Danmarks Tekniske Universitet

DTU

Assignment 3

Author

Emil Johannesen Haugstvedt - s211818

November 18, 2021

Contents

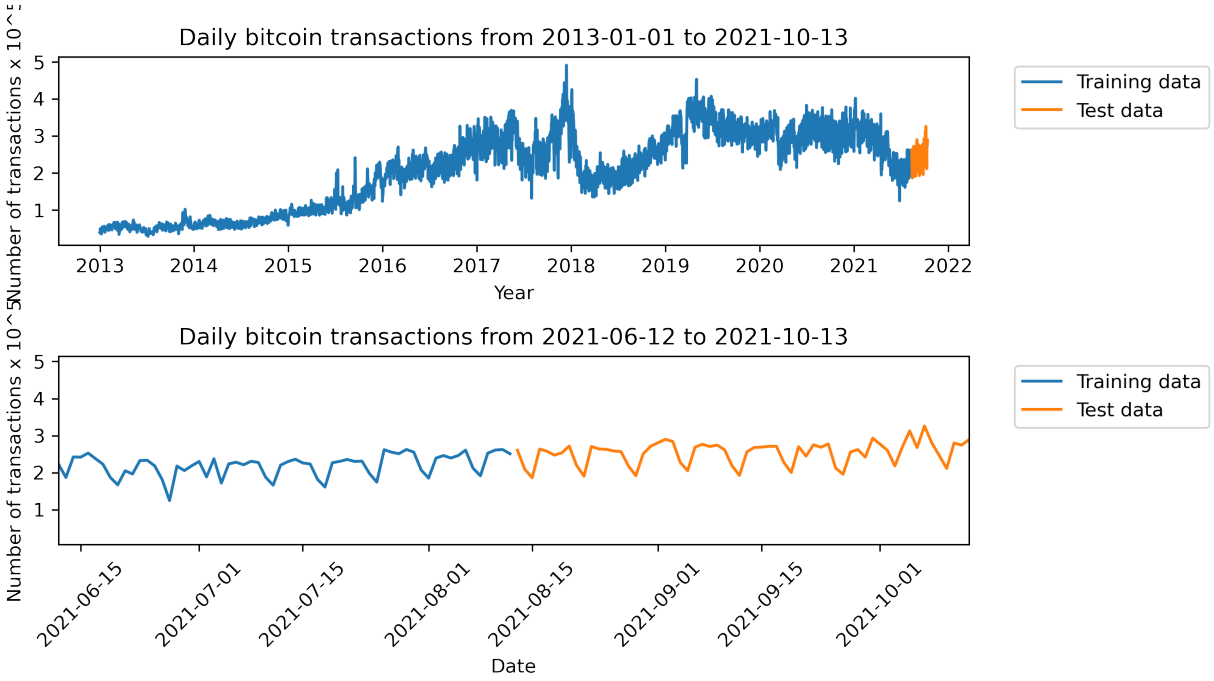

Figure 1: Daily bitcoin transactions from 01.01.2013 to 13.10.2021. The upper plot shows the whole period, while the lower shows the last four months. The training data used to fit the model is colored blue, while the test data is orange. In the upper plot, note how the data clearly is non-stationary 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.,

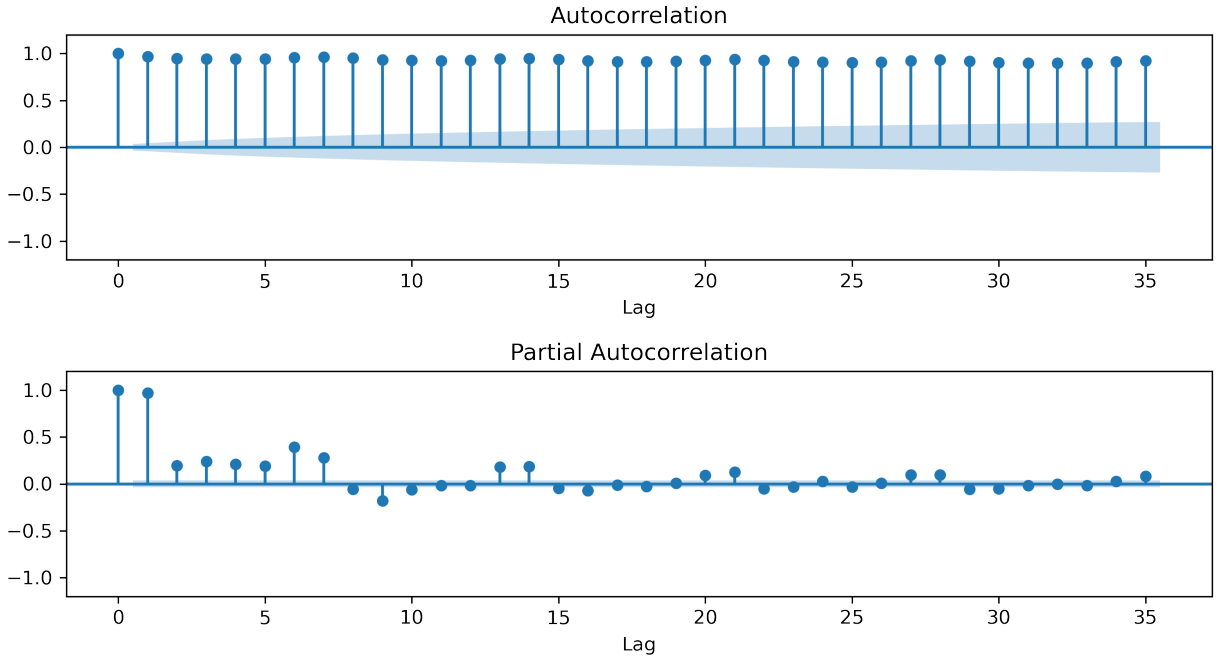

Figure 2: The autocorrelation and partial autocorrelation plot for the daily bitcoin transactions training data set. The non-stationarity of the data is supported by the consecutive positive correlations in the autocorrelation plot. This is an indication that differencing must be done in order to be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

The first order difference is shown below:

$$y'(t) = y(t) - y(t-1) \tag{1}$$

Secondly, the variance of the time series needs to be constant. A way to do this is the use the log transformation on the data set. 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.

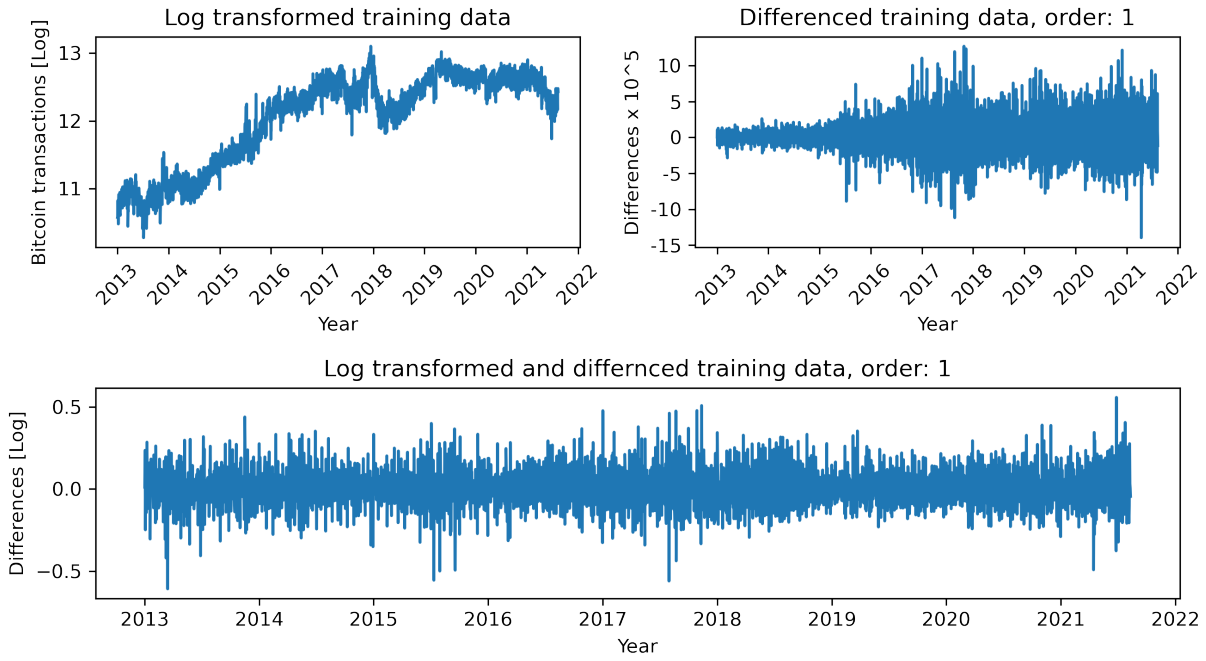

Figure 3: The transformed and differenced data set are seen in the bottom om the plot. The plot in the top left corner show the log transformed data. Note how the variance now seems to be the same through the whole time series, but the mean is changing. In the top right corner there is a plot of the differenced data of order 1. 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.

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.

After differencing and transforming the data set the autocorrelation and partial autocorrelation functions also change. The new functions are seen in Figure 4. Now both the autocorrelation and partial autocorrelation function quickly enters a sine-looking pattern, something that indicates a stationary process. From this I can conclude that it is possible to fit an ARIMA model to this time series if I difference the time series by one and do a log transformation.

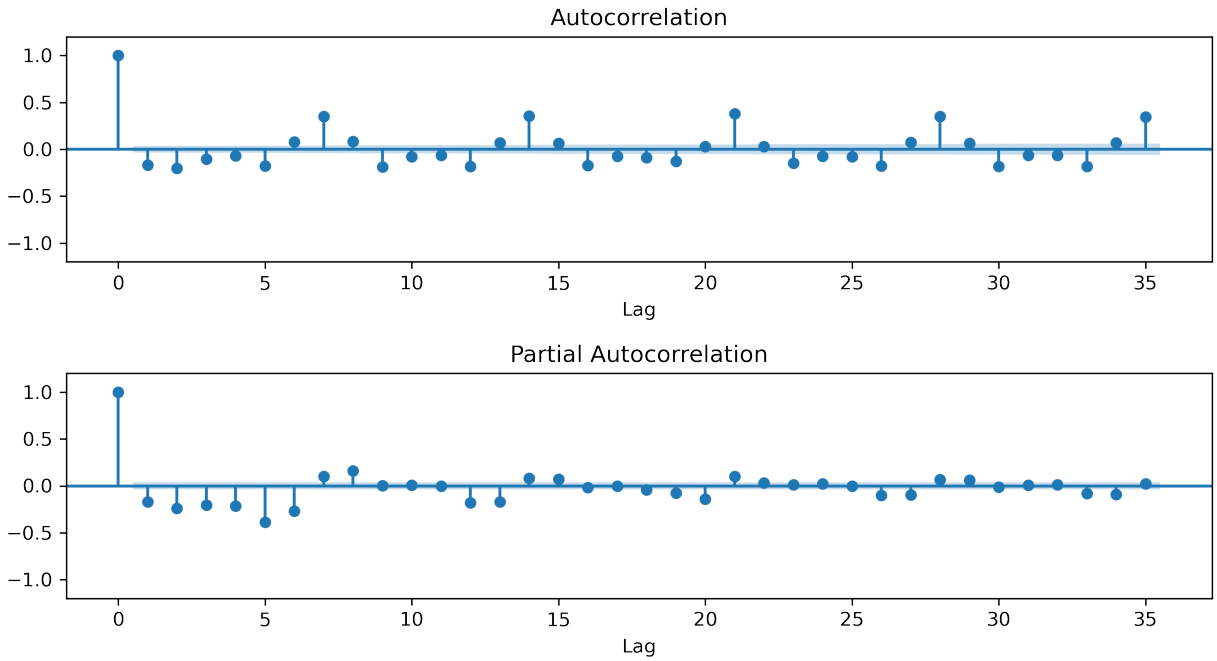

Figure 4: The autocorrelation and partial autocorrelation plots after the differencing and transformation of the data set. Now the autocorrelation plot decays quickly to zero which is a sign of non-stationarity. It also looks to be possible to identify an ARIMA model based on the plots.

Question 3.3

An ARIMA model is made by specifying the order of the AR (auto regression) part, the I (integration) and MA (moving average) part. These orders are specified as p, d and q, respectively. From the last section it is clear that the time series needs to be differenced once in order to be made stationary, this implies that d is 1, and leaves us with identifying the p and q that best fits our data.

The ACF and PACF plots in Figure 4 can be used as indicators of the order of p and q by looking at the number lags before the correlation decays to insignificant values or enters a sine-looking pattern in the PACF and ACF, respectively. Looking at Figure 4 the order of p looks to be around 6 and the order of q looks to be around 5.

Other then just looking at the ACF and PACF plots there are criteria that can be used to inspect how well the model is fitting the training data. Some popular matrics is the Akaike information criterion (AIC) and the bayesian information criterion (BIC).

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

$$BIC = kln(n) - 2ln(\hat{L}) \tag{3}$$

In both the above equations k is the number of variables and $\hat{L}$ is the likelihood function of the model. The difference between the criteria is that BIC takes the number of samples into account by introducing the n. Thus these criteria can be used to find which model fits the training data best with the least number of parameters. From this the model that fits the data best is the model with the lowst AIC and/or BIC. In some cases the AIC and BIC may disagree, then we will need to take other factors into account as well.

With this in mind I will determine the order on my model based on the AIC and BIC of the model. I will use the order suggested by the ACF and PACF as a starting point and check different models of lower and higher order, before using the model order of the model with the lowest AIC and BIC values.

Fortunately, fitting ARIMA models using statsmodels in Python does not take a long time. Therefore, to identify the right model structure, I fitted all possible models with $p, q \in [0, 10]$ and evaluated their performance. Then I saved all the scores for the different models and choose the right order based on the AIC and BIC scores.

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 beetween eachother. 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 for all 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. Lastly the QQ plot and histogram can be used as an indicator of whether residuals looks to be from the same distribution or not.

By using the approach stated earlier the best model was determined to be ARIMA(7, 1, 7). Figure 5 show the residuals of this model together with the autocorrelation of the residuals and the first 20 p-values from the Ljung-Box test. The residual of lag 1 is excluded from the plot due to a much higher value than the other residuals as a result of no prior predictions available at that stage. And since the autocorrelation for lag 0 always is 1, this is also excluded from the autocorrelation plot.

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 shows no significant correlations. This is also supported by the Ljung-Box p-value plot which shows p-values far above 0.05 for the first 20 lags, something that indicates no autocorrelation between lags. The QQ plot and histogram are shown in Figure 6 and indicates that most of the residuals looks to be coming from a normal distribution. The QQ plot indicates that there are some outliers at each tail of the distribution. 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 satisfied with how the residuals looks for the final model. The residuals are non-correlated and thus the model manages to capture most of the information hidden in the correlation structure of the model. Also the QQ plot and histogram shows that the residuals seem to be almost normal distributed. The fact that they are not perfectly normal distributed is expected 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, and thus also outliers in the QQ plot and histogram.

Trying to identify an appropriate model is a careful consideration of both model order and performance as we want the best model of the lowest order. A natural approach when first exploring the different model orders is to start with lower ordered models. By looking at the ACF and PACF plots in Figure 4 a good start could be a ARIMA(3, 1, 3) model.

In Figure 7 the residuals, autocorrelation and p-values of the Ljung-Box test of the resid-

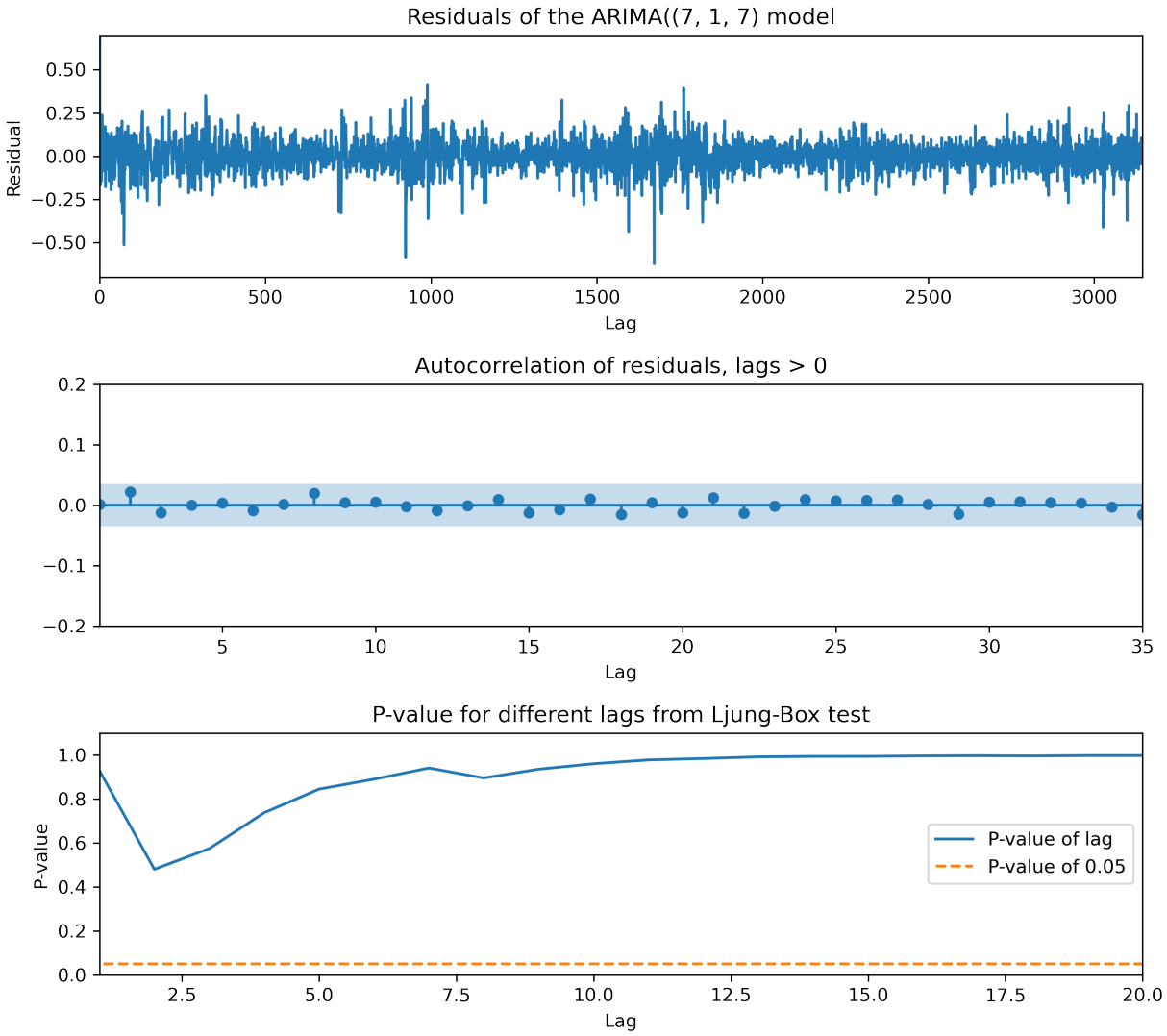

Figure 5: The residuals and autocorrelation between the residuals of the final ARIMA(7, 1, 7) model. The residuals looks to be normal distributed with the same mean and variance for all the different lags. And from the autorcorrelation plots there also looks to be small or no significant correlations between the residuals at different lags.

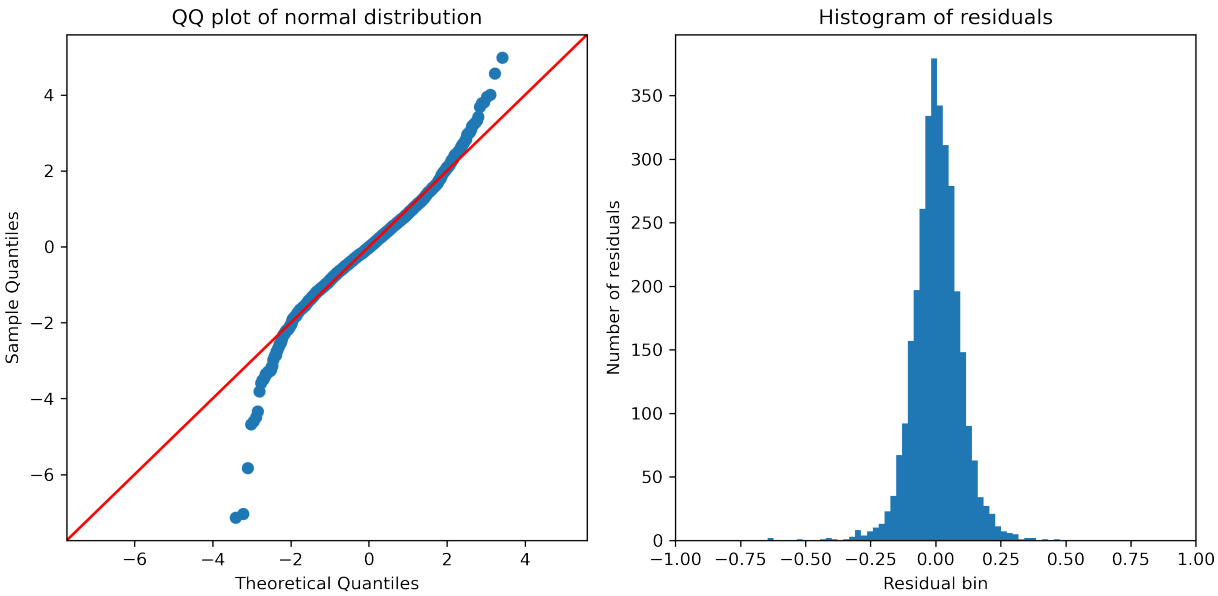

Figure 6: QQ plot and histogram of the residuals of the ARIMA(7, 1, 7) model. Both plots indicates that the residuals are mostly normally distributed.

uals of the ARIMA(3, 1, 3) are plotted. Just by looking at the residual plot the residuals still looks to be as i.i.d. as in the ARIMA(7, 1, 7) model, but when looking at the autocorrelation plot and the Ljung-Box p-values it is clear that there are significant autocorrelations between the residuals. The Ljung-Box p-value quickly drops to below 0.05 indicating correlation in the residuals. Thus there are information hidden in the correlation structure of the residuals, something that can be exploited by increasing the order of the model.

After determining the appropriate model to be ARIMA(7, 1, 7) it is also useful to see what happens when increasing the model order. Thus I tried using a model of the highest order considered in the model choosing algorithm, ARIMA(10, 1, 10). In Figure 8 you can see the residuals and autocorrelation residuals for this model. From these plots it is clear that this models performs as good as the ARIMA(7, 1, 7) model, the model determined as the best suited model. Here you can see how the AIC comes into play, we prefer the model with the lowest number of parameters in order to reduce the overfitting on the training data. Thus the ARIMA(7, 1, 7) will probably predict a more realistic future based on the training data, and is then also a better model.

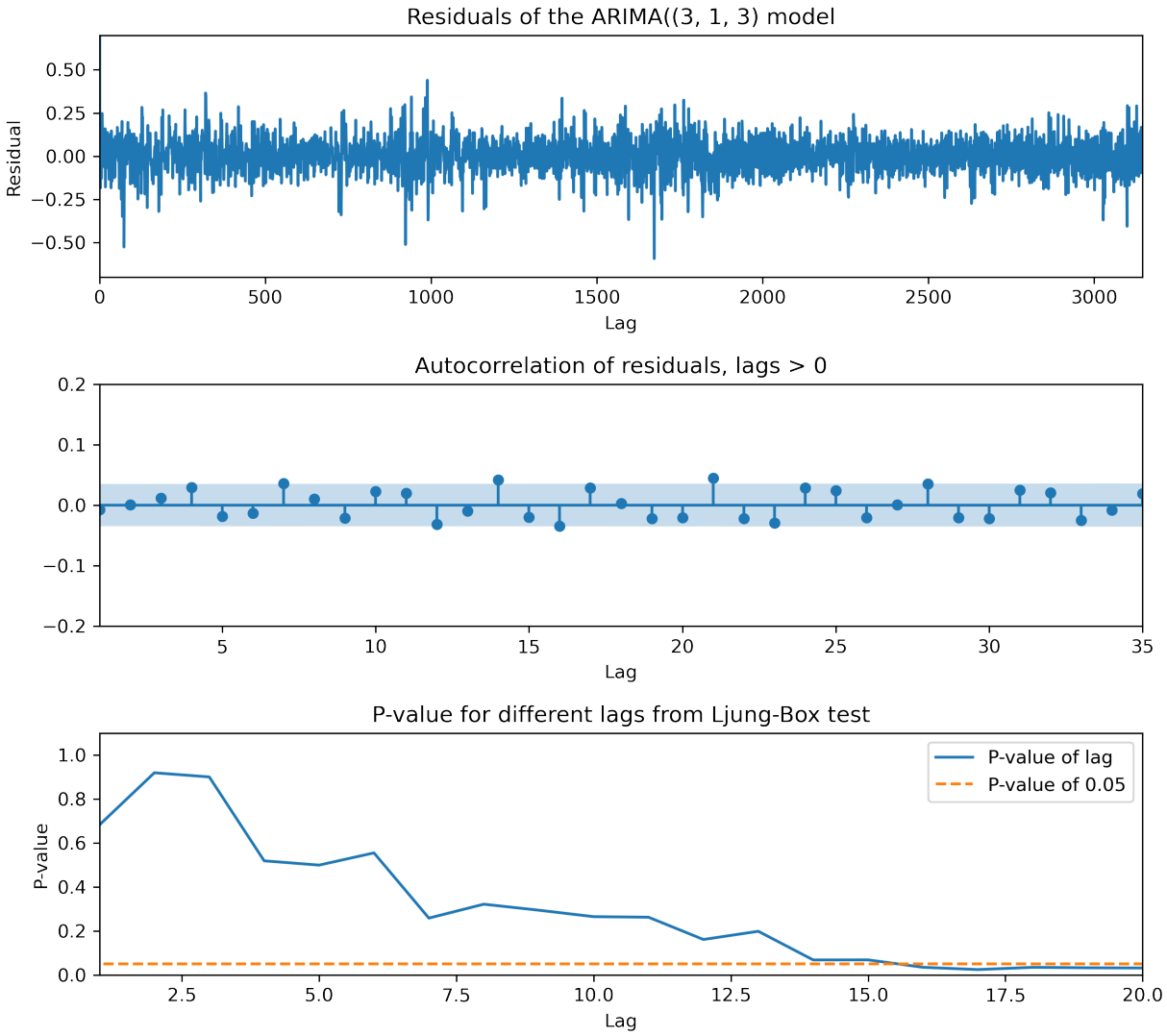

Figure 7: Residuals, autocorrelation and p-values from the first 100 lags of the Ljung-Box test of residuals for the ARIMA(3, 1, 3) model. Note how some of the lags in the residual plot shows siginficant correlations and the p-value of the Ljung-Box test quickly goes below 0.05. Thus there will be me information in the correlation structure of the residuals.

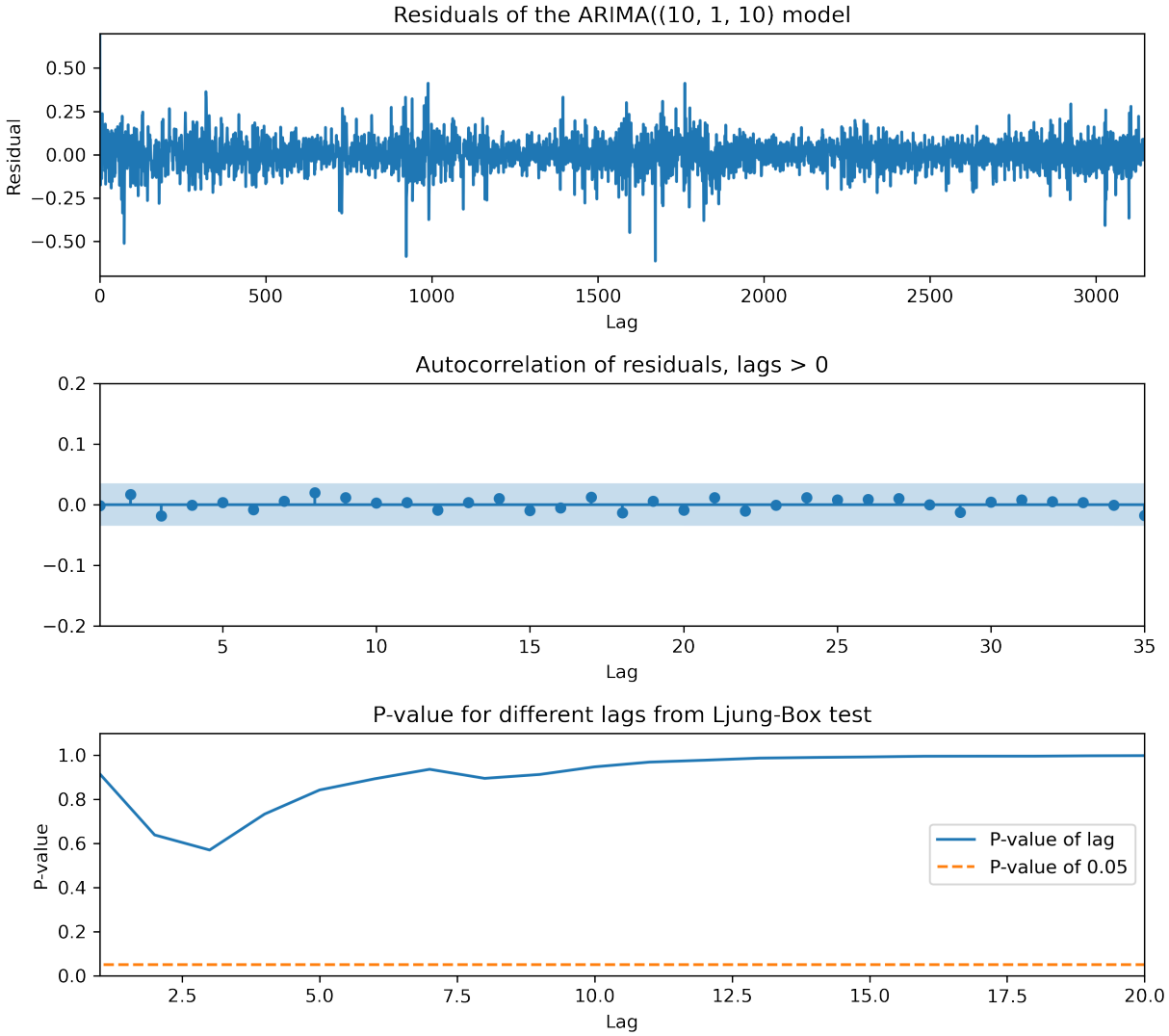

Figure 8: Residuals of the ARIMA(10, 1, 10 model. Note how the performance looks to be almost the same as the ARIMA(7, 1, 7) (Figure 5), but with much more parameters.

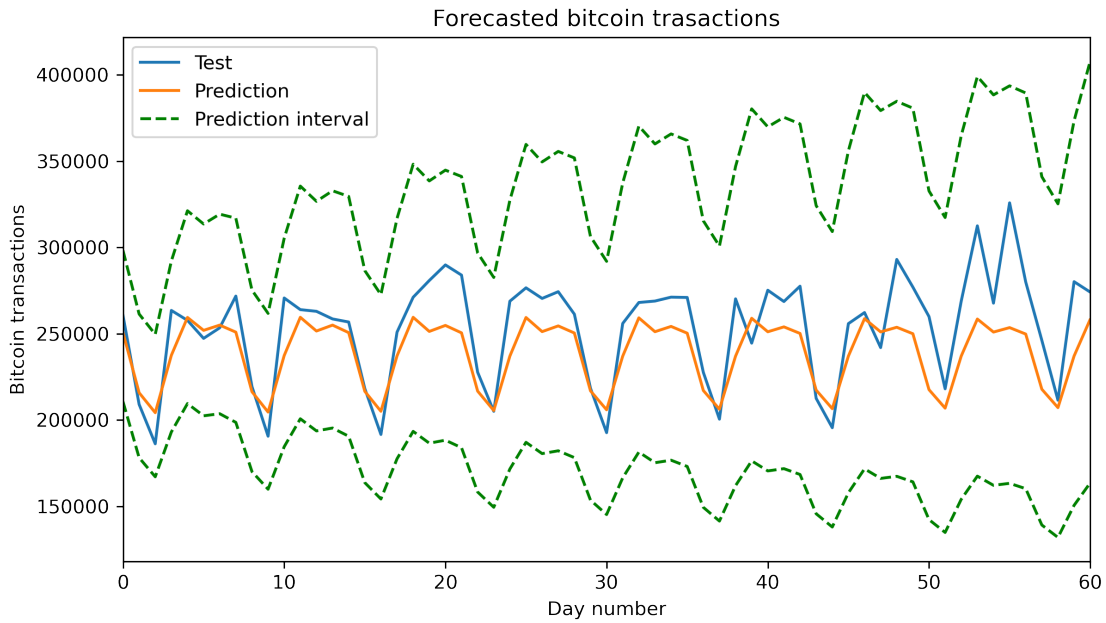

Figure 9: The true bitcoin transactions vs. the predicted transactions, including a prediction interval.

Question 3.5

In Figure 9 you can see the predicted values for the next to months using the ARIMA(7, 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. To produce normal predictions the predictions from the model needs to be transformed back to normal scale by using the exponential. This is the reason behind the upper and lower prediction interval having different distances from the prediction.

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	250666.03	210662.73	298265.66
2	215833.46	178175.43	261450.65
14	255016.37	195366.92	332877.99
31	205936.33	145246.37	291985.07
61	258350.59	163583.73	408017.52

Table 1: The predicted values of the ARIMA(7, 1, 7) model for a given day after the last day of the training set. Also including 95% confidence interval.

Question 3.6

Altough the the ARIMA(7, 1, 7) are performing good, improvements can be made. By looking at the zoomed plot in Figure 1 it is clear that there exists some weekly seasonality in the data. This could be exploited introducing a seasonal term in the model.

As in most other modelling cases increasing the number of training samples increases the accuracy in 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 accuraxy even more.

There are possibly correlations between the price of bitcoin and the number of daily transactions, and most likely correlations between the price trend and daily transactions. And the bitcoin price and trend may also be correlated with the prices and trends in the regular stock market. Including these exogenous factors into the models will give the model more information and thus maybe better predictions.

List of Figures

References

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