Predicting Bitcoin Using Twitter Tweets

Dennis Xing

August 16, 2015

1. Introduction

1.1 Bitcoin

Bitcoin is a decentralized cryptocurrency and payment system that was created in 2009 by an unknown person by the name of Satoshi Nakamoto. As a result of being decentralized and unregulated, there are benefits such as frictionless transactions –no fees and anonymity, but also drawbacks such as no buyer protection and potential risk of unknown technical flaws. Regardless of these drawbacks, the novelty and uniqueness of Bitcoin's payment protocol, media coverage, and growing adoption has the Bitcoin ecosystem gaining lots of attention from businesses, consumers, speculators, and investors. As a currency, Bitcoin can be purchased through exchanges or can be mined by computing cryptographic puzzles.

1.2 Price Prediction

There is a clear analogy between the Bitcoin market and the modern stock market. The exchanges where stocks and equities are traded such as the New York Stock Exchange, NASDAQ, and London Stock Exchange have clear analogies in the Bitcoin world such as Mt. Gox, OKCoin, etc. High frequency trading paired with machine learning algorithms drove the developments to maximize monetary profits and financial rewards. Thus it makes sense that the tools used in the Wall Street world can be replicated in the world of Bitcoin. The black boxes, algorithms that the user does not see nor need to know, that these hedge funds use to determine optimal trading practices also take into account Twitter feed. For example, when there was a fake White House bombing posted on Twitter by AP, the entire Dow tanked. In an analogous manner, I am using twitter data to predict the future price of Bitcoin.

1.3 Prior Work

Devavrat Shah and Kang Zhang from MIT EECS published a paper titled "Bayesian Regression and Bitcoin", which described their application of Bayesian regression to Bitcoin price prediction, achieving high profitability. Other research in the field including projects from Stanford University's Machine Learning Class, CS 229, also used traditional features such as Bitcoin market capitalization and Bitcoin mining speed. I sought to explore greater explanation of the price of Bitcoin through another medium, Twitter tweets.

2. Data and Observations

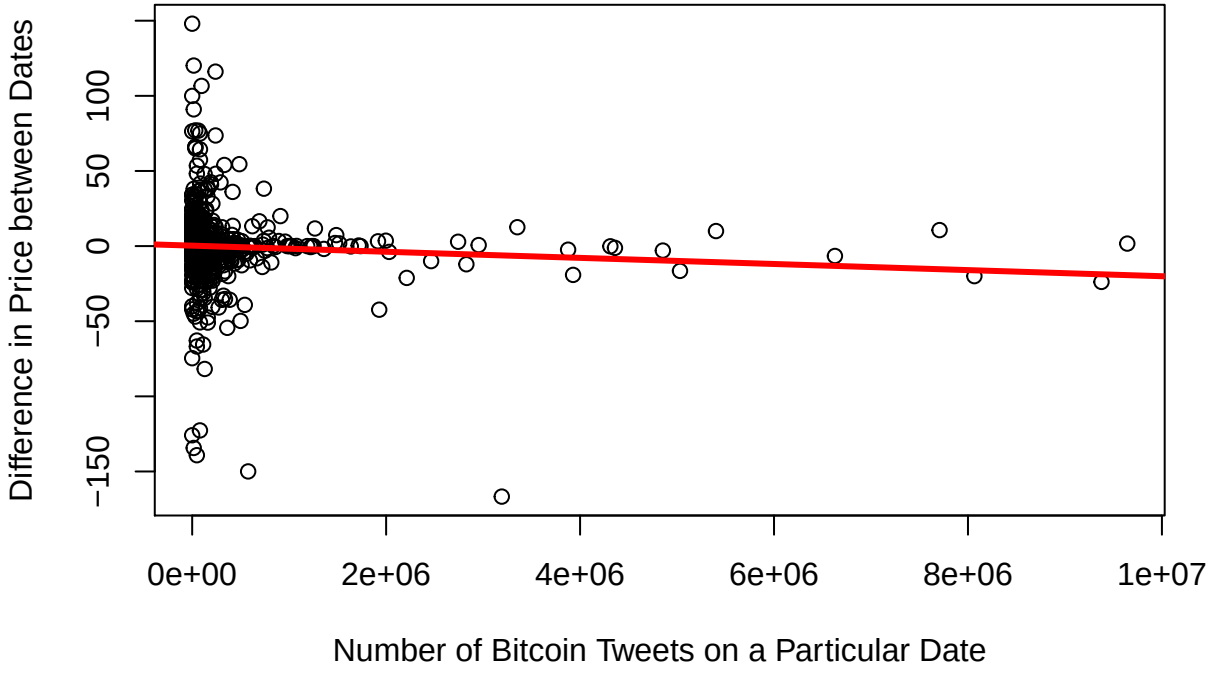
The data for my analysis came from public available datasets online. The historical price of Bitoin came from coindesk.com, a website dedicated to real-time Bitcoin price charts. The twitter aspect of my data came from an individual who scraped Twitter tweets relating to Bitcoin already. The dataset for tweets consisted of 18136 entries, however, there were many retweets by Twitter Bots. The feature I selected was the number of Bitcoin Tweets on a particular date. Instead of predicting the actual price of Bitcoin, I predicted the change in the price of Bitcoin between the dates of tweets. I hypothesized that the more Bitcoin Tweets on a particular day there are, the more positive and greater the price difference and vice versa.

3. Results

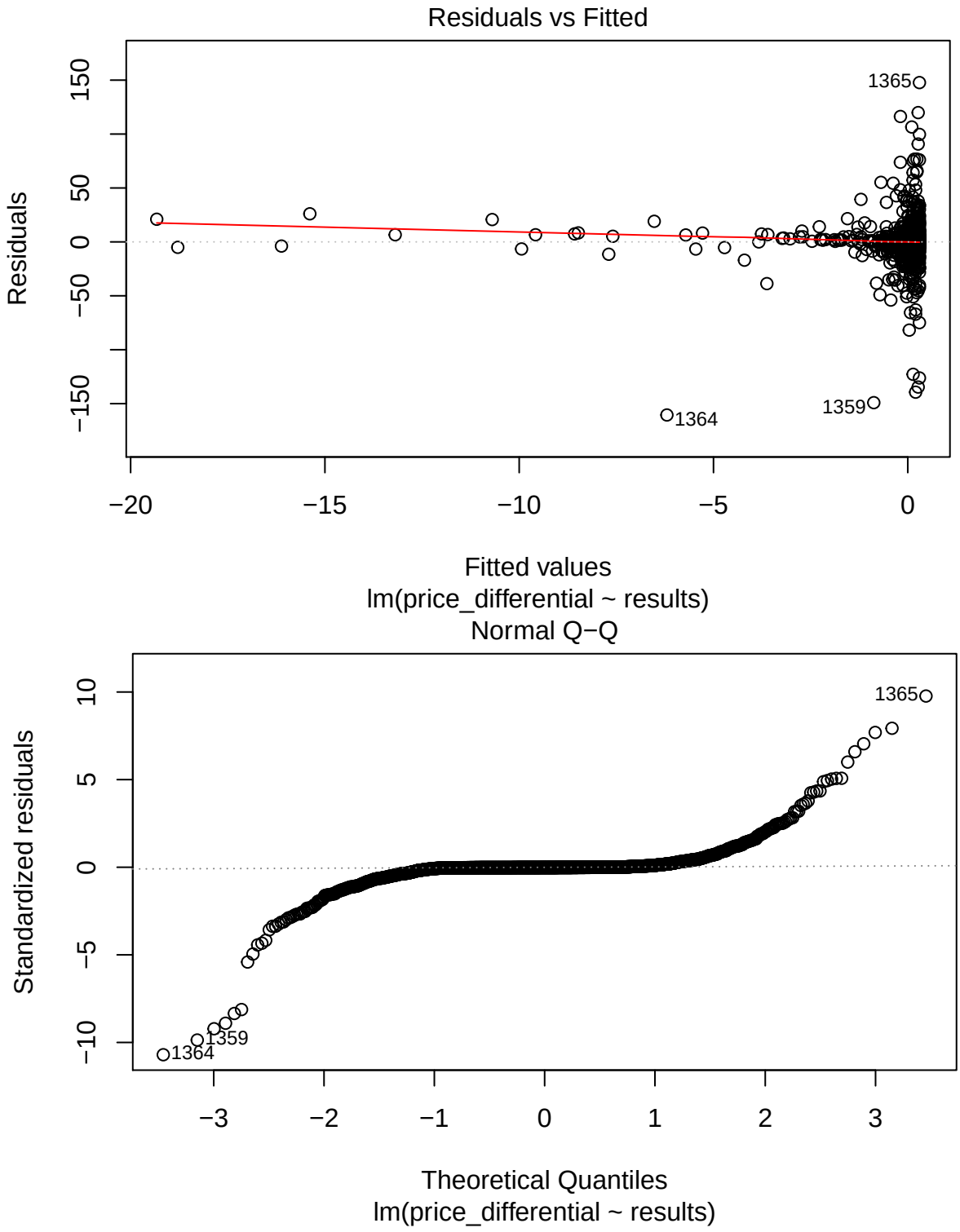
The statistical method that I used to predict the price of Bitcoin was linear regression. This was chosen over all the other complex methods that we studied in the course because I did not have the flexibility of mutliple predictors. Using Occam's razor as a principle for creating my models of the data, I chose linear regression not only because it is among the simplest, but also because of the interpretability. After generating the plot of my data and fitting the regression line through the plot, I noticed that this was strictly the opposite of what my initial inclinations suggested. The regression line had a slight negative slope with the more Twitter

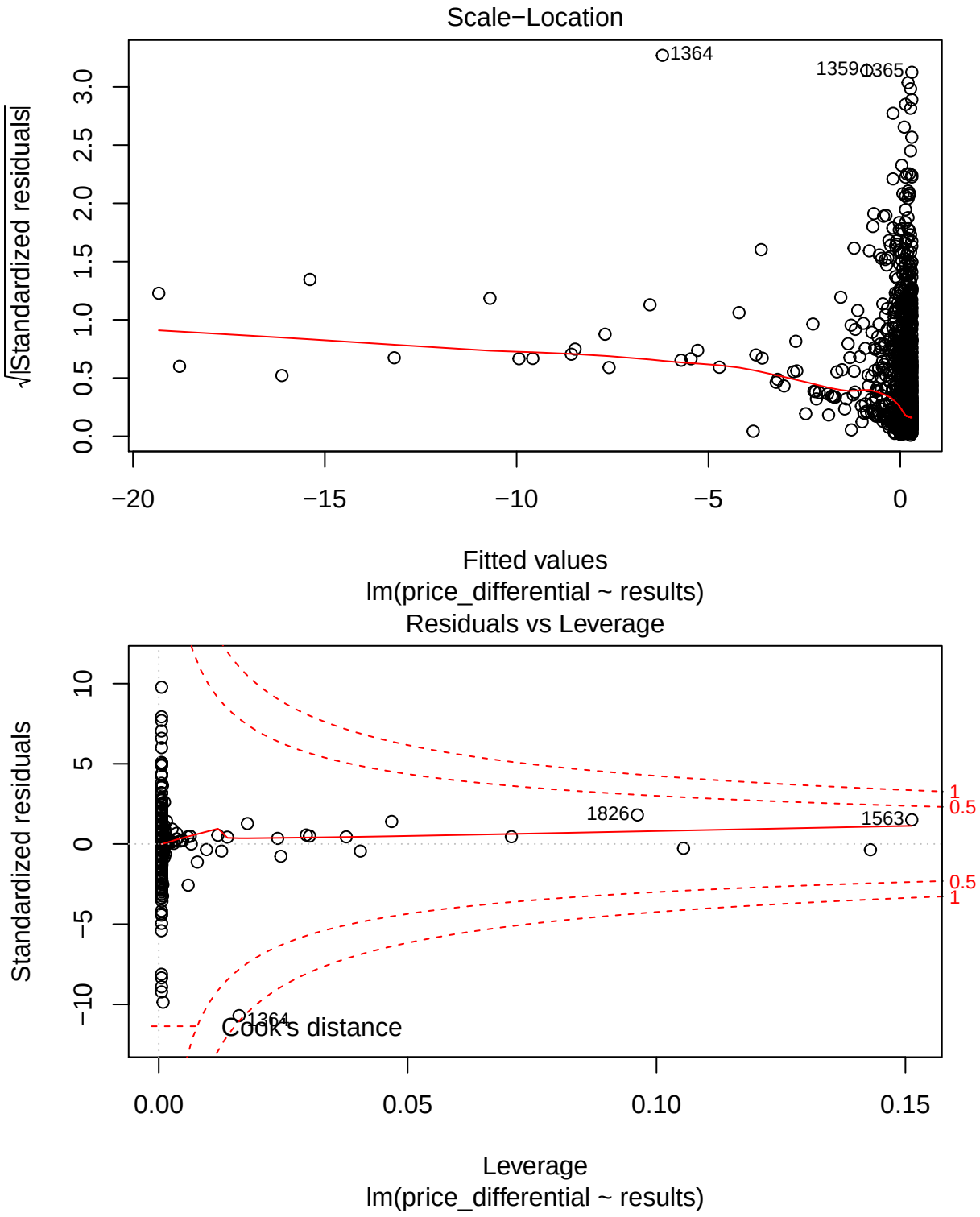
tweets there were actually having a slight decrease in the change in price. Although the price of Bitcoin is largely based on speculation and "hype", I still thought that it could be caught in the form of Twitter tweets.

I am still confident in the fact that Bitcoin could be predicted to an extent with Twitter data; however, in this project I think the issue that played a factor in the failure of predicting price is the collection of Twitter Tweets. When I was playing with the data and looking at tweets and dates to get a feel with what I would be working with, I noticed that some tweets had dates with year 2006 ! This raised a red flag with me, because Bitcoin was released in 2009 and there is no one a tweet with Bitcoin in it could be mentioned pre-2009. Even though this result appeared to hinder my research, I made sure to only use data that was factually valid. By intersecting the dates from the price data collected from Bitcoincharts with the dates from the Twitter Data, I was able to get a accurate historical timeline of Bitcoin prices. Still this was not enough as there were many sparse Bitcoin tweets on a particular date and also a very minimal change in price between dates. This caused a huge black clump around around the origin.




```
##
## Call:
## lm(formula = price_differential ~ results)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -160.512   -0.350   -0.243    0.149  147.709
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.007e-01  3.631e-01   0.828    0.408
## results     -2.036e-06  6.176e-07  -3.296    0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.12 on 1831 degrees of freedom
## Multiple R-squared:  0.005898,   Adjusted R-squared:  0.005355
## F-statistic: 10.86 on 1 and 1831 DF,  p-value: 0.0009999
```


Residuals vs Fitted
Residuals
150
50
0
−50
−150
1365
1364
1359
−20
−15
−10
−5
0
Fitted values
lm(price_differential ~ results)
Normal Q−Q
Standardized residuals
10
5
0
−5
−10
1365
1359
1364
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(price_differential ~ results)



4. Conclusion

Predicting the price of the stock market is hard. Predicting the price of Bitcoin is just as hard. I regret to inform you that I will not be making money off the Bitcoin market anytime soon. There was too much noise in the tweet predictor and in fact it was the opposite of what my hypothesis showed. The more Bitcoin related Tweets on a particular day should have an increase in the difference between Bitcoin prices. Instead, the opposite happened. There was never a clear signal that I could have captured. Whether this was because

the dataset of tweets that I gathered could have been flawed or because there is no signal to begin with is unknown. But the modelling of prices is much more complicated than just a simple predictor and for most black boxes at hedge funds, time series and Brownian motion are just the prerequisites to a good winning forumla.

4.1 Future Work

To improve on the possiblity of making money and also the predictability of Bitcoin, I would generate my own data so that it is more representative and clean. My feature space in this project was extremely small. I will have to enlarge my feature space and take in many more predictors. After doing cross-validation and making increasing gains and test data, it would be time to apply this to the real world and start cashing in.