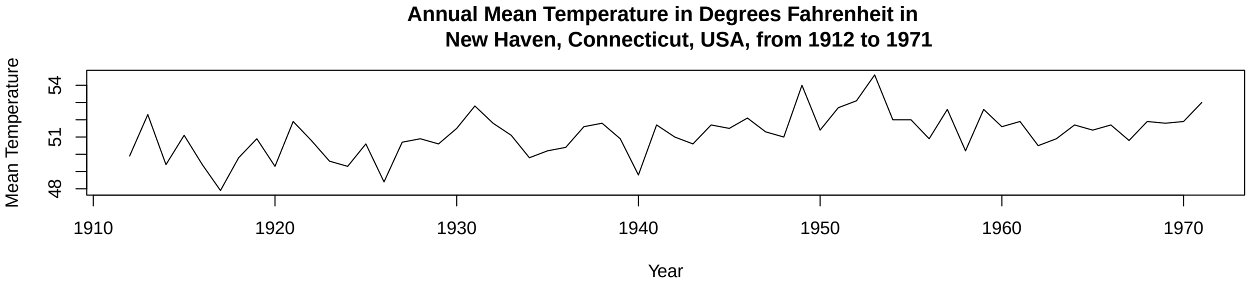
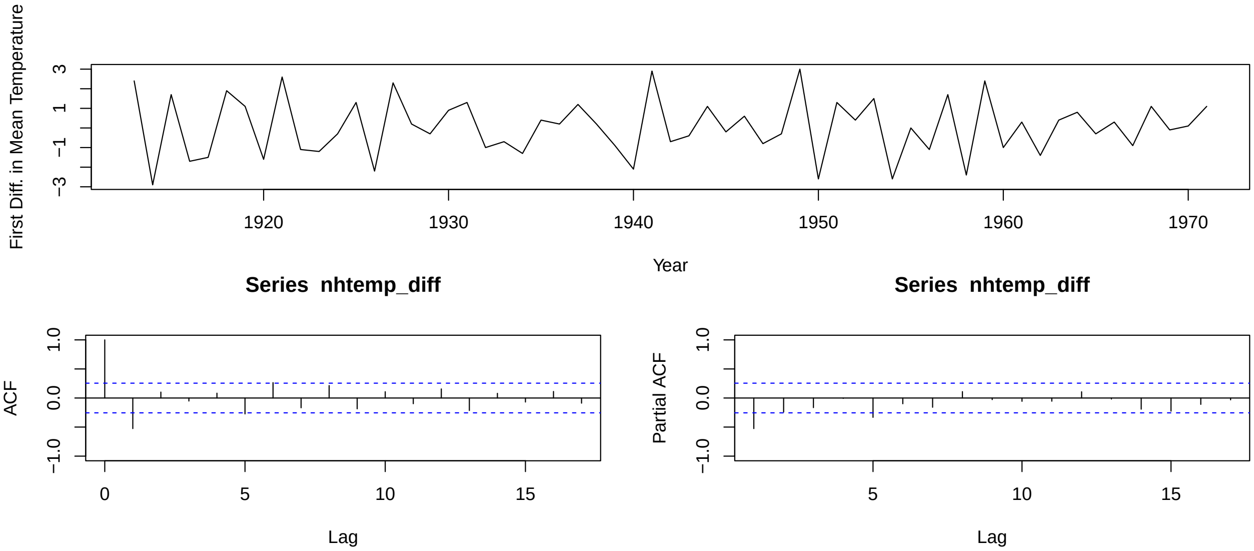
20129999 MATH3026 CW2

a) nhtemp dataset

This is the dataset regarding to a time series of the annual mean temperature in degrees Fahrenheit in New Haven, Connecticut USA from 1912 to 1971. We first plot this time series to gain a general view of it.



As we can see form the plot, there was a possible upward trend in mean temperature from 1912 to 1971 but no clear seasonality. As a result, we apply the first difference to the time series to remove the trend.



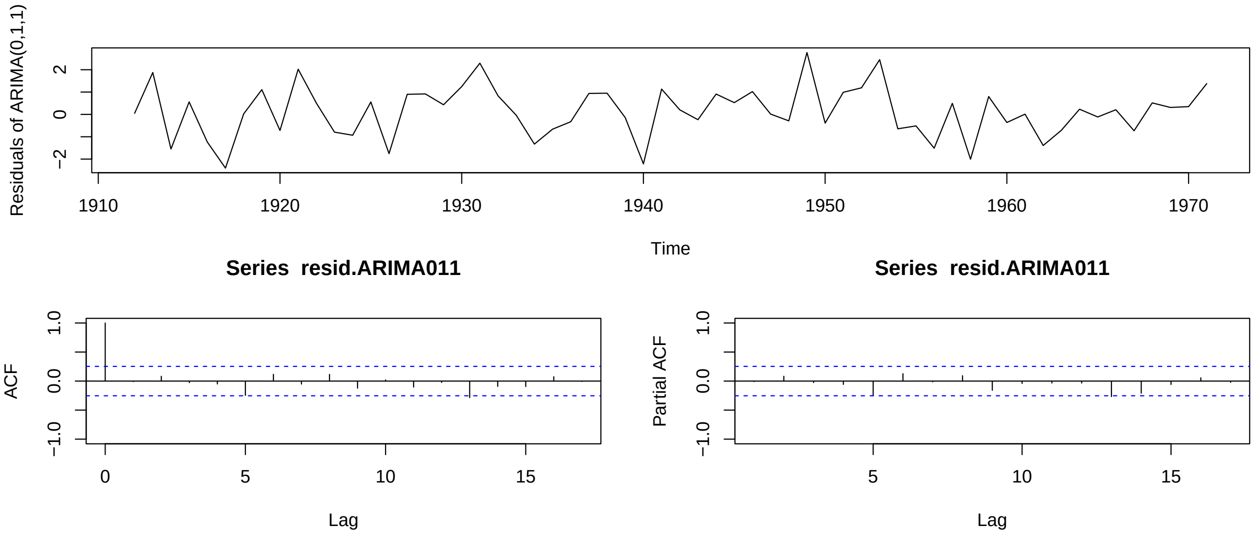
Generally speaking, the differenced data appears stationary. The mean does not change over time and the variability of the process is fairly constant. Looking at the sample ACF and PACF plot, both of them decay to zero at small lags, lag 1. Thus, the data seem reasonable to fit an ARIMA model.

We begin with an ARIMA(0,1,1) model. We have the fitted model

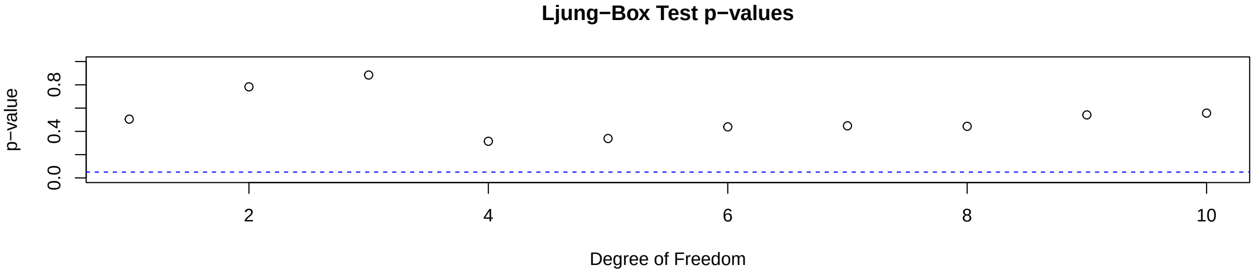
$$X_t - X_{t-1} = Z_t - 0.7983Z_{t-1}$$

where Z_t is the white noise process with mean zero and variance 1.291. This model has AIC value of 187.52.

We then check if the residuals of the model appears to be the white noise process or not.



We notice that the sample ACF of the residuals cut off after lag 0 which fits the feature of white noise process. We use the Ljung-Box test to verify whether the residuals are independent or not.



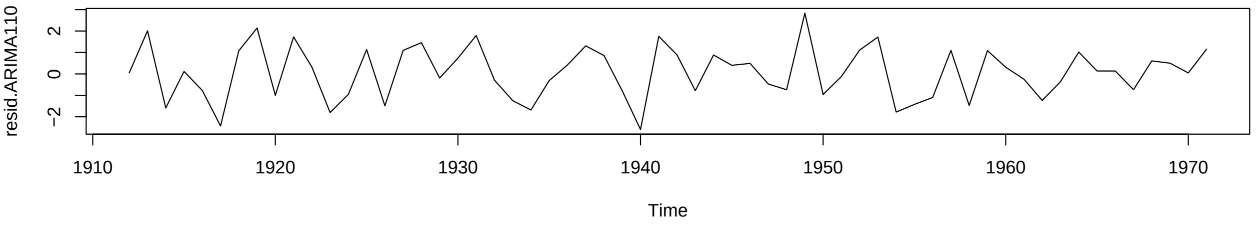
It is clear to see from the plot that the p-values are above the blue line which presents 0.05. As the p-values are large, we do not reject the null hypothesis that the residuals are independent. Therefore, an ARIMA(0,1,1) model may be a candidate for fitting the dataset.

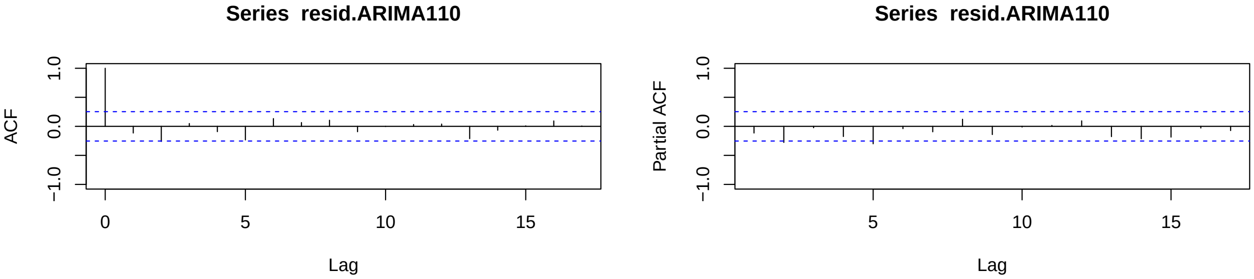
Next, we fit an ARIMA(1,1,0) model. We have the fitted model

$$X_t - 0.4533X_{t-1} + 0.5467X_{t-2} = Z_t$$

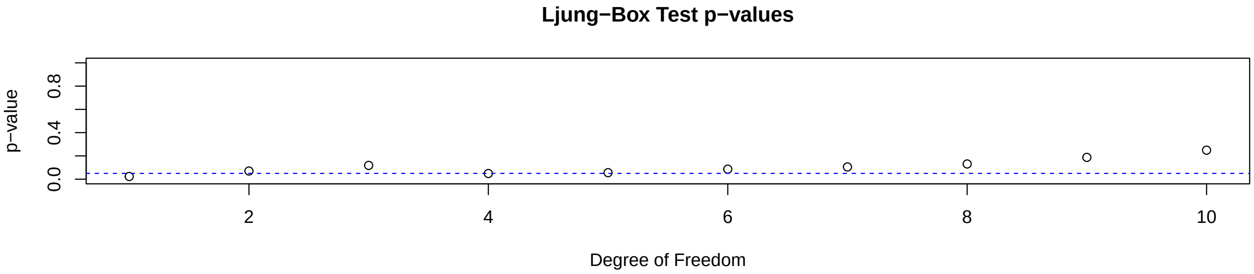
where Z_t is the white noise process with mean zero and variance 1.494. This model has AIC value of 195.46.

Again, we check if the residuals of the model appears to be the white noise process or not.





Looking at the sample ACF of the residuals, it cuts off after lag 0 which may present the residuals are the white noise process. We shall use the Ljung-Box test to verify the independence.



While the degree of freedom is one, the p-value is less than 0.05 and this suggests us to reject the null hypothesis that the process is independent. Thus, the residuals of this ARIMA(1,1,0) model do not appear completely to be the white noise process based on the result of the Ljung-Box test.

Next, we fit an ARIMA(1,1,1) model as a comparison with the ARIMA(0,1,1) model. We have the fitted model

$$X_t - 1.0073X_t + 0.0073X_{t-2} = Z_t - 0.8019Z_{t-1}$$

where Z_t is the white noise process with mean zero and variance 1.291. This model has AIC value of 189.52.

Under ARIMA(1,1,1) model, the fitted value of the autoregressive component $\hat{\phi}_1$ is 0.0073 which has a standard error of 0.1802. The approximate 95% confidence interval for ϕ_1 is $[-0.3459, 0.3605]$ which contains zero so we have an evidence at 5% level saying that ϕ_1 is not significant to the model. This result leads us to the previous ARIMA(0,1,1) model we had.

Finally, we fit an ARIMA(0,1,2) model as a comparison with the ARIMA(0,1,1) model. We have the fitted model

$$X_t - X_{t-1} = Z_t - 0.7956Z_{t-1} - 0.0042Z_{t-2}$$

where Z_t is the white noise process with mean zero and variance 1.291. This model has AIC value of 189.52.

Under this ARIMA(0,1,2) model, the fitted value of the second moving average component $\hat{\theta}_2$ is -0.0042 which has a standard error of 0.1221. The approximate 95% confidence interval for θ_2 is $[-0.2435, 0.2351]$ which contains zero so we have an evidence at 5% level saying that θ_2 is not significant to the model. This result shows that an additional moving average component is redundant.

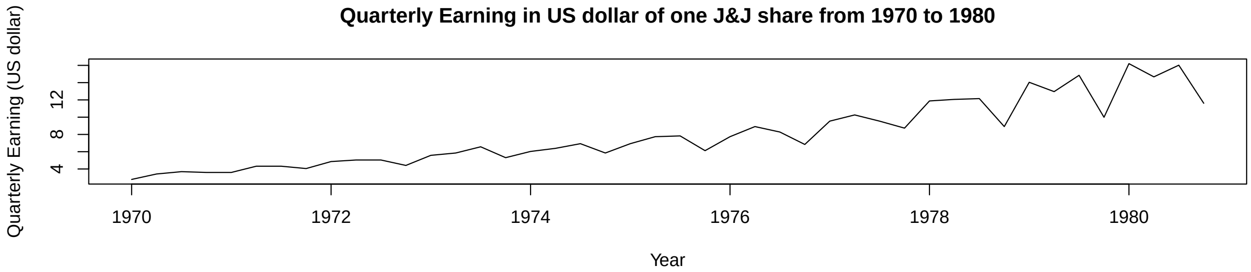
In sum, the ARIMA(0,1,1) model

$$X_t - X_{t-1} = Z_t - 0.7983Z_{t-1}$$

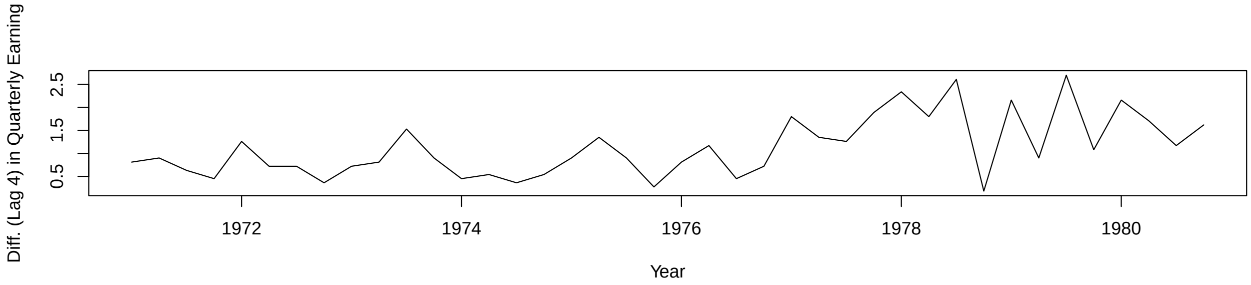
where Z_t is the white noise process with mean zero and variance 1.291 is the best choice to fit the nhtemp dataset among these four models we have. The reasoning is as following: first, the ARIMA(0,1,1) model has lowest value of AIC, 187.52; second, about the ARIMA(1,1,0) model, we have proven that the residuals of this model is not quite like the white noise process based on the Ljung-Box test; finally, we have shown that the ARIMA(1,1,1) and ARIMA(0,1,2) models are overfitting by the hypothesis test for parameters.

b) JJ dataset

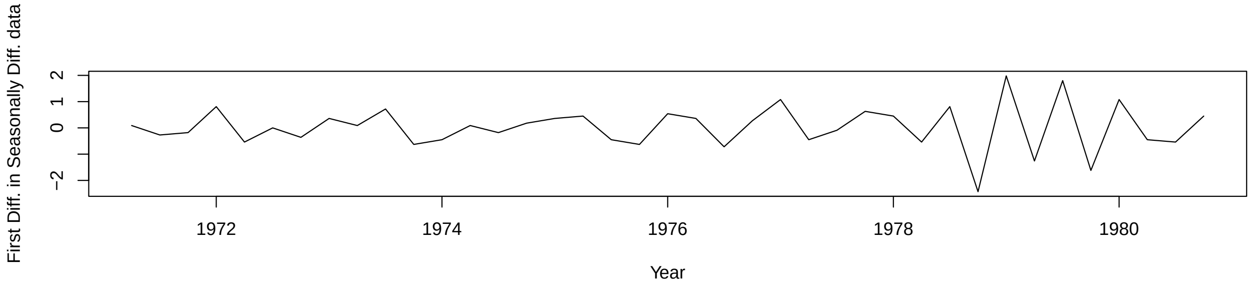
This is the dataset regarding to a time series of the quarterly earnings in US dollars of one Johnson & Johnson share from 1970 to 1980. To begin with, we plot the observations to see the features of the time series.

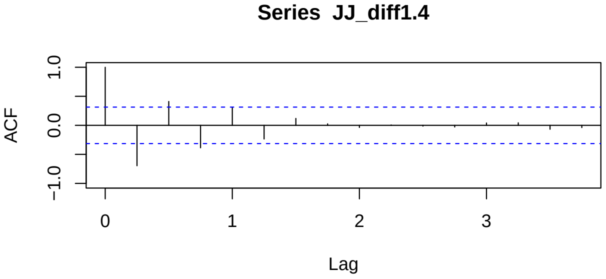


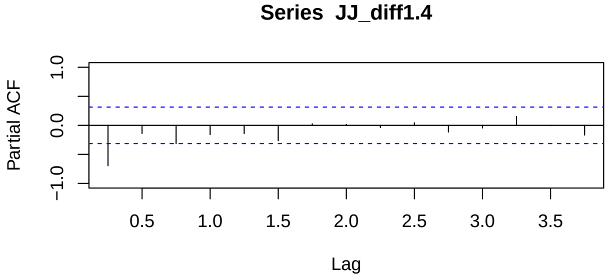
As shown, the time series is non-stationary and exhibits seasonality. Intuitively, we would difference the data with lag 4 to remove the seasonal travel trends over four quarters.



Although the seasonality appears to have been removed, this time series is still not stationary. We further compute the first difference of the seasonally differenced Quarterly Earning of One J&J Share data.







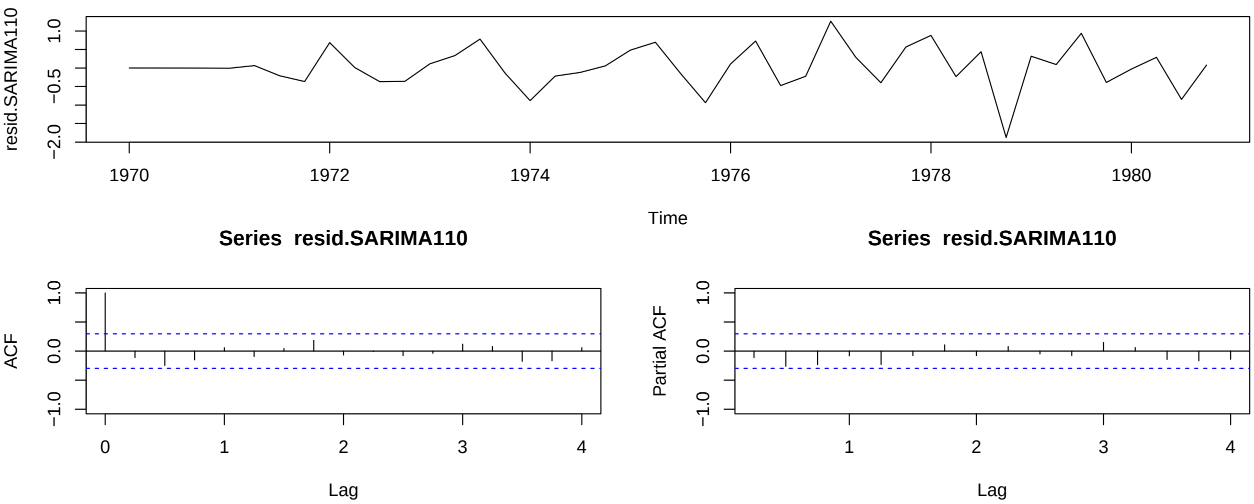
These data appear to be stationary. The above analysis suggests us to consider fitting an ARMA model on the first difference of the seasonally differenced observed data. Particularly, we see that the sample PACF cuts off after lag 1 and the sample ACF tails off so we would focus on fitting an ARMA(1,0) model and an ARMA(1,1) model.

First, we fit a SARIMA$(1,1,0) \times (0,1,0)_4$. We have the fitted model

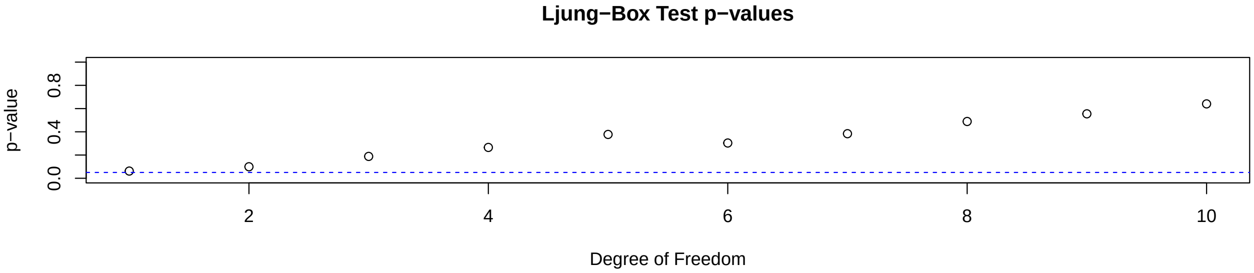
$$X_t - 0.3162X_{t-1} - 0.6838X_{t-2} - X_{t-4} + 0.3162X_{t-5} + 0.6838X_{t-6} = Z_t$$

where Z_t is the white noise process with mean zero and variance 0.349. This model has AIC value of 74.25.

We then check if its residuals appears to be the white noise process or not.



From the above plots, we see that the ACF of residuals cuts off after lag 0 which fits the feature of white noise process. We use the Ljung-Box test to verify whether the residuals are independent or not.



The p-values are all above the blue line which presents 0.05. This suggests that we do not reject the null hypothesis that the residuals are independent. As a result, this SARIMA$(1,1,0) \times (0,1,0)_4$ model could be a candidate for fitting the dataset.

Next, we fit a SARIMA$(2,1,0) \times (0,1,0)_4$ model as a comparison with the SARIMA$(1,1,0) \times (0,1,0)_4$ model. We have

$$X_t - 0.2255X_{t-1} - 0.6461X_{t-2} - 0.1284X_{t-3} - X_{t-4} + 0.2255X_{t-5} + 0.6461X_{t-6} + 0.1284X_{t-7} = Z_t$$

where Z_t is the white noise process with mean zero and variance 0.3426. This model has AIC value of 75.58.

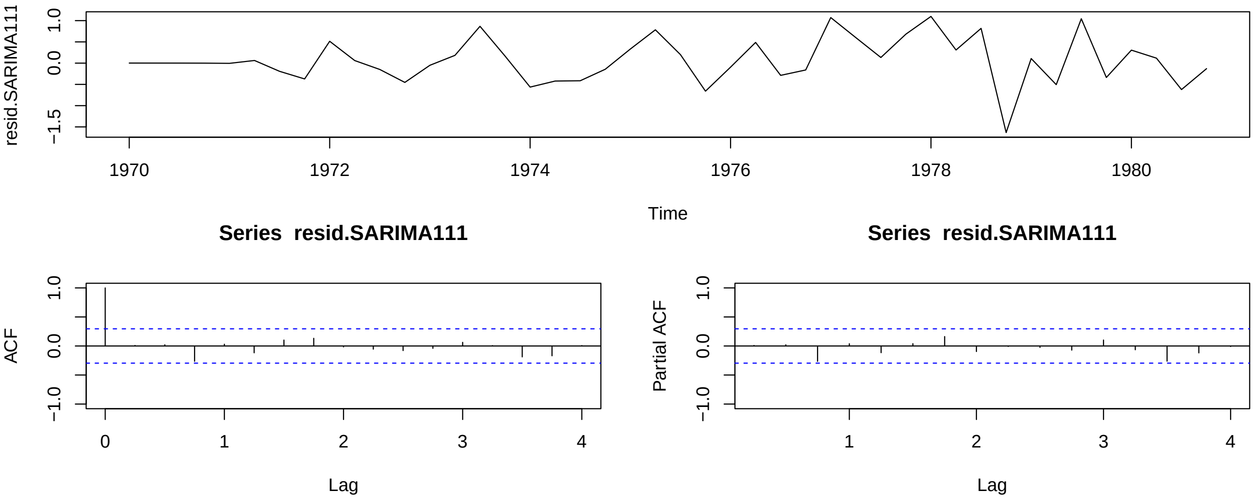
Under the SARIMA$(2,1,0)\times(0,1,0)_4$ model, the fitted value of the secondly non-seasonal autoregressive component $\hat{\phi}_2$ is -0.1284 which has a standard error of 0.1559. The approximate 95% confidence interval for ϕ_2 is $[-0.4340, 0.1772]$ which contains zero so we have an evidence at 5% level saying that ϕ_2 is not significant to the model.

Next, we fit a SARIMA$(1,1,1)\times(0,1,0)_4$ model. We have the fitted model

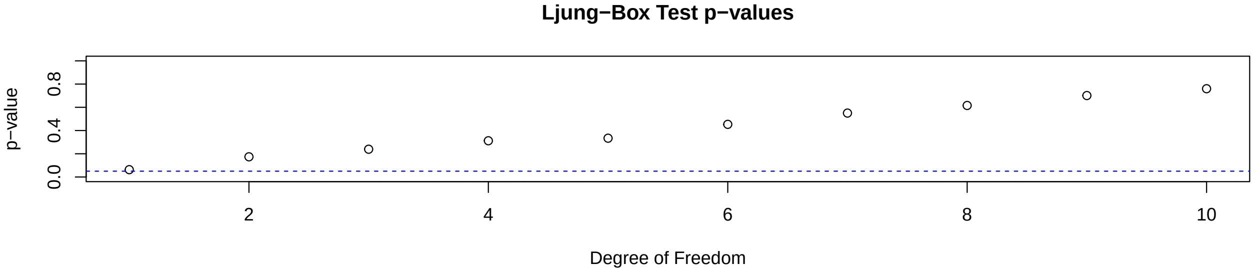
$$X_t - 0.6535X_{t-1} - 0.3465X_{t-2} - X_{t-4} + 0.6535X_{t-5} + 0.3465X_{t-6} = Z_t - 0.6308Z_{t-1}$$

where Z_t is the white noise process with mean zero and variance 0.3166. This model has AIC value of 72.86.

Again, we check if its residuals appears to be the white noise process or not.



According to the plots, the residuals of this model appear fairly like the white noise process. We would use the Ljung-Box test to verify the independence.



It is clear to see from the plot that the p-values are above the blue line which presents 0.05. As the p-values are large, we do not reject the null hypothesis that the residuals are independent. Therefore, a SARIMA$(1,1,1)\times(0,1,0)_4$ model may also be a choice for fitting the dataset.

Next, we fit a SARIMA$(2,1,1)\times(0,1,0)_4$ model as a comparison with the SARIMA$(1,1,1)\times(0,1,0)_4$ model. We have the fitted model

$$X_t - 0.8384X_{t-1} - 0.4005X_{t-2} + 0.2389X_{t-3} - X_{t-4} + 0.8384X_{t-5} - 0.4005X_{t-6} - 0.2389X_{t-7} = Z_t - 0.775Z_{t-1}$$

where Z_t is the white noise process with mean zero and variance 0.3064. This model has AIC value of 73.59.

Under the SARIMA$(2,1,1) \times (0,1,0)_4$ model, the fitted value of the secondly non-seasonal autoregressive component $\hat{\phi}_2$ is 0.2389 which has a standard error of 0.2173. The approximate 95% confidence interval for ϕ_2 is $[-0.1870, 0.6648]$ which contains zero so we have an evidence at 5% level saying that ϕ_2 is not significant to the model.

Finally, we fit a SARIMA$(1,1,2) \times (0,1,0)_4$ model as a comparison with the SARIMA$(1,1,1) \times (0,1,0)_4$ model. We have the fitted model

$$X_t - 0.2011X_{t-1} - 0.7989X_{t-2} - X_{t-4} + 0.2011X_{t-5} + 0.7989X_{t-6} = Z_t - 0.1320Z_{t-1} - 0.4026Z_{t-3}$$

where Z_t is the white noise process with mean zero and variance 0.3004. This model has AIC value of 72.86.

Under the SARIMA$(1,1,2) \times (0,1,0)_4$ model, the fitted value of the secondly non-seasonal moving average component $\hat{\theta}_2$ is -0.4026 which has a standard error of 0.2136. The approximate 95% confidence interval for θ_2 is $[-0.8213, 0.0161]$ which contains zero so we have an evidence at 5% level saying that θ_2 is not significant to the model.

In sum, the SARIMA$(1,1,1) \times (0,1,0)_4$ model

$$X_t - 0.6535X_{t-1} - 0.3465X_{t-2} - X_{t-4} + 0.6535X_{t-5} + 0.3465X_{t-6} = Z_t - 0.6308Z_{t-1}$$

where Z_t is the white noise process with mean zero and variance 0.3166 is the best choice to fit the JJ dataset among these models we have. The reasoning is as following: first, the SARIMA$(1,1,1) \times (0,1,0)_4$ model has a lower value of AIC, 72.86, than the SARIMA$(1,1,0) \times (0,1,0)_4$ model has, 74.25; about the SARIMA$(2,1,1) \times (0,1,0)_4$ model and the SARIMA$(1,1,2) \times (0,1,0)_4$ model, we have proven that these models are overfitted by the hypothesis test for the parameter.

The discussion ends here and the following part is the code as an appendix.

```r
# read data "nhtemp"
load("nhtemp.rda")

# plot the data
ts.plot(nhtemp,
        xlab = "Year",
        ylab = "Mean Temperature",
        main = "Annual Mean Temperature in Degrees Fahrenheit in
        New Haven, Connecticut, USA, from 1912 to 1971")
```

```r
# try diff_1
nhtemp_diff <- diff(nhtemp)
ts.plot(nhtemp_diff,
        xlab = "Year",
        ylab = "First Difference in Mean Temperature")

# set the R graphics device to contain two plots (1 row, 2 columns)
par(mfrow = c(1,2))
# plot the sample ACF based on nhtemp_diff
acf(nhtemp_diff, ylim = c(-1,1))
# plot the sample PACF based on nhtemp_diff
pacf(nhtemp_diff, ylim = c(-1,1))
```

```r
# fit an ARIMA(0,1,1) model to the nhtemp dataset
ARIMA011 <- arima(nhtemp, order = c(0,1,1), method = "ML")
ARIMA011
```

```r
# residuals of this ARIMA(0,1,1) model
resid.ARIMA011 <- residuals(ARIMA011)
# plot the residuals
ts.plot(resid.ARIMA011, ylab = "Residuals of ARIMA(0,1,1)")
# set the R graphics device to contain two plots (1 row, 2 columns)
par(mfrow = c(1,2))
# ACF plot of this ARIMA(0,1,1) model
acf(resid.ARIMA011, ylim = c(-1,1))
# PACF plot of this ARIMA(0,1,1) model
pacf(resid.ARIMA011, ylim = c(-1,1))
```

```r
#Function to produce P-values for the Ljung-Box test for different lags
#where an ARMA(p,q) model has been fitted.
#Note that k must be > p+q (See Lecture 9 slides)
#Number of degrees of freedom for the test = k-p-q

#Arguments for the function "LB_test"
#resid = residuals from a fitted ARMA(p,q) model.

#max.k = the maximum value of k at which we perform the test
#Note that the minimum k is set at p+q+1 (corresponding to a test with one degree
#of freedom)

#p = Order of the AR part of the model
#q = Order of the MA part of the model

#The function returns a table with one column showing the number of degrees
#of freedom for the test and the other the associated P-value.

LB_test<-function(resid,max.k,p,q){
 lb_result<-list()
 df<-list()
 p_value<-list()
  for(i in (p+q+1):max.k){
   lb_result[[i]]<-Box.test(resid,lag=i,type=c("Ljung-Box"),fitdf=(p+q))
   df[[i]]<-lb_result[[i]]$parameter
   p_value[[i]]<-lb_result[[i]]$p.value
  }
 df<-as.vector(unlist(df))
 p_value<-as.vector(unlist(p_value))
 test_output<-data.frame(df,p_value)
 names(test_output)<-c("deg_freedom","LB_p_value")
 return(test_output)
 }
```

```r
# Ljung-Box test
ARIMA011.LB <- LB_test(resid.ARIMA011, max.k = 11, p = 0, q = 1)
# plot of the p-values against degree of freedom
# add a blue line at 0.05
plot(ARIMA011.LB$deg_freedom, ARIMA011.LB$LB_p_value,
     xlab = "Degree of Freedom",
     ylab = "p-value",
     main = "Ljung-Box Test p-values",
     ylim = c(0,1))
abline(h = 0.05, col = "blue", lty = 2)
```

```
# fit an ARIMA(1,1,0) model to the nhtemp dataset
ARIMA110 <- arima(nhtemp, order = c(1,1,0), method = "ML")
ARIMA110
```

```
# residuals of this ARIMA(1,1,0) model
resid.ARIMA110 <- residuals(ARIMA110)
# plot the residuals
ts.plot(resid.ARIMA110)
# set the R graphics device to contain two plots (1 row, 2 columns)
par(mfrow = c(1,2))
# ACF plot of this ARIMA(1,1,0) model
acf(resid.ARIMA110, ylim = c(-1,1))
# PACF plot of this ARIMA(1,1,0) model
pacf(resid.ARIMA110, ylim = c(-1,1))
```

```
# Ljung-Box test
ARIMA110.LB <- LB_test(resid.ARIMA110, max.k = 11, p = 1, q = 0)
# plot of the p-values against degree of freedom
# add a blue line at 0.05
plot(ARIMA110.LB$deg_freedom, ARIMA110.LB$LB_p_value,
     xlab = "Degree of Freedom",
     ylab = "p-value",
     main = "Ljung-Box Test p-values",
     ylim = c(0,1))
abline(h = 0.05, col = "blue", lty = 2)
```

```
# fit an ARIMA(1,1,1) model to the nhtemp dataset
ARIMA111 <- arima(nhtemp, order = c(1,1,1), method = "ML")
ARIMA111
```

```
# fit an ARIMA(0,1,2) model to the nhtemp dataset
ARIMA012 <- arima(nhtemp, order = c(0,1,2), method = "ML")
ARIMA012
```

```
# read data "JJ_data"
load("JJ_data.rda")
JJ <- JJ_data

# plot the data
ts.plot(JJ,
        xlab = "Year",
        ylab = "Quarterly Earning (US dollar)",
        main = "Quarterly Earning in US dollar of one J&J share from 1970 to 1980")
```

```
# difference with lag 4 to remove seasonality
JJ_diff4 <- diff(JJ, lag = 4)
ts.plot(JJ_diff4,
        xlab = "Year",
        ylab = "Diff. (Lag 4) in Quarterly Earning")
```

```r
# first difference of the seasonally differenced data
JJ_diff1.4 <- diff(JJ_diff4)
ts.plot(JJ_diff1.4,
        xlab = "Year",
        ylab = "First Diff. in Seasonally Diff. data")
# set the R graphics device to contain two plots (1 row, 2 columns)
par(mfrow = c(1,2))
# plot the sample ACF based on JJ_diff1.4
acf(JJ_diff1.4, ylim = c(-1,1))
# plot the sample PACF based on JJ_diff1.4
pacf(JJ_diff1.4, ylim = c(-1,1))
```

```r
# fit a SARIMA(1,1,0)(0,1,0)_4 model to the JJ dataset
SARIMA110 <- arima(JJ, order = c(1,1,0),
                   seasonal = list(order = c(0,1,0), period = 4))
SARIMA110
```

```r
# residuals of this SARIMA(1,1,0)(0,1,0)_4 model
resid.SARIMA110 <- residuals(SARIMA110)
# plot the residuals
ts.plot(resid.SARIMA110)
# set the R graphics device to contain two plots (1 row, 2 columns)
par(mfrow = c(1,2))
# ACF plot of this SARIMA(1,1,0)(0,1,0)_4 model
acf(resid.SARIMA110, ylim = c(-1,1))
# PACF plot of this SARIMA(1,1,0)(0,1,0)_4 model
pacf(resid.SARIMA110, ylim = c(-1,1))
```

```r
# Ljung-Box test
SARIMA110.LB <- LB_test(resid.SARIMA110, max.k = 11, p = 1, q = 0)
# plot of the p-values against degree of freedom
# add a blue line at 0.05
plot(SARIMA110.LB$deg_freedom, SARIMA110.LB$LB_p_value,
     xlab = "Degree of Freedom",
     ylab = "p-value",
     main = "Ljung-Box Test p-values",
     ylim = c(0,1))
abline(h = 0.05, col = "blue", lty = 2)
```

```r
# fit a SARIMA(2,1,0)(0,1,0)_4 model to the JJ dataset
SARIMA210 <- arima(JJ, order = c(2,1,0),
                   seasonal = list(order = c(0,1,0), period = 4))
SARIMA210
```

```r
# fit a SARIMA(1,1,1)(0,1,0)_4 model to the JJ dataset
SARIMA111 <- arima(JJ, order = c(1,1,1),
                   seasonal = list(order = c(0,1,0), period = 4))
SARIMA111
```

```r
# residuals of this SARIMA(1,1,1)(0,1,0)_4 model
resid.SARIMA111 <- residuals(SARIMA111)
# plot the residuals
ts.plot(resid.SARIMA111)
# set the R graphics device to contain two plots (1 row, 2 columns)
par(mfrow = c(1,2))
```

```
# ACF plot of this SARIMA(1,1,1)(0,1,0)_4 model
acf(resid.SARIMA111, ylim = c(-1,1))
# PACF plot of this SARIMA(1,1,1)(0,1,0)_4 model
pacf(resid.SARIMA111, ylim = c(-1,1))
```

```
# Ljung-Box test
SARIMA111.LB <- LB_test(resid.SARIMA111, max.k = 12, p = 1, q = 1)
# plot of the p-values against degree of freedom
# add a blue line at 0.05
plot(SARIMA111.LB$deg_freedom, SARIMA111.LB$LB_p_value,
     xlab = "Degree of Freedom",
     ylab = "p-value",
     main = "Ljung-Box Test p-values",
     ylim = c(0,1))
abline(h = 0.05, col = "blue", lty = 2)
```

```
# fit a SARIMA(2,1,1)(0,1,0)_4 model to the JJ dataset
SARIMA211 <- arima(JJ, order = c(2,1,1),
                   seasonal = list(order = c(0,1,0), period = 4))
SARIMA211
```

```
# fit a SARIMA(1,1,2)(0,1,0)_4 model to the JJ dataset
SARIMA112 <- arima(JJ, order = c(1,1,2),
                   seasonal = list(order = c(0,1,0), period = 4))
SARIMA112
```