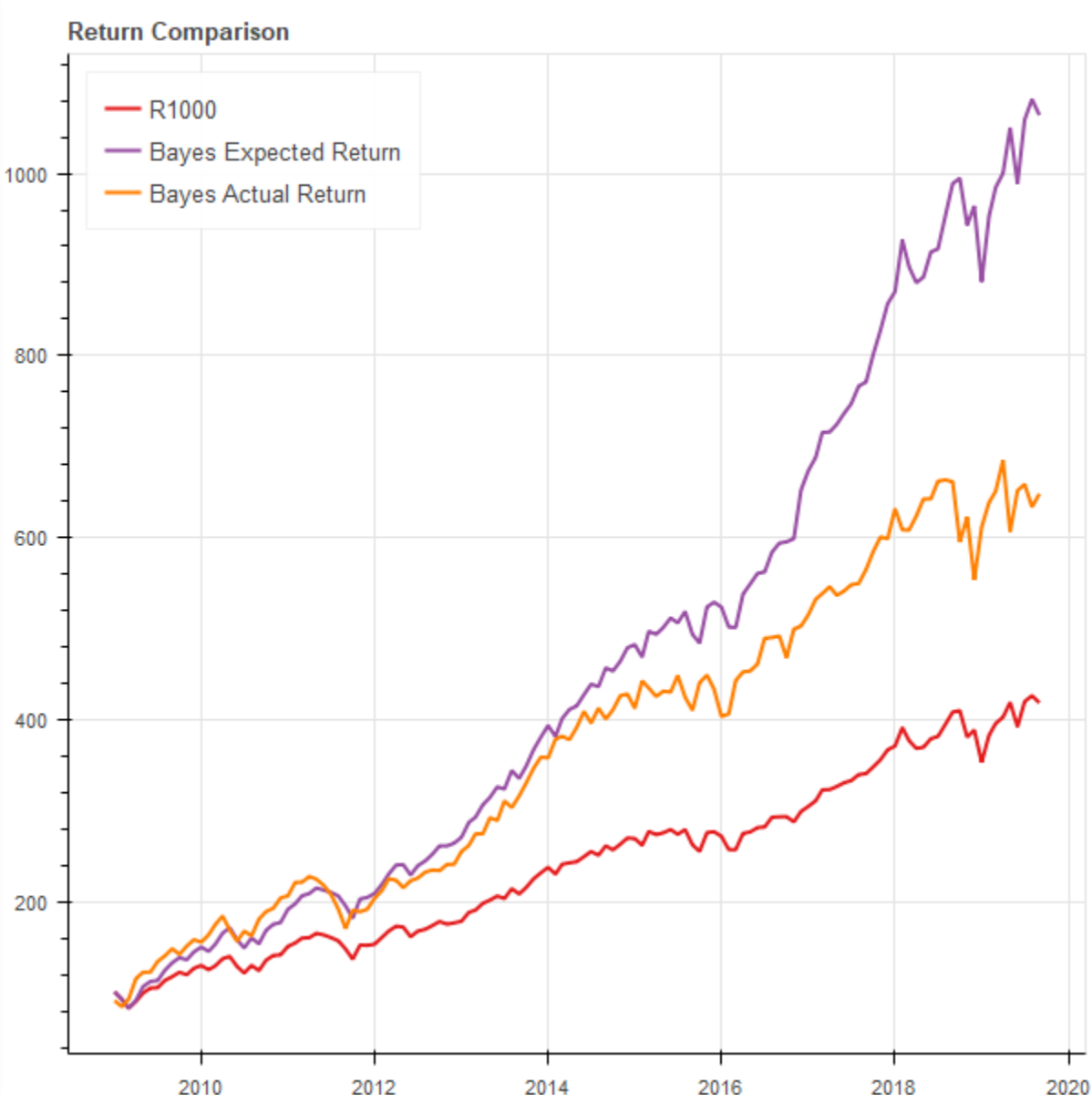
- Mean – 38 Constituents
- Min – 29 Constituents
- Max – 62 Constituents

Constituent weights

- Mean – 2.6%
- Min – 1%
- Max – 15%

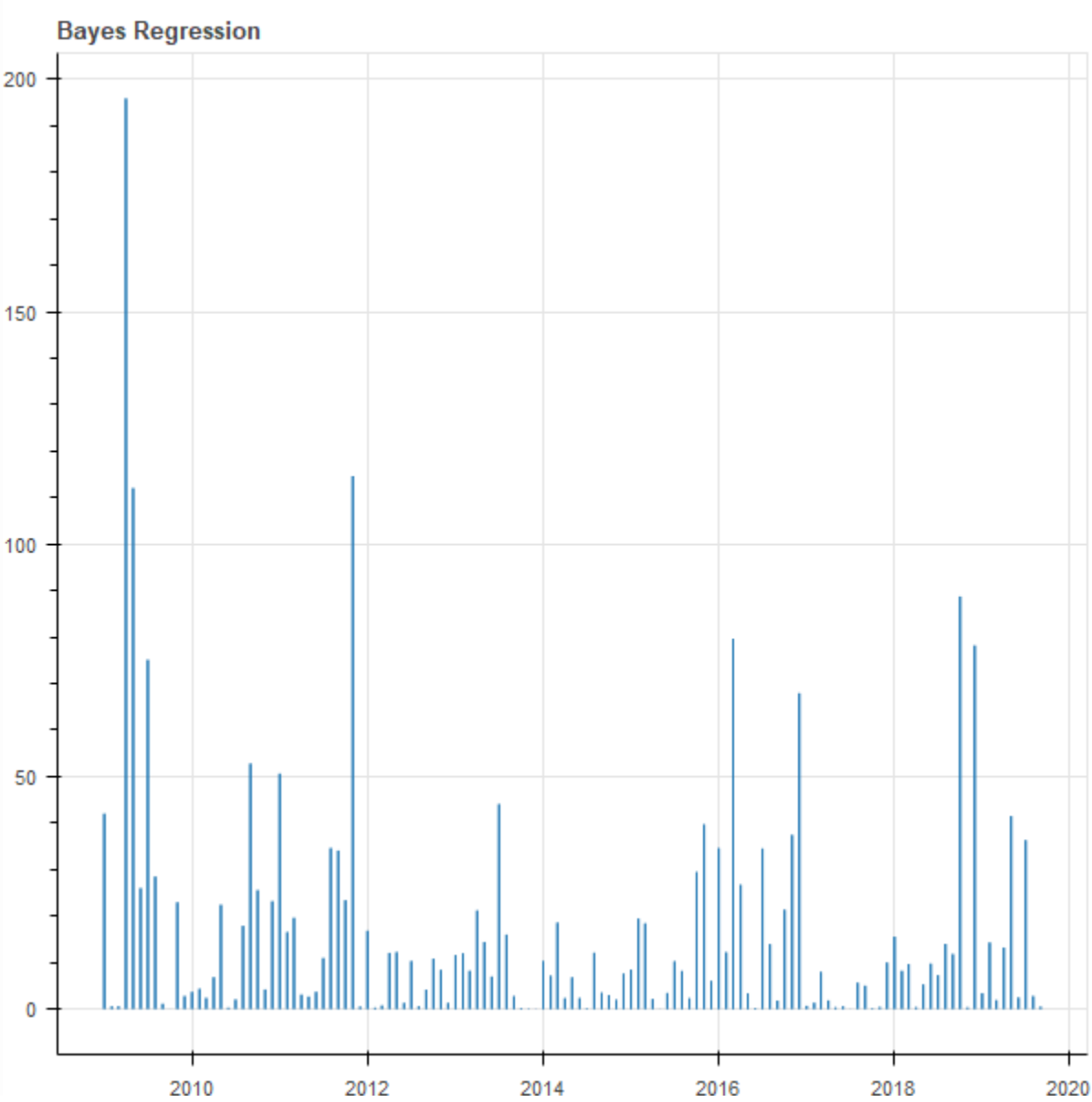
When comparing these results with the Linear Regression, they are very similar. The average portfolio size and weighting is almost identical. Demonstrating that the universes are similar in construction is important to show continuity in the analysis.

When analyzing the compounding returns between the Expected Return portfolio and the Bayes Actual return portfolio you can see that the Bayesian process of selection was much closer to reality vs the Linear Regression one.  That said the expected values are still not in line with what you would like to see.  Acording to this process the comounding return for Expected return ~1,064% vs the actual return ~648%




However, one thing to note is that the Bayesian portfolio did ultimately out perform the Linear Regression model by almost 200% over the 10 year period.

When comparing the MSE of the monthly Bayesian process you can see that like the Linear Regression it starts out with larger errors but slowly begins to level off. The MSE for the Bayesian process 42.5% over the 10 years.




Analyzing some results from this process shows that the difference between the expected values and actual values in the Bayesian analysis are much closer than those in the Linear Regression.

Bayes Values

Expected Values
- Annualized Return – 24.6%
- Annualized Volatility – 1.29%
- Sharpe Ratio – 1.5

Actual Values
- Annualized Return – 18.99%
- Annualized Volatility – 1.55%
- Sharpe Ratio – 1.02

# Comparing Results

After analyzing the two different processes for selecting portfolios, the results are a contradiction. The Bayesian process selected a portfolio that clearly outperforms the one built from Linear Regression. However, when using a standard error measurement to check for our goal of error reduction, it turns out that the Bayesian portfolio has a worse error – LR 35.8% vs Bayes 42.5%. Based upon the fact that the Bayesian portfolio has beat the Linear Regression by almost 4% annually with only 0.3% more volatility in the process means that there is something clearly better about the Bayes process but that is not being demonstrated at the portfolio level.

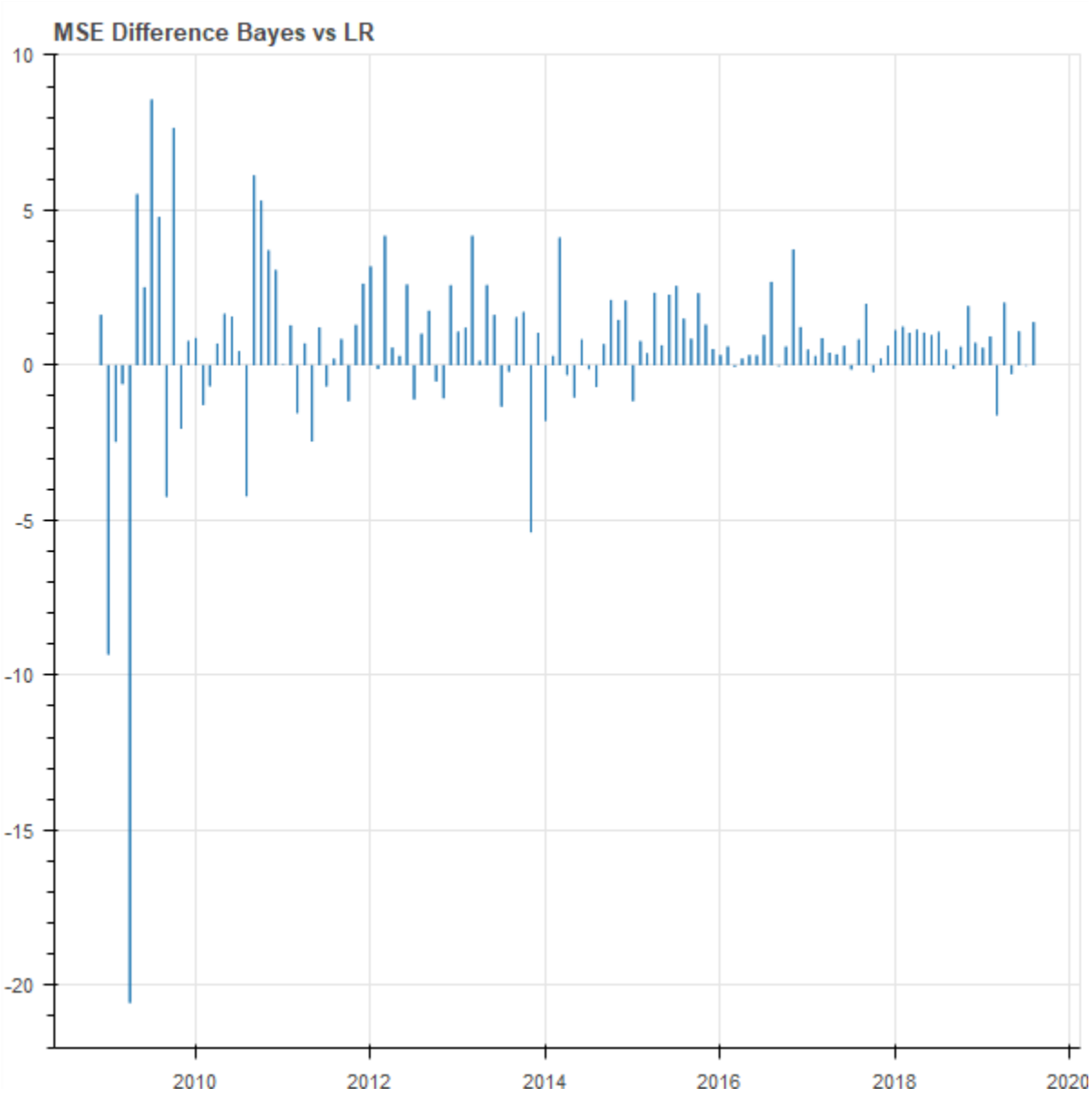
## *Individual Security results*

To calculate the Mean Squared Error of each individual security I take the actual security return for that month minus the Expected Return that is generated during the individual process of either Linear Regression or Bayesian Regression.

$$MSE_t = (ActRtn_t - EstRtn_t)^2$$

By taking the average of these values for each month, I get the individual security MSE for all the securities processed. When analyzed over time I see a result that makes more sense as to why Bayes had better return. The MSE over time for Bayes is 60.6% vs Linear Regression 61.27%. While this number is now better than the Linear Regression, less than 0.6% is well within an error estimate thus not enough proof on its own. What I discovered is as I continue to

average values, the two processes tend to converge. However, when I look at an individual value there are better results. The below graph shows the monthly value for MSE for LR – Bayes. A positive value indicates that Bayes is better. Ultimately Bayes outperformed MSE 73% of the time.



By doing a better job in predicting individual security returns, the optimizer does a better job selecting securities. When analyzing individual securities, the Bayes process picked a better return 53% of the time.

## Conclusion

After reviewing all the results, there is better performance from the Bayesian portfolio vs the Linear Regression portfolio. However, after reviewing the data and looking at various error metrics it is not clear what the benefit is coming from. It appears that at the individual security level the Bayesian process does a better job predicting security return values that are closer to reality. This ultimately leads to better optimization choices. I believe therefore the Bayesian process has performed better than the Linear Regression. But more analysis is needed including running this analysis with a larger dataset to see if there is still better performance.

## Further Analysis

Running this process for larger dataset would possibly generate better results.
I have already run this data comparing the time series differently. Since we are attempting to predict values based upon the last known factor value, running the analysis comparing security returns of [1 : t] vs factors [0 : t-1], then predicting returns using factors[t] should yield better results.

## Code

The code for this analysis is stored in https://github.com/debreceni/BayesianPortfolioAnalysis

# Bibliography

Fama, E. F., & French, K. R. (1993, 02). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics, 33*(1), 3-56. doi:10.1016/0304-405x(93)90023-5

Graham, B., & Dodd, D. L. (2009). *Security analysis: Principles and technique*. London: McGraw-Hill.

Michaud, R. O. (2014). The Markowitz Optimization Enigma: Is 'Optimized' Optimal? *SSRN Electronic Journal*. doi:10.2139/ssrn.2387669

Robust Optimisation for Factor Portfolios (2016), from https://www.maths.ox.ac.uk/system/files/attachments/593233_0.pdf

Panopoulou, E., & Plastira, S. (2011). Fama French Factors and US Stock Return Predictability. *SSRN Electronic Journal*. doi:10.2139/ssrn.1804927

(n.d.). Retrieved December 15, 2019, from https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html