2022 Fall-STAT443

Daily traffic in one Junction Analysis

Xiangru Tan

Mark Liu

Jiawei Wang

Jerry Han

Responsibilities

Xiangru Tan: Dataset finding and Data cleaning, Intro of the report and Conclusion of the report.

Jiawei Wang: Analysis of regression, Regression part of the report

Mark Liu: Analysis of Holt Winters, Holt Winters part of the report

Jerry Han: Analysis of SARIMA models, SARIMA part of the report

**Motivation**

Nowadays, with more and more people owning their own car and the expanding unban populations, traffic congestion has become a more and more serious problem. In this project, we are interested in how the traffic condition has been changed on a daily basis and how it will be like in the future. With this prediction, we can better handle future traffic congestion which can greatly save people’s fuel, time, and money.

**Data Preprocessing**

The dataset we collected comes from Kaggle ( <https://www.kaggle.com/datasets/fedesoriano/traffic-prediction-dataset?resource=download> ), it is an hourly data collected in 4 junctions of roads starting from 2015-11-01 00:00:00 to 2017-06-30 23:00:00. Since we are interested about the daily traffic data, we sum all 24 hours together to make a daily data. Since all 4 junction tends to be similar, we can randomly pick one of the junctions to analyze, in this project, we pick junction 1. After summing the data to a daily data and picking junction 1, we have in total 608 datapoints.

With so many datapoints, let’s first look at a slice of the data to get a sense of the data, the plot of the whole data can be seen in the Appendix A.

图表, 折线图

描述已自动生成

The data seems to contain a trend and a seasonality, and the variance is not constant.

The ACF of the data can be seen below.

图表, 直方图

描述已自动生成

We can see a trend and seasonality since the ACF plot has a slow decay and a periodic behavior.

Moreover, if we run Flinger’s test, the p-value is , so the variance is indeed not constant.

After doing a search for the largest p-value we found that the log transformation has transformed the data to have a constant variance, and the Flinger’s test p-value has increased to (p-value being the largest doesn't mean the best transformation, but the plot and Flinger’s test seems to show a constant variance, so we will go with this value). The plot of a slice of the data is below. We will use this transformed data for all the following modelling.

图表, 折线图

描述已自动生成

After transformation, we can reconfirm the trend and seasonality using classic decomposition. We have seen that the range of the seasonality is greater than the range of the random part, so seasonality is present. (Plot of the classic decomposition can be seen in Appendix A).

For the rest of the project, we need to split the data into training and testing set, we do an approximate 95%-5% split. So, the first 82 weeks are in training set and the rest to be in the test set.

**Modelling**

**Regression**

In the regression modelling, we will look methods of Least Squared and Elastic Net (combination of Lasso and Ridge) to fit the training data. We choose the best model base on prediction error.

**Least Square model**

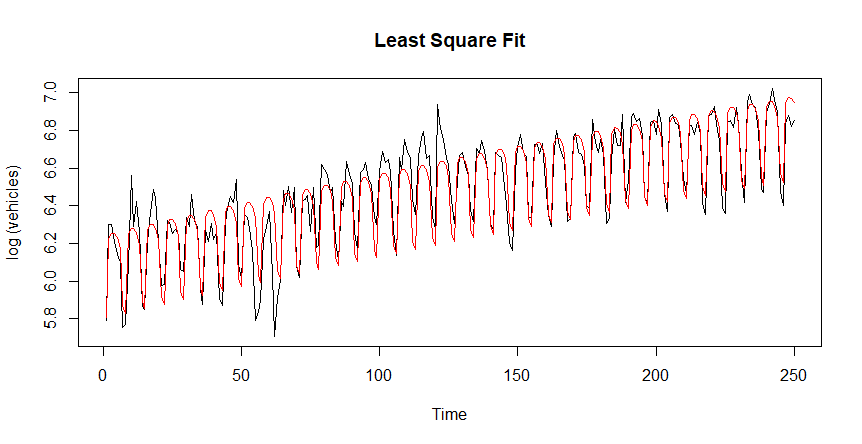
where

We also restrict the maximum polynomial degree , and based on MSE on the transformed data (as a whole) we found the second-order polynomial model with outperforms the others.

Chart, line chart

Description automatically generated

The fitted model superimposed on the transformed training data looks like it estimates the trend and seasonality, and we include a snippet of the fitted data which is the first 250 datapoints (See Appendix B for the whole fitted data).



We have the LS regression model, now we plot its residuals and its ACF. We hope the residuals looks stationary.

图表, 直方图

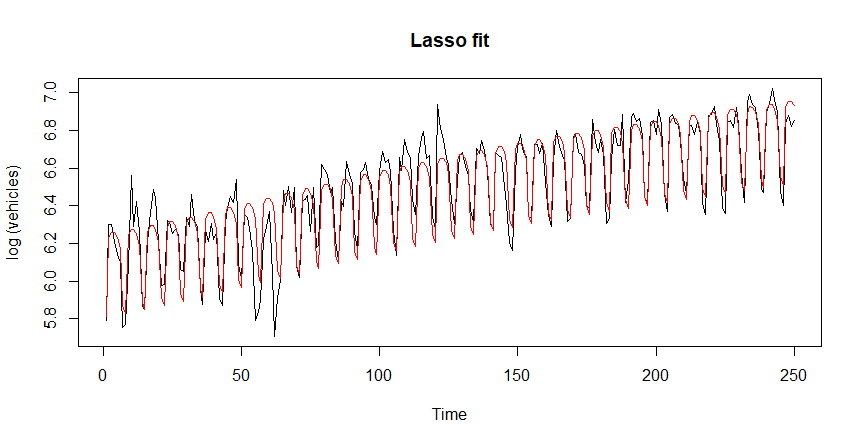
描述已自动生成

It is not clear that whether the ACF has exponential decay (stationary) or slow decay. We need to consider both probabilities. If we regard this as a slow decay, then we need to do differencing to remove the trend.

**Shrinkage Methods**

We fit LASSO, ridge and elastic net based on prediction error, we tested for from to , inclusive with step size , and we found that the LASSO has the smallest MSE. (See Appendix B for the MSE plot verses ).

Here we demonstrate the best fitted Lasso plot with polynomial degree 6 and the first 250 datapoints (See Appendix B for the whole fitted data).

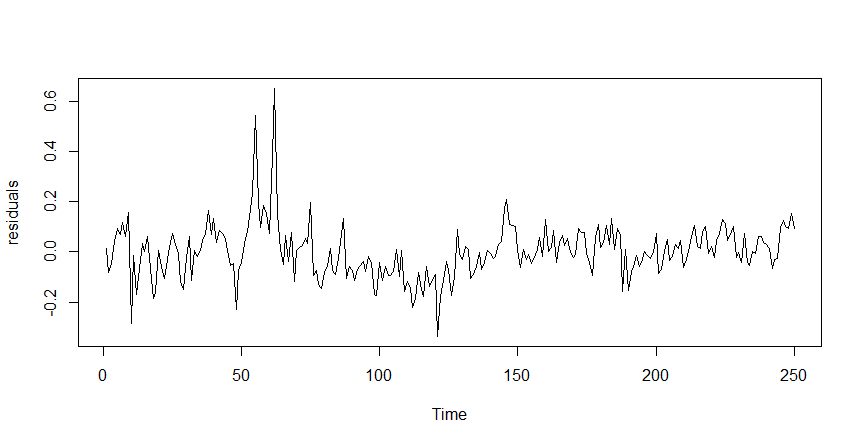


Since the fitted Lasso model looks quite like second-order polynomial regression, which is a much simpler model. By the parsimony rule, we will go with the simpler model.

**SARIMA analysis for Residuals of Regression**

We perform SARIMA Analysis on the second order least square polynomial, including the indicator for season.

If we regard ACF as an exponential decay, we do no differencing.



图表, 直方图

描述已自动生成

Since both the ACF and PACF are not significant at any seasonal lag, we propose ARIMA instead of proposing SARIMA. Since it is already stationary, it is equivalent to ARMA model:

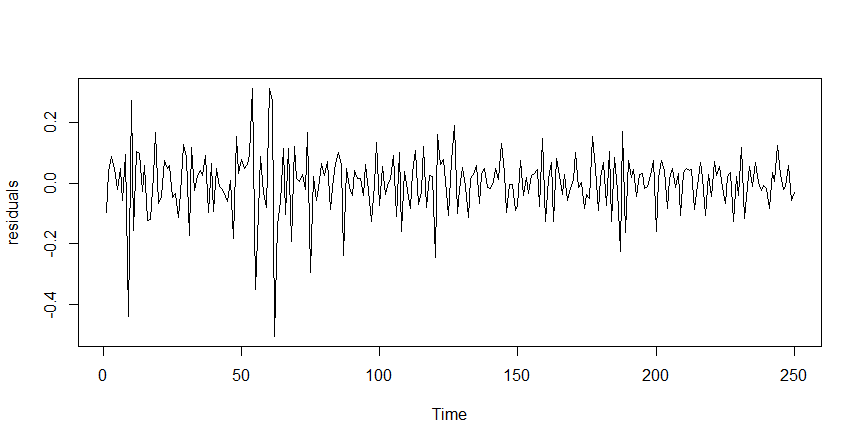
* The ACF looks like exponential decay while PACF cuts off after lag , so we propose .
* The ACF looks like exponential decay, and PACF tails off significantly so it may also be an exponential decay, so we propose .

Chart

Description automatically generated with medium confidence

We use as a demonstration. There is no obvious trend and non-constant variance in standardized residuals, ACF plot show no correlation, and all p-values for Ljung-Boxd statistic is greater than 0.05. We are not worried about the normality as we won’t do prediction interval. (See Appendix B for diagnostics for )

If we regard ACF as slow decay, we do one regular differencing.



图表

描述已自动生成

The ARIMA models we proposed are: (we do one-time regular differencing so it’s an ARIMA)

* ACF cuts off after lag and PACF looks like exponential decay, so we propose .
* Both ACF and PACF looks like exponential decay, so we propose .
* ACF looks like tailing off, thereby exponential decaying. PACF cuts off after lag , so we propose .

Note that the model does not pass the residual diagnostics. The reason is all p-values for Ljung-Box statistic below .

图表

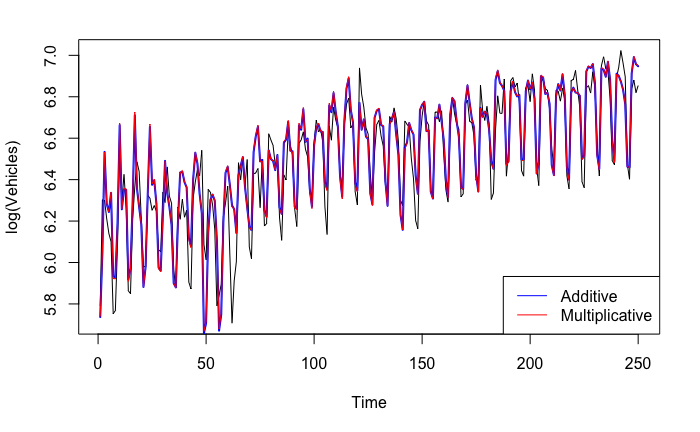
中度可信度描述已自动生成

Hence, we proposed the following models for regression:

1. Second order regression (include seasonality) followed by
2. Second order regression (include seasonality) followed by .
3. Second order regression (include seasonality) followed by .
4. Second order regression (include seasonality) followed by .

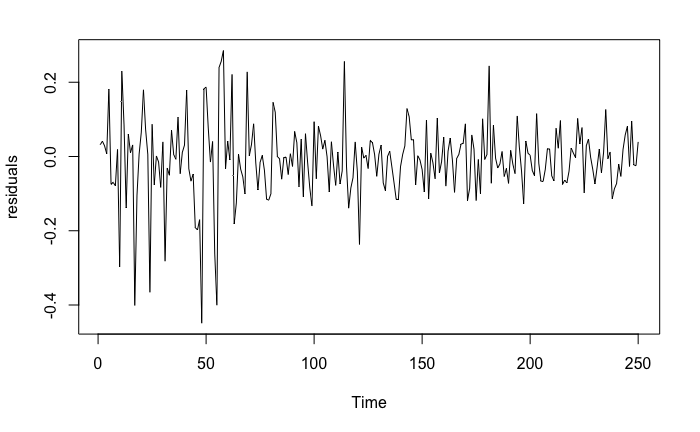
**Holt Winters**

Now we apply Holt-Winters Algorithm to estimate the trend and seasonality in the data. In either additive or multiplicative case, we found they both describe trends and seasonality well. We can see in the plot below that the additive and multiplicative Holt Winters acts similar (For visual purpose, we take the first 250 datapoints).



**SARIMA analysis for Residuals of Holt Winters**

For simplicity, in the following analysis, we will use the residuals in additive case.



图表

描述已自动生成

Since both the ACF and PACF are not significant at any seasonal lag, we propose ARIMA instead of proposing SARIMA. Since it is already stationary, it is equivalent to ARMA model:

* We might think the ACF has damped sine wave, and the PACF is cut off after lag 4, so we can propose .
* We might think the PACF has damped sine wave, and the ACF is cut off after lag 4, so we can propose .
* We might think both the ACF and PACF have damped sine wave, so we start with , and end up trying , and .

All model except for passed Residuals Diagnostics. Using as an example:

图表

中度可信度描述已自动生成

We see that standardized residuals do not show any trend or seasonality. Moreover, we have 1 out of 35 lags of ACF lies outside of 95% confidence interval, which is fine since we can expect , or 1 or 2 lags, to be outside of the CI. Moving to the Q-Q plot, we find that both tails do not lie in the shaded area. Finally, the p-values for Ljung-Box statistic are all above 5%. In conclusion, the assumption of constant mean, constant variance and uncorrelatedness holds for model, while normality of the data does not hold true. However, normality only affects prediction power of the model, so the fitting power of model is overall satisfied. (See Appendix C for Residuals Diagnostics of other models.)

is the only one that fails the Residuals Diagnostics. We see that some p-values for Ljung-Box statistic are below 5%

图表

描述已自动生成

Hence, we choose the following model:

1. Additive HW followed by .
2. Additive HW followed by .
3. Additive HW followed by .
4. Additive HW followed by .
5. Additive HW followed by .

**SARIMA(Differencing)**

From the ACF plot of first order regular differencing (See appendix D), we can see a periodic pattern. Hence, first order differencing with lag 1 is not a good choice. Let’s try first order differencing on the season.

Looking at the ACF plot of one order differencing on the season, we are not sure if there is a damped sine wave or a slow decay. We have either reached stationary or first order differencing on the lag of season is not good enough.

图表

描述已自动生成

Let’s assume we have reached stationarity.

图表, 直方图

描述已自动生成

We can propose the following 2 models since we have only done first order differencing on the lag of season.

* When ignoring the seasonal lag, the PACF and ACF both showed a damped sine wave. When looking at the seasonal lag, the PACF plot showed a damped sine wave, and the ACF plot has an exponential decay, so we can propose .
* When ignoring the seasonal lag, the PACF and ACF both showed a damped sine wave. When looking at the seasonal lag, the PACF plot shows a clear damped sine wave, and the ACF cut off after 1 on the seasonal lag, so we can propose .

Both models passed the residual Diagnostics, we will look at model as an example.

图表

描述已自动生成

From the plot above, we see that standardized residuals do not show any trend or seasonality. Moreover, we have 1 out of 35 lags of ACF lies outside of 95% confidence interval, which is fine since we can expect , or 1-2 lags, to be outside of the CI. Moving to the Q-Q plot, we find that both tails do not lie in the shaded area. Finally, the p-values for Ljung-Box statistic are all above 0.05. In conclusion, the assumption of constant mean, constant variance and uncorrelatedness holds for , while normality of the data does not hold true. However, normality only affects prediction power of the model, so the fitting power of model is overall satisfied.

Now we assume first order differencing is not good, so we need to try second order differencing.

日程表

描述已自动生成

From the ACF plot, we do not see slow decay or seasonal behavior. Hence, result from 2nd order differencing is stationary.

图表

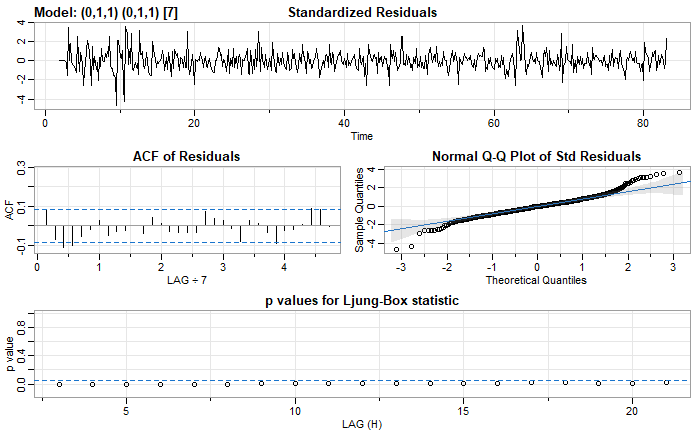
中度可信度描述已自动生成

From the plot above:

* We might regard both the ACF and PACF plot as damped sine wave ignoring the data at the period. Also, ACF is 0 after lag 1 while PACF is exponential decay only considering data at the period. Hence, we have .
* We might regard the ACF as 0 after lag 1, the PACF is damped sine wave ignoring the data at the period. Also, ACF is 0 after lag 1 while PACF is exponential decay only considering data at the period. Hence, we have .
* We might regard the ACF as 0 after lag 6, the PACF as damped sine wave ignoring the data at the period. Also, ACF is 0 after lag 1 while PACF is exponential decay only considering data at the period. Hence, we have .

Both and passed the Residuals Diagnostics. (Plots can be found in appendix D),

For , the p-values for Ljung-Box statistic are all under 0.05, which is the sign of violation of uncorrelation at every lag.



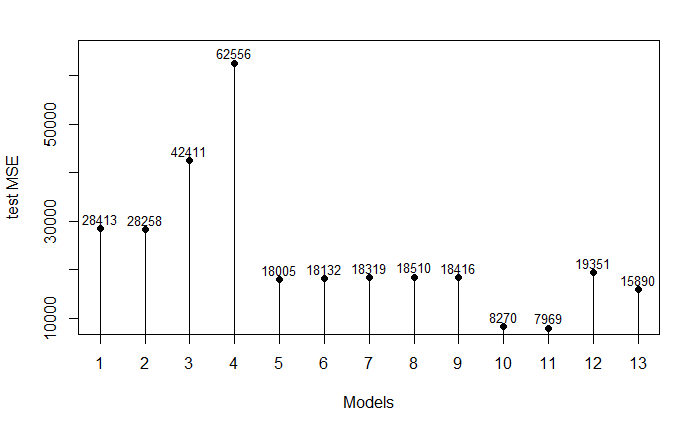
Overall, for the SARIMA modelling, we have 4 models:

**Statistical Conclusions**

We have the following 13 models to chose.

1. Second order regression (include seasonality) followed by
2. Second order regression (include seasonality) followed by .
3. Second order regression (include seasonality) followed by .
4. Second order regression (include seasonality) followed by .
5. Additive HW followed by .
6. Additive HW followed by .
7. Additive HW followed by .
8. Additive HW followed by .
9. Additive HW followed by .
10. .
11. .
12. .
13. .

We chose the best model based on test error. The plot of all 13 models is drawn below.



Even through model 11 has a smaller test error, it has 4 more parameter compared to model 10, so we picked model 10, , which has a slight larger test error but less parameter. The plot of the fit and prediction from week 30 to week 97 after transforming back the data can be seen below from (for the whole plot, see appendix E). Note that the prediction interval can be seen in the plot just for demonstration purpose, it doesn’t real make sense to talk about prediction interval when the normality assumption doesn’t hold.

图表, 折线图

描述已自动生成

**Conclusions in the context of the problem**

We found that the traffic condition has an increasing trend and a seasonality with period 7. The prediction for the next week can be seen in the table below. Note again that the prediction interval can be seen in the table just for demonstration purpose, it doesn’t real make sense to talk about prediction interval when the normality assumption doesn’t hold.

表格

描述已自动生成

We can see that the number of cars on junction 1 will be approximately 1348 on next Saturday, and 1276 on next Sunday, and so on.

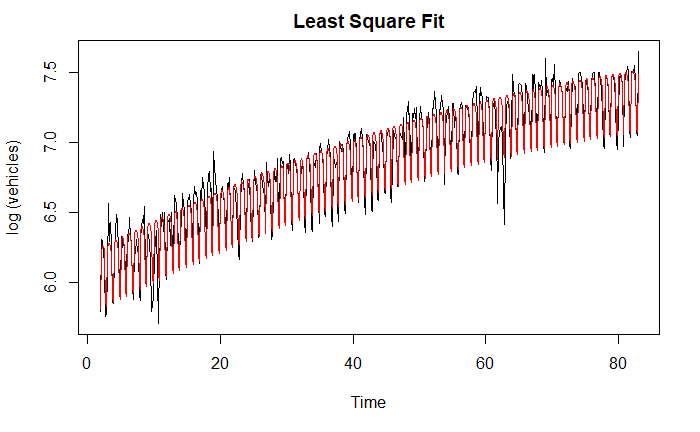
**Appendix A**

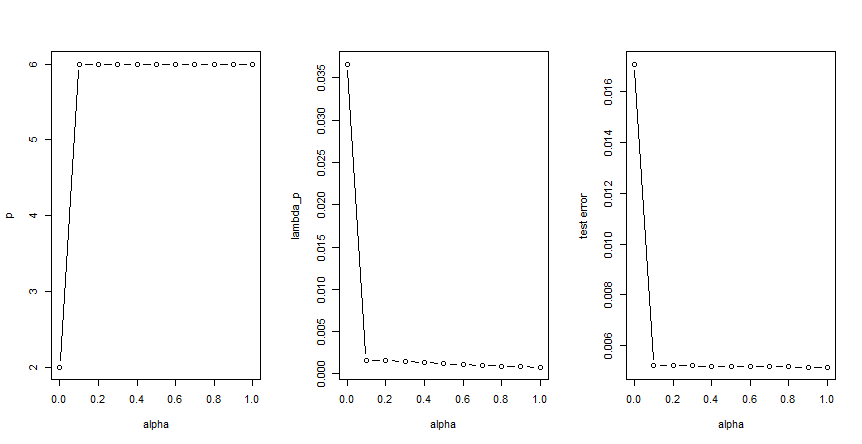
图表, 折线图

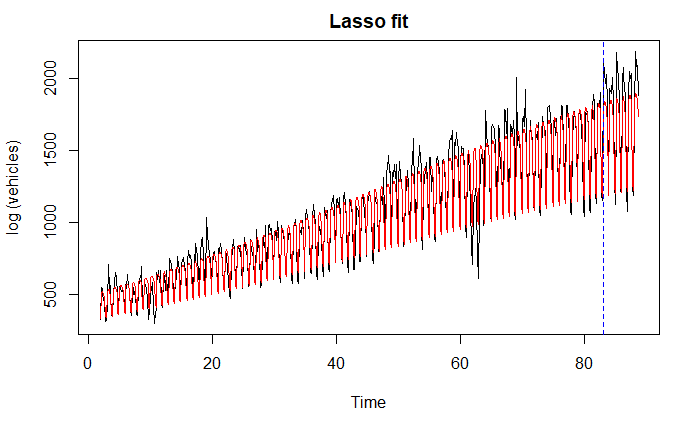
描述已自动生成图表, 直方图

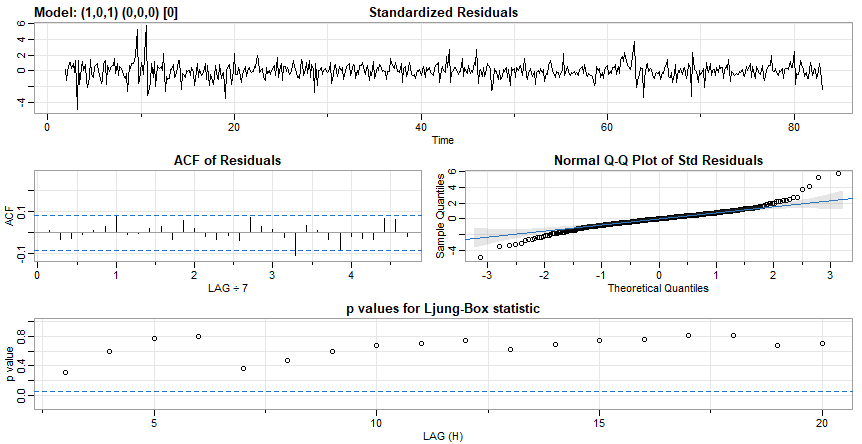
描述已自动生成

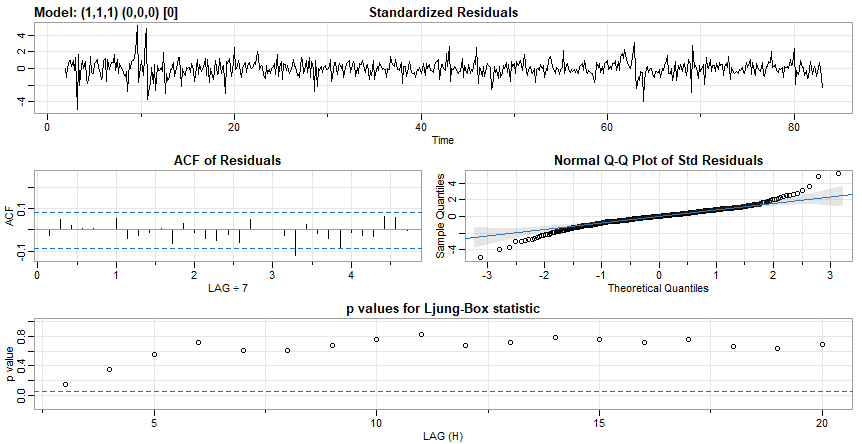
**Appendix B – Regression**

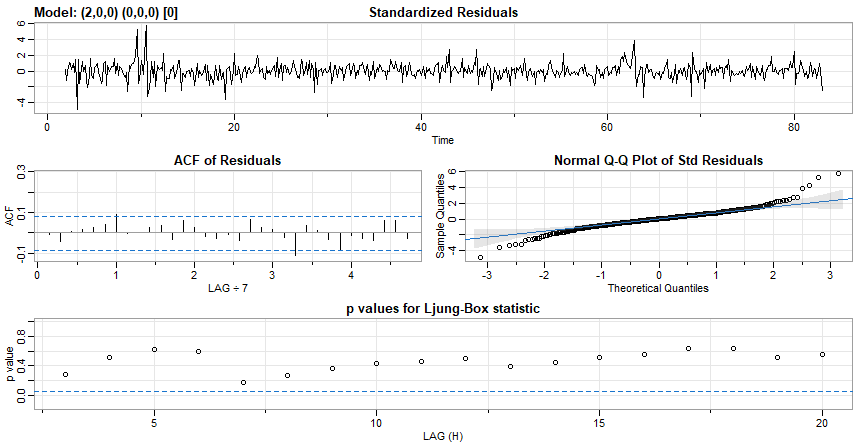
****

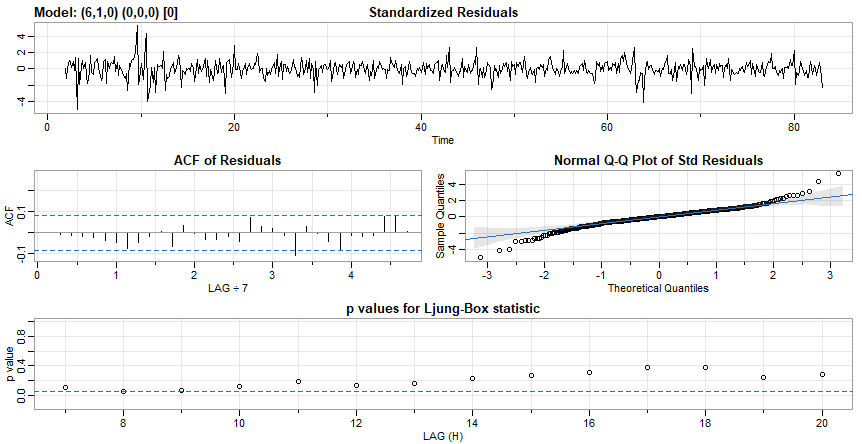




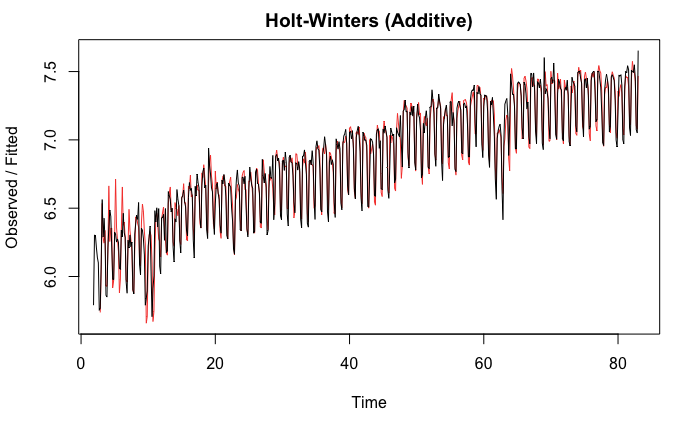


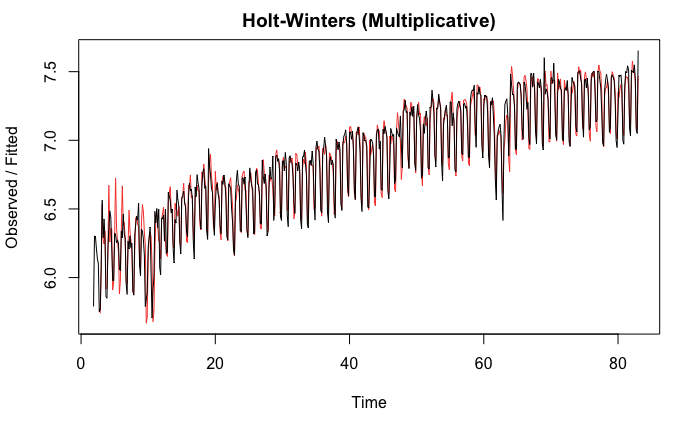


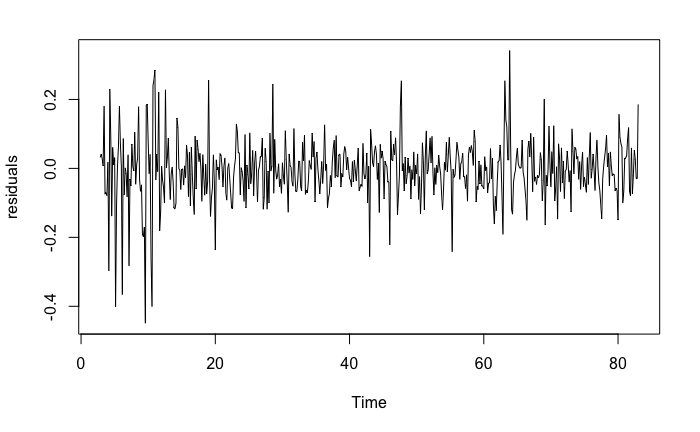


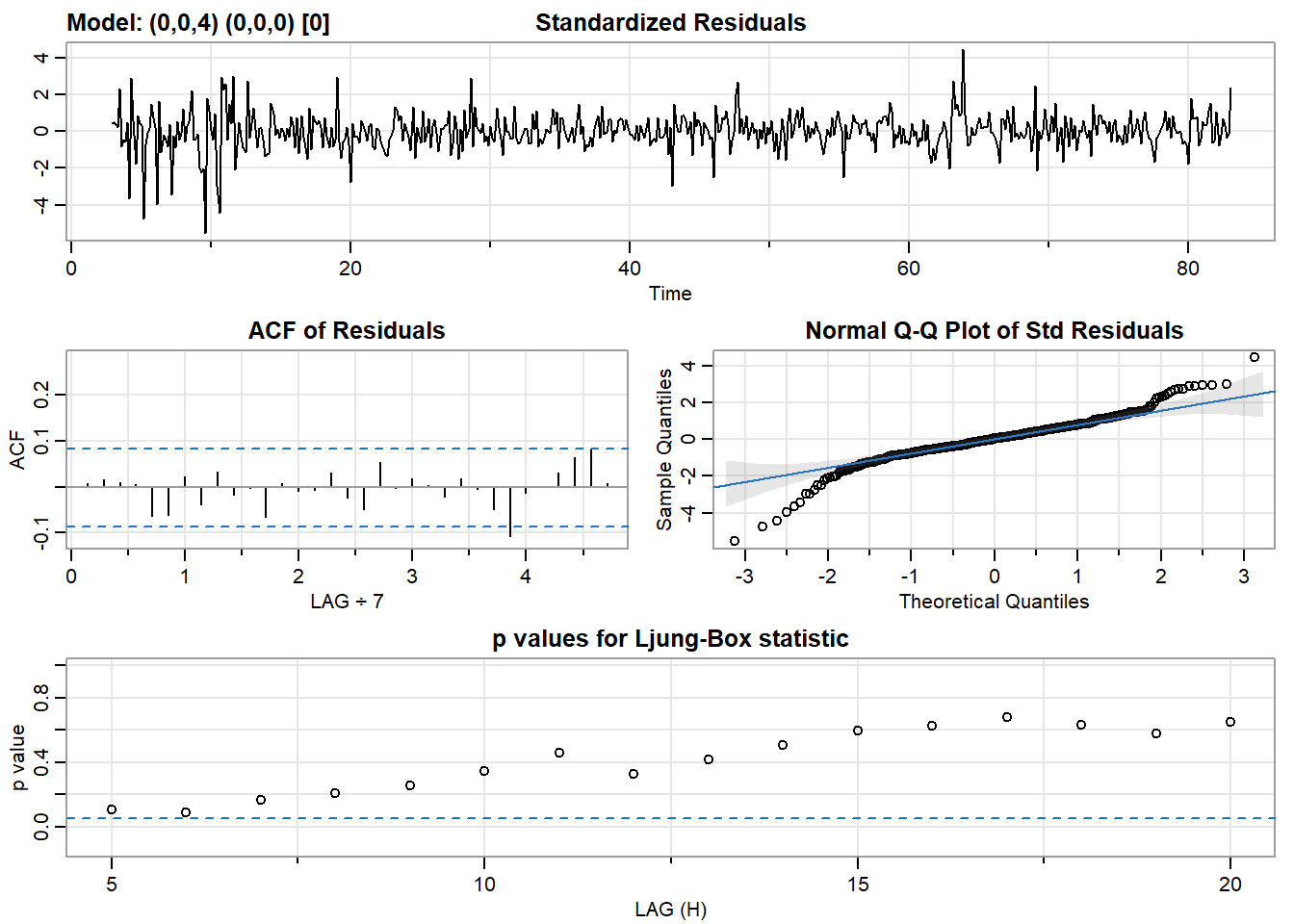
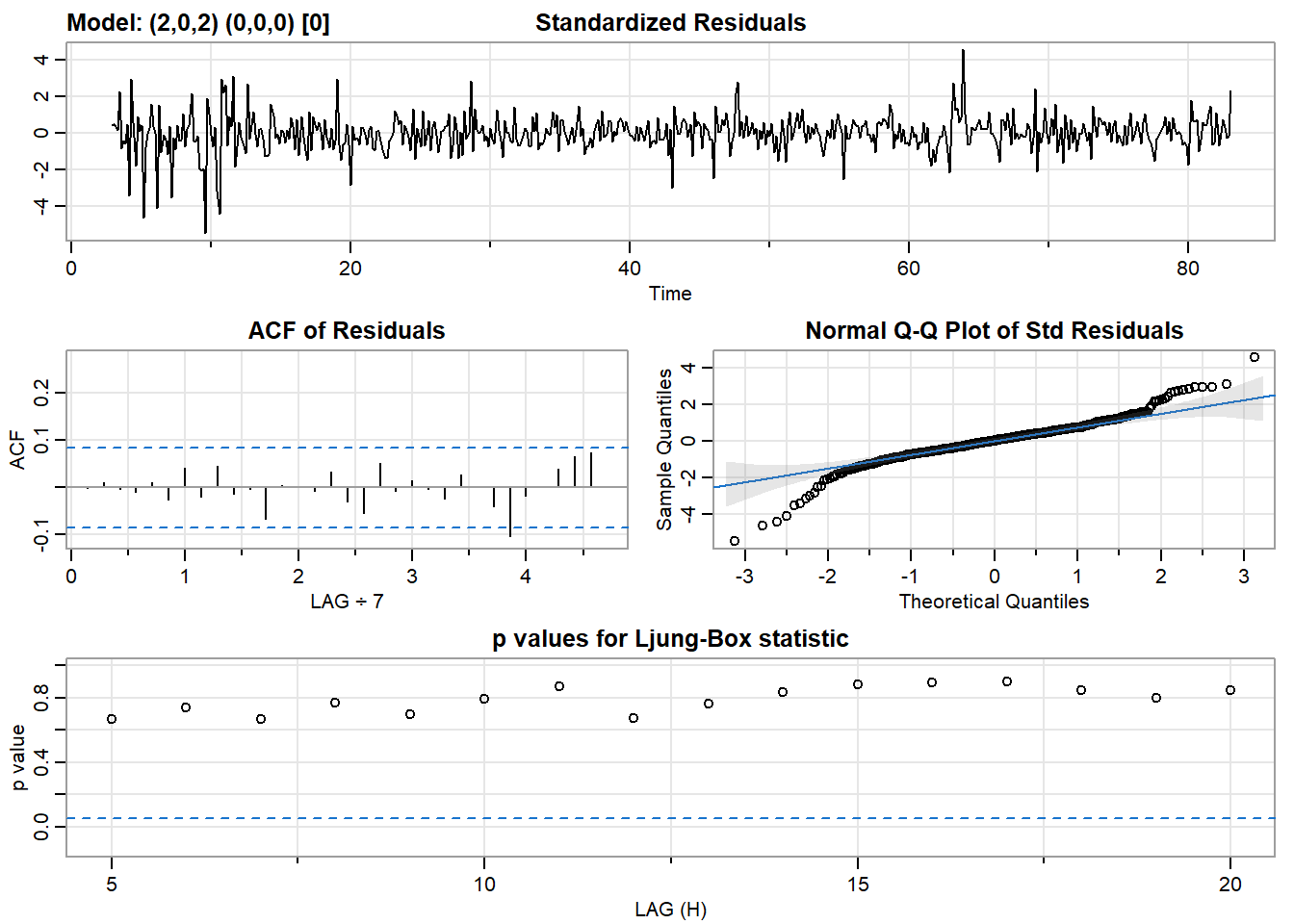
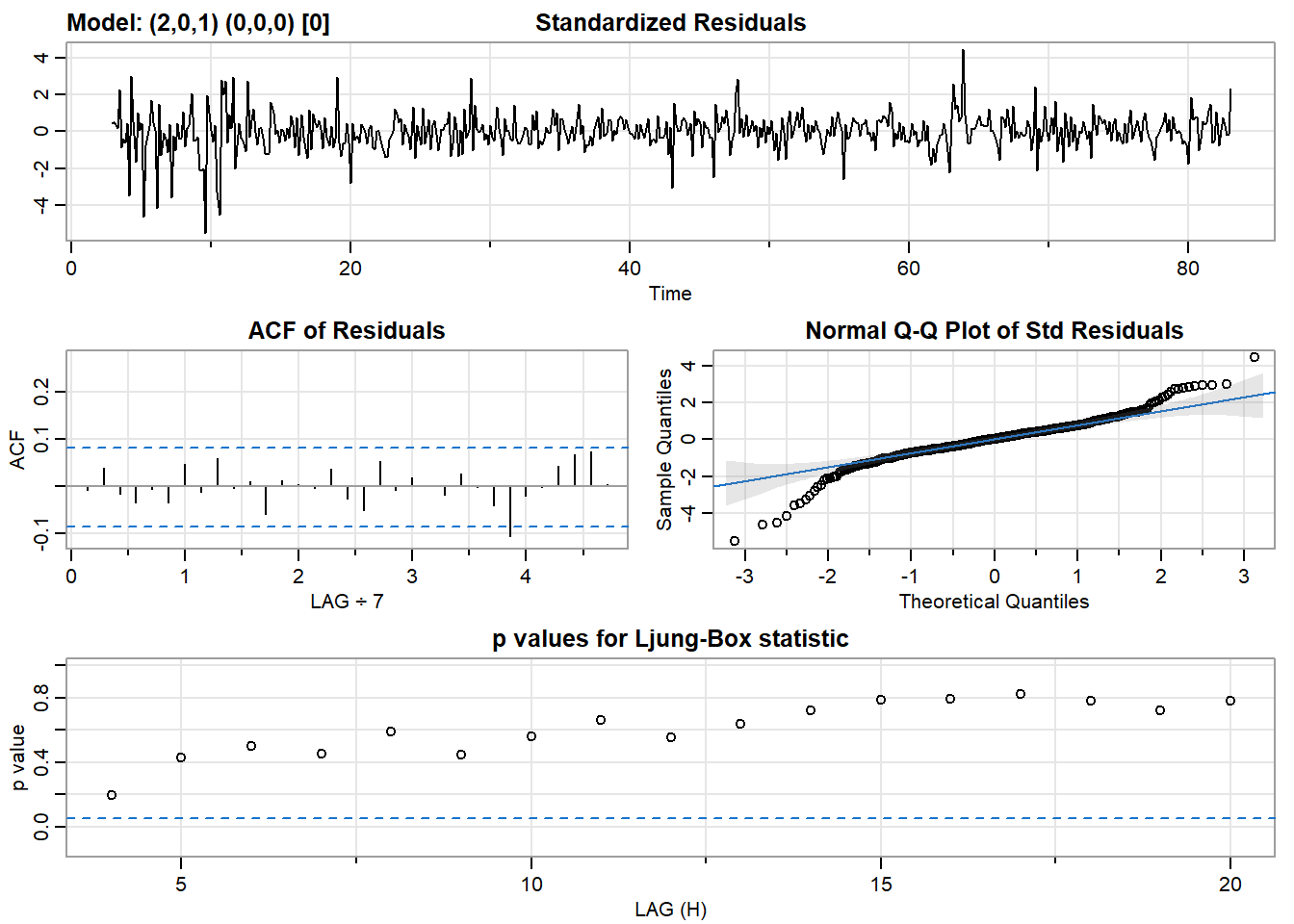
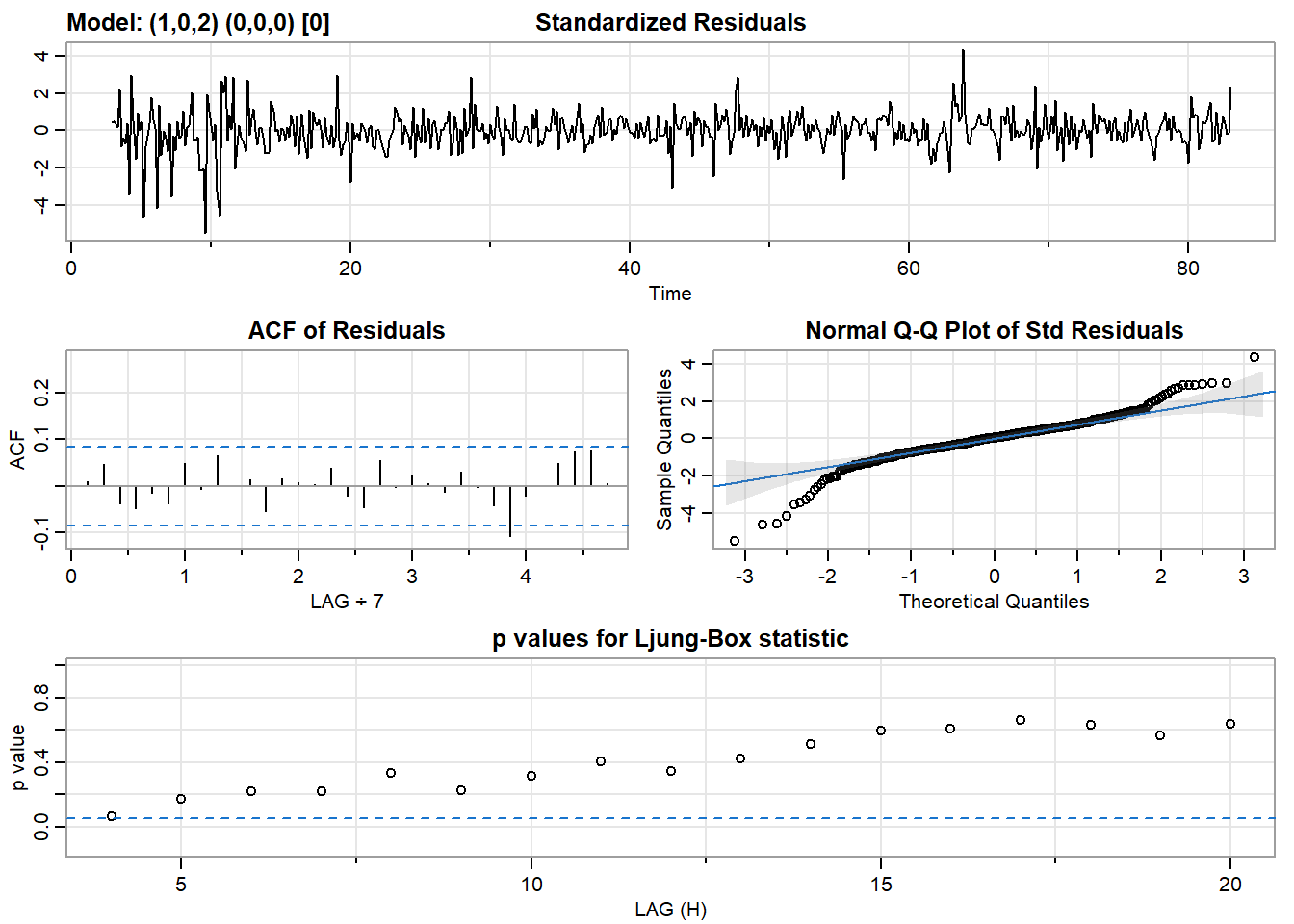
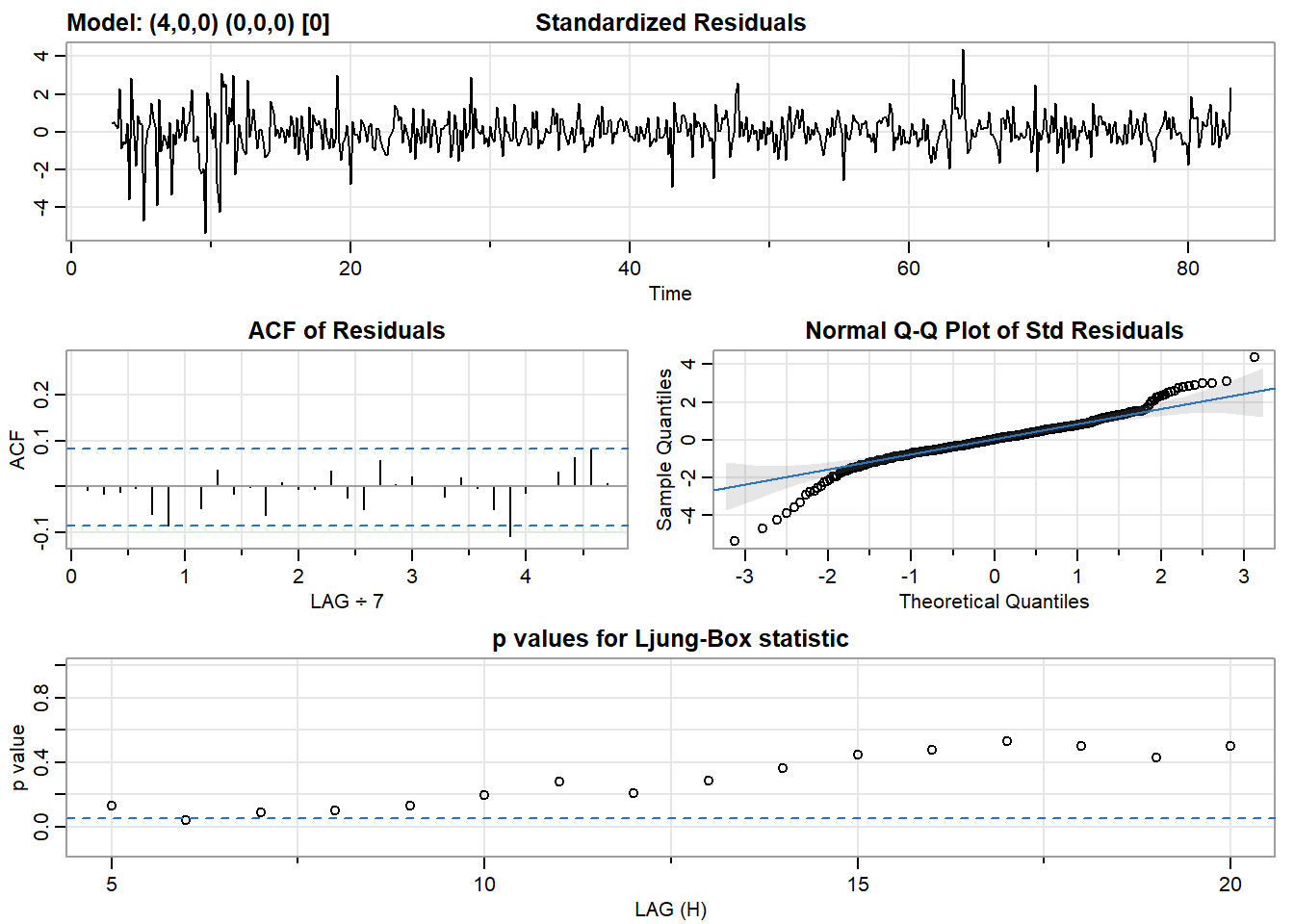


**Appendix C - Holt Winters**

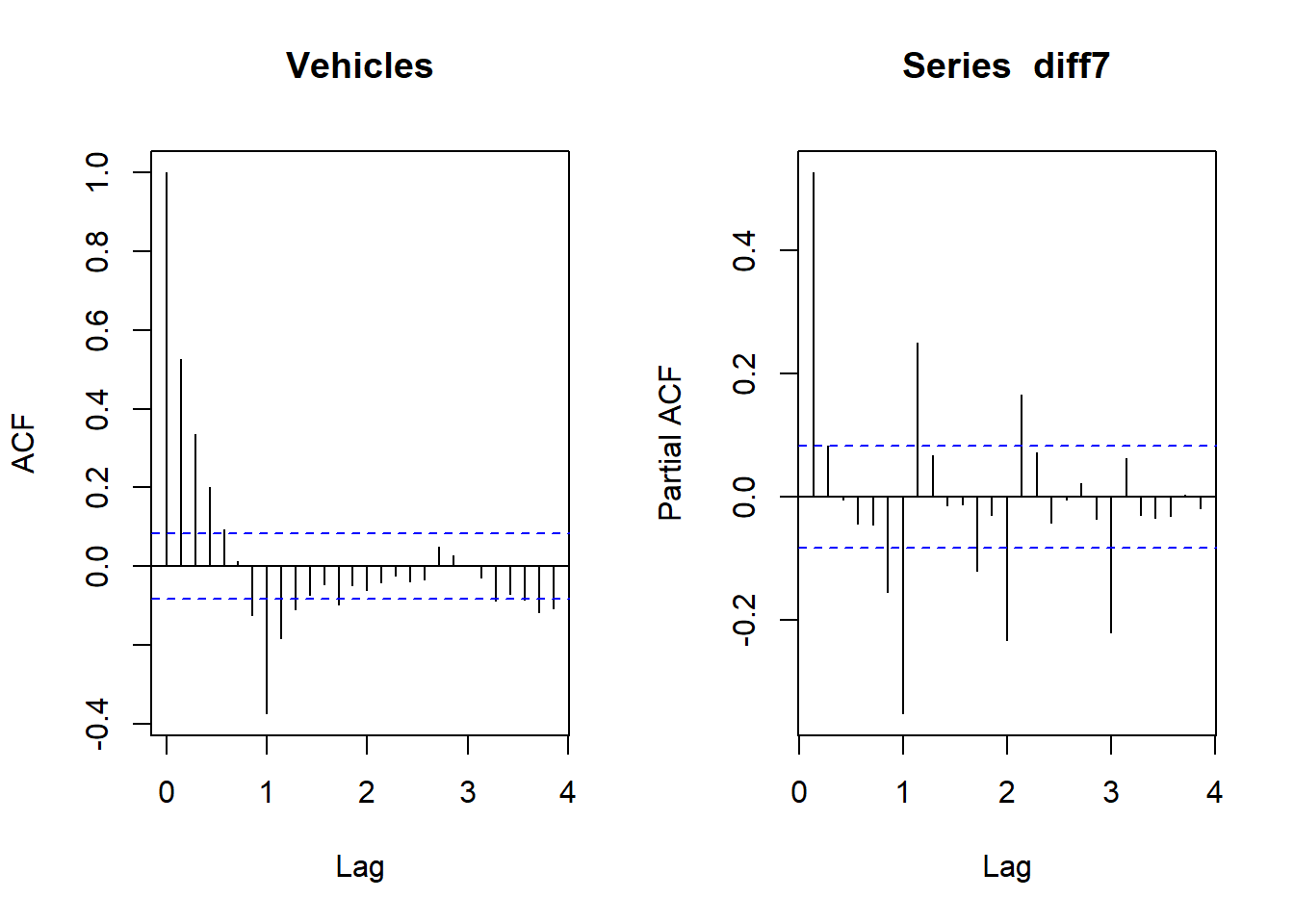
****

****

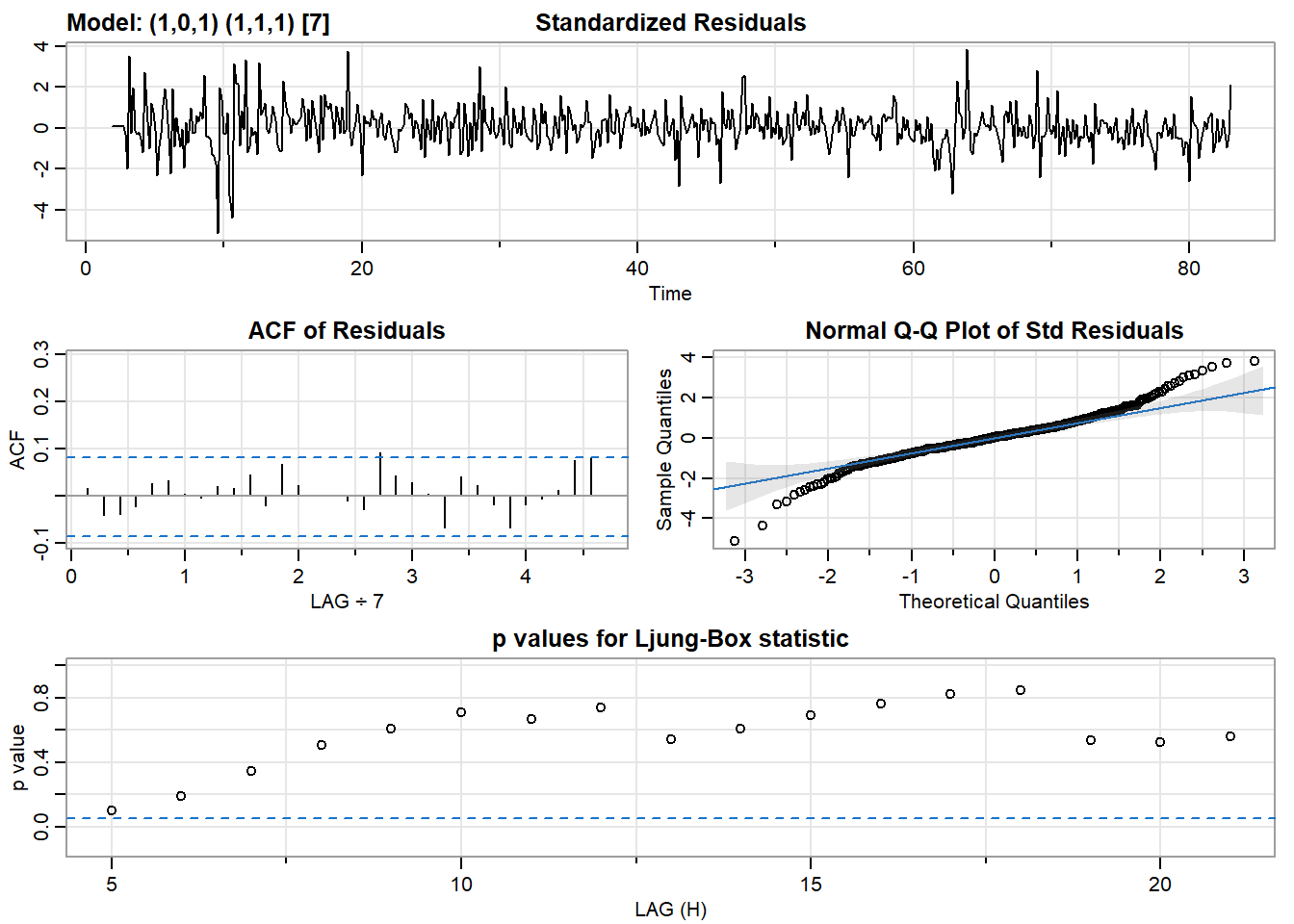
****



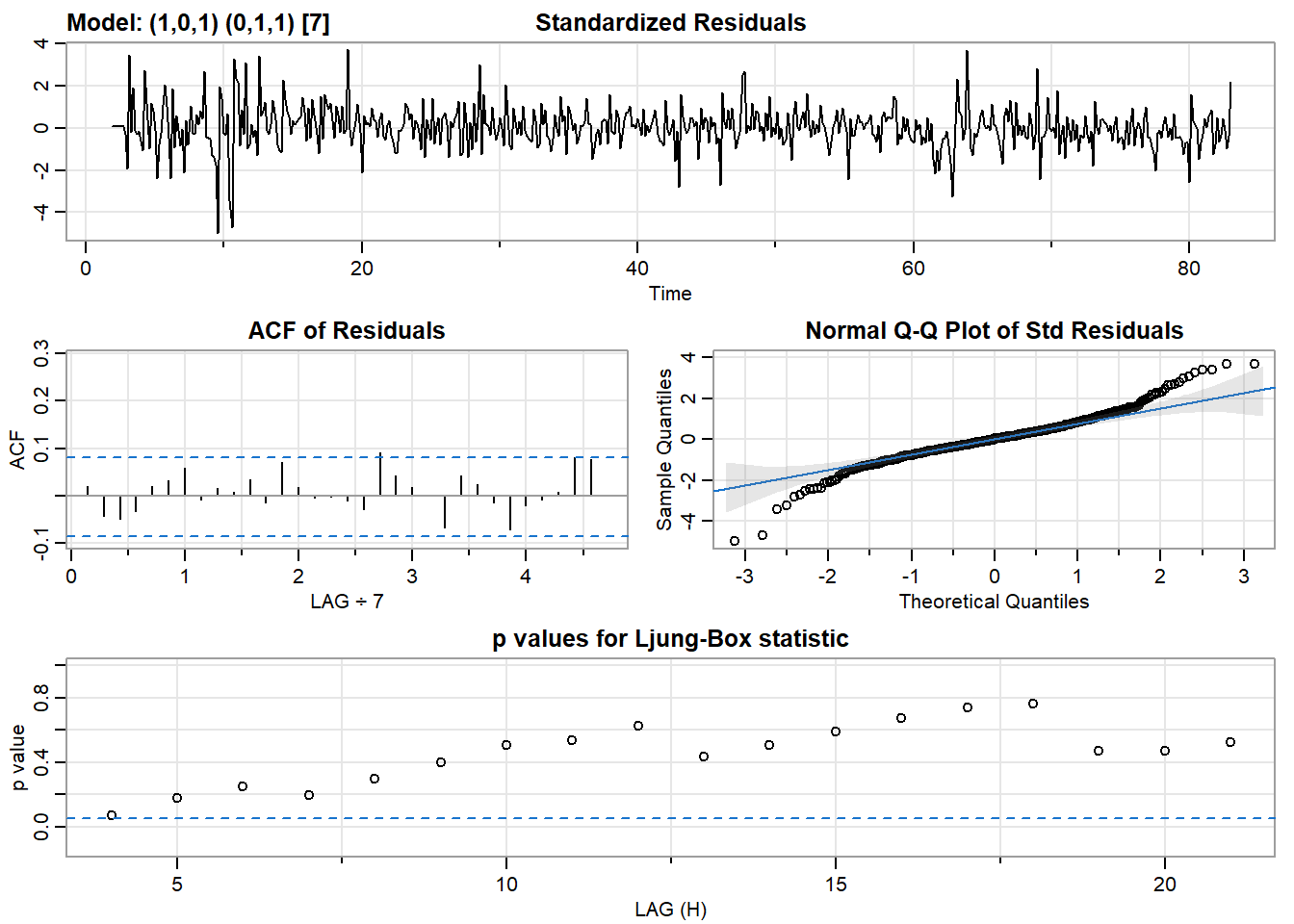
**Appendix D - Differencing**



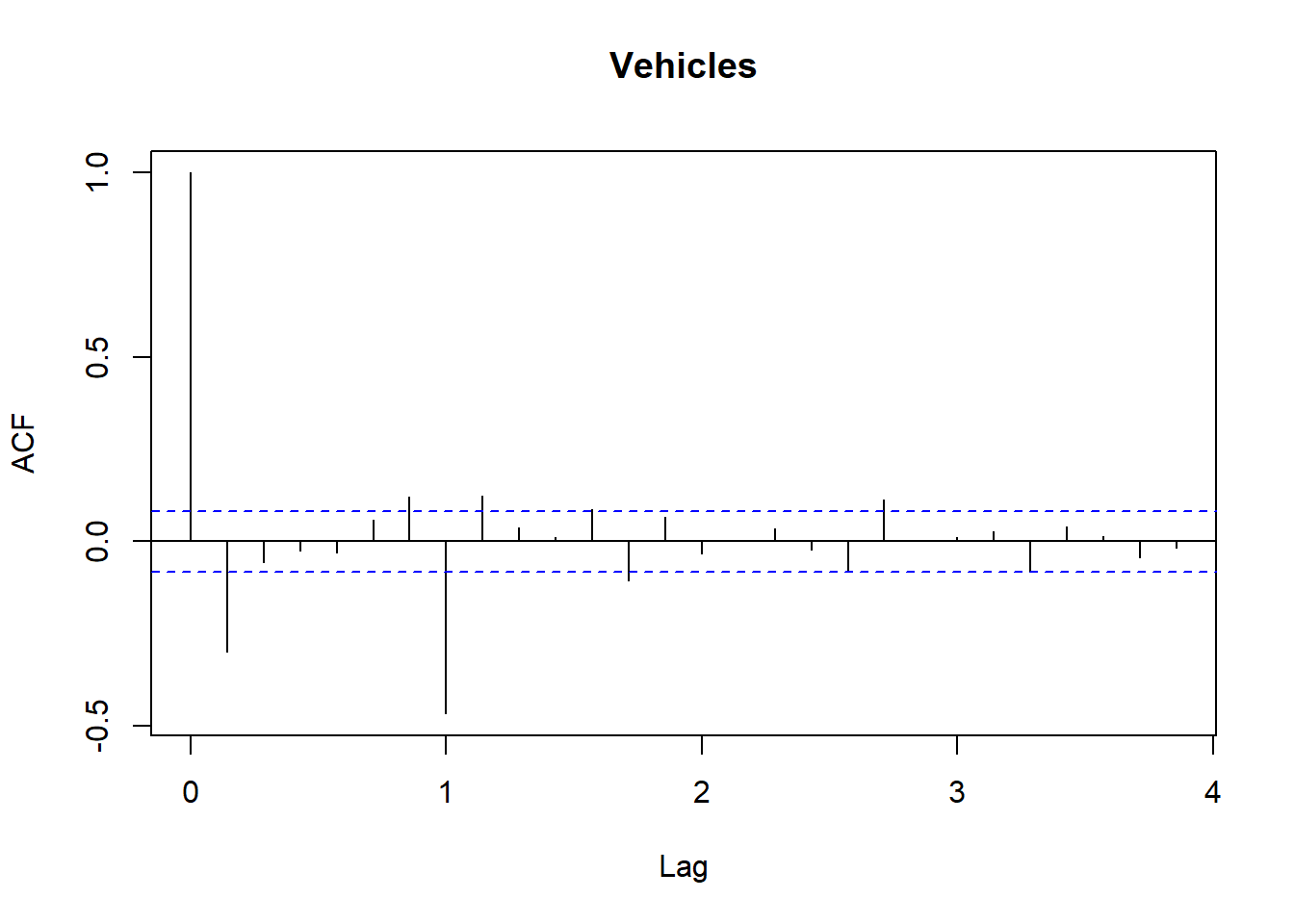
ACF&PACF of stationary



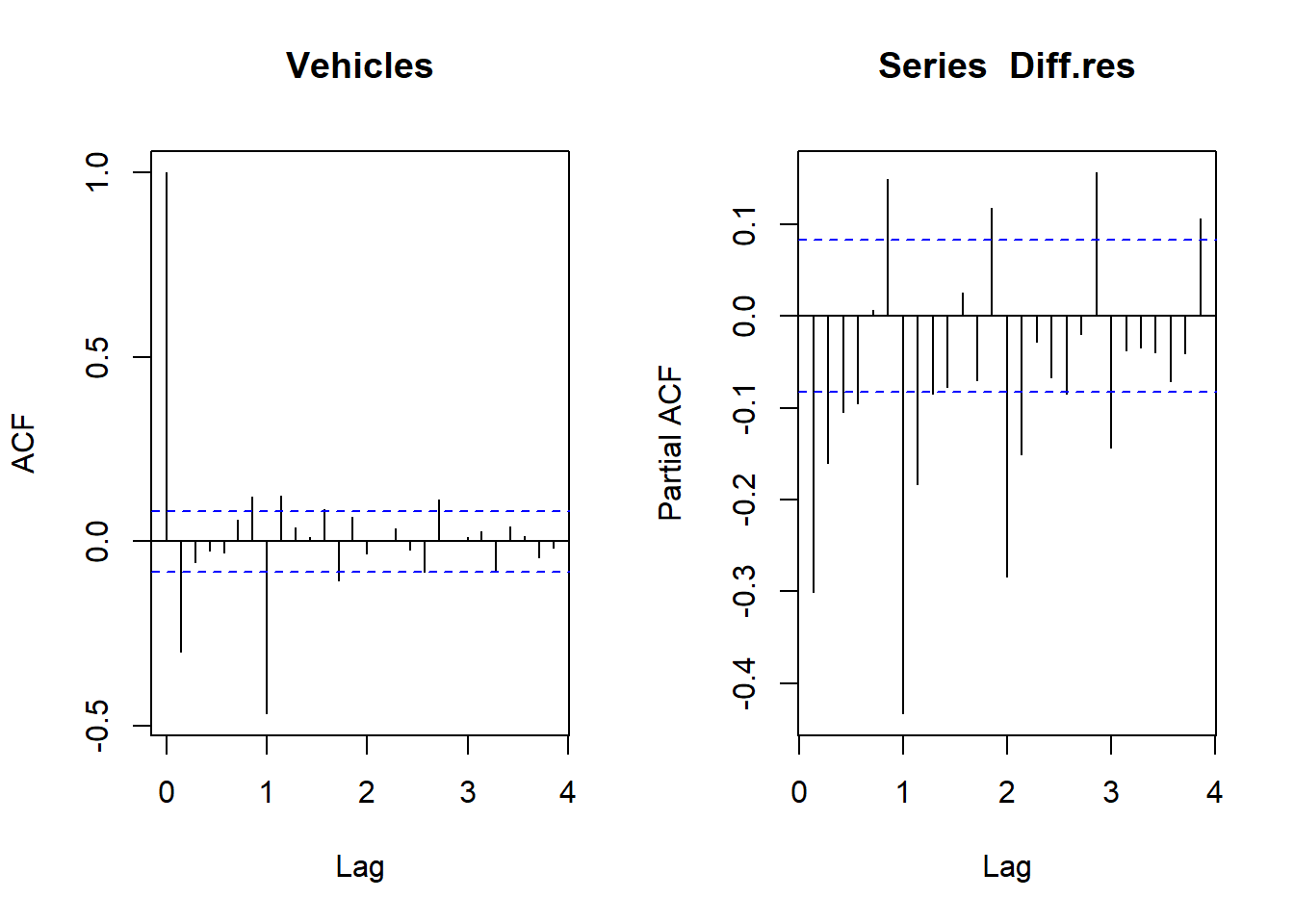
SARIMA(1,0,1)(1,1,1)



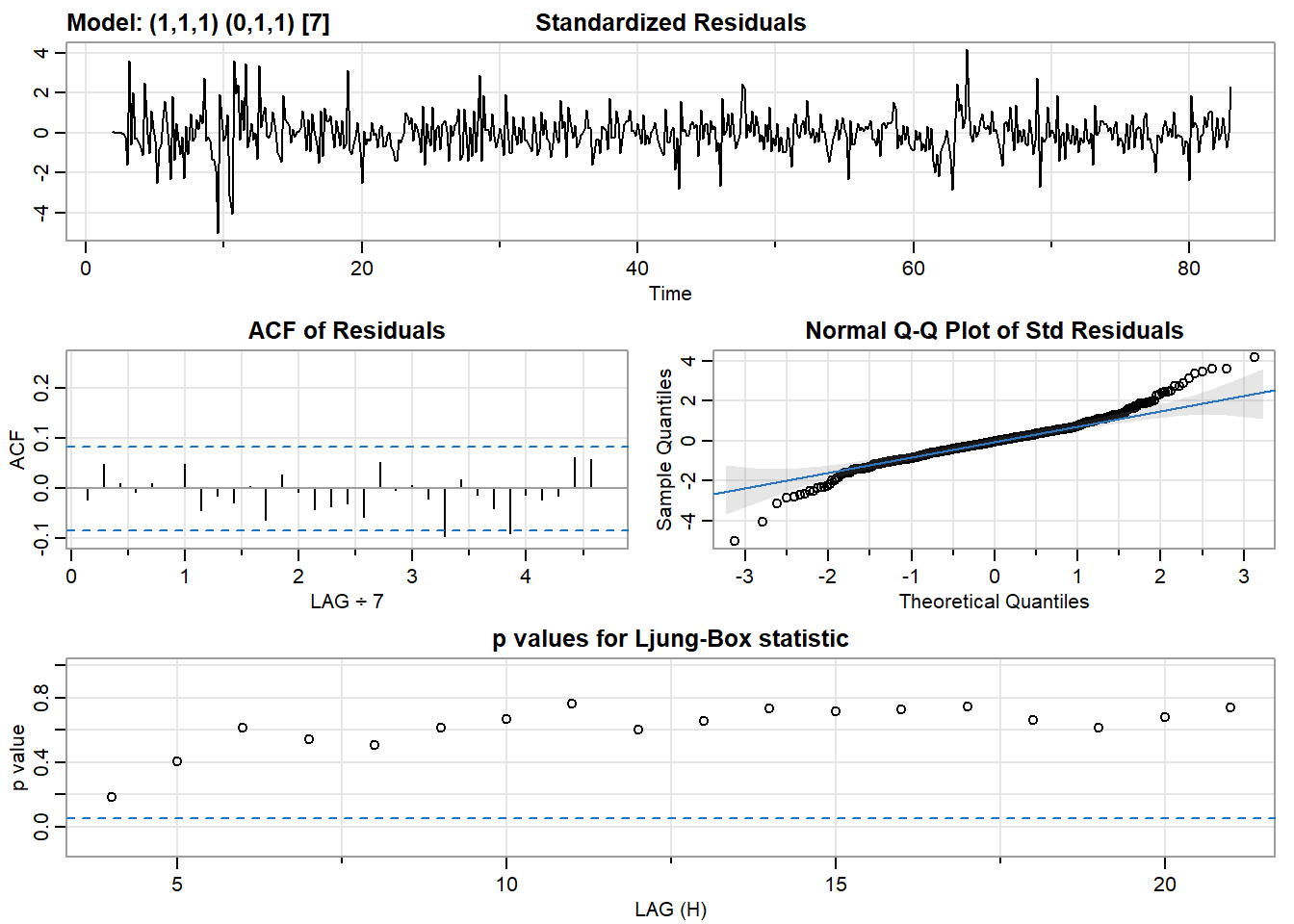
SARIMA(1,0,1)(0,1,1)



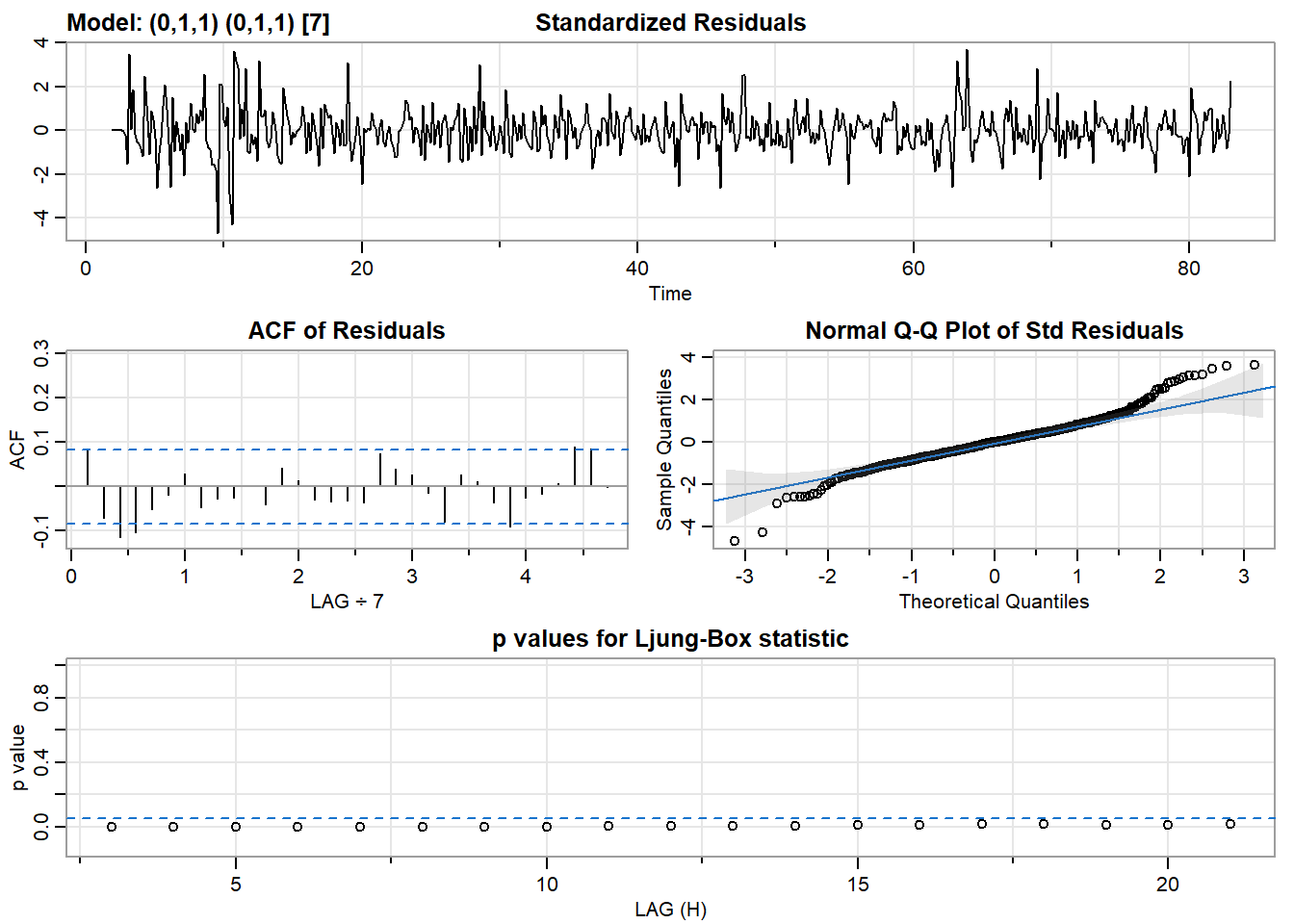
ACF of 2nd diff



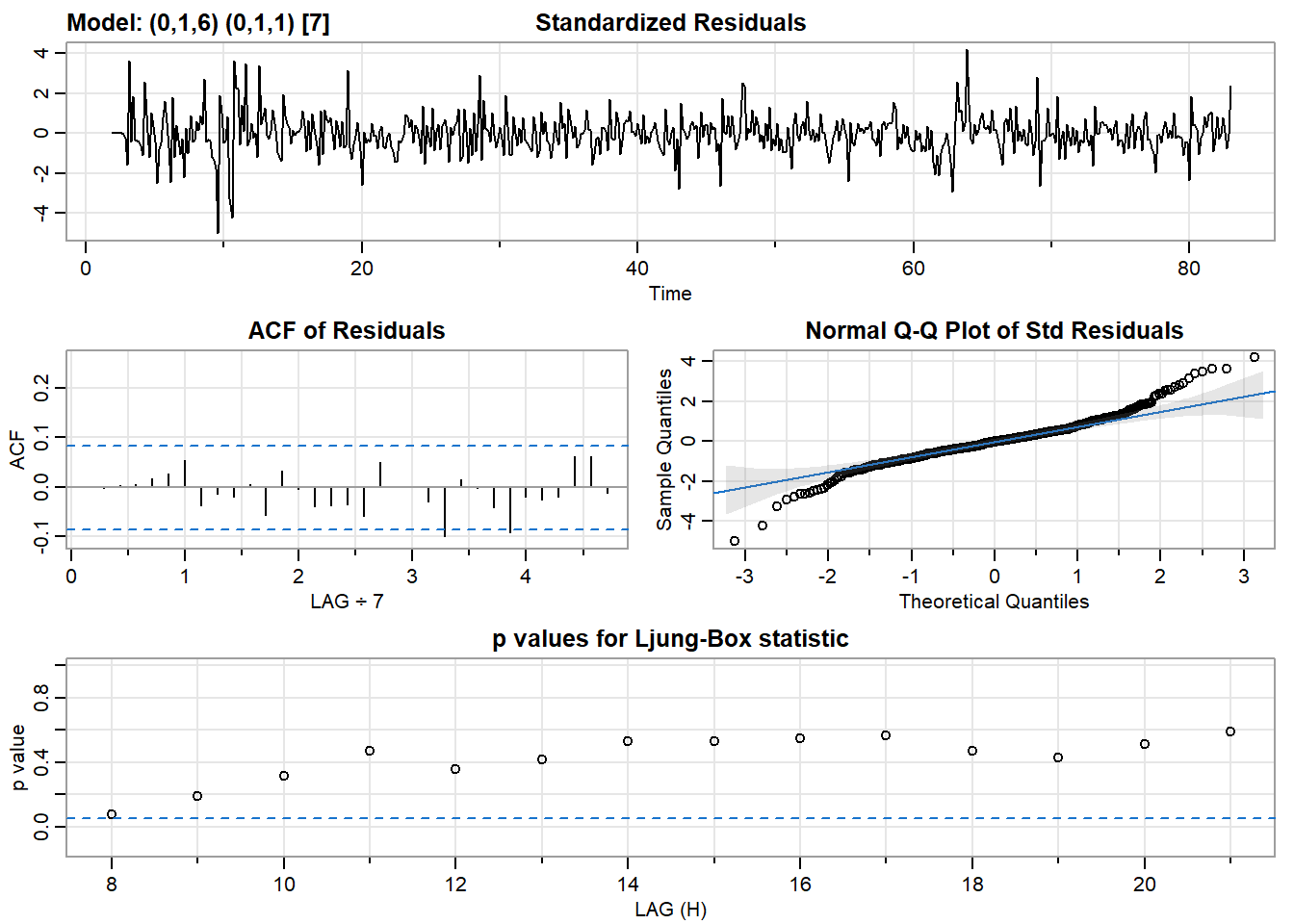
ACF&PACF of 2nd diff



SARIMA(1,1,1)(0,1,1)7



SARIMA(0,1,1)(0,1,1)7



SARIMA(0,1,6)(0,1,1)7

**Appendix E - Conclusion**

