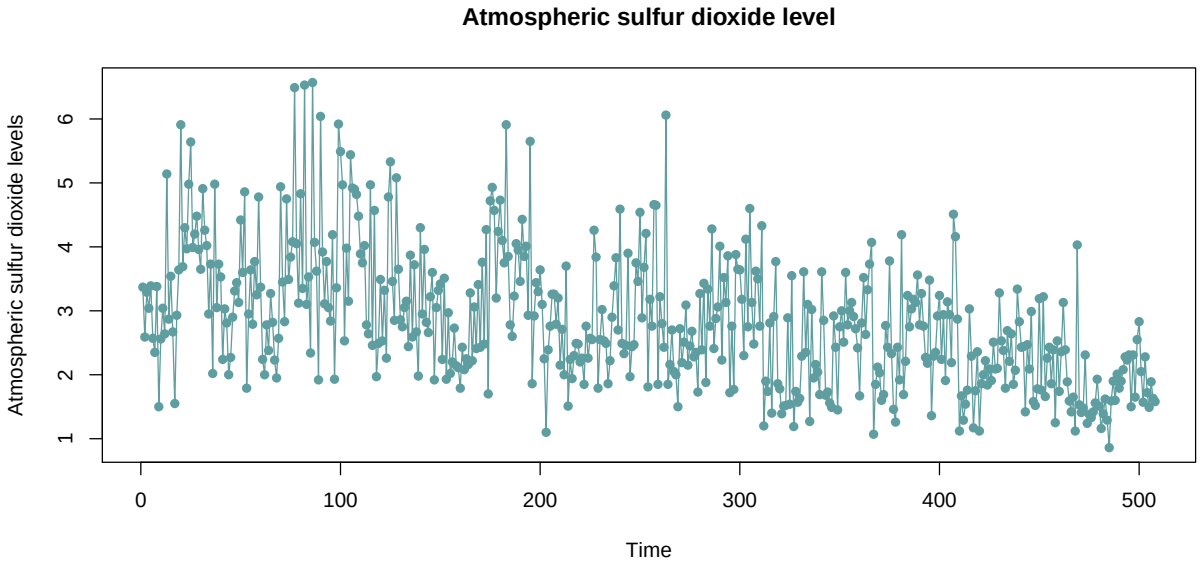
Atmospheric sulfur dioxide levels time series analysis

Dawu Liu

Part 1: Determination of d.

(a) Present a plot of the original data.



(b) Make observations about possible trends.

There seems to be a slightly downward trend, probably linear.

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

The trend model is: $a_0 + a_1 t + a_2 t^2 + a_3 t^3$

The coefficients are :
For t: $a_1 = 2.206 \times 10^{-04}$, for t^2: $a_2 = -1.330 \times 10^{-05}$, for t^3: $a_3 = 1.274 \times 10^{-08}$

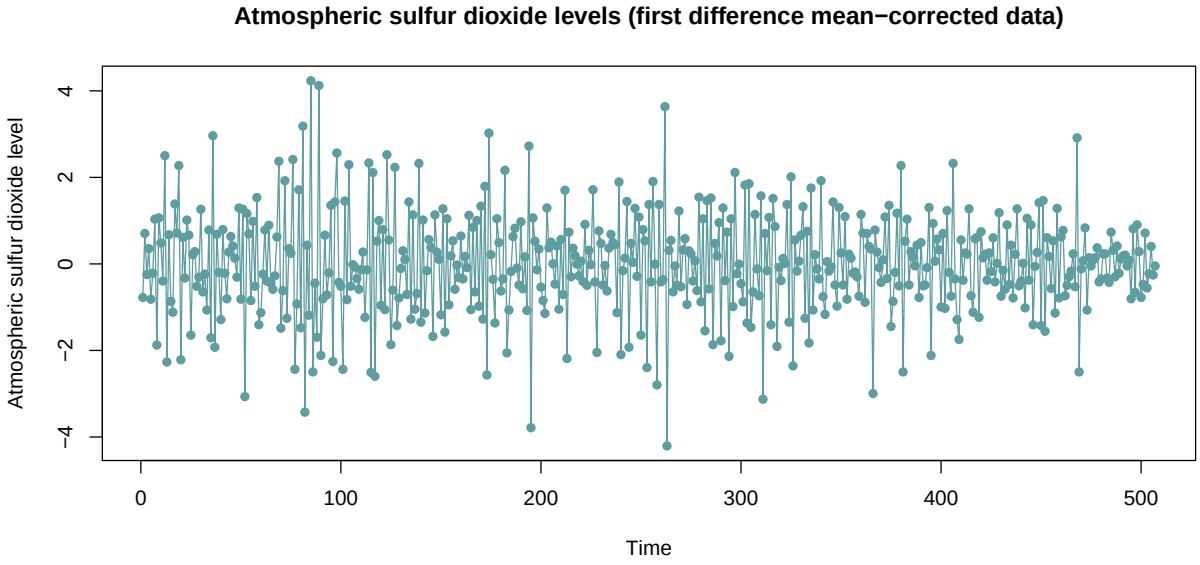
The first degree t has the largest coefficient, its about 16~17 times the magnitude of the coefficient of t^2, and more than 17000 times the magnitude of the coefficient of t^3. The coefficients for t^2 and t^3 are relatively much smaller compare to t.

```
##   (Intercept)             t        I(t^2)        I(t^3)
##  3.516166e+00  2.205607e-04 -1.329452e-05  1.273799e-08
```

(d) Specify d and explain your choice.

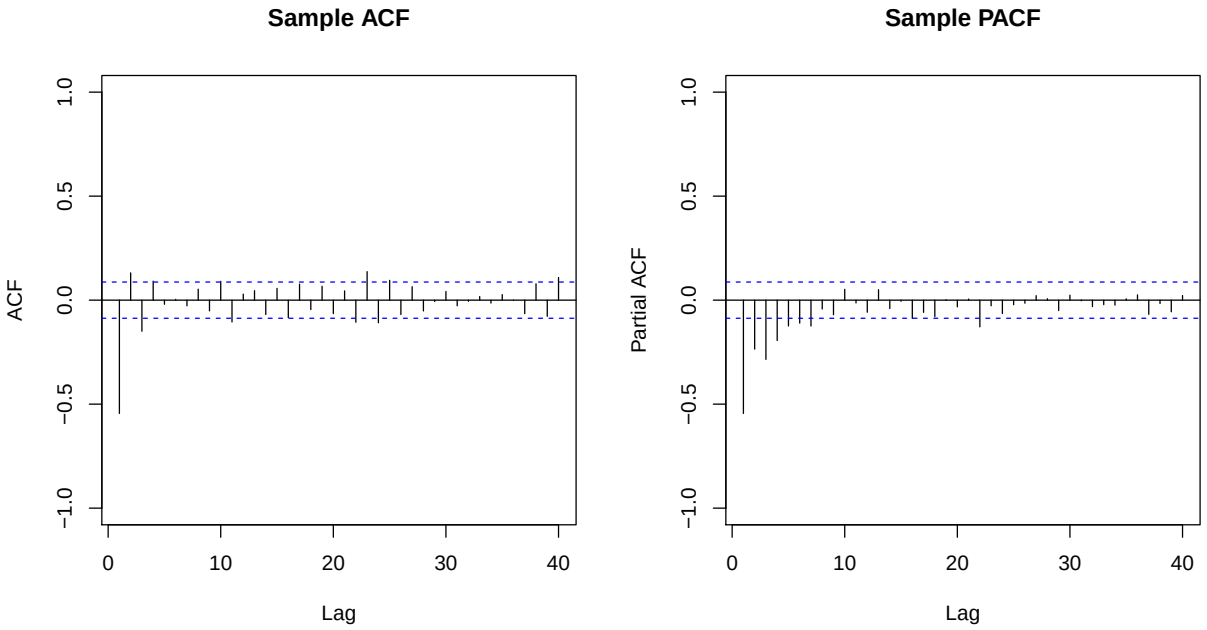
I will choose d=1. The coefficient for t (first degree) is the largest, and the coefficients of t^2 and t^3 are relatively much smaller. This indicates the trend is close to linear, first difference would do.

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



Part 2: Determination of p and q for the mean-corrected differenced data.

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

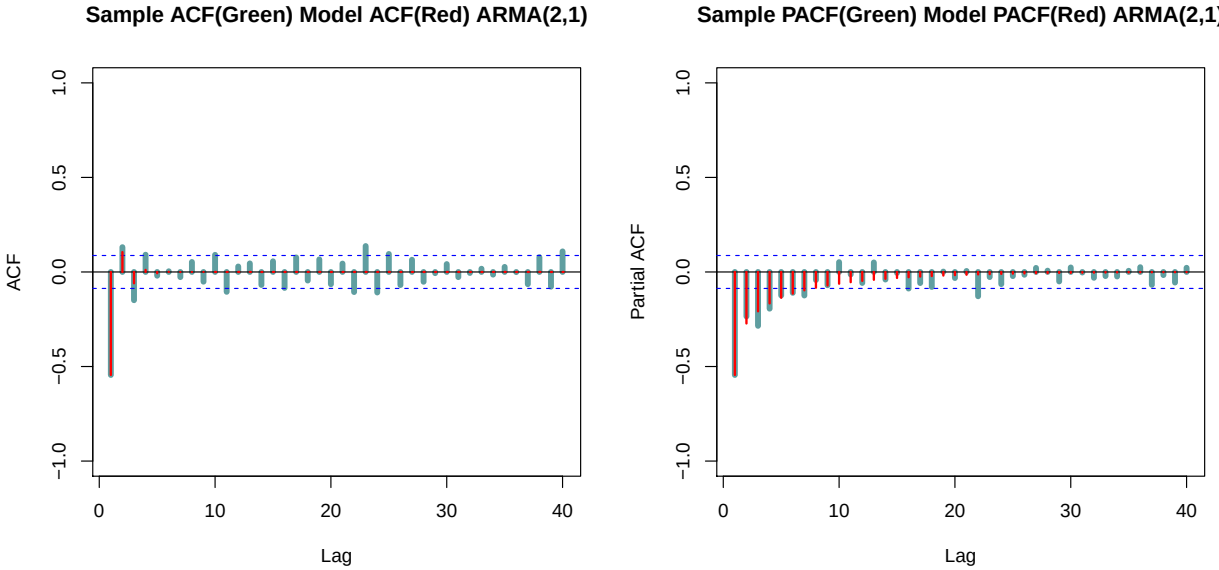


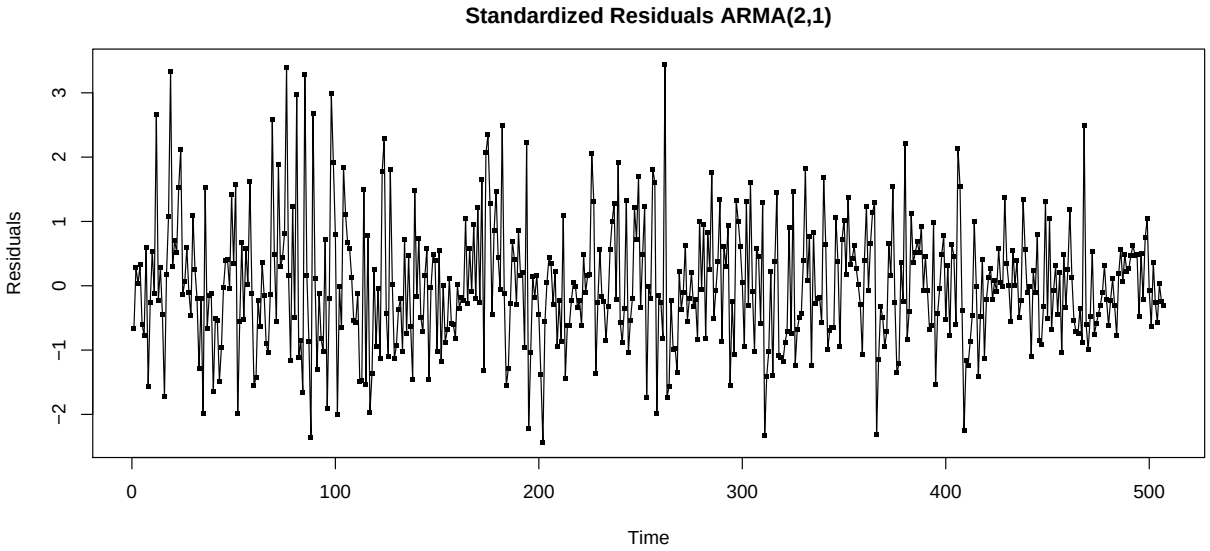
(b) Give observations on possible orders of dependency.

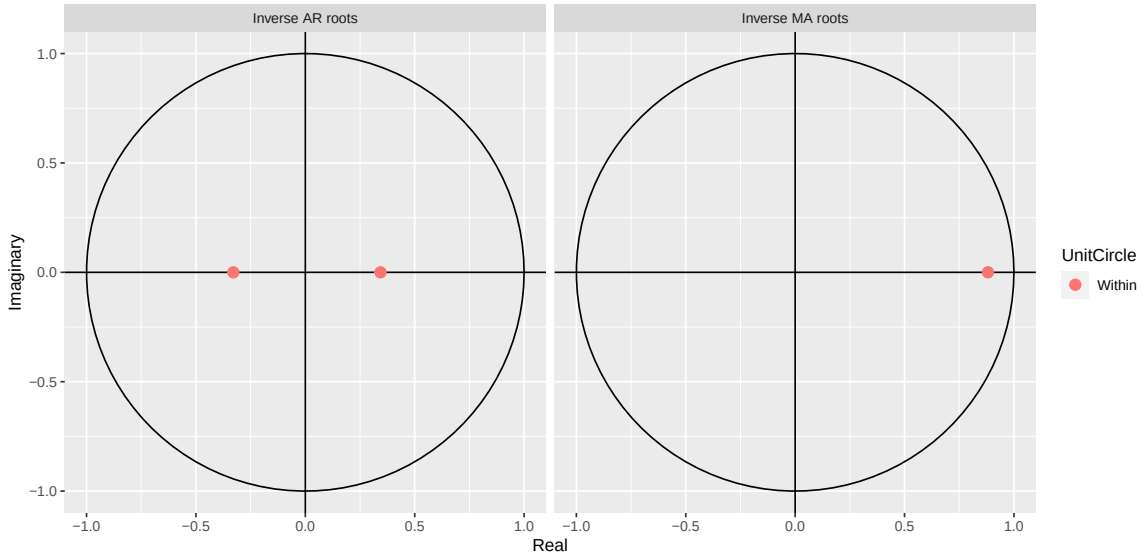
For both plots, the magnitude of acf and pacf decays. This suggest us to use an ARMA(p,q) model. The sample acf is significanct at lag = 1,2,3,4, and generally insignificant after. While the pacf plot has more significant values in the first few lags, this suggests that the order p is probably higher than q. My initial order would be $p = 2, q = 1$, the initial model would be ARMA(2,1).

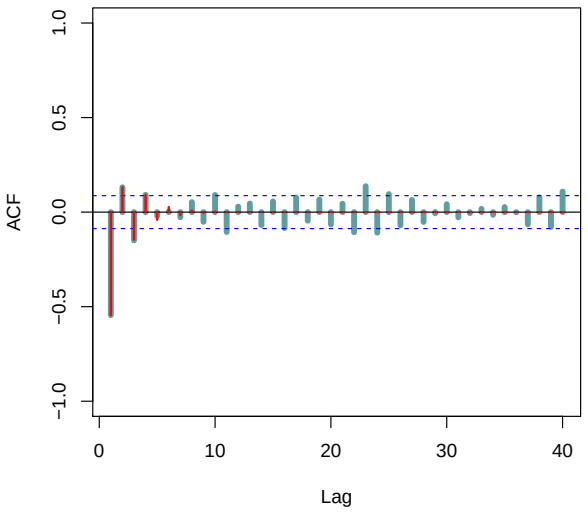
(c) For the ARMA(p, q) estimated using MLE, show plots of the model acf/pacf values together with the sample acf/pacf and plots of the model residuals for four choices of p and q.

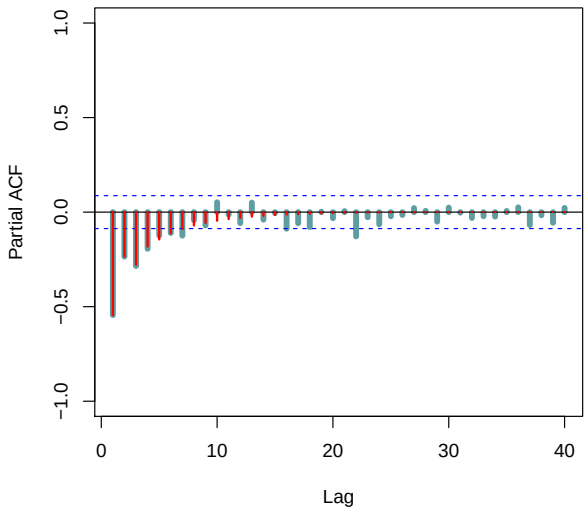
The models are ARMA(2,1), ARMA(3,2), ARMA(3,3), ARMA(4,4)

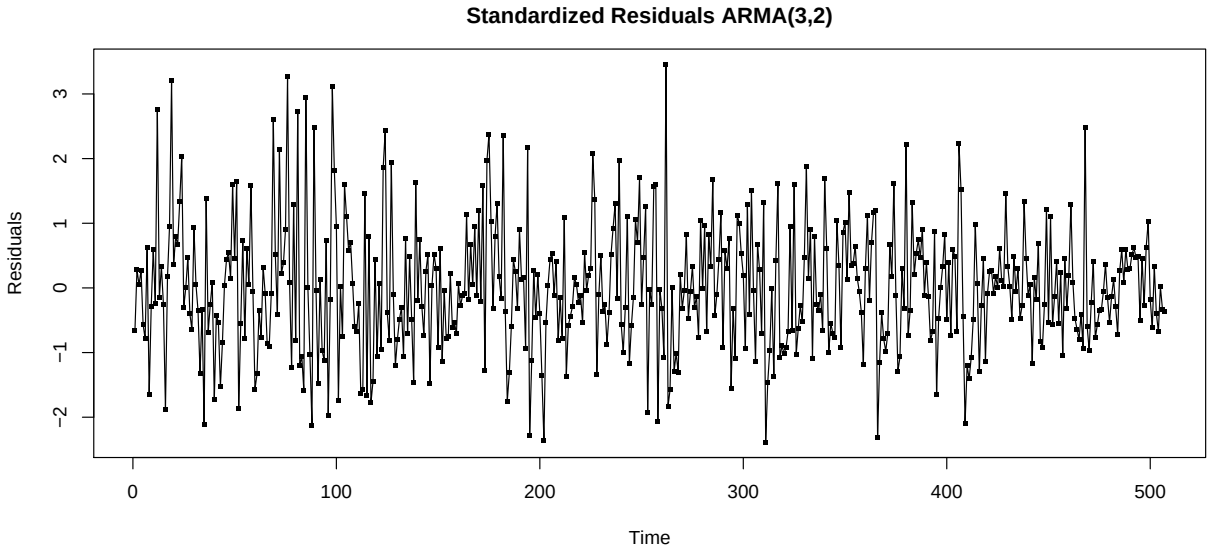


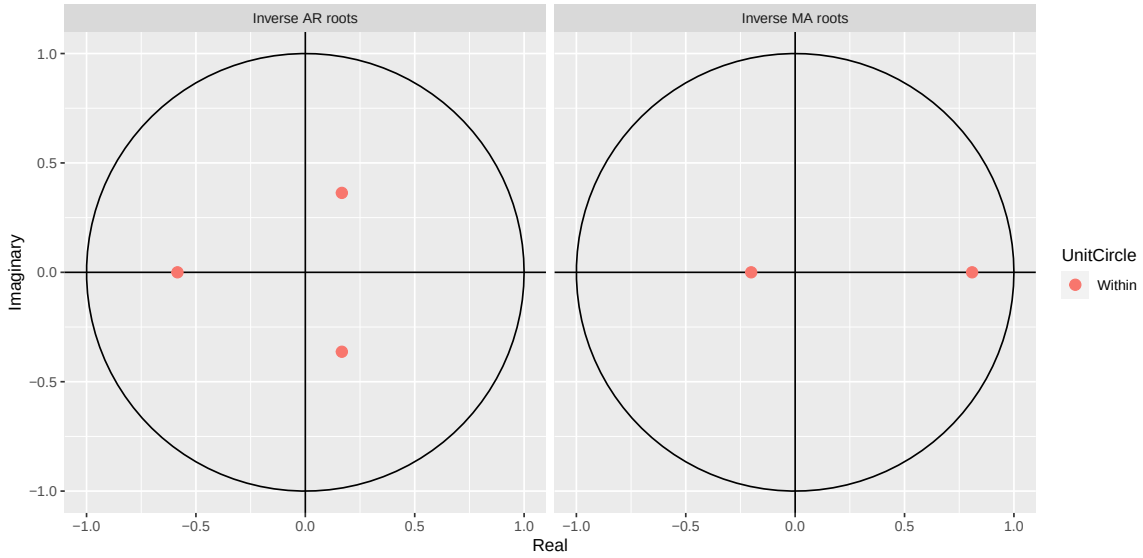

Standardized Residuals ARMA(2,1)
Residuals
Time

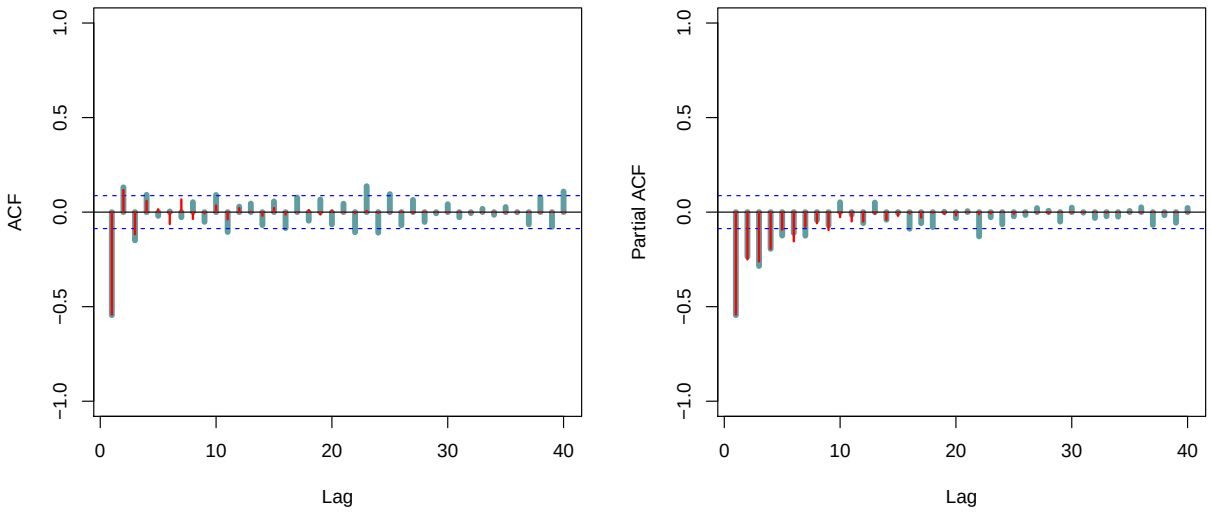

Inverse AR roots
Inverse MA roots
Imaginary
Real
UnitCircle
Within


Sample ACF(Green) Model ACF(Red) ARMA(3,2)
ACF
Lag

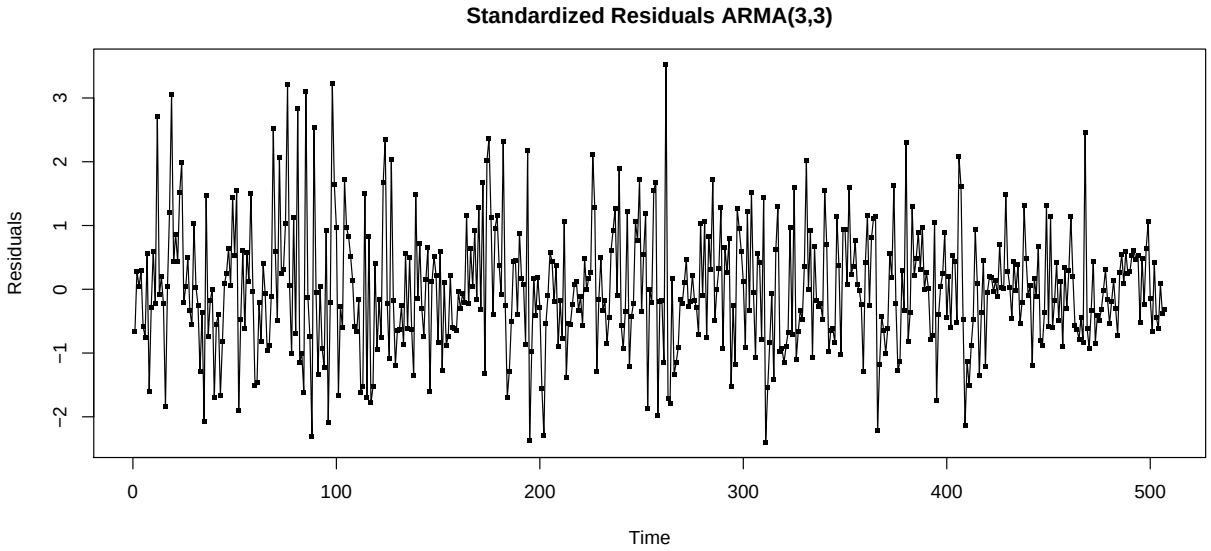

Sample PACF(Green) Model PACF(Red) ARMA(3,2)
Partial ACF
Lag

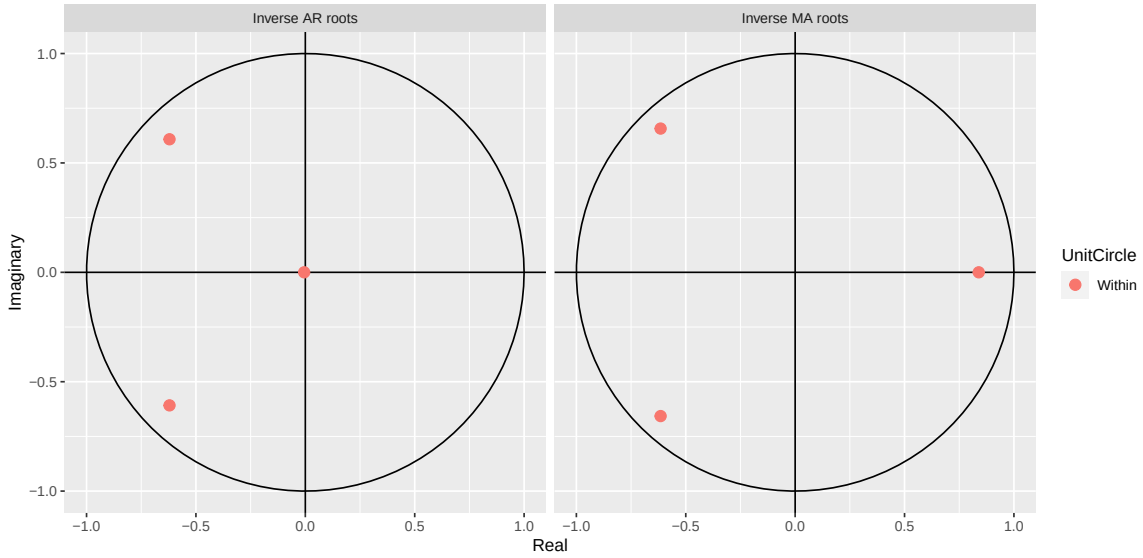

Standardized Residuals ARMA(3,2)
Residuals
Time

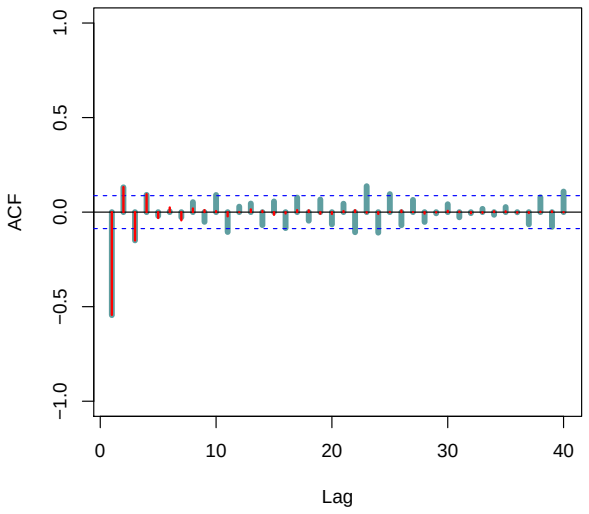

Inverse AR roots
Inverse MA roots
Imaginary
Real
UnitCircle
Within

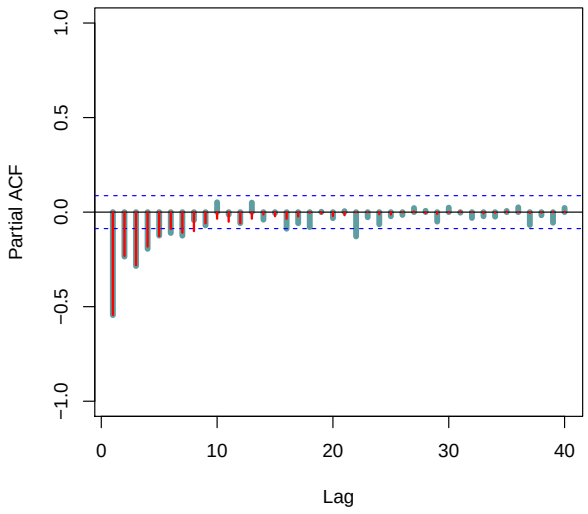

Sample ACF(Green) Model ACF(Red) ARMA(3,3)
ACF
Lag

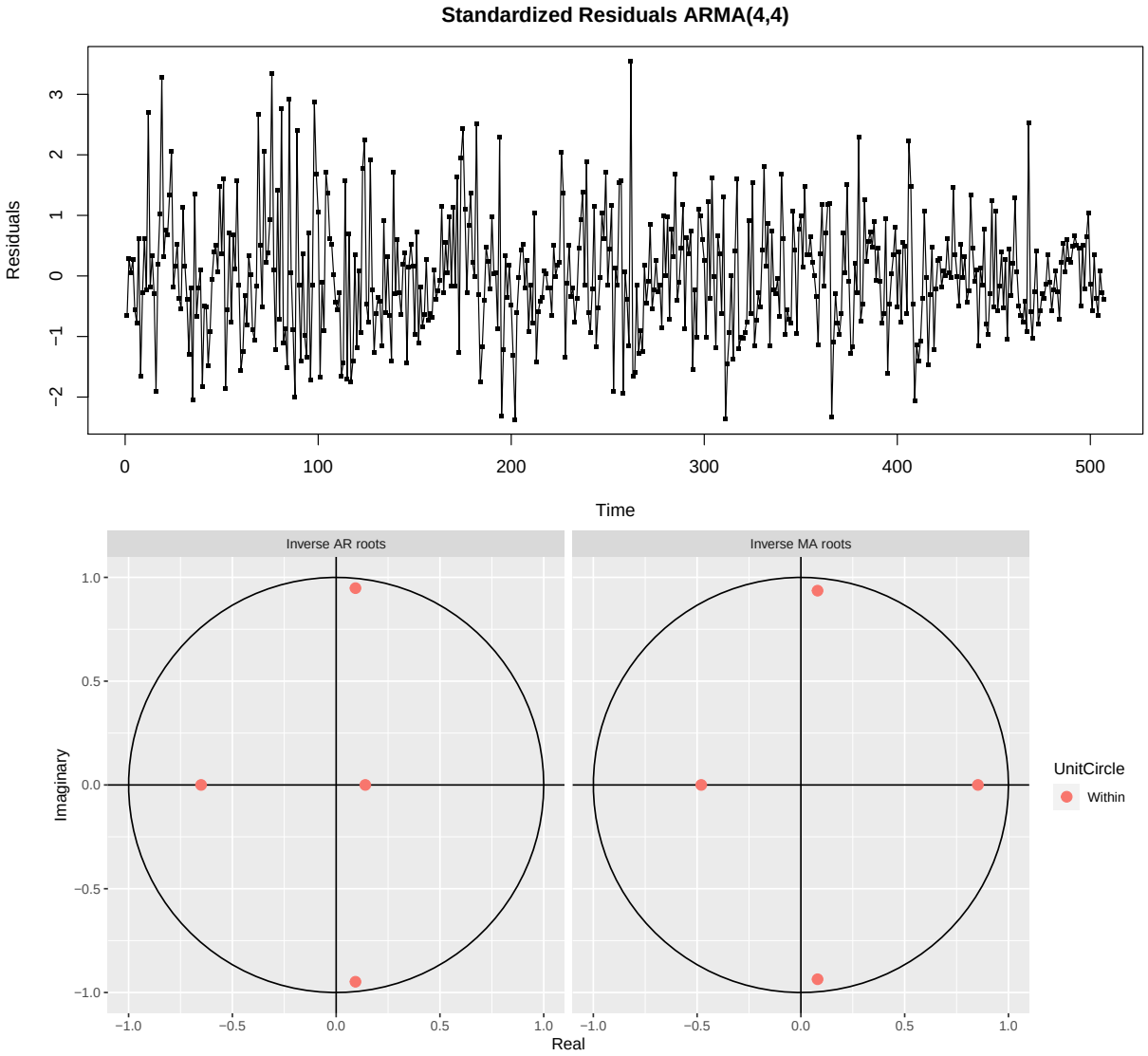
Sample PACF(Green) Model PACF(Red) ARMA(3,3)
Partial ACF
Lag


Standardized Residuals ARMA(3,3)
Residuals
Time


Inverse AR roots
Inverse MA roots
Imaginary
Real
UnitCircle
Within


Sample ACF(Green) Model ACF(Red) ARMA(4,4)
ACF
Lag


Sample PACF(Green) Model PACF(Red) ARMA(4,4)
Partial ACF
Lag



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

Model	AICC
ARMA(2,1)	1321.037
ARMA(3,2)	1321.433
ARMA(3,3)	1321.720
ARMA(4,4)	1325.774

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

My final choice is $\mathbf{p = 3}$, $\mathbf{q = 2}$, which is the **ARMA(3,2)** model for the differenced data. There doesn't appear to be any pattern in any of the residual plots. I see good agreement in the acf values up to lag 4 and good agreement in the pacf values to lag 6 for ARMA(3,2). The model acf/pacf of ARMA(3,2) visibly fit better than ARMA(2,1) and ARMA(3,3) in the plots. While ARMA(4,4) comes close in the acf/pacf plots, it has a higher order which increases the complexity of the model and higher AICC for almost no improvement. Also, the AICC value of ARMA(3,2) is the second lowest and only slightly higher than ARMA(2,1), the difference is really small(less than 0.4). Overall, ARMA(3,2) is the best out of the four models.

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

(a) Specify p, d, and q.

$p = 3, d = 1, q = 2$

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

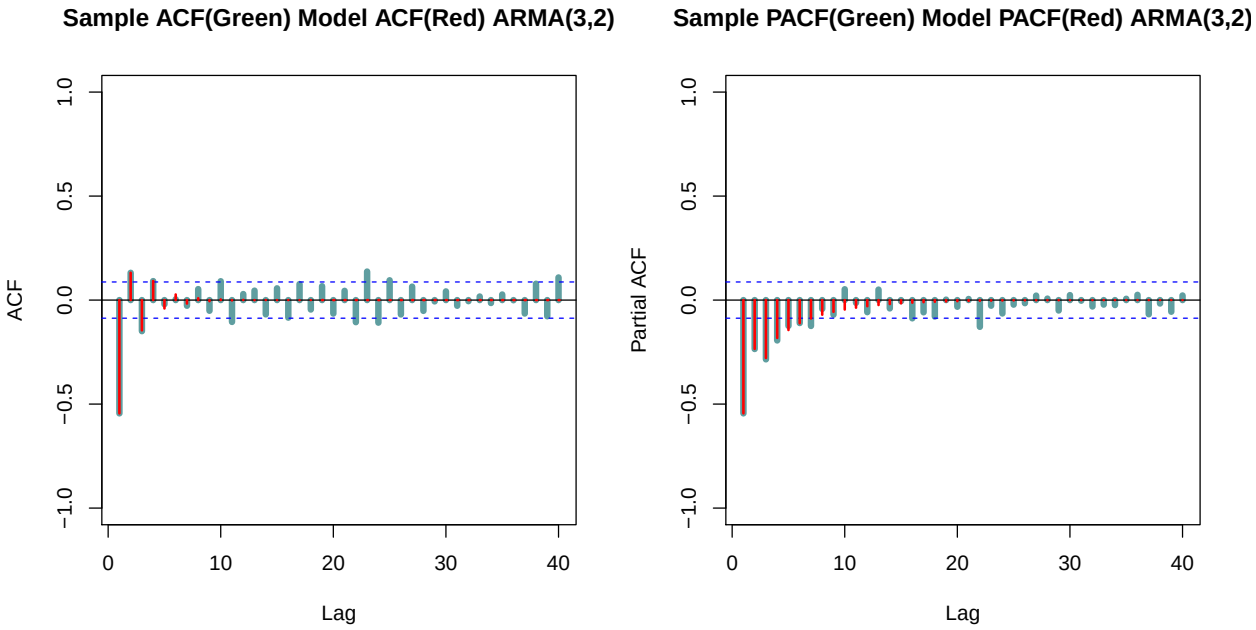
$\hat{\phi}_1 = -0.2504$, $\hat{\phi}_2 = 0.0355$, $\hat{\phi}_3 = -0.0933$

$\hat{\theta}_1 = -0.6079$, $\hat{\theta}_2 = -0.1625$, $\hat{\sigma}^2 = 0.7804$

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

The AICC value is 1321.43

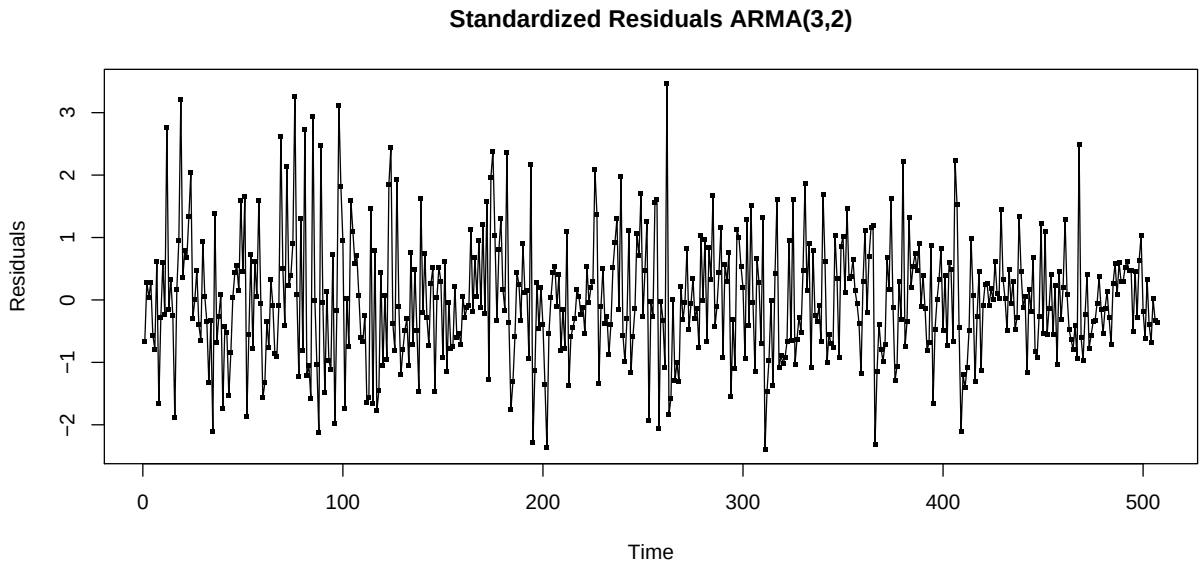
(d) Plot the model and sample acf/pacf values together.



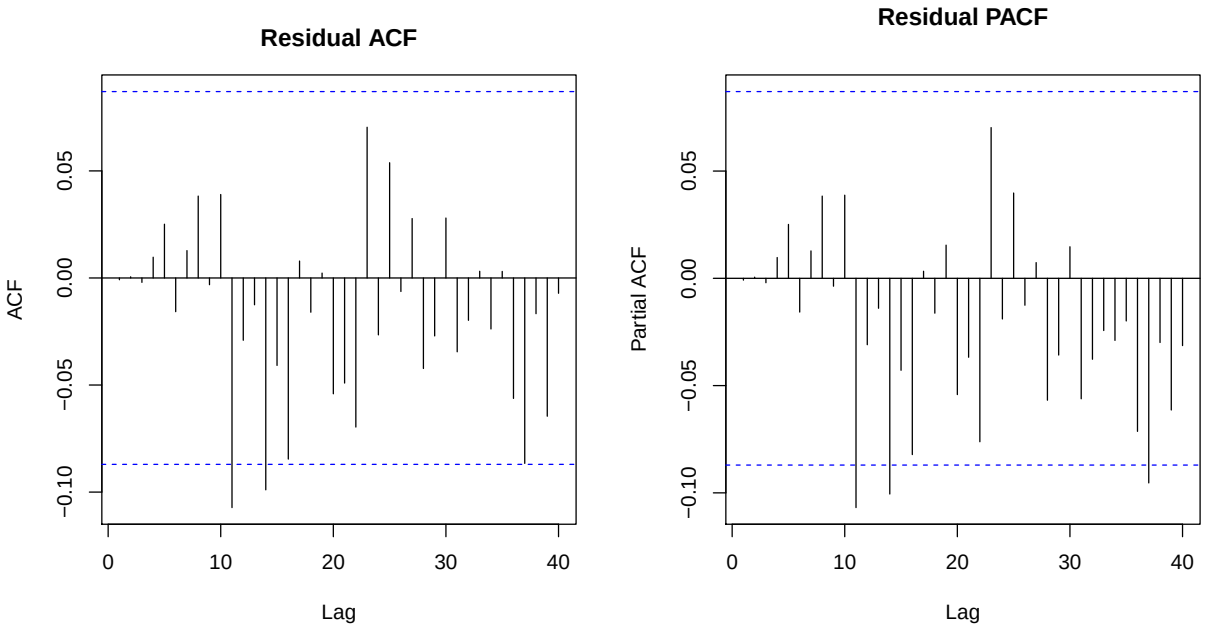
(e) Use the plot from (d) to assess the quality of the model fit.

The model acf fits the sample acf really well up to lag 4, and the model pacf fits the sample pacf really well up to lag 6. The model is a fairly well fit in the plots.

(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.

There is no pattern in the residuals plot, the variance of the residuals seems constant. 2 residual acf are slightly beyond the 95% CI bound and 1 acf is touching the bound. 3 residual pacf are slightly beyond the 95% CI bound. Both are still reasonable numbers for iid noise. The plots suggest the residual might 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.

At lag = 20, the Ljung-Box statistic $Q_{LB} = 20.126$ with p-value 0.45, which fails to reject the iid hypothesis

at level 0.05.
The Mcleod-Li statistic $QML = 51.489$ with p-value 0.0001351, which supports the rejection of the iid hypothesis at level 0.05.

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

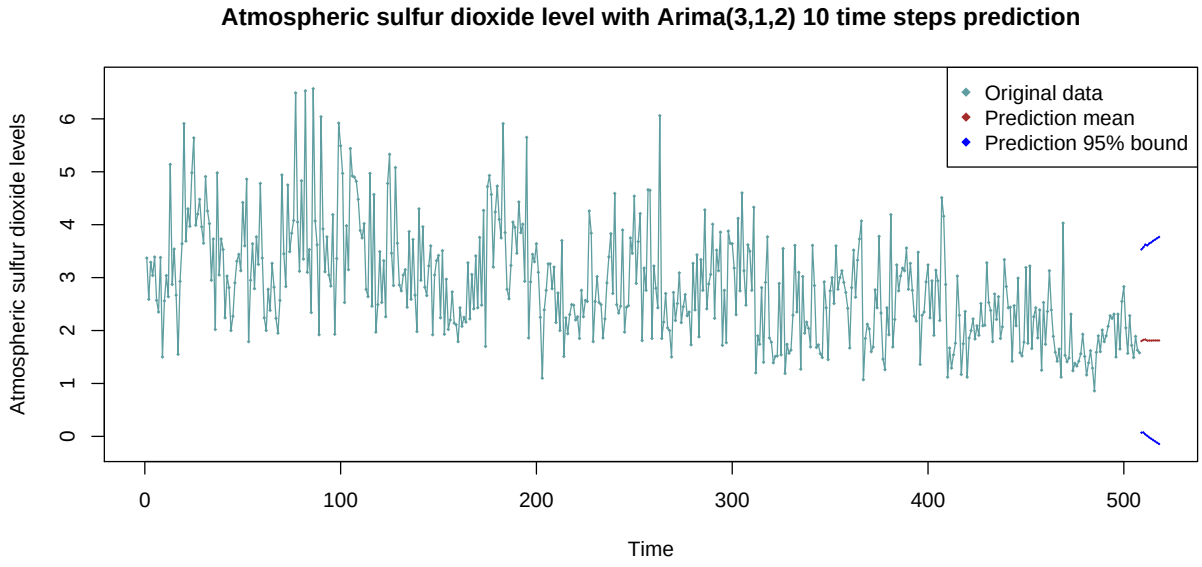
In balance, I conclude the model residuals are likely to behave like iid. The number of siginificant values in the ACF/PACF plots are 2 to 3 each, which is reasonable. There is not strong enough evidence to completely reject the iid hypothesis.

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

The model acf/pacf fits well and are in good agreement with the sample acf/pacf, and the residuals behave like iid noise. Overall, the model is a good fit.

Part 4: Use the estimated model to make a forecast.

(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.



Code used:

```
knitr::opts_chunk$set(echo=FALSE, warning=FALSE, message=FALSE,fig.width=10,fig.height=5)
library(forecast)
library(itsmr)
library(TSA)
data = read.delim("E:/2021 Fall/STAT 485/data/so2.txt", header = FALSE)
# Part 1
plot(data$V1, type="o", pch=16, col="cadetblue", xlab="Time", ylab="Atmospheric sulfur dioxide levels",
     main="Atmospheric sulfur dioxide level")
t = 1:length(data$V1)
cubic = lm (data$V1 ~ t + I(t^2) + I(t^3)) # cubic
cubic$coefficients
y = diff(data$V1, lag=1, differences=1)
y = y-mean(y)
plot(y, type="o", pch=16, col="cadetblue", xlab="Time", ylab="Atmospheric sulfur dioxide level",
     main="Atmospheric sulfur dioxide levels (first difference mean-corrected data)")
par(mar=c(4,4,3,2), mfrow=c(1,2))
acf.sample = acf(y, lag.max = 40, plot = FALSE)
plot(acf.sample, ylim=c(-1,1), main="Sample ACF")
pacf.sample = pacf(y, lag.max = 40, plot = FALSE)
plot(pacf.sample, ylim=c(-1,1), main="Sample PACF")
acf.sample = acf(y, lag.max = 40, plot = FALSE)
pacf.sample = pacf(y, lag.max = 40, plot = FALSE)

# ARMA(2,1)
arma21 = Arima(y, order = c(2,0,1), method = "ML", include.mean = FALSE)
arma21model.acf = ARMAacf(ar=c(arma21$coef[1],arma21$coef[2]),
```

```
                    ma=c(arma21$coef[3]),
                    lag.max = 40, pacf = FALSE)
arma21model.pacf = ARMAacf(ar=c(arma21$coef[1],arma21$coef[2]),
                    ma=c(arma21$coef[3]),
                    lag.max = 40, pacf = TRUE)

par(mar=c(4,4,3,2), mfrow=c(1,2))

plot(acf.sample, ylim=c(-1,1), lwd = 5, col="cadetblue", main="Sample ACF(Green) Model ACF(Red) ARMA(2,
points(arma21model.acf[-1], col="red", type="h", lwd = 2)

plot(pacf.sample, ylim=c(-1,1), lwd = 5, col="cadetblue", main="Sample PACF(Green) Model PACF(Red) ARMA
points(arma21model.pacf, col="red", type="h", lwd = 2)

par(mfrow=c(1,1))
plot(rstandard(arma21), type="o", pch=15, cex=0.5, ylab="Residuals", main="Standardized Residuals ARMA(

autoplot(arma21)
# ARMA(3,2)
arma32 = Arima(y, order = c(3,0,2), method = "ML", include.mean = FALSE)

arma32model.acf = ARMAacf(ar=c(arma32$coef[1],arma32$coef[2],arma32$coef[3]),
                    ma=c(arma32 $coef[4], arma32$coef[5]),
                    lag.max = 40, pacf = FALSE)
arma32model.pacf = ARMAacf(ar=c(arma32$coef[1],arma32$coef[2],arma32$coef[3]),
                    ma=c(arma32 $coef[4], arma32$coef[5]),
                    lag.max = 40, pacf = TRUE)

par(mar=c(4,4,3,2), mfrow=c(1,2))
plot(acf.sample, ylim=c(-1,1), lwd = 5, col="cadetblue", main="Sample ACF(Green) Model ACF(Red) ARMA(3,
points(arma32model.acf[-1], col="red", type="h", lwd = 2)

plot(pacf.sample, ylim=c(-1,1), lwd = 5, col="cadetblue", main="Sample PACF(Green) Model PACF(Red) ARMA
points(arma32model.pacf, col="red", type="h", lwd = 2)

par(mfrow=c(1,1))
plot(rstandard(arma32), type="o", pch=15, cex=0.5, ylab="Residuals", main="Standardized Residuals ARMA(

autoplot(arma32)
# ARMA(3,3)
arma33 = Arima(y, order = c(3,0,3), method = "ML", include.mean = FALSE)
arma33model.acf = ARMAacf(ar=c(arma33$coef[1],arma33$coef[2],arma33$coef[3]),
                    ma=c(arma33$coef[4],arma33$coef[5],arma33$coef[6]),
                    lag.max = 40, pacf = FALSE)
arma33model.pacf = ARMAacf(ar=c(arma33$coef[1],arma33$coef[2],arma33$coef[3]),
                    ma=c(arma33$coef[4],arma33$coef[5],arma33$coef[6]),
                    lag.max = 40, pacf = TRUE)

par(mar=c(4,4,3,2), mfrow=c(1,2))
plot(acf.sample, ylim=c(-1,1), lwd = 5, col="cadetblue", main="Sample ACF(Green) Model ACF(Red) ARMA(3,
points(arma33model.acf[-1], col="red", type="h", lwd = 2)

plot(pacf.sample, ylim=c(-1,1), lwd = 5, col="cadetblue", main="Sample PACF(Green) Model PACF(Red) ARMA
```

```
points(arma33model.pacf, col="red", type="h", lwd = 2)

par(mfrow=c(1,1))
plot(rstandard(arma33), type="o", pch=15, cex=0.5, ylab="Residuals", main="Standardized Residuals ARMA(3

autoplot(arma33)
# ARMA(4,4)
arma44 = Arima(y, order = c(4,0,4), method = "ML", include.mean = FALSE)
arma44model.acf = ARMAacf(ar=c(arma44$coef[1],arma44$coef[2],arma44$coef[3],arma44$coef[4]),
                      ma=c(arma44$coef[5],arma44$coef[6],arma44$coef[7],arma44$coef[8]),
                      lag.max = 40, pacf = FALSE)
arma44model.pacf = ARMAacf(ar=c(arma44$coef[1],arma44$coef[2],arma44$coef[3],arma44$coef[4]),
                      ma=c(arma44$coef[5],arma44$coef[6],arma44$coef[7],arma44$coef[8]),
                      lag.max = 40, pacf = TRUE)

par(mar=c(4,4,3,2), mfrow=c(1,2))
plot(acf.sample, ylim=c(-1,1), lwd = 5, col="cadetblue", main="Sample ACF(Green) Model ACF(Red) ARMA(4,4
points(arma44model.acf[-1], col="red", type="h", lwd = 2)

plot(pacf.sample, ylim=c(-1,1), lwd = 5, col="cadetblue", main="Sample PACF(Green) Model PACF(Red) ARMA
points(arma44model.pacf, col="red", type="h", lwd = 2)

par(mfrow=c(1,1))
plot(rstandard(arma44), type="o", pch=15, cex=0.5, ylab="Residuals", main="Standardized Residuals ARMA(4

autoplot(arma44)
# AICC
Model = c('ARMA(2,1)', 'ARMA(3,2)', 'ARMA(3,3)', 'ARMA(4,4)')
AICC = c(arma21$aicc, arma32$aicc, arma33$aicc, arma44$aicc)
table = data.frame(Model, AICC)
knitr::kable(table)
arma32
par(mar=c(4,4,3,2), mfrow=c(1,2))
plot(acf.sample, ylim=c(-1,1), lwd = 5, col="cadetblue", main="Sample ACF(Green) Model ACF(Red) ARMA(3,1
points(arma32model.acf[-1], col="red", type="h", lwd = 2)
plot(pacf.sample, ylim=c(-1,1), lwd = 5, col="cadetblue", main="Sample PACF(Green) Model PACF(Red) ARMA
points(arma32model.pacf, col="red", type="h", lwd = 2)
resid = rstandard(arma32)
plot(resid, type="o", pch=15, cex=0.5, ylab="Residuals", main="Standardized Residuals ARMA(3,2)")
par(mar=c(4,4,3,2), mfrow=c(1,2))
acf(resid, lag.max = 40, plot = TRUE, main="Residual ACF")
pacf(resid, lag.max = 40, plot = TRUE, main="Residual PACF")
# Ljung-Box test
Box.test(resid, type = "Ljung", lag = 20)
# McLeod-Li test
Box.test(resid^2, type = "Ljung", lag = 20)
arima312 = Arima(data$V1, order = c(3,1,2), method = "ML")
future.pred = forecast::forecast(arima312, h = 10)
plot.ts(future.pred$x, type = "o", pch = 18, col = "cadetblue", xlim = c(0, length(y)+10), cex = 0.4,
        ylab ="Atmospheric sulfur dioxide levels", ylim = c(-0.2, 6.7),
        main="Atmospheric sulfur dioxide level with Arima(3,1,2) 10 time steps prediction")
points(future.pred$mean, type = "o", pch = 18, col = "brown", cex = 0.3)
points(future.pred$upper[,2], type = "o", pch = 18, col = "blue", cex = 0.3)
```

```
points(future.pred$lower[,2], type = "o", pch = 18, col = "blue", cex = 0.3)
legend("topright", legend = c("Original data", "Prediction mean", "Prediction 95% bound"),
       pch = c(18, 18, 18), col = c("cadetblue", "brown", "blue"))
```