STAT605 Project Proposal

Augustine Tang, rtang56; Jonquil Liao, zliao42;
Yudi Mu, ymu27; Ruyan Zhou, rzhou84

1 Introduction

Our data contains two part. First part is the bit coin transaction data from Kaggle. It contains five variables: height, input, output, sum, and time. We will talk more about the variables in second part. Second part of our data is the corresponding daily bitcoin price data, which has daily open price, highest price, lowest price, mean price, and percent of price change.

Our first question is: What is the time-varying pattern of the transaction amount? Is it predictable? Is there seasonality? We integrated the transaction amount per minute and utilized extreme value analysis on its maxima, then we gain an index per week representing the trading volume of that week. We thus find the fluctuation of the index aligns with the transaction amount's time series well, which indicates it a good way to represent the very long-term data by using the much shorter extreme value distribution index. Our second part is to explore the association between daily prices and transaction amount using Spearman rank-order correlation test. In the third part, in order to catch the trends and forecast values, we tried to fit time series models to the price series andthe transaction amount series.

2 Data Analysis

2.1 Data and variable description

Our data sets which from Kaggle website have 19 csv files 19.82 GB in total. Each file contains many blocks of bitcoin, and the file name shows the scope of blocks. Every file has the same structure, and each has five columns: block height(can be regarded as block names), input (address), output (address), sum (of transaction), time (of the transaction confirmation). A block includes many transactions. We also included the daily price data of the bitcoin in the time scope (2016-09-15 to 2017-07-16) of our main data.

The variable Input can be seen as the name of user who sell out bitcoins, and the variable Output is the name of user who bought bitcoins from the input user, and it also contains the transaction of each user. The variable Sum is the sum of transaction that shown in the output variable. The unit of transaction is 1 bitcoin.

2.2 Data preparation

Firstly, the data sets contain variable "Time", but this is a character type of data. We first extracted the time from every entry and made it into Unix time format. And we ordered the data by Unix time to see the beginning and the end of the time. scope. We also formatted the time into day and minute. And then summed the transaction by day and by minute for further use. We also extracted input and output user names and summarised the number of unique users by day and by minute for analysis. In this part, we used R to write the script and used CHTC to run the parallel jobs.

2.3 Extreme Value analysis on minute transaction amount

Statistical model

This section we used the minute-sum transaction amount mentioned in the Data preparation section. We picked out the block maxima Q_t of every 60 data points, which is the maximum trading volume per minute in an hour. Then the Q_t should be one of the three extreme value distribution: Fréchet, Weibull and Gumbel. We assume that in approximately a week, the distribution of Q_t is the same. And then for

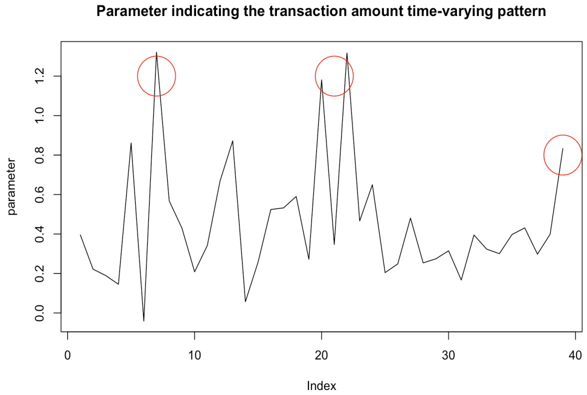

Figure 1: ξ series

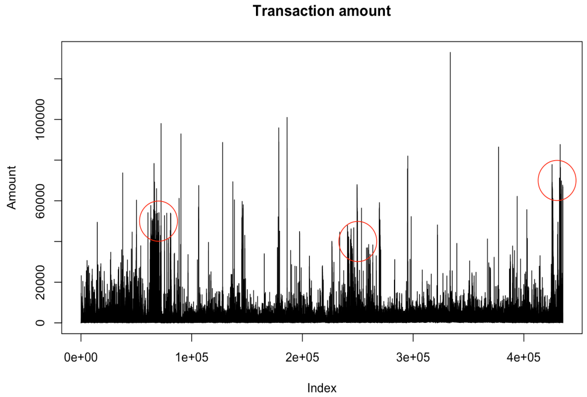

Figure 2: transaction amount

every 180 Q_ts we utilized the gev.fit function in package "ismev" in R to calculate the three parameters (σ, μ, ξ) of its distribution:

$$G(x) = exp\left\{ -\left[1 + \xi(\frac{x-\mu}{\sigma}) \right]^{-\frac{1}{\xi}} \right\}$$

The series of the shape parameter ξ_ts has a time-varying pattern and larger ξ means larger transaction amount in that week.

Computations and findings

We used CHTC to do computations in this section. Firstly script pre1splitdata.sh splitted the long-term data with 430000 entries into 10 shorter data sets. Then process.sub created 10 parallel jobs to process each data set as the procedure in section 'Statistical model' and produced 10 short ξ series. The script alpha_model.sh combined all the ξ series into one. Figure 1 and Figure 2 shows the combined ξ series and transaction series separately, from which we can see they have good alignments in regard to peaks and trends. This section extracted the pattern of the trading volume series from a very long data set and integrated it into a short ξ series and thus we can do tests and analysis on ξ and gain intuitions more conveniently.

Difficulties

1. The package "ismev" should be downloaded and added to chtc and the resource is not as easy to get as R402.tar.gz.
2. Originally we plan to model ξ with a ARMA time series model and thus predict the next several ξs. But the Ljung-Box test and acf plot showed it is a white noise series, so this method may not be applicable.

2.4 Association test

Spearman rank order correlation test
We used sumbyday.csv from data preparation part, and the original daily price data. We want to test the association between daily transaction versus daily lowest price, highest price, open price, mean price, and price change. In our data set, the time span is 11 months from 2016-09 to 2017-07. We first test the association in each month separately. However, there is not a fixed relationship between transaction with prices by month. Yet, in the next section, we want to fit a time series model on the whole period. Thus, independence in each month cannot imply independence in the whole period. So we focus on the overall association. The output is shown in the table:

	mean price	open price	highest price	lowest price	price change
Spearman ρ	0.41	0.42	0.4223093	0.40	0.10
p-value	$<10^{-13}$	$<10^{-14}$	$<10^{-14}$	$<10^{-12}$	0.08

Table 1: Association

Finding
The correlation between transaction and prices are all positive, therefore, they are positively correlated, except for price change whose Spearman rho is very small. And the p values for the first four tests are much smaller than 0.05, which means all the test are significant. Then we can say that daily transaction is positively correlated with daily mean price, open price, highest price, and lowest price. For the price change, the p-value indicates there might be association.

Method
The method we used to test for association is Spearman test. The Spearman rank-order correlation coefficient (Spearman's correlation, for short) is a nonparametric measure of the strength and direction of association that exists between two variables measured on at least an ordinal scale. And after we denote the association, then it is reasonable to fit a multivariate time series model in next step.

2.5 Time Series Model

In order to catch the trends and forecast values, we tried to fit time series models to the price series and the transaction amount series. This part is fully conducted in R and we ran the script in chtc.

VAR Model
As we know from the association test, the transaction amount is highly correlated with prices. So, we tried to fit a multivariate time series model on the daily closed price series and daily transaction amount series. After checking the stationality of these two series (We used adf test and acf plots), we found that they both fail to satisfy this property. Thus, we took differences of these two series and studied the new series. After taking differences, these two new series are stationary. We selected the perfect lag by several criteria and it was 4. Then, we fitted the vector auto-regression model. The results are: Denote the price of time t as P_t and the transaction amount at time t as W_t. Also, denote their differences as $\Delta P_t = P_t - P_{t-1}$ and $\Delta W_t = W_t - W_{t-1}$ respectively:

$$\begin{aligned}\Delta W_t =& 23348.53 - 0.40\Delta W_{t-1} - 0.42\Delta W_{t-2} - 0.25\Delta W_{t-3} - 0.29\Delta W_{t-4}\\ &+ 98.90\Delta P_{t-1} - 334.22\Delta P_{t-2} - 1073.98\Delta P_{t-3} + 558.05\Delta P_{t-4} + \epsilon_{W,t}\\ \Delta P_t =& 7.52 + 4.62\times10^{-3}\Delta P_{t-1} + 9.98\times10^{-4}\Delta P_{t-2} - 2.34\times10^{-2}\Delta P_{t-3} - 7.22\times10^{-2}\Delta P_{t-4}\\ &- 2.80\times10^{-6}\Delta W_{t-1} + 1.56\times10^{-7}\Delta W_{t-2} + 5.44\times10^{-7}\Delta W_{t-3} - 5.38\times10^{-8}\Delta W_{t-4}\\ &+ \epsilon_{P,t}\end{aligned}$$

Validation
In order to check the assumptions, we studied the residuals of our model. It showed that the residuals have mean zero and are stationary. There were no violations of the model assumptions.
The fitted values are shown in Figure 3. In each plot, the black line is the raw series and the red line is

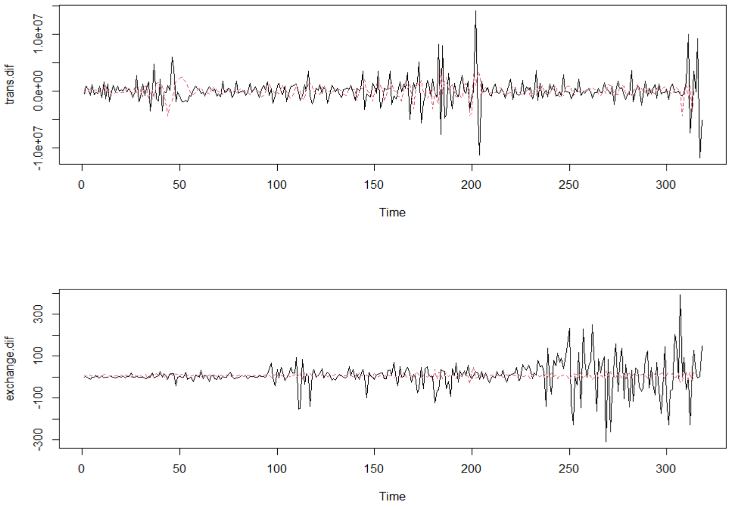

Figure 3: VAR fitted series

the fitted series. The one above is the fitted series of transaction amount differences and the other one is the fitted series of price differences. As we can see, the one above fits quite good but the later one didn't capture the high volatility of its later part.

Figure 4 shows the forecasting values of our model. The blue parts denote its 95% CI and 99% CI.

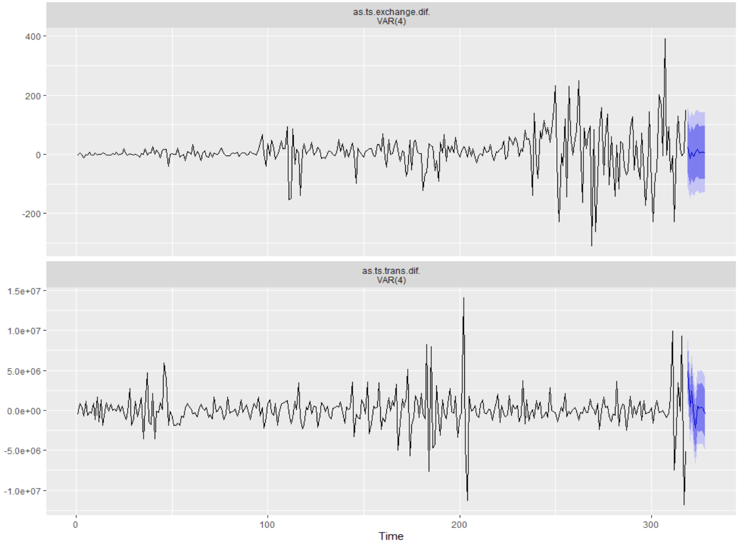

Figure 4: VAR forecasting values

3 Conclusion

1. We found that the fluctuation of the index aligns with the transaction amount's time series well. it's a good way to represent the long-term data by using the much shorter extreme value distribution index.
2. Daily transaction is positively correlated with daily mean price, open price, highest price, and lowest price. It is reasonable to fit a multivariate time series model in next step.
3. We fitted the multivariate time series model on the daily closed price series and daily transaction amount

series. The fitted series of transaction amount differences has a pretty good result.