Linear Regression on Cryptocurrency Prices

Marcon Louie C. Fikingas

Western Governors University

Research Question

The advent of cryptocurrency in the early 1990's was unremarkable with electronic cash systems such as Ecash and DigiCash ending up bankrupt without any impact on financial life. Not until a decade later in 2008, the first appearance of Bitcoin emerged that would later change the financial marketplace. Bitcoin would later be valued at over $10,000 in 2017 and continues to grow today. Other cryptocurrencies have been established since then. The blockchain technology behind the electronic transaction systems allow for secure and instant peer-to-peer transactions. Cryptocurrencies allow for a decentralized form of currency based on a peer-to-peer network free from hierarchical structures (Farell, 2015). The market value of all the world's cryptocurrency surpasses $1 trillion dollars.

This research paper aims to develop a model that can determine future prices of different cryptocurrencies. More specifically: can a model be developed to predict the prices of different cryptocurrencies based on historical data?

There have been made many attempts to predict the prices of cryptocurrencies because of the lucrative possibilities. Cryptocurrencies are still relatively new to the financial market and the data available is sparse compared to the historic data for stocks and other market tickers. The model for this study focuses on daily dates, open, high, low, volume of the cryptocurrency and the volume in terms of the United States dollar in order to predict the close price. These variables are used determine the closing price of each respective cryptocurrency. The data originates from the daily values of Bitcoin, Ethereum, Litecoin, XRP, also known as Ripple and Bitcoin Cash.

In 2010, Bitcoin began trading publicly and the first instance of a cryptocurrency being given a value had occurred. In 2013, Bitcoin reached an all-time high of a $1,000, a surprising

125,000,000% increase from Bitcoin's initial value of $0.008. Today, there are more than 1,500 cryptocurrencies in existence. Identifying potential cryptocurrencies prices would help investors decide which cryptocurrencies to invest in for profitable returns. Additional and future study could determine what other outside factors, such as social cues, affects cryptocurrency prices and why such factors have an influence on prices.

The study analyzed the data at a daily level, which provides enough observations to create a linear regression model. Another option was to use datasets on an hourly basis; however, doing so would create extremely large and redundant data. The daily data dates as far back to 2014 and makes for adequate data to create a linear regression model. Most of the data is still relatively new considering that cryptocurrencies have only begun trading in the last decade, but this allows for a vast number of opportunities to study the emerging data. One major reason cryptocurrency is relevant to the future is the use of cryptocurrencies in international business transactions (DeVries, 2016).

There are certain disadvantages to using daily basis data. Since cryptocurrencies are being developed simultaneously as they are being traded, historical data on specific cryptocurrencies will differ with each other, since no two cryptocurrencies have been developed and opened to trading on the same day. Another downside to using this data is that there are over 1,500 cryptocurrencies in circulation; this study analyzed only five major cryptocurrencies that have the most abundant data. With more data behind the cryptocurrencies, the better the linear regression model will be. One might consider using the hourly data instead for this reason; however, cryptocurrencies are prone to volatile ups and downs in small amounts of time. By using daily data, the volatility of cryptocurrencies can be somewhat contained. Overall, new emerging data will only help to improve a study of this design for future studies.

With the current understanding of the state of cryptocurrency data, this study has this proposed hypothesis:

A model based on predictive regression can be created from historical cryptocurrency datasets. The null hypothesis states that a model based on predictive regression cannot be created from historical cryptocurrency datasets.

This study contains understandable variables that are consistent with each different cryptocurrency dataset. The model developed is based on internal data and variables, dates and different pricing measures. This study does not look at outside factors that may influence cryptocurrency pricing; that would be relevant to analyze in a different study. Altogether, the hypothesis is straight forward: either a model can be developed to predict future cryptocurrency prices or not. If such a model can be developed, then this would greatly influence potential investors in the cryptocurrency market.

Data Collection

The data for this study was relatively easy to acquire. Since cryptocurrencies are publicly traded on a national and international scale, the data is recorded immediately. The datasets themselves are publicly available and are updated daily. For this study, a total of five datasets were retrieved.

The data was all collected from a single source; this allowed for a consistent format for all datasets. Cryptocurrencies are traded internationally. The source had the datasets formatted in three various ways: US dollars, Euros and in Bitcoin. This study used the datasets formatted in US dollars. The data can easily be transformed to fit the other two formatted datasets.

The advantage of collecting the data from a single source is that that the format is consistent and any errors that may be found can be easily traceable to the original source. A possible disadvantage to collecting from a single source is that the data may be unreliable if the source was not a verified source. One way this study overcame this challenge was to cross reference the data with other sources and compare the daily closing prices. In doing so, the comparison to other sources revealed that the data this study used was reliable and accurate.

The individual datasets are based on the five key cryptocurrencies. Bitcoin is the first dataset and contains the largest number of observations. The second dataset is Litecoin, which was created in 2011. Litecoin is an alternative cryptocurrency that has a faster block generation time, which translate to faster mining, and has a larger coin limit than Bitcoin. The third dataset in this study is Ethereum. Ethereum was created in 2015 and is currently the second highest valued coin behind Bitcoin. The fourth dataset used is based on XRP, also known as Ripple. Ripple is one of the more controversial cryptocurrencies with a pending class action lawsuit still being waged. The fifth dataset is based on Bitcoin Cash, an offshoot of Bitcoin. The key difference between Bitcoin and Bitcoin Cash, is that Bitcoin Cash has a faster transaction process than Bitcoin. Overall, all datasets are updated daily and are reliable.

Data Extraction and Preparation

All the datasets need to be prepared before analysis. Since there are no missing values, data scarcity is zero. For the data to be efficiently analyzed, redundant variables need to be eliminated. A single variable in each dataset needed to be renamed to have the variable consistent throughout the analysis. After all the preparation, the datasets were imported into SAS for further analysis. SAS is preferred over Python and R when it comes to handling large datasets

(Brittain, Cendon, Nizzi, & Pleis, 2018). Overall, these were the only changes needed to be made.

For this study, Microsoft Excel was used to extract the data from online, import the data as csv files and prepare the data. Excel was an easy way to drop and change variables in each individual dataset without too much complexity. Each dataset ranged from around 1,000 observations to around 2,000 observations. Using Excel allowed the small datasets, like the ones used for this study, to be quickly searched for any missing values. Using other programs, such as R, would have made the preparation process more complex. In R, the same preparations are available; however, the steps required are not as straightforward. Multiple import command lines and drop and rename lines would have been necessary. Along with that, R would have needed to create separate worksheets for each dataset and then would have needed to create a single spreadsheet containing all sperate worksheets. This would have taken multiple lines of code to implement. In the end, Microsoft Excel was the better choice for such small datasets.

The preparation needed were efficiently done. In this study, the variables were dropped due to redundancy purposes, or they contained unnecessary information. Removing data redundancies ensures data integrity (Hong, 1994). The unix variable, which contained a unix timestamp to convert to local time, was dropped due to redundancy, since there is already a data variable sufficient present. Another variable dropped was the symbol variable, which contained the cryptocurrency's abbreviated symbol. The only other variable that needed to be prepared was each datasets volume by cryptocurrency to "Volume Crypto" so that the datasets maintain their consistency. After preparation, the data was imported into SAS for analysis.

SAS was used because most major businesses use SAS as their preferred choice of program. Not only that, but SAS is also a closed source program, which ensures data

confidentiality (Anurag, 2020). Another advantage of SAS over R and Python is the ability to perform statistical analysis with an easy-to-use interface that allows for graphs, plots and libraries.

date	open	high	low	close	Volume Crypto	Volume USD
2021-02-25 0:00	49715.13	49980	49705.38	49853.78	48.79198139	2432464.706
2021-02-24 0:00	48927.33	51459.51	47000	49754	8774.366743	436559843
2021-02-23 0:00	54173.64	54196.58	44845.72	48887.93	17766.39815	868562429.1
2021-02-22 0:00	57485.74	57564.19	47400	54173.65	17002.83273	921105509.4
2021-02-21 0:00	55919.75	58354.14	55537.99	57492.91	3562.96951	204845485.3
2021-02-20 0:00	56000.59	57553.81	54000	55936.04	5187.250463	290154249.4
2021-02-19 0:00	51560.46	56399.99	50627.32	55988.58	9229.782504	516762416.1
2021-02-18 0:00	52141.23	52566.98	50869.61	51579.54	6330.142919	326505859.9
2021-02-17 0:00	49136.7	52640	48896.19	52174.28	9571.647969	499393841.2
2021-02-16 0:00	47954.05	50602.53	47036.02	49166.53	8376.076088	411822596.3
2021-02-15 0:00	48620.48	49048.82	45914.75	47942.57	6924.334089	331970371.7
2021-02-14 0:00	47258.66	49714.66	47081.02	48662.5	5201.693926	253127430.7
2021-02-13 0:00	47424.25	48220	46133	47228.48	4271.672151	201744582.8
2021-02-12 0:00	48044.88	49000	46231.16	47395.84	5970.135903	282959606
2021-02-11 0:00	44854.63	48696.84	44040.96	47981.48	9390.816355	450585267.1
2021-02-10 0:00	46493.9	47364.09	43746.06	44854.63	9483.328956	425371211.5
2021-02-09 0:00	46445.51	48216.09	45000	46505.2	12505.08248	581551361.9
2021-02-08 0:00	38870.36	46712.12	38050.97	46416.45	19053.71344	884405737
2021-02-07 0:00	39274.8	39727.44	37412.93	38858.39	5973.240161	232110495.7
2021-02-06 0:00	38320	41025.48	38230.81	39282.1	8635.660207	339226867.8

Figure 1 – An excerpt from the Bitcoin dataset after initial data preparation. Note that the date contains default time of midnight.

Within the SAS program, the datasets are individually manipulated to clean the date column. The original data column was a datetime variable, which contains both date and time as a single numeric variable. In SAS, each dataset converted the date column with the datepart function to extract only the date part of the original variable. The extracted data was originally formatted as the number of days elapsed since January 01, 1960. Within the procedure, the format statement was used to convert the number of days elapsed to a more readily readable

format month, day and year. Overall, the date variable in each dataset was first converted from datetime to date and then formatted accordingly.

date	open	high	low	close	Volume_Crypto	Volume_USD
02/25/21	49715.13	49980	49705.38	49853.78	48.79198139	2432464.706
02/24/21	48927.33	51459.51	47000	49754	8774.366743	436559843
02/23/21	54173.64	54196.58	44845.72	48887.93	17766.39815	868562429.1
02/22/21	57485.74	57564.19	47400	54173.65	17002.83273	921105509.4
02/21/21	55919.75	58354.14	55537.99	57492.91	3562.96951	204845485.4
02/20/21	56000.59	57553.81	54000	55936.04	5187.250463	290154249.4
02/19/21	51560.46	56399.99	50627.32	55988.58	9229.782504	516762416.1
02/18/21	52141.23	52566.98	50869.61	51579.54	6330.142919	326505859.9
02/17/21	49136.7	52640	48896.19	52174.28	9571.647969	499393841.2
02/16/21	47954.05	50602.53	47036.02	49166.53	8376.076088	411822596.3
02/15/21	48620.48	49048.82	45914.75	47942.57	6924.334089	331970371.8
02/14/21	47258.66	49714.66	47081.02	48662.5	5201.693926	253127430.7
02/13/21	47424.25	48220	46133	47228.48	4271.672152	201744582.8
02/12/21	48044.88	49000	46231.16	47395.84	5970.135903	282959606
02/11/21	44854.63	48696.84	44040.96	47981.48	9390.816355	450585267.1
02/10/21	46493.9	47364.09	43746.06	44854.63	9483.328956	425371211.5
02/09/21	46445.51	48216.09	45000	46505.2	12505.08248	581551361.9
02/08/21	38870.36	46712.12	38050.97	46416.45	19053.71344	884405737
02/07/21	39274.8	39727.44	37412.93	38858.39	5973.240161	232110495.7
02/06/21	38320	41025.48	38230.81	39282.1	8635.660207	339226867.8

Figure 2 – An excerpt from the Bitcoin dataset after date column conversion. The column has been cleaned to be easily readable.

Analysis

Before fitting a multiple linear regression model, the data must be checked for the four assumptions of linear regression. The four assumptions are independence, homoscedasticity, normality and linearity.

For this study, a linear regression model was used in this study because of the trend cryptocurrencies are showing in real time scenarios. A similar study was conducted in support of a linear regression model predicting Bitcoin by Lee (2018), which described the impact of target features in a linear model. This study focused on internal features of individual cryptocurrency datasets.

This study first looked at the Bitcoin dataset. Scatterplots are used to check the assumption of linearity. The relationship between the outcome variable and the independent variables may either be linear or curvilinear. Bitcoin emerged as the leading cryptocurrency in terms of price; however, all cryptocurrencies began with relatively small volume and even smaller starting prices, as depicted in the Volume Crypto and Volume USD graphs. Most of independent variables displayed a linear relationship with the dependent variable, apart from the date variable.

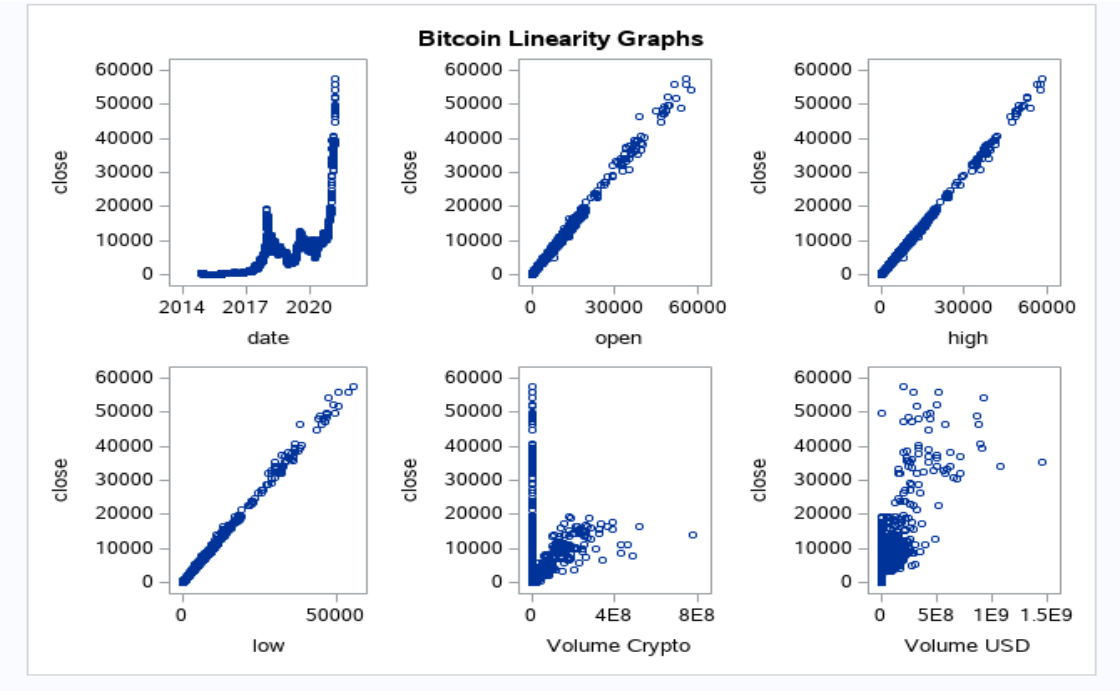

Figure 3 – Scatterplot of Bitcoin independent variables and the dependent variable.

The next assumption that was tested, was the assumption of normality. There are two graphs to verify normality: the residual Q-Q plot and residual histogram plot. SAS produces these two plots, along with other diagnostic plots using the proc reg procedure. The histogram plot displayed rejected the assumption of normality. For the second graph, the Q-Q plot showed a visible pattern, which is explained by the date variable, confirming a time-series correlation with the dependent variable. Overall, the Bitcoin data does not pass the assumption of normality.

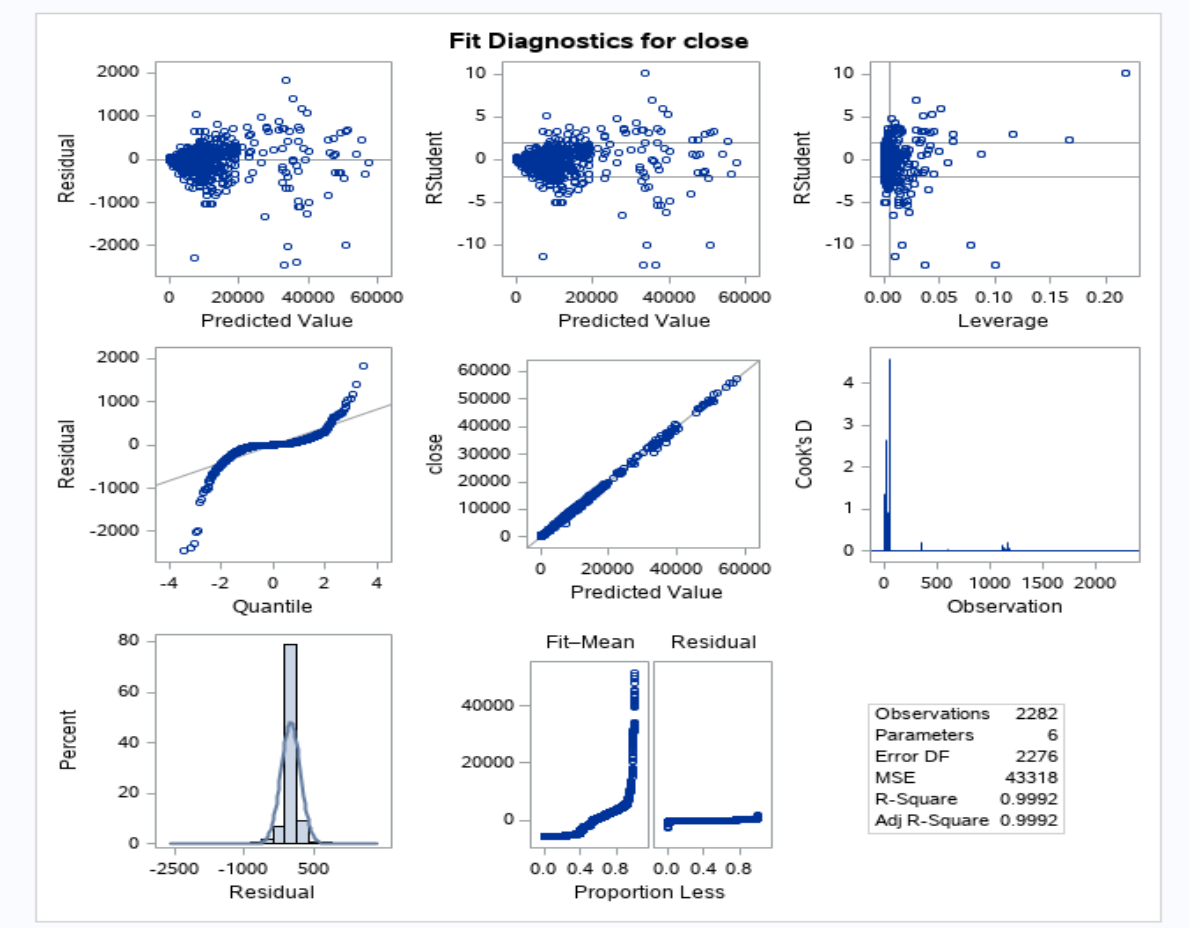

Figure 4 – Dianostic Plots for Bitcoin containing the Q-Q plot and histogram plot.

The assumption of homoscedasticity was tested using a couple different methods. One of the methods used was the White Test, which test the variance of errors in a regression model.

The test determined that the model was heteroscedastic, which means that the variance of errors is not approximately equal.

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
20	175.49	<.0001

Figure 5 – The Test of First and Second Moment Specification displays the results of the White Test. With a p-value of less than 0.5, the study rejects the null of homoscedasticity.

Another method used to determine homoscedasticity was the graph of the residuals by the predicted values. The graph displayed a cone pattern that indicates the violation of homoscedasticity. Overall, the two methods both indicate an unequal variance of errors.

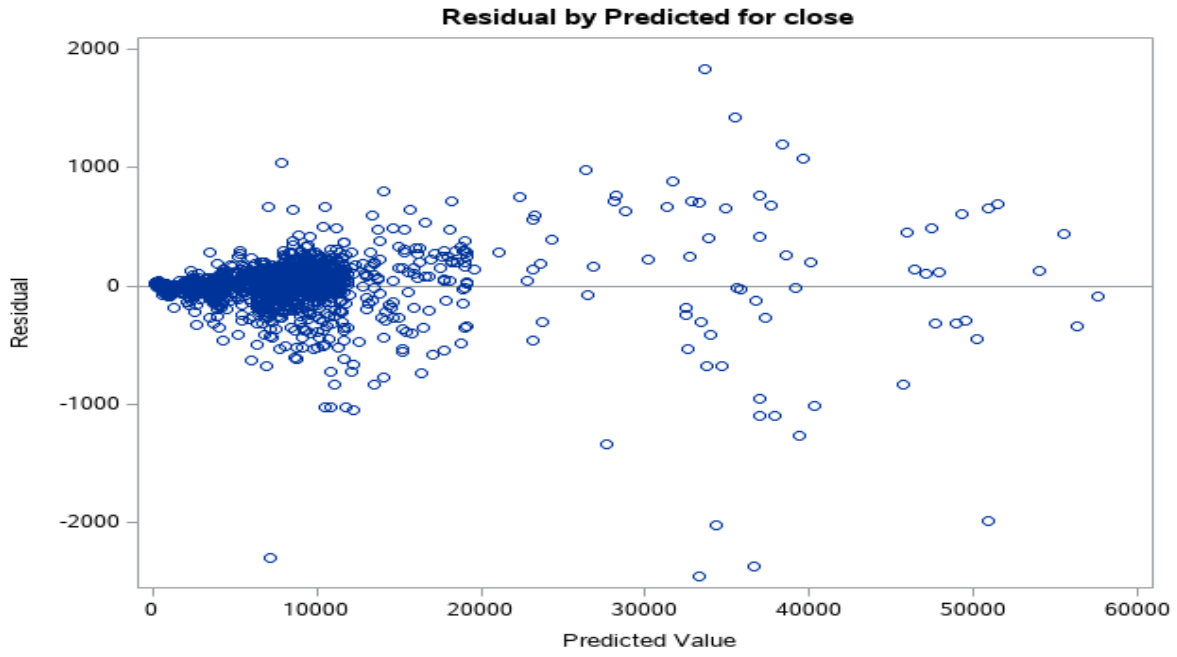

Figure 6 – The graph of the Bitcoin dataset Residuals by Predicted Values. Note the cone shape pattern, as the predicted values increase, so does the range of the residuals, thus a clear indication of a pattern.

The fourth assumption that needs to be verified is the assumption of independence. This assumes that there does not exist a relationship between the residuals and the variable. This study used the Durbin-Watson coefficient to test for autocorrelation. The Durbin-Watson coefficient for the Bitcoin dataset was 2.095. This indicated an absence of first-order autocorrelation; thus, the coefficient indicated no violation of independence.

Durbin-Watson D	2.095
Number of Observations	2282
1st Order Autocorrelation	-0.049

Figure 7 – The Bitcoin dataset Durbin-Watson coefficient of 2.095. A value of less than 1.5 indicates a positive correlation, where a value greater than 2.5 indicates a negative correlation.

One other factor this study considered was the prominence of multicollinearity. The variance of inflation coefficients was used to determine if multicollinearity existed in the model. A variance of inflation coefficient greater than 5 indicated multicollinearity.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-1432.96563	203.31746	-7.05	<.0001	0
date	date	1	0.07042	0.00976	7.21	<.0001	2.17860
open	open	1	-0.31779	0.01441	-22.05	<.0001	574.50690
high	high	1	1.29883	0.01408	92.23	<.0001	593.47493
Volume_Crypto	Volume Crypto	1	-0.00000284	9.946304E-8	-28.59	<.0001	1.32835
Volume_USD	Volume USD	1	-0.00000282	8.134572E-8	-34.65	<.0001	3.19156

Figure 8 – Bitcoin dataset parameter estimates. Note the Variance Inflation column.

The parameter estimates tables indicated that there existed a multicollinearity issue with the variables open and high. This was not surprising, since the data is based on daily values, where the previous day values would be expected to correlate with the next day value.

Based on the four critical assumptions, a linear regression model would not work; however, the study continued with the model due to the fact the violations were known to stem from the data being based on time. The assumptions of linearity and independence were verified; however, the assumptions of normality and homoscedasticity were violated. Another issue found was the presence of multicollinearity in the model, which was to be expected from a time-based dataset.

Although the dataset does not confirm to the assumptions of a linear regression model, the data does fit well in predicting values. The R-Square value for the dataset was 0.9992. This indicated that the model explains 99.92% of the variability of the response data around the mean. The independent variables explain and predict the response variable accurately. This study used the R-Square values as the goodness of fit measure over other methods because the model is based on an individual dataset. In contrast to the adjusted R-Square goodness of fit measure, this study would be hindered using the adjusted R-Sqaure method. The adjusted R-Square would compares different models to one another; however, the initial model already has a R-Square coefficient of 99.92% which is more than viable.

The Bitcoin dataset was scored and used to create a linear regression model despite the violation of two assumptions. For this study, backwards selection was used along with the Schwarz-Bayesian criterion to determine which independent variables are to remain in the linear regression model. The model used the best average squared error for the validation data. Bitcoin was used to create the initial data and Bitcoin Cash was used to validate the linear model. Based

on backward selection, the only variable removed was the date variable. The linear regression model that was selected contained the variables: intercept, open, high, Volume_Crypto and Volume_USD. According to the Analysis of Variance table, the model had an R-Square value of 0.9992, which means the model explains 99.92% of the variation in the response variable around its mean.

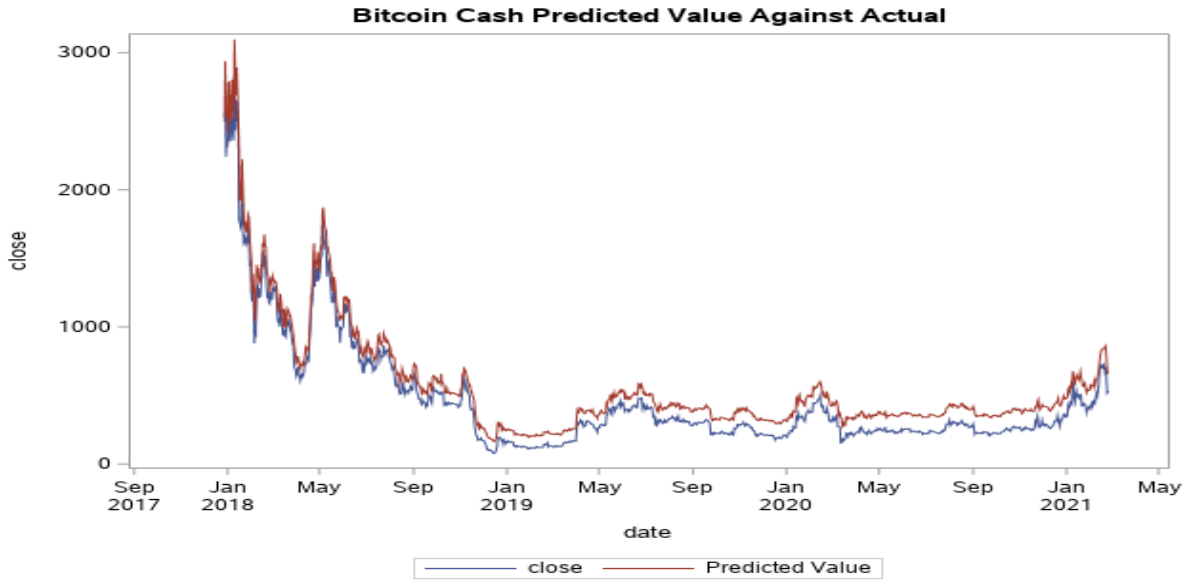

Figure 9 – Bitcoin Cash plot of predicted data along with the actual data.

The linear regression model based on the Bitcoin dataset was used to score the Bitcoin Cash dataset. Predicted values were plotted along with the actual values of the Bitcoin Cash dataset and the linear regression model does a good job of closely predicting the true values. The minimum percent error difference between the predicted value and actual value was 1.4%, where the maximum value of 129.0%. The overall average percent error difference was 36.1%.

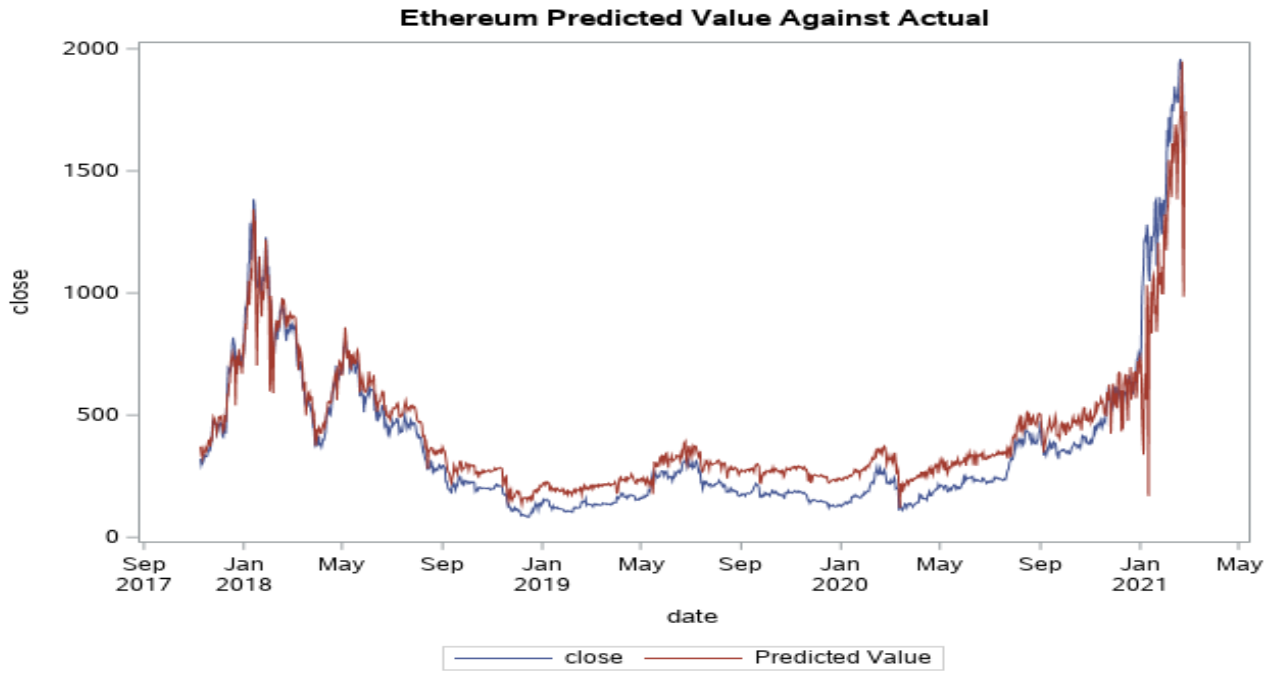

Figure 10 – Ethereum data plot of predicted data along with the actual data.

The Ethereum data was scored with the linear regression model created by the Bitcoin dataset. Similar to the scored Bitcoin Cash dataset, the predicted values are close to the actual close values. The predicted values overestimate the response variable from May 2018 to around January 2021. According to the data, the maximum percent error between the close and predicted values was 96.7%, occurring on 12/16/18. The minimum percent error was less than 1% occurring on 01/20/18. The overall average percent error across the whole dataset was 32.73%.

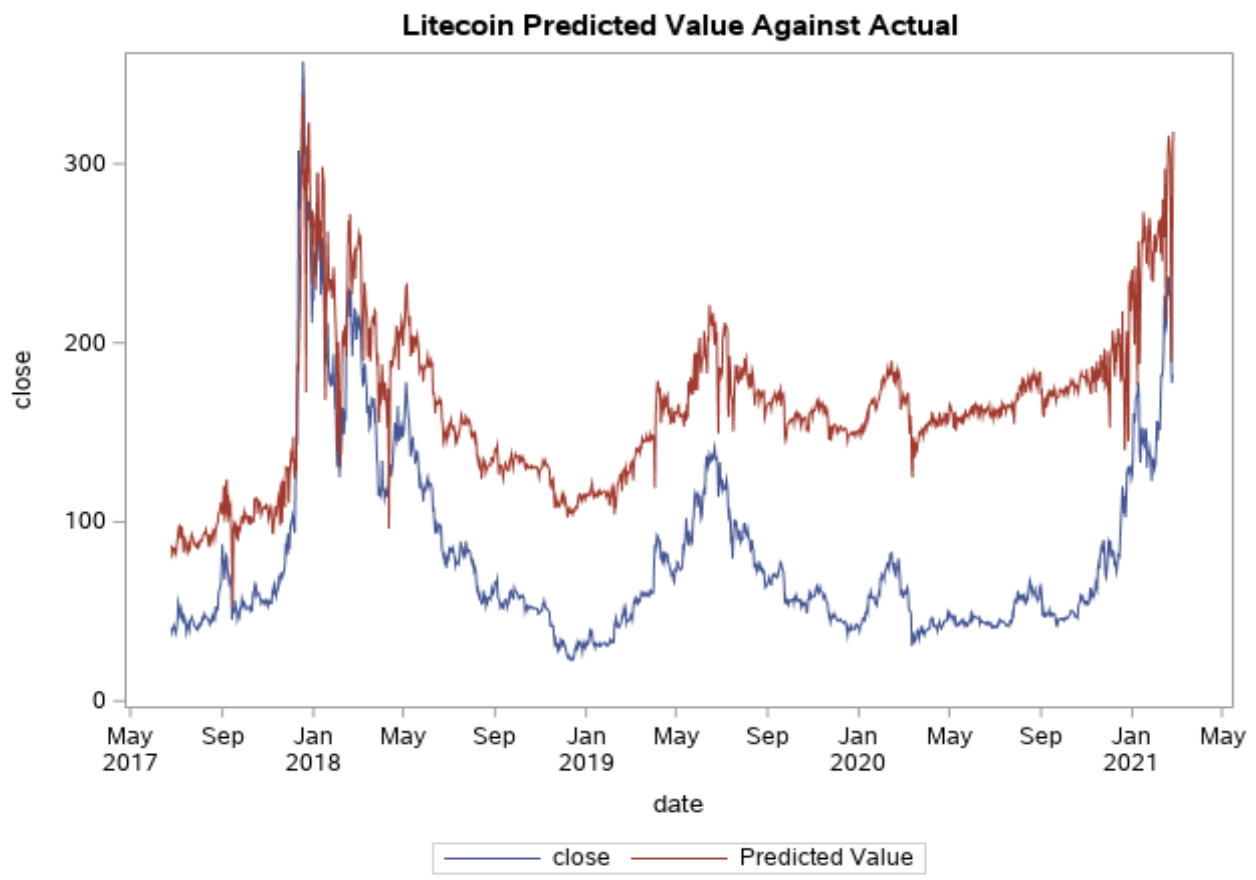

Figure 11 – Litecoin Plot of the predicted values along with the actual values. Note the more notable difference between the close and predicted value.

The scored Litecoin data was not as accurate as the previous Ethereum and Bitcoin Cash data. One contributing factor is the range of the y-axis. The previous scored data had close values reaching up to the thousands. Litecoin's maximum close value was $357.12. The maximum percent error difference in the scored data was 360.5% and the minimum percent error difference was 0.8%. The overall average percent error difference was 164.1%. A linear regression model based on Bitcoin would do a poor job of accurately predicting the price of Litecoin. Although the percent error differences are large, the linear regression model does a decent job of predicting the

trend of Litecoin's volatility. Both the close and Predicted Value graphs show similar dips and spikes when compared side by side.

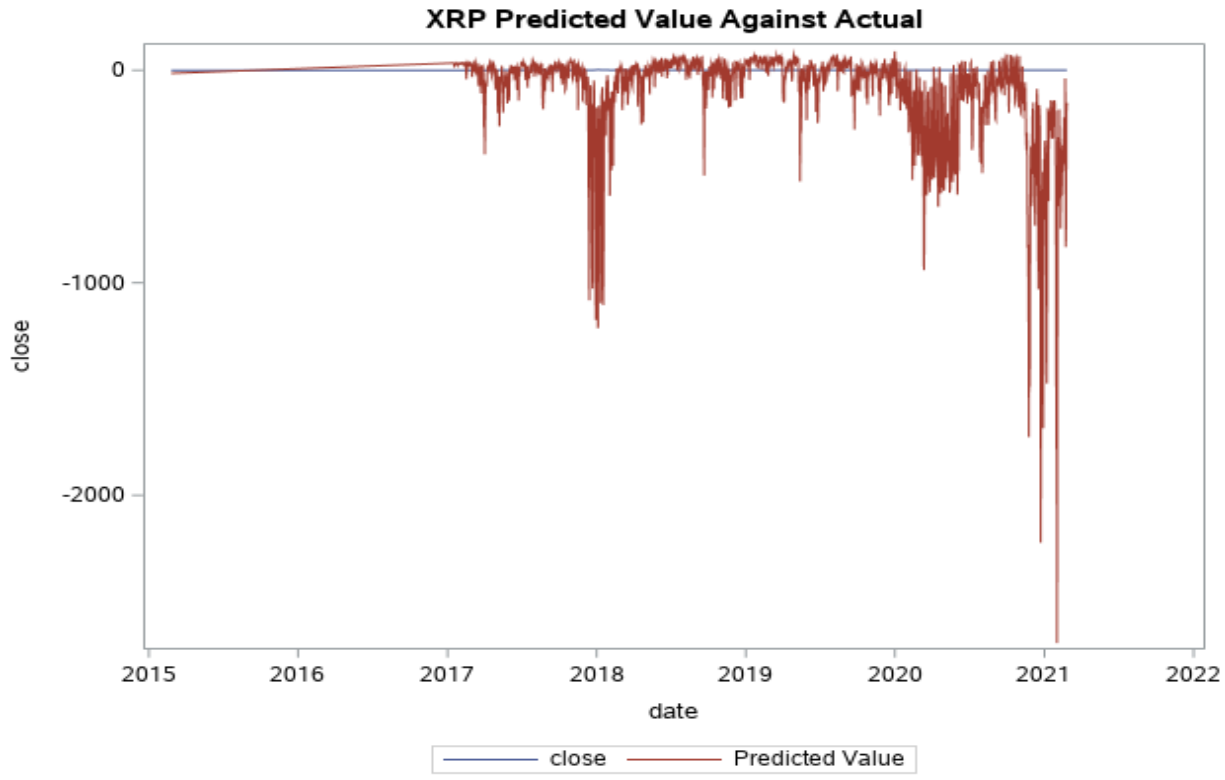

Figure 12 – Ripple, more commonly known as XRP, scored data plot of predicted values and actual values.

The scored data of XRP immediately points out the flaw of this linear regression model. XRP is a relatively newer cryptocurrency with price ranges of less than a dollar to a maximum of $2.8. Even more evident in a cryptocurrency of less worth, the linear regression model based on Bitcoin does a poor job of predicting the actual close price.

Overall, the linear regression model based on Bitcoin was created in SAS using the backward selection with the Schwarz-Bayesian Criterion to determine which independent variables stayed in the model. Average square error was used to validate the data on the Bitcoin

Cash dataset. Altogether, the model removed the date variable and had an R-square value of 99.92%.

Data Summary and Implications

The study was conducted to determine whether or not a model can be created to determine cryptocurrency prices in the future. The hypothesis questioned if a predictive regression model can be created from historical cryptocurrency datasets. This study found that there can be a predictive linear regression model that can be created from historical datasets. However, the accuracy of the model depends on whether the data to be scored is in the same value range of the data the model was based on.

Although the model worked for the Bitcoin Cash and Ethereum datasets, the model failed with the Litecoin and XRP datasets. The original Bitcoin dataset had values of upwards to $50,000, whereas the other datasets were nowhere near that value.

The variables used to create the linear regression model were open, high, Volume_Crypto and Volume_USD. These variables did a explain approximately 99.92% of the variability in the data. The study did end up modeling a predictive regression model that would work well given the right datasets.

The study and the following results do have limitations. This study focuses on a relatively small number of cryptocurrencies when there exist over 1,500 traded cryptocurrencies. Another limitation that resulted from the study, suggests that an entirely different model can be created using a different dataset for the original predictive model. One other limitation to the study, is that the variables involved were all internal variables that corresponded to each individual dataset. There are other outside factors that may contribute to the prices of cryptocurrencies such

as, social media, potential big investors, and how other markets affect the price. Overall, the predictive model that was created may be useful as a standard to more in-depth studies.

The original null hypothesis stated that a model based on predictive regression cannot be created from historical cryptocurrency datasets. This study showed that the null hypothesis was false, as a model was created based on the Bitcoin dataset and validated and tested on the remaining datasets. Overall, future cryptocurrency prices can be determined with a linear regression model.

The results of this study would benefit potential investors in deciding whether or not to buy certain cryptocurrencies. A course of action that would be advised, there should be an increase in purchase volume for cryptocurrencies that are modeled to increase in price. This study cannot suggest which cryptocurrencies to potentially invest in, but rather creates a template model for which can be replicated for other studies.

The basis of this study was formed around a relatively small number of cryptocurrencies. Future research should increase the number of cryptocurrencies studied and investigate other factors that may contribute to the changing prices of cryptocurrencies. One example would be to investigate how Bitcoin prices affect other cryptocurrency prices. Bitcoin is the leading cryptocurrency in the market and has influence on other cryptocurrencies. Another research direction that may be taken, is to create a time series data on cryptocurrencies and determine whether cryptocurrencies, in general, increase in price overtime. The financial market has been thoroughly analyzed with technical and quantitative analysis to predict price trends (Norris, 2020). With the ever-increasing amount of data that is being generated, cryptocurrencies may be researched as well as the financial market one day. In summary, cryptocurrencies are relatively new to the market and have a wide range of potential studies available.

Sources

Anurag. (2020, September 17). SAS review - what is IT, pros, cons, and suitability. Retrieved February 23, 2021, from https://www.newgenapps.com/blog/sas-review-what-is-it-pros-cons-suitability/

Brittain, J., Cendon, M., Nizzi, J., & Pleis, J. (2018). Data Scientist's Analysis Toolbox: Comparison of Python, R, and SAS Performance. SMU Data Science Review, 1(2), 7th ser., 1-20. doi:https://scholar.smu.edu/cgi/viewcontent.cgi?article=1021&context=datasciencereview

DeVries, P. D. (2016, September). An Analysis of Cryptocurrency, Bitcoin, and the Future. Retrieved February 20, 2021, from https://ijbmcnet.com/images/Vol1No2/1.pdf

Farell, R. (2015, May). An Analysis of the Cryptocurrency Industry. Retrieved February 20, 2021, from https://archive.org/details/AnAnalysisOfTheCryptocurrencyIndustry

Hong, S. (1994, March). A Method for Analyzing and Reducing Data Redundancy in Object-Oriented Databases. Retrieved February 22, 2021, from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.17.4191&rep=rep1&type=pdf

Lee, W. (2018, August 07). Bitcoin linear regression: Correlation exploration - by Brian McMahon. Retrieved February 25, 2021, from https://medium.com/hashreader/bitcoin-linear-regression-correlation-exploration-f16c0b22afbe

Norris, E. (2020, August 29). The Linear Regression of Time and Price. Retrieved February 22, 2021, from https://www.investopedia.com/articles/trading/09/linear-regression-time-price.asp