MSDS 6372 - Project 2

Jostein Barry-Straume
Laura Ludwig
David Tran

11/1/2017

Time Series Analysis of Bitcoin

MSDS 6372 - Section 403

Project 2

Data Science @ Southern Methodist University

Figure 1: Source: bitcoin.com

Table of Contents

- Team Members
- Introduction
- Problem Statement
- Constraints and Limitations
- Data Set Description
- Exploratory Data Analysis
- Variable Screening
- Model Selection
- Serial Correlation
- Conclusion
- Appendix

Team Members

- Jostein Barry-Straume
- Laura Ludwig
- David Tran

Introduction

Cryptocurrency is a digit currency and acts as a medium for exchanges/transactions. Cryptocurrencies are decentralized, which means it is not processed by any banking system and goes straight to the consumers. The transactions are posted on an online ledger for transparency. Users' identities are protected through an encryption key, which is a feature that Bitcoin has. Bitcoin is one of the popular choices of cryptocurrency. Since its introduction into the market in 2009, it has drastically increased and decreased in value. The analysis below will offer insights on the characteristics of the cryptocurrency and its projected value and trend.

Problem Statement

Develop a time series model based on an observed set of explanatory variables that can be utilized to predict future price of Bitcoin.

Constraints and Limitations

Add text here

Data Set Description

Add data set description here

Variable	Variable Type	Summary
Date		
Open		
High		
Low		
Close		
Volume		
Market Cap		
Time		

Snapshot of the data set

```
## 'data.frame':    1620 obs. of  8 variables:
##  $ Date      : Factor w/ 1620 levels "Apr 01, 2014",..: 109 114 119 1069 1074 1079 1084 1089 1094 109
##  $ Open      : num  135 134 144 139 116 ...
##  $ High      : num  136 147 147 140 126 ...
##  $ Low       : num  132.1 134 134.1 107.7 92.3 ...
##  $ Close     : num  134 145 139 117 105 ...
##  $ Volume    : Factor w/ 1378 levels "-","1,002,120,000",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Market.Cap: Factor w/ 1616 levels "1,000,070,000",..: 130 125 158 142 75 37 16 64 74 62 ...
##  $ Time      : Date, format: "2013-04-28" "2013-04-29" ...
```

```
## [1] 1620    8
```

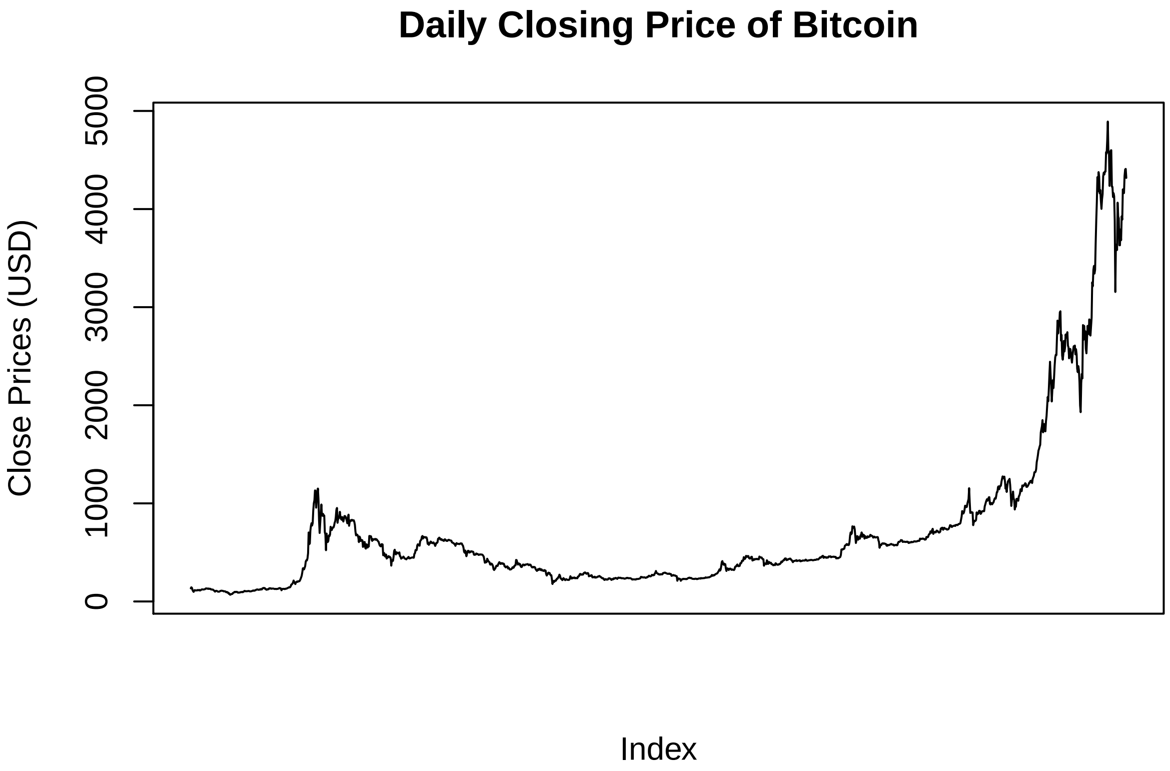
```
##              Date   Open   High    Low  Close Volume    Market.Cap
## 1620 Apr 28, 2013 135.30 135.98 132.10 134.21      - 1,500,520,000
## 1619 Apr 29, 2013 134.44 147.49 134.00 144.54      - 1,491,160,000
## 1618 Apr 30, 2013 144.00 146.93 134.05 139.00      - 1,597,780,000
## 1617 May 01, 2013 139.00 139.89 107.72 116.99      - 1,542,820,000
## 1616 May 02, 2013 116.38 125.60  92.28 105.21      - 1,292,190,000
## 1615 May 03, 2013 106.25 108.13  79.10  97.75      - 1,180,070,000
##            Time
## 1620 2013-04-28
## 1619 2013-04-29
## 1618 2013-04-30
## 1617 2013-05-01
## 1616 2013-05-02
## 1615 2013-05-03
```

The above output shows the structure, dimension, and head of the data set. There are 1,630 observations with 8 explanatory variables.

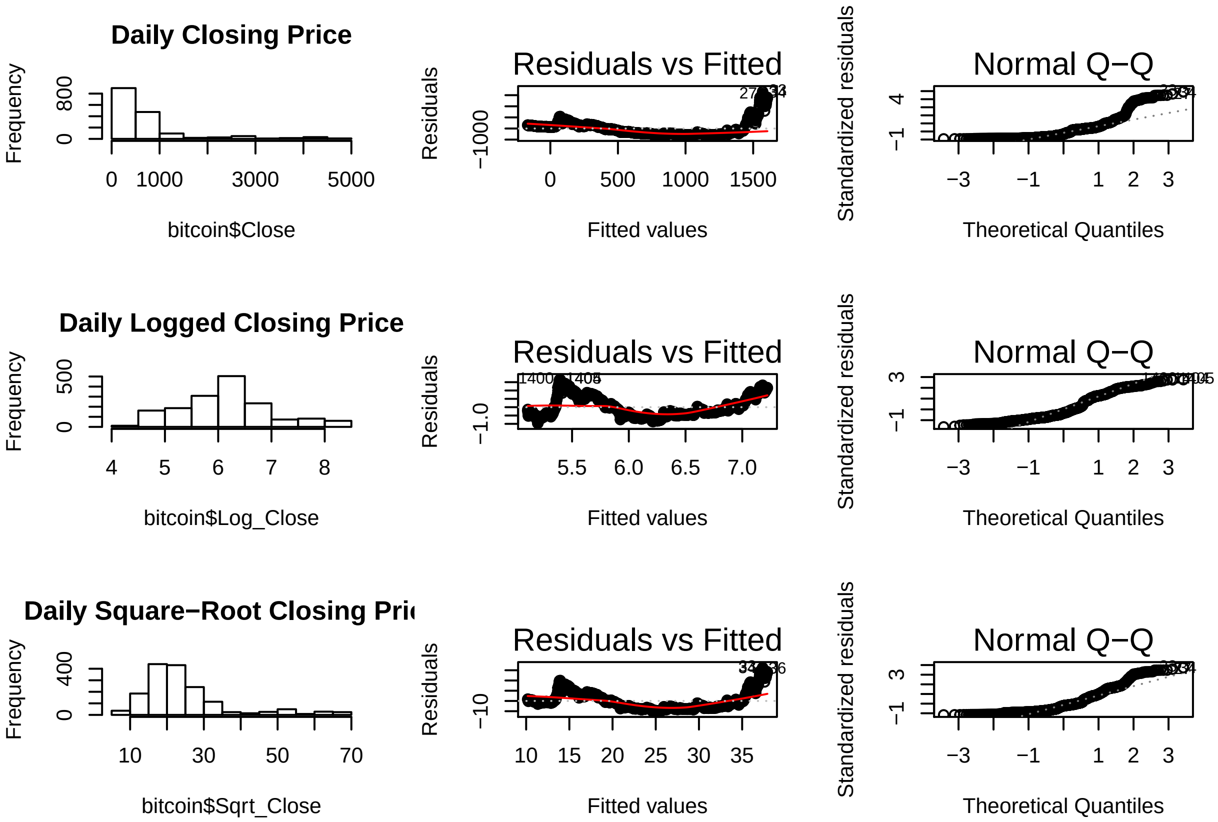
Summary statistics of daily closing price of bitcoin:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   68.43  261.41  448.19  718.80  705.28 4892.01
```

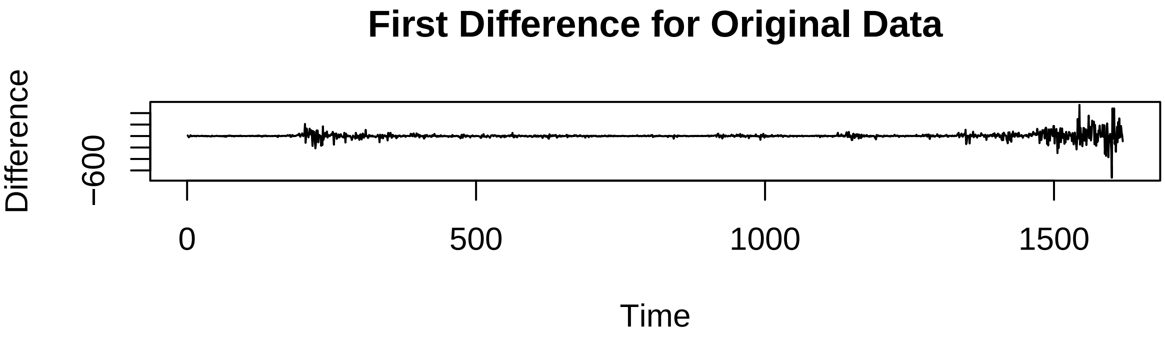
Exploratory Data Analysis



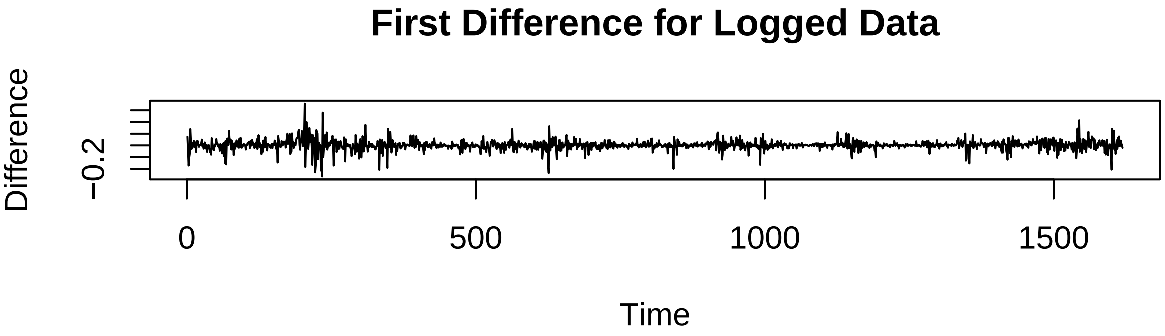
The above plot reflects the daily closing price of bitcoin from to April 28th, 2013 to October 3rd, 2017. Although there appears to be no pattern in the change of the closing price, a general increase in price over time is apparent. Increasing variance over time necessitates transformation of the original data.



The above diagnostic plots confirm the need for transformation, as well as give insight into which transformation is most appropriate. The histograms of both the original data and logged data are heavily right skewed, with the former to a larger degree. Additionally, the Q-Q plots for the original data and logged data venture far of the path of diagonal line. In contrast, the logged data displays a normal distribution for its histogram, as well as a fairly good Q-Q plot. The tail ends of the logged Q-Q plot indicate some skewness at both ends, which the corresponding histogram supports. However, the size of our data set should ease any concern we might have. The residual diagnostic plot of the logged data reflects non-constant variance. This will be addressed by taking the first degree difference of the logged daily closing price.



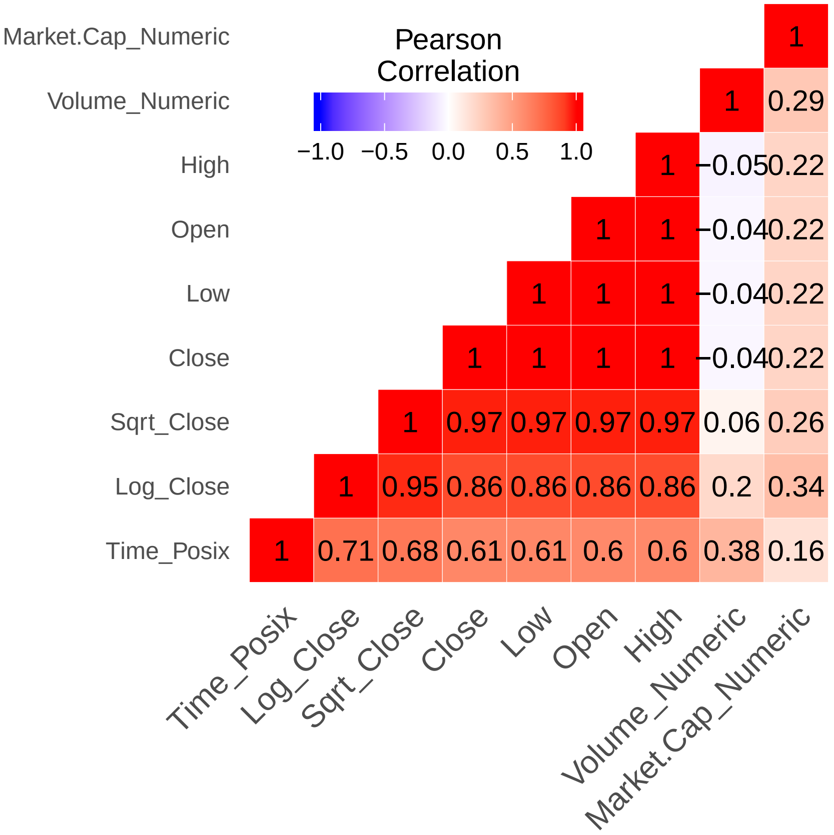
First Difference for Original Data



First Difference for Logged Data

The variance of the first difference between the original and logged data are vastly different. In the original data the increasing variance as time goes one is visually clear, whereas the variance of the logged data is reasonably constant with no apparent patterns.

Variable Screening



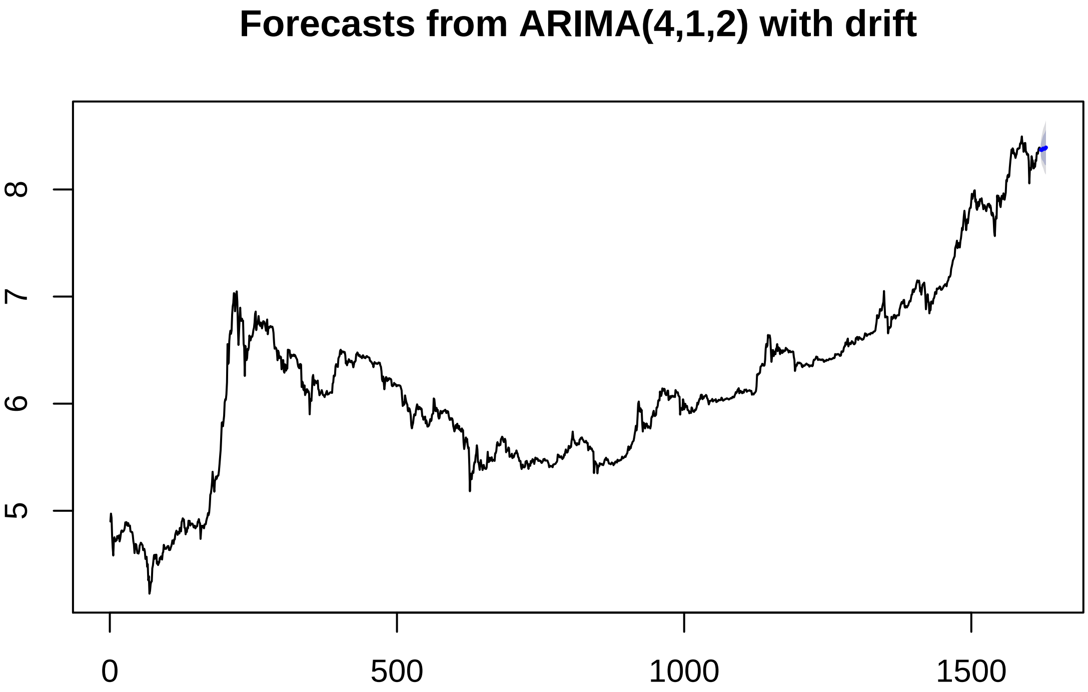
The above heat map correlation matrix offers limited new comprehension of the bitcoin data set, but is still helpful nonetheless. Volume of daily bitcoin trades has a weak positive correlation (R = 0.20) with logged closing price, and a moderate positive correlation (R = 0.38) with time. This suggests that as time goes on, the volume of trades increases and might have a impact on the closing price of bitcoin. Of note, the total market cap of mined bitcoins has a moderate positive correlation with logged closing prices. In other words, the total value of mined bitcoins possibly influences the closing price.

The original and logged closing prices have strong positive correlations with time (R = 0.61, and R = 0.71 respectively). This suggests the need to address auto correlation issues. Moreover, the following variables are 100% colinear with each other: High, Low, Open, and Close. This makes sense as all of the said variables pertain to the price of bitcoin. To reduce redundancy, only the closing price of bitcoin will be utilized for a time series model.

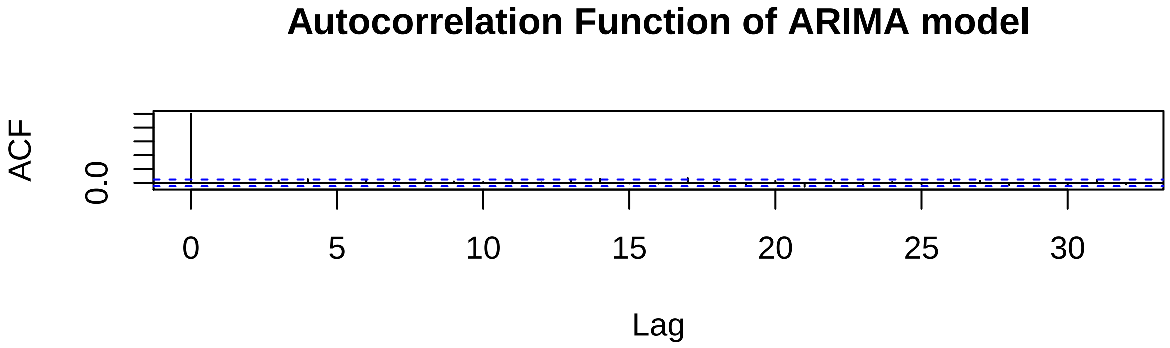
Model Selection

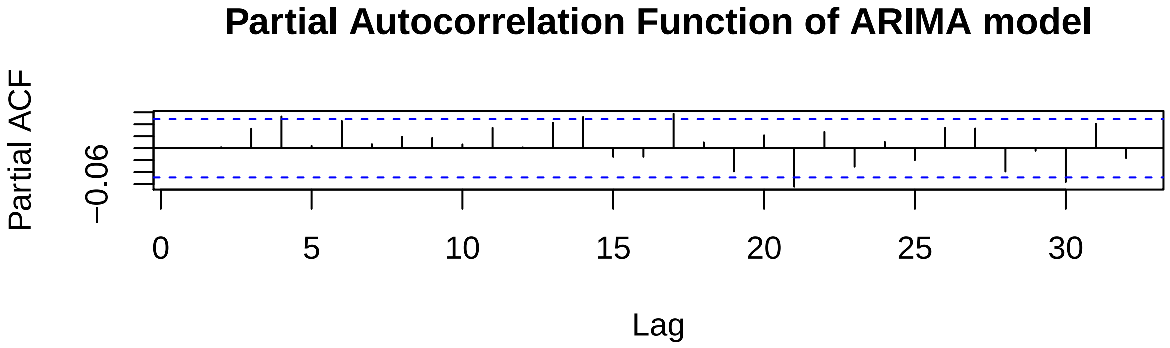
Analysis of the daily closing price of bitcoin can now be carried out with the following model:

```
## Series: bitcoin$Log_Close
## ARIMA(4,1,2) with drift
##
## Coefficients:
##          ar1      ar2      ar3      ar4      ma1     ma2   drift
##       0.6941  -0.9619  -0.0242  -0.0091  -0.7099  0.9456  0.0022
## s.e.  0.0431   0.0362   0.0309   0.0293   0.0352  0.0198  0.0010
##
## sigma^2 estimated as 0.001821:  log likelihood=2812.69
## AIC=-5609.39   AICc=-5609.3   BIC=-5566.27
##
## Training set error measures:
##                         ME       RMSE        MAE          MPE     MAPE
## Training set -1.133225e-05 0.04256813 0.02627099 -0.005240175 0.431615
##                  MASE         ACF1
## Training set 1.000085 0.0002758099
```



Serial Correlation





```
##
##  Box-Pierce test
##
## data:  residuals(arima_fit)
## X-squared = 0.00012324, df = 1, p-value = 0.9911
```

Add serial correlation here

Conclusion

Possible take on project, combine close prices with google trends.

Appendix

Add R code and pertinent graphs here