**A Statistical model comparison in predicting Bitcoin Price**

**SEN VARGHESE**

**A Statistical model comparison in predicting Bitcoin Price**

Sen Varghese

# **Introduction**

This study was undertaken to build the optimal model to predict the share price of Bitcoin. Given the high volatile nature of Bitcoin and the limited research on them, this study is significant in that it builds models for predicting Bitcoin prices given information that is available in the market. The model can be employed by investors and fund managers that seek to include Bitcoin in their portfolios.

# **Previous Research**

Given the abrupt rise in Bitcoin as an investable alternative asset, previous research in this area has not been conclusive.

# **Methodology**

This study employs market data on Bitcoin price in USD (btc), the 1oz price of gold (gold), the S&P 500 index (sp500) and the volume of Bitcoin transitions (btctr). The data was collect from the databases of Quantl for the period Jan 2017 to Apr 2018. During this period Bitcoin became a significant investment option for many investors.

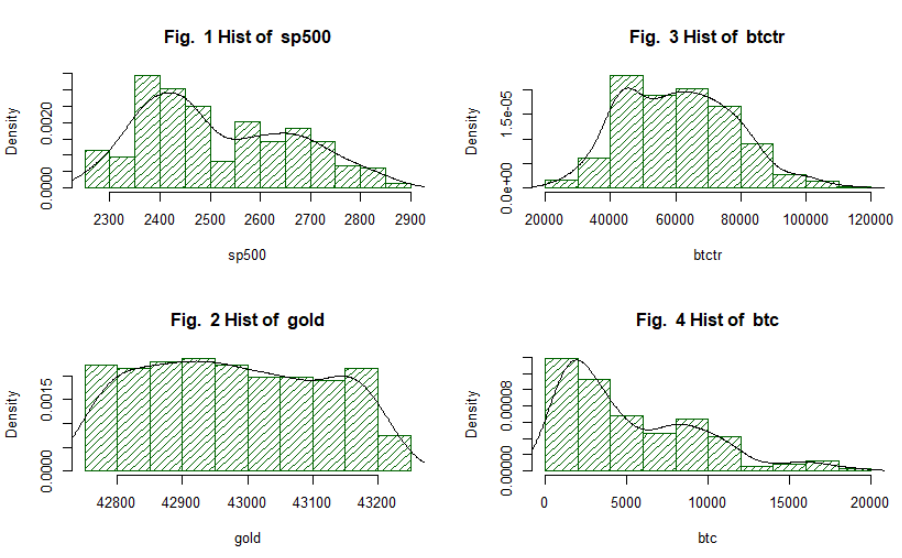
This study is a time series study to predict the price of Bitcoin given the variables mentioned above. The study uses, graphical techniques, correlation analysis, linear regression model, log-linear model, generalized additive model and auto regressive model to predict the depended variable (Price). The dataset has 4 variables and 296 observations, and the analysis is done using the statistical package R. The functional specification is as follows:

|  |  |  |
| --- | --- | --- |
|  | Linear model: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ +\_\_\_+\_\_\_+s |  |
|  |  | Eqn.1  Eqn.2  Eqn.3 |
|  | log-linear model: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ +\_\_\_+\_\_\_+s |  |
|  |  | Eqn.4  Eqn.5  Eqn.6 |
|  | AR (1) model: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_+s |  |
|  |  | Eqn.7  Eqn.8  Eqn.9 |

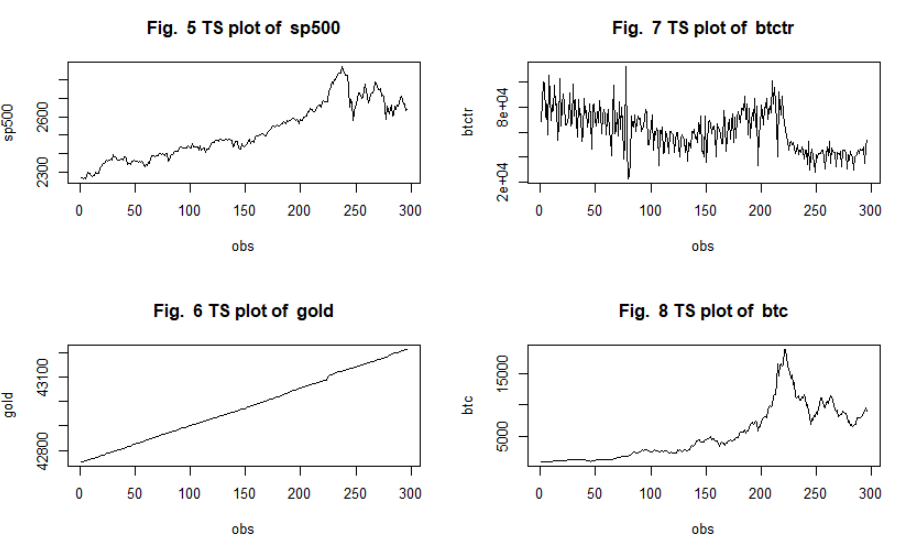
# **Results**

The histogram of variables shows in Figure 1 – 4, indicates that sp500 and btc tend to be positively skewed, while btc transactions (btctr) is normally distributed. the descriptive statistics in table 1 indicates that sp500, btctr and gold are platykurtic. The time series plots shown in figure 5-8, indicates that the price of gold increases over time, while sp500 and btc are both increasing at an increasing rate over time.

## Histograms

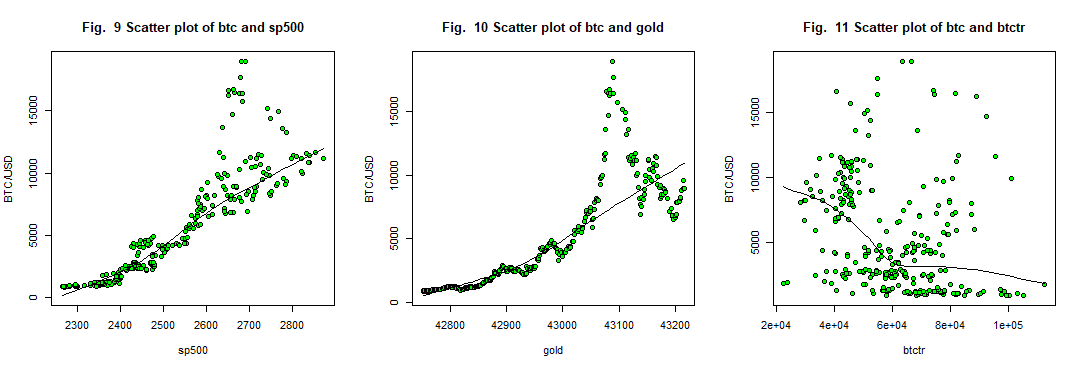


## Time series plot



## Scatterplots

The scatter plots of dependent variable btc against each of the 3 independent variables shown in Fig. 9 – 11 indicate that btc has a strong, positive linear relationship with sp500 and gold as mentioned in the functional specification, while the relationship between btc and btctr remain inconclusive.



Descriptive statistics for the data as shown in Table 1 includes the mean, median, standard deviation (sd), skewness and kurtosis for each variable.

### Table 1: Descriptive Statistics

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variables | n | mean | sd | median | min | max | range | skew | kurtosis |
| sp500 | 296 | 2515 | 149 | 2473 | 2263 | 2872 | 609 | 0.41 | -0.87 |
| gold | 296 | 42978 | 135 | 42970 | 42752 | 43215 | 463 | 0.07 | -1.19 |
| btctr | 296 | 60733 | 17007 | 60066 | 22332 | 112714 | 90382 | 0.3 | -0.4 |
| btc | 296 | 5233 | 4208 | 3975 | 831 | 18949 | 18118 | 1.02 | 0.36 |

The correlation matrix in table 2 indicates that sp500 and gold are positively correlated with Price. The correlation of 0.886 and 0.825 indicates a strong positive correlation between sp500, btctr and Price.

### Table 2: Correlation matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | btc | btctr | gold | sp500 |
| sp500 | 0.886 | -0.475 | 0.935 | 1 |
| gold | 0.825 | -0.532 | 1 | 0.935 |
| btctr | -0.297 | 1 | -0.532 | -0.475 |
| btc | 1 | -0.297 | 0.825 | 0.886 |

## Regression results:

The results of the regression analysis resulted in the following sample regression equations (excluding GAM):

|  |  |  |
| --- | --- | --- |
| Linear Regression |  | Eqn. 10 |
| Log linear model |  | Eqn. 12 |
| AR (1) model |  | Eqn. 13 |

The summary results for the four models employed in this study are shown in Table x. The results indicate that while all models the AR (1) model and log-linear model achieve the highest R-squared, however, the log-linear model has a Durbin–Watson statistic less than 2 indicating the presence of positive serial correlation.

### Table 3: Summary Results

|  |  |  |  |
| --- | --- | --- | --- |
| Summary | Multiple R-squared | F-statistic | DW |
| Linear | 0.8068 | 406 | 0.2827 |
| log-linear | 0.9161 | 1063 | 0.2341 |
| GAM | 0.8476 | n/a | 0.2853 |
| AR (1) | 0.9864 | 21310 | 1.9353 |

## Interpretation of Regression Results:

### T-statistic

The t-test is intended to test the significance of the coefficient of each independent variable. The hypothesis for the t-test is as follows:

H0: B1 = 0, the variable is not significant in explaining the dependent variable, ceteris paribus.

HA: B1 ≠ 0, the variable is significant in explaining the dependent variable, ceteris paribus.

The cut off values for the t-stat pertaining to the study are shown in table 4:

### Table 4: t-statistic cut off values

|  |  |  |
| --- | --- | --- |
| Level of significance | Hypothesis | Cut off for t-stat (df = 292) |
| 10% | H0: β = 0 | 1.28 |
|  | HA: β > 0 |  |
| 5% | H0: β = 0 | 1.65 |
|  | HA: β > 0 |  |
| 1% | H0: β = 0 | 2.33 |
|  | HA: β > 0 |  |

The t-statistic (appendix I) for the linear model indicates that sp500 and btctr are significant at a 1% level of significance, while the t-statistic for the log-linear model indicates that all variables are significant at a 1% level of significance. The t-statistic of the one period lag of btc is significant at a 1% level of significance in predicting btc in the AR (1) model.

### Adjusted R-squared

The adjusted R-squared (appendix I) of 0.9864 for the AR (1) indicates that 98.64% of variation in the Price of Bitcoin today can explained by the variation in the price of bitcoin yesterday. The adjusted R-squared of 0.8068, 0.9161, 0.8476 for the linear, log-linear and GAM models indicates that 80.68%, 91.61% and 84.76% of variation in the Price of Bitcoin can explained by the variation in sp500, gold and volume of bitcoin transactions.

### F-statistic

The F-statistic is intended to test the significance of the coefficient of all independent variable collectively. The hypothesis for the F-stat is as follows:

H0: B1 = B2 = B3 = 0, the variables are collectively not significant in explaining the dependent variable, ceteris paribus.

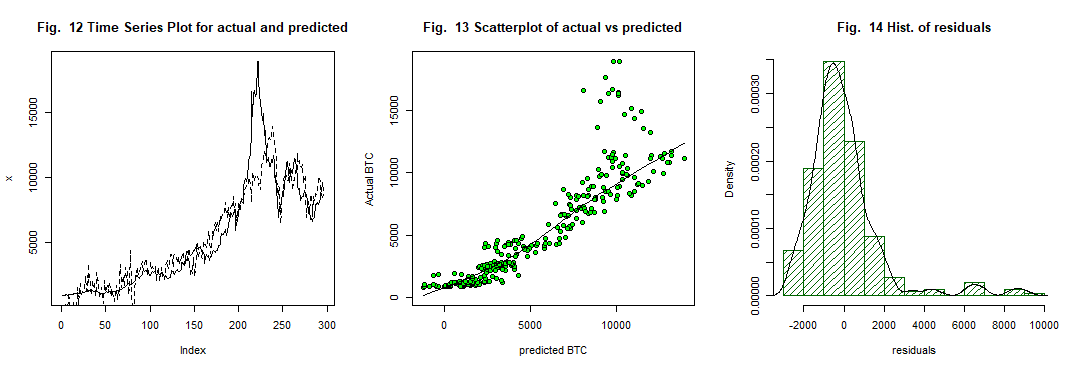
HA: at least one variable is significantly different from 0.

The cut off values for the F-stat pertaining to the study are 2.63 (df1 = 3, df2 = 292) for the linear and log-linear models and 3.88 (df1 = 1, df2 = 292) for the AR (1) model. The F-statistic (appendix I) of the models studied are above the cut off limit. Thus, the alternative hypothesis (*HA*) can be accepted, implying that at least one variable is significantly different from 0 for each model.

### Residual analysis:

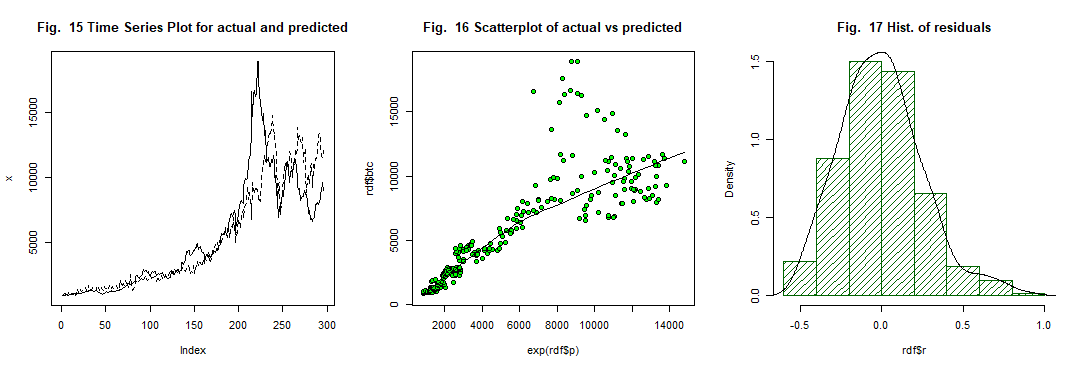
#### Residual analysis of linear model

The time series and of actual btc vs predicted btc (Fig. 12) indicate that the predicted model closely trends the actual btc price, the scatterplot of actual vs predicted (Fig. 13) indicates a strong, positive, linear relationship between predicted price of btc with the actual price of btc. However, there are several outliers in the scatterplot, and the histogram of Residuals are skewed to the right (shown in Fig 14).



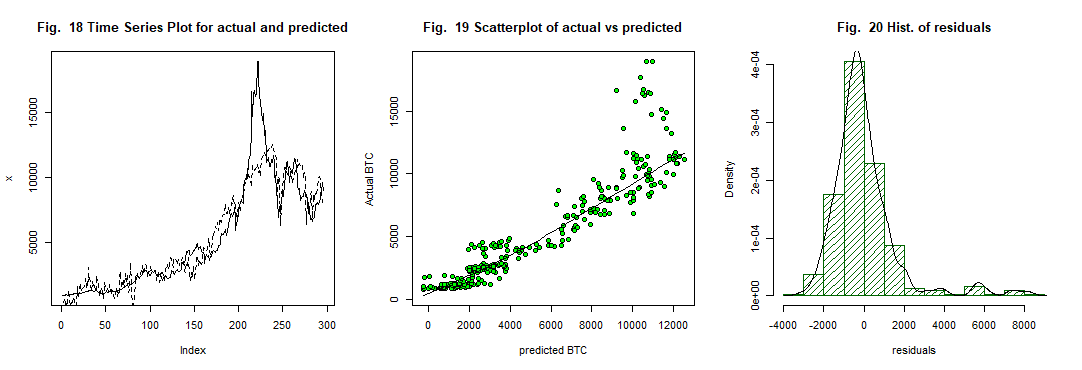
#### Residual analysis of log-linear

The time series and of actual btc vs predicted btc (Fig. 15) for the log-linear model indicate that the predicted model closely trends the actual btc price early on but tends to deviate on more recent periods, the scatterplot of actual vs predicted (Fig. 16) indicates semi-strong, positive, non-linear relationship between predicted price of btc with the actual price of btc with several outliers. The histogram of Residuals (Fig. 17) are slightly skewed to the right.



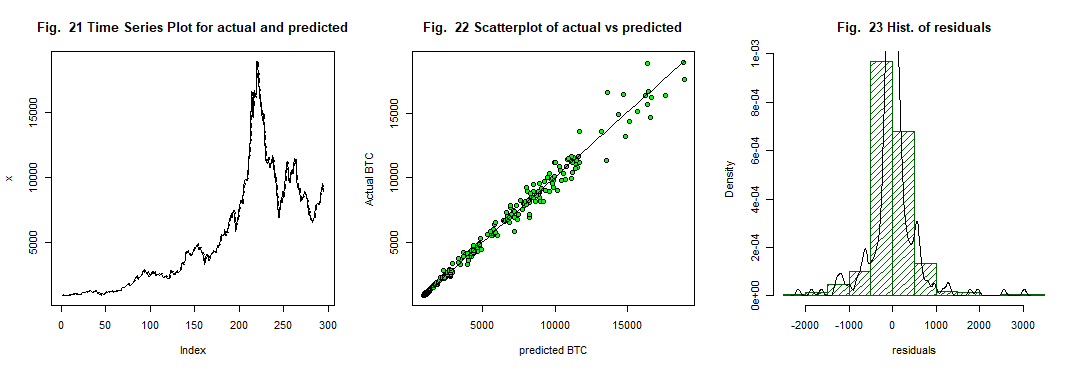
#### Residual analysis of GAM

The rug of the GAM model (shown in appendix 1) indicate that sp500 is the most significant variable in predicting the price of BTC, the time series plot and of actual btc vs predicted btc (Fig. 18) for GAM indicate that the predicted model closely trends the actual btc price, the scatterplot of actual vs predicted (Fig. 19) indicates a strong, positive, linear relationship between predicted price of btc with the actual price of btc. The outliers in the scatterplot, and right skewness of histogram of Residuals (shown in Fig 20) and a DW test statistic of 0.2853 make this model biased.



#### Residual analysis of AR(1)

The time series and of actual btc vs predicted btc (Fig. 21) for the one period lagged model indicate that the predicted model very closely trends the actual btc price, the scatterplot of actual vs predicted (Fig. 22) indicates a very strong, positive, linear relationship between predicted price of btc with the actual price of btc with no outliers. The histogram of Residuals (Fig. 23) are normally distributed.



# **Conclusions**

Given the rise of Bitcoin as a serious alternative investment to tradition stocks and bonds, the reason for researching analytical techniques therein is ever more prevalent. The study employed the linear model, log-linear model, GAM and the AR (1) model in predicting Bitcoin prices.

This research is successful in comparing various statistical models in the prediction of Bitcoin prices. The results indicate that the AR (1) model, with a R-square of 0.986 and DW 1.935 is the most appropriate model in predicting Bitcoin price. This research can be further improved by exploring more models that might have explanatory power and using more data as it becomes available in future.

# **Appendix I**

## Linear Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients: | Estimate | Std. error | t val | Pr(>|t|) |
| (Intercept) | -176300 | 96970 | -1.818 | 0.0701 |
| sp500 | 24.99 | 2.056 | 12.157 | <2.00E-16 |
| gold | 2.702 | 2.363 | 1.143 | 0.2538 |
| btctr | 0.04232 | 0.007541 | 5.612 | <4.64E-08 |
|  |  |  |  |  |
| Residual standard error | 1859 on 292 degrees of freedom | |  |  |
| Multiple R-squared | 0.8068 |  |  |  |
| F-statistic | 406.4 on 3 and 292 DF |  |  |  |
| p-value | < 2.2e-16 |  |  |  |
| DW | 0.28271 |  |  |  |
| Adjusted R-squared | 0.8048 |  |  |  |

## Log-linear model

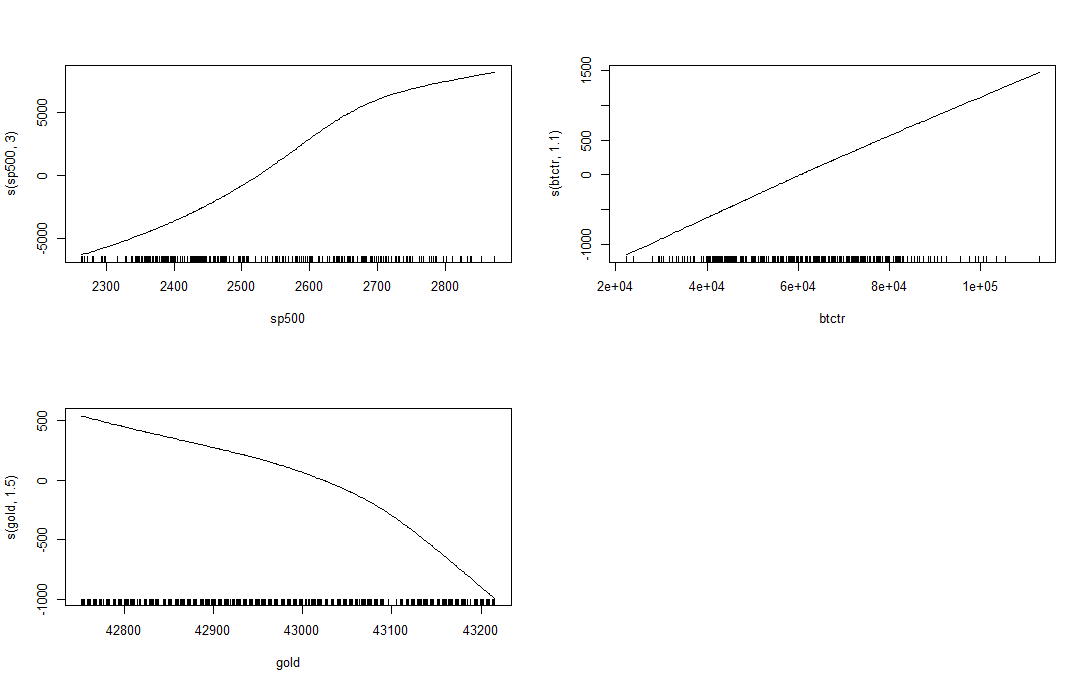
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients: | Estimate | Std. error | t val | Pr(>|t|) |
| (Intercept) | -166.000 | 1.34E+01 | -12.423 | 2.00E-16 |
| sp500 | 0.003 | 2.83E-04 | 8.982 | 2.00E-16 |
| gold | 0.004 | 3.26E-04 | 11.965 | 2.00E-16 |
| btctr | 0.000 | 1.04E-06 | 5.918 | 9.12E-09 |
| --- |  |  |  |  |
| Residual standard error | 0.2561 on 292 degrees of freedom | |  |  |
| Multiple R-squared | 0.9161 |  |  |  |
| F-statistic | 1063 |  |  |  |
| p-value | < 2.2e-16 |  |  |  |
| DW | 0.23411 |  |  |  |
| Adjusted R-squared | 0.9152 |  |  |  |

## AR (1)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients: | Estimate | Std. error | t val | Pr(>|t|) |
| (Intercept) | 65.8188 | 45.5748 | 1.444 | 0.15 |
| btc.lag1 | 0.9927 | 0.0068 | 145.979 | 2.00E-16 |
| --- |  |  |  |  |
| Residual standard error | 490.9 on 293 degrees of freedom | |  |  |
| Multiple R-squared | 0.9864 |  |  |  |
| F-statistic | 2.13E+04 |  |  |  |
| p-value | 2.20E-16 |  |  |  |
| DW | 1.9353 |  |  |  |
| Adjusted R-squared | 0.9864 |  |  |  |

## GAM model

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Anova for Parametric Effects | Df | Sum Sq | Mean-sq | F-val | Pr(>F) |
| s(sp500,3) | 1 | 4099876262 | 4099876262 | 1487.052 | 2.20E-16 |
| s(gold,1.5) | 1 | 21249380 | 21249380 | 7.7073 | 0.005858 |
| s(btctr,1.1) | 1 | 52253960 | 52253960 | 18.9528 | 1.86E-05 |
| Residuals | 289.4 | 797890468 | 2757050 |  |  |
| --- |  |  |  |  |  |
| Residual standard error |  |  |  |  |  |
| Multiple R-squared | 0.8475859 |  |  |  |  |
| DW | 0.28271 |  |  |  |  |



# **Appendix II**

## R-Script

rm(list = ls())

#

# 1. Data and Preprocessing

#

# Load Packages - highlight and run these lines EACH time you launch R.

library("psych")

library("Hmisc")

library("moments")

library("lmtest")

library("gam")

library("readxl")

library("robust")

#import data from local dir

datadump <- read\_excel("C:/Users/Sen/Downloads/DS 633 (Statistics)/datadump.xlsx")

#View(datadump)

#convert to dataframe without na's

mdf <- datadump[1:322,]

mdf <- mdf[order(as.Date(mdf$Date, format = "%Y-%m-%d")),] #order by date

mdf\_clean <- subset(mdf,!is.null(mdf))

mdf\_clean <- as.data.frame(na.omit(mdf\_clean))

dim(mdf\_clean)

names(mdf\_clean)

#Custom functions used in the Analysis

ftsplot2 <- function(x,y, counter){

# displays time series plot with 2 lines

# x is in solid line and y is in dotted line

plot(x,type="l",lty=1, main = paste("Fig. ", counter,"Time Series Plot for actual and predicted")); lines(y,lty=2)}

residAnalysis <- function(rdf, dependent){

'Plots the hist of residuals,

actual vs predicted,

scatterplot of actual vs predicted

and %change if resid over time

'

par(mfcol=c(1,3))

ftsplot2(dependent,rdf$p,counter) ## Time Series Plot for aactual and predicted

counter <<- counter + 1

scatter.smooth(rdf$p,dependent, main = paste("Fig. ",counter,"Scatterplot of actual vs predicted")) #scatterplot of actual vs predicted

counter <<- counter + 1

hist(rdf$r, main = paste("Fig. ",counter,"Hist. of residuals"),prob=1,density = 20, col = "dark green")

lines(density(rdf$r))

counter <<- counter + 1 #add to global variable counter

'

rdf$pctResid <- rdf$r/dependent

ts.plot(rdf$pctResid, main = paste("Fig. ",counter,"Percentage change in resid"));abline(h=0) #percentage change in resid

counter <<- counter + 1

'

}

#

# 2. GRAPHS

#

#Histograms

counter = 1

par(mfcol=c(2,2))

for (i in 2:5){

index = names(mdf\_clean[i])

hist( mdf\_clean[,i],xlab = index, main= paste("Fig. ",counter, "Hist of ",index),prob=1,density = 20, col = "dark green")

lines(density(mdf\_clean[,i]))

counter = counter + 1

}

#Time-series plot

par(mfcol=c(2,2))

for (i in 2:5){

index = names(mdf\_clean[i])

ts.plot(mdf\_clean[,i],xlab = 'obs',ylab = index, main= paste("Fig. ",counter, "TS plot of ",index))

counter = counter + 1

}

#Scatter-plot of all var vs BTC/$

par(mfcol=c(1,3))

for (i in 2:4){

index = names(mdf\_clean[i])

scatter.smooth(mdf\_clean[,i], mdf\_clean$btc, xlab=index, ylab="BTC/USD",main= paste("Fig. ",counter, "Scatter plot of btc and",index))

counter = counter + 1

}

#

# 3. Descriptive stats

#

describe.by(mdf\_clean[,2:5])

#Correlarion

round(cor(mdf\_clean[,2:5]),3)

#

# 4. Analyisis

#

# 4.1. Simple Regression Model

fit = lm(btc~sp500+gold+btctr, data=mdf\_clean)

summary(fit)

dwtest(fit)

# residual analysis

rdf <- data.frame(mdf\_clean, p= fit$fitted.values, r= fit$residuals)

residAnalysis(rdf,rdf$btc)

# 4.2. log-linear regression

mdf\_clean$btclog <- log(mdf\_clean$btc)

ts.plot(mdf\_clean$btclog)

ts.plot(mdf\_clean$btc)

fit = lm(btclog~sp500+gold+btctr, data=mdf\_clean)

summary(fit)

dwtest(fit)

# residual analysis

rdf <- data.frame(mdf\_clean, p= fit$fitted.values, r= fit$residuals)

#residAnalysis(rdf, rdf$btc)

par(mfcol=c(1,3))

ftsplot2(rdf$btc,exp(rdf$p),counter) ## Time Series Plot for aactual and predicted

counter = counter + 1

scatter.smooth(exp(rdf$p),rdf$btc, main = paste("Fig. ",counter,"Scatterplot of actual vs predicted")) #scatterplot of actual vs predicted

counter = counter + 1

hist(rdf$r, main = paste("Fig. ",counter,"Hist. of residuals"),prob=1,density = 20, col = "dark green")

lines(density(rdf$r))

counter = counter + 1

# 4.3. GAM

fit<-gam(btc~s(sp500,3)+ s(gold,1.5)+ s(btctr,1.1), data = mdf\_clean)

summary(fit)

#plot GAM

par(mfcol=c(2,2))

plot.gam(fit)

rdf <- data.frame(mdf\_clean, p= fit$fitted.values, r= fit$residuals)

cor(rdf$btc,rdf$p)^2 #Psudo r^2

dwtest(fit)

# residual analysis

residAnalysis(rdf,rdf$btc)

# 4.4. AR(1) linear regression

n <- length(mdf\_clean$btc)

mdf\_clean$btc.lag1 = Lag(mdf\_clean$btc,+1) #Hmic

mdf\_clean <- na.omit(mdf\_clean)

fit <- lm(btc~btc.lag1, data=mdf\_clean)

summary(fit)

dwtest(fit)

# residual analysis

rdf <- data.frame(mdf\_clean, p= fit$fitted.values, r= fit$residuals)

residAnalysis(rdf,rdf$btc)

# **Appendix III**

## Sample Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Date | sp500 | gold | btctr | btc |
| 4/26/2018 | 2666.94 | #N/A | 43932.00 | 9240.00 |
| 4/25/2018 | 2639.40 | 43215.00 | 53089.00 | 8950.10 |
| 4/24/2018 | 2634.56 | 43214.00 | 48325.00 | 9570.01 |
| 4/23/2018 | 2670.29 | 43213.00 | 35156.00 | 8933.79 |
| 4/20/2018 | 2670.14 | 43210.00 | 48539.00 | 8831.56 |
| 4/19/2018 | 2693.13 | 43209.00 | 45773.00 | 8255.22 |
| 4/18/2018 | 2708.64 | 43208.00 | 45790.00 | 8165.22 |
| 4/17/2018 | 2706.39 | 43207.00 | 46236.00 | 7908.49 |
| 4/16/2018 | 2677.84 | 43206.00 | 42989.00 | 8023.22 |
| 4/13/2018 | 2656.30 | 43203.00 | 45132.00 | 7890.00 |
| 4/12/2018 | 2663.99 | 43202.00 | 42166.00 | 7877.00 |
| 4/11/2018 | 2642.19 | 43201.00 | 41272.00 | 6923.22 |
| 4/10/2018 | 2656.87 | 43200.00 | 40077.00 | 6804.43 |
| 4/9/2018 | 2613.16 | 43199.00 | 29423.00 | 6692.00 |
| 4/6/2018 | 2604.47 | 43196.00 | 40207.00 | 6550.39 |
| 4/5/2018 | 2662.84 | 43195.00 | 41346.00 | 6832.02 |
| 4/4/2018 | 2644.69 | 43194.00 | 44651.00 | 6763.59 |
| 4/3/2018 | 2614.45 | 43193.00 | 37282.00 | 7396.88 |
| 4/2/2018 | 2581.88 | #N/A | 27949.00 | 7028.91 |
| 3/29/2018 | 2640.87 | 43188.00 | 42188.00 | 7163.99 |
| 3/28/2018 | 2605.00 | 43187.00 | 43411.00 | #N/A |
| 3/27/2018 | 2612.62 | #N/A | 42390.00 | 7963.49 |
| 3/26/2018 | 2658.55 | 43185.00 | 30027.00 | 8188.52 |
| 3/23/2018 | 2588.26 | 43182.00 | 43862.00 | 8651.51 |
| 3/22/2018 | 2643.69 | 43181.00 | 45291.00 | 8705.84 |
| 3/21/2018 | 2711.93 | 43180.00 | 45750.00 | 8840.38 |
| 3/20/2018 | 2716.94 | 43179.00 | 45007.00 | 8940.10 |
| 3/19/2018 | 2712.92 | 43178.00 | 32383.00 | 8510.09 |
| 3/16/2018 | 2752.01 | 43175.00 | 44906.00 | 8499.63 |
| 3/15/2018 | 2747.33 | 43174.00 | 41871.00 | 8332.01 |

# Bibliography

Data source:

Quantl. (2018). Retrieved from: <https://www.quandl.com/>.