UNIVERSITY OF COPENHAGEN
DEPARTMENT OF ECONOMICS
FACULTY OF SOCIAL SCIENCES

# BSc Thesis

Adam Frederik Ingwersen Linnemann

# Cryptocurrency Economics

## - A Demand Side Cointegration Analysis of Bitcoin Price Formation

Supervisor: Rasmus Søndergaard Pedersen
ECTS credits: 15 - Typed Units: 43.528 excl. whitespace - Normal Pages: 19,76
4th of May 2016

**Abstract**

This paper investigates demand mechanics in the new field of cryptocurrency economics. A multivariate time-series analysis of Bitcoin exchange rates against the US-dollar is provided. In particular, this paper considers possible cointegrating relations of web-based proxies for public interest such as search activity on Wikipedia and Google with the Bitcoin exchange rate.

The fundemental mechanics of supply in the Bitcoin economy reveal that Bitcoin supply is roughly constant - and expectations towards inflation are approximately static. Data collected using programmatic web-crawling techniques reveal several cointegrating interrelations, and it is found that cointegrating relations are inconsistent across a daily sample as compared to a weekly sample. In constructing an appropriate model using a dynamic single-equation error-correction model for a sample of daily observations it is argued, that no single-equation model specification can capture the complex mechanics at work due to multiple cointegrating relations within the regressors and regressand. It is found that a multivariate dynamic error-correction model can be applied in estimating a dynamically well-specified model for the exchange rate, employing weekly aggregated search queries on Wikipedia and weekly trade frequency as regressors. The proposed model suffers severe misspecification in terms of linearity and normality, and it is argued that the parameter estimates are at risk of being biased.

A thorough discussion of the issues which arise in describing the data using the proposed model is conducted - and suggestion towards more appropriate, advanced modelling techniques are argued to possibly yield a better description of the intricate relationships observed in the data.

# Contents

# 1 Introduction

The emergence of the Internet has spawned a multitude of technologies at a rapid pace. During the past ten years, we have become buyers of smartphones, users of Facebook and customers of Amazon. E-commerce and digital communication have altered the way we interact and think. Digital technologies have not only altered behaviour, but also had an immense impact on how real economy works. Commodities, stocks and cash change hands at the click of a button. Large quantities of real-time data is avaliable to market analysts. Although these technologies allow for accessible, transparent markets; whether buying shoes or financial derivatives - the sheer mass of information available makes it difficult for non-experts to fully utilize. As markets have evolved, so have the ways in which agents are able to pay for goods and services - the first major breakthrough in this area was PayPal; a service which allow users to utilize a streamlined payment protocol regardless of their credit card or bank. PayPal acts as a middle-man between any bank, costumer and seller. The most recent innovation in digital payment is Bitcoin.

The initial intention of Bitcoin was not to undermine services like PayPal, but rather to avoid the necessity of banks in the market for secure web-based payment. The complementary invention of the Blockchain allows for secure, direct transfers between any two parties - with transaction costs significantly below what banks charge. Besides the underlying technology which facilitates transactions, Bitcoin is also a pseudo-currency; it is a medium of exchange but not a store of value. Although cryptocurrencies in general seek independence from centrally governed currencies, they are not isolated from real economies. The drivers of both supply and demand in the Bitcoin paradigm have some real foundation. Since the conception of Bitcoin, the dollar-value has exhibited highly volatile behaviour by almost any standard. Various speculations about what caused these fluctuations have been proposed - many criticising the inherent instability of the Bitcoin protocol and the ungoverned nature of the currency. The economic indicators, which economists typically rely on in order to asses real value is not present in the context of Bitcoin.

A broad spectrum of users use the Bitcoin platform - some use the cryptocurrency as digital, untracable cash, whilst other agents seek a quick return on a risky investment. Others still use the currency for web-based trade of good and services exclusively. Due to the fact, that users are anonymous, money-flows are nearly impossible to trace, and there is little solid theoretical foundation on the subject. Here, explorative quantitative analysis could prove useful in understanding the mechanics at work. The aim of this paper is to identify possible interrelations of the exchange rate of Bitcoin with proxies for public interest regarding the currency based on web-searches.

## 1.1 Methodology

In pin-pointing variables, that might explain some variation in the dollar-value of Bitcoin, an economic analysis of the features of supply in the Bitcoin economy is conducted. From here, some speculative guesses towards the complicated mechanics of demand in the economy are made in - some covarying measures are presented, and a short description of the data collection process is presented. Having identified relevant data, the econometric tools for cointegration analysis are presented along with a brief discussion of the theoretical limitations. The theoretical section also outlines the circumstances under which, an optimal model can be constructed. Based on the theoretical framework, each variable is analysed and tested in order to ensure prescence of unit-roots. Upon having described the variables in the data, model construction is discussed - and a final model is presented, tested and evaluated.

## 2 Bitcoin from an economic perspective

Bitcoin is the most widespread and popular cryptocurrency. The fundemental ideas and algorithmic protocols behind Bitcoin and the underlying concepts of the Blockchain technology was developed by the alias Satoshi Nakamoto in 2009[1]. Bitcoin was not the first cryptocurrency, but the first decentralized public ledger currency platform, which is characterized by being without distinctive ownership, developed open-source with infrastucture maintenance and transaction processing being handled by peer-to-peer networking. As a cryptocurrency, Bitcoin utilizes a SSH-256 cryptographic protocol developed by the U.S. National Security Agency (NSA) to facilitate secure payment and proof of ownership - this is maintained in a public ledger known as the Blockchain. Any holding of Bitcoins can be stored as a private key, typically consisting of 256 hexadecimal digits. Tthe quantity of Bitcoins in circulation is controlled, and cannot be altered significantly by any single entity. The issuance of Bitcoins is controlled by a publicly viewable algorithm, that requires "mining" in order for new Bitcoins to be minted. Whoever participates in the mining process and provides computational power to the network will on average be rewarded accordingly by recieving newly minted Bitcoins - this is the only way in which new Bitcoins enter the economy. The total number of possible Bitcoins and the approximate rate at which they enter the system is defined a-priori and this information is publicly available.

### 2.1 Public ledger & the Blockchain

The Blockchain is a digital public ledger of the transaction history on the Bitcoin network. The name Blockchain stems from the fact that, when a number of transactions have been processed and validated by the network, they are contained within a block of X transactions. The blocks are added to the digital chain in a linear, chronological order - and the entire chain is stored on each miner's computer, called a node. The fact that each record contain a complete history of transactions and thus balances for each user, results in a system in which double spending is practically impossible. The Blockchain was invented alongside with Bitcoin - but these are two seperate technologies and the Blockchain technology could be implemented outside Bitcoin in a variety of ways[2]. The main pitfall of the Blockchain is that the validation and security, which the system requires, is an immense task in terms of computing resources. As the blockchain grows, more data has to be stored on each node on the network, thus increasing the computational burden over time. In the remainder of this paper, Blockchain will not be discussed explicitly

[1]Nakamoto, S. (2009). Bitcoin: A Peer-to-Peer Electronic Cash System

[2]Ali, R. & Barrdear, J. (2014). Innovations in payment technologies and the emergence of digital currencies

and the analysis is restricted to the seperate economic environment of Bitcoin.

### 2.2 Bitcoin as a currency

A currency is defined as having three functionalities: A medium of exchange, a unit of account and a store of value[3]. Initially government-issued currencies acted as representative money as they were directly linked to some other store of value such as gold or silver. The government-issued currencies that are available today - since the abandonment of the gold-standard, 1971 - are mostly fiat money in the sense that they are not worth their value in any other store of value. Essentially, fiat money has value due to the fact that the central bank or government guarantees its value. Fiat currencies such as the US-dollar is both printed as cash, and is also issued as digital currency. These conventional forms of currency, in most developed countires, serve as a liquid medium of exchange, a unit of account, and given stable inflation, a store of value.

The disctinction between digital and virtual currency[4] is as follows; digital currency is a type of currency which has some link to a real entity, but exists within computers - whilst stricly virtual currency also only exists in cyberspace it does not allow for trading real goods. Bitcoin is both digital and virtual in this sense; it only exists in cyberspace, but real costs are linked to maintaining the system and real goods can, with limitations, be purchased using Bitcoin[5]. Bitcoin serves to some extent as a medium of exchage on the web - internet businesses are increasingly accepting Bitcoin as payment. Bitcoin is could potentially serve as a unit of account - although this is seldomly the case; few individuals would consider denominating tax-calculations in Bitcoins. As a store of value, Bitcoin is very volatile, and does not serve as a proper alternative to currency in countries with well-established central banks and stable inflation.

### 2.3 Supply mechanics

New Bitcoins are added to circulation at every completed block. These Bitcoins are awarded to the node, that solved a particular combinatorial problem. The reward of mining is diminishing - the reward of completing a block is halved with the completion of every 210.000 blocks. For the first time since 2012, in June 2016, the block reward will be reduced from 25BTC to 12,5BTC[6]. This mechanic was built in to the system in order to control supply and thus inflation. Whilst the incentive for miners to continue mining will diminish if the price of Bitcoin does not rise accordingly, the protocol allows for users of Bitcoin to add a small fee in order for miners to handle transactions swiftly, which may keep miners incentivised to supply their computing power. Furthermore, the drop in block reward has been anticipated for a long time - and as such is incorporated into the BTC price. In understanding the factors determining supply, one should look towards the incentives of the miners - as this is the only way in which Bitcoins

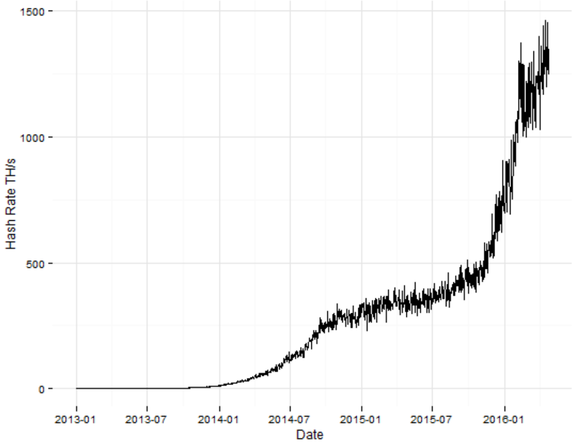
[3] Mankiw, Gregory (2007). Principles of Economics
[4] Wagner, A. (2014). Digital vs. Virtual Currencies
[5] Kancs, C. et al. (2015). The Digital Agenda of Virtual Currencies: Can BitCoin Become a Global Currency?
[6] Estimated by http://bitcoinblockhalf.com

can be added to the Bitcoin ecosystem. The Bitcoin miner can be thought of as solving an optimization problem, in which he has to make expectations about the future prices of energy, computing power, difficulty of the combinatorial problem and value of Bitcoin.

### 2.3.1 Bitcoin mining

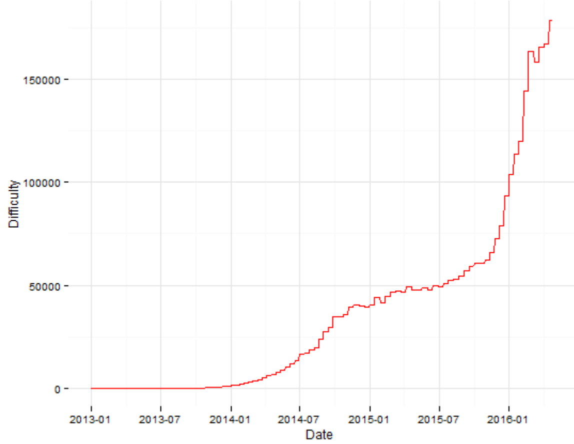
Mining is in practice done by a relatively small group of individuals and organizations with large amounts of computing power. Bitcoin mining serves as a way to maintain the underlying system.

Figure 2.1: Computing Power on The Bitcoin Network




Source: http://blockchain.info

Figure 2.2: Difficulty of Cryptographic Hash




Source: http://blockchain.info

Mining is conducted by a variety of individuals, collectives (known as pools), and businesses. In the past two years the total computing power on the Bitcoin network has increased drastically. Whilst computing power on nodes in the network is crucial for facilitating transactions, an increase in power will, by construct, be complemented by an increased difficulty of the cryptographic hash, which the computer has to solve. Effectivly the relative measure between the hash rate and difficulty is by approximation a stationary process[7]. This suggests, that the supply is stable. Any increase in the amount of computing power does not imply that the number of Bitcoin put into circulation will increase. This is due to built-in stability features, which evaluates the speed at which 2.016 blocks are solved - if these are solved faster than 2 weeks, the underlying algorithm corrects the difficulty of the cryptographic hash, such that the speed is approximately constant.

Computers' processing power is increasing over time - this is also known as Moore's Law, which states that the number of transistors fitted onto a computer chip doubles every two years, in turn implying more efficient computing. Moore's Law has shown to be a good approximation of the evolution of CPU-power, empirically. Engineers and coumputer scientists have figured out elaborate ways to efficiently utilize GPU's (Graphics Processing Units) for mining instead of CPU's, which is faster. The transition towards GPU's instead of CPU's may have sped up the increase in network hash rates at a faster pace than Moore's law dictates.

With a diminishing mining reward and an approximately constant rate of block solving, the rate at which Bitcoins enter the economy is by construction diminishing - this is inteded,

[7]See Figure A.3 in Appendix

as reliable and predictable supply is necessary in order not to skew the holdings of Bitcoin in favor of the miners. The rate at which new Bitcoin are minted is stable except for the reward drop, which occurs approximately every 4 years. This is expected. As a result bitcoin supply is almost perfectly inelastic and does not respond significantly to exogenous factors - the number of Bitcoins in circulation have thus far increased in a linear fashion.

## 2.4 Demand mechanics

The Bitcoin economy does in many ways rely on the surrounding economic environment[8] - but Bitcoin does not directly respond to any known demand-side stimuli from the outside environment. Due to the fact that Bitcoin is not isolated from the outside economic environment - it is possible that real or digitally based drivers for demand exist. Due to inelastic supply, the price of Bitcoin is primarily determined by demand for the currency[9]. Possible drivers are presented below.

### 2.4.1 Trade activity

Monitoring Bitcoin markets for user-level activity proves a difficult task due to the inherent anonymity of the system. One approach is to look for trading activity on different nodes of the network; this could however prove misleading as one user may have multiple active nodes, and multiple users may share one node. The cummulative trade activity can however easily be accessed as the entire blockchain is public. Inspecting the evolution of daily trades conducted on the Bitcoin Platform[10], some fluctuations occur - and at several points in time, trading largely exceeds the trend level. Noticable, is a clear upward trend in the number of trades conducted on any given day since the start of 2013. This increase could be explained by more users. Alternatively, this could be the result of a general increase in attention from speculants, who trade more frequently for financial gains.
Whilst the idea of increasing trade activity being a proxy for overall use of Bitcoin is an appealing one, it should be noted that investments linked to Bitcoin has become more popular as e.g. hedgefunds have incorporated minor ammounts of Bitcoin in their portfolios in order to e.g. short the Chinese Yuan. Albeit, the amount of trades conducted over the past three years have more than doubled, which in some sense may be an expression of the Bitcoin market maturing and in turn an indication that market liquidity has increased. Inspecting the daily traded volume[11] on the platform yields different insight; traded volume seems to revert to some mean, with four major spikes in the traded volume, which seems to correspond well to the four large, persistent spikes observed in the USD/BTC exchange rate[12], in April 2013, November and December 2013, June 2014 and November 2015, respectively. This would seem to suggest that the total value of trade conducted as measured in units of Bitcoin has not increased. The increase in trade

[8] Ali, R. & Barrdear, J (2014). The economics of digital currencies
[9] Kancs et al. (2014). The Economics of BitCoin Price Formation
[10] Appendix, Figure A.1
[11] Appendix, Figure A.2
[12] Appendix A, Figure A.4

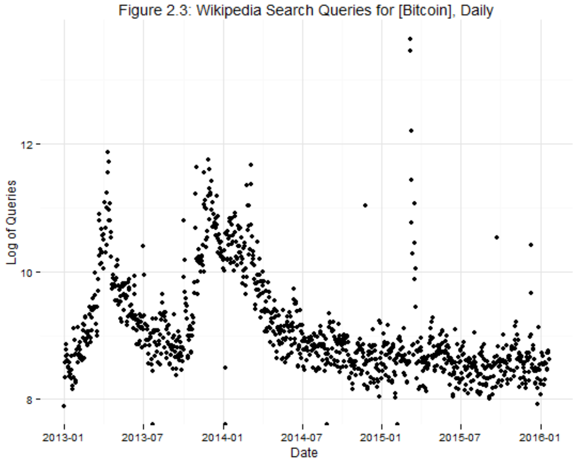
frequency may indicate one of two things; either it is becoming more widely accepted as a means of payment in retail and internet services; individuals are conducting more types of regular trade on the platform. Alternatively, investors may be using Bitcoin for high-frequency trading as the platform becomes more agile. The latter does not seem plausible due to the fact that transactions are, on average, processed within 5-10 minutes and do not hold up to the standards of the 21st century fincancial sector in terms of transaction speed. It does not seem entirely implausible that the acceptance of Bitcoin has become more widespread. Microtransactions in crowdfunding, digital art and journalism are becoming increasingly popular. Bitcoins may also serve as an internationally acceptable currency for creative minds, who rely payment in the form of pay-per-view. Bitcoin has the ability to facilitate such transactions, as they are without costs to the user - as opposed to e.g. bank-transactions. Minor transactions from one end of the world to another is not facilitated by regular transaction services, and transferring larger amounts is costly.

### 2.4.2 Proxies for public interest

As seen above, Bitcoin derives some value from the liquidity and obviously from any increase in widespread usability. Besides these factors, researchers in multiple fields have suggested that search queries on web-based services such as Google and Wikipedia may prove a useful indicator of overall public interest that can be linked to real entities[13][14]. Public interest may serve as a usful proxy; it is plausible that any measure of searches of a topic on e.g. Wikipedia captures both general movements in interest, but also occurences of related news articles and trends on social media. As such, search engine queries are very rough measures of public interest, but could likely capture large movements due to e.g. significant increases in news reporting.

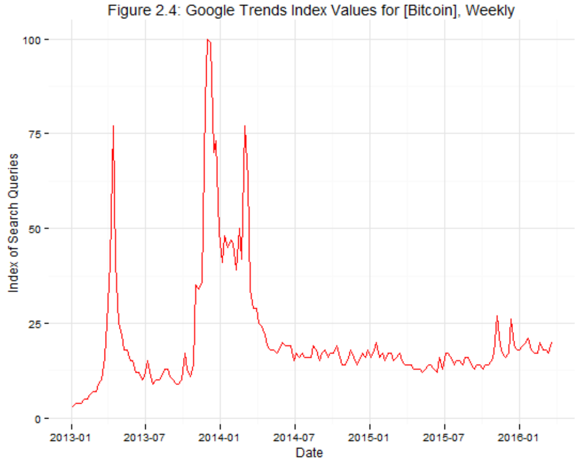
If search queries capture overall movements in public interest, one would expect them to be impacted by the same underlying data-generating process.

Figure 2.3: Wikipedia Search Queries for [Bitcoin], Daily




Source: http://stats.grok.se

Figure 2.4: Google Trends Index Values for [Bitcoin], Weekly




Source: https://google.trends.com

Figures 2.3 and 2.4 exhibits common features, apart from a surge in Wikipedia searches in March 2015 - the Wikipedia article for Bitcoin was shown 847.614 March 8th 2015. Also, a noticable

[13] Kristoufek, L. (2013). BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era

[14] Varian, H. & Choi, H. (2009). Predicting the Present with Google Trends

commonality of the Public Interest proxies appear when comparing with the trade volume in Figure A.2. The apparent commonality suggests that the series may be subject to influence of a common underlying data-generating process. In particular, inspecting the three major increases in both the Wikipedia search queries and Google Trends, these seem to co-move with the actual USD/BTC exchange rate over time.

## 3 Data Collection

The aim of this section is to gather data which may help with understand the evolution in Bitcoin prices. The data is gathered from seperate websites with asynchronized time-stamps which has had to be aligned. In completing this task, intermediate scripting in the programming language R has proven useful. The data is as reported - and a further discussion of the validity will not be undertaken.

The time-series of primary interest are the Bitcoin Exchange Rates as reported on http://coindesk.com, an index of search frequencies for "Bitcoin" on Google as well as the weekly number of search queries on the Wikipedia page for Bitcoin. The motivation for using web-search data, is that this have proven, in a variety of cases, to capture major changes in attention amongst the public towards a particular topic[15]. The samples have been shortened to only include data from 2013-01-01 until 2016-01-04 for the sake of consistency of the samples.

### 3.1 Bitcoin exchange rates

The Bitcoin exchange rates are reported on CoinDesk, a website, which provides exchange rates from the 5 most popular exchanges (itBit BPI, BitStamp, BitFinEx, CoinBase, OKCoin). The data provided by CoinBase is an average of the bid/ask spread of the 5 exchanges. The CoinBase exchange rate is not volume-weighted due to the historical lack of liquidity on the Bitcoin exchange market. Due to the fact that the potential cointegrating relationship between exchange rates and primarily western search-engines are considered, the Bitcoin exchange rate against the Chinese Yuan, which is the second-largest, is disregarded as it is not deemed sensible to justify, that demand drivers are common across these economies. Furthermore, recent research suggest, that Bitcoin exchange rates abide by the same arbitrage-laws, that hold in conventional exchange-rate markets[16] - it is found that despite some illiquidity in the market, relative Bitcoin prices adjust rapidly to restore the exchange-rate parity and that market efficiency is maintained by a deep and active pool of traders.

[15] Garcia et. al (2014). The Digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy

[16] Smith, J. (2015). An Analysis of Bitcoin Exchange Rates

### 3.2 Public interest proxies

Data from Wikipedia was gathered from a third party website [17], which has actively mirrored the Wikipedia.org website since 2007. The website provides daily data on the absoulte count of search queries to english Wikipedia articles. The time series, which is of interest is the historical number of visits to main article for Bitcoin on Wikipedia.org. The procedure of scraping the website has been documented thoroughly and is publicly accessible[18]. The raw data reports the daily number of visits to the main Wikipedia page for Bitcoin.
Data from Google was been collected using Google's Trend service, which provides an index for searches including the term "Bitcoin", and is as such a normalized measure, compared to the Wikipedia search queries. The procedure of fetching Google data is explained and suitable for reproduction[19]. Issues does ensue in analysing the Google Trends data as the reported index is measured in weekly averages - and as such reduces sample size and obfuscates fluctuations.

In utilizing the samples fully, two data sets are constructed; one includes daily data from all sources providing daily data - Wikipedia, Exchange Rate, Trade Volume & Trade Frequency. The other is aggregated data such that the sample reports weekly data including the Google Trends index. The weekly data cointains 152 observations. The daily data contains 1098 observations.

## 4 Cointegration Analysis

Having identified some variables of interest, the tools necessary for an econometric analysis should be discussed. The particular case of interest is when two or more variables are cointegrated, such that they have a common, or partly common underlying process. The interrelation of such series can be defined as one or multiple linear combinations of parameter values. For series to cointegrate, the DGP must be equivalent in order of integration - cointegration analysis can exclusively be conducted of a regressor with an underlying I(1) process, if the regressand is also I(1), also known as a unit root. In order to assess the prescence of unit-roots in the time-series data, a formal test is needed. The approaches described in the succeeding sections rely on correct dynamic specification. Ensuring satisfactory dynamic specifcation relies on general-to-specific stepwise backwards selection of lagged regressors.

### 4.1 Unit Root Testing

Intermediate econometric timeseries analysis is primarily concerned with stationary time-series[20]. In many applications, series exhibit unit root behaviour - an I(1) process. A unit root process of order one is a process in which the characteristic polynomial has 1 as the solution to it's characteristic equation. Unit root processes behave differently compared to stationary processes - most noticable, shocks have permanent effects and do not diminish or revert back to a mean. It is possible to identify a unit root process by conducting Dickey-Fuller testing of a univariate

[17] http://stats.grok.se/
[18] https://github.com/adamingwersen/CryptoCurrencyEconomics/blob/master/Wikipedia_STATS.R
[19] https://github.com/adamingwersen/CryptoCurrencyEconomics/blob/master/Bitcoin_GoogleTrends.R
[20] Verbeek, M. (2012). A Modern Guide To Econometrics, p.p. 282

autoregressive model. In the case of an AR-model of e.g. 3 lags, the test can be described as follows:

$$y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \theta_3 y_{t-3} + \epsilon_t$$

This equation is characterized as having a unit root in the case where $\theta(z) = 1 - \theta_1 z - \theta_2 z^2 - \theta_3 z^3$, which corresponds to $\theta(1) = 0$. The model can be rewritten as the augmented dickey-fuller specification:

$$\Delta y_t = \pi y_{t-1} + c_1 \Delta y_{t-1} + c_2 \Delta y_{t-2} + \epsilon_t$$

with a null hypothesis of $H_0 : \pi = \theta_1 + \theta_2 + \theta_3 - 1 = 0$ against $H_A : -2 < \pi < 0$. The relevant critical values for $\pi$ are derived from the DF-distribution.

In the prescence of a unit root, the behaviour of constant and trend terms does not behave normally over time: Both constants and trends change the asymptotics of the DF distribution towards the left - and the critical values for either a trend or constant are smaller. Therefore special augmentations of the DF distribution have been constructed in order to compensate for this. It should also be noted, that null-hypotheses, to which the alternative is stationarity, are not naturally extended. The relevant alternative hypotheses for a unit root process with no constant, is a stationary process with a constant; the alternative hypothesis for a unit root with drift(constant) is a stationary series with a trend.

Testing the prescence of e.g a constant term, a Likelihood-Ratio test for the restricted and unrestricted version of the model should be undertaken. The relevant test-statistics are the squared DF-critical values for constant or trend.

### 4.2 Cointegration and Error Correction

If two time-series exhibit properties of a unit root, their potential cointegrating relationship can be explored. A cointegrating relationship between two I(1) variables is present if a linear transformation of the two variables results in a stationary process. This linear transformation is known as the cointegrating vector $\beta = (1, -\beta_2)'$ in the model

$$y_t = \alpha + \beta_2 x_t + \epsilon_t$$

Cointegration of two variables imply, that the random-walk component in the two seperate unit-root processes cancel each other out when including the cointegration vector. The residuals of the above stated regression model is an I(0) process if the $x_t$ and $y_t$are in fact cointegrated and I(1).

If there is only one cointegrating relationship, and this cointegrating relationship is assumed not to be bi-directional, so that $y_t$ does not provide feedback into $x_t$, a single-equation model can succesfully describe the cointegrating relationship. Assessing whether two variables cointegrate can be done by using one of two approaches. Both rely on the proper testing of unit roots, as described in section 4.1.

### 4.2.1 Engle-Granger Two-step procedure for Static Regression Models

In most applied situations, the cointegrating vector, $\beta$, is unknown - the solution here is to estimate it, which can be done by OLS. $\hat{\beta}_2$ is consistent for the true parameter, $\beta_2$, due to the fact that the variance of $\hat{\beta}_2$ converges to zero as $T$ increases - this convergence is faster for cointegrated series of integrated order 1, even when accounting for omitted dynamics. This phenomenon is known as super-consistency.
Testing for no-cointegration is essential to test the prescence of a cointegration vector - this can be done by estimating whether the residuals of the equation stated above contain a unit root. If we define $\hat{\epsilon}_t = \hat{u}_t$, and note, that we can write the error-correction term, as the deviation from an equilibrium given the cointegration vector:

$$\hat{u}_t = y_t - \hat{\alpha} - \hat{\beta}x_t$$

If the series are cointegrated, $\hat{u}_t$is a stationary proces. The relevant test-statistics for $H_0$of a unit root in $\hat{u}$ can be estimated by running the regression in ADF-specification:

$$\Delta\hat{u}_t = \pi u_{t-1} + \sum_{i=1}^{k-1} c_i \Delta\hat{u}_{t-i} + \eta_t$$

The parameter-estimates are DF-distributed. The null-hypothesis is again $H_0 : \pi = 0$, i.e. a unit root - if the hypothesis of a unit root is rejected, then one cannot also reject that the process is stationary, and thus, the series must cointegrate, c.f. the alternative hypothesis of $H_A : \pi < 0$. Given that $\hat{u}_t$is stationary, $\hat{u}_{t-1}$ can be incorporated into an error correction model as a fixed regressor variable - and the following error correction model can be estimated:

$$\Delta y_t = \delta + \lambda_1 \Delta y_{t-1} + \kappa_0 \Delta x_t + \kappa_1 \Delta x_{t-1} + \alpha\hat{u}_{t-1} + \epsilon_t$$

This regression can be consistently estimated by OLS. All parameter estimates are stationary, so normal inference applies - and the parameter estimates' t-ratios follow standard normal distributions asymptotically. The main drawback of the Engle-Granger Two Step procedure is, that if a dynamic model better represents the DGP, this approach is insufficient, as no dynamic terms are included in the estimation of the cointegrating vector, $\beta$ - one would expect the series considered to exhibit some autocorrelation. Thus, the approach proposed by Engle and Granger may omit important dynamic features of the DGP.

### 4.2.2 Dynamic Regression and the Autoregressive Distributed Lag Model

An appropriate alternative to the Enge-Granger Two Step procedure is specifying a dynamic model, which captures the dynamic structure of an autoregressive unit root series. Provided that the variables in question are I(1), the approach suggested below could provide better representation of the cointegrating relationship and thus error-correction term. If a properly specified Autoregressive Distributed Lag (ADL) model can be specified, such that lag-lengths are set to eliminate residual autocorrelation, this approach should be preferred over the Engle-Granger approach. Suppose both variables should have two lags, e.g. an ADL(2,2) model:

$$y_t = \delta + \theta_1 y_{t-1} + \theta_2 y_{t-2} + \phi_0 x_t + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \epsilon_t$$

The unrestricted ADL model can be reformulated to an ECM by taking first differences and rewriting:

$$\Delta y_t = \delta + \lambda_1 \Delta y_{t-1} + \kappa_0 \Delta x_t + \kappa_1 \Delta x_{t-1} + \gamma_1 y_{t-1} + \gamma_2 x_{t-1} + \epsilon_t$$

With $\gamma_1$ being the adjustment parameter. Here, the cointegrating coefficient is given by the long-run solution i.e. the estimated parameters of the non-stationary terms:

$$\hat{\beta}_2 = \frac{\hat{\phi}_0 + \hat{\phi}_1 + \hat{\phi}_2}{1 - \hat{\theta}_1 - \hat{\theta}_2} = -\frac{\hat{\gamma}_2}{\hat{\gamma}_1}$$

Given cointegration, $\hat{u_t} = y_{t-1} - \mu - \beta_2 x_{t-1}$ , with $\mu = -\delta/\gamma_1$, is an error-correcting term. The equation above can be modeled in general as the ECM(p,q) of an ADL(p,q)[21].

$$\Delta y_t = \delta + \sum_{i=1}^{p} \lambda_i \Delta y_{t-i} + \sum_{j=0}^{q} \kappa_j \Delta x_{t-j} + \gamma_1 \hat{u}_{t-1} + \epsilon_t$$

Standard inference apply to some parameters - defining $\Psi$ as a vector of the stationary parameters' t-values for e.g. $p = q = 1$, i.e. 1 lag of the first differences of $x_t$ and $y_t$:

$$\Psi \equiv \begin{pmatrix} \lambda_1 & \kappa_0 & \kappa_1 \end{pmatrix} \overset{a}{\sim} N(0,1)$$

The ADL and ECM are different representations of the same underlying model, therefore there is a 1-to-1 correspondence in the model in terms of parameter values, which have been rewritten as stated above. Both representations are suitable for testing no-cointegration and identifying the parameter estimates $\hat{\beta}_2$, $\gamma_1$. Testing for no-cointegration in the ECM specification is conducted by constructing the so-called PcGive test for no-cointegration, with a null hypothesis of no-cointegration: $H_0 : \gamma_1 = 0$ and an alternative of $H_A : \gamma_1 < 0$. The critical values of the PcGive distribution depends on the number of regressors included in the model and is closely related to the Dickey-Fuller distribution. The relevant test-statistic is computed as $\hat{t}_{\gamma_1=0} = \frac{\hat{\gamma}_1}{se(\hat{\gamma}_1)}$, which is simply the t-statistic.
Explicit assumptions has to be made when estimating the ECM proposed above: 1) The explanatory variable in the ECM is determined conditionally on the regressor(s), which by construction results in the regressors being predetermined, exogenous and the model does as such only facilitates a one-way correspondence from $x_t$ to $y_t$ - i.e. no feedback is assumed. 2) It is also important to note, that the approach described in this section assumes that only $y_t$ cointegrates, and therefore does not allow for multiple equation systems with more than one cointegrating relation. The approach readily extends to multiple regressors - however, as more regressors are included in the model, the assumption of only one cointegrating relationship and no feedback becomes gradually more unlikely.

## 5 Variables

In this section, the time-series properties of each variable used in modelling is described. The process of describing the time-series properties of each variable involves testing for unit roots and cointegrating relationships.

[21] Bohn H. (2012). Econometrics II Lecture Note: Cointegration and Common Trends

## 5.1 Regressand

In explaining movments in the Bitcoin exchange rate as described in section 4.1 by a cointegrating or error-correcting relationship with explanatory variables such as public interest proxies or trade activity/volume, it is essential that a unit root of equivalent order is present. This subsection investigates the properties of the Bitcoin exchange rate variable and shows that a unit root is indeed present by conducting Augmented Dickey-Fuller (ADF) testing and Likelihood-Ratio testing[22]. The $rate_t$ -variable is log-transformed as large level fluctuations occur - and in order to normalize the time-series without flattening or obscuring patterns, the log-transformation is found to be useful.

Table 1: Descriptive Statistics of ADF-tests on daily $rate_t$ and $log(rate_t)$

| Rate AR(3) | $\pi rate_{t-1}$ | $\pi l(rate_{t-1})$ | $\pi rate_{t-1} + \delta$ | $\pi l(rate_{t-1}) + \delta$ | $\pi rate_{t-1} + \delta + \gamma t$ | $\pi l(rate_{t-1}) + \delta + \gamma t$ |
|---|---|---|---|---|---|---|
| Coefficient, $\pi$ [t-value] | -0,001 [-0,227] | 0,000 [1,43] | -0.005 [-0,929] | -0,007 [-3,75]*** | -0,005 [-0,831] | -0,006 [-2,79] |
| Log. Likelihood | -4822,04 | 1661,26 | -4820,33 | 1669,29 | -4820,29 | 1669,55 |

* 10% Critical Value, ** 5% Critical Value, *** 1% Critical Value, DF-test

The time-series appear to exhibit non-normality of error-terms, but is otherwise well-specified. It is found that $H_0 : \pi = 0$ is not rejected on a 1%-confidence inteval in the ADF-distribution, with a t-value of -0,227 over the alternative of $H_A : -2 < \pi < 0$ Thus, the prescence of a unit root cannot be rejected and testing for a trend and/or constant is conducted as follows:

$$LR_c(\pi = \delta = 0) : -2(LogLik_U - LogLik_c) = -2(-4822,04 - (-4820,33)) = 3,42 \sim DF_c^2(5\% = 9,13)$$

$$LR_t(\pi = \gamma = 0) : -2(LogLik_c - LogLik_t) = -2(-4820,33 - (-4820,29)) = 0,08 \sim DF_t^2(5\% = 12,39)$$

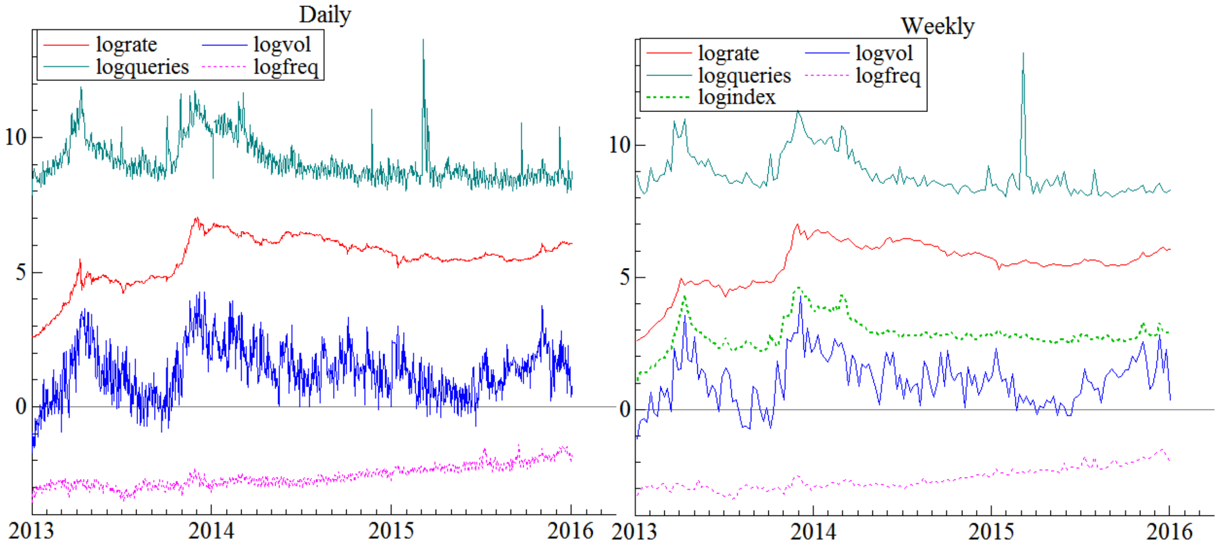
The Likelihood-Ratio tests show, that the null cannot be rejected and thus there is no evidence in favor of a trend or constant in $rate_t$. For the weekly data, the same procedure has been conducted[23]. It is found that, in the weekly data, the null of a unit root cannot be rejected in $rate_t$ or $log(rate_t)$. The null of either ($H_0 : \pi = \delta = 0$ or $H_0 \pi = \gamma = 0$) in $rate_t$, cannot be rejected. For the $log(rate_t)$ variable, the null of no-constant and a unit root is rejected at the 5% critical level. This implies, that the unit-root process is a random walk with drift.

## 5.2 Regressors

Testing for unit roots in each variable ensures that the subsequent cointegration analysis is valid. Using an identical approach as described above, tests for non-stationarity in each variable across samples is conducted and the essential results are presented. Figure 5.1 illustrates the log-transformed time-series of daily and weekly observations, respectively.

[22] Appendix B
[23] Appendix B.1

Figure 5.1: Log-transformed series




Sources: https://google.trends.com, http://stats.grok.se, https://coindesk.com, https://blockchain.info

### 5.2.1 Outliers

The log-transformed series presented in Figure 5.1 does not in general depict obvious outliers - other than a significant increase $log(queries_t)$ during the span of 6 days; March 7th to March 12th in 2015. One such observation should under normal circumstances be explicitly modelled as a dummy - however as this increase is persistent and strictly decreasing in the period, starting on the 8th of March 2015, it is difficult to present an argument which would justify exlcuding these observations based on them being unlikely or simply a data error. That being said; having considered the features of the series, it is clear, that any model would not succeed in describing the particular period - and in order to obtain the best possible model for the data in general, any models including $queries_t$, will explicitly have these features explicitly modeled using dummy-variables in the remainder of the paper.

### 5.2.2 Sample of daily observations

Within the data sample of daily observations, three possible cointegrating variables have been seleceted. Firstly, suitable lags are included in the ADF-specification by including lags according to Autocorrelation Functions into AR-models of various lag-length. From here, using general-to-specific backwards selection have been identifiied and written as AR(p)-models. Having specified an appropriate AR-model for each univariate model, the ADF-specification of each model have been estimated - standard log-transformations are included. The reported test-statistics[24] are compared to the relevant Dickey-Fuller distributions, with or without trend/constants, and evaluated at the 5% critical values. It is immideately found that $H_0$ of a unit root cannot be

[24] Appendix B.2, Table 6

rejected for $queries_t$, $log(queries_t)$, $frequency_t$ and $log(frequency_t)$, whilst both $volume_t$ and $log(volume_t)$ appear to be non-I(1) series, and thereby the null is rejected in favor of the $H_A$ of stationarity. Non-unit root variables will be disregarded for cointegration analysis.

Having identified the time-series which exhibit unit root properties, LR-tests for the inclusion of trends and constants are conducted[25]. In all but one LR-test, the hypothesis of either $LR(\pi = \delta = 0)$ and $LR(\pi = \gamma = 0)$ is not rejected, and thus, it cannot be concluded that the trend or constant term is equal to zero at a 5% confidence interval - alas these should not be included. In the case of the $log(frequency_t)$ variable, $H_0 : LR(\pi = \gamma = 0)$ is rejected, and given the LR-test, both a trend and constant term should be included, as the series exhibit both a linear- and quadratic trend in the prescence of the unit root. Here it should be noted, that the ADF-test for a unit-root is significant at the 10% critical value, and such, this conclusion is stated with some degree of uncertainty as to the prescence of a unit root when a trend and constant is included.

### 5.2.3 Sample of weekly observations

It appears, that the regressors based on weekly data[26] take on less significant values of $\pi$ as compared to the daily data. The only regressor, which one might deem stationary and thus not a unit root is the $log(volume_t)$, which appears to be significant at the 10% level. LR-test values have been computed[27]. The LR-test values does not appear significant at the 5% confidence interval, and thus, we reject that any of the unit-root processes should be modeled with either drift or linear and quadratic trend.

# 6 Modelling

This section is concerned with constructing well-specified models given the data, and relies on the single-step cointegration analysis based on ADL-to-ECM approach described in section 3.2.2. This is due to the fact, that most of the regressors - and the regressor appear to exhibit dynamic properties and these processes are best modeled using multiple lags - which the Engle-Granger Two Step procedure does not allow for. In order to ensure robust results, both the weekly and daily data should be modeled, such that any deviating results should compromise the confidence in the chosen model and a less concise conclusion can be drawn from the analysis. As the chosen approach is restrictive in terms of the assumptions necessary for optimality, special attention should be given to the number of cointegrating relationships modelled - it is assumed that there only exist one cointegrating relationship in any proposed model, and one drawback from this is that the approach does not allow feedback between the regressors.

[25] Appendix B.3, Table 8
[26] Appendix B.2, Table 7
[27] Appendix B.3, Table 8

## 6.1 Specification

It proves difficult to model the raw $rate_t$ variable as the lags are very persistent and over 20 lags should be included in the ADL-model if one wishes to model the dynamic structure of $rate_t$ properly. Instead modelling the log-transformed variable proves more sensible - and in order to provide meaningful interpretations, the log-transformation of all other variables, expcept $Index_t$ are used to construct models in this section. In selecting a model, one must first assess the possible interrelations of the variables. In this particular case, no intiutive explanation can be given as to why any of the variables should or should not be cointegrated - without any formal testing, that is. Therefore tests for no-cointegration of well-specified ADL(p,q) models have to be computed, in order to identify cointegrating relationships. In order to determine, whether or not the assumptions, stated in section 4.2.2 are feasible, the ADL/ECM's should be tested with interchangeable regressors and regressands to ensure, that there are not more than one cointegrating relation.

### 6.1.1 Model Specification and Cointegration Testing based on daily data

Significant test-values shown in Table 2 depicts cointegrating relationships. The strongest cointegrating relationship in terms of signficance is $log(freq_t)$, which cointegrates on $log(rate_t)$. This is a two-way relationship - the variables cointegrate with each other. The effect from $log(freq_t)$ is on $log(rate_t)$ and vice-versa is therefore unclear. There is also a two-way cointegrating relationship from the daily traded volume to the exchange rate - here it is, too, difficult to grasp the mechanics with the approach of single-equation test for no-cointegration as elaborated in section 3.2.2. No cointegration relation between $log(rate_t)$ and $log(queries_t)$ is found. Upon excluding $log(rate_t)$ from further cointegration analysis, it is postulated, that no concurrent model can be set up to explain price movements in the USD/BTC exchange rate using daily data. In conclusion, none of the models proposed in Table 2 would be sufficient nor optimal in a single-step cointegration analysis given the data.

Table 2: Cointegration Test Values, Daily Data

| Regressand $\rightarrow$ | $log(rate_t)$ | $log(queries_t)$ | $log(freq_t)$ | $log(vol_t)$ |
|---|---|---|---|---|
| $log(rate_t)$ | ... | ADL(5, 3): $\gamma_1$=-1,88 | ADL(3, 2): $\gamma_1$=-3,45** | ADL(8, 7): $\gamma_1$=-4,92*** |
| $log(queries_t)$ | ADL(4, 4): $\gamma_1$=-2,34 | ... | ADL(6, 2): $\gamma_1$=-2,36 | ADL(7, 5): $\gamma_1$=-3,97*** |
| $log(freq_t)$ | ADL(4, 7): $\gamma_1$=-3,95*** | ADL(5, 4): $\gamma_1$=-2,05 | ... | ADL(8, 5): $\gamma_1$=-1,34 |
| $log(vol_t)$ | ADL(6, 4): $\gamma_1$=-3,86*** | ADL(8, 5): $\gamma_1$=-1,34 | ADL(7, 1): $\gamma_1$=-1,80 | ... |

* 10% Critical Value, ** 5% Critical Value, *** 1% Critical Value, PcGive-test with 2 variables

### 6.1.2 Model Specification and Cointegration Testing based on weekly data

Given the statistics in Table 3, it is clear, that some cointegrating relations exist. The strongest, in terms of significance, is the error correction from $log(vol_t)$ to $log(rate_t)$. The traded volume as measured in dollars is, by construction, dependent on the exchange rate; as prices rise, the dollar-value of accumuluated daily trades will as well - this is clear when inspecting Figure 2.4 and Figure A.2. The cointegrating relation between these variables are not interesting as such, due to the fact traded volume directly depends on the exchange rate. It is also apparent, that $Index_t$ has a cointegrating relation with $rate_t$ - but direct infrerence may be difficult, as the single-equation ADL-to-ECM approach does not allow for feedback in the error-correction model, and therfore may be inclined to over-simplify the relationship between these variables. The most interesting relation in the weekly data, is the cointegrating relation for $log(rate_t)$ on $log(queries_t)$, which appears, at the 5% significance level to be a one-way relation. The same goes for $log(freq_t)$, which does not appear cointegrate with $log(queries_t)$.

Table 3: Cointegration Test Values: Weekly Data

| Regressand→ | $log(rate_t)$ | $log(queries_t)$ | $Index_t$ | $log(freq_t)$ | $log(vol_t)$ |
|---|---|---|---|---|---|
| $log(rate_t)$ | ... | ADL(4, 3): $\gamma_1$=-3,10* | ADL(1, 3): $\gamma_1$=-5,50*** | ADL(3, 5): $\gamma_1$=-0,40 | ADL(3, 3): $\gamma_1$=-4,13*** |
| $log(queries_t)$ | ADL(1, 3): $\gamma_1$= -3,21** | ... | ADL(1, 0): $\gamma_1$=-5,88*** | ADL(5, 3): $\gamma_1$=-0,27 | ADL(2, 4): $\gamma_1$=-4,37*** |
| $Index_t$ | ADL(1, 2): $\gamma_1$=-3,50** | ADL(2, 1): $\gamma_1$=-5,17*** | ... | ADL(5, 2): $\gamma_1$=-0,01 | ADL(2, 1): $\gamma_1$=-5,73*** |
| $log(freq_t)$ | ADL(4, 5): $\gamma_1$=-3,08* | ADL(3, 4): $\gamma_1$=-2,97* | ADL(1, 4): $\gamma_1$=-3,70** | ... | ADL(2, 2): $\gamma_1$=3,60** |
| $log(vol_t)$ | ADL(1, 1): $\gamma_1$=-4,04*** | ADL(3, 1): $\gamma_1$=-3,04* | ADL(1, 4): $\gamma_1$=-4,02*** | ADL(5, 1): $\gamma_1$=0,179 | ... |

* 10% Critical Value, ** 5% Critical Value, *** 1% Critical Value, PcGive-test with 2 variables

Given that any preferrable model should not include variables, which exhibit feedback mechanics, an appropriate model should consider variables, which only has a one-way cointegrating relationship with $rate_t$. One such model could include 1 or more regressors. Any of the models in Table 3 could be exmined further, but an ADL(1, 3, 5)-model which include both $log(queries_t)$ and $log(freq_t)$ as regressors is proposed. This model can be represented by an ECM, which should include a constant:

$$\Delta l(rate_t) = \delta + \sum_{i=0}^{3} \kappa_i \Delta l(queries_{t-i}) + \sum_{i=0}^{5} \iota_i \Delta l(freq_{t-i}) + \gamma_1 l(rate_{t-1}) + \gamma_2 l(queries_{t-1}) + \gamma_3 l(freq_{t-1}) + D_M + \epsilon_t \tag{6.1}$$

The model proposed above depends on two other, possibly cointegrating, relationships. The relevant ECM's to consider and estimate to ensure no-cointegration are:

$$\Delta l(queries_t) = \delta + \sum_{i=1}^{3} \lambda_i \Delta l(queries_{t-i}) + \kappa_0 \Delta l(freq_t) + \sum_{i=0}^{3} \iota_i \Delta l(rate_{t-i}) + \gamma_1 l(queries_{t-1}) + \gamma_2 l(freq_{t-1}) + \gamma_3 l(rate_{t-1}) + D_M + \epsilon_t \tag{6.2}$$

$$\Delta l(freq_t) = \delta + \sum_{i=1}^{2} \lambda_i \Delta l(freq_{t-i}) + \kappa_0 \Delta l(queries_t) + \sum_{i=0}^{3} \iota_i \Delta l(rate_{t-i}) + \gamma_1 l(freq_{t-1}) + \gamma_2 l(queries_{t-1}) + \gamma_3 l(rate_{t-1}) + D_M + \epsilon_t \tag{6.3}$$

The reduced forms of these model specifications are not considered other than the no-cointegration tests performed in Table 3.

## 6.2 Regression Output

The raw regression output obtained from running regressions in section 6.1.2[28], is presented in a condensed format in Table 4 below:

Table 4: Summary of Regression Output

| Model $\rightarrow$ | 6.1 | 6.2 | 6.3 |
|---|---|---|---|
| $\gamma_1$ [t-value] | -0,053 [-3,56]** | -0,195 [-2,41] | -0,026 [-0,81] |
| $\gamma_2$ [t-value] | 0,034 [1,86] | -0,236 [-2,59] | -0,020 [-1,07] |
| $\gamma_3$ [t-value] | 0,028 [0,82] | -0,052 [1,09] | 0,025 [1,56] |
| $\beta_2$,$\beta_3$ | $\frac{-\gamma_2}{\gamma_1}$ =0,577, $\frac{-\gamma_3}{\gamma_1}$ =0,415 | $\frac{-\gamma_2}{\gamma_1}$ =-1,21, $\frac{-\gamma_3}{\gamma_1}$ =-0,27 | $\frac{-\gamma_2}{\gamma_1}$ =-0,769, $\frac{-\gamma_3}{\gamma_1}$ =0,961 |
| AR-test | 0,125 [0,88] | 6,606 [0,00]** | 2,005 [0,13] |
| Normality-test | 38,291 [0,00]** | 23,910 [0,00]** | 8,033 [0,18]* |
| - Excess Kurtosis | 2,876 | 2,834 | 1,064 |
| - Skewness | 0,484 | 0,697 | 0,445 |
| Heteroskedasticity-test | 3,916 [0,00]** | 3,061 [0,00]** | 0,704 [0,96] |
| RESET-test | 7,009 [0,00]** | 12,874 [0,01]** | 1,384 [0,25] |

* 10% Critical Value, ** 5% Critical Value, *** 1% Critical Value, PcGive-test with 3 variables

In models 6.2 and 6.3 the PcGive test for no-cointegration with 2 variables, with $H_0 : \gamma_1 = 0$ cannot be rejected at the 10% critical level - and thus, they do not cointegrate. Issues with specification ensue.

Criticism should be directed at model 6.2, which appears to be misspecified. Estimating model 6.1, the test for no-cointegration is rejected at the 5% critical level, and variables $log(queries_t)$ and $log(queries_t)$ appear to error-correct towards $log(rate_t)$ without exhibiting cointegrating relations with any other variables in the model.

The assumption of only one cointegration relations, alas only one ECM is upheld - model 6.1 is evidently the only ECM of the three variables considered. The assumption of exogeneity $E[\epsilon_t|\Delta l(queries_t)] = E[\epsilon_t|\Delta l(freq_t)] = 0$, so that the regressors are predetermined. This cannot be fully investigated. It could be argued, that queries and exchange rate have some feedback effects; an increase in price may incline users to search more.

Model 6.1 is dynamically well-specified. Normality, Heteroskedasticity and Functional Misspecification appears to be an isse. In addressing the issue of non-normality, the PcGive reports a normality test compared to a $\chi^2$-distribution with 2 degrees of freedom. This test is sensitive to anomalies in smaller samples. The Jarque-Bera test is a joint test for normality, considering both skewness and excess kurtosis; $\xi_{JB} = \xi_K + \xi_S$ against the null of $H_0 : S = K_{excess} = 0$. It is clear, that what causes rejection of the null is the excess kurtosis in the distribution of residuals. Inspecting the distribution of residuals[29], it is clear that most of the mass of lies around zero - and it is also evident, that the distribution has fat tails. The heteroskedasticity-test reports critical values, which is partly due to large residual outliers from 2013-2014. Given, that the only outliers considered initially, was the ones in March 2015. Heteroskedasticiy is also evident when inspecting the cross-plot of fitted values against residuals; the dispersion appears somewhat truncated towards the right. Critical values in the Ramsey RESET-test indicates, that there are issues with functional misspecification - and that the model may have issues with

[28] Appendix C
[29] Appendix D.1

non-linearity and thus estimation using OLS; this could potentially be solved by transforming the ECM to a non-linear representation and estimate this model using Maximum Likelihood. It is clear, that the model specification is not ideal. Given awareness of the shortcomings of the specified model, it may prove a useful representation of the small sample of weekly data; the long-run solution to the underlying ADL(1, 3, 5)-model can be written as:

$$ECM = log(rate_t) - 1,903 - 0,577 log(queries_t) - 0,415 log(freq_t) \tag{6.4}$$

### 6.3 Interpretation

When interpreting coefficients, the explicit assumption is made, that coefficients do not change across time or samples. Ideally, one would want the coefficients to be stable or constant, but in a highly variant time-series, one would allow for some time to stabilize. The rate of convergence towards stability of the regressor-coefficients[30] appear to converge towards a constant value - for $\Delta log(freq_t)$ this change occurs within the span of a few weeks to a couple of months consistently. Contrarily, $\Delta log(queries_t)$ exhibits unstable convergence, and generally the adjustment takes place in early-2015; this is caused by the extreme observation in March 2015. Inspecting the $\gamma$-coefficients, they appear to stabilize rather quickly, and do not appear to behave sporadically after 2014. The $\kappa_0$-coefficient should be interpreted with some caution, whilst all other coefficients appear to be reasonably stable over time.
Given, that a significant cointegrating relationship has been identified, such that the null of no-cointegration is rejected, it is evident that any deviation from the equilibrium state is corrected by the error-correction term. Equation 6.4 describes the error-correction parameters, and describes the long-run mechanics. Equation 6.4 should be interpreted as follows; assume equilibrium and a 1% 1-period shock to $log(queries_t)$ occurs - then $log(rate_t)$ instantaneously responds by increasing by 0,034% in the corresponding period, all else being equal. From here, no more shocks occur, and $log(rate_t)$ will converge back towards the equilibrium level with adjustment rate 0,053 - and the cummulated impact of the 1-period shock would be 0,577%. The same general logic goes for $log(freq_t)$.

## 7 Discussion

In evaluating the findings of this paper, it proves useful to compare with other litterature, in which alike analyses have been carried out. The article written by L. Kristoufek "BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era", in which corresponding data from 2011-2013 is analysed utilizing the Johansen Cointegration analysis with Vector-Error-Correction Models (VECM) and Vector Autoregressive Models (VAR), it is found that a significant cointegrating relationship exists with Wikipediea search queries and USD/BTC exchange rate in both daily and weekly data. The findings presented above provide contrasting results; when eliminating the extreme observations in $queries_t$ using a dummy, no significant cointegrating relationship, in either direction, is found between $rate_t$ and $queries_t$ in the more recent data. One explanation to the lack of similar results is obviously

[30] Appendix D.3

that the data considered provides observations for two different periods. Much has happened since the beginning of Bitcoin, and as the currency has stabilized compared to previous periods, it might be that the cointegrating relationship has faded and is no longer present. Bitcoin has become more widely spread as a currency, and it might be that the fraction of variation in the exchange rate explained by e.g. search queries on Wikipedia diminishes, as prices are becoming more heavily influenced by trading on Chinese or South American online exchanges, regions which typically prefer different search engines. This might be the case; the research conducted by J. Smith "An Analysis of Bitcoin Exchange Rates" suggets, that the Bitcoin exchange rate aligns with different conventional currencies almost instantaneously.
Another explanation of the different findings might simply be caused by limitations of the techniques applied. It would undoubtebly prove useful to model the multiple cointegration relationships by using vectorizations of the ECM, which would allow for the modelling of multiple cointegrating relations. Furthermore, the VECM framework allows for feedback between variables, which would improve the number of modelling options, which proved restrictive in section 6.1.2.

# 8 Conclusion

Section 2 was concerned with justifying, that the construction of Bitcoin implies that the primary driver for price fluctuations is demand. Demand in this context is difficult to measure, and proxies were introduced in attempting to identify a cointegrating relationship. Having thoroughly analysed the possibilites for non-dual cointegrating relationships amongst 4-5 variables in both daily and weekly data, a single candidate model has been identified. Having tested for unit roots, the candidate model was constructed based on a one-step single equation cointegration analysis instead the Engle-Granger two-step procedure. It has been found that a model for $log(rate_t)$ is best described by having both $log(queries_t)$ and $log(frequency_t)$ as regressors in the weekly data; other combinations was discarded due to multiple cointegrating relations, which the assumptions of the single-equation framework does not allow for. For most variables in the weekly data - and all possible regressors in the daily data, multiple cointegrating relationships are present - and whichever results provided by models cointaining these variables, are at risk of being misleading.
Having specified the best model given the data and tools of introductory cointegration analysis, multiple issues arise. The final model is, inarguably not well-specified in terms of normality, heteroskedasticity and appears to be functionally misspecified. The proposed model's coefficients converge reasonably well towards stability, except for the parameter $\kappa_0$, which is heavily influenced by a large outlier in the search queries on Wikipedia in March 2015. Ideally, the model would clearly and consistently describe the effect of shocks in the regressors on the regressand in both the short- and long term, but due to the shortcomings described above, inference on the parameter estimates should be done very cautiously. The described cointegrating relationship on $log(rate_t)$, i.e. the $\gamma_1$-parameter is not particularly strong and is very close to the 5% critical value, whereby one should conclude that a cointegrating relation exists - however, due to a relatively small sample of weekly observations, this relationship may be compromised at the face of a few added observations. The model should be interpreted carefully and does as such not depict

a clear picture of exactly how the USD/BTC exchange rate is affected by changes in Wikipedia search queries or trade frequency on the Bitcoin platform. Considering, that no-cointegration in the log-transformed queries variable in the daily data, and in general, any appropriate model could not be specified, it should be stated, that given the data and modelling approach, there does not appear to exist a one-directional cointegrating relationship on the Bitcoin exchange rate with proxies for public interest as regressors.

## References

Ali, Robleh Barrdear, J. (2014a). *The economics of digital currencies.* Bank of England, Quarterly Bulletin 2014Q3.

Ali, Robleh Barrdear, J. (2014b). *Innovations in payment technologies and the emergence of digital currencies.* Bank of England, Quarterly Bulletin 2014Q3.

Bohn, H. (2012a). *Lecture Note for Econometrics II: Lecture Note 5 - Non-Stationary Time Series and Unit Root Testing.* Department of Economics, University of Copenhagen.

Bohn, H. (2012b). *Lecture Note for Econometrics II: Lecture Note 6 - Cointegration and Common Trends.* Department of Economics, University of Copenhagen.

Choi, H. & Varian, H. (2009). *Predicting the Present with Google Trends.* Google Inc.

Garcia, David Tessone, C. M. P. & Perony, N. (2014). *The Digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy.* Royal Society Publishing.

Kancs, d'Artis Caian, P. & Rajcaniova, M. (2014). *The Economics of BitCoin Price Formation.* Cornell University.

Kancs, d'Artis Caian, P. & Rajcaniova, M. (2015). *The Digital Agenda of Virtual Currencies: Can BitCoin Become a Global Currency?* JRC Technical Report. Eurpean Commission.

Kristoufek, L. (2013). *BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era.* Scientific Reports 3, Article number: 3415.

Mankiw, G. (2012). *Principles of Economics, 5.th Edition.* South-Western Cengage Learning.

Nakamoto, S. (2009). *Bitcoin: A Peer-to-Peer Electronic Cash System.* Bitcoin.org.

Smith, J. (2015). An analysis of bitcoin exchange rates.

Verbeek, M. (2012). *A Modern Guide to Econometrics.* John Wiley and Sons, Ltd.

UNIVERSITY OF COPENHAGEN
DEPARTMENT OF ECONOMICS
FACULTY OF SOCIAL SCIENCES

# BSc Thesis

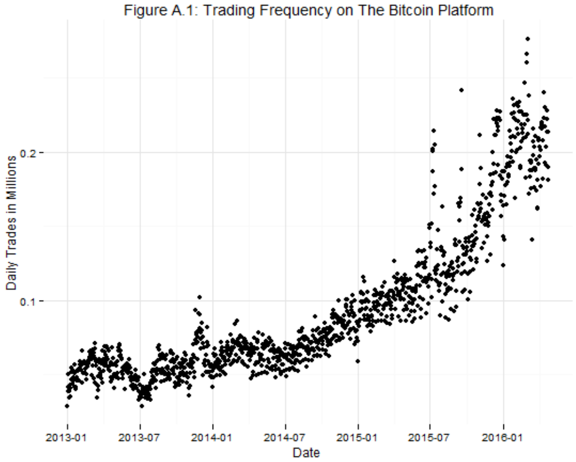
Adam Frederik Ingwersen Linnemann

# Appendix
- A Demand Side Cointegration Analysis of Bitcoin Price Formation

Supervisor: Rasmus Søndergaard Pedersen
ECTS credits: 15 - Typed Units: 43.528 excl. whitespace - Normal Pages: 19,76
4th of May 2016

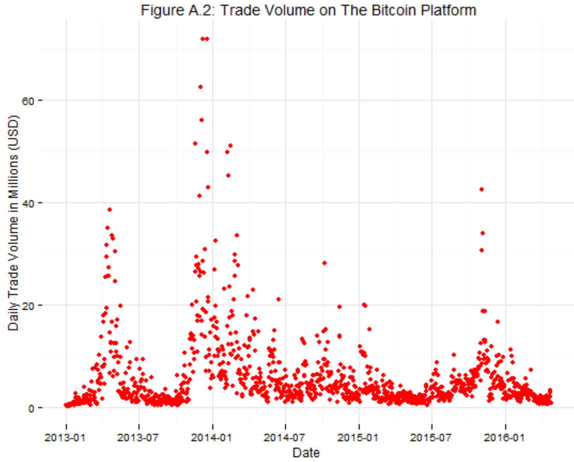
# Appendix A  Transaction mechanics in the Blockchain paradigm

Figure A.1: Trading Frequency on The Bitcoin Platform



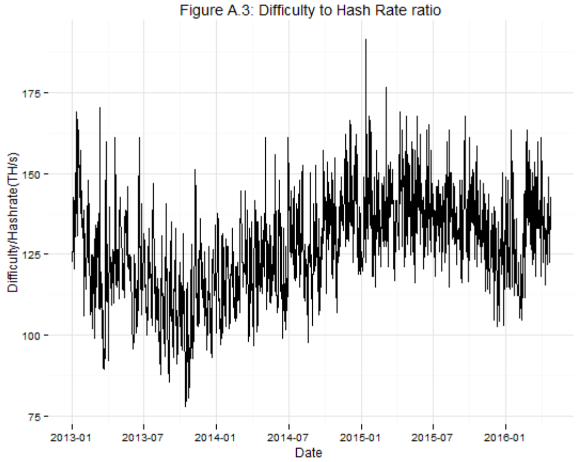

Source: http://blockchain.info

Figure A.2: Trade Volume on The Bitcoin Platform



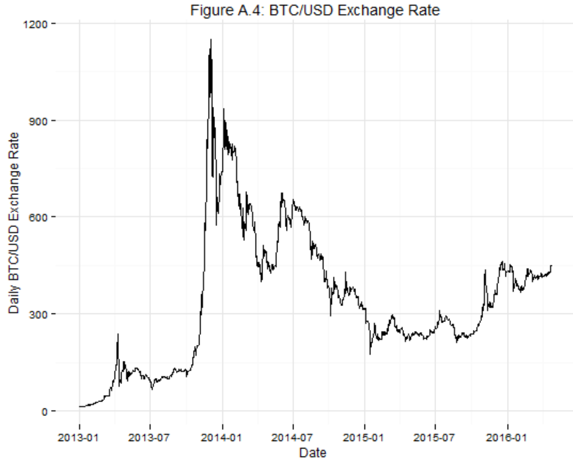

Source: http://blockchain.info

Figure A.3: Difficulty to Hash Rate ratio




Source: http://blockchain.info

Figure A.4: BTC/USD Exchange Rate




Source: http://blockchain.info

# Appendix B  Unit Root Testing

## B.1  Testing for a Unit Root in $rate_t$, weekly data

Table 5: Descriptive Statistics of ADF-tests on weekly $rate_t$ and $log(rate_t)$

| Rate AR(3) | $\pi rate_{t-1}$ | $\pi l(rate_{t-1})$ | $\pi rate_{t-1}+\delta$ | $\pi l(rate_{t-1})+\delta$ | $\pi rate_{t-1}+\delta+\gamma t$ | $\pi l(rate_{t-1})+\delta+\gamma t$ |
|---|---|---|---|---|---|---|
| Coefficient, $\pi$ [t-value] | -0,011 [-0,380] | -0,001 [-1,07] | -0.055 [-0,933] | -0,047 [-2,44] | -0,055 [-0,864] | -0,045 [-2,09] |
| Log. Likelihood | -859,161 | 83,2039 | -856,867 | 89,3409 | -856,861 | 89,3932 |

* 10% Critical Value, ** 5% Critical Value, *** 1% Critical Value.

$$rate_t)LR_c(\pi=\delta=0): -2(LogLik_U - LogLik_c) = -2(-859,161-(-856,867)) = 4,59 \sim DF_c^2(5\% = 9,13)$$

$$rate_t)LR_t(\pi=\gamma=0): -2(LogLik_c - LogLik_t) = -2(-856,867-(-856,861)) = 0,01 \sim DF_t^2(5\% = 12,39)$$

$$log(rate_t))LR_c(\pi=\delta=0): -2(LogLik_U - LogLik_c) = -2(83,2039-89,3409) = 12,27^{**} \sim DF_c^2(5\% = 9,13)$$

$log(rate_t))LR_t(\pi = \gamma = 0) : -2(LogLik_c - LogLik_t) = -2(89,3409 - 89,3932) = 0,10 \sim DF_t^2(5\% = 12,39)$

## B.2 Unit Root tests of regressors

Table 6: Descriptive Statistics for ADF-test on regressors, daily observations

| Search Queries - AR(3) | $\pi queries_{t-1}$ | $\pi l(queries_{t-1})$ | $\pi queries_{t-1} + \delta$ | $\pi l(queries_{t-1}) + \delta$ | $\pi queries_{t-1} + \delta + \gamma t$ | $\pi l(queries_{t-1}) + \delta + \gamma t$ |
|---|---|---|---|---|---|---|
| Coefficient, $\pi$ [t-value] | -0,046 [-0,947] | 0,000 [0,110] | -0,091 [-1,38] | -0,042 [-1,16] | -0,091 [-1,37] | -0,037 [-0,923] |
| Log. Likelihood | -2035,99 | -7,35025 | -2033,71 | -6,06159 | -2033,7 | -5,83058 |
| Trade Frequency - AR(5) | $\pi freq_{t-1}$ | $\pi l(freq_{t-1})$ | $\pi freq_{t-1} + \delta$ | $\pi l(freq_{t-1}) + \delta$ | $\pi freq_{t-1} + \delta + \gamma t$ | $\pi l(freq_{t-1}) + \delta + \gamma t$ |
| Coefficient, $\pi$ [t-value] | -0,003 [-0,337] | -0,000 [-0,080] | -0,129 [-2,24] | -0,013 [-1,21] | -0,175 [-3,20]* | -0,086 [-3,36]* |
| Log. Likelihood | 727,411 | 942,421 | 729,806 | 943,454 | 732,528 | 951,382 |
| Trade Volume - AR(6) | $\pi vol_{t-1}$ | $\pi l(vol_{t-1})$ | $\pi vol_{t-1} + \delta$ | $\pi l(vol_{t-1}) + \delta$ | $\pi vol_{t-1} + \delta + \gamma t$ | $\pi l(vol_{t-1}) + \delta + \gamma t$ |
| Coefficient, $\pi$ [t-value] | -0,068 [-1,98]** | -0,022 [-2,05]** | -0,137 [-2,86]** | -0,091 [-4,88]*** | -0,141 [-2,99] | -0,091 [-4,85]*** |
| Log. Likelihood | -3394,46 | -895,628 | -3387,68 | -887,698 | -3387,1 | -887,486 |

* 10% Critical Value, ** 5% Critical Value, *** 1% Critical Value, DF-test

Table 7: Descriptive Statistics for ADF-tes on regressors, weekly observations

| Search Queries - AR(2) | $\pi queries_{t-1}$ | $\pi l(queries_{t-1})$ | $\pi queries_{t-1} + \delta$ | $\pi l(queries_{t-1}) + \delta$ | $\pi queries_{t-1} + \delta + \gamma t$ | $\pi l(queries_{t-1}) + \delta + \gamma t$ |
|---|---|---|---|---|---|---|
| Coefficient, $\pi$ [t-value] | -0,005[-0,809] | -0,002 [-0,277] | -0,978 [-1,45] | -0,176 [-1,94] | -0,980 [-1,35] | -0,274 [-2,11] |
| Log. Likelihood | -1885,52 | -135,399 | -1881,27 | -131,807 | -1881,2 | -129,152 |
| Google Trends Index - AR(2) | $\pi index_{t-1}$ | ... | $\pi index_{t-1} + \delta$ | ... | $\pi index_{t-1} + \delta + \gamma t$ | ... |
| Coefficient, $\pi$ [t-value] | -0,050 [-0,999] | ... | -0,146 [-1,71] | ... | -0,154 [-1,84] | ... |
| Log. Likelihood | -541,841 | ... | -537,621 | ... | -537,077 | ... |
| Trade Frequency - AR(3) | $\pi freq_{t-1}$ | $\pi l(freq_{t-1})$ | $\pi freq_{t-1} + \delta$ | $\pi l(freq_{t-1}) + \delta$ | $\pi freq_{t-1} + \delta + \gamma t$ | $\pi l(freq_{t-1}) + \delta + \gamma t$ |
| Coefficient, $\pi$ [t-value] | 0,012 [0,691] | -0,004 [-1,09] | -0,006 [-0,112] | -0,015 [-0,449] | -0,138 [-1,41] | -0,210 [-3,12] |
| Log. Likelihood | 448,35 | 94,5361 | 448,58 | 94,6193 | 452,297 | 100,552 |
| Trade Volume - AR(3) | $\pi vol_{t-1}$ | $\pi l(vol_{t-1})$ | $\pi vol_{t-1} + \delta$ | $\pi l(vol_{t-1}) + \delta$ | $\pi vol_{t-1} + \delta + \gamma t$ | $\pi l(vol_{t-1}) + \delta + \gamma t$ |
| Coefficient, $\pi$ [t-value] | -0,207 [-0,786] | -0,069 [-1,67]* | -0,394 [-1,16] | -0,228 [-2,90]** | -0,407 [-1,21] | -0,229 [-2,88] |
| Log. Likelihood | -505,042 | -161,467 | -501,028 | -157,643 | -500,665 | -157,541 |

* 10% Critical Value, ** 5% Critical Value, *** 1% Critical Value, DF-test

## B.3 Likelihood Ratio testing for constant and trend terms

Tabel 8: LR-tests: Daily/Weekly Data

| Daily | $LR_c(\pi = \delta = 0)$ | $LR_l(\pi = \gamma = 0)$ |
|---|---|---|
| $queries_t$ | 4,56 | 0,02 |
| $log(queries_t)$ | 2,58 | 0,46 |
| $frequency_t$ | 4,79 | 5,44 |
| $log(frequency_t)$ | 2,07 | 15,856** |

| Weekly | $LR_c(\pi = \delta = 0)$ | $LR_l(\pi = \gamma = 0)$ |
|---|---|---|
| $queries_t$ | 8,50* | 0,14 |
| $log(queries_t)$ | 7,18 | 5,31 |
| $frequency_t$ | 0,46 | 7,43 |
| $log(frequency_t)$ | 0,17 | 11,87* |
| $Index_t$ | 8,44* | 1,09 |
| $volume_t$ | 8,03* | 0,73 |
| $log(volume_t)$ | 7,65* | 0,20 |

* 10% Critical Value, ** 5% Critical Value, *** 1% Critical Value, $DF^2$-distribution

# Appendix C Regression Output

Figur C.1: Raw regression output from PcGive; Models 6.1-6.3

C:\Users\Adam\Documents\GitHub\CryptoCurrencyEconomics\Results1.out 05/03/16 15:36:02

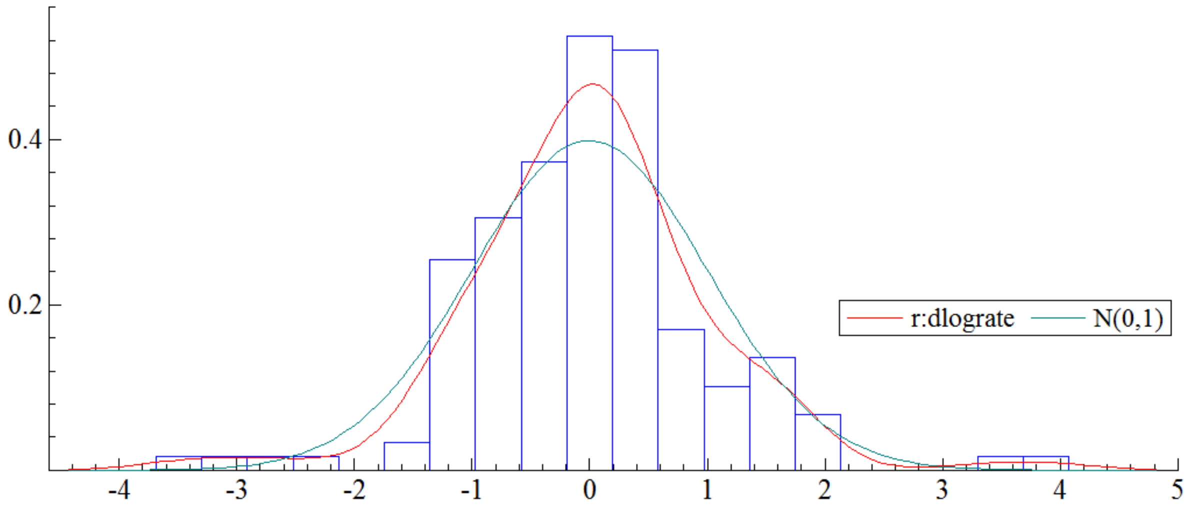
```
EQ(6.1): Modelling dlograte by OLS
       The estimation sample is: 2013-02-09 - 2016-01-02

                  Coefficient  Std.Error  t-value  t-prob Part.R^2
Constant            0.0871160     0.1484    0.587  0.5582   0.0025
dumm20150308        -0.193369     0.1907    -1.01  0.3122   0.0073
dlogqueries         0.0617916    0.02767     2.23  0.0271   0.0346
dlogqueries_1       0.0561925    0.02431     2.31  0.0223   0.0370
dlogqueries_2       0.0456043    0.01953     2.33  0.0210   0.0377
dlogfreq             0.172421    0.08793     1.96  0.0519   0.0269
dlogfreq_1          0.0426337    0.09797    0.435  0.6641   0.0014
dlogfreq_2          0.0338713     0.1022    0.332  0.7407   0.0008
dlogfreq_3         -0.0695274    0.09694   -0.717  0.4744   0.0037
dlogfreq_4          -0.192702    0.08705    -2.21  0.0285   0.0341
logfreq_1           0.0287465    0.03490    0.824  0.4116   0.0049
logqueries_1        0.0344471    0.01848     1.86  0.0645   0.0244
lograte_1          -0.0532648    0.01496    -3.56  0.0005   0.0836

sigma                0.130342  RSS                 2.361459
R^2                  0.234643  F(12,139) =   3.551 [0.000]**
Adj.R^2              0.168569  log-likelihood       100.831
no. of observations       152  no. of parameters         13
mean(dlograte)      0.0198923  se(dlograte)        0.142945

AR 1-2 test:      F(2,137)  =  0.12466 [0.8829]
ARCH 1-1 test:    F(1,150)  =   4.9042 [0.0283]*
Normality test:   Chi^2(2)  =   38.291 [0.0000]**
Hetero test:      F(22,128) =   3.9489 [0.0000]**
Hetero-X test:    F(77,73)  =   3.9159 [0.0000]**
RESET23 test:     F(2,137)  =   7.0885 [0.0012]**


EQ(6.2): Modelling dlogqueries by OLS
       The estimation sample is: 2013-03-02 - 2016-01-02

                  Coefficient  Std.Error  t-value  t-prob Part.R^2
dlogqueries_1       -0.477758    0.06443    -7.41  0.0000   0.2864
dlogqueries_2       -0.290892    0.06496    -4.48  0.0000   0.1277
dlogqueries_3       -0.139232    0.05406    -2.58  0.0111   0.0462
Constant              1.17396     0.4212     2.79  0.0061   0.0537
dlogfreq             0.493243     0.2266     2.18  0.0312   0.0334
dlograte             0.435773     0.2290     1.90  0.0591   0.0258
dlograte_1           0.929559     0.2204     4.22  0.0000   0.1149
dlograte_2           0.813929     0.2208     3.69  0.0003   0.0903
dumm20150308          4.89793     0.3669     13.3  0.0000   0.5653
logfreq_1           -0.236016    0.09119    -2.59  0.0107   0.0466
lograte_1           0.0524595    0.04807     1.09  0.2770   0.0086
logqueries_1        -0.195208    0.08088    -2.41  0.0170   0.0304

sigma                0.363359  RSS               18.0880999
R^2                  0.748031  F(11,137) =   36.97 [0.000]**
Adj.R^2                0.7278  log-likelihood      -54.3243
no. of observations       149  no. of parameters         12
mean(Y)           -0.00385296  se(Y)               0.696453

AR 1-2 test:      F(2,135)  =   6.6064 [0.0018]**
ARCH 1-1 test:    F(1,147)  =   3.1651 [0.0773]
Normality test:   Chi^2(2)  =   23.910 [0.0000]**
Hetero test:      F(20,127) =   3.0614 [0.0001]**
Hetero-X test:    F(65,82)  =   2.0996 [0.0008]**
RESET23 test:     F(2,135)  =   12.874 [0.0000]**

EQ(6.3) Modelling dlogfreq by OLS
       The estimation sample is: 2013-03-02 - 2016-01-02

                  Coefficient  Std.Error  t-value  t-prob Part.R^2
dlogfreq 1          -0.402976    0.08565    -4.70  0.0000   0.1382
dlogfreq 2          -0.215128    0.08451    -2.55  0.0120   0.0449
Constant           -0.0258624     0.1501   -0.172  0.8635   0.0002
dlogqueries         0.0350839    0.02490     1.41  0.1611   0.0142
dlograte             0.222434    0.07695     2.89  0.0045   0.0571
dlograte 1           0.141233    0.07859     1.80  0.0745   0.0229
dlograte 2         -0.0177483    0.08094   -0.219  0.8268   0.0003
dumm20150308        -0.215160     0.1773    -1.21  0.2271   0.0106
logqueries 1       -0.0202992    0.01901    -1.07  0.2875   0.0082
lograte 1           0.0257755    0.01650     1.56  0.1206   0.0174
logfreq 1          -0.0263714    0.03245   -0.813  0.4178   0.0048

sigma                0.125997  RSS               2.19077269
R^2                  0.268102  F(10,138) =   5.055 [0.000]**
Adj.R^2              0.215066  log-likelihood       102.945
no. of observations       149  no. of parameters         11
mean(dlogfreq)     0.00589339  se(dlogfreq)        0.142214

AR 1-2 test:      F(2,136)  =   2.0057 [0.1385]
ARCH 1-1 test:    F(1,147)  =   1.0954 [0.2970]
Normality test:   Chi^2(2)  =   8.0333 [0.0180]*
Hetero test:      F(18,129) =  0.70359 [0.8024]
Hetero-X test:    F(54,93)  =  0.64914 [0.9572]
RESET23 test:     F(2,136)  =   1.3843 [0.2540]
```

Page: 1 of 1

## Appendix D Evaluation of model

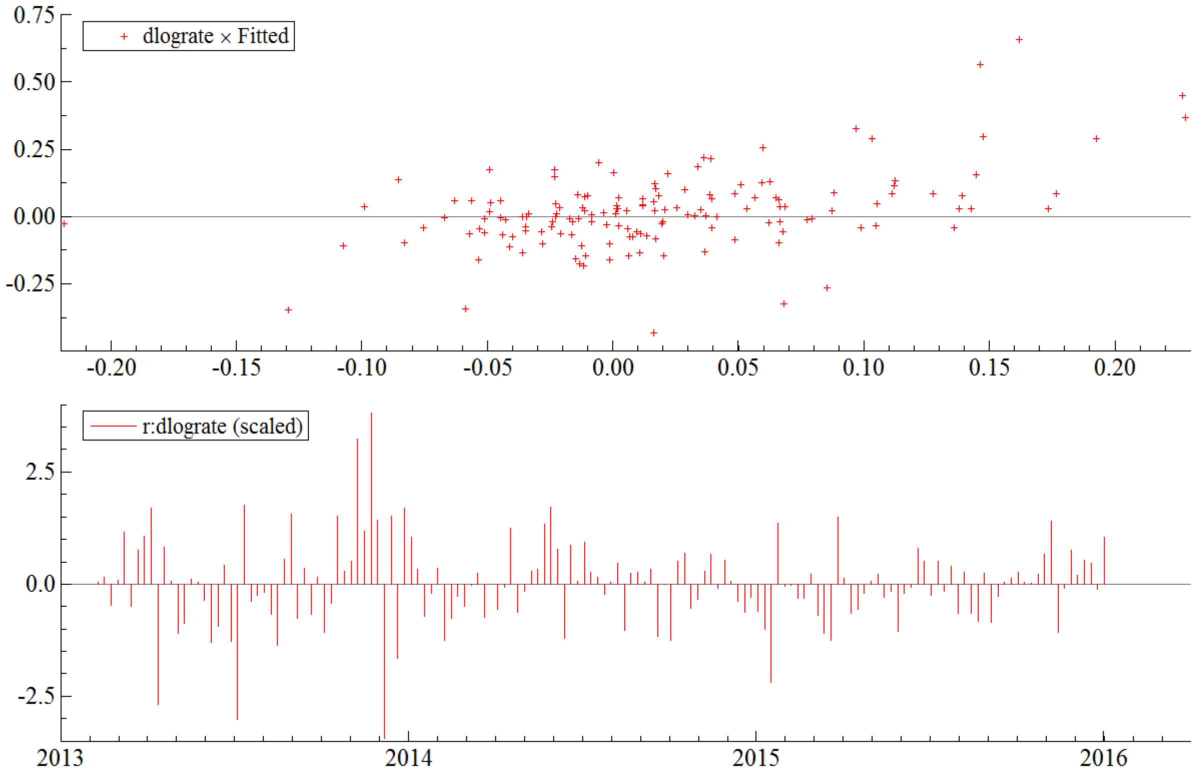
### D.1 Normality of residuals

Figur D.1: Residual density and distribution




### D.2 Heteroskedasticity

Figur D.2: Residual Crossplot and Scaled Residuals



## D.3 Stability of Coefficients

Figur D.3: Adjustment towards stability in coefficients