Danmarks
Tekniske
Universitet

DTU

Assignment 3

Author

Emil Johannesen Haugstvedt - s211818

November 23, 2021

Contents

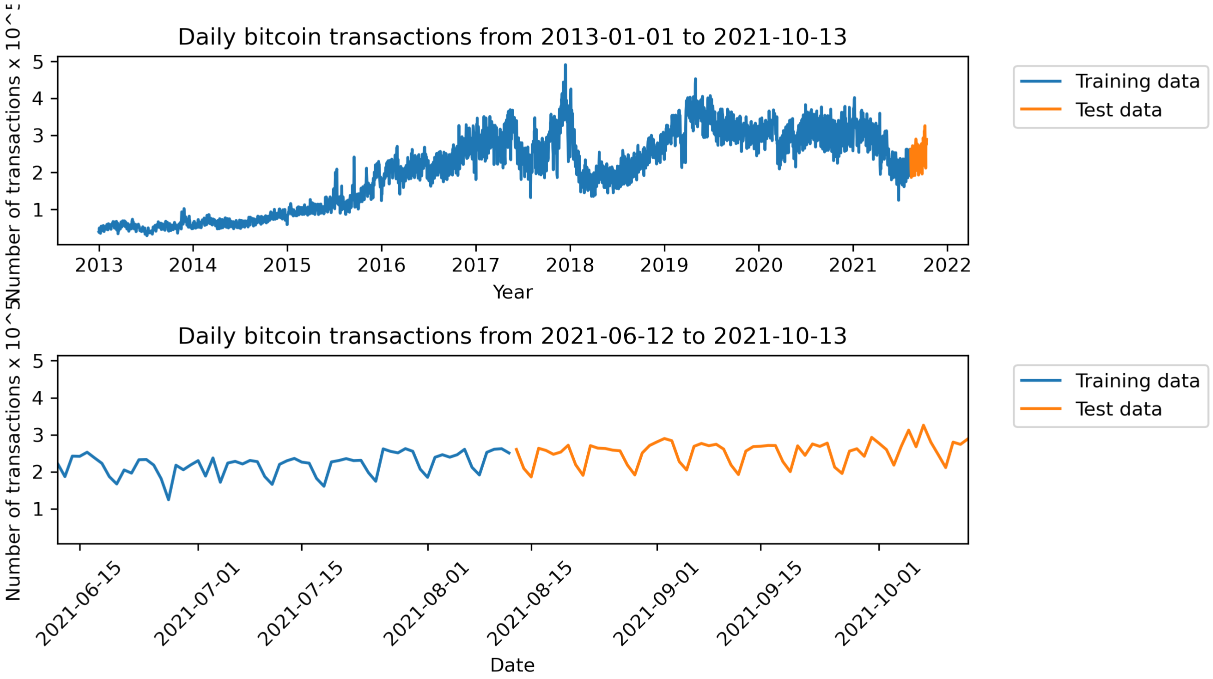

Figure 1: Daily bitcoin transactions from 01.01.2013 to 13.10.2021. The upper plot shows the whole period, while the lower shows the last four months. The training data used to fit the model is colored blue, while the test data is orange. In the upper plot, note how the data clearly is non-stationary with a change in both the variance and mean over time. For the bottom plot a seasonality of seven can clearly be seen.

Question 3.1

Figure 1 shows a plot of the daily bitcoin transactions from 01.01.2013 to 13.10.2021. The data is separated into two parts, the training data (blue) and the test (orange) data. The upper plot shows the whole data set and the lower is a zoomed plot showing only the last 4 months.

By looking at the upper plot it is clear that the time series is non-stationary. This is due to both the variance and mean changing over time. Both the mean and variance are lower in the beginning than in the end. In the lower plot the daily pattern can be examined. Here you can see how there is a clear pattern each week with more transactions in the normal weekdays than in the weekends. There are also clear correlations between the different weekdays, a Tuesday has much the same number of transactions as other Tuesday.

This suggest that a seasonal ARIMA model (SARIMA) may be a good model for the data set. This model will be good at taking the seasonal correlation into account and thus be able to predict well for the future.

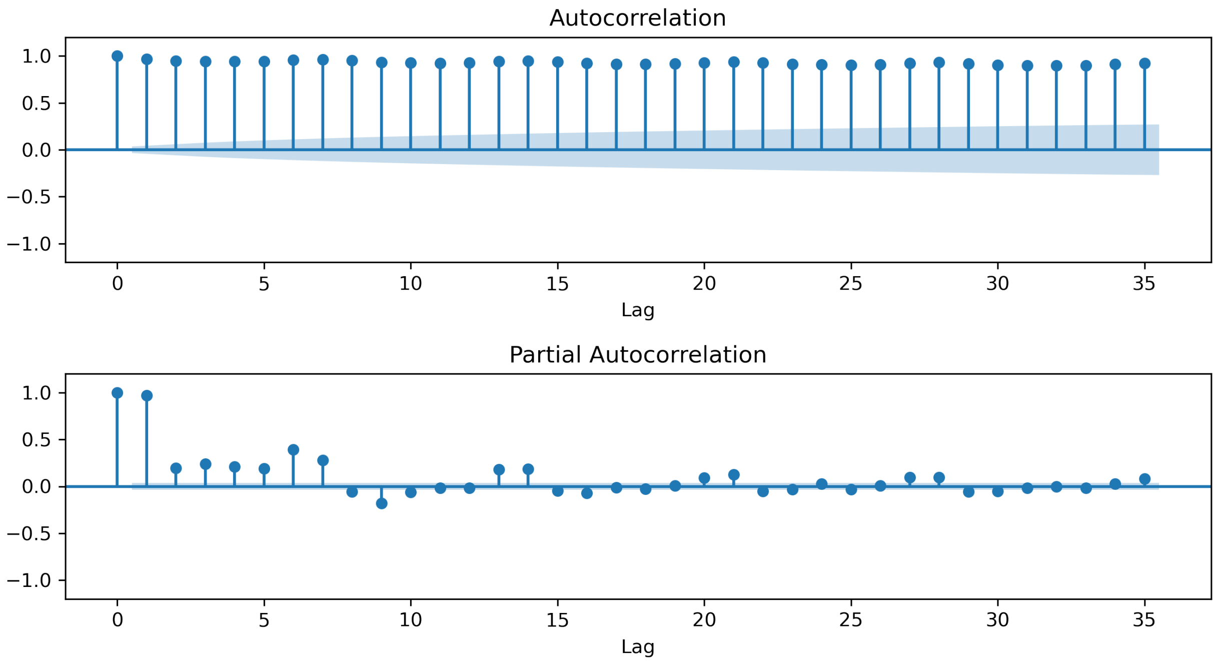

Figure 2: The autocorrelation and partial autocorrelation plot for the daily bitcoin transactions training data set. The non-stationarity of the data is supported by the consecutive positive correlations in the autocorrelation plot. This is an indication that differencing must be done in order to be able to forecast.

Question 3.2

In Figure 2 the autocorrelation and partial autocorrelation plots are shown. The autocorrelation plot shoes large positive correlations between close lags, something that is typical for a non-stationary time series like this.

In order to be able to model the time series using a SARIMA model the data set needs to be stationary. There are several of techniques for making a non-stationary time series stationary. First the mean needs to be made constant throughout the time series. This can be done by differencing the data set. Differencing is to look at the difference between two samples instead of their actual value and use that value to predict. By doing this the mean can be made constant without any loss of information.

Secondly, the variance of the time series needs to be constant. A way to do this is the use a transformation on the data set. I will use the log transformation. This will make the large spikes in the variance smaller without reducing the magnitude of the lower spikes significantly. An important thing to remember when using the log transformation is that the predictions from the model also will be log transformed. Thus the prediction needs to be transformed back using the exponential function.

Just differencing the data once proved to not be sufficient as it left significant correlations at each 7th lag in the autocorrelation plot as seen in Figure 4. This suggested one more

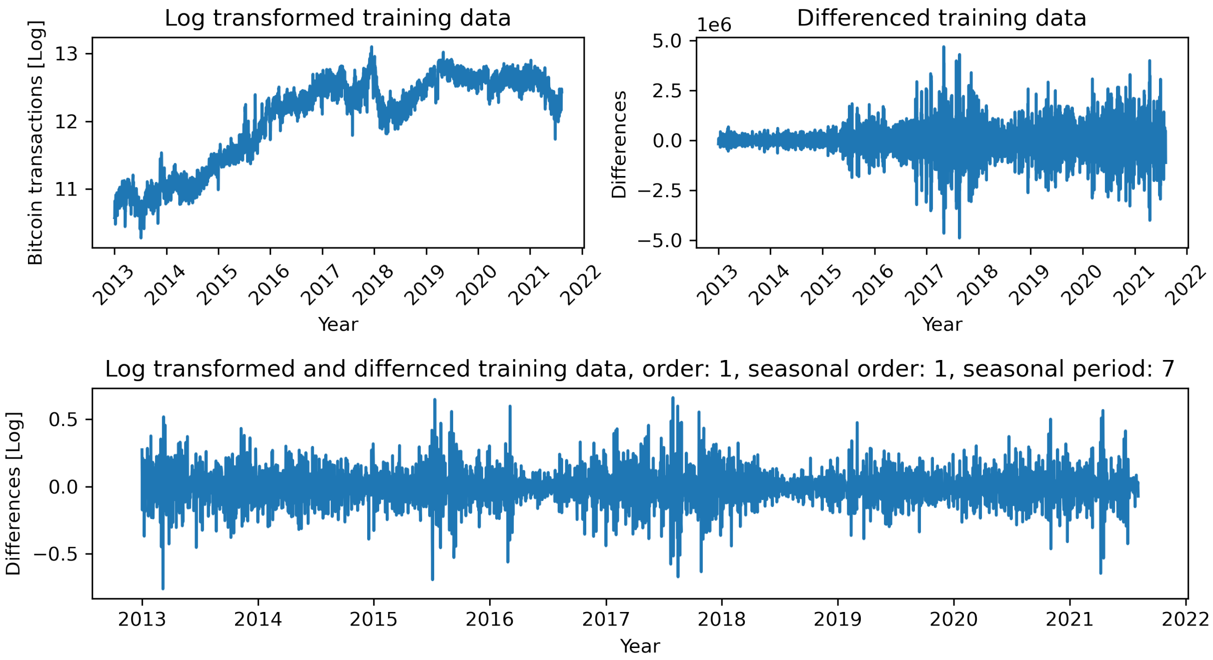

Figure 3: The transformed and differenced data set are seen in the bottom om the plot. The plot in the top left corner show the log transformed data. Note how the variance now seems to be the same through the whole time series, but the mean is changing. In the top right corner there is a plot of the differenced data. After the differencing of the data the mean is now at zero, but the variance is non-stationary. The bottom plot shows the data set after both transformation and differencing, now the time series looks to be stationary.

difference on the data, this time a seasonal difference of 1 with a seasonality of 7. This resulted in the autocorrelation plot decaying to insignificant values after lag 8 and the partial autocorrelation plot looking like an exponentially decaying sine. This suggest that the time series now is stationary and suitable for prediction. These plots can be seen in Figure 5.

In Figure 3 the differenced, transformed and the combination of both are shown. Here you can clearly see how the transformation is making the variance more constant and the difference is making the mean center around 0.

At this stage I know that both the normal and the seasonal part of the model needs to be differenced once in order to achieve stationary data. In the next section I will comment more on potential values for the auto regressive and moving average parts of the model.

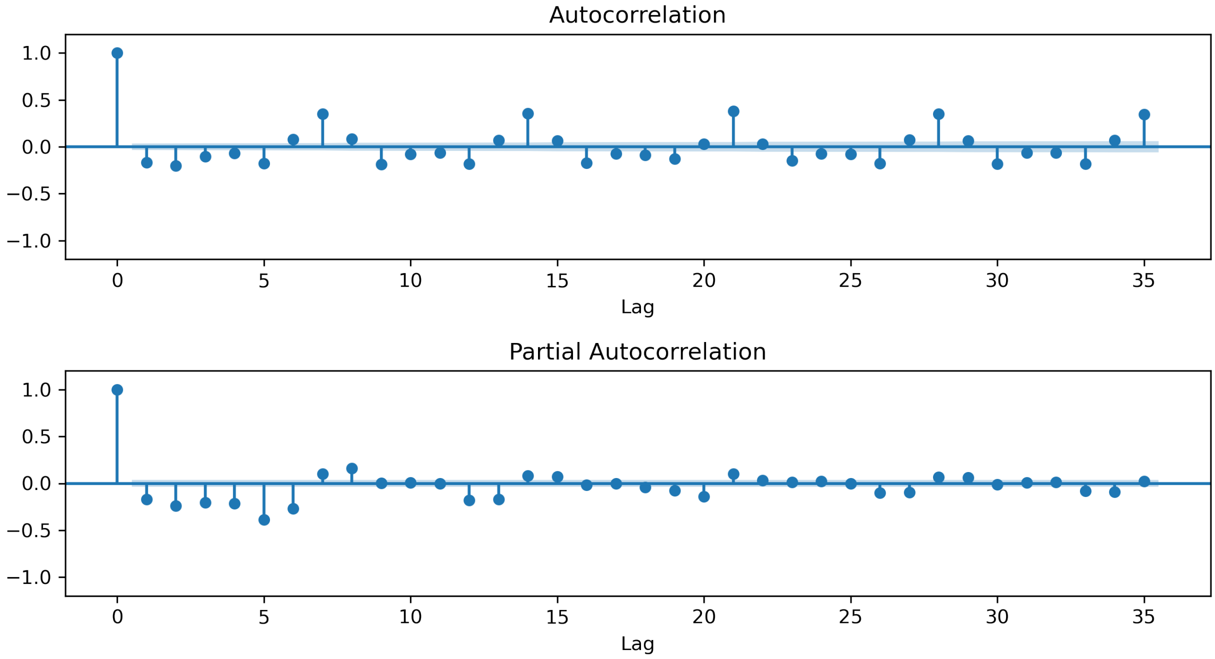

Figure 4: The autocorrelation and partial autocorrelation plots after the differencing of 1 and transformation of the data set. Now the autocorrelation plot decays quickly to zero which is a sign of non-stationarity, still has high spikes at a period of 7.

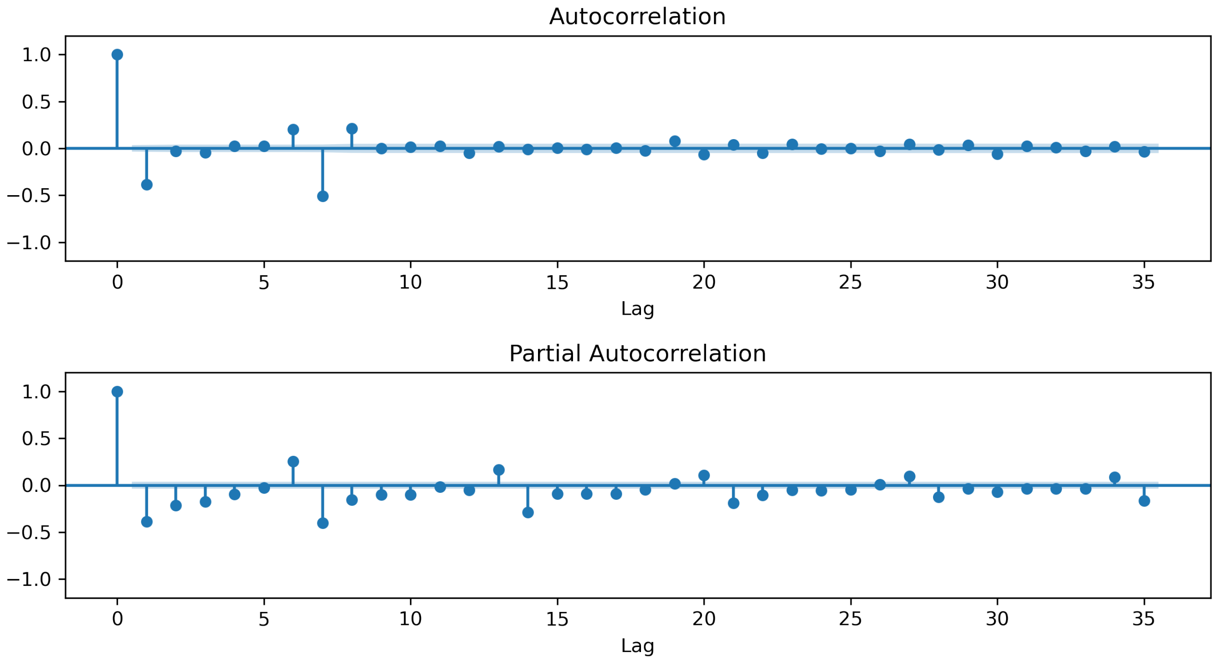

Figure 5: The autocorrelation and partial autocorrelation plots after the differencing of 7 and transformation of the data set. Now both plots decays quickly to zero and with no large correlations for later lags.

Question 3.3

In order to chose the best model order I will use table 6.1 in Time Series Analysis by Henrik Madsen as a starting point and check different models around that starting point, this is a procedure called grid search. Comparing Figure 5 and table 6.1 it seems like the model should be a SARIMA with only MA parts. This is because the PACF plot looks like a damped sine and the correlations in the ACF plot decays to zero after a number of lags.

A SARIMA model consists of two parts, one "normal" and one seasonal, both including an auto regressive part (AR), an integration part (I) and a moving average part (MA). The order of the AR is given by p for the "normal" part and P for the seasonal, the same with d and D for the I part and q and Q for the MA part. For this SARIMA model, with only MA parts the following equation should try to be fulfilled for the q and Q with a seasonality of 7:

$$q + 7Q \leq k, \text{where } k \text{ is ACF}(k) \approx 0 \tag{1}$$

From Figure 5 k looks to be 8 witch suggests both q and Q to be 1. Using this as a starting point I will search models with $p, P = 0$, $d, D = 1$ and $q, Q \in [1, 4]$ and try to find the best fitting model.

To identify the best fitting model I will look at two things, the Akaike information criterion (AIC) and the residuals on the training data. The reason why I will look at the residuals of the training data and not of the test data is because when fitting a SARIMA model the only known data is the training data and test data is not usually available.

The AIC is a popular criterion to evalute the performance of a model, it compares the number of parameters and the log likelihood of the model and favors the models with a low number of parameters and a high log likelihood. The AIC is given as:

$$AIC = 2k - 2ln(\hat{L}) \tag{2}$$

Here k is the number of variables and $\hat{L}$ is the likelihood function of the model. Thus the model with the lowest AIC will probably fit the data best.

For the residual part I will mainly look at the correlation of the residuals using both the autocorrelation plot and the Ljung-Box test to determine the best model order. We want no significant correlation between the residuals for as many lags as possible. If the models suggested by the AIC has a bad correlation structure in the residuals I will try to adjust them in order to get less correlation.

Question 3.4

Analysing the residuals is a good way of inspecting the fit of a model. The residuals of a good fitting model should be identically independently distributed (i.i.d.). Which means that they should look to be drawn from the same distribution with no correlation between each other. To inspect these properties I will plot the residuals, the autocorrelation function, use the Ljung-Box test, make a QQ plot and plot a histogram of the residuals.

The plot of the residuals of each lag can be used to inspect how the mean and variance of the residuals are. If the residuals are i.i.d., the residual plot should show the same mean and variance through all the lags. I.i.d also implies no correlation between the different residuals, thus the autocorrelation plot can be used to look for significant correlations. The Ljung-Box test can accompany the autocorrelation plot by providing a statistical test showing whether any of the autocorrelations of a time series are different from zero, statistically. Lastly the QQ plot and histogram can be used as an indicator of whether residuals looks to be from the same distribution or not.

Using the approach stated earlier the model with the lowest AIC was $\text{SARIMA}(0,1,3) \times (0,1,1)_7$. Figure 6 show the residuals of this model together with the autocorrelation of the residuals and the first 10 p-values from the Ljung-Box test.

By just looking at the plot of the residuals they seem to have a mean at zero for all lags. The variance looks to be constant for most of the lags, but looks to be bigger on the negative side than on the positive. The autocorrelation only shows one significant correlation at lag 7, other than that there are only insignificant correlations. This is also supported by the Ljung-Box p-value plot which shows p-values far above 0.05 for the first 7 lags.

The QQ plot and histogram are shown in Figure 7. When plotting the QQ plot it was clear that there were two maajor outliers in the residuals, these are excluded from the QQ plot to make it easier to read for the other residual. The QQ plot indicates that most of the residuals looks to be coming from the same normal distribution but with some outliers on the tails and also some extreme outliers that are excluded. The histogram looks to be shaped like you expect from a normal distribution with some bins of residuals far from the main bins, supporting the long tails seen in the QQ plot.

All in all I am not that satisfied with how the residuals looks for the final model. The residuals are only non-correlated for the first 6 lags and thus there are some information hidden in the correlation structure of the residuals that the model does not manage to capture. The fact that the QQ plot and histogram are showing that the residuals are not exactly normally distributed is also a concern. But it is important to take into account that when working on a time series like this, although the data are differenced to be made stationary, there are still days that differs significantly from other days, causing outliers in the residuals. Thus my major concern is the correlation of the residuals.

In Figure 8 the residuals, autocorrelation and p-values of the Ljung-Box test of the residuals of the $\text{SARIMA}(0,1,1) \times (0,1,1)_7$ model are plotted. This was the model I suggested just by looking at the ACF and PACF of the transformed and differenced data, Figure 5. As you can see this model performs quite similar to the chosen model. The difference is that the resiudals seems to be a bit more correlated, this can be seen in the Ljung-Box plot

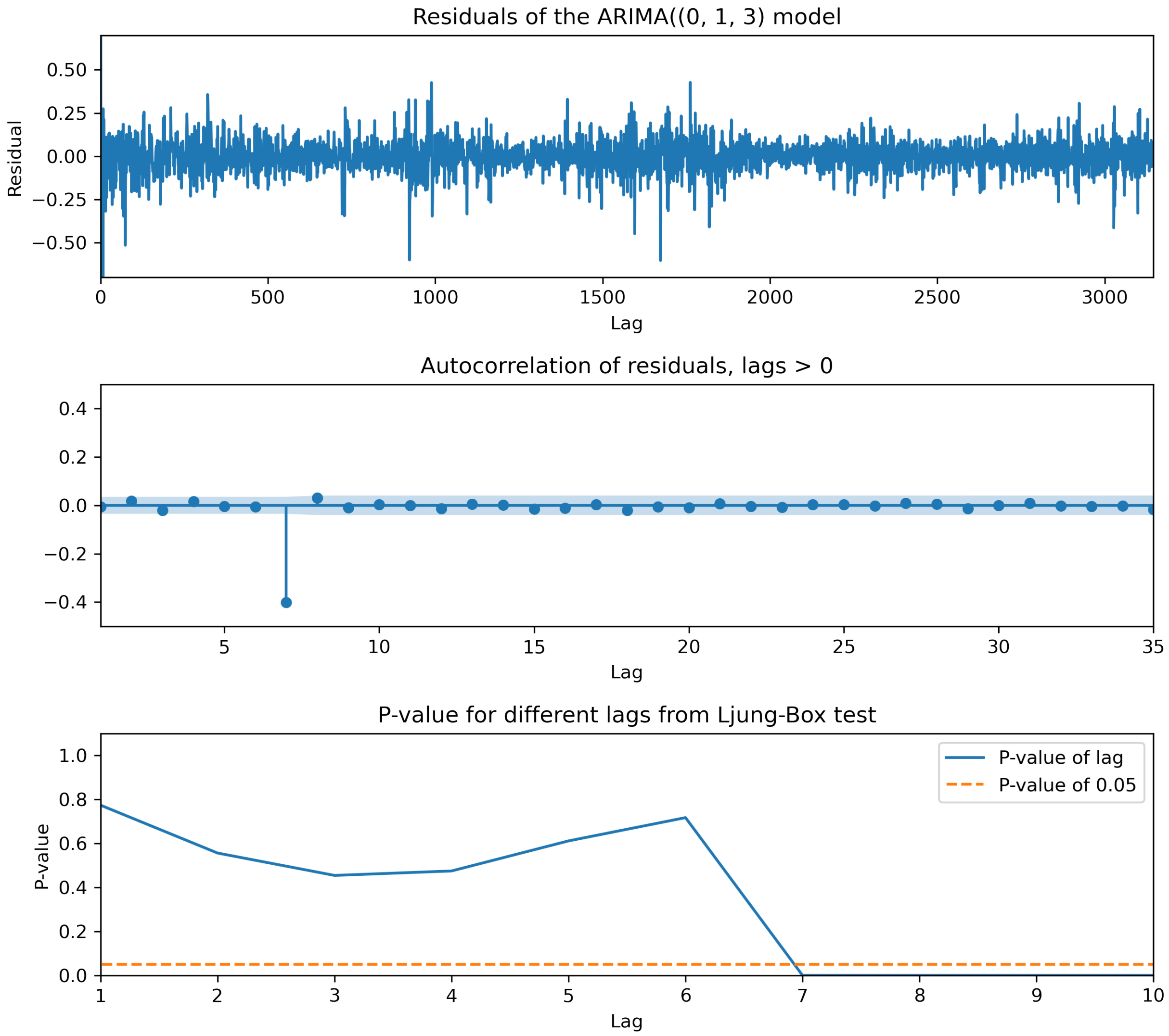

Figure 6: The residuals and autocorrelation between the residuals of the final SARIMA$(0, 1, 3) \times (0, 1, 1)_7$ model. The residuals looks to be normal distributed with the same mean and variance for all the different lags. And from the autorcorrelation plots there also looks to be small or no significant correlations between the residuals at different lags.

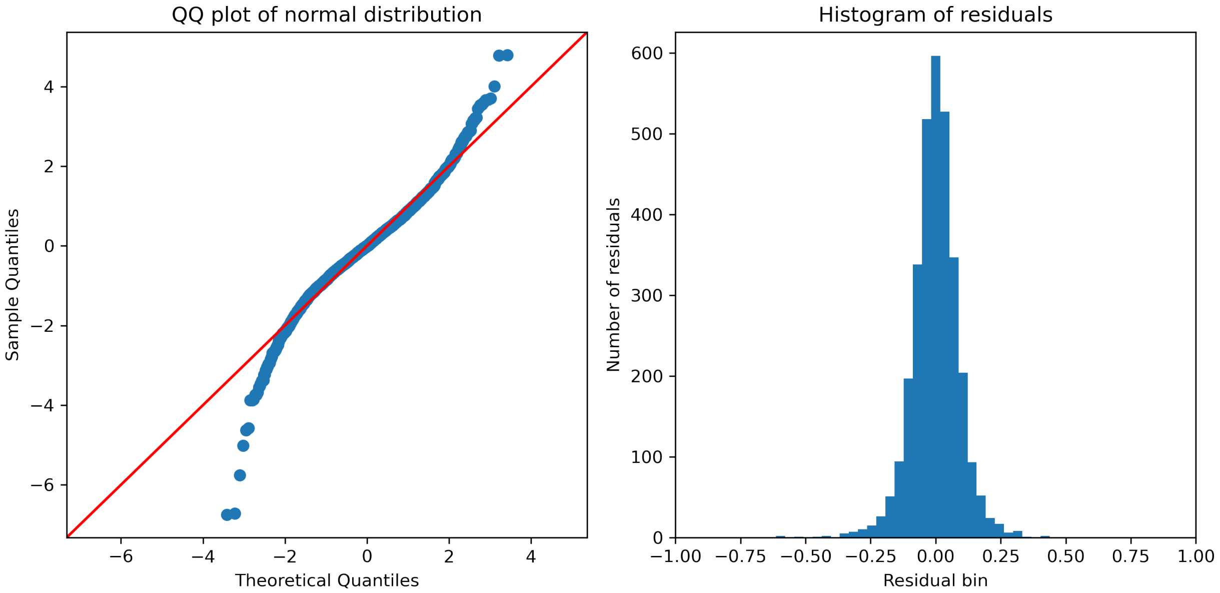

Figure 7: QQ plot and histogram of the residuals of the SARIMA$(0, 1, 3) \times (0, 1, 1)_7$ model. Both plots indicates that the residuals are mostly normally distributed.

showing lower p-values for all the different lags before 7. Thus I would say that the first model performs slightly better.

Why not include an AR term? In Figure 9 you can see the residuals, autocorrelations and Ljung-Box plot of SARIMA$(1, 1, 1) \times (0, 1, 1)_7$ here you can see that including an AR term in the model results in the Ljung-Box test showing correlations in the residuals already at lag 1. Thus there are lots of information still hidden in the residuals of the model.

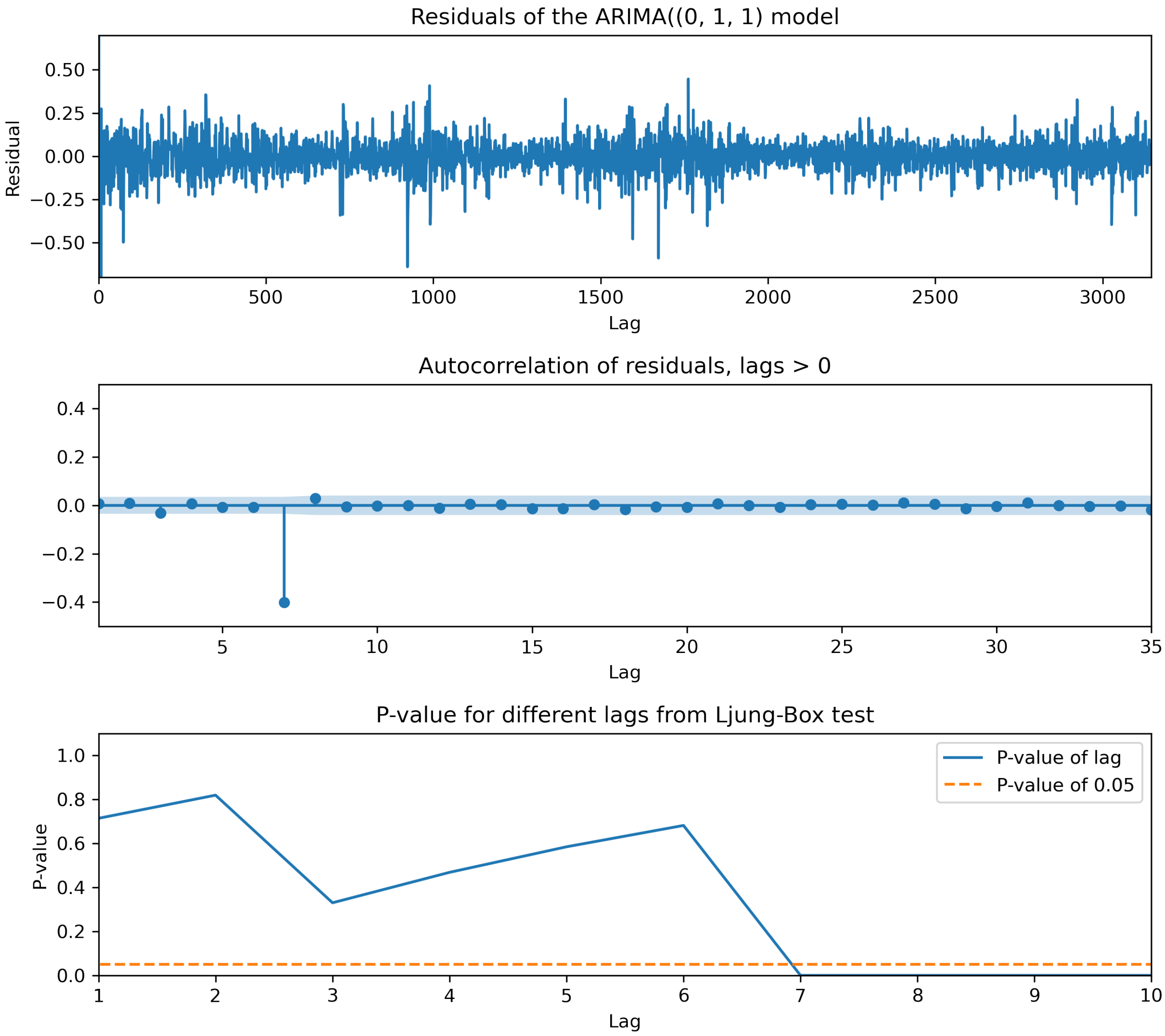

Figure 8: The residuals, autocorrelations and Ljung-Box plot of the SARIMA$(0, 1, 1) \times (0, 1, 1)_7$ model. Here you can see that the residuals seems to be a bit more correlated than in the final model.

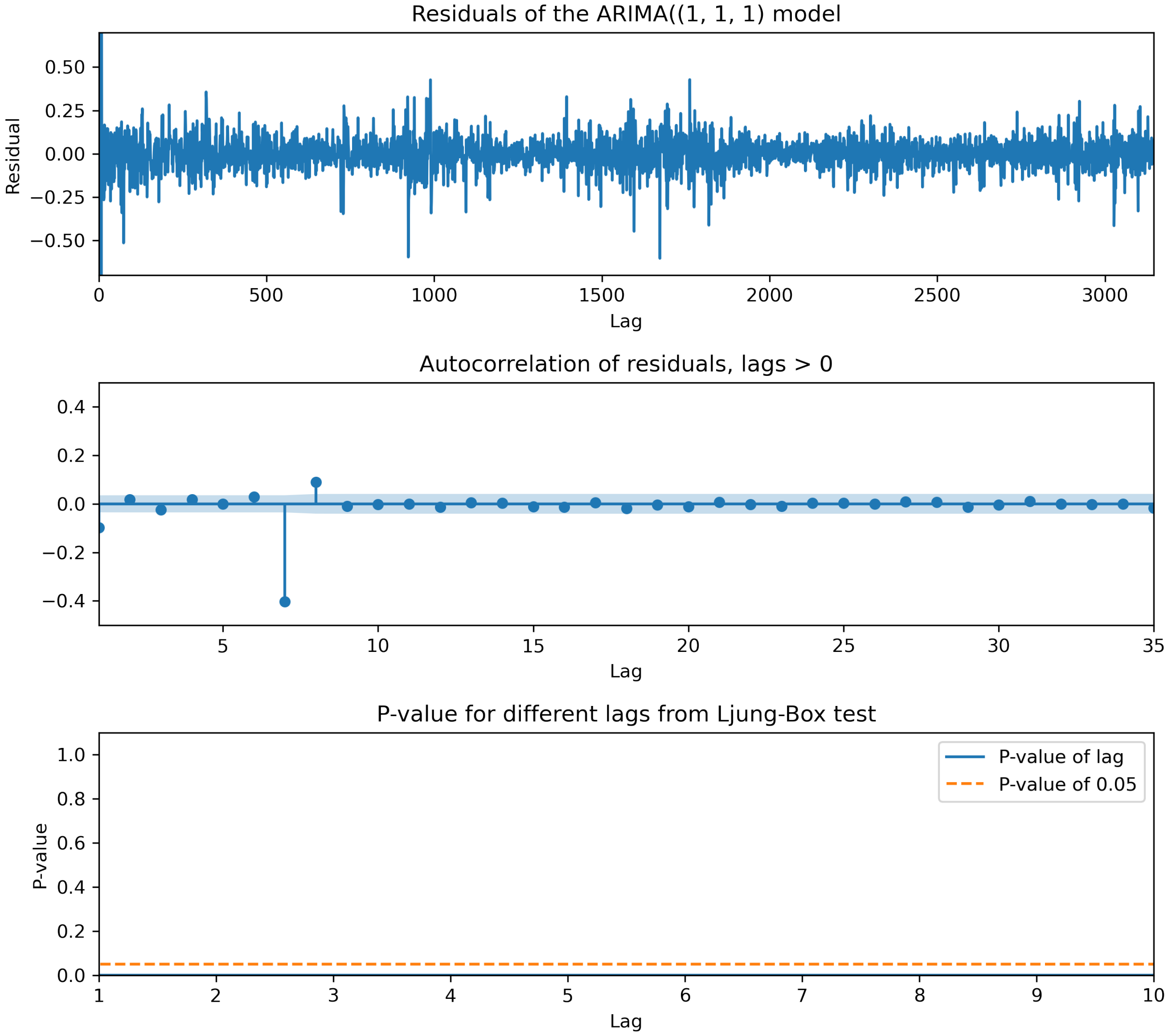

Figure 9: Residuals, autocorrelation plot of the residuals and the Ljung-Box plot of the SARIMA$(1,1,1) \times (0,1,1)_7$ model. This shows correlations between the residual indicating that the model is missing lots of information.

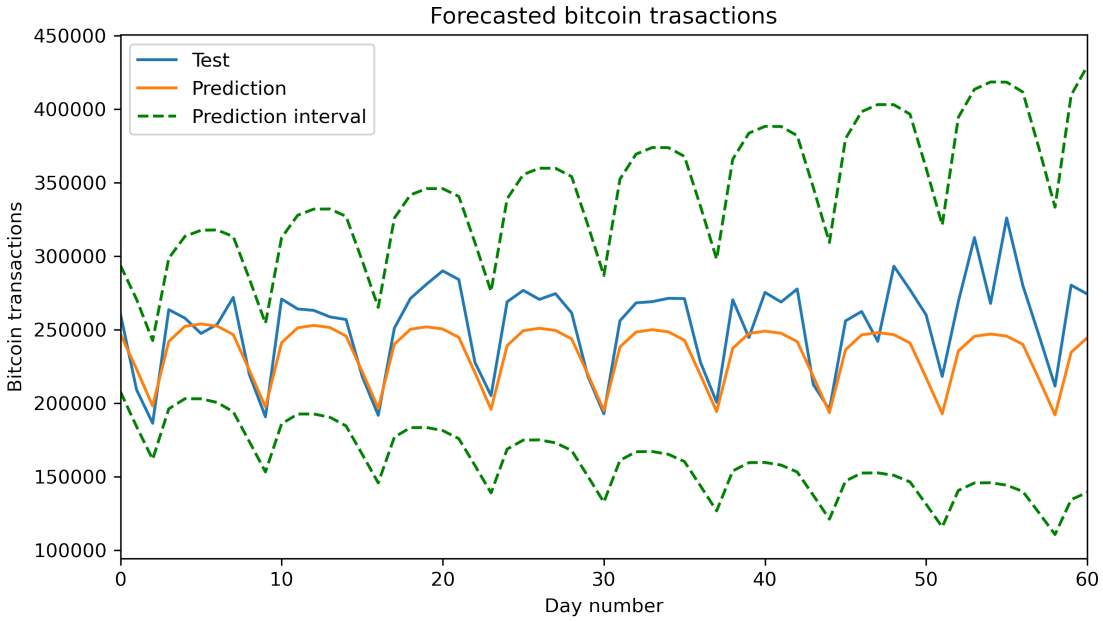

Figure 10: The true bitcoin transactions vs. the predicted transactions, including a prediction interval using the $\text{SARIMA}(0,1,3) \times (0,1,1)_7$ model.

Question 3.5

In Figure 10 you can see the predicted values for the next two months using the $\text{SARIMA}(0,1,3) \times (0,1,1)_7$ model. You can also see the true number of bitcoin transactions as well as a 95% prediction interval. Something to note from this is that the model is training on logarithmic transformed data, thus the prediction will also be logarithmic transformed resulting in different sized upper and lower prediction intervals.

The predicted values for day 1, 2, 14, 31 and 61 can are seen in Table 1. The upper and lower limits of a 95% prediction interval is also shown.

Day:	Prediction:	Lower:	Upper:
1	246936.06	207675.29	293619.02
2	223117.34	184003.60	270545.51
14	251360.87	190319.82	331979.55
31	194973.87	132580.29	286730.47
61	244418.79	139214.07	429127.22

Table 1: The predicted values of the $\text{SARIMA}(0,1,3) \times (0,1,1)_7$ model for a given day after the last day of the training set. Also including 95% confidence interval.

Question 3.6

The SARIMA$(0,1,3) \times (0,1,1)_7$ are performing ok, and improvements can be made. As in most other modelling cases increasing the number of training samples also increases the accuracy of the prediction. Thus including the test data and update the training data set daily will result in more and more accurate predictions. More training data is also achieved by increasing the resolution of the data, i.e. using hourly data instead of daily. This will also make it possible to find hourly trends in the data, increasing the accuracy even more.

Another way of increasing the performance of the model would be to introduce exogenous data. As the number of daily bitcoin transactions would be correlated with the bitcoin price, including information about the bitcoin price could be an idea. There may also be correlation between the price trend of bitcoin and the number of transactions. If the price is increasing or decreasing rapidly, the number of transactions is most likely higher than if the price remains stationary. Thus information about the trend of the bitcoin price may also result in better predictions.

List of Figures

References

[1] Madsen H.(2007) *Time Series Analysis*, Chapman & Hall/CRC.