BITCOIN REGRESSION ANALYSIS

Marija Stanojcic

STAT 6509 Theory and Application of Regression

June 2018

Marija Stanojcic
NetID: gk3964

TABLE OF CONTENTS

WHAT IS BITCOIN

Bitcoin is one of the new cryptocurrencies. It was created in 2009 by an unknown person who used the alias Satoshi Nakamoto. Bitcoin is only a digital token, it does not have physical backing, and it can be sent electronically from one user to another, anywhere in the world. The bitcoin network is not like traditional payment methods, it is run by decentralized network of computers around the world that keep track of all bitcoin transactions. The record of all bitcoin transactions, that these computers are constantly updating, is known as the blockchain. (Popper, 2017) The system is controlled entirely by a software, which will release a total of twenty-one million bitcoins, almost all of them over the next twenty years. (Davis, 2011) You can use bitcoin exchanges like Coinbase, Bitstamp and Bitifinex to buy or sell bitcoins using different currencies. Bitcoins can be send using mobile apps or computers.

What makes bitcoin popular? A public record of every coin's movement is published across entire network, while buyers and sellers remain anonymous. (Davis, 2011) This of course, makes it hard to steal bitcoins, and also there is no track of personal transactions. International payments are easy, cheap and fast, because bitcoin is not tied to any country or subject to regulation, and it can be moved in minutes, while some international money transfers can take weeks.

Bitcoin mining – it is the process of creating new bitcoins. People compete to "mine" bitcoins using computers to solve complex math puzzles. Currently, the winner gets 12.5 bitcoins, and generally there is a new winner every ten minutes. (Yellin, Aratari, and Pagliery, n.d.)

Bitcoin price – The price of bitcoin fluctuates constantly, and it is determined by open-market bidding on bitcoin exchanges. (Popper, 2017) Currently bitcoin market price is around 7500 USD.

Marija Stanojcic
NetID: gk3964

After learning about bitcoins one of the interesting topics is how to explain the bitcoin market price. The goal of this project is to fit a Linear Regression model that can explain fluctuations in bitcoin market price. Before that, because this data is time series data, and regression model should be fitted, in further analysis *Date* will not be considered.

ABOUT DATA

Data is collected from https://blockchain.info/stats and it has values for every other day starting on January 3rd, 2009 and ending on May 16th, 2018. There are no any missing values.

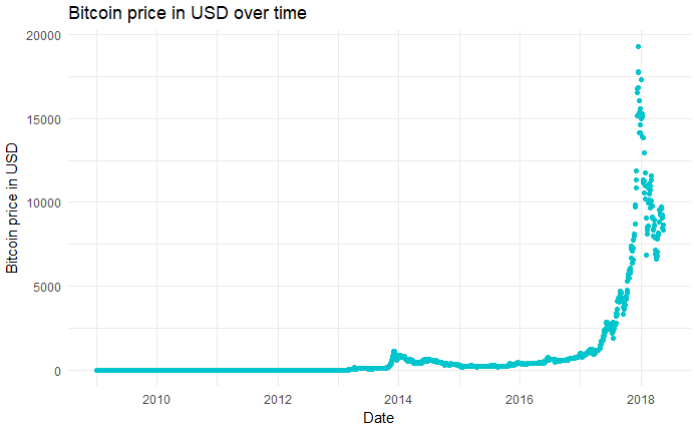
Date	Day	Week	Price	Transaction fees (BTC)	No of transactions	Output value (BTC)	Estimated Transaction Value	Miners Revenue	Cost per transaction	Difficulty	Hash rate	Trade volume
01-03-09	3	1	0	0	1	0	0	0	0	1	4.97E-08	0
01-05-09	5	2	0	0	0	0	0	0	0	0	0	0
01-07-09	7	2	0	0	0	0	0	0	0	0	0	0
01-09-09	9	2	0	0	14	0	0	0	0	1	6.96E-07	0
01-11-09	11	3	0	0	106	328.5714	0	0	0	1	5.27E-06	0
1/13/09	13	3	0	0	116	1743	0	0	0	1	5.72E-06	0
1/15/09	15	3	0	0	136	3468.429	0	0	0	1	6.31E-06	0
1/17/09	17	3	0	0	109	4977.143	0	0	0	1	5.37E-06	0

Table - *Preview of the first few rows*

Descriptions of variables: *Price* - average USD market price across major bitcoin exchanges; *Transaction Fees (BTC)* - total value of all transaction fees paid to miners; *No of transactions* - number of daily confirmed Bitcoin transactions; *Output value (BTC)* - The total value of all transaction outputs per day; *Estimated Transaction Value* - estimated transaction value in USD; *Miners Revenue* - total value of coin base block rewards and transaction fees paid to miners; *Cost per transaction* - miners revenue divided by the number of transactions; *Difficulty* - a relative measure of how much hashing power has been deployed by the network of miners; *Hash rate* -estimated number of tera hashes per second (trillions of hashes per second) the Bitcoin network is performing; *Trade volume* - total USD value of trading volume on major bitcoin exchanges. All features are quantitative.

Marija Stanojcic
NetID: gk3964

On the right is the plot of bitcoin market price in USD over time. From the beginning of bitcoin to 2014 people were not familiar with bitcoin, and the price of bitcoin was only in hundreds. From 2014 bitcoin price slightly started to change, then from 2016 there is an exponential growth of price. In mid-2017 bitcoin price jumps to a thousands. The highest average price of bitcoin in USD was in December of 2017. Next is the summary statistics of bitcoin price:



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	2.528	128.100	986.565	586.330	19289.785

. Detailed summary statistics for all variables is in the Appendix – Table 1.

ANALYSIS

Looking into the correlations between dependent variable (price) and other independent variables - all variables have a positive correlation. Features that have strong positive correlation (correlation ≥ 0.75) are estimated transaction value, miner's revenue, cost per transaction, difficulty, hash rate and trade volume. [1]

MODELS AND RESULTS

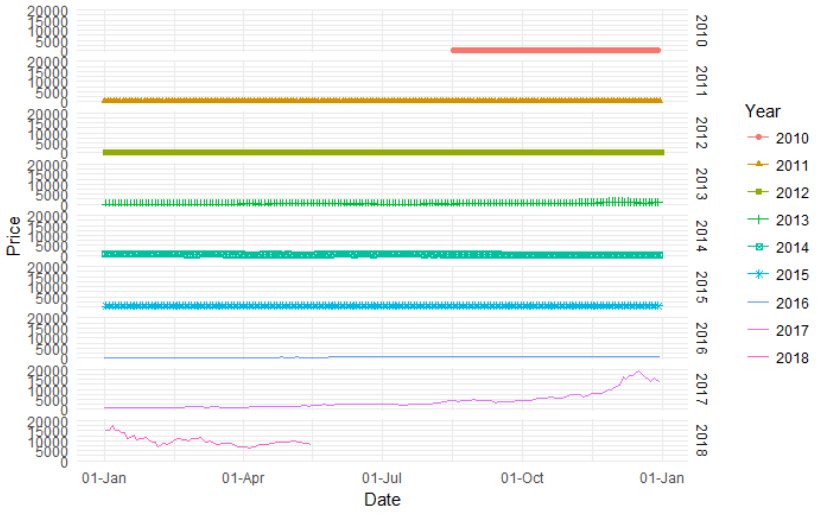
Starting with the simplest additive model with all features, all variables were significant, but regression assumptions of constant variance and normality of residuals were violated [2]. Log transformation only on the response variable, log transformation on predictors, as well, and also

[1] Look at the Plot 1 in Appendix
[2] Look at the Plot 2 in Appendix

applying stepwise regression to get the model with the lowest AIC value, did not help with heteroscedasticity. After looking at residuals vs predictors plot polynomial transformation on some predictors was done, still regression assumptions were violated. One more transformation that can help with inconstant variance is weighted regression. Weighted regression with weight = 1/(fitted values)^2 was applied, but it didn't help much. For time series data doing a percentage change transformation helps with making a data stationary. As this is a financial data, and features whose measurement unit is USD have high variance, percentage change transformation was applied on them, to achieve smaller variability. Using stepwise method and eliminating insignificant features led to the next model.

price (% change) = transaction fees + no of transactions + difficulty + hash rate + miners revenue (% change) + cost per transaction

This model showed by far the best residuals versus fitted plot [3], but it still didn't pass the regression assumptions. After all these transformations the next step was to investigate data more closely. On the right is the plot of bitcoin market price, per moths, for all years. Bitcoin price started to increase from mid-2017. In order to consider a good 24 months data with a seasonal variance subset of data from mid-2016 is taken for further analysis.



[3] Look at the Plot 3 in Appendix

FINAL MODEL

Continuing with the last model for full data, but now on the data starting from June 2016. The residual versus fitted plot looked better. Residuals versus predictors plots[4] indicated that adding polynomial features for transaction fees and miners is needed. The polynomial transformation with the second degree were done for both features. This model turned out to be the final model. The regression equation of the final model:

market price (% change) = - 21.89
- 29.68 transaction fees - 26.05 (transaction fees)^2
+ 0.000014 no of transactions
+ 0.098 cost per transaction
- 360.6 hash rate
+ 65.87 miners revenue (% change) - 25.76 (miners revenue (% change))^2
+ 1.4^{-11} difficulty

All predictors were significant. R^2 is 31%, and as response variable is now in percentage difference, R^2 indicates a good strength of the relationship between this model and market price in percentage change. Residuals plots looked good. [5]

Checking for regression assumptions - residuals vs fitted plot indicates that residuals do not have a pattern and one can assume constant variance. Breush - Pagan test confirmed assumption (p = 0.06815 is greater than 0.05). What about normality of residuals? Even though p value for Shapiro Wilkson test is small, W is close to 1 (W = 0.97965) and on the Q-Q plot [6] most of the points are on the normal line, so one can assume normality as well.

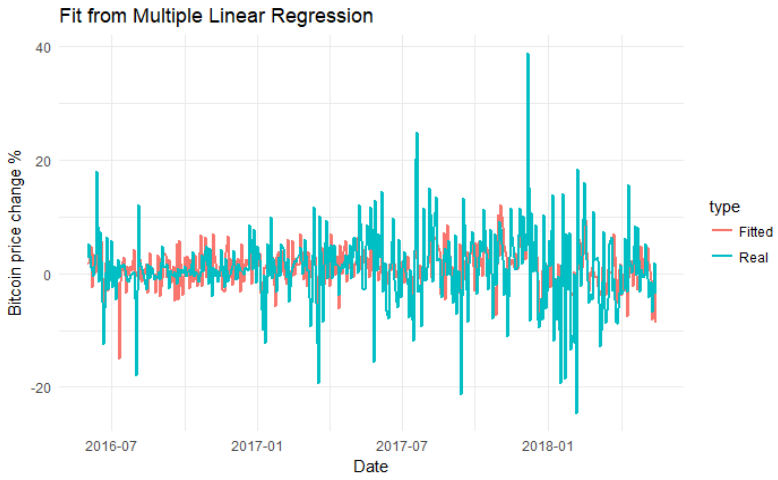
Furthermore, test for outliers[7] and influential points were done. There was only one point that was a potential outlier, and that point was the highest value for bitcoin price. As there are

[4] Look at the Plot 4 in Appendix
[5] Look at the Plot 5 in Appendix
[6] Look at the Plot 6 in Appendix
[7] Look at the Plot 7 in Appendix

different websites and a lot of information on bitcoin price, it can be found that even though that value is higher than others, it isn't a typo or mistake, it is an actual value of bitcoin price that day. As this is an actual value and the data is real time financial data it is better to keep it in the model.

CONCLUSION

On the graph is a real bitcoin price from 2016 (green line), and bitcoin price fitted with the model (red line). Red line is generally following the green line which indicates that the model is following the trend of the data, except for a few spikes. In a conclusion, the predictors that were used in the model - transaction fees, no of transactions, cost per transaction, hash rate, miner's revenue and difficulty explained a fluctuation in the bitcoin market price, starting from June 2016.



In this project, only a few features were considered for determining a bitcoin market price. The next step to make model better would be to add new factors, for example, a Google search index for bitcoin, news announcement, etc. Collecting additional data - daily or hourly data can also help. Exploring a time series analysis would take this project even further and may be able to forecast the future bitcoin price. Right now, some bitcoin analysis are predicting that bitcoin price will reach $25,000 by the end of 2018.

REFERENCES

[1] Popper, N. (2017, October 1). What is bitcoin, and how does it work? *New York Times*. Retrieved from https://www.nytimes.com/2017/10/01/technology/what-is-bitcoin-price.html

[2] Davis, J. (2011, October 10). The crypto-currency. Bitcoin and its mysterious inventor. *The New Yorker*. Retrieved from https://www.newyorker.com/magazine/2011/10/10/the-crypto-currency

[3] Yellin, T., Aratari, D., & Pagliery, J. (n.d.). What is bitcoin? *CNN Money*. Retrieved from http://money.cnn.com/infographic/technology/what-is-bitcoin/

[4] Bitcoin Stats (n.d.). Retrieved from https://blockchain.info/stats

[5] Robert N. (n.d.) What's a good value for R-squared? Retrieved from https://people.duke.edu/~rnau/rsquared.htm

[6] Jang, H., & Lee J. (2017, December 4). An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information. *IEEE Access*. Retrieved from https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8125674

Marija Stanojcic
NetID: gk3964

APENDIX

MODEL EVALUATION

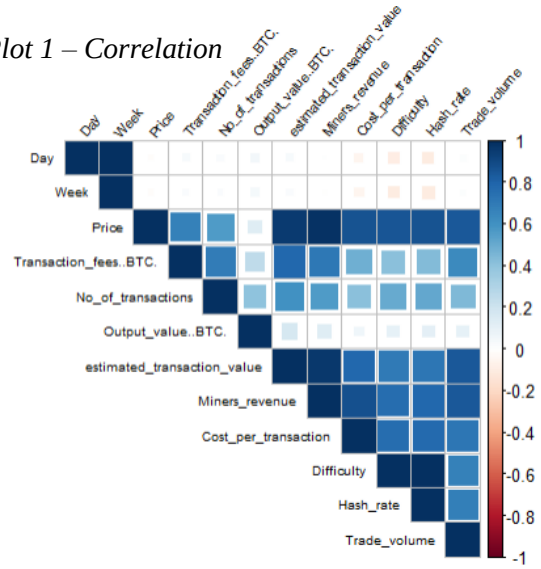
For evaluating the model, we randomly divided data into the training (80% of the data) and test set (remaining 20% of the data). We trained the model on the training set and evaluated on the test set. The mean absolute error between real values and predicted one was 4.74 in units of percentage change.

TABLES AND PLOTS

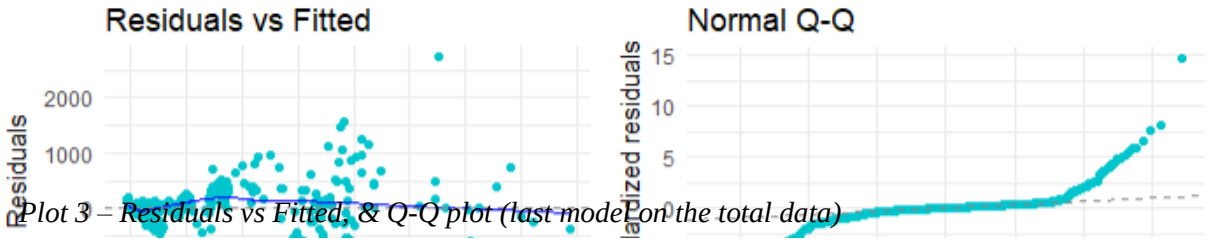
Table 1 – Summary Statistics

```
      Date                 Day              Week            Price          Transaction_fees..BTC. No_of_transactions Output_value..BTC.
 Min.   :2009-01-03   Min.   :  1.0   Min.   : 1.0   Min.   :    0.000   Min.   :   0.000       Min.   :     0     Min.   :       0
 1st Qu.:2011-05-08   1st Qu.: 86.0   1st Qu.:13.0   1st Qu.:    2.528   1st Qu.:   3.252       1st Qu.:  4380     1st Qu.:  280139
 Median :2013-09-09   Median :176.0   Median :26.0   Median :  128.100   Median :  16.172       Median : 55371     Median :  850113
 Mean   :2013-09-09   Mean   :178.7   Mean   :26.4   Mean   :  986.565   Mean   :  52.955       Mean   : 92739     Mean   : 1171801
 3rd Qu.:2016-01-12   3rd Qu.:271.0   3rd Qu.:40.0   3rd Qu.:  586.330   3rd Qu.:  44.526       3rd Qu.:176783     3rd Qu.: 1658677
 Max.   :2018-05-16   Max.   :366.0   Max.   :53.0   Max.   :19289.785   Max.   :1128.762       Max.   :425008     Max.   :21158969
 estimated_transaction_value Miners_revenue     Cost_per_transaction   Difficulty          Hash_rate         Trade_volume
 Min.   :0.000e+00           Min.   :       0   Min.   :  0.000      Min.   :0.000e+00   Min.   :       0   Min.   :0.000e+00
 1st Qu.:2.704e+05           1st Qu.:   15842   1st Qu.:  2.293      1st Qu.:1.097e+05   1st Qu.:       1   1st Qu.:5.952e+04
 Median :2.264e+07           Median :  607245   Median :  7.191      Median :8.693e+07   Median :     834   Median :6.144e+06
 Mean   :1.782e+08           Mean   : 2387408   Mean   : 15.239      Mean   :2.436e+11   Mean   : 1857248   Mean   :8.856e+07
 3rd Qu.:1.111e+08           3rd Qu.: 1800091   3rd Qu.: 14.148      3rd Qu.:1.085e+11   3rd Qu.:  830330   3rd Qu.:2.620e+07
 Max.   :3.987e+09           Max.   :53191582   Max.   :146.595      Max.   :4.140e+12   Max.   :36872804   Max.   :5.352e+09
```

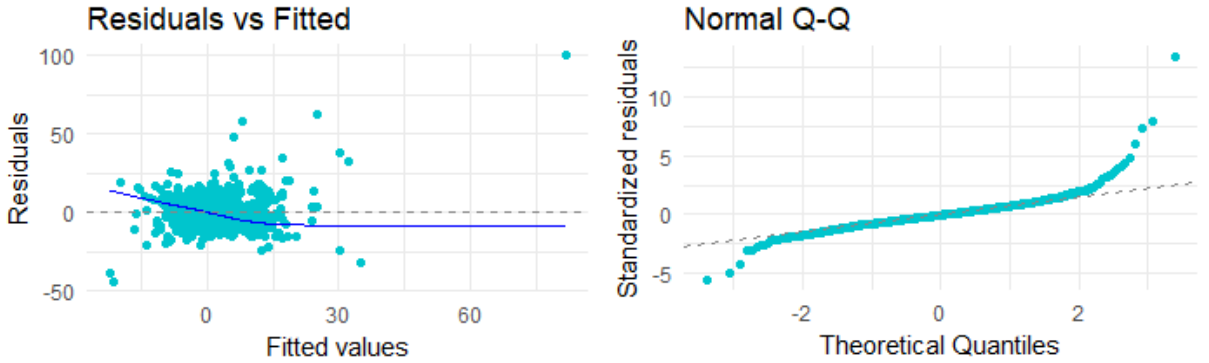
Plot 1 – Correlation



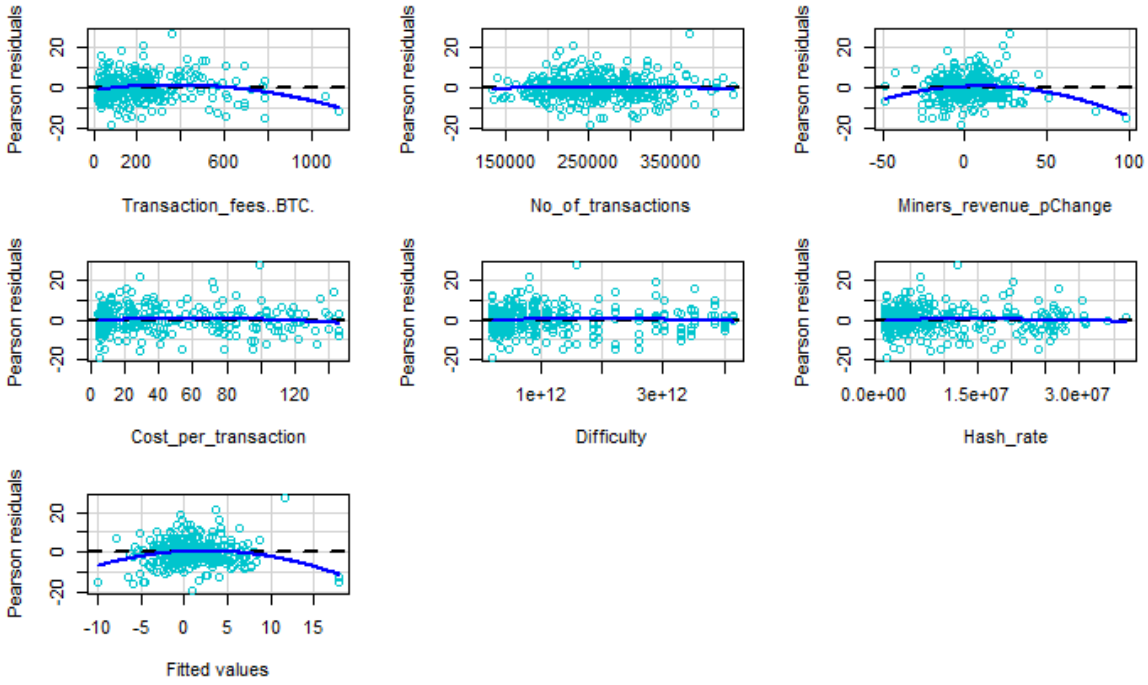
Plot 2 – Residuals vs Fitted, & Q-Q plot (additive model)



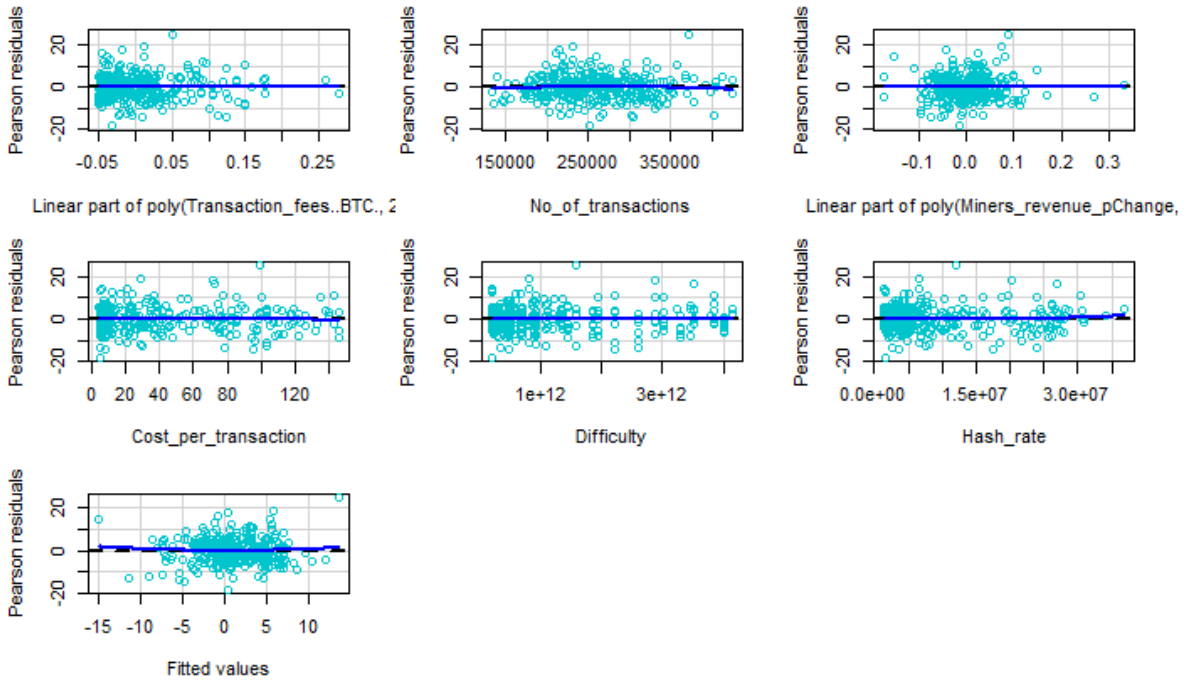
Plot 3 – Residuals vs Fitted, & Q-Q plot (last model on the total data)



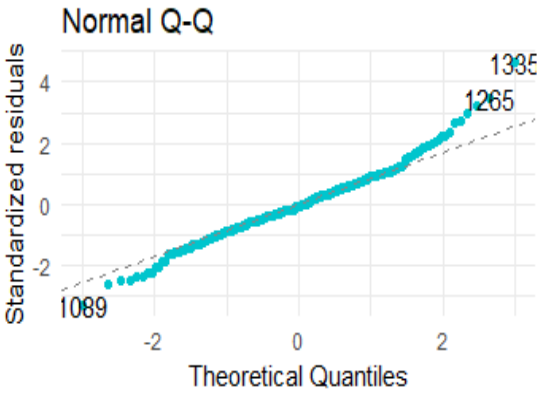
Plot 4 – Residuals plots (model for data from June 2016 with % change in price and miner's revenue)



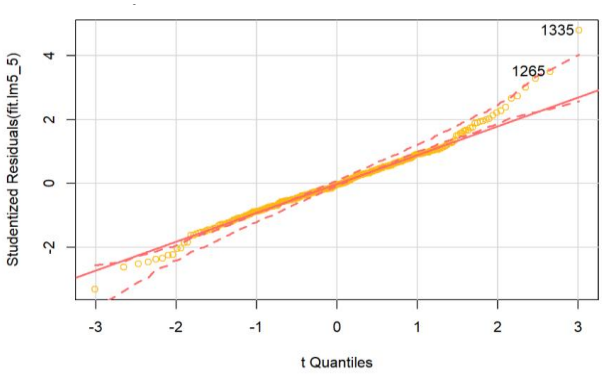
Plot 5 – Residuals plots (final model)



Plot 6 – Normal Q-Q plot (final model)



Plot 7 – Q-Q outliers test (final model)



R CODE

```
library(corrplot)
library(ggplot2)
library(ggfortify)
library(zoo)
library(dplyr)
library(lubridate)
library(car)
library(plotly)
library(reshape2) # for melt
library(data.table)
library(animation)
library(feather)

bitcoin = read.csv("bitcoin_data.csv", header = TRUE)
bitcoin <- bitcoin[1:13]
str(bitcoin)

bitcoin <- bitcoin[1:1711,]
bitcoin$Date <- as.Date(bitcoin$Date, format = "%m/%d/%y")
head(bitcoin)

anyNA(bitcoin)

summary(bitcoin)

library(ggplot2)
ggplot(bitcoin) +
  geom_histogram(aes(x= Price), col = "goldenrod3", fill = "goldenrod1") +
  theme_minimal()

ggplot(bitcoin, aes(bitcoin$Date, bitcoin$Price)) +
  geom_point(color = "turquoise3") +
  ggtitle("Bitcoin price in USD over time") +
  labs ( x="Date" , y="Bitcoin price in USD") +
  theme_minimal()

cor = cor(bitcoin[,c(2:13)])
corrplot(cor, method = "square", type="upper", tl.srt = 50, tl.col = "black",
tl.cex = 0.6, title = "")

correlation = as.data.frame(cor)
highly_correlated = correlation[correlation$Price>= 0.75, ]
rownames(highly_correlated)

fit.lm1 = lm(Price ~., data=bitcoin[,-1])
summary(fit.lm1)

autoplot(fit.lm1, label = 0, col = "turquoise3") +
  theme_minimal()
```

```
step(fit.lm1)

residualPlots(fit.lm1)

wts = 1/fitted(lm(abs(residuals(fit.lm1)) ~ fitted(fit.lm1)))^2

fit.lm3= lm(formula = Price ~  Day + Week + Transaction_fees..BTC. +
    No_of_transactions + Output_value..BTC. + estimated_transaction_value +
    Miners_revenue + Cost_per_transaction + Difficulty + Hash_rate, data =
bitcoin, weights=wts)

autoplot(fit.lm3, label = 0, col = "goldenrod1") +
  theme_minimal()

bitcoin_final <- subset(bitcoin, bitcoin$Price>0)
bitcoin_final = mutate(bitcoin_final, pChange=(bitcoin_final$Price-
lag(bitcoin_final$Price))/lag(bitcoin_final$Price)*100)
bitcoin_final = mutate(bitcoin_final,
Miners_revenue_pChange=(bitcoin_final$Miners_revenue-
lag(bitcoin_final$Miners_revenue))/lag(bitcoin_final$Miners_revenue)*100)
is.na(bitcoin_final) = sapply(bitcoin_final, is.infinite)
bitcoin_final[is.na(bitcoin_final)]=0
summary(bitcoin_final)

fit.lm4= lm(formula = pChange ~ ., data = bitcoin_final[,-c(1,4,9)])

summary(fit.lm4)

step(fit.lm4)

fit.lm5=lm(formula = pChange ~ Transaction_fees..BTC. +
    Cost_per_transaction_value + Difficulty + Hash_rate +
Miners_revenue_pChange,
    data = bitcoin_final[, -c(1, 4, 9)])
summary(fit.lm5)

autoplot(fit.lm5, label = 0, col = "turquoise3") +
  theme_minimal()
lmtest::bptest(fit.lm5) # Breusch-Pagan test
car::ncvTest(fit.lm5)
shapiro.test(resid(fit.lm5))
bitcoin_final$Year <- format(bitcoin_final$Date, "%Y")
bitcoin_final$Month <- format(bitcoin_final$Date, "%b")
bitcoin_final$Day <- format(bitcoin_final$Date, "%d")
bitcoin_final$CommonDate <- as.Date(paste0("2000-",format(bitcoin_final$Date,
"%j")), "%Y-%j")
head(bitcoin_final)
(g <- ggplot(data = bitcoin_final,
       mapping = aes(x = CommonDate, y = Price, shape = Year, colour = Year))
+    geom_point() +
    geom_line() +
```

```
    facet_grid(facets = Year ~ .) +
    scale_x_date(labels = function(x) format(x, "%d-%b")) +
   xlab("Date") +
  theme_minimal())
ggplotly(g)
bitcoin_new = subset(bitcoin_final, bitcoin_final$Date >= '2016-06-01')
bitcoin_new$Day = as.numeric(bitcoin_new$Day)
head(bitcoin_new)
ggplot(bitcoin_new, aes(bitcoin_new$Date, bitcoin_new$Price)) +
  geom_point(color = "turquoise3") +
  ggtitle("Trend of bitcoin price from 2016") +
  labs ( x="Date" , y="Bitcoin price in USD") +
  theme_minimal()
cor = cor(bitcoin_new[,c(2:12)])
corrplot(cor, method = "square", type="upper", tl.srt = 50, tl.col = "black",
tl.cex = 0.6, title = "Correlation of Variables")
correlation = as.data.frame(cor)
highly_correlated = correlation[correlation$Price>= 0.75, ]
rownames(highly_correlated)
bitcoin_new_set = bitcoin_new[,c(1:13)]
fit.lm1_1 = lm(Price ~., data=bitcoin_new_set[,-1])
summary(fit.lm1_1)
autoplot(fit.lm1_1, label = 0, col = "goldenrod1") +
  theme_minimal()
step(fit.lm1_1)
fit.lm2_2= lm(formula = Price ~ Week + Transaction_fees..BTC. +
No_of_transactions +
    Output_value..BTC. + estimated_transaction_value + Miners_revenue +
    Cost_per_transaction + Difficulty + Hash_rate + Trade_volume,
    data = bitcoin_new_set[, -1])
summary(fit.lm2_2)
autoplot(fit.lm2_2, label = 0, col = "goldenrod1") +
  theme_minimal()
residualPlots(fit.lm2_2, col = "skyblue")
wts = 1/fitted(lm(abs(residuals(fit.lm2_2)) ~ fitted(fit.lm2_2)))^2

fit.lm3_3= lm(formula = Price ~ Transaction_fees..BTC. +
    No_of_transactions + Miners_revenue + Cost_per_transaction + Difficulty +
Hash_rate, , data = bitcoin_new_set, weights=wts)

autoplot(fit.lm3_3, label = 0, col = "goldenrod1") +
  theme_minimal()
residualPlots(fit.lm3_3, col = "skyblue")
fit.lm4_4= lm(formula = pChange ~ Transaction_fees..BTC. +
    No_of_transactions + Miners_revenue_pChange + Cost_per_transaction +
Difficulty + Hash_rate , data = bitcoin_new)

summary(fit.lm4_4)
residualPlots(fit.lm4_4, col = "turquoise3")
```

```
fit.lm5_5= lm(formula = pChange ~ poly(Transaction_fees..BTC.,2) +
    No_of_transactions + poly(Miners_revenue_pChange,2) +
Cost_per_transaction + Difficulty + Hash_rate , data = bitcoin_new)

summary(fit.lm5_5)
autoplot(fit.lm5_5, col = "turquoise3") +
  theme_minimal()
lmtest::bptest(fit.lm5_5) # Breusch-Pagan test
car::ncvTest(fit.lm5_5)
shapiro.test(residuals(fit.lm5_5))
residualPlots(fit.lm5_5, col = "turquoise3")
qqPlot(fit.lm5_5, id.n = 2, col = "blue", col.lines = "darkgoldenrod1")
outlierTest(fit.lm5_5)
influencePlot(fit.lm5_5)
influenceIndexPlot(fit.lm5_5)
bitcoin_new= as.data.table(bitcoin_new)
datas <- rbindlist(list(bitcoin_new[, .(pChange, Date)],data.table(value =
fit.lm5_5$fitted.values, data_time = bitcoin_new[, Date])))
datas[, type := rep(c("Real", "Fitted"), each = nrow(bitcoin_new))]

ggplot(data = datas, aes(Date, pChange, group = type, colour = type)) +
  geom_line(size = 0.8) +
  labs(x = "Date", y = "Bitcoin price change %",
       title = "Fit from Multiple Linear Regression") +
  theme_minimal()
set.seed(123)
indexes=sample(1:nrow(bitcoin_new), size=0.2*nrow(bitcoin_new))
test = bitcoin_new[indexes,]
train = bitcoin_new[-indexes,]
final_model = lm(formula = pChange ~ poly(Transaction_fees..BTC.,2) +
    No_of_transactions + poly(Miners_revenue_pChange,2) +
Cost_per_transaction + Difficulty + poly(Hash_rate,1) , data = train)
test$pred = predict(final_model, test)
error = mean(abs(test$pred - test$pChange))
error
```