Time Series Project

Hoang Sam Bui

03/01/2022

1. In this project, you are asked to find, analyze, and use an ARIMA(p, d, q) model for

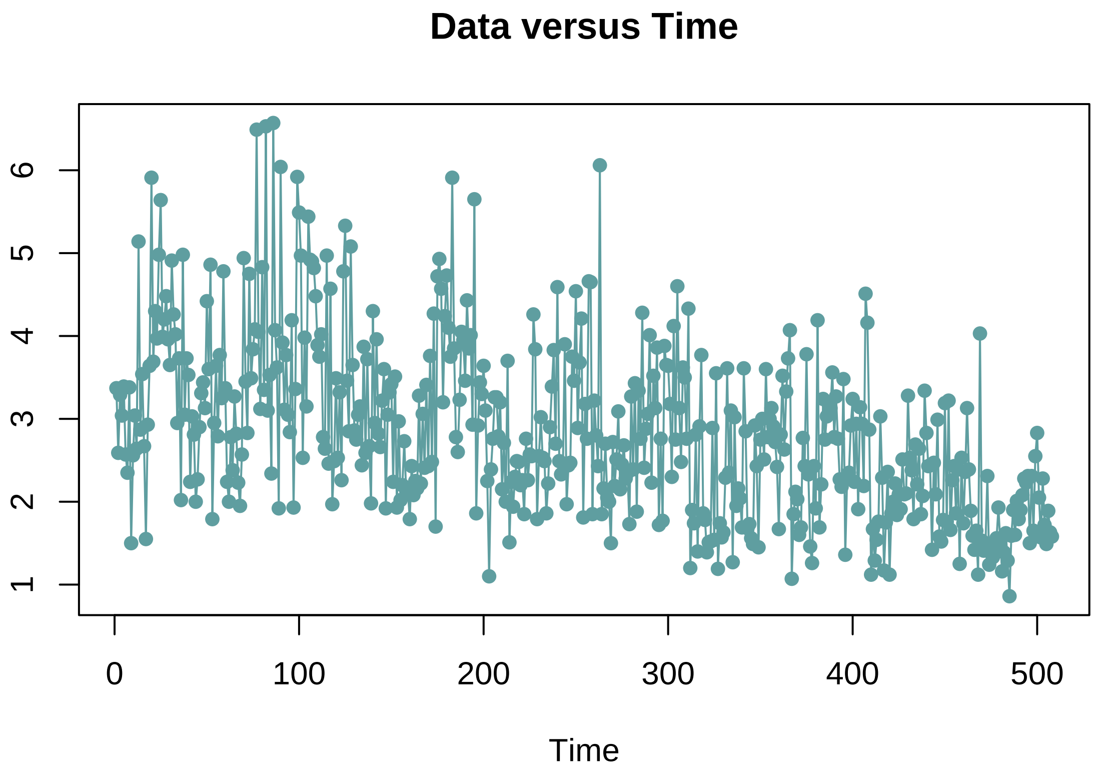
the data set so2.txt. This is data from monitoring atmospheric sulfur dioxide levels from *S. Mazumdar and N. Sussman, Relationships of Air Pollution to Health: Results from the Pittsburg Study, Arch. Env.Health., 38: 17-24, 1983.*

2. Part 1: Determination of d.

You should consider a plot of the data and fit a cubic polynomial using least squares then compare relative sizes of the coefficients. Use the results to choose a d.

Note that overdifferencing, or choosing d too high, will result in significant loss of points.

#(a) Present a plot of the original data.

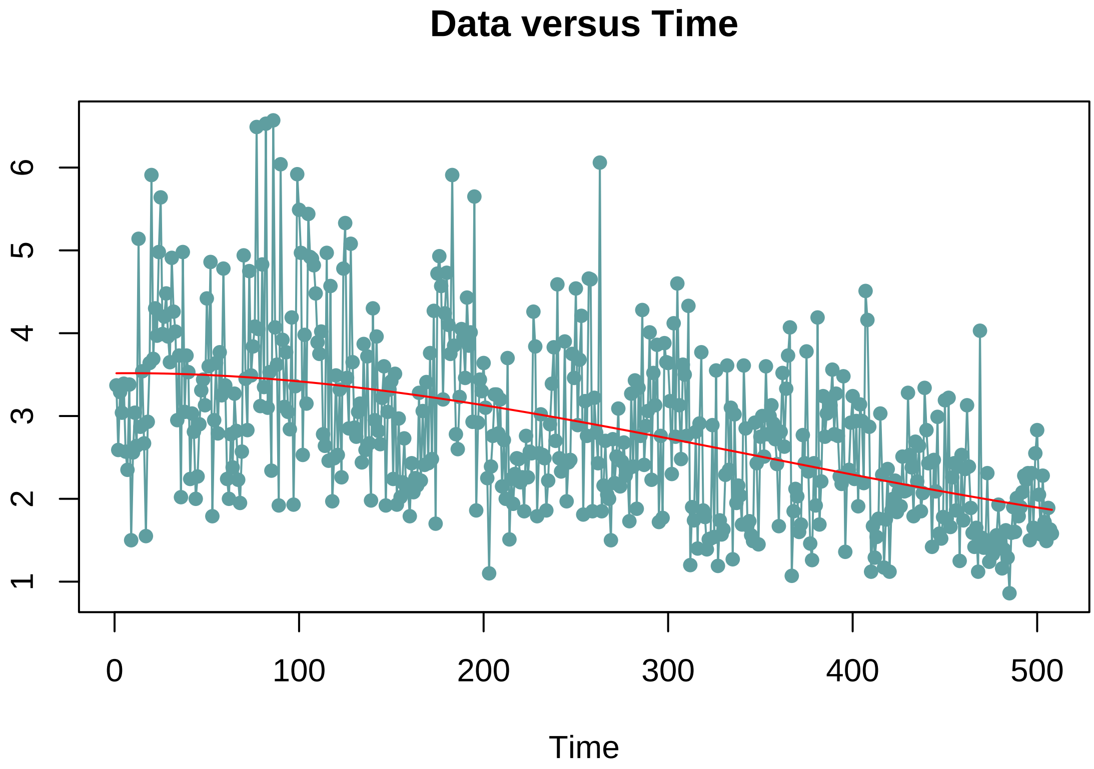


#(b) Make observations about possible trends.

- We observe there is decreasing trend.

#(c) Report the results of the least square polynomial fit and the relative sizes of coefficients.

```
##
## Call:
## lm(formula = Y ~ t1 + t2 + t3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0196 -0.6509 -0.0949  0.5125  3.1737
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.516e+00  1.626e-01  21.625   <2e-16 ***
## t1           2.206e-04  2.764e-03   0.080    0.936
## t2          -1.329e-05  1.261e-05  -1.054    0.292
## t3           1.274e-08  1.629e-08   0.782    0.435
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.9094 on 504 degrees of freedom
## Multiple R-squared:  0.2551, Adjusted R-squared:  0.2507
## F-statistic: 57.54 on 3 and 504 DF,  p-value: < 2.2e-16
```



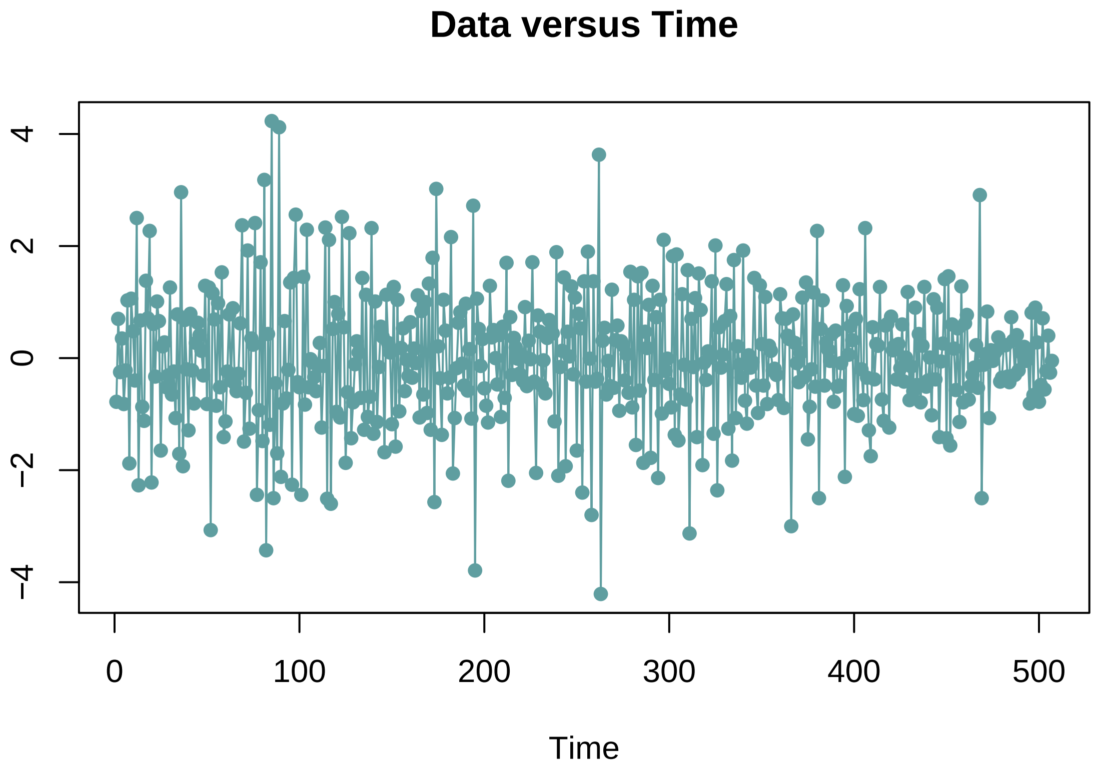
- The coefficients: $a_0 = 3.516 * 10^0$, $a_1 = 2.206 * 10^{-04}$, $a_2 = -1.329 * 10^{-05}$, $a_3 = 1.274 * 10^{-08}$.

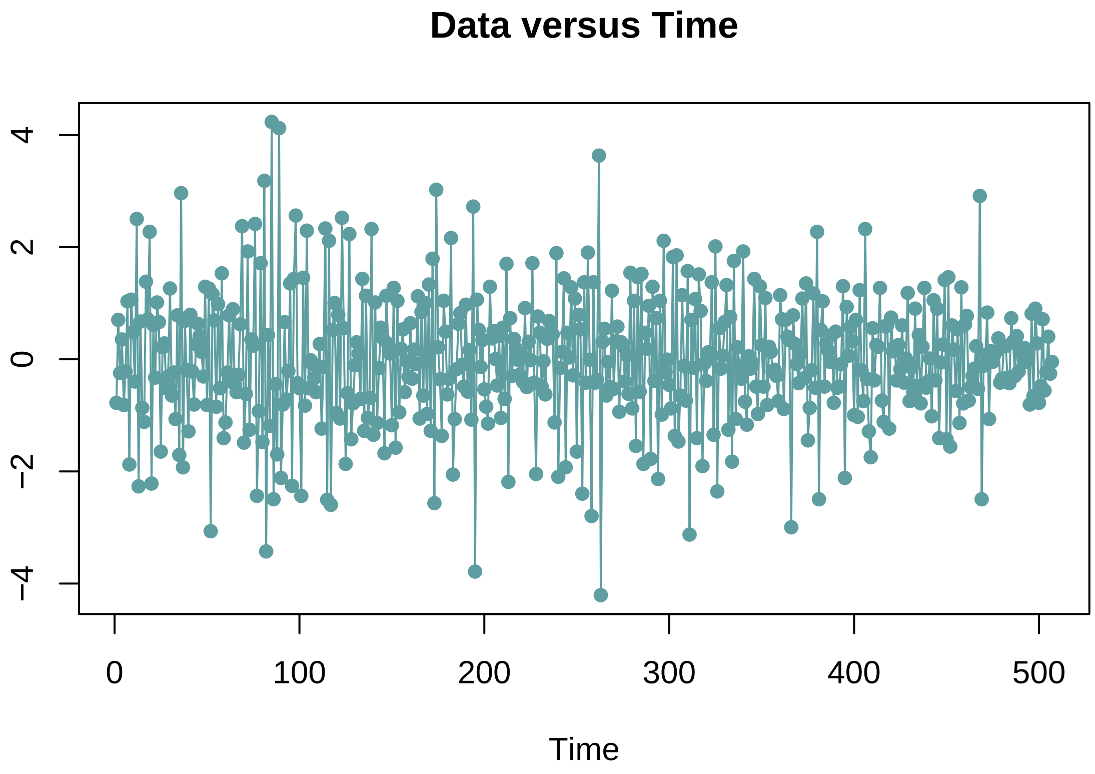
- We observe that the size of a_0 is relatively larger than the others three.

#(d) Specify d and explain your choice.

- We choose d = 1 since the coefficient size of each explanatory variable are relatively small. It appears that d = 1 is good enough to remove the trend of the data.

#(e) For the chosen d, display a plot of the mean-centered differenced data.



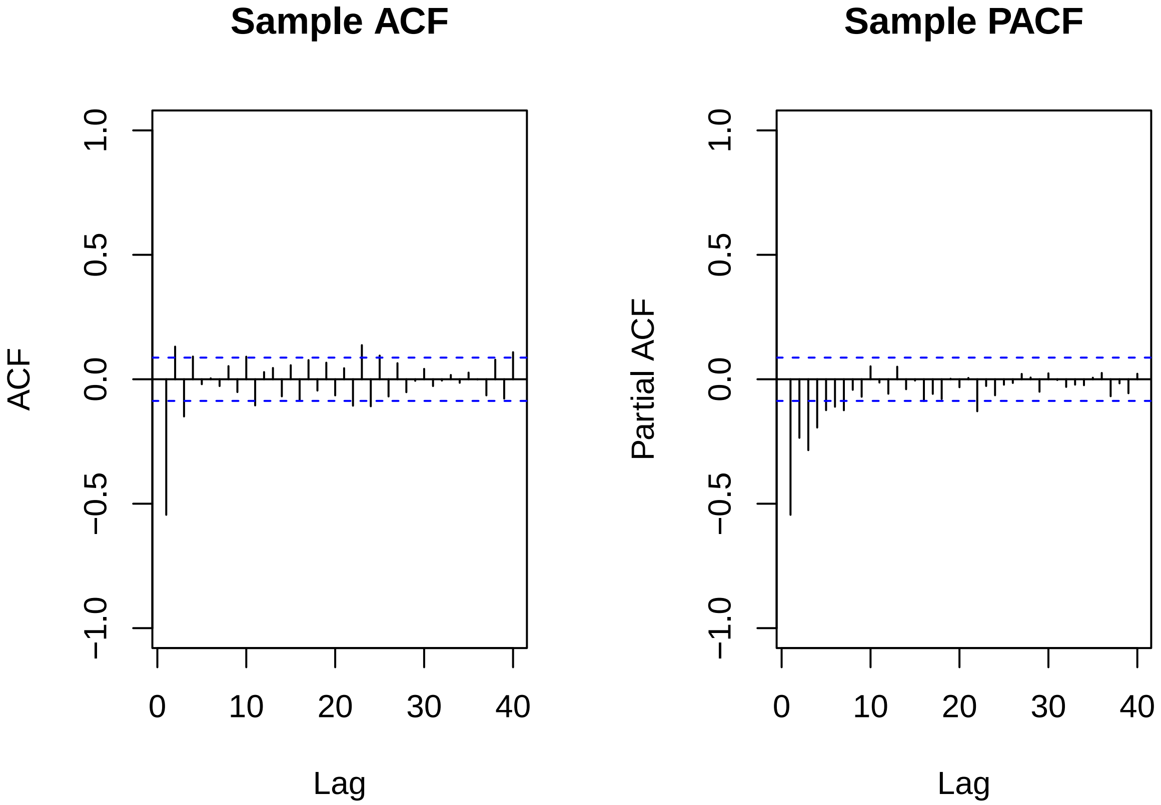


3. (44 points)__ Part 2: Determination of p and q for the mean-centered differenced data.You should begin by using a plot of the sample acf/pacf to make observations about possible orders of dependency. Then, you must use MLE to fit an ARMA(p, q) model for at least four combinations of p and q. Compare the plots of the ARMA(p, q) model together with sample acf/pacf values, plots of the model residuals, and the aic or aicc values to choose p and q.

Hint: The best model has $p > 1$ and $q > 1$. Some of the assigned points will depend on how close you get to the optimal values.

For your answer:

(a) Show the plot of the sample acf/pacf for the mean-centered differenced data.

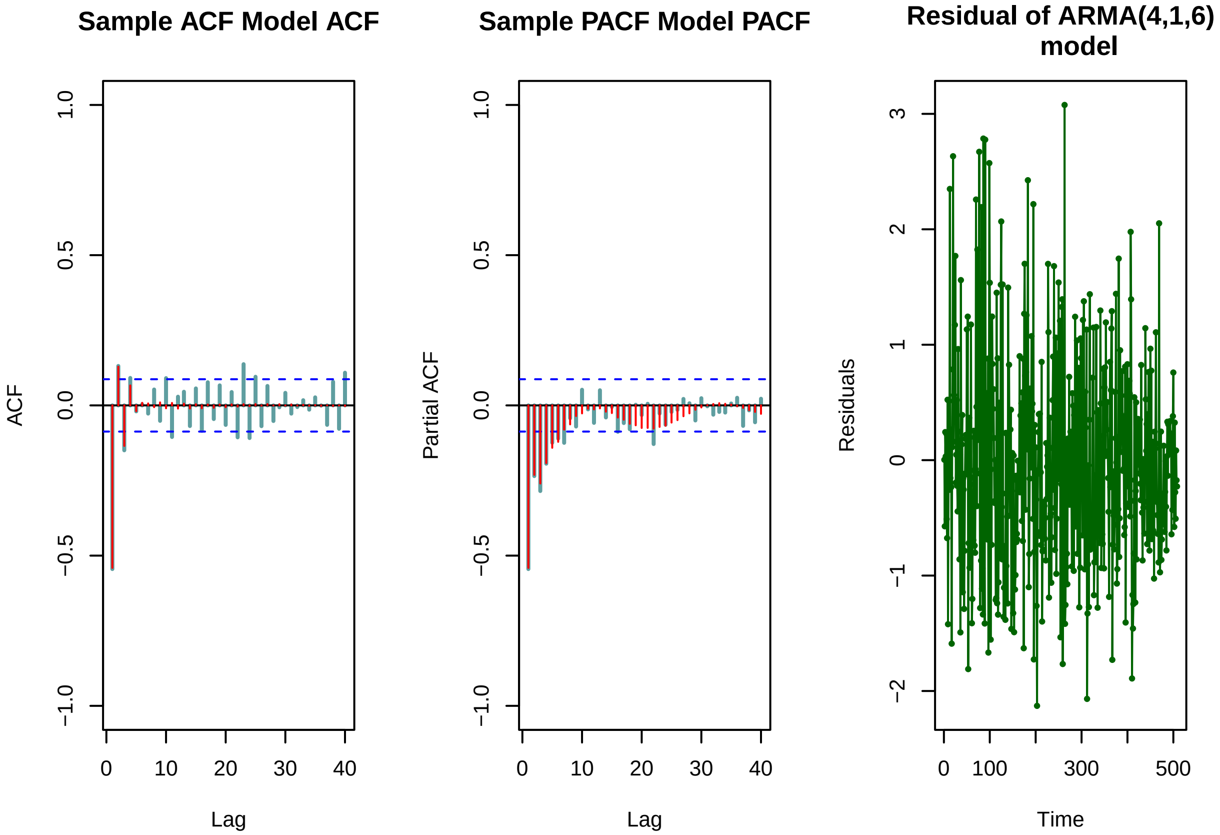


(b) Give observations on possible orders of dependency.

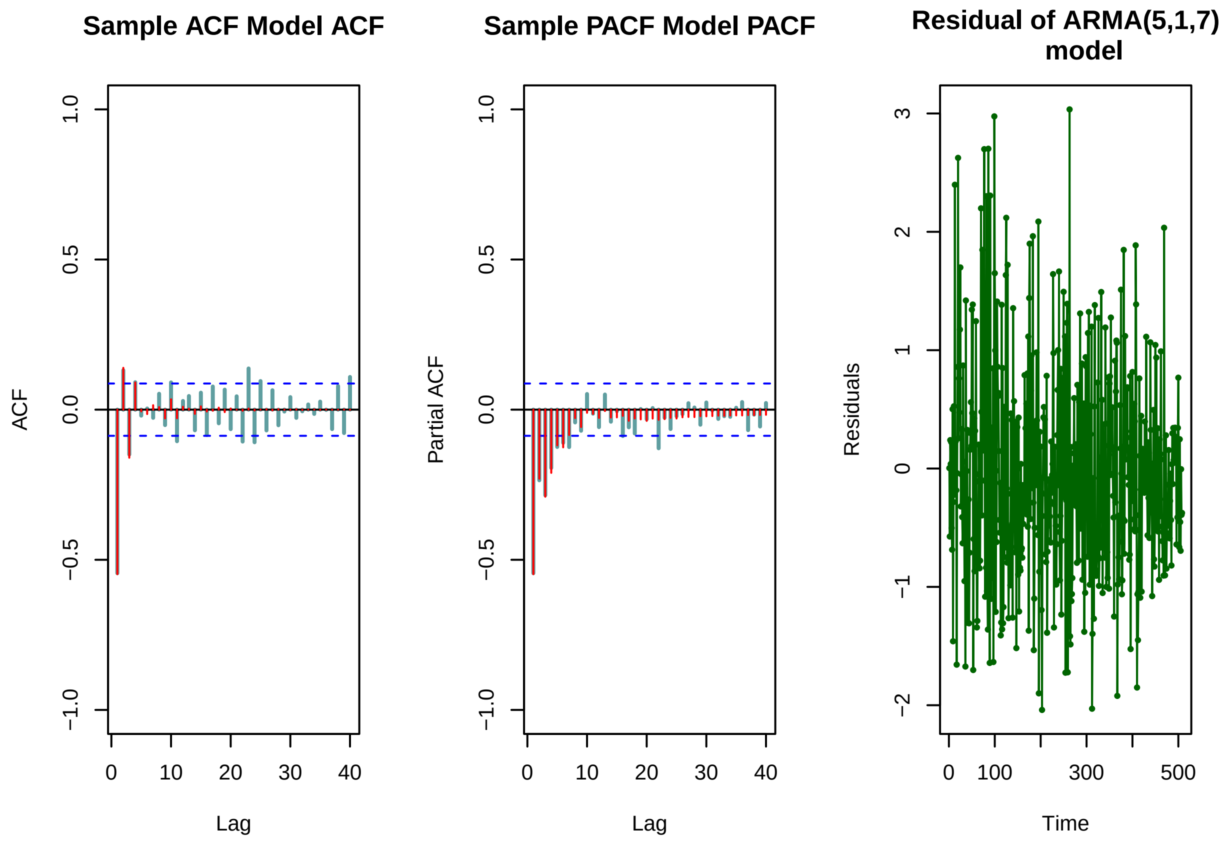
- In the ACF plot, we observe there is a cut off at the lag 2.
- In the PACF plot, we observe that it appears that there is exponential decay in first 6 lags.
- We think that the possible orders of dependency is $q = 2$.

(c) For the ARMA(p, q) estimated using MLE, show plots of the model acf/pacf values together with the sample acf/pacf and plots of the model residuals for four choices of p and q. *Display plots for only four choices even if you try more. If you try more, display results for values that help justify your final choice.*

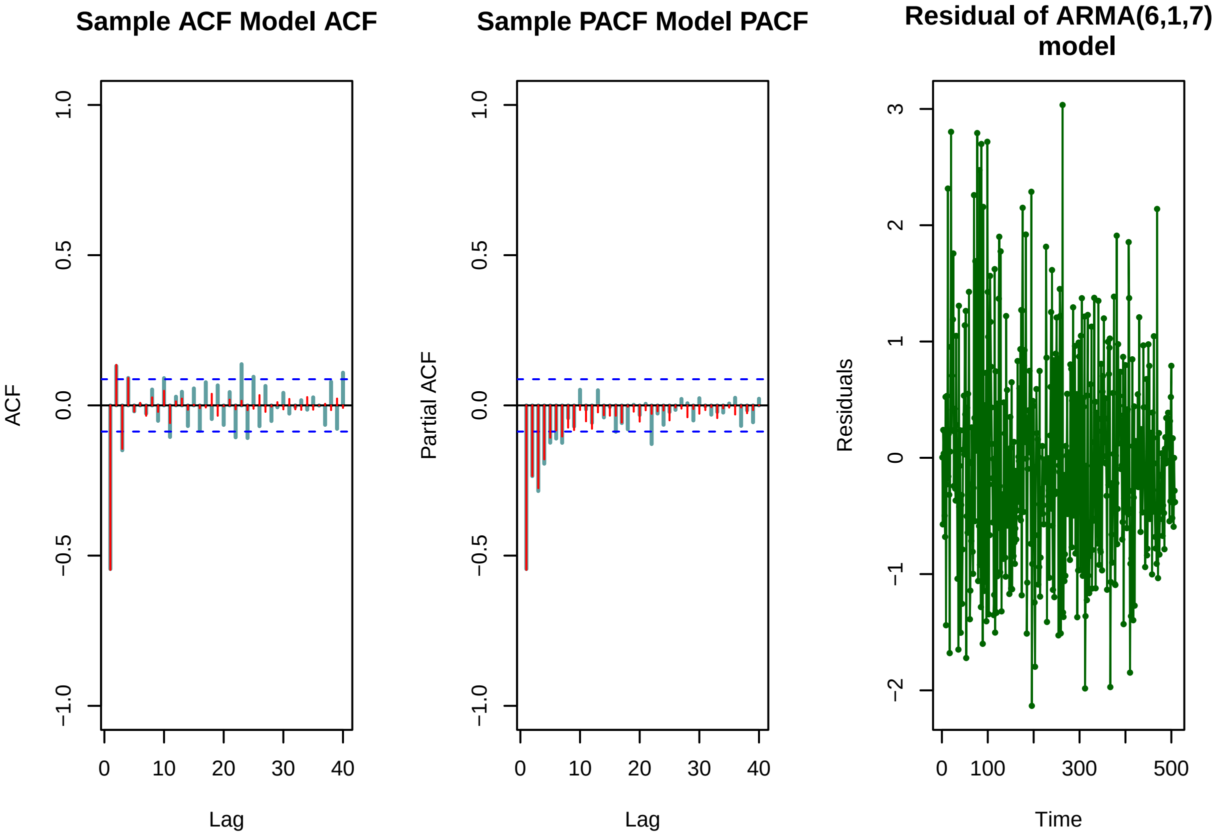
(i)First model ARMA(4,1,6)



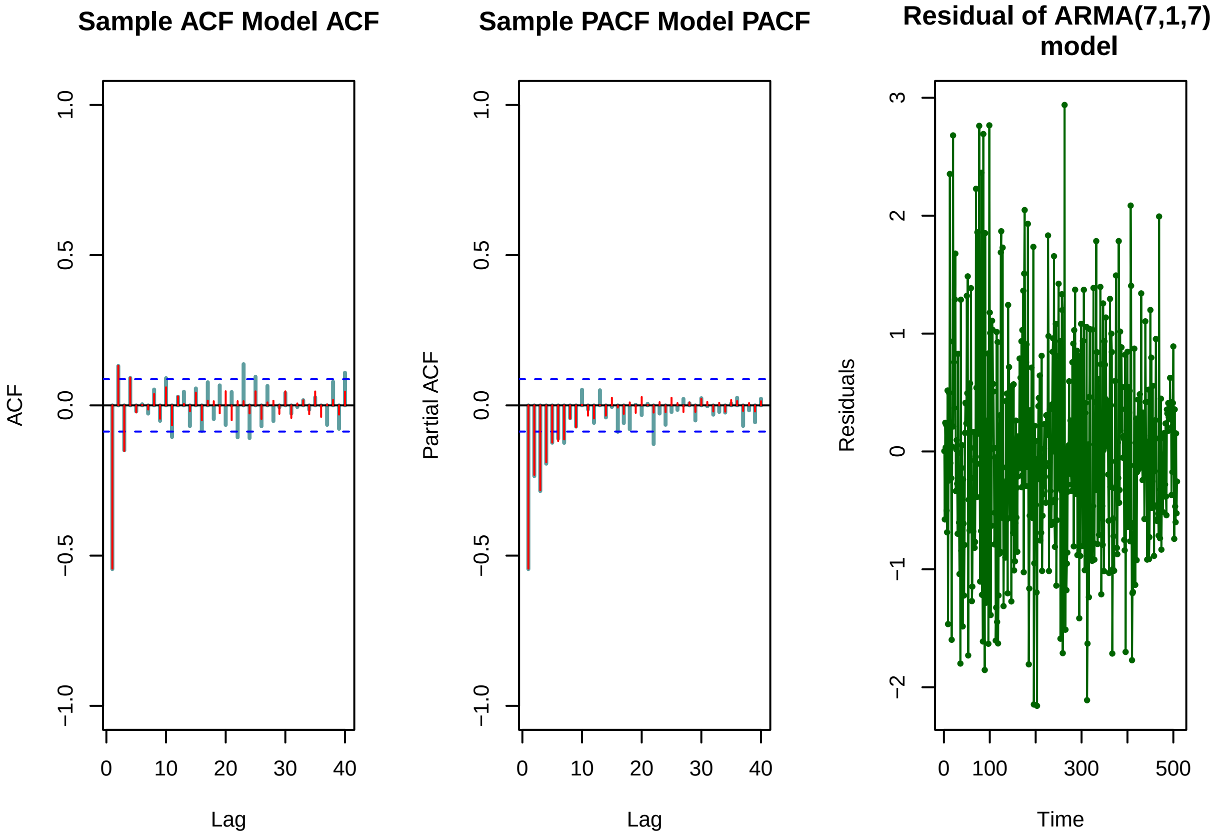
(ii) Second model ARMA(5,1,7).



(iii) Third model ARMA(6,1,7).



(iv) forth model ARMA(7,1,7).



(d) Give the aic or aicc values for each of the estimated models in (c).

```
##   Model Name       AIC
## 1 ARMA(4,1,6) 1313.340
## 2 ARMA(5,1,7) 1318.701
## 3 ARMA(6,1,7) 1318.347
## 4 ARMA(7,1,7) 1323.160
```

(e) Specify the p and q values you choose and give the reason.

- We choose $p = 5, andq = 7$ since the model ARMA(5,1,7) has a low AIC value among all the models we tried. Although the AIC of model ARMA(5,1,7) is not the lowest one, but we believe that this AIC value is still a sufficient.
- Although, we observe the ARMA(4,1,7) has the lowest AIC value, the PACF of the ARMA(4,1,6) shows a slow decay with a possibility of having a periodic trend in the model. Meanwhile, the ACF/PACF of sample and model ARMA(5,1,7) plots show the model value is more fitted with the sample value.

4. (36 points) Part 3: Use MLE to fit the ARMA(p, q) model for the chosen p and q and analyze the model.

You have already displayed the original and mean-centered differenced data in 1. You are working with that data!

For your answer,

(a) Specify p, d, and q.

- $p = 5, d = 1, q = 7$

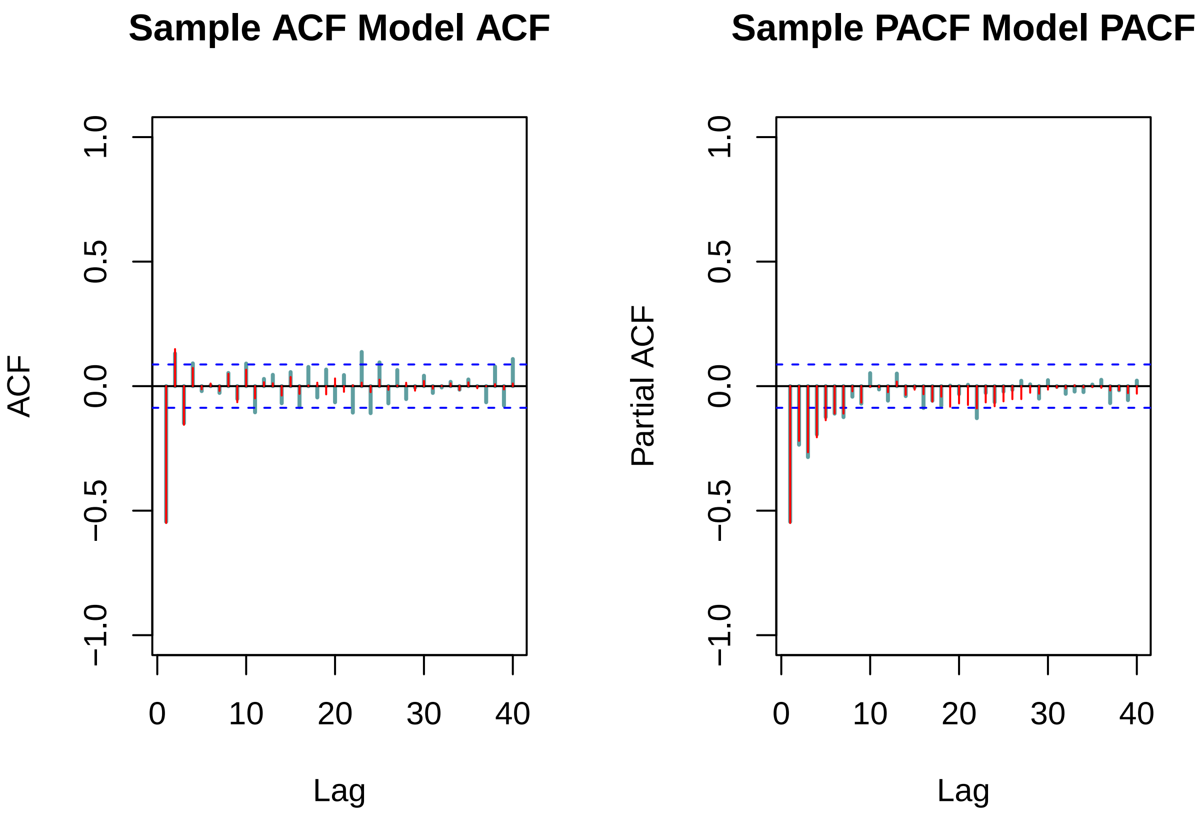
(b) Give the estimated coefficients for the MLE fit.

```
## Series: diff.data
## ARIMA(5,0,7) with non-zero mean
##
## Coefficients:
##           ar1     ar2     ar3      ar4      ar5      ma1      ma2     ma3
##       -0.3672  1.2272  0.9835  -0.6204  -0.5624  -0.5566  -1.4751  0.0017
## s.e.   0.1278  0.0943  0.0673   0.0973   0.1631   0.1336   0.2206     NaN
##          ma4     ma5      ma6      ma7     mean
##       1.4789  0.1526  -0.4378  -0.1635  -0.0036
## s.e.     NaN  0.2510   0.1810   0.0502   0.0005
##
## sigma^2 estimated as 0.7394:  log likelihood=-639.24
## AIC=1306.48   AICc=1307.33   BIC=1365.68
```

(c) Give the value of the AIC or AICC.

```
## [1] 1306.481
```

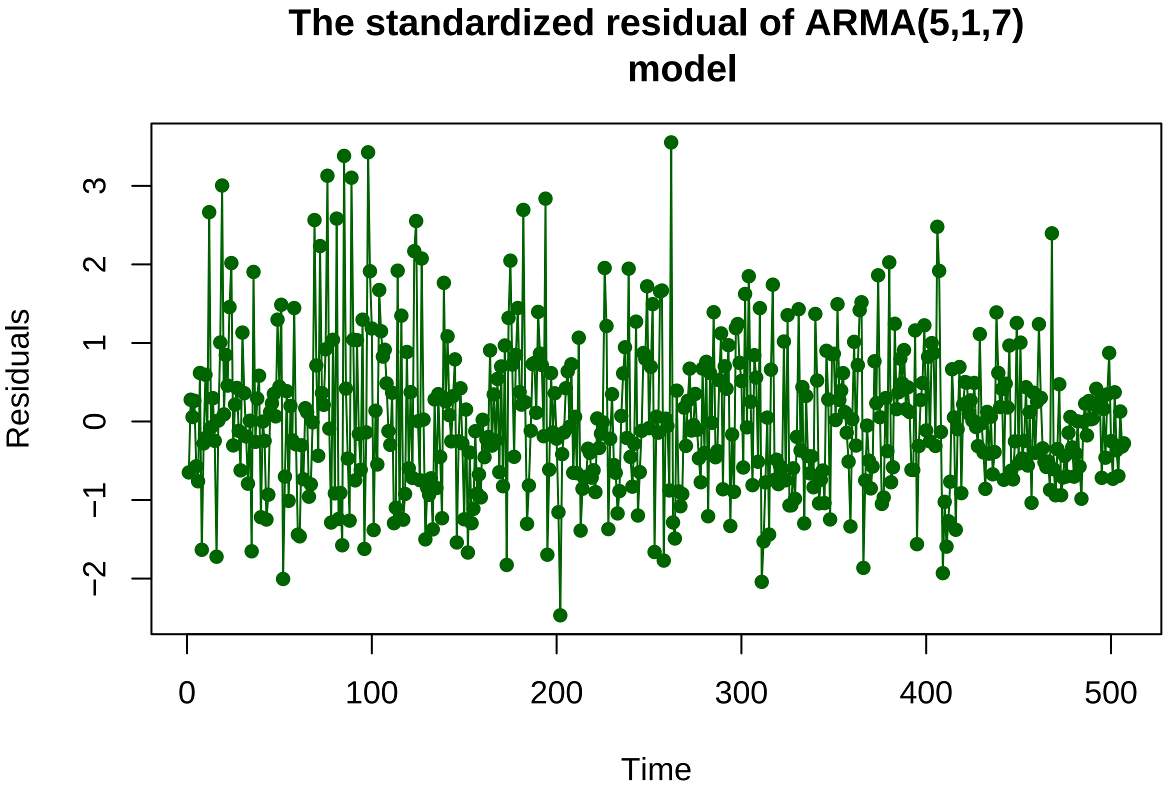
(d) Plot the model and sample acf/pacf values together.



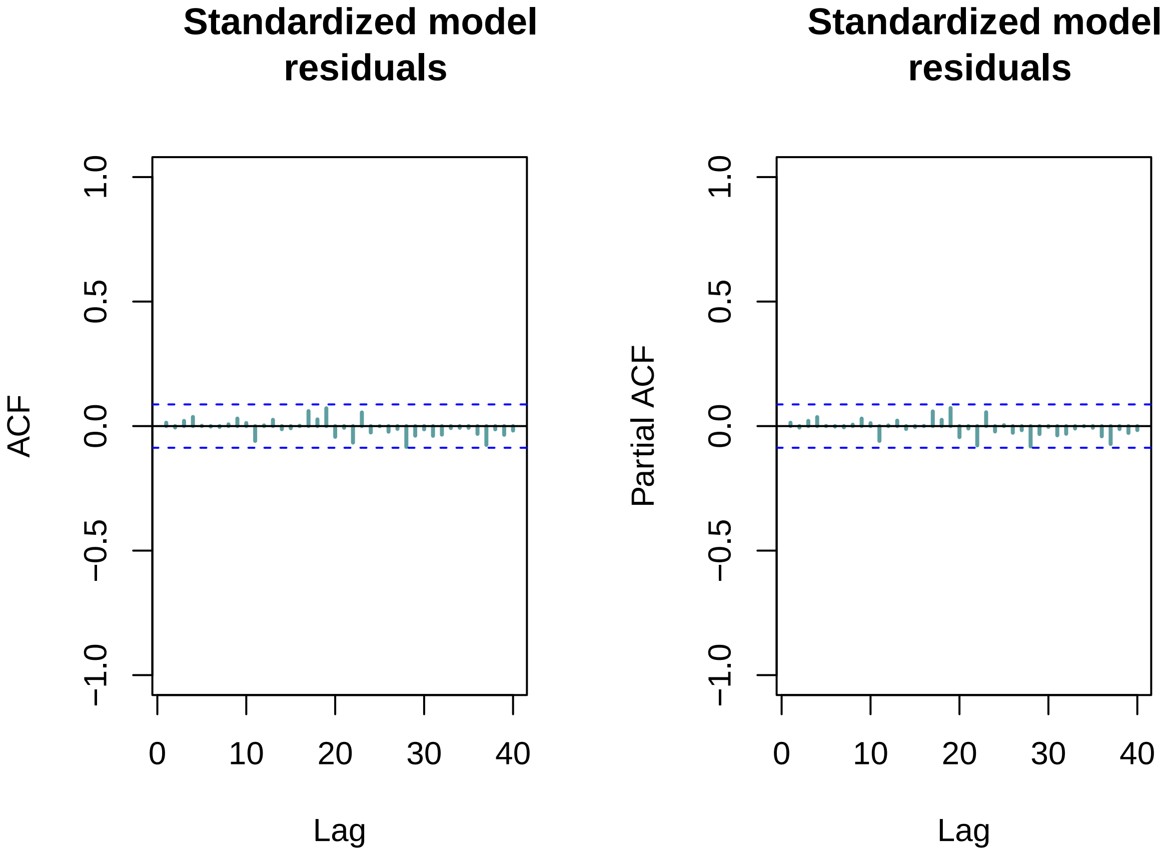
(e) Use the plot from (d) to assess the quality of the model fit.

- In ACF plot, the model fitted approximately well with the sample at the first 6 lags.
- In the PACF plot, the model fitted approximately well with the sample at the first 9 lags.

(f) Plot the standardized model residuals.



(g) Plot the sample acf/pacf for the standardized model residuals.



(h) Assess the plots from (f) and (g) with respect to the hypothesis that the model residuals behave like iid noise.

- In the ACF plot, we do not observe any value that is outside of the bound. The ACF plot support the hypothesis that the model residuals behave like iid noise.
- In the PACF plot, we do not observe any value that lies outside of the bound. The PACF support the hypothesis that the model residuals behave like iid noise.

(i) Evaluate the Ljung-Box and McLeod-Li statistics and indicate if they support rejection of the hypothesis that the model residuals behave like iid noise.

```
##
##  Box-Ljung test
##
## data:  arma517.stad.res
## X-squared = 9.8828, df = 20, p-value = 0.9702
```

```
##
##  Box-Ljung test
##
## data:  arma517.stad.res^2
## X-squared = 55.871, df = 20, p-value = 3.04e-05
```

- In the Ljung-Box test, the test statistic is 9.88 and the p-value is 0.9702, which is larger than 0.05. The Ljung-Box test supports the hypothesis that the model residuals behave like iid noise.
- In the McLeod-Li test, the test statistic is 55.87 and the p-value is $3.04 * 10^{-05}$, which is much lower than 0.05. The McLeod-Li test supports the rejection of the hypothesis that the model residuals behave like iid noise.

(j) Using (h) and (i), give a final assessment on the validity of the hypothesis that the model residuals behave like iid noise.

- Both of the ACF and PACF plots support the hypothesis that the model residuals behave like iid noise. The Ljung-Box test support the hypothesis that the model residuals behave like iid noise while the McLeod-Li test does not support the rejection of the hypothesis. In conclusion, we do not reject the hypothesis that the model residuals behaves as iid noise.

(k) Use the results from (e) and (j) to give a summary evaluation about the quality of the fitted model.

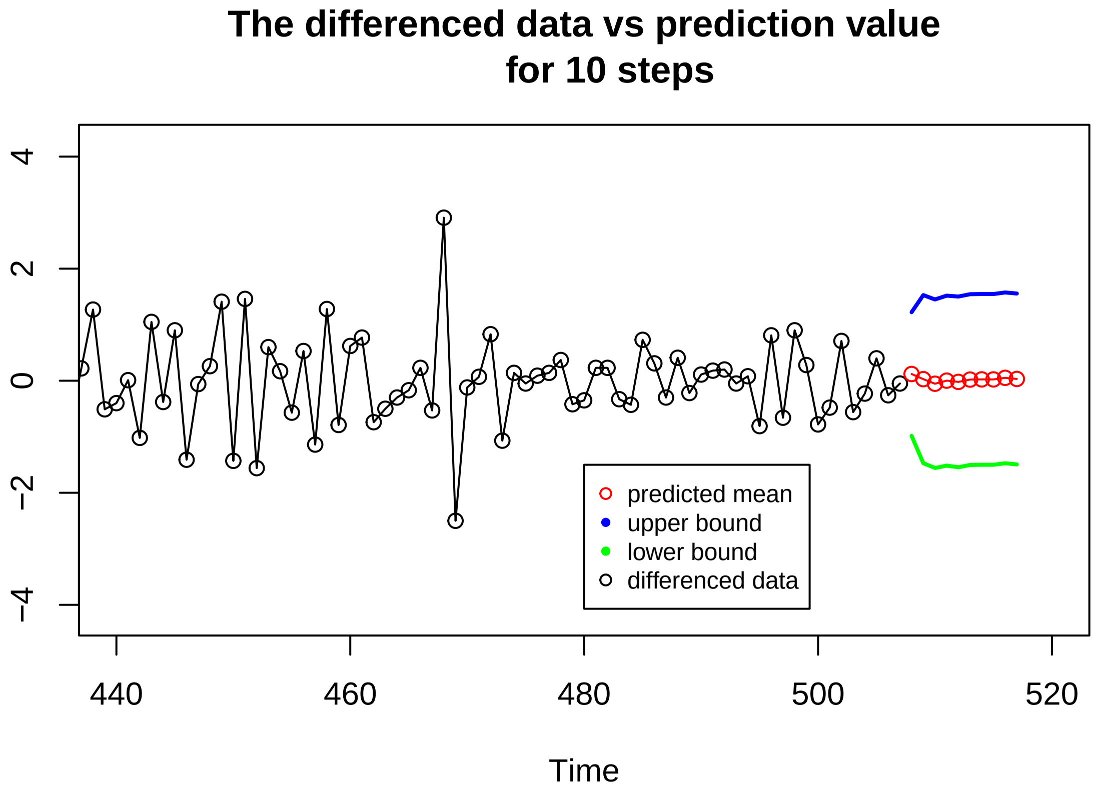
In 3., you compare the plots of model/sample acf/pacf and model residuals for different p and q to choose best values for p and q. In this question, you are asked to assess how well the model for the chosen p and q fits the data. The model corresponding to the best value of p and q may or may not be a good model!

- We observe that the AIC of the ARMA(5,1,7) is sufficient among others models.
- In the ACF plot, the model captures the cutoff well at the first 2 lags. In the second lag, the value between the model and sample is completely accurate but the difference between the model and sample value is acceptable.
- In the PACF plot, the model captures the exponential decay pattern of the sample quite well. In the first lag, the model value is approximately accurate with the sample value. From the second to the ninth lag, the model value is not as highly accurate as the sample value but the difference between model and sample values are not significant.
- In conclusion, the model corresponding to the best value of p and q might be a good model.

5. (3 points) Part 4: Use the estimated model to make a forecast.

For your answer,

(a) Plot the data together with prediction of values for 10 time steps past the last time of the data together with the confidence bounds.


The differenced data vs prediction value
for 10 steps
Time
○ predicted mean
● upper bound
● lower bound
○ differenced data