University of St.Gallen

School of Management, Economics, Law, Social Schiences and International Affairs

The effect of Twitter activity on Bitcoin price fluctuation

Software Engineering for Economists
(7,610,1.00)

Alen Stepic - 11-475-258
Dimitrios Koumnakes - 10-613-370
Joël Sonderegger - 11-495-488
Severin Kranz - 13-606-355
Chi Xu - XX-XXX-XXX

Fall Term 2017

Supervisor:
Prof. Dr. Philipp Zahn
FGN HSG
Varnbüelstrasse 19
9000 St. Gallen

28th December 2017

Abstract

This paper, examines the dynamics between the amount of twitter activity regarding bitcoin and the actual price fluctuation of bitcoin. For the analysis two sets of data have been collected. One reflecting twitter activity and the other one showing the bitcoin prices for the same period. Based on that an econometric model is used for the statistical analysis and the interpretation of the results.

INPUT ALEN: ADD JUST KEY FINDING THE REST I LIKE

This work was created in the context of a programming course for economists at the university of St. Gallen. To not exceed the framework of this study we require extensive knowledge in econometrics and take concepts as known. The underlying goal was to get familiar with software engineering tools and project management. Therefore, the academic content of this work is neither completed nor concluding.

Contents

List of Figures

List of Abbreviation

VAR	Vector Autoregression
EMH	Efficient Market Hypothesis

1 Introduction

Malkil and Fama introduced in 1970 the Efficient Market Hypothesis (EMH), where actual prices include all information and investors act rational. This hypothesis is broadly accepted in the financial world. However, the digitalisation leads to an increased network effect where information can be share within second over the globe. Furthermore, the EMH fails to address the behavioural and emotional role of investors. The large fluctuation of the bitcoin prices in the last year (mostly increasing) and the fact that Twitter has become a very popular social network, where people interact, led to different research in the field of sentimental analysis with the focus on bitcoin and twitter. Mao, Counts and Bollen (2015), claim that sentiment analysis using twitter tweets which contain the buzzword bullishness can indeed be used as an sentiment indicator for bitcoin prices. As digitalization is expected to continue in the future, crypotocurrencies will hypothetical get more important. Thus, the topic is relevant.

1.1 Research Question and Goal of the Paper

As existing research focus mostly on the sentimental analyses of tweet content, this paper aims to examine existence of a correlation between the price development of bitcoin and the amount people talk about bitcoin on twitter. The goal should be reached by answering the following research question:

In what extend does the amount of twitter tweets influence the price of bitcoin?

1.2 Methodology

To answer the above mentioned research question a scientific approach has been applied. The paper consists out of three parts. The first part contains an introduction into the topic by pointing out the relevance of the topic, the research question, methodology and the scope. This is conducted in chapter 1. The second part is characterized as the theory part. The theory aims to provide to the reader the necessary background information by an introduction of relevant bitcoin and twitter information in context of the paper. The conducted research is based on forward research and using relevant sources. Furthermore, it contains a short introduction into the Vector Autoregression approach (VAR) and the constrains in context with the paper. Chapter XYZ The third part contains the discussion of the results and a conclusion. The conclusion is based on the theory and the own conducted mathematical computation. Those computations are based on a data set, which was gathered by own coded python scripts for the timeperiod ADD TIME PERIOD. The data set contains out of two data sources (1) twitter tweets which contain the buzzword "bitcoin" and (2) historical bitcoin prices. Those two sources has been aggregated with a separate python script. The output of aggregation was further proceeded with stata to conduct the VAR on a daily basis. Chapter 4 and 5

INPUT: IS IT NOW DAILY? WHAT IS THE REASON?

1.3 Scope

The paper's focus lies on the scientific discussion and answer of the research question mentioned above. The research question has consciously been design very tiny, as the focus of the course Software Engineering for Economists is the application of different tools, the documentation and reproducibility of the results, rather than writing a high quality scientifc paper. The approach and technical issues (like code scripts etc.) are not addressed. Those points are explained in the separate documentation paper. It is not a deep description of the code as the code itself as the code is documented separately. Nevertheless, important lines of code are discussed.

2 Brief Background and Data Collection

Bitcoin

(Text for Bitcoin) Input Alen: Explain why bitcoin is of relevance, short intro what bitcoin is

Twitter

(Text for Twitter) Input Alen: Explain why twitter is the right source and and some facts about it

Data Collection

For our dataset, we used 2 different sources. We first collected historical data on bitcoin prices to USD and secondly, we gathered twitter massages related to bitcoin. We fulfil this relationship by selecting tweets that contain the term "bitcoin" in the text message. The time period for both collected data sets was from 21.12.2017 16:00:00 UTC to 26.12.2017 16:00:00 in an hourly basis. We chose an hourly basis to be able to do a statistical analysis with a certain amount of observation for the restricted processing time of this homework.
For the bitcoin prices, a publicly available API from CoinDesk was used (Source? See documentation?). The prices collected represent the final price level of bitcoin before every hour changes. For example, our first listed price of bitcoin was on 21.12.2017 at 16:59:59 and our last value on 26.12.2017 at 15:59:59. These values construct our first time series data set and are shown in the following graph.

3 Econometric Modelling and Results

As stated in the beginning, we want to examine the underlying relationship between twitter activity and bitcoin price fluctuations. From our data collection, we are left with two time series datasets representing two variables. A broadly used statistical method to simultaneously analyze multiple time series is the VAR (Vector Autoregression) approach. In this approach the endogenous variables are determined both by their own historical values and by the historical values of the other endogenous variables (Lütkepohl, 2005, p. 4-5).
To generate the econometric results that follow we used the statistical software STATA.

Stationarity

Before estimating a VAR model, the time series data must be checked for stationary. Thus, the means and variances are constant over time and the datasets do not show any trending behavior. Non-stationary data can lead to an inaccurate model which is undesirable. To test for stationarity, we use the Augmented Dickey-Fuller (ADF) test and get following results from STATA.
Interpreting the results and accepting or rejecting stationarity.

Lag Specification

To select the optimal number of lags in our VAR regression we check for various information criteria. Information criteria are measuring the tradeoff between model fit and parsimony, giving use the optimal number of lag to use (Brandt and Williams, 2007, p. 27). The calculations of these criteria for our given data set is easily done by a statistical software and given in the following table.

VAR Regression Model

In this paper, we use a basic unrestricted VAR model which consists of two endogenous variables, T for aggregate tweets and B for bitcoin prices. The selected time lag is 3 and this choice will be discussed later. Given that the model equations can be written as follows:

Granger Causality

To assess the causal relationship between our two endogenous variables and interpret the result we use the Granger causality. This test confirms if one variable is statistically useful to predict the other variable. If this is given the dynamics stated by the calculated coefficient above can be assumed to be valuable.
A second analysis to interpret our results is the impact response analysis. This evaluates the impact of changes in the one variable to the other variable. This is also provided by STATA and the results can be seen in the following graphs.

Impulse Response Analysis

4 Conclusion

(Insert text)