Cryptocurrency Final Project Report

Gerasimos Mouikis, Spencer Pope, Jia Lin Hau

Monkie Business

DS 768.02

5/7/18

1. Introduction

There is a very good chance that somewhere, somehow, everyone has heard about Bitcoin by now, and consequently, it is beginning to impact the way the public views trading currencies. Cryptocurrencies are "a digital currency in which encryption techniques are used to regulate the generation of units of currency and verify the transfer of funds, operating independently of a central bank."(1). These cryptocurrencies have been around for a long time now, but recently gained great popularity, because Bitcoin's stock price skyrocketed and in the course of only one year, grew by 2,000%. It was worth approximately $20,000 per share in December of 2017. Due to these events, trying to predict the next big cryptocurrency and how high it's price will rise, are some of the most talked about topics in the stock trading world today. Everyone wants to be the first one to buy the next big cryptocurrency before it blows up like Bitcoin did.

The practical consequence of solving this problem is that Bitcoin gives us, for the first time, a way for one Internet user to transfer a unique piece of digital property to another Internet user, such that the transfer is guaranteed to be safe and secure, everyone knows that the transfer has taken place, and nobody can challenge the legitimacy of the transfer. The consequences of this breakthrough are hard to overstate. (2)

For our project, we will be creating a forecasting method for predicting the prices of cryptocurrencies in the future. Our data comes from a dataset on Kaggle.com called "Cryptocurrencies: Historical Data for 1200 Cryptocurrencies (excluding bitcoin)". The dataset starts in early 2013, but due to the fact that the 12 cryptocurrencies we picked only started overlapping in 2016, our data will go from the end of October 2016, to the end of June 2017. The data was observed and recorded daily for this 8-month period for all 12 different cryptocurrencies and there is an obvious positive trend that appears around halfway through our data set around March of 2017.

We used time series methods, to help us forecast the future prices of these 12 different cryptocurrencies. We used multiple methods to predict the data trends, such as: ensemble

averaging, LASSO regression, vector autoregression, and benchmark methods. We split the data into a training set and test set and analyzed the test set to see how well it was working. To better validate the methods, we calculated a few different kinds of errors such as MAPE, MSE, and RMSE.

2. Data

The data in our forecasting study is calculated using a collection of time series that record the open, close, high and low price for the day, volume traded and date. From among the selected data, we selected closing price in terms of bitcoin as our major factor, as it is closely correlated between the different cryptocurrencies. We focused on 12 different highly traded and famous crypto currencies with a base unit of BTC (bitcoin). Our data covers an 8 month period starting from October 2016 through June 2017. The main reason that we only focused on these eight months, was because this was the period that data in all cryptocurrencies was concurrently available according to our chosen dataset. Using this information, we tried to retrieve relevant and important information showcasing the relationship between all 12 cryptocurrencies.

Figure 1 presents time series plots of 4 different cryptocurrencies price (Ethereum, Augur, Lisk and MaidSafeCoin) with a base unit of BTC(bitcoin). Our full dataset consists of 12 cryptocurrencies with their price at the y-axis and date at the x-axis of time series. 17,110 depicts the starting date and 17,350 depicts the end date of our data set.

Figure 1:

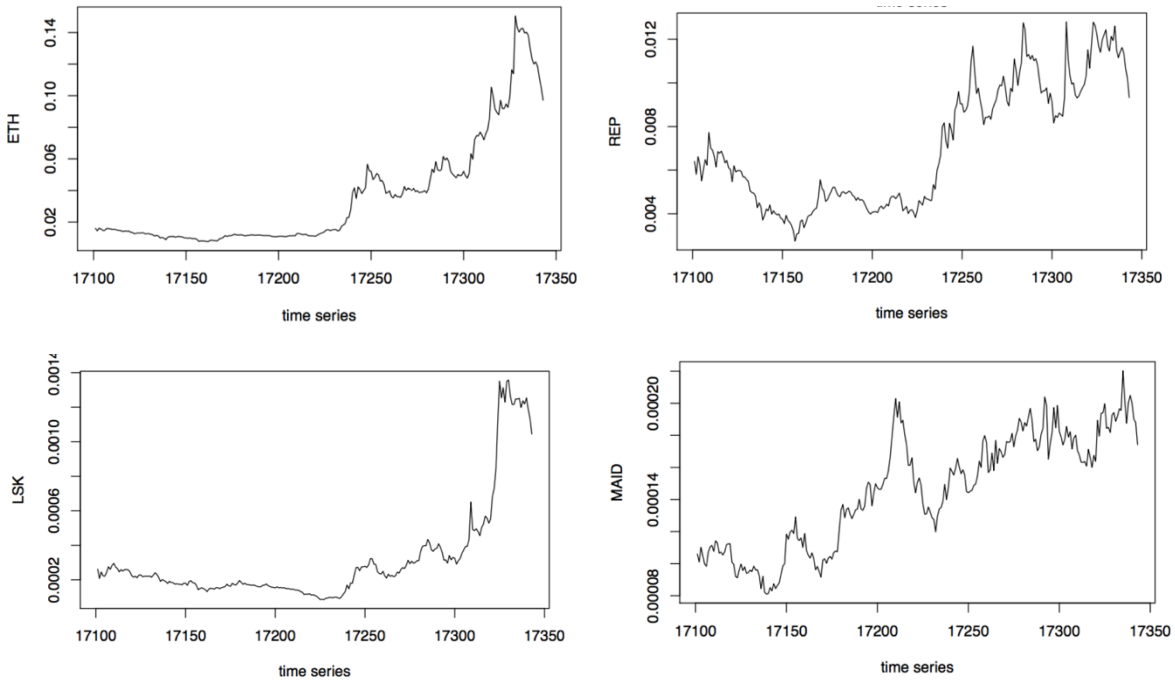
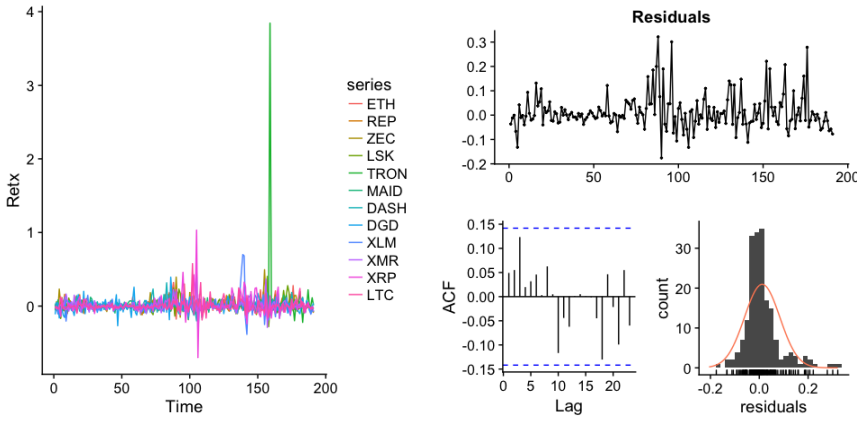


Figure 2 (left) represents the daily return of each currency, and Figure 2 (right) represents the analysis of the daily return for Ethereum only. From the analysis of the daily return, we can see that its pattern of the daily return is very randomly distributed along the time series, and there seems to be no significant relationship between the lag, while the return is normally distributed around zero. The return of all other currencies looks very similar to the return of Ethereum. Since it is a random normal distribution, the distribution of return is more similar throughout the currencies compared to the price distribution because the distribution of the price just reflects to the period of time the currency price stayed at a level.

Figure 2:



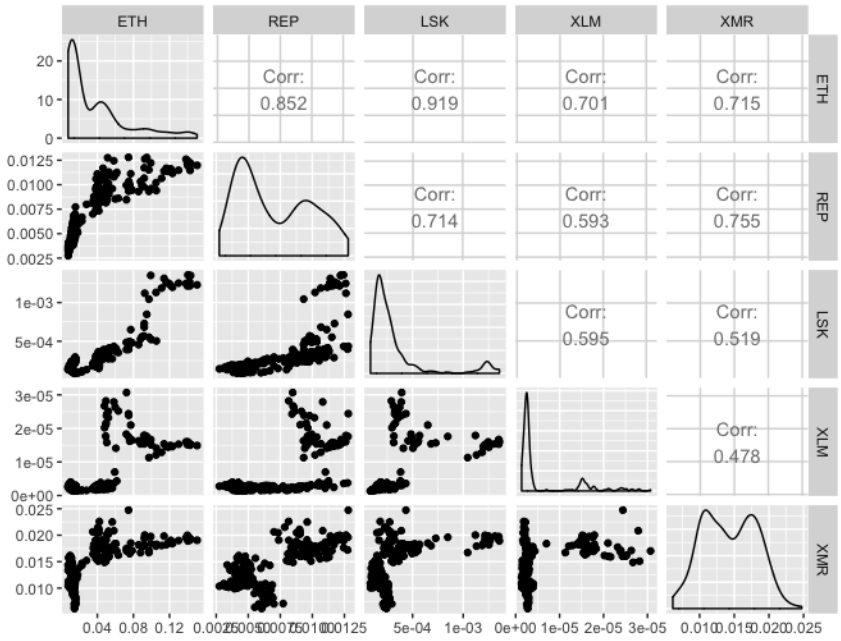
The following formula is the formula we used to compute the return from price and back:

$$Return_{k-1} = \log log\left(\frac{t_k}{t_{k-1}}\right) \qquad [1.1]$$

$$P_{k+n} = e^{P_k} * \prod_{t=0}^{n-1} Return_{k+t} \qquad [1.2]$$

Figure 3 shows the correlation between 5 of our cryptocurrencies. The plot shows each combination of possible scatterplot and distribution of each price of cryptocurrency and the correlation of each combination of price of cryptocurrency. Based on our prediction model, it appears that if the price of bitcoin goes up, the price of all other cryptocurrencies will go down relatively, this is because the price of our cryptocurrencies is in base unit of bitcoin. Thus, it appears that the 5 cryptocurrencies chosen could be intercorrelated.

Figure 3: Correlation between all cryptocurrencies



There were several challenges in making accurate conclusions, based on our chosen dataset.

One of the challenges in working with this data was that we only had a total of only 243 days worth of data when the 12 different currencies were concurrently available. This could have easily resulted in an overfitting in the training set, which consequently could result in inaccurate predictions. It is possible that the limited amount of data available is skewing our results.

Another challenge was that we are working with an ambitious objective that was difficult to achieve. We tried to identify/predict a cryptocurrency price based on previous market values. This is an obvious problem and the main reason why a lot of people trying to make similar predictions keep their algorithm secret, especially if they have found a way to accurately predict prices. Thus, it is extremely hard to find detailed descriptions and reliable strategies or methods of how this can be achieved and without delving deeper into the limited available literature.

3. Methods

1. VAR (Vector Autoregression)

After trying multiple data analysis methods including: Ensemble Averaging, LASSO and Vector Autoregression (VAR), we found that VAR appears to be the best fit for our data, because our data is a time series and VAR is able to keep the information that is produced by previous lags. We used the VAR function in R to explore which currency has the most significant effect on each lag or which of the lags observed was the most significant for each currency. Based on the VAR results, we reduced the dimension of the model and run a multivariate linear regression.

$$y_{1,t} = \beta_0 + \sum_{k=1}^{n} \sum_{i=1}^{12} \beta_{i+12(k-1)} y_{i,t-k} + \varepsilon$$

To use this function, we had to assume three factors:

- That the price of a cryptocurrency had a linear relationship between lags for all currencies.
- Every variable was assumed to influence every other variable in the system.
- A zero correlation was implied between error terms as a desired method.

There was no statistical evidence according to our dataset, that showed any of our assumptions might be violated since the prices are correlated and we assume they influence each other.

The objective of the VAR function is to maximize the likelihood of the condition. Thus, we used a maximum likelihood estimator. It is also equivalent as multivariate least square (MLS) approach for estimating the coefficient. MLS is similar to OLS but for more dimensions.

2. Benchmark

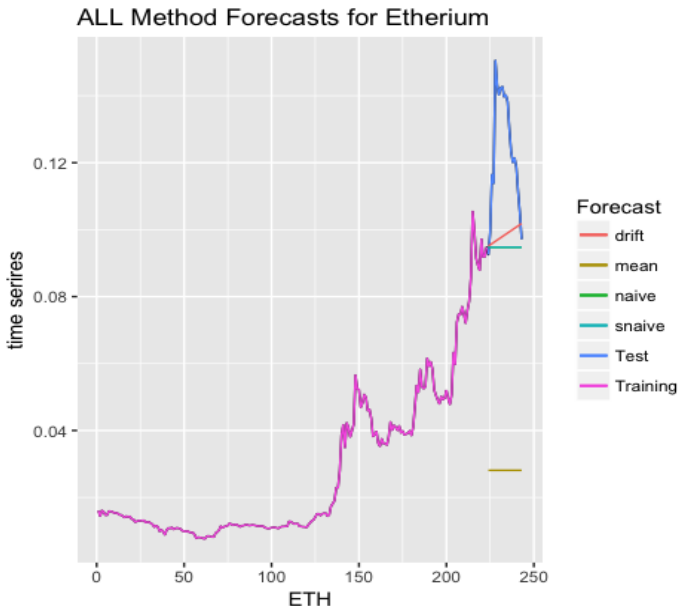


Figure 4. ALL Method Forecast for Etherium

When we tested our benchmark methods, we originally tried plugging in drift, mean, naïve, and seasonal naïve methods into our training and test set plots, and we then compared them with one another. Based on our observations, it appears that the drift method was the most accurate of the four methods we used (Figure 4).

Our benchmark method included drift, mean, naïve and snaive. Both the naïve method and snaive methods use the most recent data provided as a prediction of the future. In our case snaive is equal to the naïve because there is no apparent and significant seasonal pattern observed $thus\ P_{t+n} = P_t$. The mean method uses the mean of all the data in our dataset to predict future prices. $P_{t+n} = \overline{P_{1:t}}$ (the P bar is the average P from 1 to t). The drift method takes the first and last point on the data and expands it further as the prediction $P_{t+n} = P_t + \frac{P_t - P_1}{t-1} n$.

None of these 4 methods appears to accurately predict future prices of cryptocurrency. The mean in particular, appeared to be notably the least suitable of the methods explored in this assignment, while the drift method appeared to be the most suitable. Since drift gives us the best prediction out of all the common benchmark, we will let drift to be our benchmark (Figure 5).

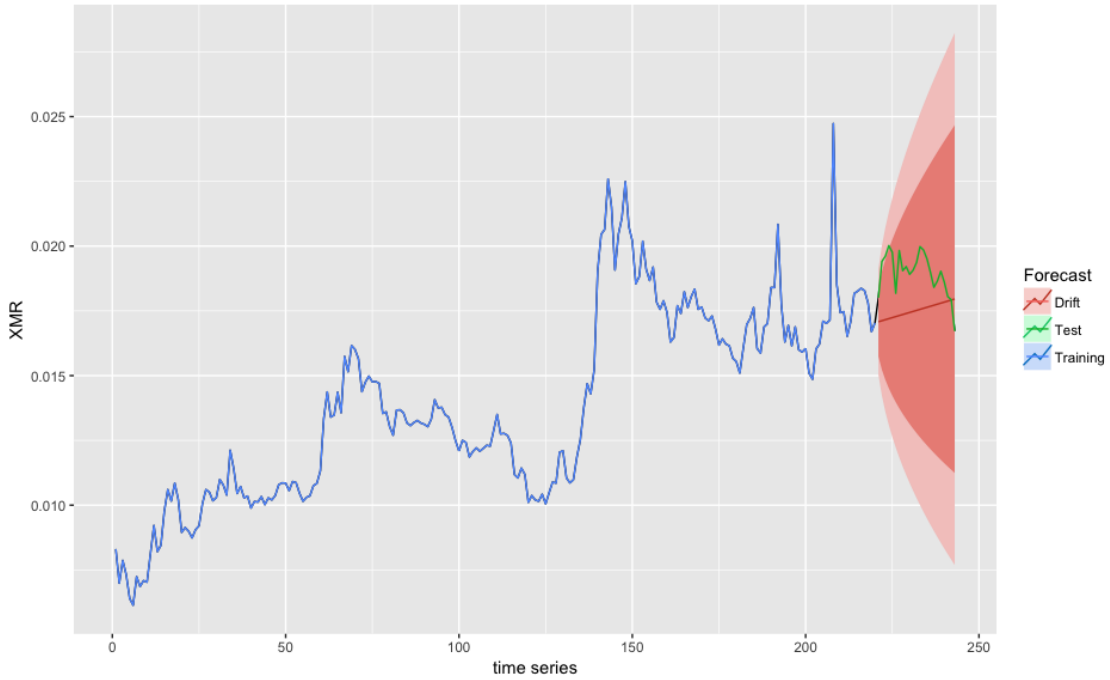


Figure 5. Drift method on price of XMR with 80 and 95 percent confidence intervals.

3. Results

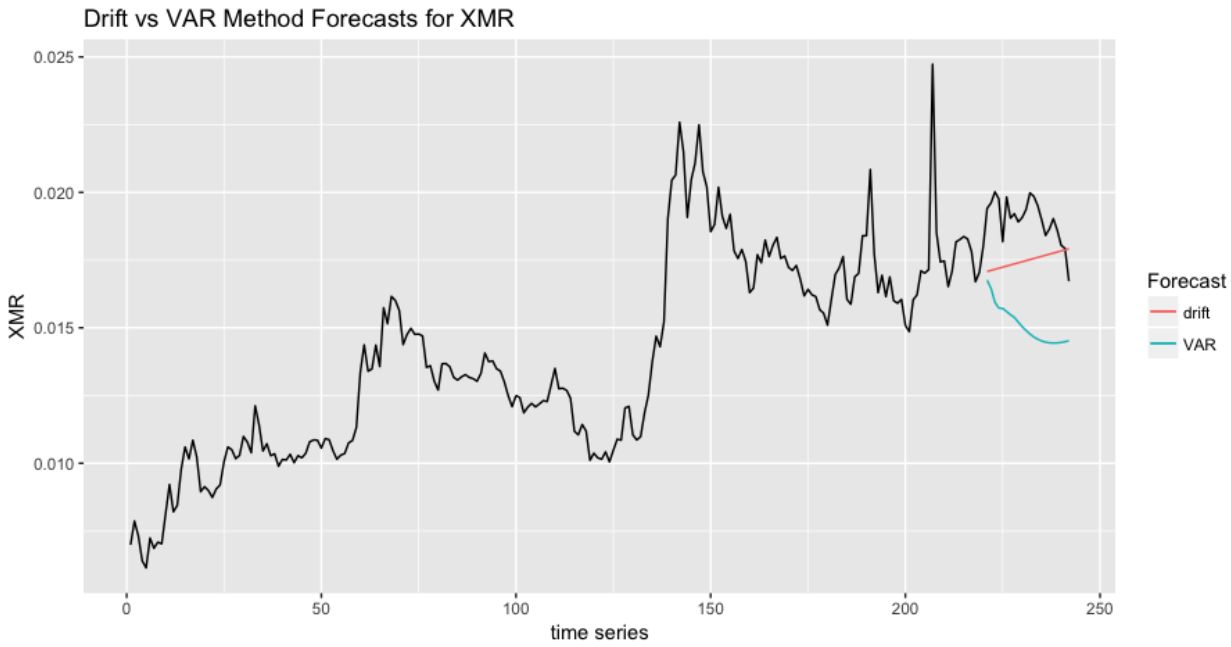


Figure 6. Drift vs Method Forecast for XMR

We initially used VAR on the price of the cryptocurrencies, however, the VAR of the prices did not appear to work well at all. It was even worse than our benchmark method (drift method) according to Figure 6. We believe this is mainly because most of the lags and prices of currencies are highly correlated. With multiple lags of prices trying to predict future prices, it easily over fit the model and thus produced a model that does not work well.

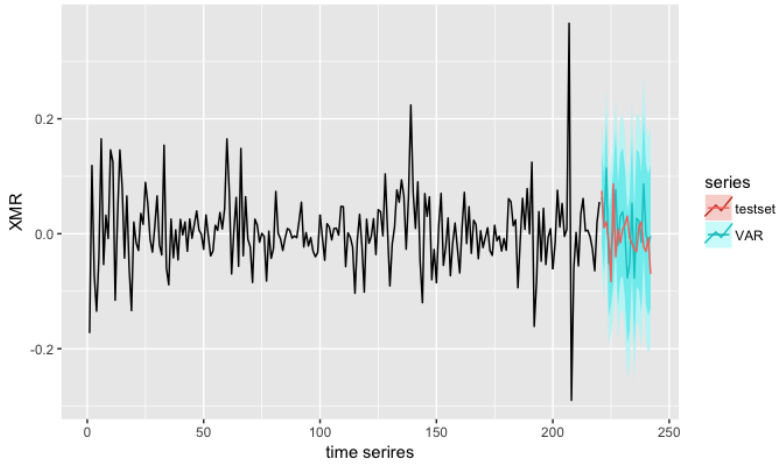


Figure 7. Time vs XMR

We used the return formula in the data section [1.1] to transform the price to daily return. Then, we used VAR to predict the daily return instead of the price. The result of the return prediction is given on the figure above. Note that the actual test data and the prediction by the VAR of the return stay very close to each other, as well as within the confidence intervals. We then used the formula in data section [1.2] to transform the predicted return back to price level and used that as the prediction of future price.

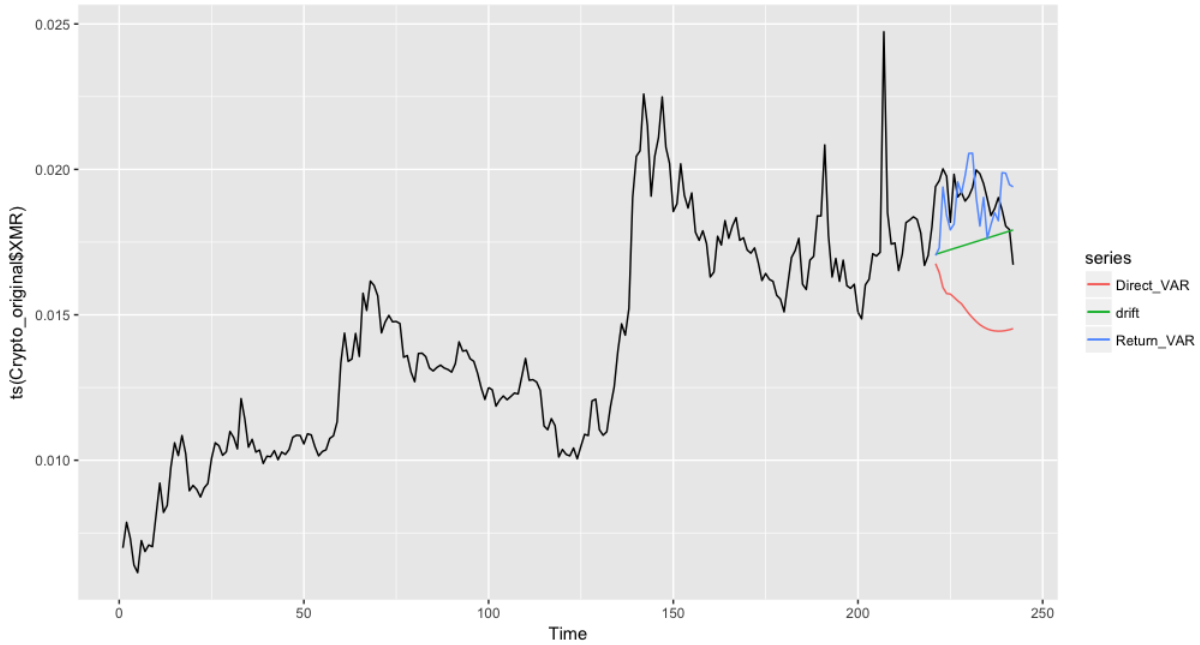


Figure 8. Prediction of 2 different VAR results and the drift method.

	ME	RMSE	MAE	MPE	MAPE
Direct_VAR	0.0039739540	0.004070316	0.003973954	20.7662904	20.766290
drift	0.0015667148	0.001749994	0.001566715	8.0933620	8.093362
Return_VAR	0.0001994602	0.001210431	0.001011989	0.9248174	5.290679

Table 1 Evaluation of the accuracy of 2 different VAR results and the drift method.

Based on Figure 8 it is clear that the VAR prediction that we transformed from the return prediction works the best. The drift method was the second best and the prediction of VAR based on the price performed the worst of the three. We also were able to numerically evaluate this result (Table 1). We can see in the table that the VAR of return calculated has a Mean Absolute Percent Error (MAPE) of 5.29 which outperformed the drift method that has an 8.09 MAPE and the price VAR which has a MAPE of 20.77.

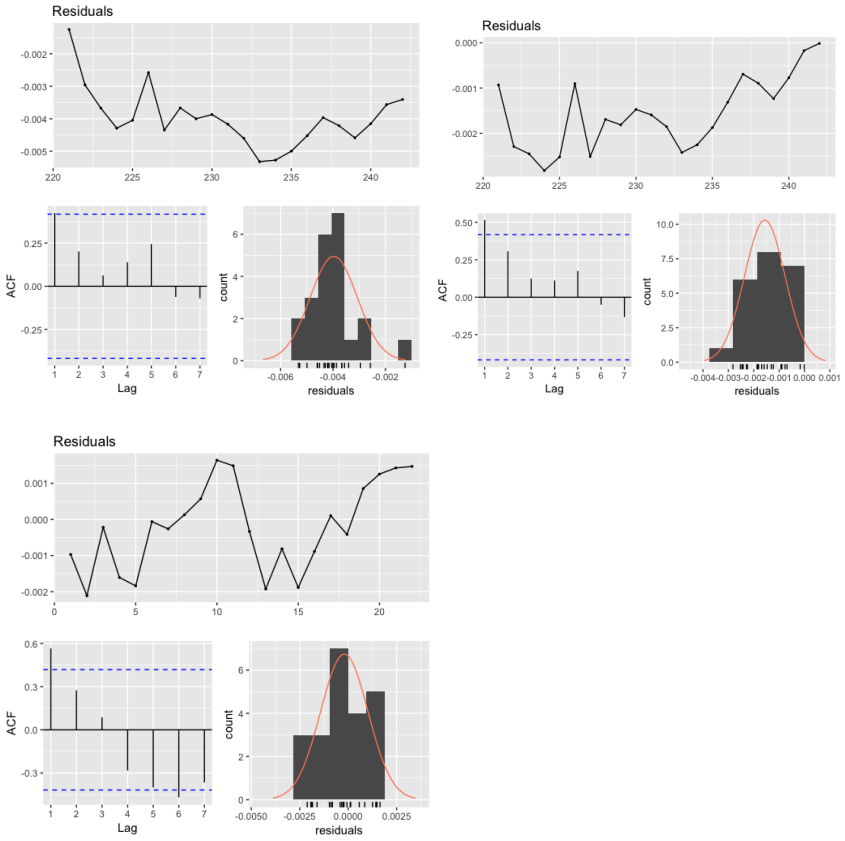


Figure 9. Residual plots of 2 different VAR results and the drift method.

According to the histogram of the residuals (Figure 9), the VAR price prediction gave a residual with a mean/median around -0.004, which is not zero. The residual of the drift method has a mean and median around -0.0015, which is closer to zero and thus better than the VAR price prediction. However, the residual of the VAR return method is almost zero. Although, the drift and VAR of return has a significant correlation in the first lag, it is not too far from the significance limit, so it is acceptable.

In conclusion, the VAR works better on the return compared to the price level. With a reasonable and good transformation model, we could improve the prediction significantly with the same methods. Overall. the VAR of return worked the best, drift second, and VAR of price was the worst.

5. Discussion

At the beginning of this project, we quickly noticed several interesting things about our dataset. Even though data was available for a total of 1,200 cryptocurrencies, only a few were notable in terms of their price fluctuation, while the rest were so irrelevant that their stock prices never even changed throughout the 8 months we included in our dataset.

As mentioned previously, our dataset only covered the period between the end of October 2016 to the end of June 2017. Our forecasting predictions would have been more accurate have we had data covering a longer period of time. Considering that interest in cryptocurrencies is a relatively recent phenomenon, it would have been useful to not only have more data, but also more recent data, especially covering the period between June, 2017 to the present time. In addition, the sudden burst in cryptocurrency/ bitcoin prices negatively impacted our ability to accurately forecast future cryptocurrency pricing, as this sudden burst was not going to continue forever and would eventually come back down.

There are many different forecasting methods available that can be used to explore this topic. Due to time limitations we were not able to investigate each one of them and find the one that would best fit our objectives and dataset. We also had to use daily data, which sounds favorable, but for our data, the prices didn't fluctuate significantly from day to day. If instead, we

had a larger amount of data, and used weekly data, we could have predicted the future prices much better. 24 hours is not a long period of time in situations like these.

The above-mentioned data related issues, negatively impacted our ability to accurately forecast the price of cryptocurrency in the future.

If we could continue this project in the future, we would most likely try to find more recent data that shows the burst of prices, as well as them coming back to more realistic levels. This would give us a lot more data to work with and increase our ability to make more accurate predictions.

Works Cited:

1:

"Cryptocurrency | Definition of Cryptocurrency in English by Oxford Dictionaries." Oxford Dictionaries | English, Oxford Dictionaries, en.oxforddictionaries.com/definition/cryptocurrency.

2:

Andreessen, Marc. “Why Bitcoin Matters.” The New York Times, The New York Times, 21 Jan. 2014, dealbook.nytimes.com/2014/01/21/why-bitcoin-matters/.