STAT 153 Project Checkpoint 4

Danny Wu

4/23/2021

1. Executive Summary

COVID-19 has had a drastic impact to the daily lives of everyone around the globe. Having a prediction of the number of daily COVID-19 cases would be informative to both the health care sector as well as policy makers. Through our investigation, we fitted ...

2. Exploratory Data Analysis

Daily COVID-19 cases have dramatically increased since March of last year. As seen in the left panel of Fig. 1, there is a very strong trend in the data. Cases grew from March to August then slightly declined until November, where it rapidly grew and exceeded the levels before the decline. There is also some seasonality based on the day of the week. We can see it in the fluctuations in the left panel as well as the right panel of Fig. 1, where on average, cases are lower on Sunday and Monday. It is also clear that the data exhibits heretoscedasticity as the variance of daily COVID-19 cases has been increasing over time.

The are a few anomalies in the data where a day has zero counts and the following day has a spike in cases (marked by the red and orange points respectively in the left panel of Fig. 1). These anomalies are likely due to cases being miscounted and accidentally moved to the next day. There are four of these instances. To correct for them, we divide the number of counts on the second date in half and set that as the number of counts for both dates.

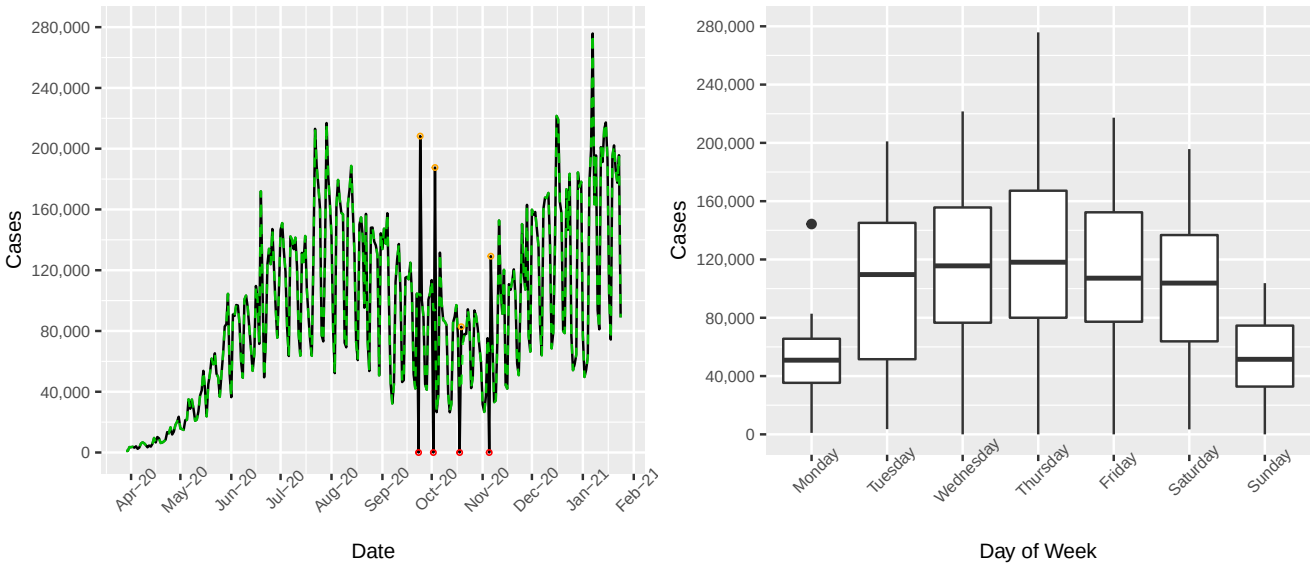

Figure 1: Left panel: Daily COVID-19 cases from March 29, 2020 to January 24, 2021. Dashed green line represents corrected dataset with red and orange points indicating erroneous data points. Right panel: Box-plot of the distribution of daily cases grouped by the day of the week.

3. Models Considered

3.1 Parametric Signal Model

First we consider a parametric signal model. Using a periodogram we see that there are two dominant fourier frequencies at 2/302 and 43/302 corresponding to a period of 151 and 7.02 days. Thus we create a sinusoid with frequency 1/151 to model the larger trend, and then use indicators for the day of week to model the weekly seasonality. We include an interaction term between the sinusoid and the indicators to caputure the fluctuation of the sinusoud's magnitude. Lastly, we also interact time, day of week indicator, and the larger sinusoid to capture any effects that might be from all three at the same time. There exists great heteroskedasticity, so we also log our data in order to help with the fitting process. Our parametric model is as follows:

$$log(\text{Cases}_t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \sum_{i=0}^{5} \left[\beta_{3+6i} I_{\text{weekday}_{it}} + \beta_{4+6i} t I_{\text{weekday}_{it}} + \beta_{5+6i} I_{\text{weekday}_{it}} \cos\left(\frac{2\pi t}{151}\right) \right. \tag{1}$$

$$\left. + \beta_{6+6i} I_{\text{weekday}_{it}} \sin\left(\frac{2\pi t}{151}\right) + \beta_{7+6i} t I_{\text{weekday}_{it}} \cos\left(\frac{2\pi t}{151}\right) + \beta_{8+6i} t I_{\text{weekday}_{it}} \sin\left(\frac{2\pi t}{151}\right) \right] \tag{2}$$

$$+ \sum_{j=0}^{2} \left[\beta_{39+2j} \cos\left(\frac{2\pi t}{151}\right) + \beta_{40+2j} \sin\left(\frac{2\pi t}{151}\right) \right] \tag{3}$$

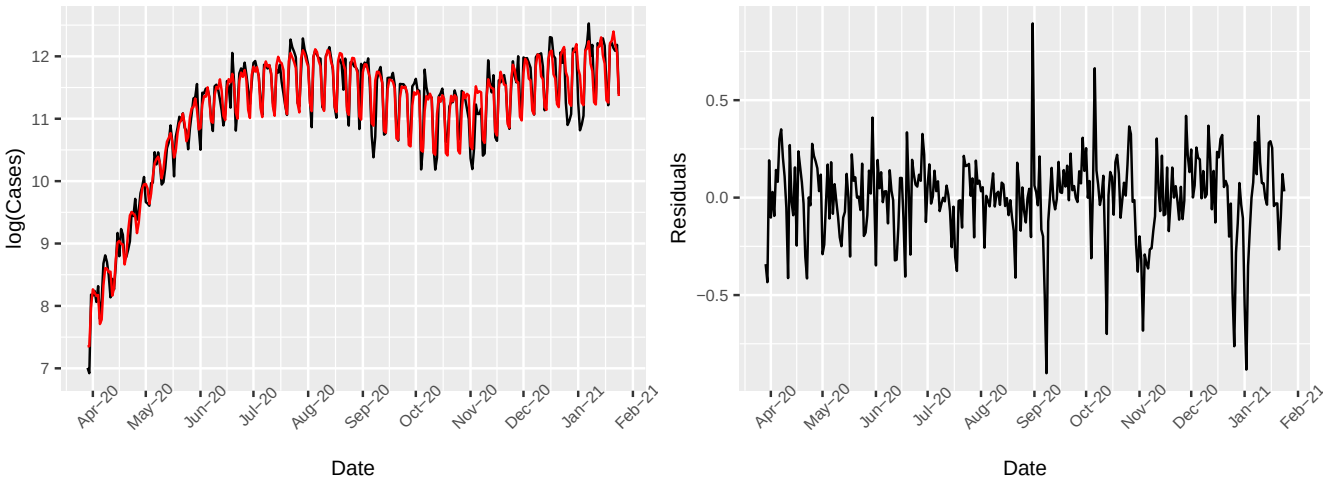

Figure 2: Left panel: Our parametric signal model with the fitted values plotted in red and the logged data plotted in black. Right panel: Residuals of the aforementioend parametric model

3.1.1 Parametric Signal Model with ARMA(0,3)x(2,1)[7]

Looking at the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) of our residuals in Fig. 3, we observe large magnitudes at lags 1, 2, 3, 14 and 28 on the ACF plot and at lags 1, 14, and 28 in our PACF plot. The significant values in lags 1, 2, and 3 in the ACF plot the significant value at lag 1 in the PACF suggest an ARMA(1,3) fit. In addition, the 2 large magnitudes at lags 14 and 28 in both ACF and PACF plots suggest SARMA(2,2). Though the seasonal autocorrelation appears to begin at lag 14, through trial and error, we found that a period of 7 works much better, and through additional tweaking, we arrive at an ARMA(0,3)x(2,1)[7]. Looking at the SARIMA diagnostic plots in Fig. 4, we see that not only does the values of our ACF of Residuals lie within the confidence interval for white noise, but our p values for the Ljung-Box statistics are all very high, indicating that this model fits decently well.

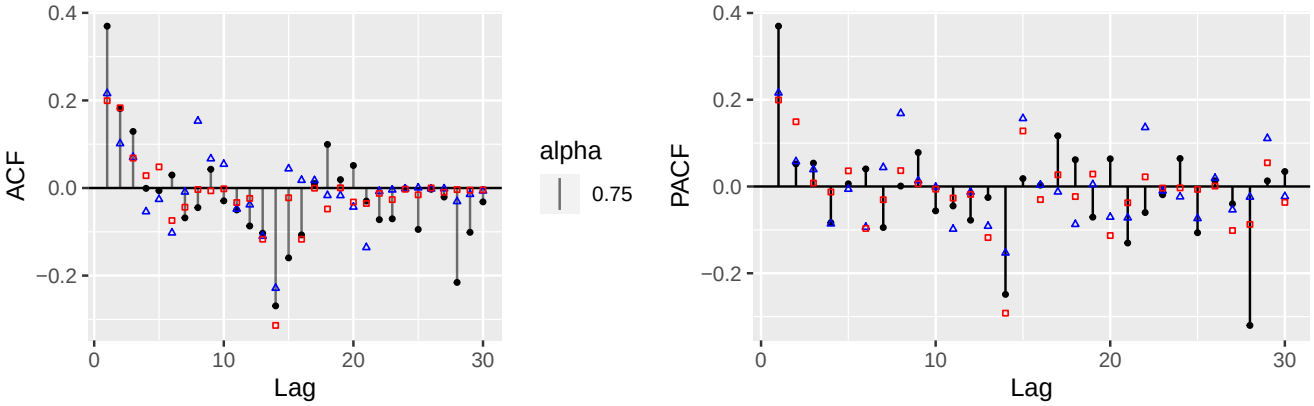

Figure 3: Sample ACF and PACF values for our logged parametric model (in black). Theoretical distributions for ARMA(0,3)x(2,1)[7] in blue triangles, and ARMA(2,1)x(1,2)[7] in red squares.

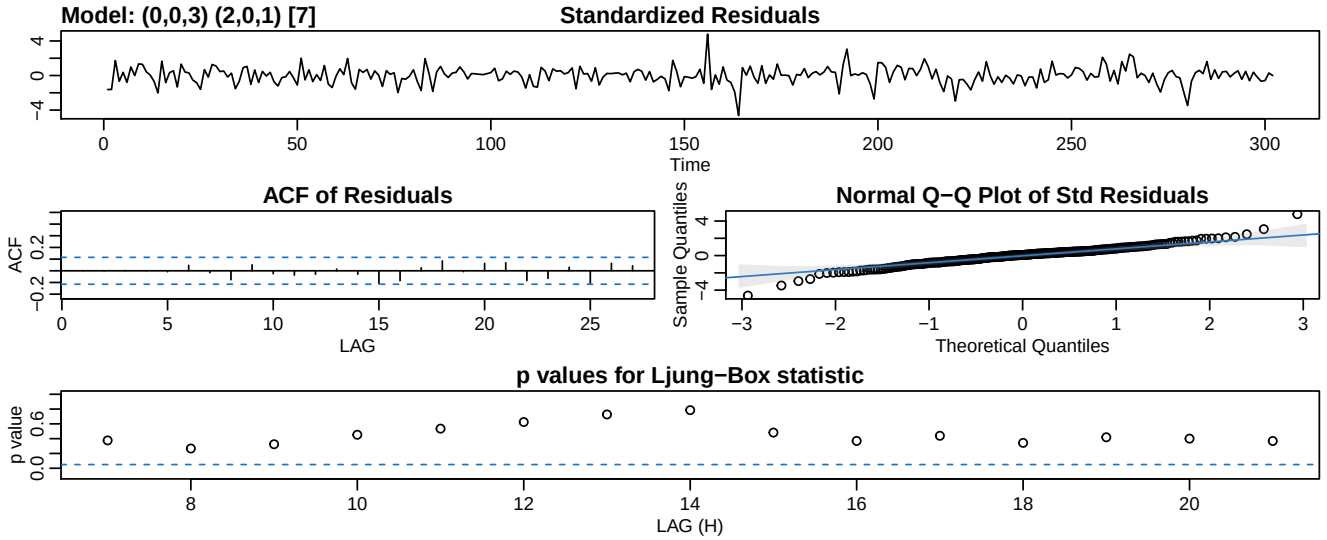

Figure 4: SARIMA diagnostics plots for Parametric Signal Model with ARMA(0,3)x(2,1)[7]

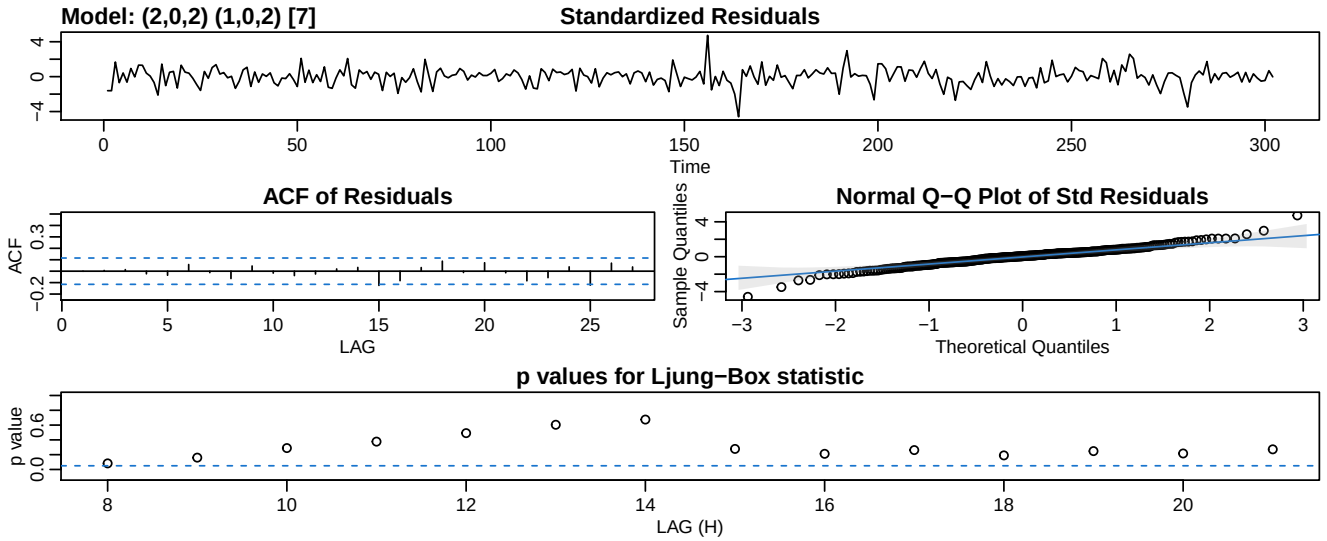

Figure 5: SARIMA diagnostics plots for Parametric Signal Model with ARMA(2,2)x(1,2)[7]

3.1.2 Parametric Signal Model with ARMA(2,2)x(1,2)[7]

The R function auto.arima() suggests an ARMA(2,2)x(1,2)[7] model. This model is plausible, as the Ljung-Box statistics are all almost all of large magnitude as seen in Fig. 5. In addition, in Fig. 3, the theoretical ACF and PACF points from the ARMA(2,2)x(1,2)[7] model appear to fit the sample ACF and PACF slighly better than the prior model.

3.2 Differencing model

We now try a differencing approach. Because there exists heteroskedasticity, we apply a VST by logging all of the values. Since there is weekly seasonality, we applying differencing with a lag of 7. Looking at the residuals there is still a slight downward trend so we apply differencing once again to get rid of that trend. Our resulting differenced data is shown in Fig. 6.

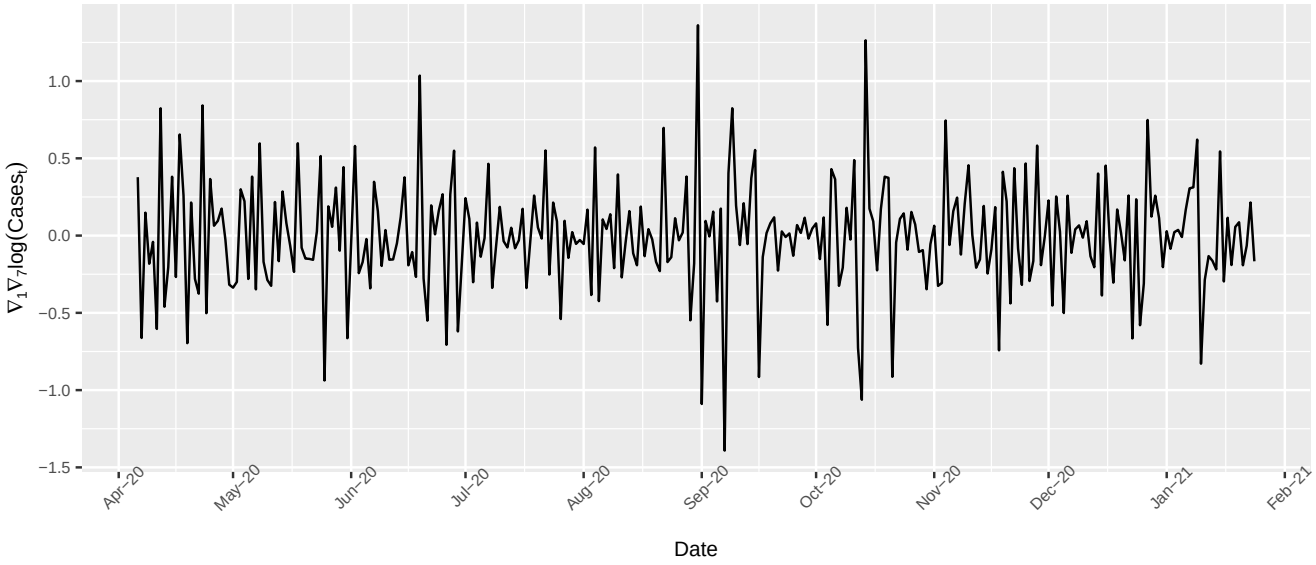

Figure 6: Plot of VST transformed data with lag 7 and lag 1 differencing applied. The result looks relatively stationary.

3.2.1 Differencing Model With ARMA(2,1)x(2,1)[7]

Looking at the ACF and PACF values for our differenced data in Fig. 7, we see that there are large magnitudes at lags 1 and 7 for the ACF plot. This suggests q = 1 and Q = 1 with a seasonal period of 7. In the PACF plot we see large magnitudes at lags 1, 2, 7 and 14, which reaffirms the seaonal period of 7, and suggests p = 2 and P = 2. As seen through the SARIMA diagnostics plots for this model in Fig. 8, the Ljung-Box statistics are decent and the ACF of residuals also does not have any significant magnitudes at any lags, thus this model is a good for for the differenced data.

3.2.2 Differencing Model With ARMA(1,1)x(0,1)[7]

For the second model, rather than basing it off of any signs from the ACF or PACF, I tried to adjust the values from the prior ARMA(2,1)x(2,1)[7] model to make it simpler while still satisfying our SARIMA diagnostics tests. After some trial and error, I ended up with an ARMA(1,1)x(0,1)[7]. As seen in the SARIMA diagnostics in Fig. 9, this model's Ljung-Box statistics are all satisfactory, while also being relatively less complex than the ARMA(2,1)x(2,1)[7] model.

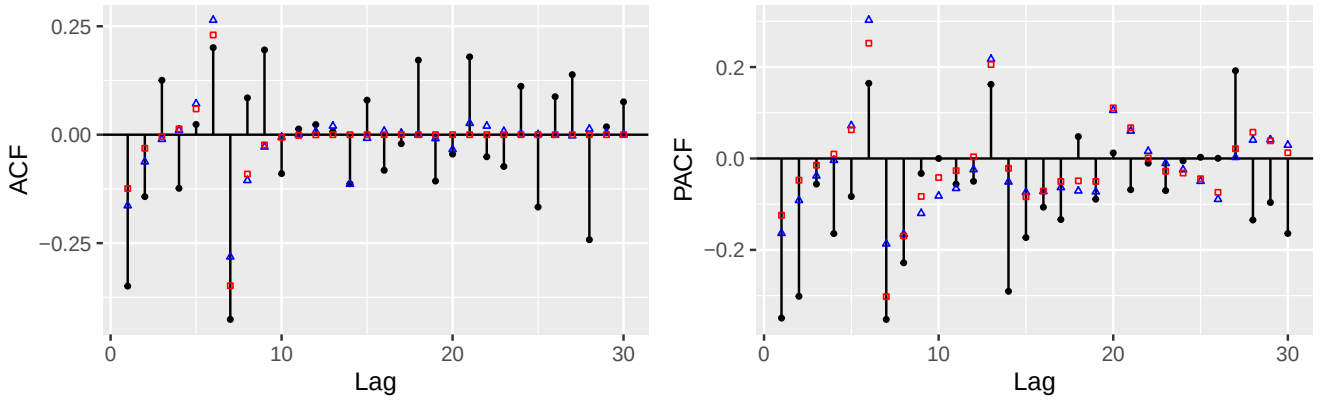

Figure 7: Sample ACF and PACF values from our differencing model. Theoretical distributions for ARMA(2,1)x(2,1)[7] in blue triangles, and ARMA(1,1)x(0,1)[7] in the red squares

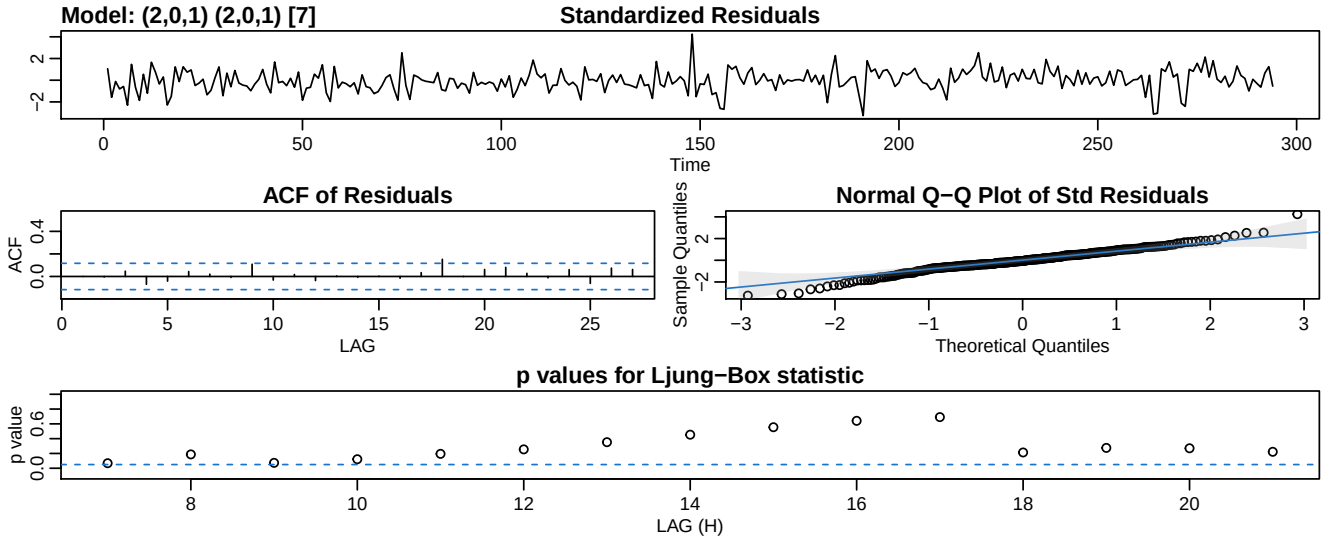

Figure 8: SARIMA diagnostics plots for Differencing Model with ARMA(2,1)x(2,1)[7]

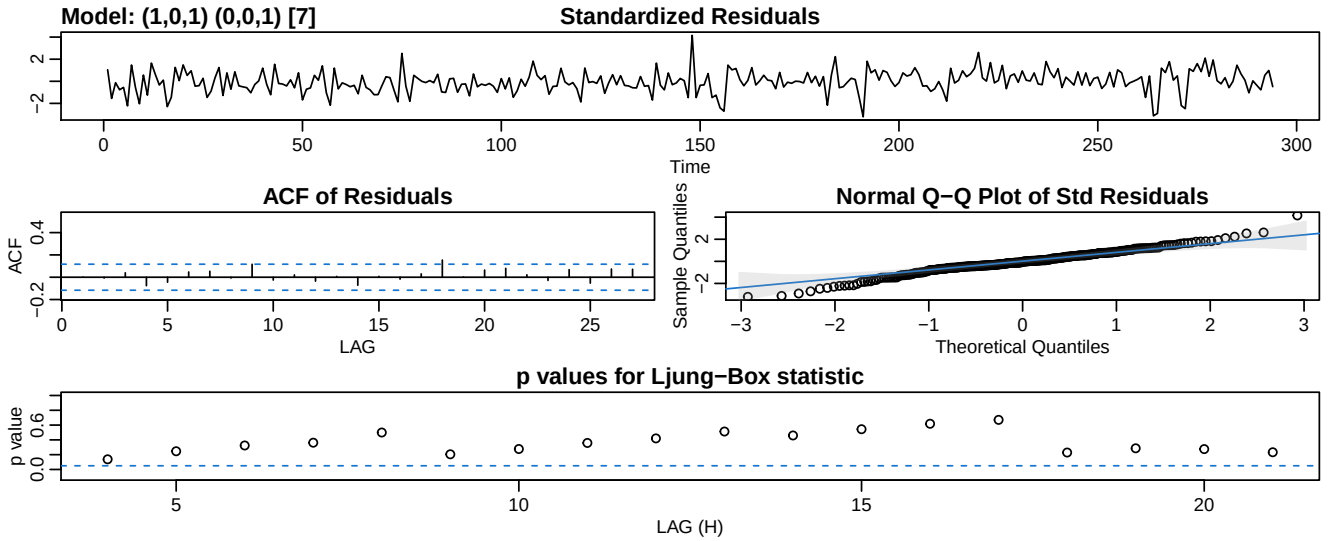

Figure 9: SARIMA diagnostics plots for Differencing Model with ARMA(1,1)x(0,1)[7]