STAT605 Final Draft

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 is from Kaggle website, which has 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 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

At the very beginning, we want to get a general intuition about the time-varying transaction amount. How does the transaction amount change over time? Is there specific pattern? Can we get a rough prediction of it?

Statistical model

In this section, we used the data of minute-sum transaction amount (univariate time series, each entry is the sum of the transaction amount in a minute) mentioned in the Data preparation section and split it into several blocks (This block is different from which mentioned in Data preparation section. It is chosen manually by separating a long-term data series into several short ones with appropriate length for model fitting.). By fitting static extreme value model on each block, we got one parameter value on each block which roughly represents the transaction amount level in this block relatively. And then the plot of these values can roughly show the time-varying pattern of the transaction amount. The detailed steps are:

1. We split the whole minute-sum transaction amount series into 40 blocks. Each block has 10800 data points. The length of the block is chosen for convenience, because in the next step we are going to fit extreme value distribution on each block, and this length ensure that the fitting is almost accurate and also not too long.

2. We picked out the maxima Q_t of every 60 data points, which is the maximum trading volume per minute in an hour. So there are 180 Q_ts in a block. Then the Q_t should be one of the three extreme value distribution: Fréchet, Weibull and Gumbel. We assume that in the same block, the distribution of Q_t is the same. So for the Q_ts in one block we utilized the gev.fit function in package "ismev" in R to calculate the three parameters (σ, μ, ξ) of its distribution:

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

   $G_{Q_t}(x)$ is the cumulative distribution function of extreme value distribution. So now we got the parameters (σ, μ, ξ) in each block, and ξ is what we want to focus on.

3. From the property of extreme value distribution, a large ξ indicates that this series has a high tendency to have large value, which means in our minute-sum transaction amount time series, if a block's ξ is larger than another block's ξ, we may conclude that this block is more possible to have large transaction amount in a minute.

4. If we index each block by t (t=i means in the ith block), as we had 40 ξs, we had a series of ξ_t, t=1,2,...,40, and if we plot these ξ_ts in a figure, by looking at how the ξ_ts' values change over time, we can easily have a insight about in which time period (which block), the transaction amount is large, and about how the amount level will change in the next block.

Computations and findings

We used CHTC to do computations in this section.

1. Firstly script pre1splitdata.sh split the long-term data with 430000 entries into 10 shorter data sets. Since after integrating the transaction amount in each minute in Data preparation section, the series is not as long as it was. So 10 parallels each with 43000 data points can be done in a short time.

2. Then process.sub created 10 parallel jobs to process each data set following the procedure in section 'Statistical model' and produced 10 short ξ_t series.

3. The script alpha_model.sh combined all the ξ series into one series with 40 ξ_ts. Figure 1 and Figure 2 shows the combined ξ series and minute-sum transaction series separately. "Index" in Figure 1 is the block number, which is in time order. And "Index" in Figure 2 is the time of the transaction amount happens, each index represents one minute.

4. Each job runs about 3 minutes and the whole process approximately takes about 10 minutes.

From Figure 1 we can see there are 3 peaks shown on the plot at around the 8th, block, the 21st block and the 40th block. And in Figure 2, the transaction amount also shows peaks in roughly the corresponding time period of the 8th 21st and 40th blocks. So we conclude, in terms of extreme values, they have good alignments. So by doing this process, we can obtain a rough intuition of the trend of the transaction amount, even when we do not want to plot the whole series of the minute transaction amount. Because

the amount series is quite long sometimes, it may froze our laptop if we insist to plot the whole series to see the pattern by eye. Even in our data set (bitcoin), which is not that long after Data preparation, plotting Figure 2 also froze my R for at least 1 minute and the plot can not be saved (Each time I want to export the plot results in my R collapse at last.). So this extreme value model is a good try to shorten the series when we want to get a intuition of the data series' time varying pattern and especially want it to reflect when the relatively largest values happen.

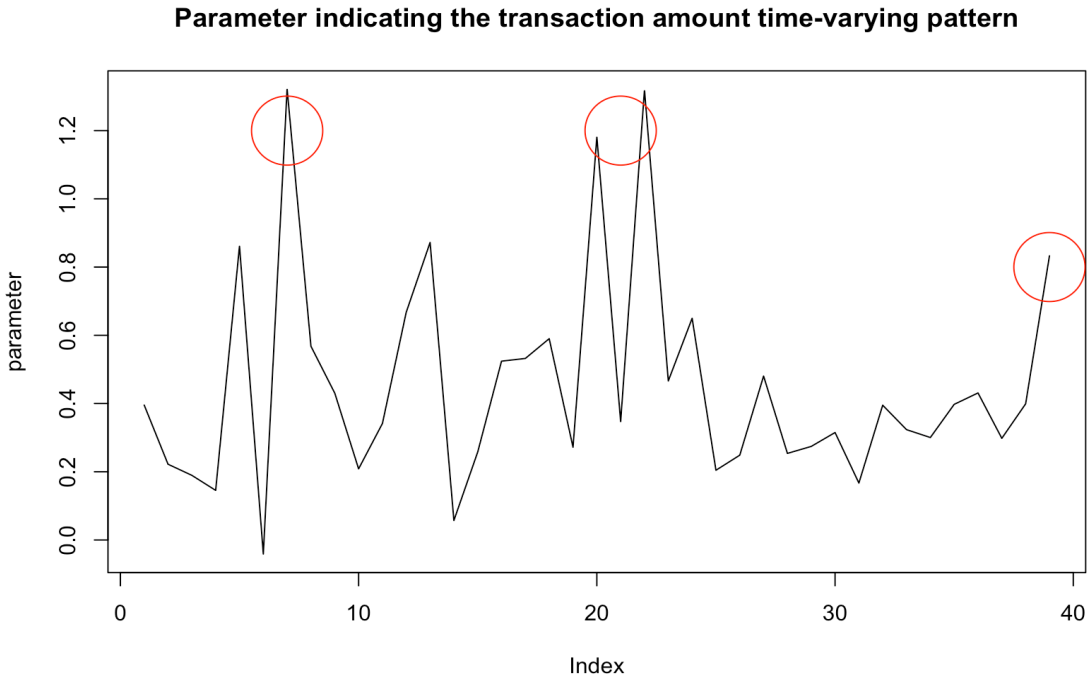

Figure 1: The shape parameter ξ series: each point represents the distribution in one block

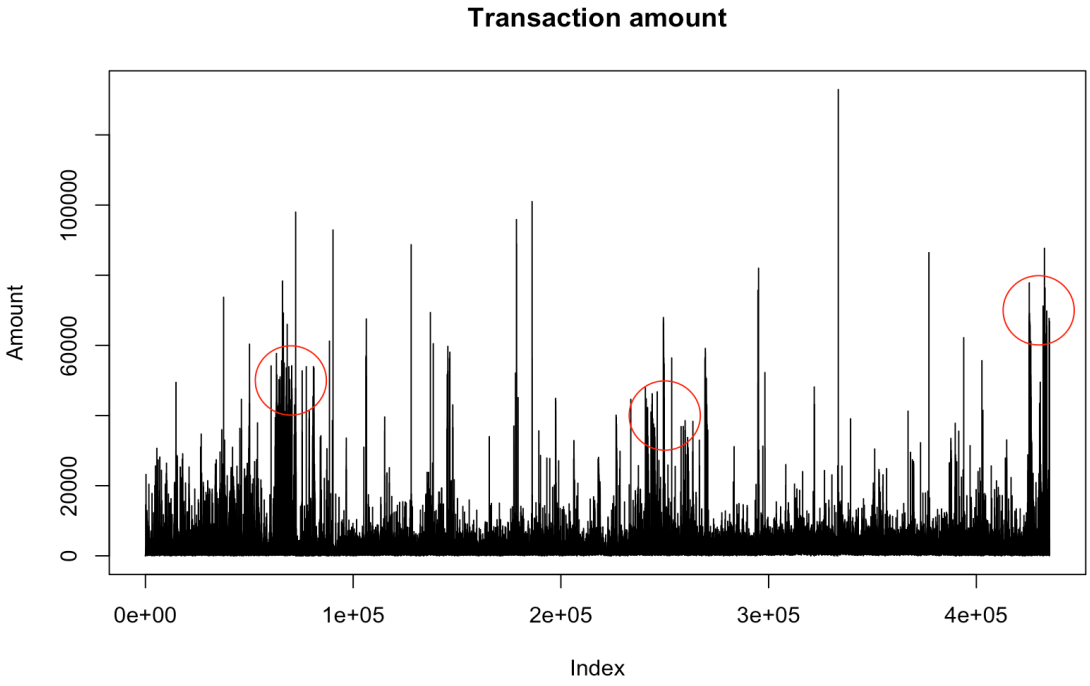

Figure 2: The minute-sum transaction amount: each point is the sum of transaction amount in a minute

Pros and Cons

1. The Q_ts distribution in one block may not be the same in reality. We assumed they are the same and fit static extreme value distribution on each block, which is not appropriate some times. A better choice is using a latent process to get ξ for each Q_t, and then the recovered series of ξ is more accurate and efficient. But it has a lot of difficulties in both theoretical and computational aspects.

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 in this circumstance we can not predict the next ξ.

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 amount 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 between daily values 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 method we choose 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. We choose it because of its robustness.

Finding
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

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.
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($) and the transaction amount series($\times$1k). 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($) and daily transaction amount series($\times$1k).
Denote the price of time t as P_t and the transaction amount at time t as W_t. First, we need to check if these series are stationary. The method we use here is the Augmented Dickey–Fuller test. The results are summarized in Table 2. We found that they both fail to satisfy this property. Thus, we took differences of these two series and studied the new series. Denote their differences as $\Delta P_t = P_t - P_{t-1}$ and $\Delta W_t = W_t - W_{t-1}$

Series	lag	p-value
W_t	0	<0.01
	1	<0.01
	2	0.03
	3	0.06
	4	0.21
	5	0.24

Series	lag	p-value
ΔW_t	0	<0.01
	1	<0.01
	2	<0.01
	3	<0.01
	4	<0.01
	5	<0.01

Series	lag	p-value
P_t	0	0.98
	1	0.98
	2	0.98
	3	0.98
	4	0.98
	5	0.98

Series	lag	p-value
ΔP_t	0	<0.01
	1	<0.01
	2	<0.01
	3	<0.01
	4	<0.01
	5	<0.01

Table 2: Results of the Augmented Dickey–Fuller test. The alternative hypothesis is that the tested series is stationary. According to adf tests, the original series are not stationary while the difference of them are stationary.

respectively. After taking differences, these two new series are stationary.

We selected the suitable lag by several criteria including AIC(n), HQ(n), SC(n) and FPE(n). Then, we fitted the vector auto-regression model. The results are:

$$\begin{aligned}\Delta W_t =& 23.35 - 0.40\Delta W_{t-1} - 0.42\Delta W_{t-2} - 0.25\Delta W_{t-3} - 0.29\Delta W_{t-4} \\ &+ 0.1\Delta P_{t-1} - 0.33\Delta P_{t-2} - 1.07\Delta P_{t-3} + 0.56\Delta P_{t-4} + \epsilon_{W,t} \\ \Delta P_t =& 7.52 + 4.62 \times 10^{-3}\Delta P_{t-1} + 9.98 \times 10^{-4}\Delta P_{t-2} - 2.34 \times 10^{-2}\Delta P_{t-3} - 7.22 \times 10^{-2}\Delta P_{t-4} \\ &- 2.80 \times 10^{-3}\Delta W_{t-1} + 1.56 \times 10^{-4}\Delta W_{t-2} + 5.44 \times 10^{-4}\Delta W_{t-3} - 5.38 \times 10^{-5}\Delta W_{t-4} \\ &+ \epsilon_{P,t}\end{aligned}$$

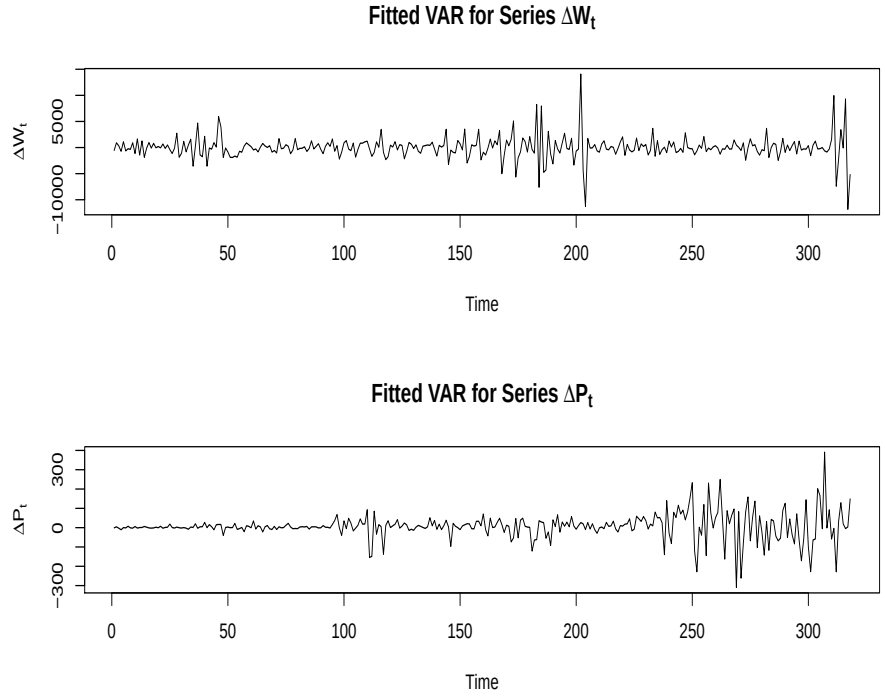

Figure 3: Fitted VAR series. The black lines are the original series and the red lines are the fitted series. The model moderately captures the trend of transaction amounts but fail to capture the high volatility of the price series.

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 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 grey fill areas denote its 95% CI and 99% CI. According to these two plots, the transaction amounts tend to remain the same while the prices are increasing in the near future.

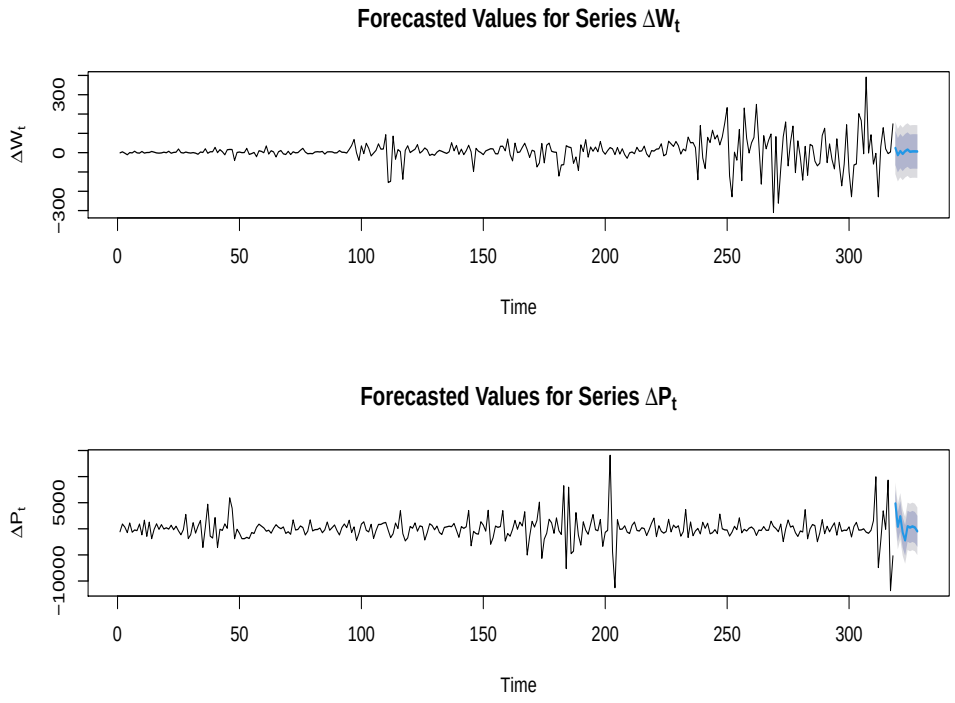

Figure 4: Forecasted Values based on VAR model. The grey fill areas denote its 95% CI and 99% CI. According to these two plots, the transaction amounts tend to remain the same while the prices are increasing in the near future.

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 model moderately captures the trend of transaction amounts but fail to capture the high volatility of the price series. According to the forecast of our model, the transaction amounts tend to remain the same while the prices are increasing in the near future.