Cryptocurrency Price Trend Analysis

Can Google search trends be used an indicator for the price of Bitcoin?

Alexander Turner - s3643665

Last updated: 22 October, 2017

RPubs Link

- http://rpubs.com/alexturner1991/321133

Intro

- Recently, the world of cryptocurrencies (virtually currencies that use blockchain technology) has exploded into the mainstream
- This increasded attention has sent the price of Bitcoin, the largest cryptocurrency in terms of market capitalisation, skyrocketing as investors rush to get a piece of the action:
- https://www.theguardian.com/technology/2017/oct/12/bitcoin-price-5000-cryptocurrency-gold-bubble
- In less than a year, Bitcoin has gone from something few people were aware of to one the hottest investment propositions on the planet
- This final point got me thinking

Problem Statement

- The fact Bitcoin has come from relative obscurity to being one of the top investment choices poses a very interesting question:
- Is there a relationship between online search trends relating to crytocurrencies and the price of Bitcoin?
- The idea behind this question is that people around the world are doing some quick research online before investing themselves
- This question is a bit broad though, so it needs to be narrowed down for statistical analysis:
- Can Google Trends 'Interest over time' index for the search term 'cryptocurrency' be used to predict Bitcoin price in USD?
- Bitcoin has been selected as it is the most well-known cryptocurrency with the highest market cap and 'cryptocurrency' as the search term as it is the fundamental concept behind Bitcoin (making it a logical thing to be searched when looking for an understanding of the topic)
- In order to answer this question a linear regression will be fitted, the model parameters and assumptions will be tested and, finally, the correlation direction and strength will be examined

Data

- This analysis will utilise 2 data sets: one pertaining to historical Bitcoin prices and the other to Google search trends for the word ‘cryptocurrency’
- Both were obtained separately and then combined using Microsoft Excel before importing to RStudio
- Data source #1 is historical Bitcoin prices (in USD) obtained via Kaggle
- Closing price (which is the end of each day in the USA, given cryptocurrencies can be traded 24/7) was taken for for the 80 days from 16/07/2017 to 03/10/2017
- Data source link:
- https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory/data

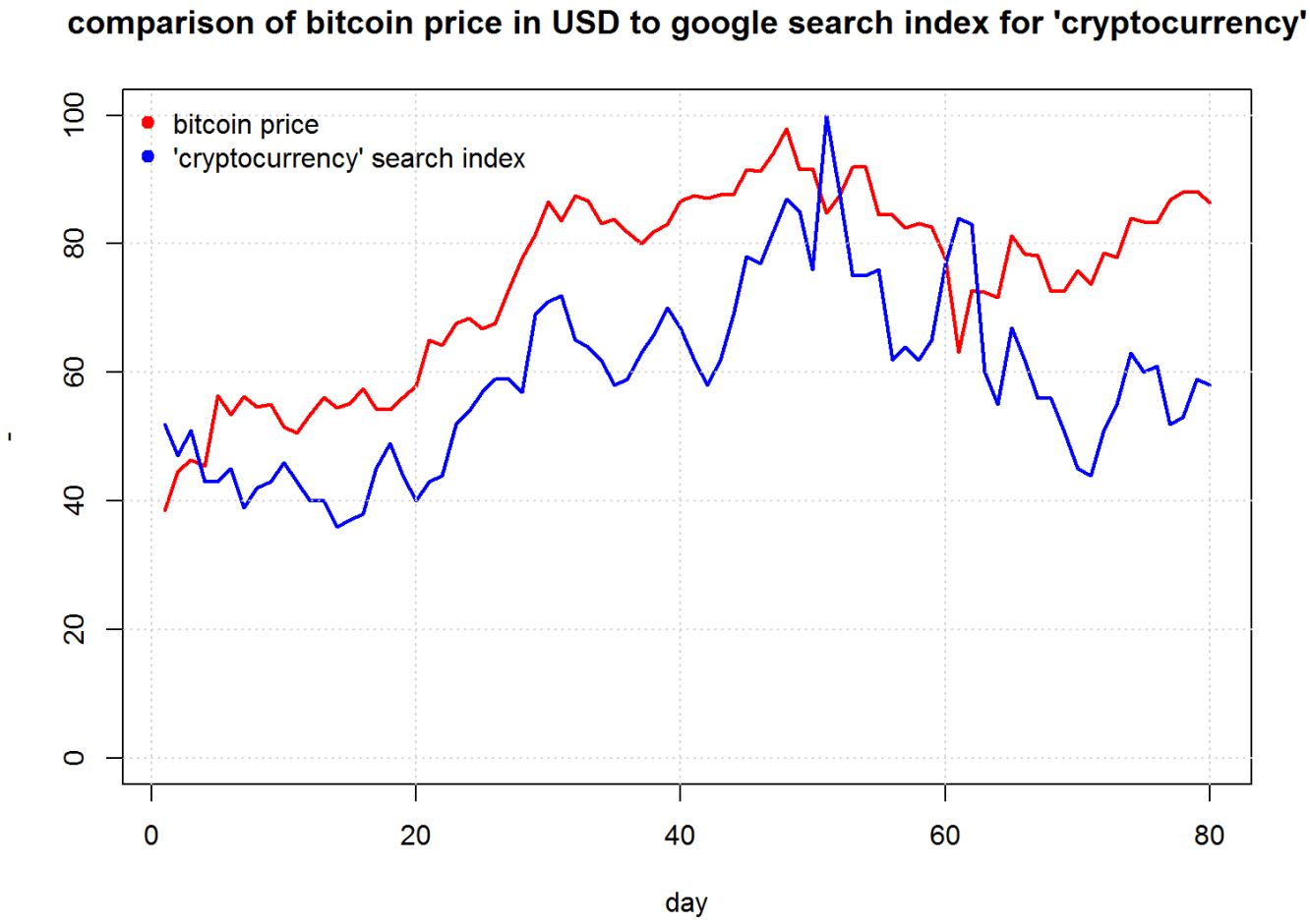
Data Cont.

- Data source #2 is the Google Trends 'Interest over time' index for the word 'cryptocurrency' worldwide
- The daily index was taken for the 80 days from 16/07/2017 to 03/10/2017 (matching the price data set)
- Google's explanation of their 'Interest over time' index:
- 'Numbers represent search interest relative to the highest point on the chart for the given region and time. A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. Likewise a score of 0 means the term was less than 1% as popular as the peak'
- The key here is that the search index is a ratio scale which is appropriate for linear regression
- Data source link:
- https://trends.google.com/trends/explore?date=today%203-m&q=%2Fm%2F0vpj4_b

Descriptive Statistics

- Firstly, let's see how well the search trends and price data match up using a line chart. (The Bitcoin price will be divided by 50 so both variables can be plotted together. This results in the Y axis being the actual search index and 1/50 of the Bitcoin price in USD).
- It seems the overall trends match up fairly well for the period 16/07/2017 to 03/10/2017

```
plot(
  bitcoin_price_search$price/50,
  type = "l",
  col = "red",
  lwd = 2,
  ylim = c(0,100),
  xlab = "day",
  ylab = "-",
  main = "comparison of bitcoin price in USD to google search index for 'cryptocurrency'",
  cex.main = 1.2
)
lines(bitcoin_price_search$search_index, col = "blue", lwd = 2)
grid()
legend("topleft", legend = c("bitcoin price","'cryptocurrency' search index"), col = c("red","blue"), bty = "n", pch
```

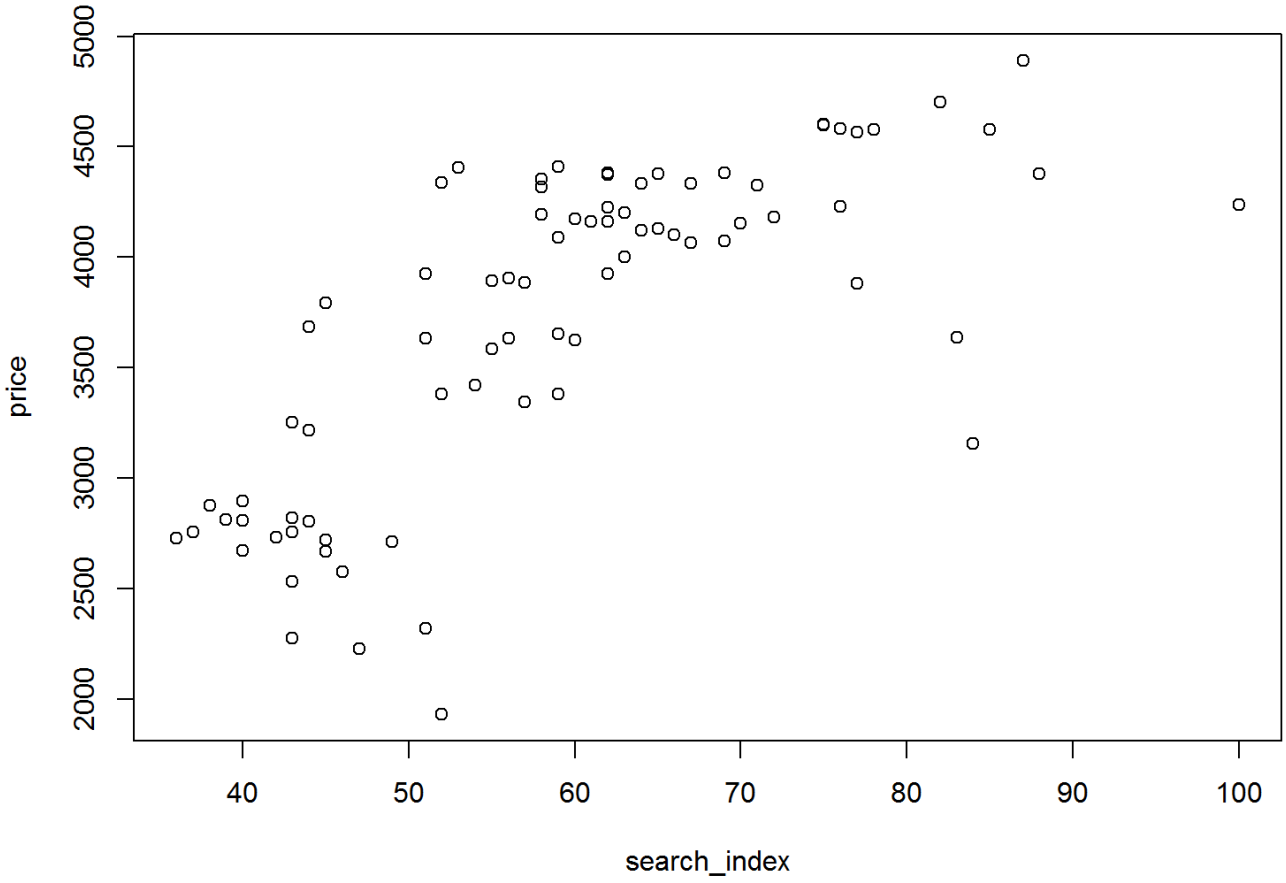


- Interestingly, this period captures the high point of both variables of interest (excluding data later than 03/10/2017)
- Next, let's examine if the relationship is in fact linear

Decsriptive Statistics Cont.

- Scatter plots can help determine the level of linearity (ideally the dots move from bottom left to top right or top left to bottom right):

```
plot(price ~ search_index, data = bitcoin_price_search)
```

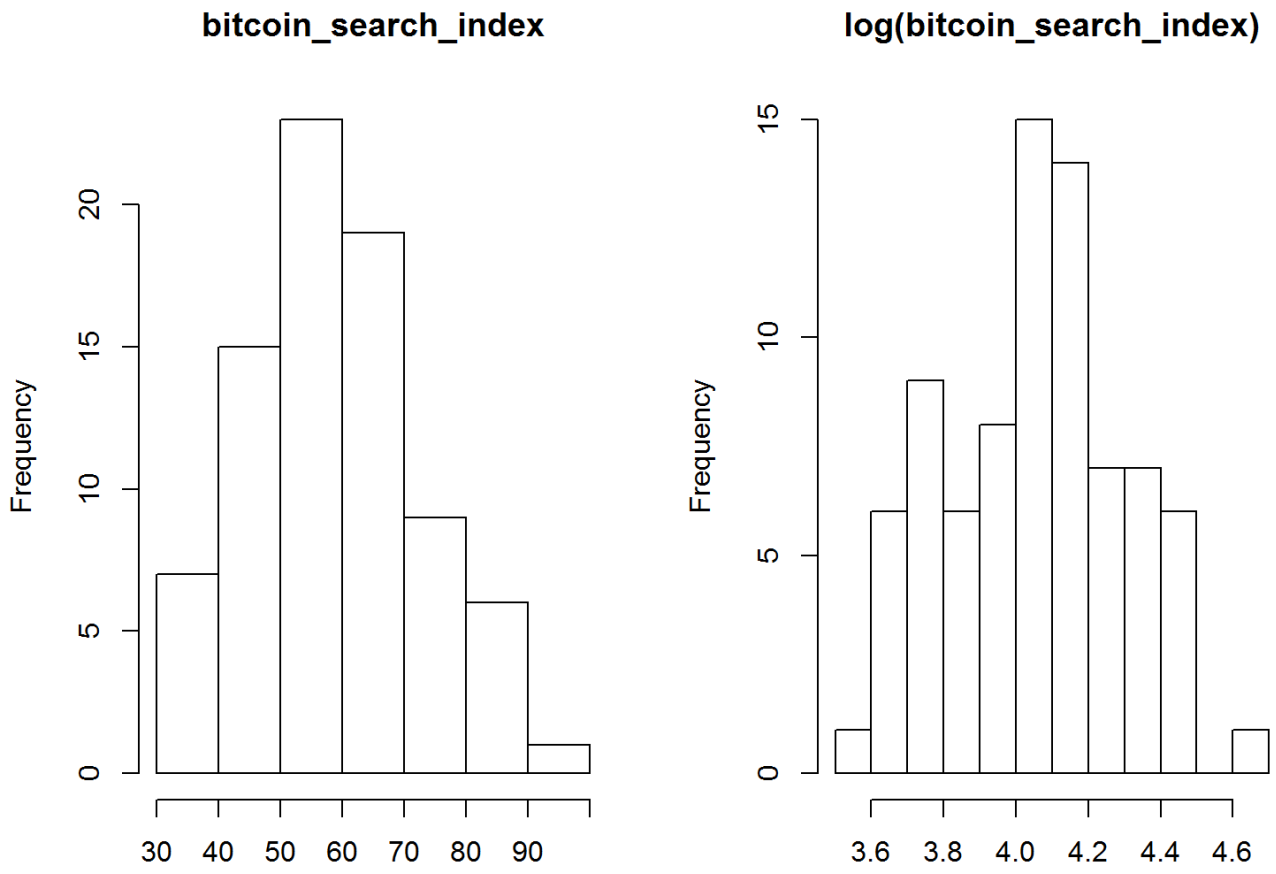


- It doesn't seem to be quite linear, perhaps transforming the X axis variable will help correct for skewness

Decsriptive Statistics Cont.

- Taking the logarithm of a variable can help achieve normality in the distribution of the data, let's do this for the X axis (using the natural log):

```
par(mfrow = c(1,2))
hist(bitcoin_price_search$search_index, main = "bitcoin_search_index", xlab = NULL)
hist(log(bitcoin_price_search$search_index), main = "log(bitcoin_search_index)", xlab = NULL)
```

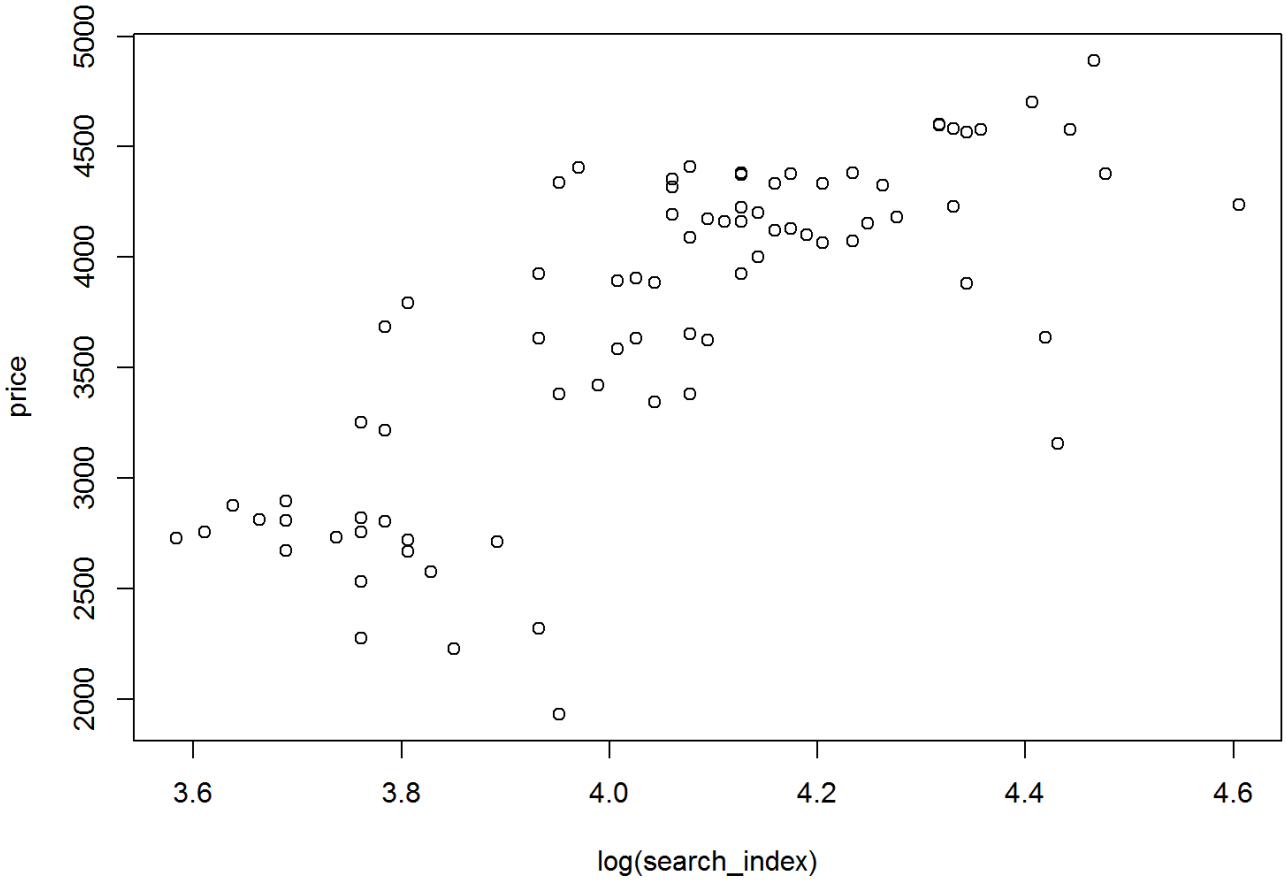


- The distribution looks more normal now

Decsriptive Statistics Cont.

- Now the log function has done it's work we can re-examine the linearity:

```
plot(price ~ log(search_index), data = bitcoin_price_search)
```



- That looks better, we can move on to model fitting and testing

The Model Overall

- Hypotheses test for the model:
  - H_0: *The data does not fit the linear regression model*
  - H_A: *The data does fit the linear regression model*
- Assumptions:
  - *Independence of variables (this is evidently true from previous explanations of data on previous slides)*
  - *Linearity (after the log transformation was applied to search index data a linear relationship was found)*
  - *Normality of residuals (to be checked in further slides after model fitting)*
  - *Homoscedasticity (to be checked in further slides after model fitting)*
- Decision Rules and Conclusion:
  - *Reject* H_0 *if* p*-value is < 0.05*
  - *Model will be statistically significant if we reject* H_0

Testing and Interpreting the Model

```
price_search_model <- lm(price ~ log(search_index), data = bitcoin_price_search)
summary(price_search_model)
```

```
##
## Call:
## lm(formula = price ~ log(search_index), data = bitcoin_price_search)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1542.9  -190.9   102.5   272.4   886.0
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5863.9      913.3  -6.421 9.75e-09 ***
## log(search_index)   2363.0      225.0  10.501  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 472.5 on 78 degrees of freedom
## Multiple R-squared:  0.5857, Adjusted R-squared:  0.5804
## F-statistic: 110.3 on 1 and 78 DF,  p-value: < 2.2e-16
```

- The p-value for the F-test is $< .001$, meaning the model is statistically significant as H_0 is rejected
- The line of best fit is $price = -5863.9 + 2363.0 \times log(search\ index)$
- Intercept: when $log(search\ index) = 0, price = -5863.9$. This intercept doesn't really have a meaningful interpretation though as a Bitcoin price in USD below 0 is impossible
- Slope: as $log(search\ index)$ increases by 1 unit, $price$ changes by 2363.0 on average

Testing Model Parameters

- Both the intercept and slope need to be tested for statistical significance using t-tests
- Hypotheses for linear regression model parameters:
  - *Intercept:*
    - $H_0 : \alpha = 0$
    - $H_A : \alpha \neq 0$
  - *Slope:*
    - $H_0 : \beta = 0$
    - $H_A : \beta \neq 0$
- Assumptions:
  - *Same as overall model*
- Decision Rules and Conclusion:
  - *Reject H_0 if 95% CI of the parameter does not capture $H_0 : \alpha / \beta = 0$*
  - *Parameters will be statistically significant if we reject H_0 for each*

```
confint(price_search_model)
```

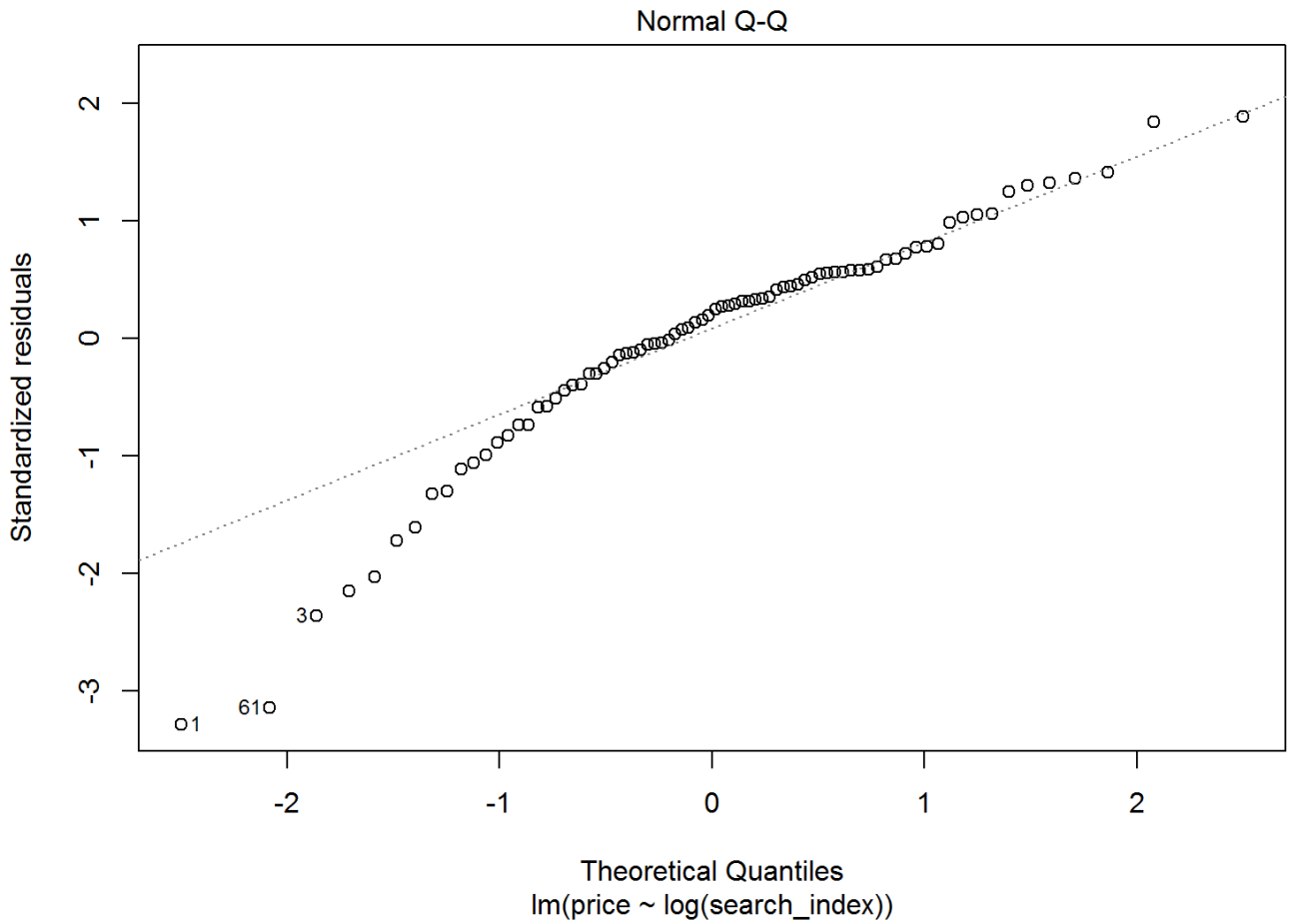
```
##                       2.5 %    97.5 %
## (Intercept)       -7682.195 -4045.666
## log(search_index)  1914.987  2810.952
```

- In both cases the 95% CI does not capture H_0, so it is rejected. The parameters are statistically significant
- All the tests have proved to be statisticall significant so far. Next the assumptions, influential cases and correlation will be looked at

Testing Normality of Residuals

- In a linear regression model the residuals are the difference between an observed score and the corresponding predicted score by the line of best fit
- An assumption of the model is that the residuals will be normally distributed
- The Q-Q plot can be used to test the normality (ideally most of the dots will fall close to the line):

```
plot(price_search_model, which = 2)
```

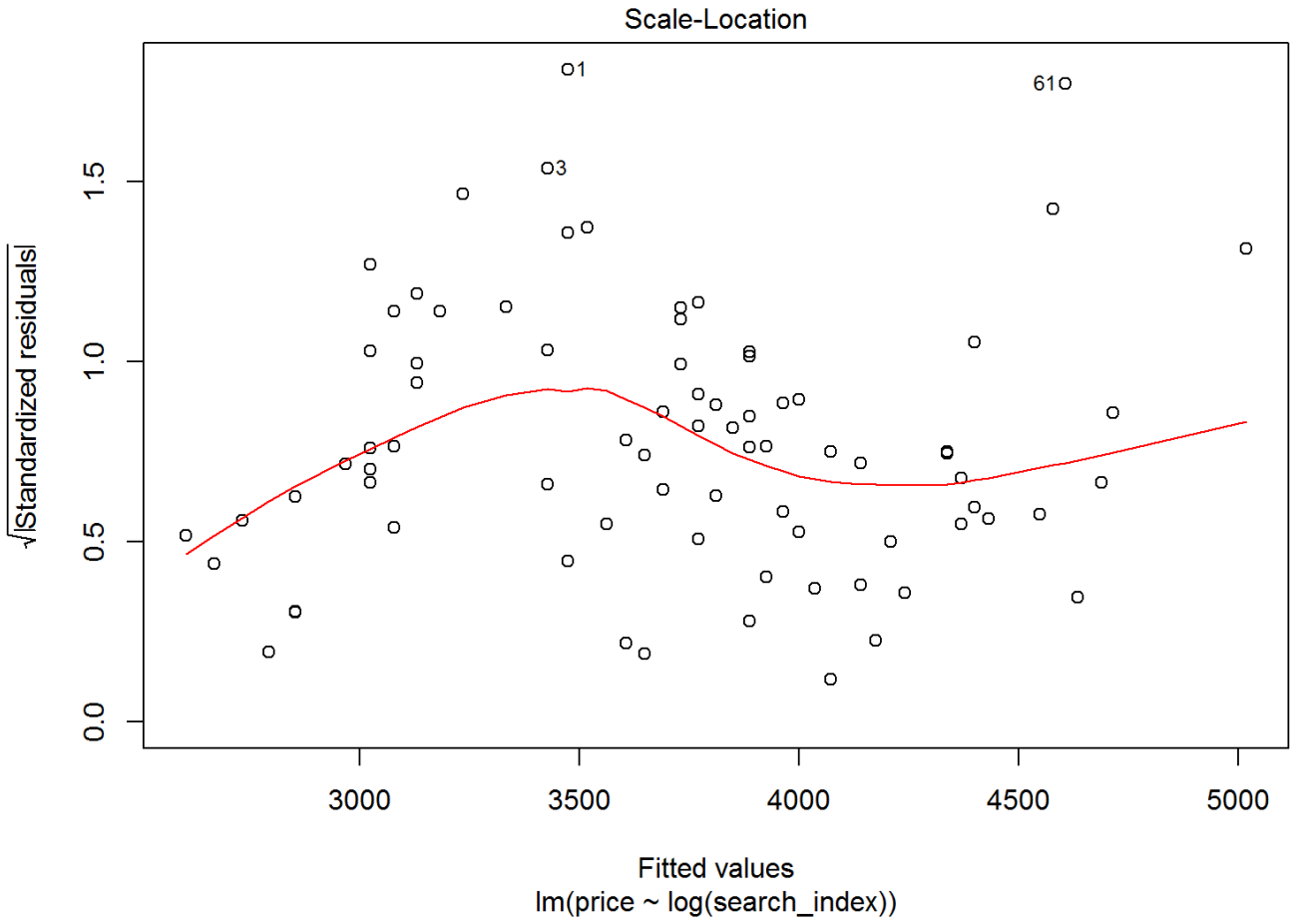


- It appears there are no major deviations from normality as most of the residuals do fall close to the line

Testing Homoscedasticity

- Homoscedasticity is the assumption that the variance around the regression line is the same for all values of the predictor variable ($log(search\ index)$)
- The Scale-Location plot can be used to test homoscedasticity (the red line should be flat due to consistent variance in the residuals):

```
plot(price_search_model, which = 3)
```

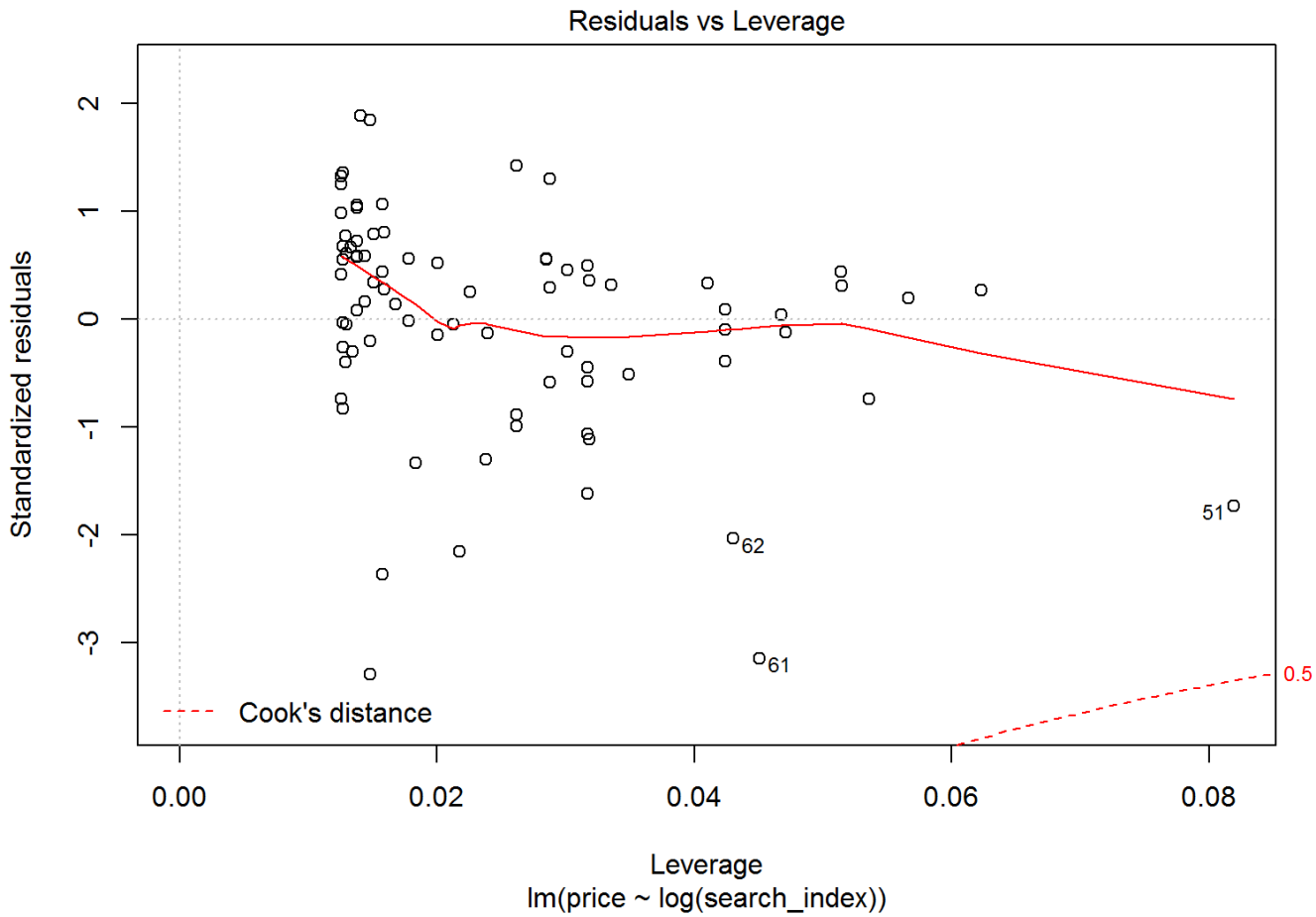


- A bit of a bend in the line, but nothing too dramatic, indicating homoscedasticity is present in the model

Influential Cases

- After the model, its parameters and its assumptions have been tested we can move on to reviewing data points that may have a disproportionally large impact on the fit of the model
- This seems pertinent as there were a few potential outliers in the in the initial scatter plot that was used to check linearity
- We can check for influential using a Residuals vs Fitted plot (the red line should be flat and no dots should fall beyond the red bands):

```
plot(price_search_model, which = 5)
```



- A few outliers, but nothing that should be flagged for removal due to having a large influence on the model

Correlation

- Finally, we need to measure the strength and direction of the linear relationship between the two variables using the correlation coefficient r
- r can vary from -1 (perfect negative correlation) to 1 (perfect positive correlation)
- 0 indicates there is no correlation
- r is not to be confused with the slope of the line!
- Let's calculate r (it is the r in R^2 reported earlier):

```
r <- cor(bitcoin_price_search$price, log(bitcoin_price_search$search_index))
r
```

```
## [1] 0.7653157
```

- It appears we have reasonably strong positive correlation
- Let's see if it is statistically significant:
  - $H_0 : r = 0$
  - $H_A : r \neq 0$

```
CIr(r, n = 80, level = 0.95)
```

```
## [1] 0.6558882 0.8432395
```

- The 95% does not capture H_0, meaning it is rejected and r is statistically significant

Summary

- Linear regression summary
  - *Linearity: check*
  - *Normality of residuals: check*
  - *Homoscedasticity: check*
  - $r = 0.77$ *95% CI* $(0.66, 0.84), r^2 = 0.59$
  - *Model ANOVA,* $F(1, 78) = 110.3, p < 0.001$
  - $\alpha = -5863.9, p < 0.001,$ *95% CI* $(-7682.2, -4045.7)$
  - $\beta = 2363.0, p < 0.001,$ *95% CI* $(1915.0, 2811.0)$
- Resulting linear model equation:
  - $price = -5863.9 + 2363.0 \times log(search\ index)$
- Decisions:
  - *Overall model: reject* H_0
  - *Intercept: reject* H_0
  - *Slope: reject* H_0
- Conclusion:
  - *There was a statistically significant positive linear relationship between the Google Trends 'Interest over time' search index for the word 'cryptocurrency' and the price of Bitcoin in USD for the 80 days from 16/07/2017 to 03/10/2017*

Discussion

- Major findings:
  - *Internet search trends appear to be a reliable metric for predicting the price of Bitcoin*
  - *One could do deduce that as Google Trends 'Interest over time' index rose, so too would the price of Bitcoin in USD (and vice versa)*
  - *Of course, this does not mean that both variables will always rise and fall together as linear regression does not prove this*
- Strengths of analysis:
  - *The reliability of the data: it can be said with near certainty that the data was correct as correct Bitcoin prices in any currency are easily obtained and the search data was from Google itself*
  - *Recency of the data: the data set was up to 03/10/2017, less than two weeks before undertaking this analysis*
- Weaknesses of analysis:
  - *Slightly small data set: there were only 80 observations due to the Google only being able in daily increments for the last 90 days*
  - *Possibility of the relationship holding in future: as mentioned earlier, there appears to have been a fairly unique set of circumstances that resulted in this trend. This leads to the belief that things may change as things settle*
- In terms of futher work, it could certainly be the case that this phenomenon occurs in the future for other variables meaning other linear regression models could be created

References

- Course materials