CW2

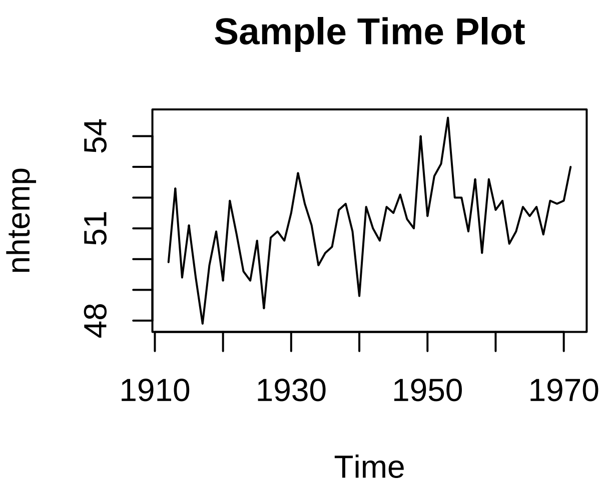
Liangxiao LI

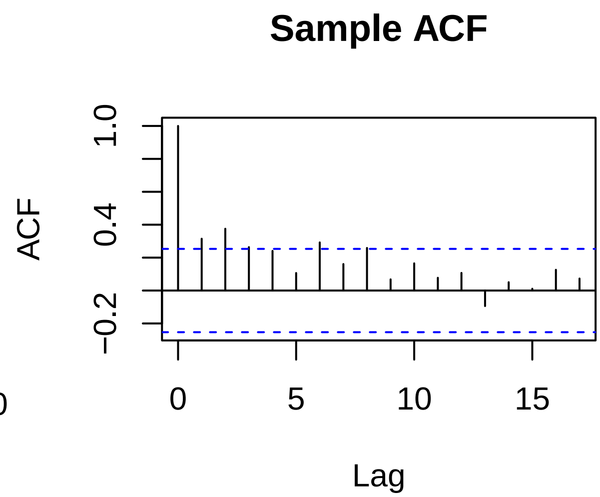
2024-04-10

Q1: nhtemp

Part1: Check Stationarity and Seasonality

First we produce the time plot and ACF plot from the given data:

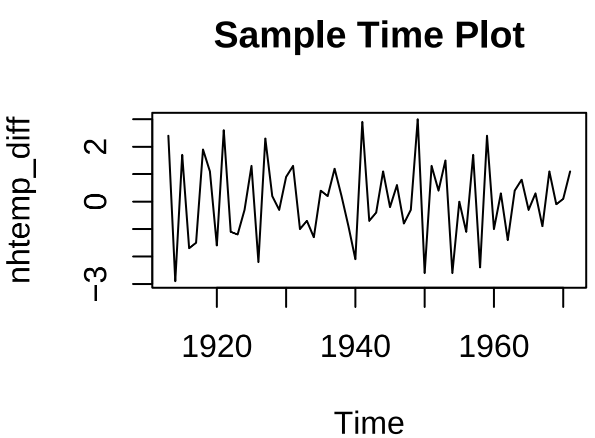


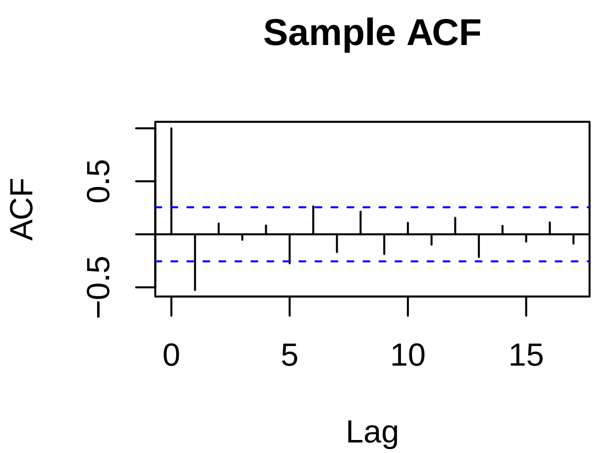


From the plots above, we conclude that the series is non-stationary and non-seasonal due to following reasons:

1) Time plot: the mean of the series appears higher between 1940-1970 to the period between 1910-1940.
2) Sample ACF plot: doesn't decline rapidly, therefore it's not stationary.

To remove non-stationarity, we take the first difference of the time series nhtemp as nhtemp_diff and plot again:





From the plots above, we conclude that the series is stationary without seasonality due to following reasons:

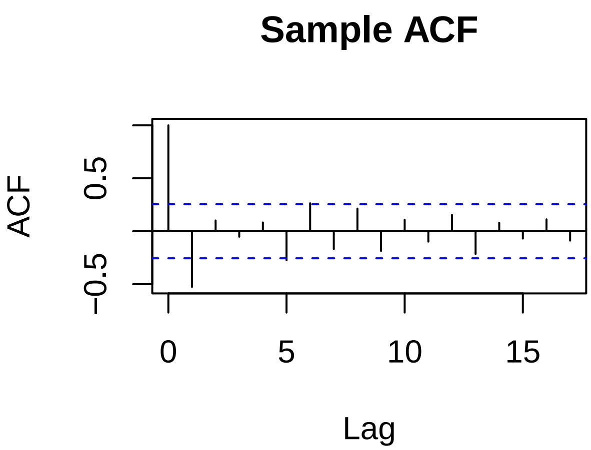
1) Time plot: has a mean equal to zero and shows constant variability over time.
2) Sample ACF plot: declines rapidly to zero as the lag increases, cut off after lag 1

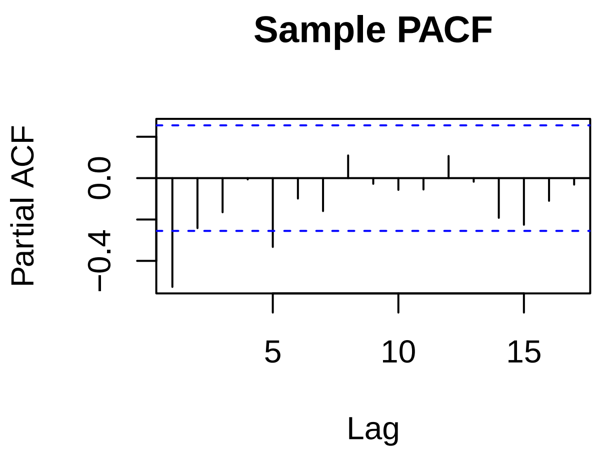
In conclusion, we'll explore models fitted to the data nhtemp_diff which has been differenced once.

Part2: Model fitting

Parameter analysis

The analysis begins by analyzing the sample ACF and PACF plot for nhtemp_diff.

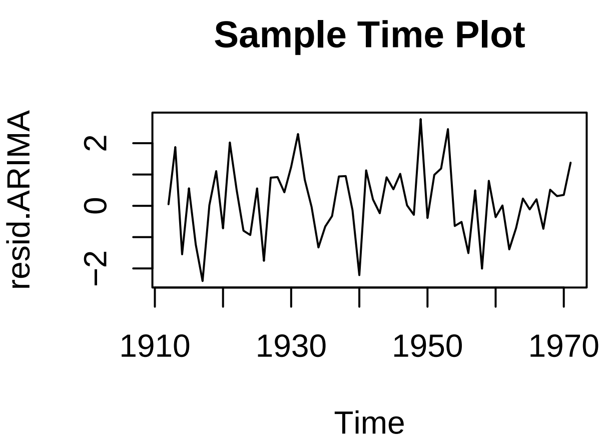


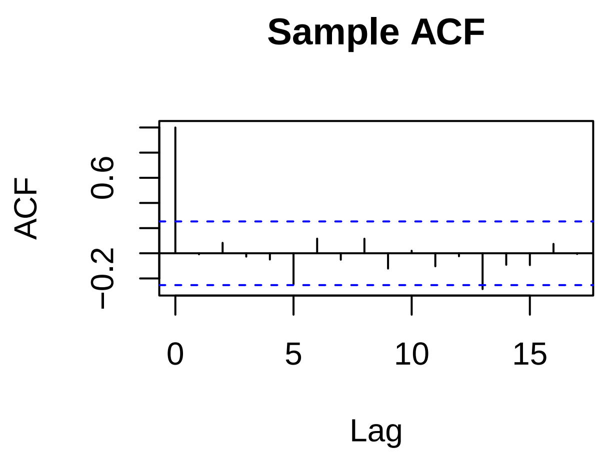


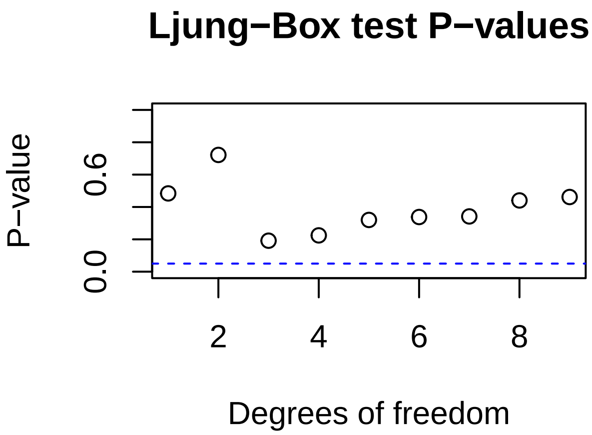
1) Since ACF cut off after lag 1, this suggest that we should begin by fitting an ARIMA(0,1,1) model
2) Since PACF doesn't cut off, this suggest the time series doesn't contain an AR component.

Attempt1: ARIMA(0,1,1)

After fitting the model, we perform goodness of fit check on ARIMA(0,1,1) base on the following three plots:







From the plots above, we conclude that ARIMA(0,1,1) is a good fit due to following reasons:

1) Time plot of the model residuals:

The time plot of the residuals looks similar to white noise, with mean zero and constant variance.

2) A plot of the sample ACF of the model residuals

For all lags > 0, the sample ACF are all close to zero. This suggests that the residuals are independent(uncorrelated).

3) A plot of the first ten P-values for the Ljung-Box test

All p-values are greater than 0.05(non-significant), this suggests the ARIMA(0,1,1) is a good fit to the data.

However, it's still worth checking if adding an AR(p) component would be a better fit. Therefore we fit the model again with ARIMA(1,1,1)

Comparison: ARIMA(0,1,1) vs. ARIMA(1,1,1)

```
##
## Call:
## arima(x = nhtemp, order = c(0, 1, 1), method = "ML")
##
## Coefficients:
##           ma1
##       -0.7983
## s.e.   0.0956
##
## sigma^2 estimated as 1.291:  log likelihood = -91.76,  aic = 187.52
```

```
##
## Call:
## arima(x = nhtemp, order = c(1, 1, 1), method = "ML")
##
## Coefficients:
##          ar1      ma1
##       0.0073  -0.8019
## s.e.  0.1802   0.1285
##
## sigma^2 estimated as 1.291:  log likelihood = -91.76,  aic = 189.52
```

From the summary above, we conclude that ARIMA(0,1,1) is better than ARIMA(1,1,1) due to following reasons:

1) AIC for ARIMA(0,1,1) is 187.12 is less than AIC for ARIMA(1,1,1), which is 189.52.

2) Perform hypothesis test: $H_0 : \phi_1 = 0$ vs. $H_1 : \phi_1 \neq 0$. The test statistic $= \frac{0.0073}{0.1802} < 2$, therefore we don't reject the null hypothesis and thus ARIMA(0,1,1) is better than ARIMA(1,1,1) model.

3) Overall we'd prefer a parsimonious model, thus ARIMA(0,1,1) is better than ARIMA(1,1,1)

##Conclusion

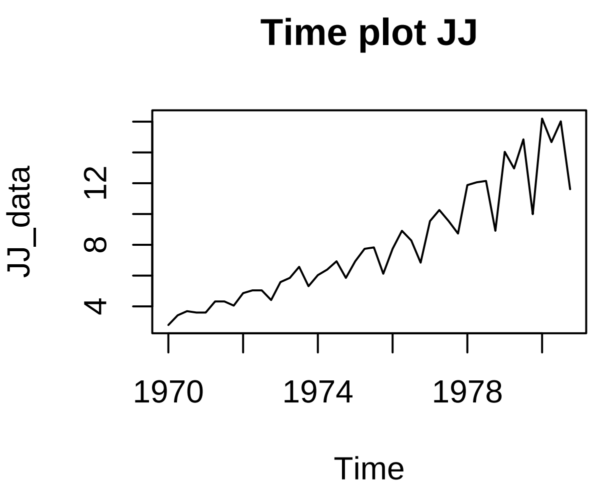
For question 1, the equation for the final fitted model is included below:

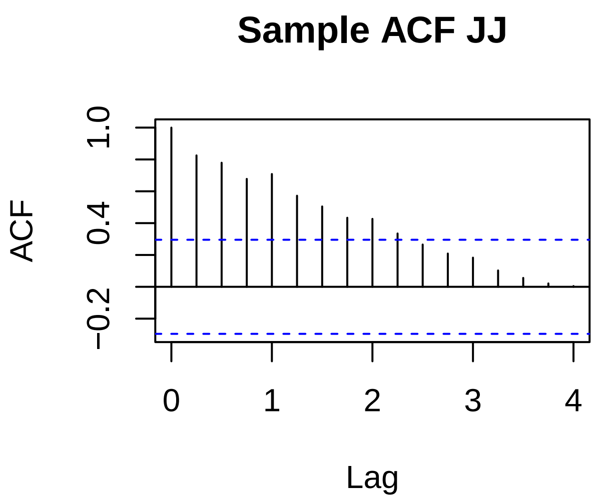
$$(1 - B)X_t = (1 - 0.7983B)Z_t$$

Q2: JJ_data

Part1: Check Stationarity and Seasonality

First we produce the time plot and ACF plot from the given data:





From the plots above, we conclude that the series is non-stationary and seasonal due to following reasons:

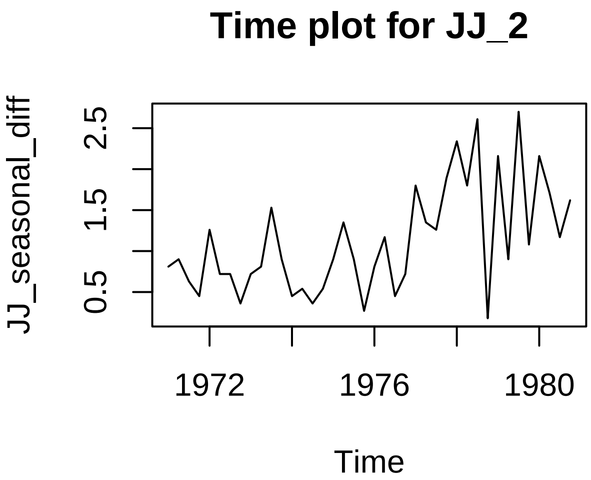
1) Time plot: both the mean and variance of the series appears to increase overtime, which indicate non-stationarity.

2) Sample ACF plot: doesn't decay rapidly, therefore it's not stationary.

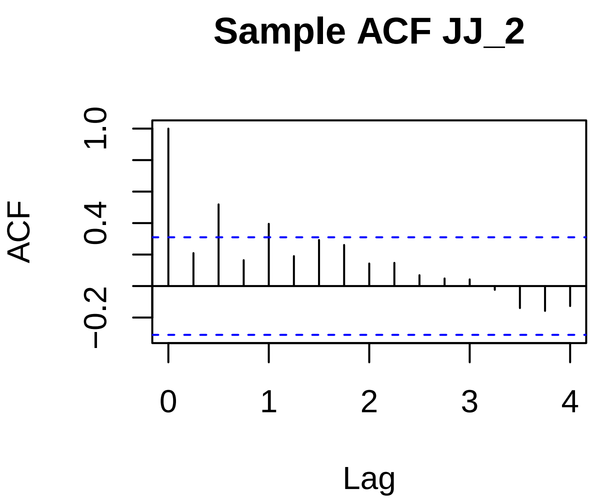
3) Time plot: the data shows seasonality, as the earnings are higher in Qtr 2,3 and lower in Qtr 1,4

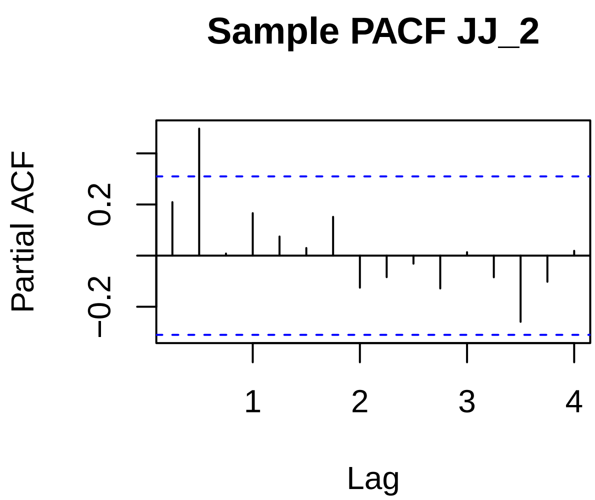
Therefore would need to apply a SARIMA model for JJ_data.

According to the data description, JJ is a time series of the quarterly earnings between years, so the seasonal difference lag should be set to $h = 4$. There fore if JJ_1 denotes our original time series, we define the lag 4 difference time series JJ_2 as $JJ_2 = \nabla_4 JJ_1 = (1 - B^4)JJ_1$



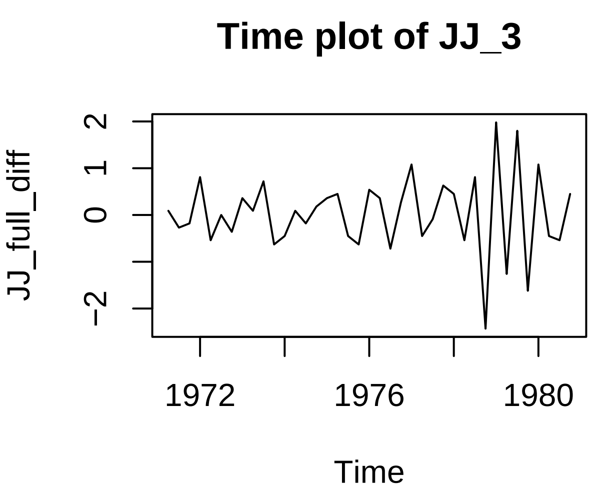
From the time plot, it seems that the seasonality has been removed in JJ_2.

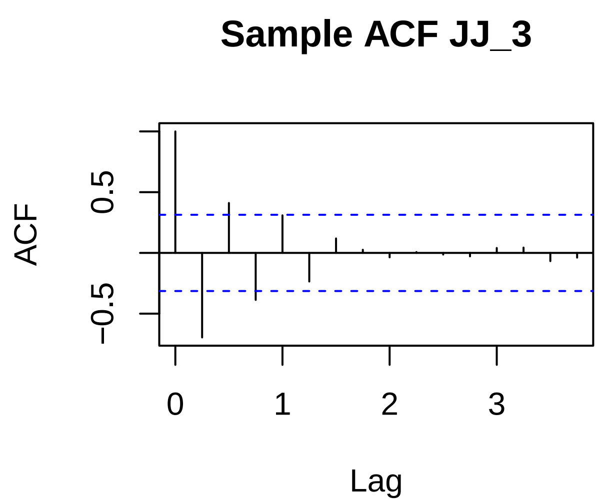




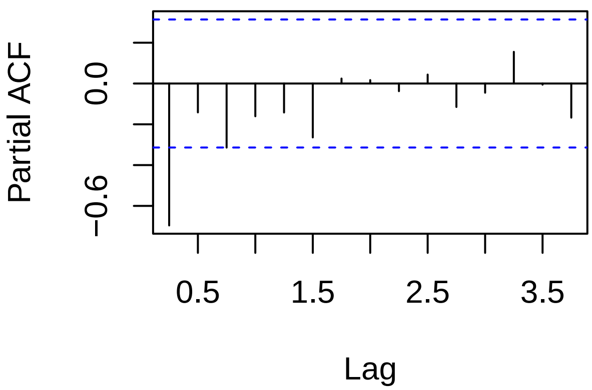
However, according to the sample ACF and sample PACF for the seasonally differnecd data, it suggest non-stationarity, because the ACF decays slowly.

Therefore we'll take the first difference of JJ_2 and obtain $JJ_3 = \nabla^1 JJ_2 = (1 - B)JJ_2$





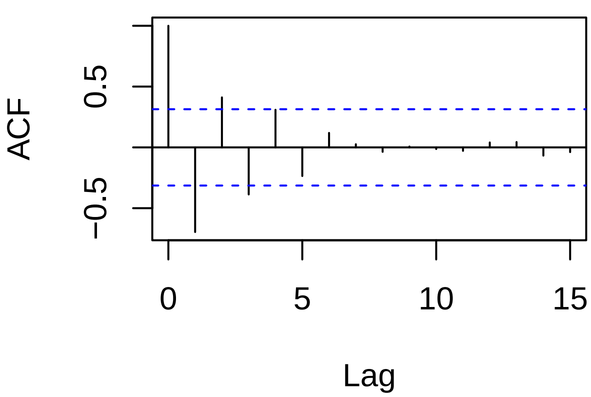
Sample PACF JJ_3



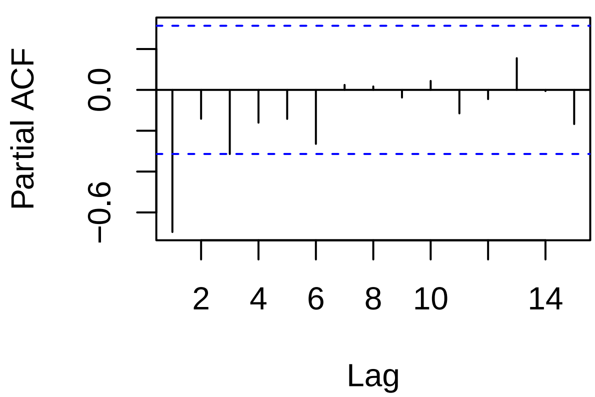
Now JJ_3 appear to be stationary without seasonality. Therefore we start our fitting attempt with SARIMA(p,1,q)x(P,1,Q)[4].

Parameter analysis

Sample ACF JJ_3



Sample PACF JJ_3



Based on the sample ACF and sample PACF figure, the best model to begin should be SARIMA(1,1,1) x (0,1,0)[4] due to following reasons:

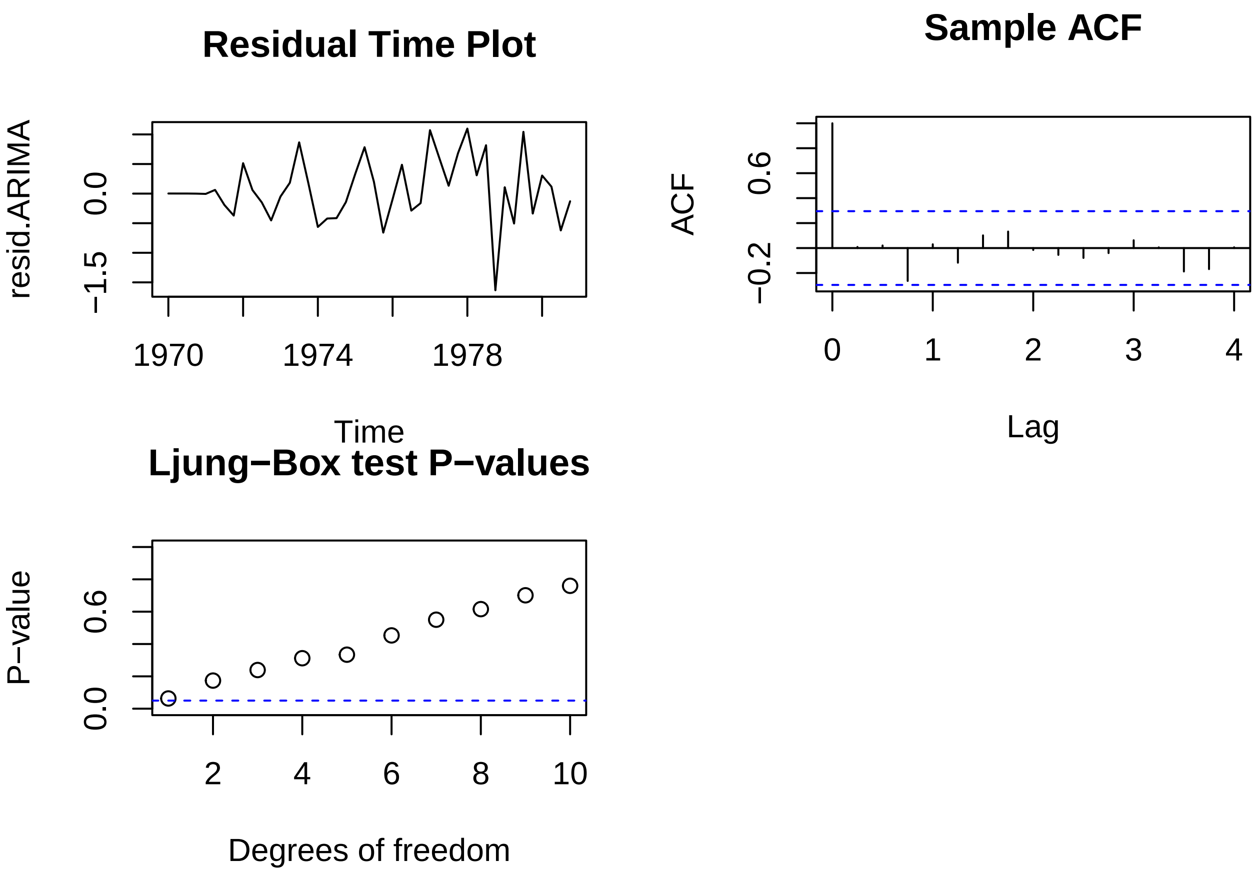
Seasonal components: (P,Q)

1) P: Check PACF at lag = 4,8,12 ... PACF cut off already at lag = 4, therefore we choose P = 0.
2) Q: Check ACF at lag = 4,8,12 ... ACF cut off already at lag = 4, therefore we choose Q = 0

Non Seasonal components : (p,q)

3) p: PACF cut off after lag = 1, therefore we choose p = 1
4) q: ACF cut off after lag = 1, therefore we choose q = 1.

Model fitting



From the plots above, we conclude that SARIMA(1,1,1)x(0,1,0)[4] is a fairly good fit due to following reasons:

1) Time plot of the model residuals:

The time plot of the residuals looks similar to white noise, with mean zero, but the variance increases overtime.

2) A plot of the sample ACF of the model residuals

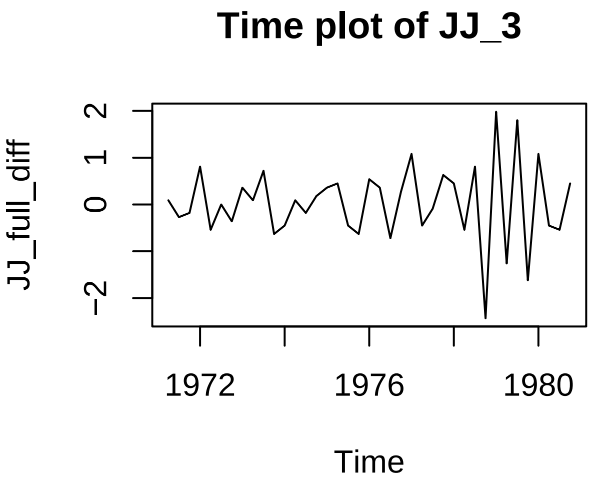
For all lags > 0, the sample ACF are all close to zero except at lag $= 1$. This suggests that the residuals are almost independent(uncorrelated).

3) A plot of the first ten P-values for the Ljung-Box test

Although the first p-value is fairly significant, all other p-values are greater than 0.05(non-significant), this suggests a fairly good model for JJ_data.

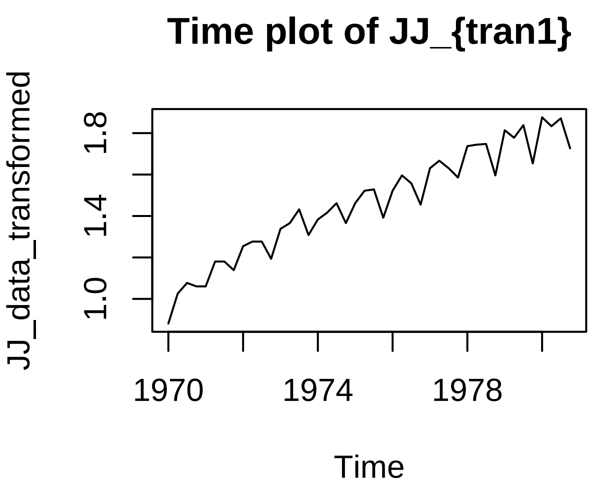
Transformed JJ_data

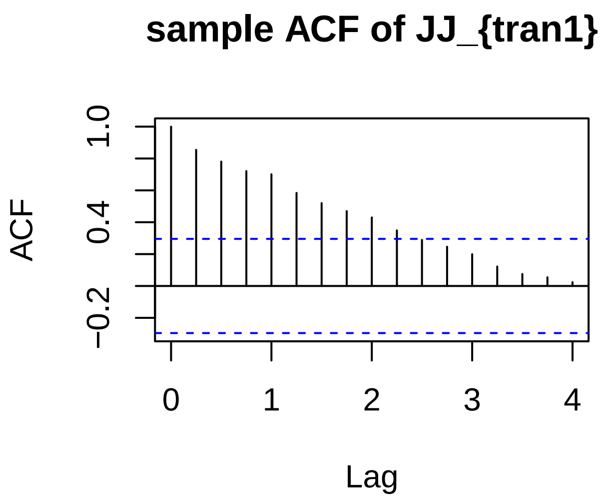
Look back to the time plot for JJ_3, it seems that the final part of the time series has greater variance compared with earlier part.



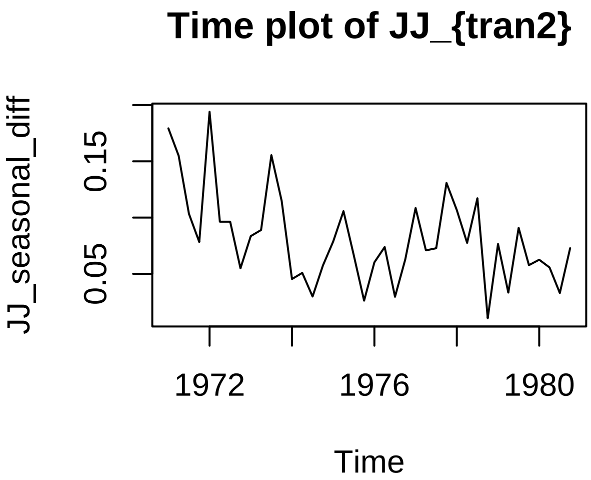
Therefore we perform transformation on JJ_1 to remove non-constant variance. I've applied both log and sqrt transformation but the final fitted model but the final fit doesn't perform well. Therefore I implement box-cox transformation on JJ_1, where the optimal lambda is chosen to be -0.305

$$JJ_{tran1} = boxcox(JJ_1)$$

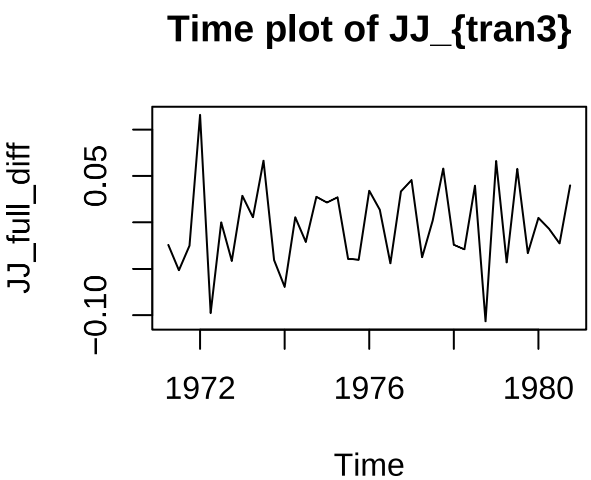


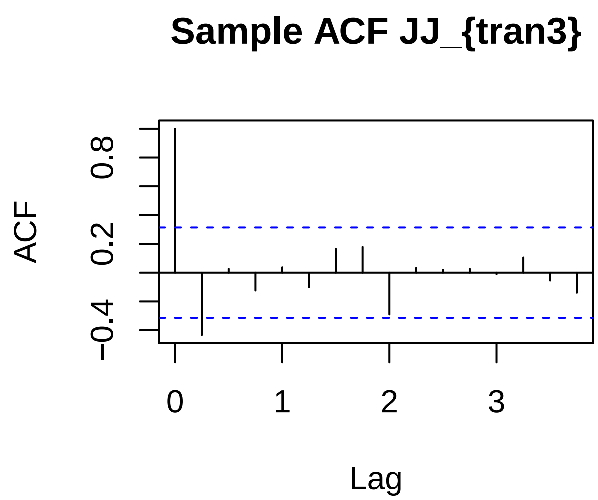


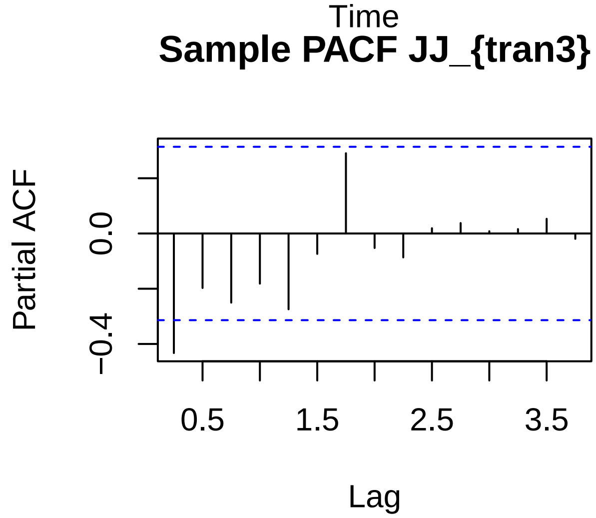
After that we carry on the same process to remove the non-stationarity. We difference JJ_{tran1} with a seasonal difference lag $h = 4$ and gain $JJ_{tran2} = \nabla_4(JJ_{tran1}) = (1 - B^4)(JJ_{tran1})$



From the time plot, it seems that the seasonality has been removed in JJ_{tran2}, however the data is still non-stationary, so we take the first difference on JJ_{tran2} and obtain $JJ_{tran3} = \nabla^1 JJ_{tran2} = (1-B)JJ_{tran2}$

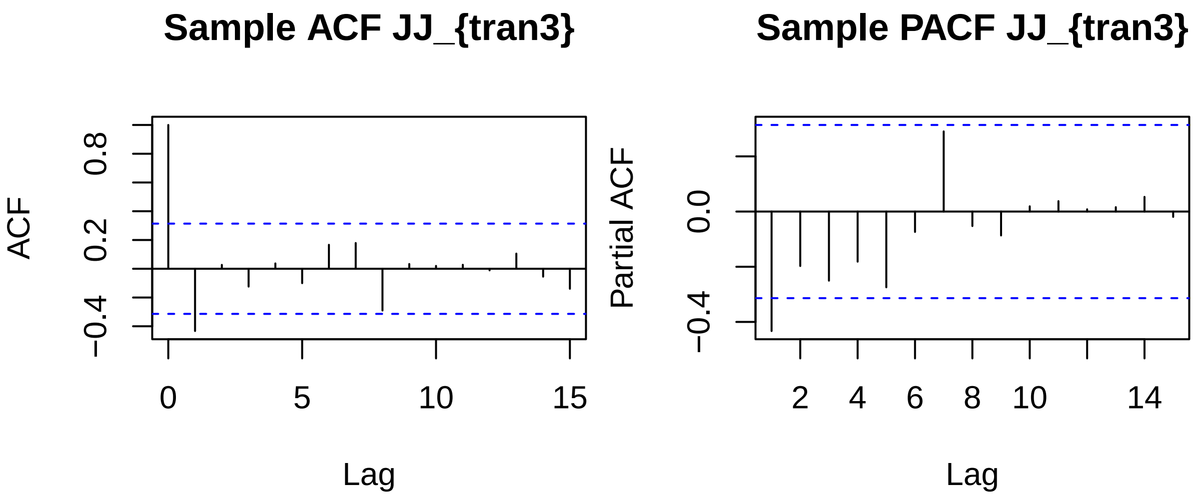






Now the data is stationary after performing box-cox transformation, seasonal difference at lag 4 and first difference. Now we start our fitting attempt with SARIMA(p,1,q)x(P,1,Q)[4].

Parameter analysis



Based on the sample ACF and sample PACF figure, the best model to begin should be SARIMA(0,1,1) x (0,1,0)[4] due to following reasons:

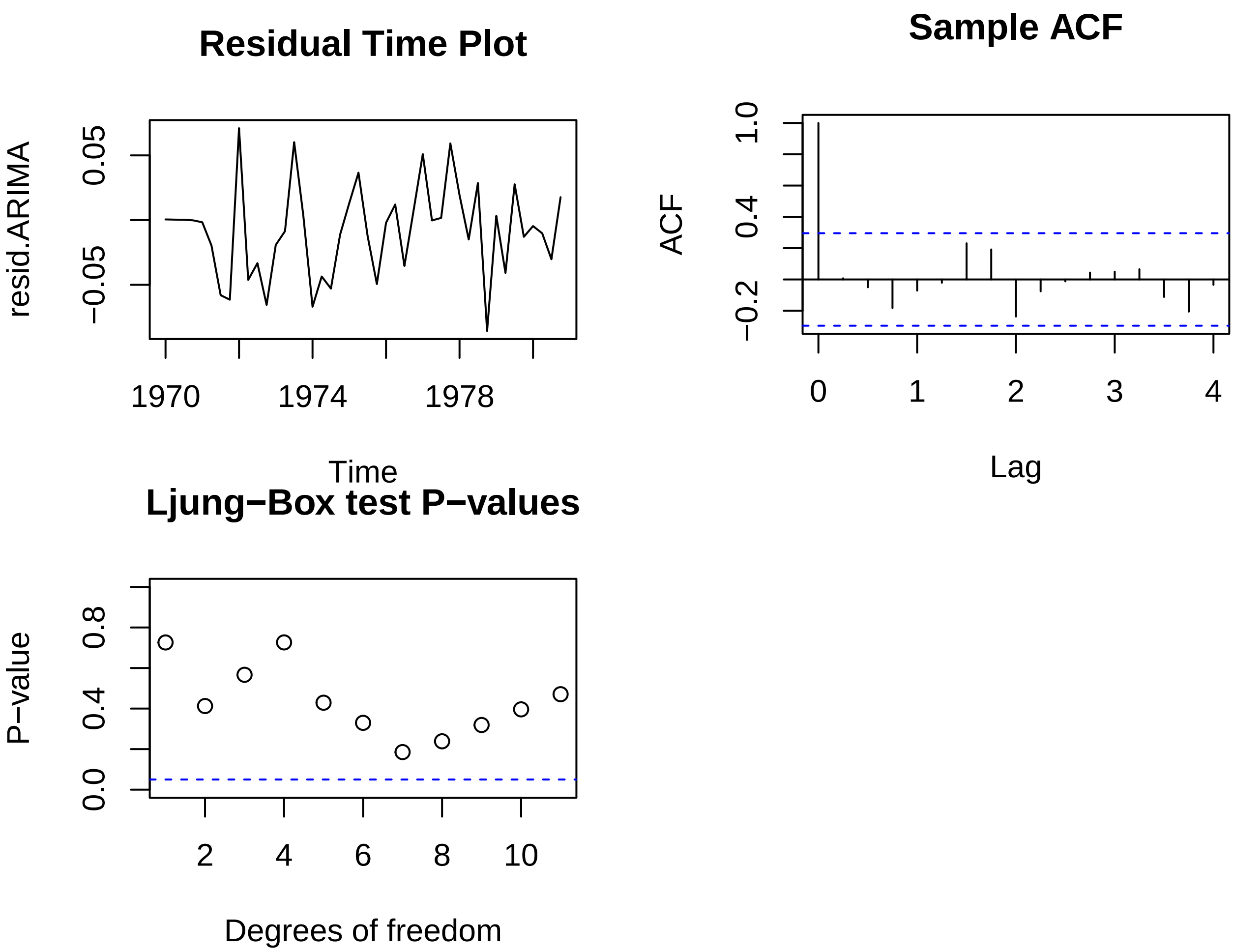
Seasonal components: (P,Q)

1) P: Check PACF at lag = 4,8,12 … PACF cut off already at lag = 4, therefore we choose P = 0.
2) Q: Check ACF at lag = 4,8,12 … ACF cut off already at lag = 4, therefore we choose Q = 0

Non Seasonal components : (p,q)

3) p: PACF cut off after lag = 0, therefore we choose p = 0
4) q: ACF cut off after lag = 1, therefore we choose q = 1.

Model fitting



From the plots above, we conclude that SARIMA(0,1,1)x(0,1,0)[4] is a good fit due to following reasons:

1) Time plot of the model residuals:

The time plot of the residuals looks to be white noise, with mean zero and constant variance.

2) A plot of the sample ACF of the model residuals

For all lags > 0, the sample ACF are all close to zero except at lag $= 1$. This suggests that the residuals are almost independent(uncorrelated).

3) A plot of the first ten P-values for the Ljung-Box test

Although the first p-value is fairly significant, all other p-values are greater than 0.05(non-significant), this suggests a fairly good model for JJ_data.

Conclusion

Let X_t denote the original JJ_data

For the non-transformed data, the best model is SARIMA(1,1,1)x(0,1,0)[4], and the equation is:

$$(1 + 0.3465B)(1 - B)(1 - B^4)X_t = (1 - 0.6308B)Z_t$$

For the boxcox transformed data, the best model is SARIMA(0,1,1)x(0,1,0)[4], and the equation is:

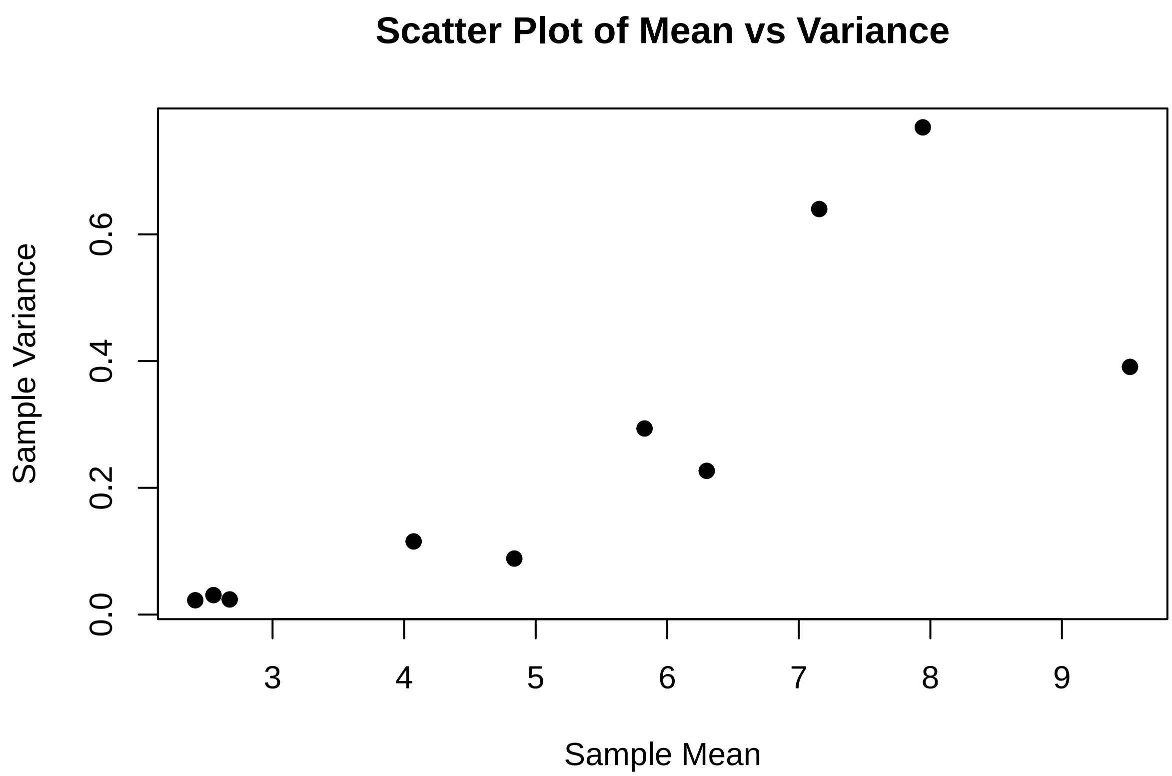
$$(1 - B)(1 - B^4)boxcox(X_t) = (1 - 0.7325B)Z_t$$

here $boxcox()$ denotes the transformation performed on the original JJ_data.

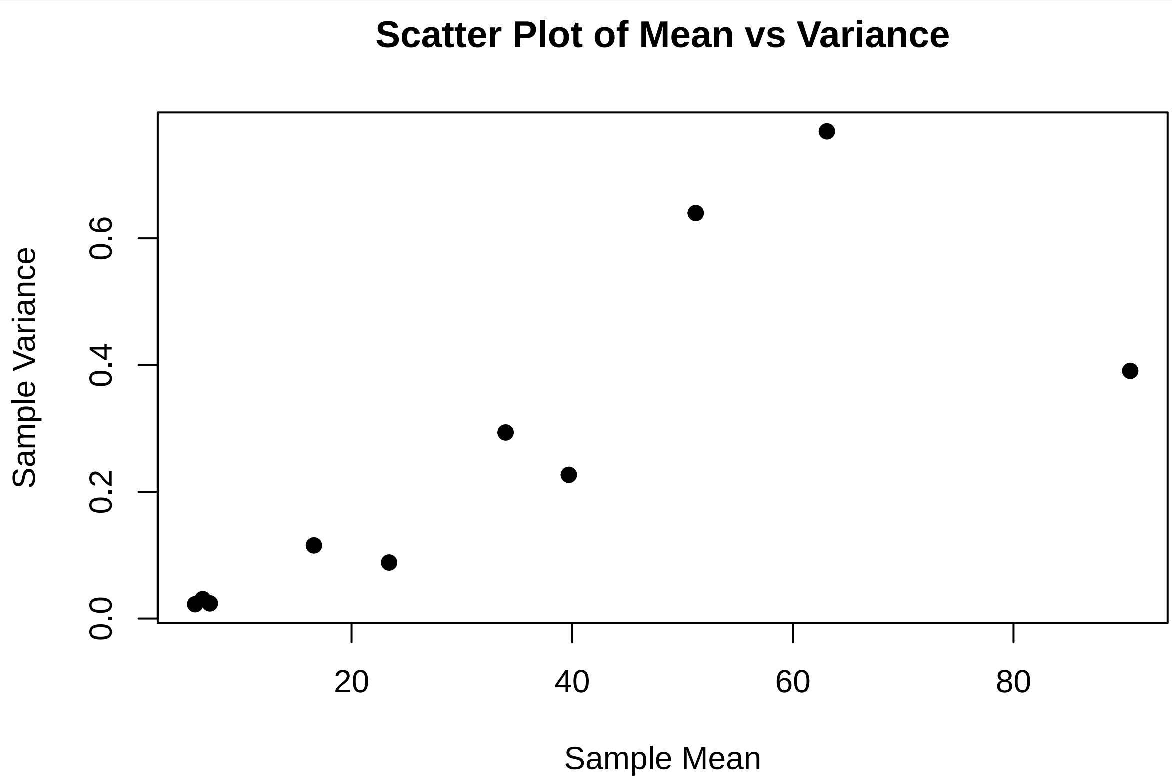
#trash content

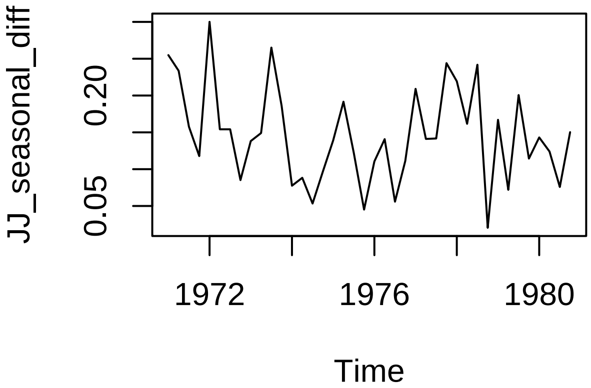
```
last_12 <- tail(JJ_data, 12)

# Apply a logarithmic transformation to the last 12 elements
transformed_last_12_log <- log(last_12)

# Replace the last 12 elements in the original time series with the transformed values
# Calculate the starting index for the last 12 elements
start_index <- length(JJ_data) - length(transformed_last_12_log) + 1
end_index <- length(JJ_data)

JJ_transform <- JJ_data

# Replace the elements
JJ_transform[start_index:end_index] <- transformed_last_12_log
```

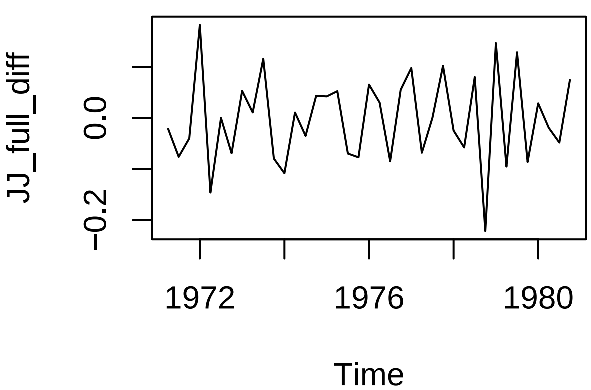
```
subset <- window(JJ_transform, start = c(1971,1), end = c(1971,4))
mean_subset = list()
var_subset = list()
mean_subset <- mean(subset)
var_subset <- var(subset)

for (k in 1:9){
  subset <- window(JJ_transform, start = c(1971+k,1), end = c(1971+k,4))

  mean_subset[[k+1]] <- mean(subset)
  var_subset[[k+1]] <- var(subset)
}

mean_vector <- unlist(mean_subset)
var_vector <- unlist(var_subset)

plot(mean_vector, var_vector, main="Scatter Plot of Mean vs Variance",
     xlab="Sample Mean", ylab="Sample Variance", pch=19)
```



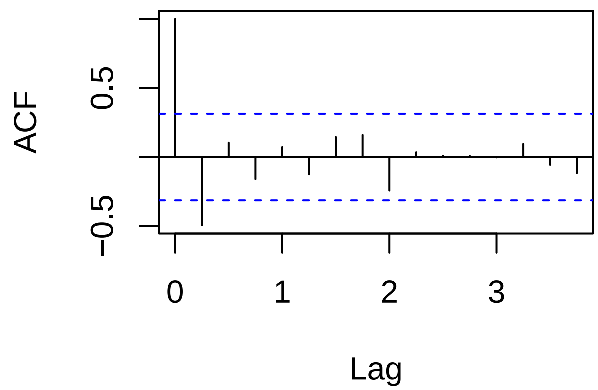

```r
plot(mean_vector^2, var_vector, main="Scatter Plot of Mean vs Variance",
     xlab="Sample Mean", ylab="Sample Variance", pch=19)
```



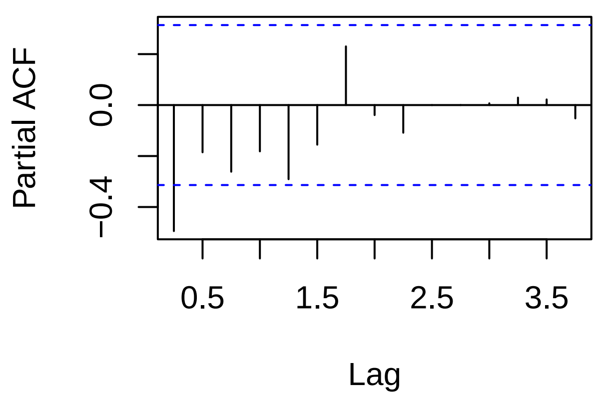




Series JJ_full_diff



Series JJ_full_diff



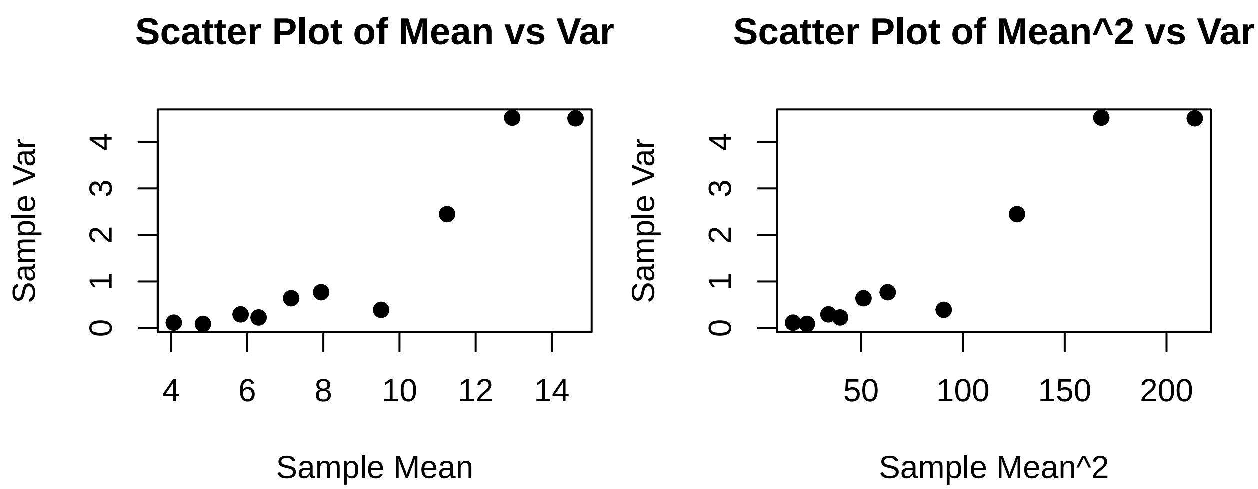

```r
fit <- auto.arima(log(JJ_data))
summary(fit)
```

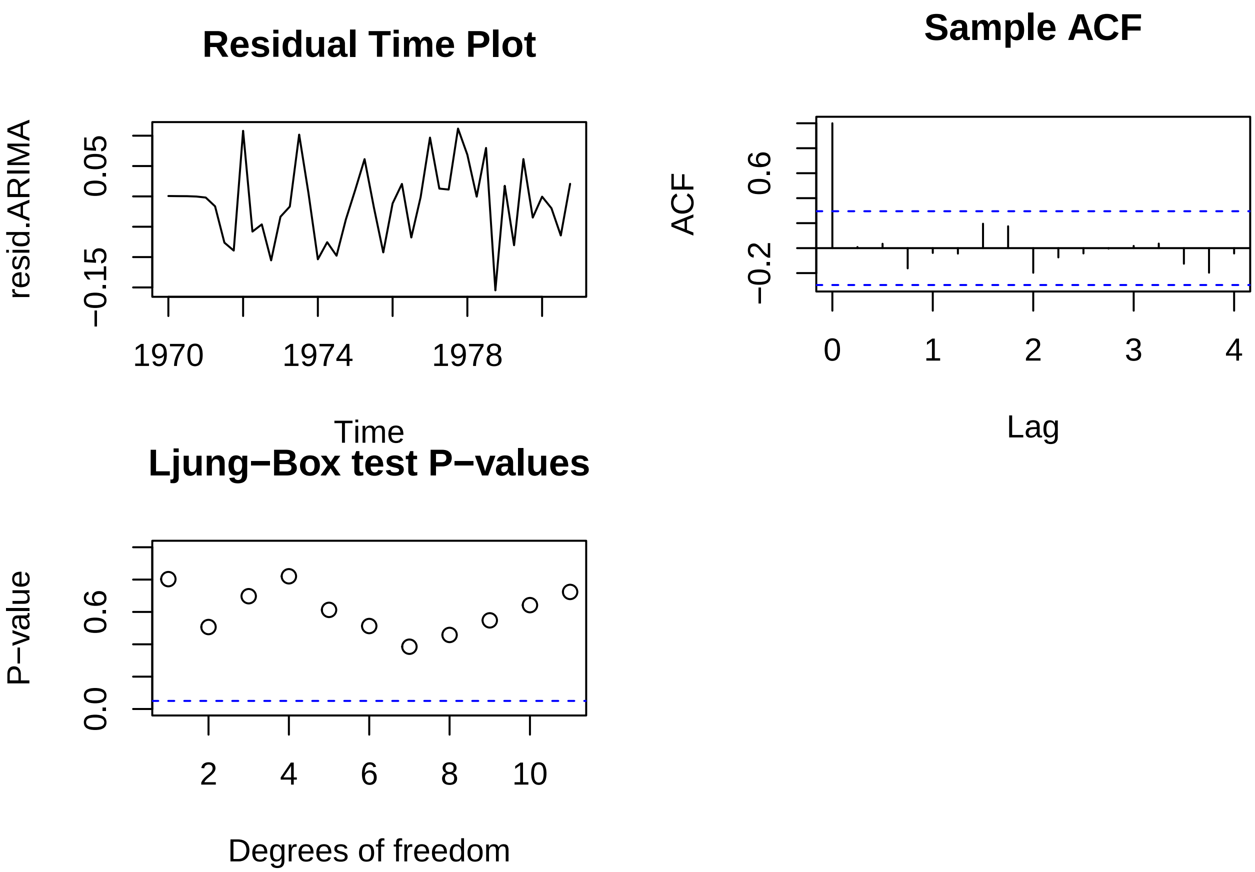
```
## Series: log(JJ_data)
## ARIMA(0,0,0)(0,1,0)[4] with drift
##
## Coefficients:
##         drift
##        0.0366
## s.e.   0.0026
##
```

```
## sigma^2 = 0.004255:  log likelihood = 52.94
## AIC=-101.88   AICc=-101.56   BIC=-98.51
##
## Training set error measures:
##                        ME       RMSE        MAE       MPE     MAPE      MASE
## Training set 0.0001017352 0.06141291 0.04580905 0.1018794 2.394819 0.3129954
##                   ACF1
## Training set 0.1119376
```

```r
fit <- auto.arima(JJ_data_transformed)
summary(fit)
```

```
## Series: JJ_data_transformed
## ARIMA(0,1,1)(0,1,0)[4]
##
## Coefficients:
##           ma1
##       -0.7325
## s.e.   0.1219
##
## sigma^2 = 0.001524:  log likelihood = 71.27
## AIC=-138.54   AICc=-138.21   BIC=-135.21
##
## Training set error measures:
##                        ME       RMSE        MAE        MPE     MAPE      MASE
## Training set -0.008508696 0.03627806 0.02730114 -0.6725556 1.910621 0.3345961
##                     ACF1
## Training set 0.007332271
```





Appendix

```
knitr::opts_chunk$set(echo = TRUE)

library(forecast)
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)
}
load("nhtemp.rda")
# Time Series Plot
ts.plot(nhtemp, main="Sample Time Plot")

# ACF Plot
```

```r
acf(nhtemp, main="Sample ACF")

# PACF Plot
#pacf(nhtemp, main="Sample PACF")
nhtemp_diff<-diff(nhtemp)
ts.plot(nhtemp_diff, main="Sample Time Plot")
acf(nhtemp_diff, main="Sample ACF")
#pacf(nhtemp_diff)
acf(nhtemp_diff, main="Sample ACF")
pacf(nhtemp_diff, main = "Sample PACF")
ARIMA<-arima(nhtemp,order=c(0,1,1),method="ML")
ARIMA
resid.ARIMA<-residuals(ARIMA)
ts.plot(resid.ARIMA, main = "Sample Time Plot")
acf(resid.ARIMA, main = "Sample ACF")
ARIMA.LB<-LB_test(resid.ARIMA,max.k=11,p=0,q=2)
#To produce a plot of the P-values against the degrees of freedom and
#add a blue dashed line at 0.05, we run the commands
plot(ARIMA.LB$deg_freedom,ARIMA.LB$LB_p_value,xlab="Degrees of freedom",ylab="P-value",main="Ljung-Box t
abline(h=0.05,col="blue",lty=2)
ARIMA
ARIMA<-arima(nhtemp,order=c(1,1,1),method="ML")
ARIMA
load("JJ_data.rda")
#JJ_data_ts <- ts(JJ_data, start=c(1970, 1))
LB_test_SARIMA<-function(resid,max.k,p,q,P,Q){
 lb_result<-list()
 df<-list()
 p_value<-list()
  for(i in (p+q+P+Q+1):max.k){
   lb_result[[i]]<-Box.test(resid,lag=i,type=c("Ljung-Box"),fitdf=(p+q+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)
 }
ts.plot(JJ_data, main = "Time plot JJ")
acf(JJ_data,main = "Sample ACF JJ")
#pacf(JJ_data)
JJ_seasonal_diff <- diff(JJ_data,lag=4)
ts.plot(JJ_seasonal_diff,main ="Time plot for JJ_2")
acf(JJ_seasonal_diff,main = "Sample ACF JJ_2")
pacf(JJ_seasonal_diff, main = "Sample PACF JJ_2")
JJ_full_diff <- diff(JJ_seasonal_diff)
ts.plot(JJ_full_diff,main = "Time plot of JJ_3")
acf(JJ_full_diff, main = "Sample ACF JJ_3")
pacf(JJ_full_diff, main = "Sample PACF JJ_3")
JJ_data_ts <- ts(JJ_full_diff, start=c(1970, 1))
acf(JJ_data_ts, main = "Sample ACF JJ_3")
```

```r
pacf(JJ_data_ts, main = "Sample PACF JJ_3")
ARIMA<-arima(JJ_data,order=c(1,1,1),seasonal=list(order=c(0,1,0),period=4),method="ML")
ARIMA
resid.ARIMA<-residuals(ARIMA)
ts.plot(resid.ARIMA, main = "Residual Time Plot")
acf(resid.ARIMA, main = "Sample ACF")
ARIMA.LB<-LB_test_SARIMA(resid.ARIMA,max.k=12,p=1,q=1,P=0,Q=0)
#To produce a plot of the P-values against the degrees of freedom and
#add a blue dashed line at 0.05, we run the commands
plot(ARIMA.LB$deg_freedom,ARIMA.LB$LB_p_value,xlab="Degrees of freedom",ylab="P-value",main="Ljung-Box 
abline(h=0.05,col="blue",lty=2)
ts.plot(JJ_full_diff,main = "Time plot of JJ_3")
# Estimate the optimal lambda for the Box-Cox transformation
lambda <- BoxCox.lambda(JJ_data)
# Apply the Box-Cox transformation
JJ_data_transformed <- BoxCox(JJ_data, lambda)
ts.plot(JJ_data_transformed, main ="Time plot of JJ_{tran1}")
acf(JJ_data_transformed, main = "sample ACF of JJ_{tran1}")
#pacf(JJ_data)
JJ_seasonal_diff <- diff(JJ_data_transformed,lag=4)
ts.plot(JJ_seasonal_diff,main ="Time plot of JJ_{tran2}")
JJ_full_diff <- diff(JJ_seasonal_diff)
ts.plot(JJ_full_diff,main = "Time plot of JJ_{tran3}")
acf(JJ_full_diff, main = "Sample ACF JJ_{tran3}")
pacf(JJ_full_diff, main = "Sample PACF JJ_{tran3}")
JJ_data_ts <- ts(JJ_full_diff, start=c(1970, 1))
acf(JJ_data_ts, main = "Sample ACF JJ_{tran3}")
pacf(JJ_data_ts, main = "Sample PACF JJ_{tran3}")
fit <- auto.arima(JJ_data_transformed)
summary(fit)
ARIMA<-arima(JJ_data_transformed,order=c(0,1,1),seasonal=list(order=c(0,1,0),period=4),method="ML")

#ARIMA

resid.ARIMA<-residuals(ARIMA)
ts.plot(resid.ARIMA, main = "Residual Time Plot")
acf(resid.ARIMA, main = "Sample ACF")
ARIMA.LB<-LB_test_SARIMA(resid.ARIMA,max.k=12,p=0,q=1,P=0,Q=0)
#To produce a plot of the P-values against the degrees of freedom and
#add a blue dashed line at 0.05, we run the commands
plot(ARIMA.LB$deg_freedom,ARIMA.LB$LB_p_value,xlab="Degrees of freedom",ylab="P-value",main="Ljung-Box 
abline(h=0.05,col="blue",lty=2)
library(forecast)
fit <- auto.arima(JJ_data)
summary(fit)

last_12 <- tail(JJ_data, 12)

# Apply a logarithmic transformation to the last 12 elements
transformed_last_12_log <- log(last_12)

# Replace the last 12 elements in the original time series with the transformed values
# Calculate the starting index for the last 12 elements
```

```r
start_index <- length(JJ_data) - length(transformed_last_12_log) + 1
end_index <- length(JJ_data)

JJ_transform <- JJ_data

# Replace the elements
JJ_transform[start_index:end_index] <- transformed_last_12_log
subset <- window(JJ_transform, start = c(1971,1), end = c(1971,4))
mean_subset = list()
var_subset = list()
mean_subset <- mean(subset)
var_subset <- var(subset)

for (k in 1:9){
  subset <- window(JJ_transform, start = c(1971+k,1), end = c(1971+k,4))

  mean_subset[[k+1]] <- mean(subset)
  var_subset[[k+1]] <- var(subset)
}

mean_vector <- unlist(mean_subset)
var_vector <- unlist(var_subset)

plot(mean_vector, var_vector, main="Scatter Plot of Mean vs Variance",
     xlab="Sample Mean", ylab="Sample Variance", pch=19)

plot(mean_vector^2, var_vector, main="Scatter Plot of Mean vs Variance",
     xlab="Sample Mean", ylab="Sample Variance", pch=19)

JJ_seasonal_diff <- diff(log(JJ_data),lag=4)
ts.plot(JJ_seasonal_diff)
JJ_full_diff <- diff(JJ_seasonal_diff)
ts.plot(JJ_full_diff)
acf(JJ_full_diff)
pacf(JJ_full_diff)
fit <- auto.arima(log(JJ_data))
summary(fit)
fit <- auto.arima(JJ_data_transformed)
summary(fit)
subset <- window(JJ_data, start = c(1971,1), end = c(1971,4))
mean_subset = list()
var_subset = list()
mean_subset <- mean(subset)
var_subset <- var(subset)

for (k in 1:9){
  subset <- window(JJ_data, start = c(1971+k,1), end = c(1971+k,4))

  mean_subset[[k+1]] <- mean(subset)
  var_subset[[k+1]] <- var(subset)
}

mean_vector <- unlist(mean_subset)
```

```
var_vector <- unlist(var_subset)

plot(mean_vector, var_vector, main="Scatter Plot of Mean vs Var",
     xlab="Sample Mean", ylab="Sample Var", pch=19)

plot(mean_vector^2, var_vector, main="Scatter Plot of Mean^2 vs Var",
     xlab="Sample Mean^2", ylab="Sample Var", pch=19)

ARIMA<-arima(log(JJ_data),order=c(0,1,1),seasonal=list(order=c(0,1,0),period=4),method="ML")
ARIMA
resid.ARIMA<-residuals(ARIMA)
ts.plot(resid.ARIMA, main = "Residual Time Plot")
acf(resid.ARIMA, main = "Sample ACF")
ARIMA.LB<-LB_test_SARIMA(resid.ARIMA,max.k=12,p=0,q=1,P=0,Q=0)
#To produce a plot of the P-values against the degrees of freedom and
#add a blue dashed line at 0.05, we run the commands
plot(ARIMA.LB$deg_freedom,ARIMA.LB$LB_p_value,xlab="Degrees of freedom",ylab="P-value",main="Ljung-Box
abline(h=0.05,col="blue",lty=2)
```