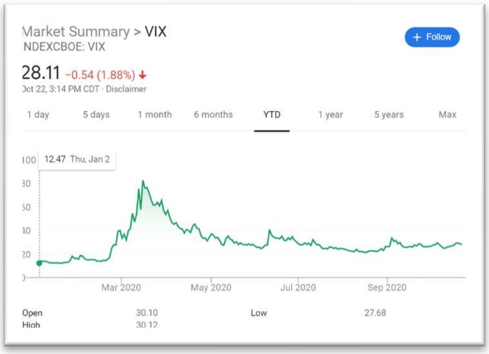
Abstract:

This project seeks to gain insight as to what drives the volatility of a well-known stock options index, VIX, which serves as a proxy for implied volatility within US financial markets. The question this analysis seeks to answer is, do a small subset of portfolios drive the majority of the daily returns of VIX. IE is market volatility motivated by only a few firms having exceptionally bad days? This remains unclear. The analysis used included Ordinary Least Squares, Ridge, and LASSO Regressions. Further, with the exception of OLS (because of its deterministic nature), the final regressions were selected through cross validation. The analysis showed that while the resulting models were fairly accurate in determining the general direction of deviation from the mean, they failed to model the magnitude of the direction very well. More simply put, the models predicted the right direction, but not the right size.

Background:

A constant problem that has plagued academic finance-and more broadly finance as a whole-is how to better understand how the market values volatility throughout time. Volatility is referring to the changes in prices for stocks on the market. Typically, these price movements are relatively small-think the stock tickers you see on news channels. But sometimes these movements can be very large and seemingly discontinuous. Options traders value this volatility inherent in the market to make educated trades. A popular proxy for volatility is a measure called VIX or the Volatility Index. Created by the Chicago Board Options Exchange (CBOE), it serves as a metric to measure the expectations of the market regarding future volatility. It is often called the *Fear Index* because, when it has high values, there is often panic within the market. Take for instance our recent circumstances. The Index peaked in mid-March at the peak of uncertainty relating to the pandemic. Additionally, it has stayed at an elevated value since then. The index is generated by looking at the forward volatility within the options traded on the stocks that make up the S&P500


Market Summary > VIX
NDEXCBOE: VIX

28.11 -0.54 (1.88%)
Dct 22, 3:14 PM CDT · Disclaimer

12.47 Thu, Jan 2

Open	30.10	Low	27.68
High	30.12		

index. This project seeks to better understand how options traders (who are the ones who dictate the value of the future volatility when they trade) come up with their volatility measures (which they use to trade). Mainly, I want to see if a few stocks having an uncommonly bad/good day have a disproportionate effect on market expectations of volatility. Why then, is it useful to better understand the drivers of volatility? Plainly put, if I were to create a model with significant predictive power, I could develop a trading strategy around it and make money. Indeed, there are many firms within Wallstreet that develop these types of models for various industries, stocks, derivatives, and almost any financial instrument possible. One of the most prosperous of which would be Renaissance Technologies, the founder of which was originally a mathematician who worked with the US government developing cryptography methods. As for how lucrative this is when done well, that mathematician is named Jim Simons. He has a net worth of $23.5 billion.

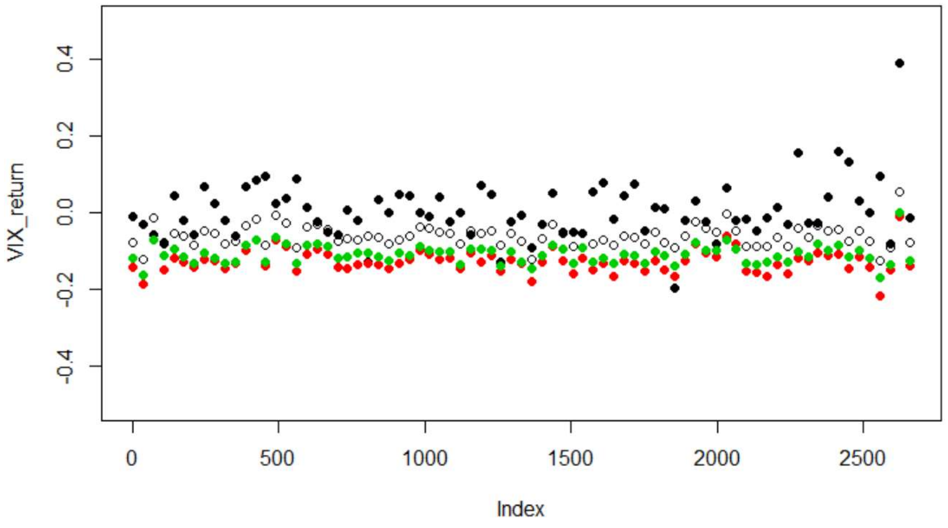
Data and Methodology:

This analysis will be using WRDS or Wharton Research Database System which is the academic finance data standard from which most academic finance papers published derive their data. Not to be overly specific, this project will be using their data on stock returns and index composition to generate novel data. Specifically, the novel data will be created using the daily pricing data on every stock publicly traded in the United States of America between the period of January 2010 and December 2019. This dataset which was over 2.7 gigabytes of memory which included over 42,000 distinct companies was very difficult to coerce and required dozens of hours of compute time alone to rework into an error checked usable format. Then, this data was used to create several portfolios that required daily returns between two thresholds i.e., between a 0% daily return and a 2.5% daily return. These portfolios were renewed daily meaning the same firm was rarely in the same portfolio on back-to-back days. In total there were 10 portfolios that were created. From each portfolio five variables were generated: total market capitalization or “tmc” (this is what people are referring to when they say Apple is worth $2 trillion), average market capitalization or “amc”, weighted average return or “war” (if a portfolio has company A with a market cap of $2 with a return of 3% and company B has a market cap of $1 with a return of 6% the weighted average return is found through the process: $\frac{2}{2+1} * 3\% + \frac{1}{1+2} * 6\% = 4\%$), average return “ar”, and the number of companies in the portfolio “nc”. These altogether form our independent variables. There is a big caveat here, the way the data was constructed likely violates some of the assumptions of linear regressions mainly independence and collinearity. This is due to several factors influencing the returns of companies. For more information on this look up the Fama-French three factor model. For the sake of simplicity, I myopically assume the data does not violate the assumptions required for linear regression. Moving on, we have our independent variables, but are missing our dependent variable. Our dependent variable is constructed using VIX pricing information. The value is equal to the natural logarithm of the daily return, as were all prior returns. This data was then used to run an Ordinary Least Squares (OLS) linear regression, 10,000 cross-validated Ridge Regression, and 10,000 cross-validated LASSO regressions. Because linear regressions are deterministic and can be found in closed form no cross-validation was needed. Further, while ridge regressions can be found just as easily in closed form, it was decided that cross validation coupled with iterative descent would be a better demonstration for this project. As for the LASSO regression, because there does not exist a closed form solution cross validation and gradient descent were required. I performed traditional cross-validation to determine which weights produced the best outcome in the training sample, and then tested 10,000 different weighting schemes in the holdout part of the dataset. The scheme with the lowest average mean squared error from the non-training data was chosen to represent the general model. And then ASME was calculated for each model using the new performance data hidden from the prior processes.

Results:

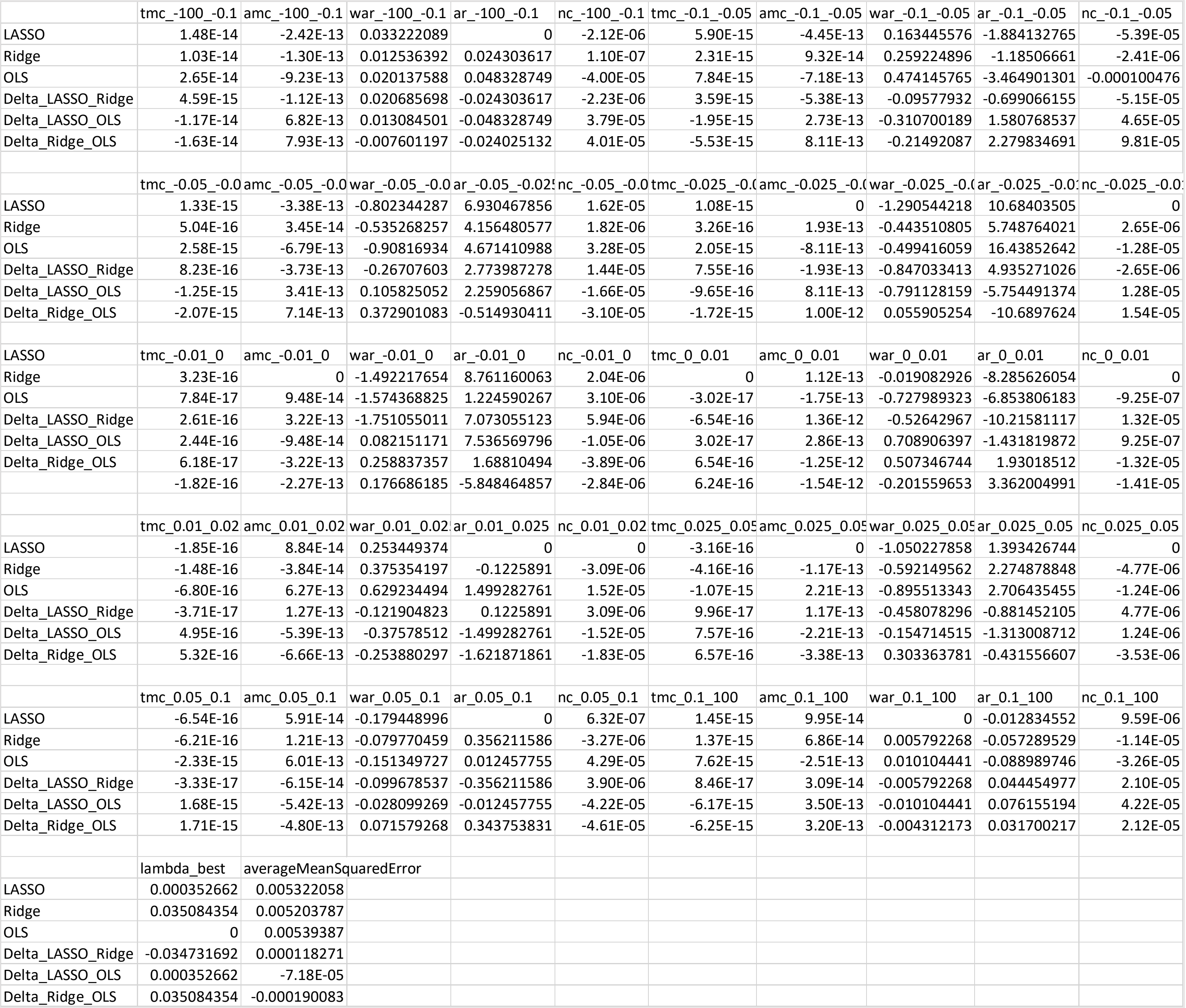
Here are how the models predicted VIX returns along with actual VIX returns (1/35 points shown):

Black = Actual; **Red** = OLS; **Green** = Ridge; **White** = LASSO



ECE 532 **Final Project** **Trent McKinnon**

Here are the results for the three models weights along with the differences between the weights:



	tmc_-100_-0.1	amc_-100_-0.1	war_-100_-0.1	ar_-100_-0.1	nc_-100_-0.1	tmc_-0.1_-0.05	amc_-0.1_-0.05	war_-0.1_-0.05	ar_-0.1_-0.05	nc_-0.1_-0.05
LASSO	1.48E-14	-2.42E-13	0.033222089	0	-2.12E-06	5.90E-15	-4.45E-13	0.163445576	-1.884132765	-5.39E-05
Ridge	1.03E-14	-1.30E-13	0.012536392	0.024303617	1.10E-07	2.31E-15	9.32E-14	0.259224896	-1.18506661	-2.41E-06
OLS	2.65E-14	-9.23E-13	0.020137588	0.048328749	-4.00E-05	7.84E-15	-7.18E-13	0.474145765	-3.464901301	-0.000100476
Delta_LASSO_Ridge	4.59E-15	-1.12E-13	0.020685698	-0.024303617	-2.23E-06	3.59E-15	-5.38E-13	-0.09577932	-0.699066155	-5.15E-05
Delta_LASSO_OLS	-1.17E-14	6.82E-13	0.013084501	-0.048328749	3.79E-05	-1.95E-15	2.73E-13	-0.310700189	1.580768537	4.65E-05
Delta_Ridge_OLS	-1.63E-14	7.93E-13	-0.007601197	-0.024025132	4.01E-05	-5.53E-15	8.11E-13	-0.21492087	2.279834691	9.81E-05
	tmc_-0.05_-0.0	amc_-0.05_-0.0	war_-0.05_-0.0	ar_-0.05_-0.02	nc_-0.05_-0.0	tmc_-0.025_-0.	amc_-0.025_-0.	war_-0.025_-0.	ar_-0.025_-0.0	nc_-0.025_-0.0
LASSO	1.33E-15	-3.38E-13	-0.802344287	6.930467856	1.62E-05	1.08E-15	0	-1.290544218	10.68403505	0
Ridge	5.04E-16	3.45E-14	-0.535268257	4.156480577	1.82E-06	3.26E-16	1.93E-13	-0.443510805	5.748764021	2.65E-06
OLS	2.58E-15	-6.79E-13	-0.90816934	4.671410988	3.28E-05	2.05E-15	-8.11E-13	-0.499416059	16.43852642	-1.28E-05
Delta_LASSO_Ridge	8.23E-16	-3.73E-13	-0.26707603	2.773987278	1.44E-05	7.55E-16	-1.93E-13	-0.847033413	4.935271026	-2.65E-06
Delta_LASSO_OLS	-1.25E-15	3.41E-13	0.105825052	2.259056867	-1.66E-05	-9.65E-16	8.11E-13	-0.791128159	-5.754491374	1.28E-05
Delta_Ridge_OLS	-2.07E-15	7.14E-13	0.372901083	-0.514930411	-3.10E-05	-1.72E-15	1.00E-12	0.055905254	-10.6897624	1.54E-05
LASSO	tmc_-0.01_0	amc_-0.01_0	war_-0.01_0	ar_-0.01_0	nc_-0.01_0	tmc_0_0.01	amc_0_0.01	war_0_0.01	ar_0_0.01	nc_0_0.01
Ridge	3.23E-16	0	-1.492217654	8.761160063	2.04E-06	0	1.12E-13	-0.019082926	-8.285626054	0
OLS	7.84E-17	9.48E-14	-1.574368825	1.224590267	3.10E-06	-3.02E-17	-1.75E-13	-0.727989323	-6.853806183	-9.25E-07
Delta_LASSO_Ridge	2.61E-16	3.22E-13	-1.751055011	7.073055123	5.94E-06	-6.54E-16	1.36E-12	-0.52642967	-10.21581117	1.32E-05
Delta_LASSO_OLS	2.44E-16	-9.48E-14	0.082151171	7.536569796	-1.05E-06	3.02E-17	2.86E-13	0.708906397	-1.431819872	9.25E-07
Delta_Ridge_OLS	6.18E-17	-3.22E-13	0.258837357	1.68810494	-3.89E-06	6.54E-16	-1.25E-12	0.507346744	1.93018512	-1.32E-05
	-1.82E-16	-2.27E-13	0.176686185	-5.848464857	-2.84E-06	6.24E-16	-1.54E-12	-0.201559653	3.362004991	-1.41E-05
	tmc_0.01_0.02	amc_0.01_0.02	war_0.01_0.02	ar_0.01_0.025	nc_0.01_0.02	tmc_0.025_0.05	amc_0.025_0.05	war_0.025_0.05	ar_0.025_0.05	nc_0.025_0.05
LASSO	-1.85E-16	8.84E-14	0.253449374	0	0	-3.16E-16	0	-1.050227858	1.393426744	0
Ridge	-1.48E-16	-3.84E-14	0.375354197	-0.1225891	-3.09E-06	-4.16E-16	-1.17E-13	-0.592149562	2.274878848	-4.77E-06
OLS	-6.80E-16	6.27E-13	0.629234494	1.499282761	1.52E-05	-1.07E-15	2.21E-13	-0.895513343	2.706435455	-1.24E-06
Delta_LASSO_Ridge	-3.71E-17	1.27E-13	-0.121904823	0.1225891	3.09E-06	9.96E-17	1.17E-13	-0.458078296	-0.881452105	4.77E-06
Delta_LASSO_OLS	4.95E-16	-5.39E-13	-0.37578512	-1.499282761	-1.52E-05	7.57E-16	-2.21E-13	-0.154714515	-1.313008712	1.24E-06
Delta_Ridge_OLS	5.32E-16	-6.66E-13	-0.253880297	-1.621871861	-1.83E-05	6.57E-16	-3.38E-13	0.303363781	-0.431556607	-3.53E-06
	tmc_0.05_0.1	amc_0.05_0.1	war_0.05_0.1	ar_0.05_0.1	nc_0.05_0.1	tmc_0.1_100	amc_0.1_100	war_0.1_100	ar_0.1_100	nc_0.1_100
LASSO	-6.54E-16	5.91E-14	-0.179448996	0	6.32E-07	1.45E-15	9.95E-14	0	-0.012834552	9.59E-06
Ridge	-6.21E-16	1.21E-13	-0.079770459	0.356211586	-3.27E-06	1.37E-15	6.86E-14	0.005792268	-0.057289529	-1.14E-05
OLS	-2.33E-15	6.01E-13	-0.151349727	0.012457755	4.29E-05	7.62E-15	-2.51E-13	0.010104441	-0.088989746	-3.26E-05
Delta_LASSO_Ridge	-3.33E-17	-6.15E-14	-0.099678537	-0.356211586	3.90E-06	8.46E-17	3.09E-14	-0.005792268	0.044454977	2.10E-05
Delta_LASSO_OLS	1.68E-15	-5.42E-13	-0.028099269	-0.012457755	-4.22E-05	-6.17E-15	3.50E-13	-0.010104441	0.076155194	4.22E-05
Delta_Ridge_OLS	1.71E-15	-4.80E-13	0.071579268	0.343753831	-4.61E-05	-6.25E-15	3.20E-13	-0.004312173	0.031700217	2.12E-05
	lambda_best	averageMeanSquaredError								
LASSO	0.000352662	0.005322058								
Ridge	0.035084354	0.005203787								
OLS	0	0.00539387								
Delta_LASSO_Ridge	-0.034731692	0.000118271								
Delta_LASSO_OLS	0.000352662	-7.18E-05								
Delta_Ridge_OLS	0.035084354	-0.000190083								

Analysis:

Out of all the regressions, the one with the best AMSE was the ridge regression. That said, all the regressions had very similar AMSEs. This makes sense when the optimal lambdas for the two involving it were very close to zero to begin with. This means that the coefficients were able to be closer to the OLS regression than if a larger lambda had been found to be optimal. Looking at the graph that shows the predictions versus the actual values, we can see that the predictions seem to be tamed in a sense, there is not much volatility for all the models when compared to the real data. As mentioned before, all of the models seem to get the general direction from the mean correct, but fall short on magnitude. We can clearly see that using our LASSO model many coefficients are found to be zero, which is exactly what we would expect from such a model. When I was running the program with a smaller number of trials to pick the best from, I found that many of the best models had only a dozen or so nonzero coefficients. Our best model out of 10,000 trials has 38 nonzero coefficients. This is potentially something worth exploring further, to see how models with only a few nonzero coefficients perform on the hidden dataset. As for the rest of the results, the coefficients tended to be large for the variables related to returns, but small for the ones related to market cap and number of companies, this is to be expected because total market cap is many orders of magnitude larger than the number of companies and the corresponding returns. Likewise, coefficients relating to number of companies are larger than ones relating to return, as was expected. Returns are in the few percent, market caps are in the billions, and number of companies are in the hundreds. This confirms the validity of the regressions because their coefficients should be inversely related to their size which is exactly what we see.

Potential Further Exploration:

We could further expand the number of portfolios that we included. Additionally, the bounds that were selected for the portfolios could also be adjusted and might produce better results. Also, if this project were to be something that would be up for publication, we would need to address our collinearity issues.

Conclusion:

This project sought to better understand the drivers of volatility, and I believe that was accomplished. Clearly, this idea that a few portfolios can drive volatility is not the whole picture, but it did show that they do play some role in it. Our predictions were able to get the correct direction from the mean the returns changed over time, but they fell short when predicting the scale of this movement. The biggest hurdle encountered in this project was in creating usable data from a massive and error ridden dataset, this took many hours over many days to accomplish, but in the end it was worth it.

GithubLink:

https://github.com/trentmckinnon/Matrix-Methods-of-Machine-Learning/

(Unfortunately, cannot post dataset because it is proprietary to WRDS and would violate TOS)