Final Project Report Time Series Analysis

STAT 5664 – Fall22

# Title:
# Predicting Death of Heart Disease in US from 1969 through 2020 for Men and Women

**Data Reference:**

https://www.cdc.gov/pcd/issues/2016/16_0211a.htm#2

Professor Marco Ferreira

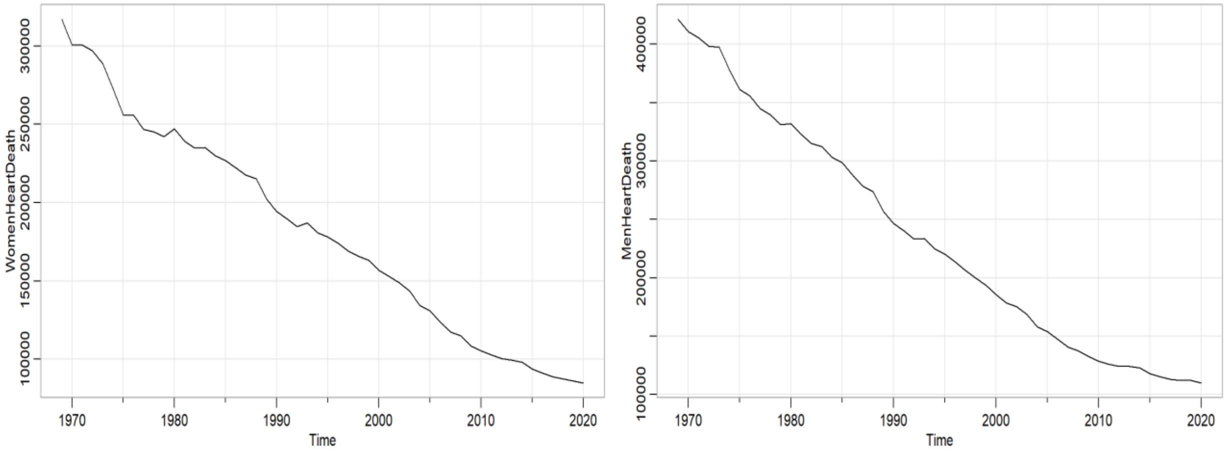
Elham Nasarian

December 2022

In this project I am analyzing yearly data collected for the Death of Heart Disease from 1969 to 2020. I have applied the steps from Data Activity 1, 2,3 and 4 in the following sections. All codes, is provided at the end of this report.

## 1. Plotting Heart Disease Time Series Datasets:

**Death of Heart Disease for Women & Men in the US from 1969 to 2020**



## 2. Heart Disease Time Series Datasets Stationary Status:

Both plot show, this time series dataset is **non- stationary** for women and men, because it depends on the time, and during the time it has a little increasing in 1972, 1980, 1987, and 1993 for women, also a little increasing in 1973, 1980, and 1993 for men, and then again decreasing.

## 3. Heart Disease Time Series Datasets Notable Trends:

Both plot show, if they have some short-term increasing changes over the time for women and men, but they have decreasing trends during the time from 1969 to 2020.

## 4. Heart Disease Time Series Datasets cyclical behaviors:

Plots demonstrate that short-term (every year) and long-term (1969 - 2020) cyclical behaviors for women and men in US.

## 5. Heart Disease Time Series Datasets Heteroscedasticity Status:

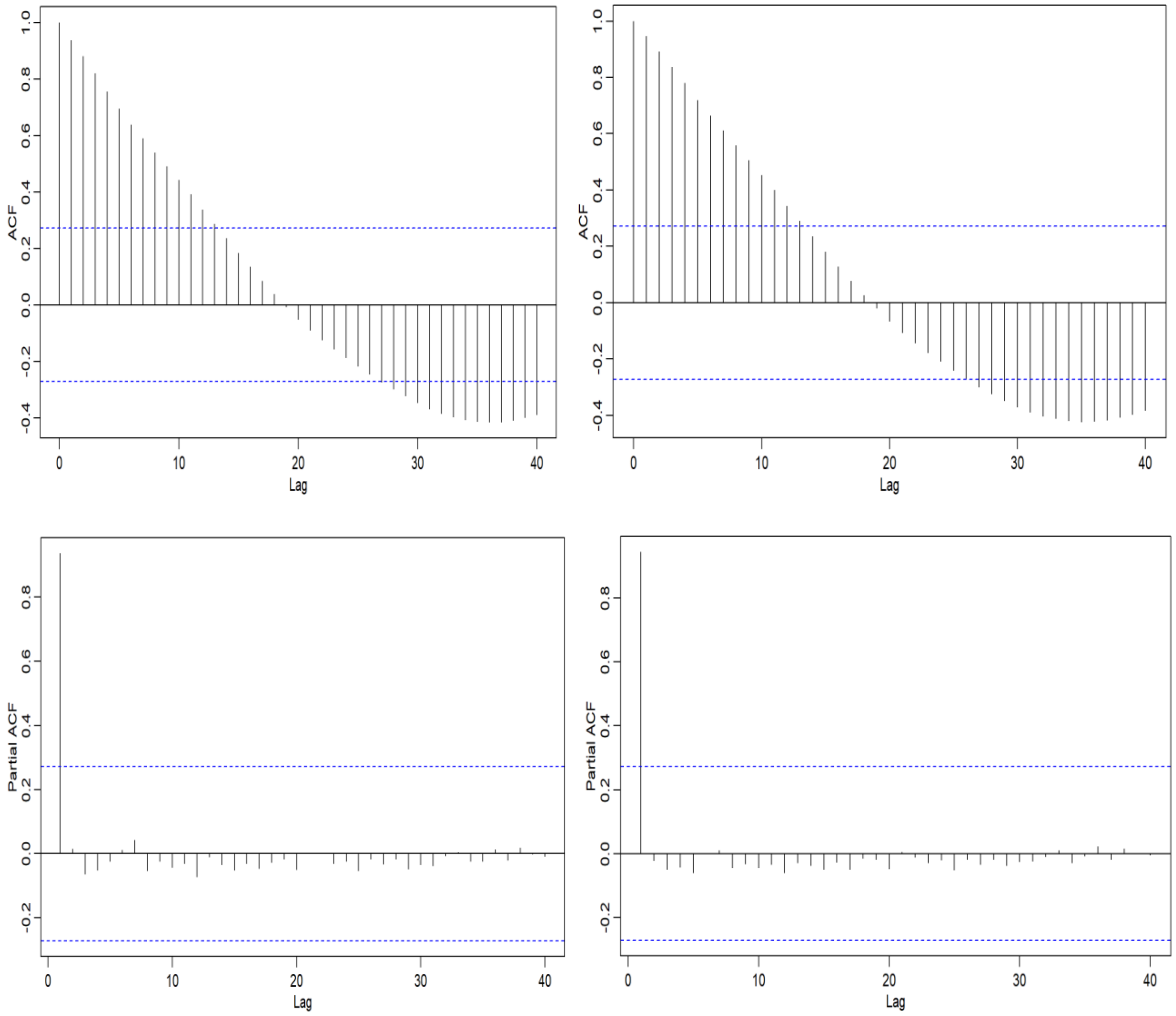
Looking at the plots, we can easily find that the variance in the data does change over available data and it is not constant and non-linear, so, we can argue that there is a heteroscedasticity in the dataset over time for women and men.

## 6. Heart Disease Time Series Datasets Seasonality Status:

By looking at the plots, we can find seasonality, because we can see some short-term increasing and decreasing like in 1972, 1980, 1987, and 1993 for women, also a little increasing in 1973, 1980, and 1993 for men. However, in general both plots have decreasing changes over the time.

## 7. PCF & PACF Plot for Heart Disease Time Series Datasets:

**Death of Heart Disease for Women & Men in the US from 1969 to 2020**

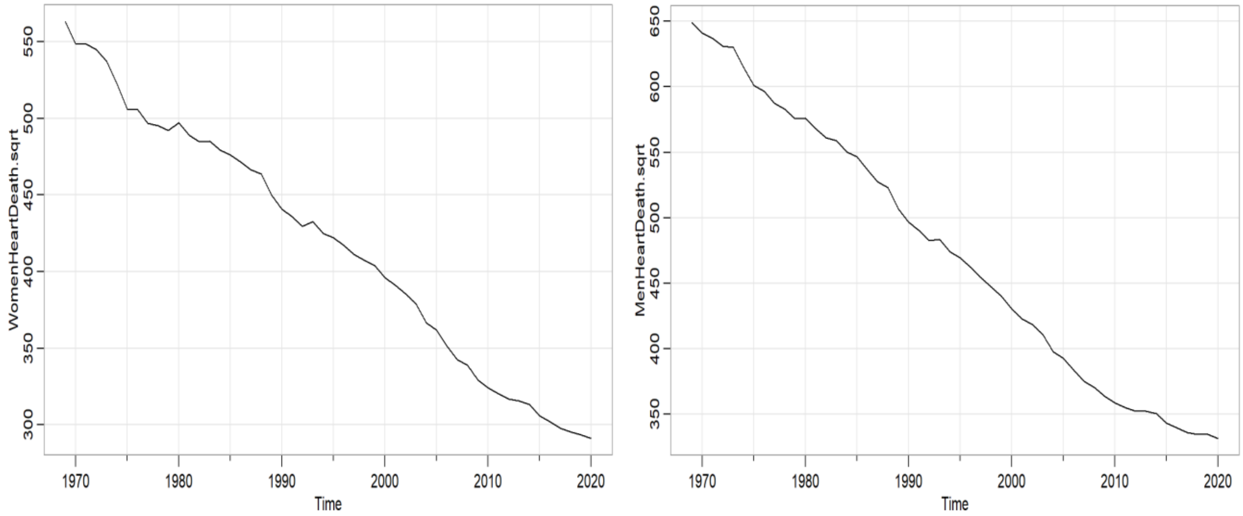


We can make the following observations:

- There are several autocorrelations that are significantly non-zero.
- Therefore, the time series is non-random.
- There is a large peak in the first lag (lag = 1), followed by a decreasing wave that alternates between positive and negative correlations. Which can mean an autoregressive term of higher order in the data.

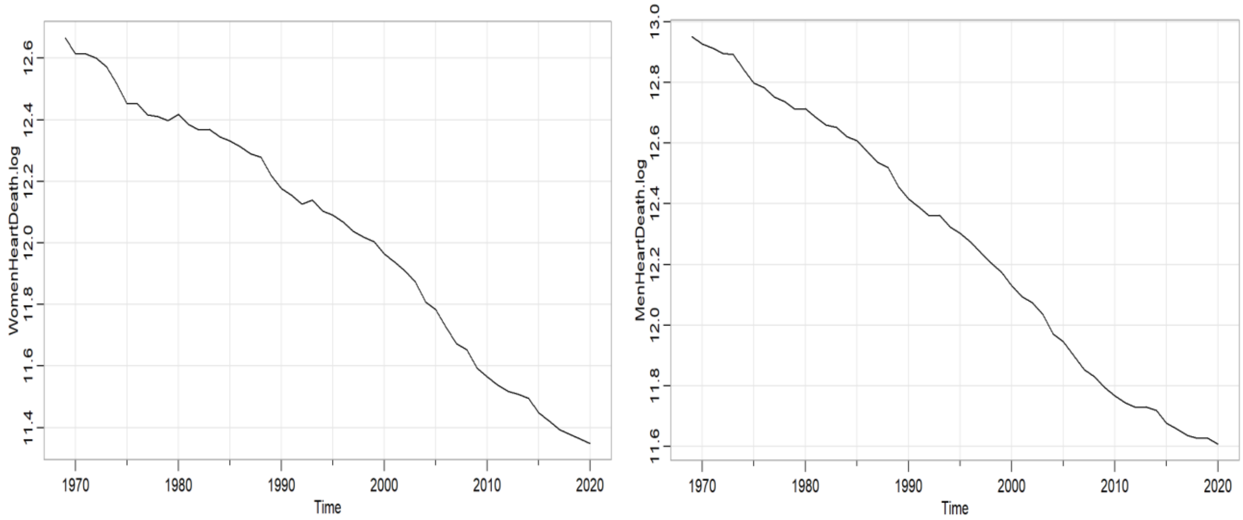
## 8. Removing trends and Analysis the Change in Heteroscedasticity for Heart Disease Time Series Datasets:

- **Square-Root Transformation plot**



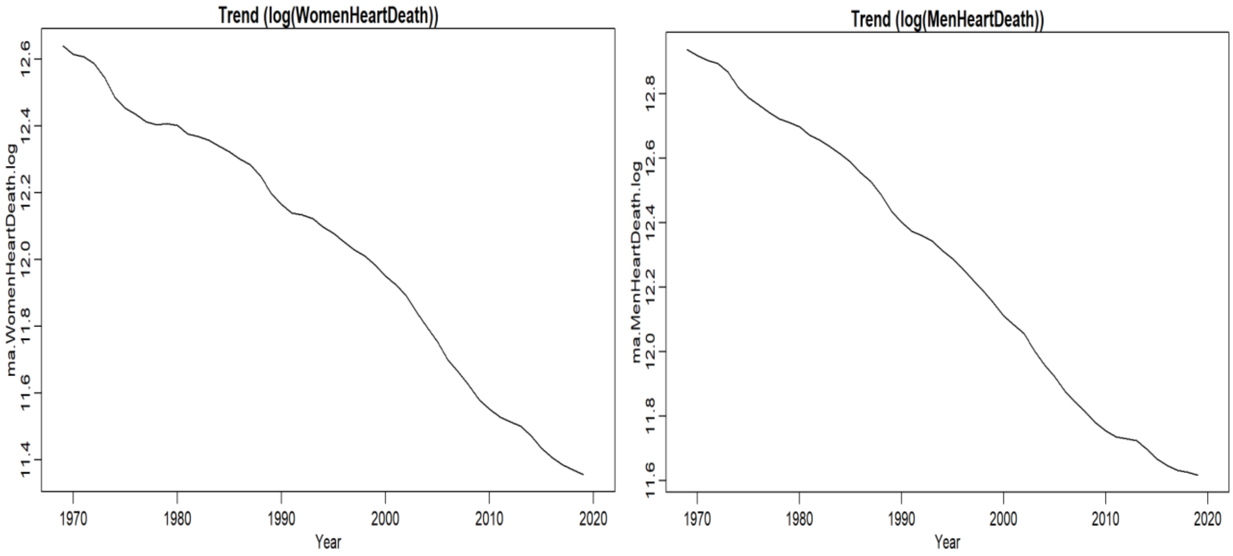
By looking at the square root transformation plot, we can see that square root transformation did not make the data homoscedastic since variance is changing over the time and not constant, so let's look at logarithmic transformation plot.

- **Logarithmic Transformation plot**



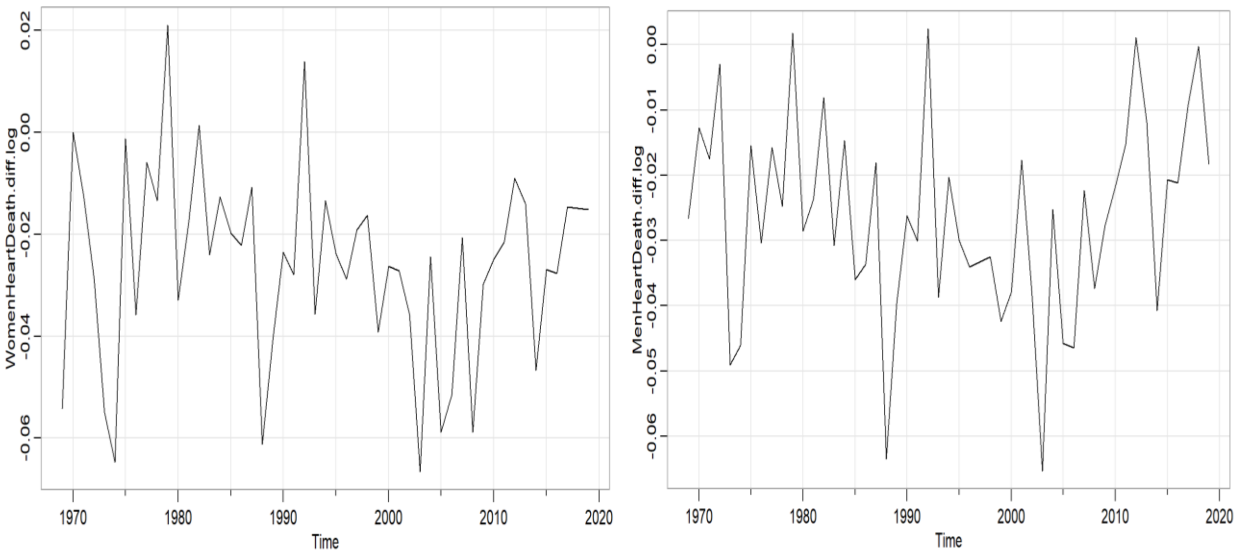
By Observing the plot from log transformation, we cannot see a constant variance over time. As a result, the logarithm transformation did not make the data homoscedastic, and still, we have decreasing trends in both women and men plots.

- **Trend Estimation Using 2 year Moving Average plot**



By observing both logarithmic plots, we can see a decreasing trends in time series. It has a short-term in creasing of death for both men and women, but in general, we can see decreasing the number of deaths over time.

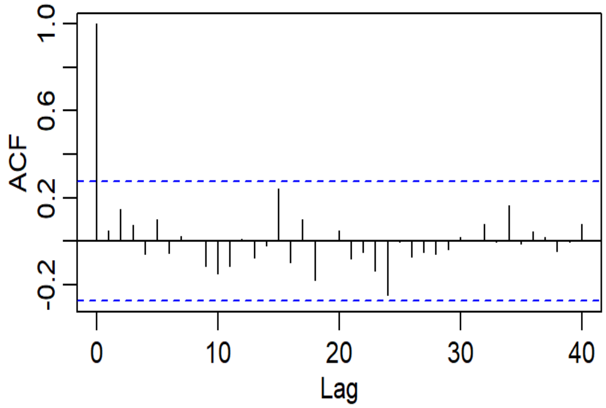
- **Removing the trend by the first-order Differencing the Logarithmic Transformed Timeseries**

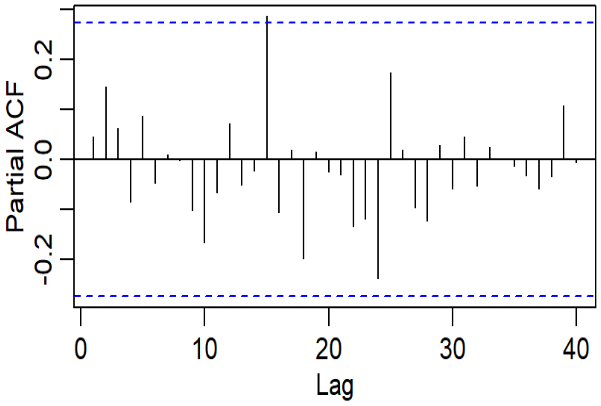


By looking at the plot, we cannot see a constant variance in the data, on the other words, it doesn't have any clear patterns, and we can say that it is a time series stationary.

## 9. Finding the Order of Autoregressive Model By plotting ACF & PACF:

- **Women:**

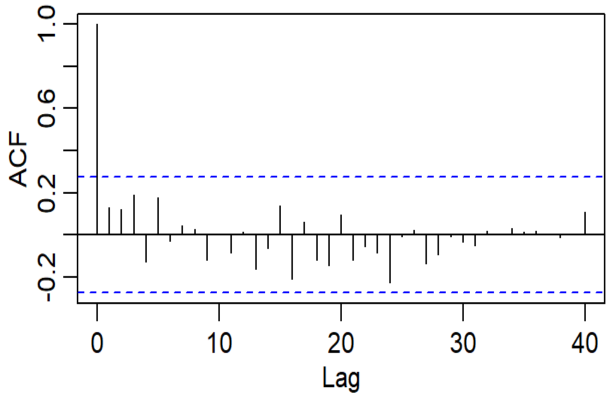


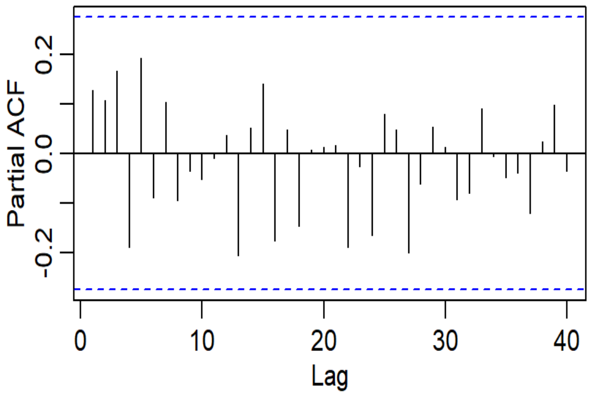


To find the best order of an AR model, we need to look at the PACF Plot and identify the significant spike in it. During last steps, we tried to make the time series data stationary. Now let's find the best order for Autoregressive model or AR.

From the Women PACF plot of the first-order Differencing the Logarithmic Transformed Timeseries plot (above plot), we can see that there is a significant spike at lag 8 and 15 in the plot. So, at this step, we ch oose 15, and we want to perform more analysis based on BIC and AIC.

- **Men:**





From the Men PACF plot of the first-order Differencing the Logarithmic Transformed Timeseries plot (ab ove plot), we can see that there is a significant spike at lag 5 and 13 in the plot. So, at this step, we choose 13, and we want to perform more analysis based on BIC and AIC.

## 10. AIC and BIC Model Selection:

We want to apply BIC and AIC model selection based on maximum order of 15 for WOMEN, AND 13 FOR MEN. The result is as bellow:

- **Women/AIC:**

```
    0            1           2          3          4          5          6          7
0.000000     1.888821    2.756037   4.570272   6.109899   7.709623   9.575081   11.572306
    8            9          10         11         12         13         14         15
13.571260   14.888775   14.998450  16.625549  17.951481  19.899495  21.899475  16.262226
```

- **Women/BIC:**

```
    0            1           2          3           4           5           6           7
1.931826     5.752473    8.551514   12.297575   15.769027   19.300577   23.097860   27.026911
    8            9          10          11          12          13          14          15
30.957691   34.207032   36.248532   39.807457   43.065214   46.945054   50.876860   47.171436
```

- **Men/AIC:**

```
    0           1          2          3          4          5          6          7
0.000000    1.171567   2.538208   2.890132   3.249277   3.121383   4.524050   5.917322
    8           9          10         11         12         13
 7.651082   9.526317   11.304986  13.305550  15.127929  13.019402
```

- **Men/BIC:**

```
    0           1          2          3           4           5           6           7
1.931826    5.035219   8.333685   10.617435   12.908405   14.712337   18.046830   21.371927
    8           9          10          11          12          13
25.037513   28.844574  32.555068   36.487458   40.241662   40.064961
```

**We can see in women and men, that both AIC and BIC values are the lowers for order 0. So, we select AR (0) for the model.**

## 11. Maximum likelihood estimation:

I want to fit an AR (0) model with maximum likelihood.

- **Women:**

  Order selected: 0   sigma^2 estimated as: 0.0003593

  Intercept: -0.02586755

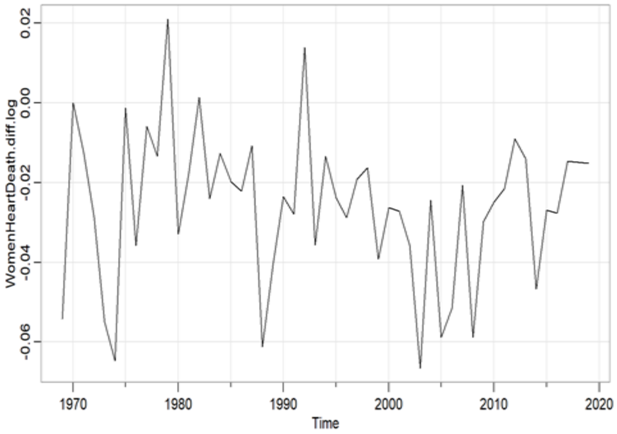
- **Men**:

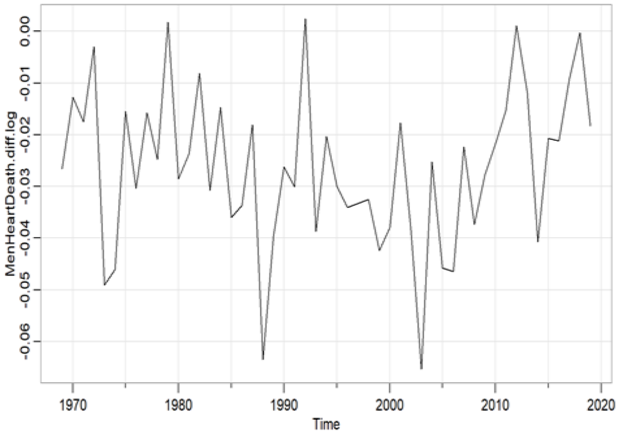
  Order selected: 0   sigma^2 estimated as: 0.0002295

  Intercept: -0.02634886

## 12. At the following steps, I want to find Q, P, p, and q for applying SARIMA model and make Prediction.

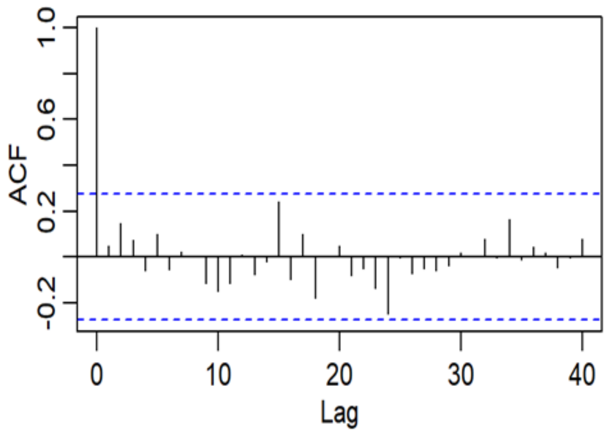
As I described in the previous steps, our time series dataset for death of heart disease in US, is non-stationary, and we applied log transformer and first order difference transformer to remove this trend, and after applying the first order difference transformer, we saw, the time series didn't have any trends, and we can say that it is a stationary time series.

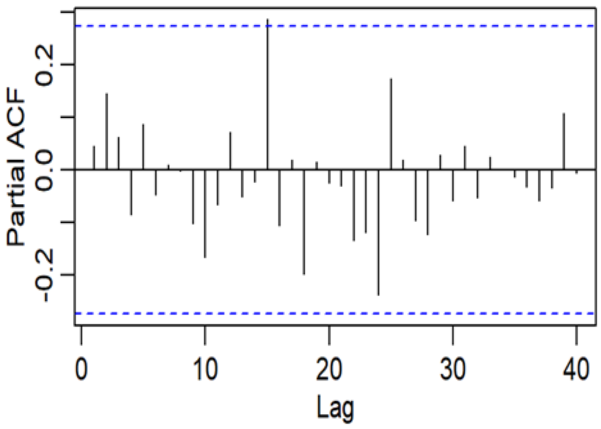




For the next step, I want to look at ACF and PACF plots again to finding Q, P, q, and p.

- **Women: Q, P, q, and p**





To find the non-seasonal part for the p, and q, we need to look at the lower lags, the ACF tails off at lag = 1, and the PACF cuts off. So, the non-seasonal part seems to be an AR (1). Thus, let us use:

**max.p = 1** and **max.q = 0**

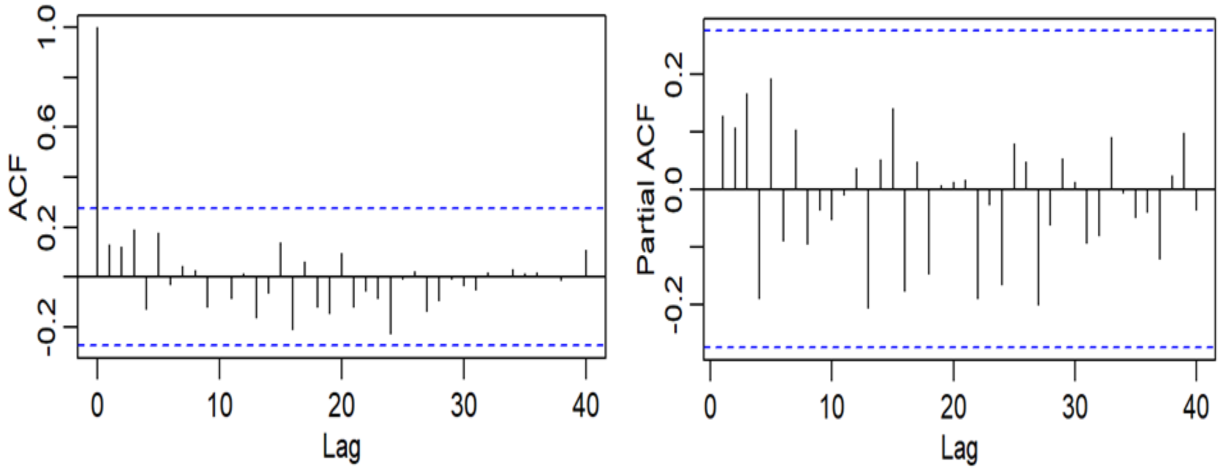
At the seasonal lags, the PACF cuts off at lag = 2, and lag = 3, and the ACF either cuts off at lag = 2 or tails off. So, let's us use:

**max.Q = 2 and max.P = 2**

Based on the differences that we have already computed; we would say that:

**max.d = 1 and max.D = 1**

- **Men: Q, P, q, and p**



To find the non-seasonal part for the p, and q, we need to look at the lower lags, the ACF tails off at lag = 1, and the PACF cuts off. So, the non-seasonal part seems to be an AR (1). Thus, let us use:

**max.p = 1 and max.q = 0**

At the seasonal lags, the PACF cuts off at lag = 2, and the ACF either cuts off at lag = 2 or tails off. So, let's us use:

**max.Q = 2 and max.P = 2**

Based on the differences that we have already computed; we would say that:

**max.d = 1 and max.D = 1**

## 13. BIC and AIC for model selection:

- **Women:**

```
> best.bic
[1] -242.4299
> best.fit

Call:
arima(x = x.ts, order = c(p, d, q), seasonal = list(order = c(P, D, Q), frequency(x.ts)),
    method = "CSS-ML")

Coefficients:
        sar1      sma1
      0.9991   -0.9431
s.e.  0.0058    0.1671

sigma^2 estimated as 0.0003789:  log likelihood = 127.14,   aic = -248.28

> best.model
[1] 0 0 0 1 1 1
```

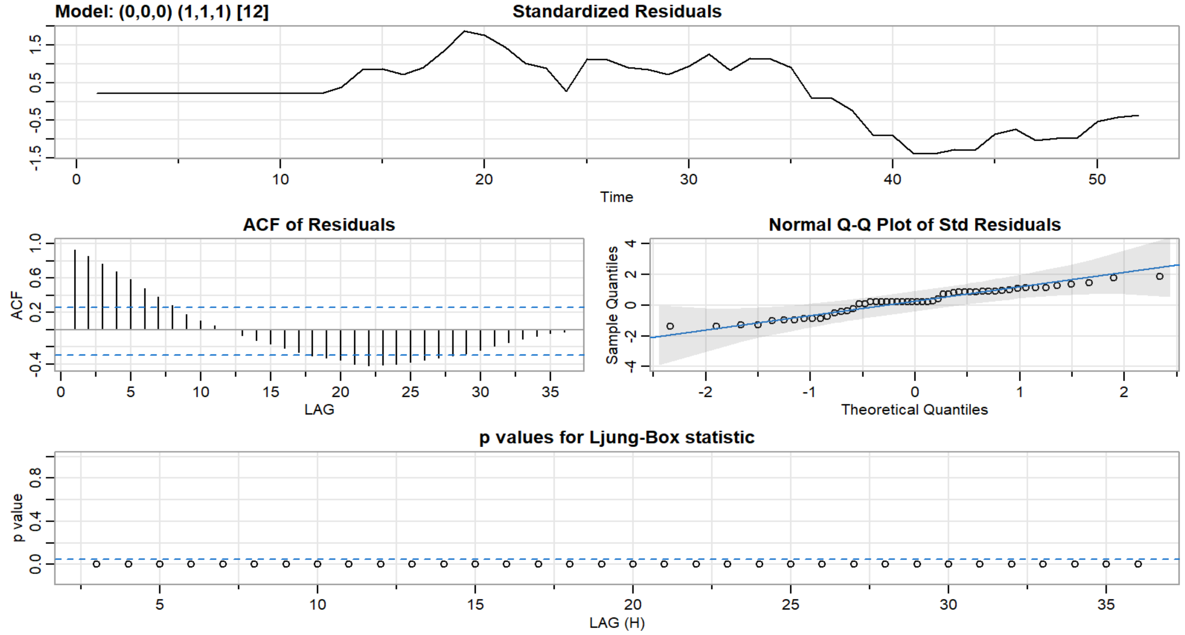
- **Men:**

```
> best.bic
[1] -267.4287
> best.fit

Call:
arima(x = x.ts, order = c(p, d, q), seasonal = list(order = c(P, D, Q), frequency(x.ts)),
    method = "CSS-ML")

Coefficients:
        sar1      sma1
      0.9911   -0.8041
s.e.  0.0134   0.1016

sigma^2 estimated as 0.0002362:  log likelihood = 139.64,   aic = -273.28

> best.model
[1] 0 0 0 1 1 1
```

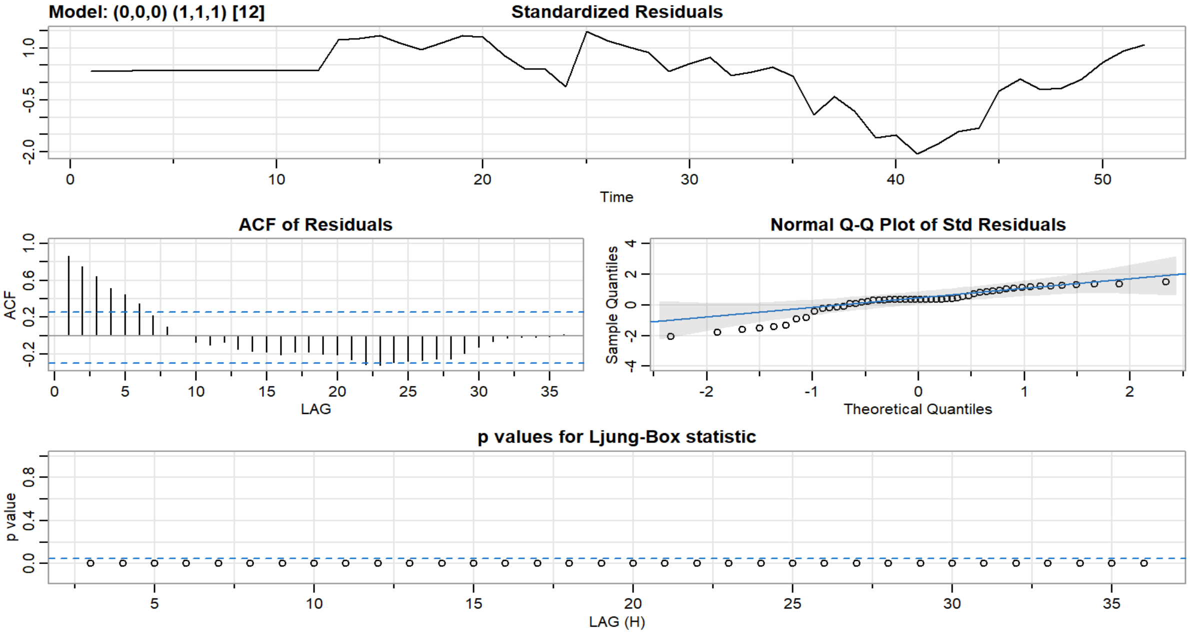
## 14. Perform model diagnostics for the ARIMA with the SARIMA function:

In this step, we are going to apply model diagnostics to observe how the residuals are in the model that we found the best model in previous step which is SARIMA (0,0,0,1,1,1).

- **Women:**



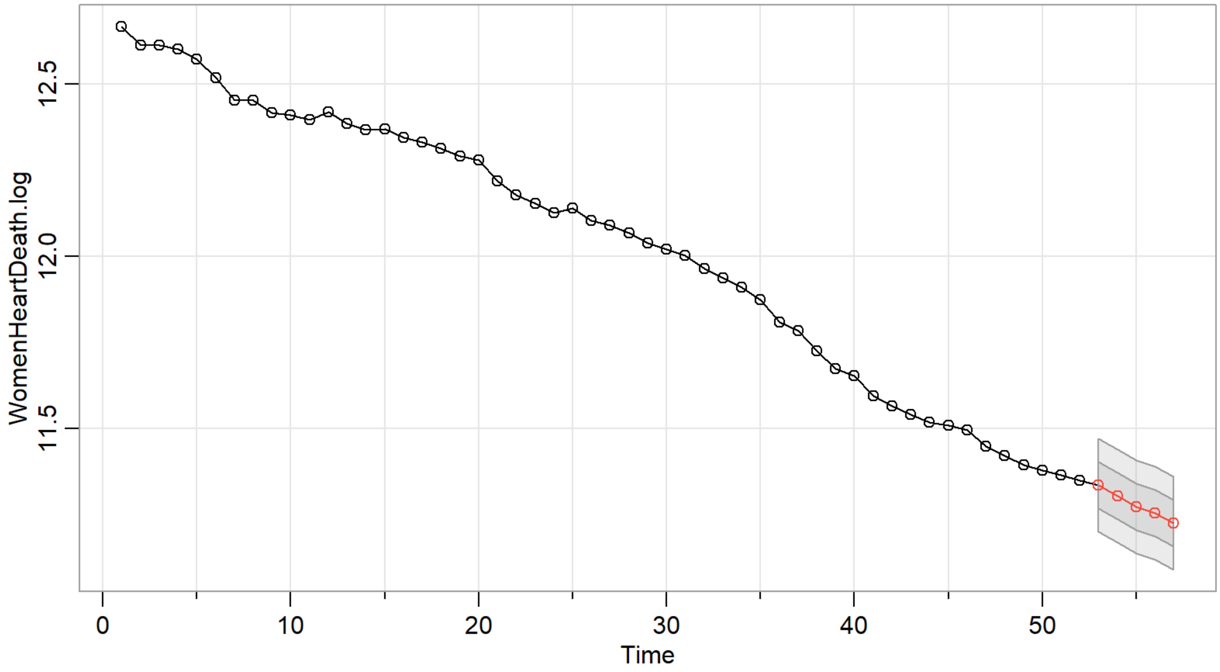
- **Men:**



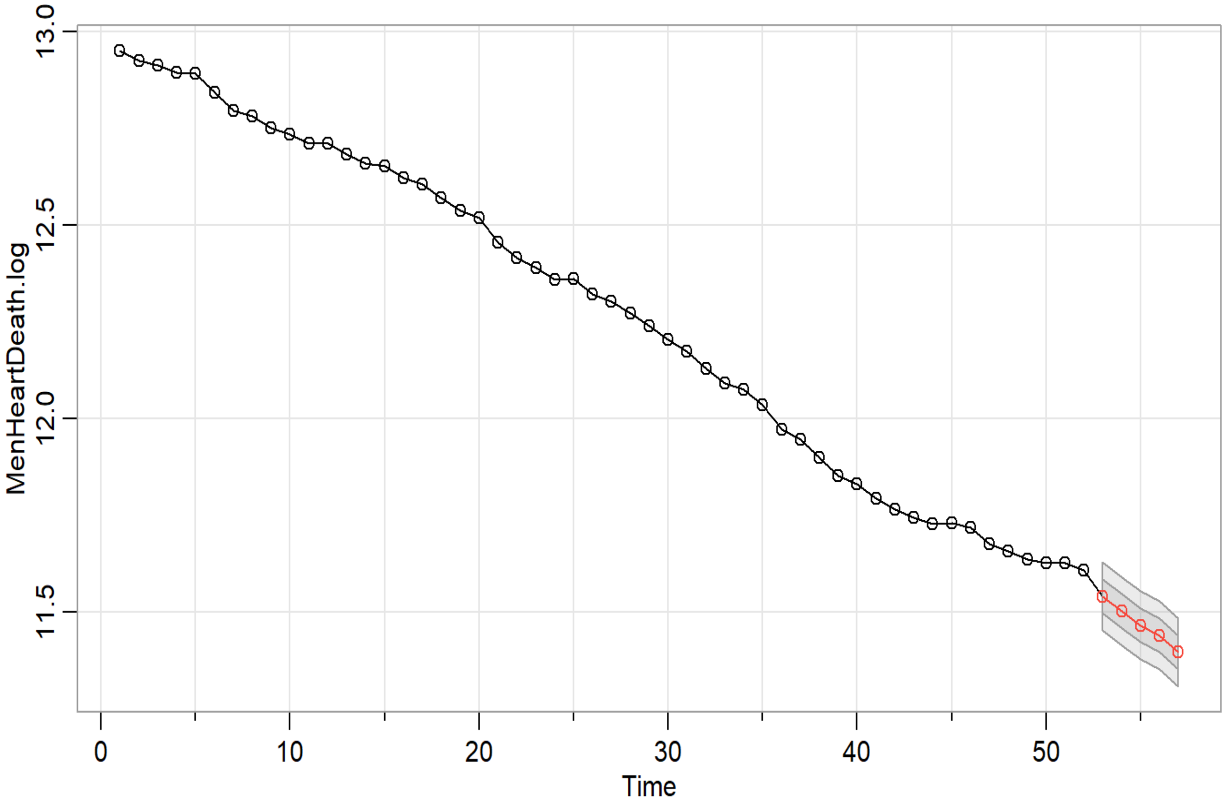
The ACF plot demonstrate that there are a few lags in which ACF is out of the boundary. Since we can see that the residuals have some autocorrelation. In addition, by looking at the QQ plot, the data are very close to the normal distribution, and we can tell that the assumption of normality of the residuals are reasonable. Moreover, all of the p-values of the Ljung-Box statistic are at or below the 5% significance threshold, indicating existence of autocorrelation in the residuals.

## 15. Perform forecasting for 5 years using the ARIMA model:

- **Women:**



- **Men:**



### 16. Predictions for next 5 years (2021-2025):

- **Women:**

```
       2021        2022        2023        2024        2025
[1] 83755.46    81162.30    78619.39    77201.13    74988.78
```

- **Men:**

```
        2021        2022        2023        2024        2025
[1] 102898.11    99010.73    95463.57    93034.44    89028.44
```

### 17. 95% prediction intervals:

- **Lower bounds/Women:**

```
        2021        2022        2023        2024        2025
[1] 73356.80   71085.59   68858.40   67616.22   65678.55
```

- **Upper bounds/Women:**

```
        2021        2022        2023        2024        2025

[1] 95628.18     92667.43   89764.05   88144.74     85618.79
```

- **Lower bounds/Men:**

```
        2021        2022        2023         2024         2025

[1] 94418.62     90851.59    87596.73   85367.78     81691.90
```

- **Upper bounds/Men:**

```
        2021        2022         2023         2024         2025

[1] 112139.13    107902.63   104036.91   101389.63    97023.86
```

**#Appendix:**

```
#Final project: Final Report time series Project time series analysis Fall22
#Elham-Nasarian
#US-heart-failure-cases(1969-2020)

library(astsa)
library(TTR)

#Read the data into R
USdeath <- read.csv("C:\\Users\\elhamn20\\Documents\\fall2022\\Applied time sries\\final project\\FINALREPORT\\USdeath.csv")
WomenHeartDeath <- USdeath$WomenHeartDeath
MenHeartDeath <- USdeath$MenHeartDeath
Year <- USdeath$Year

#plot the data
tsplot(Year, WomenHeartDeath)
tsplot(Year, MenHeartDeath)

#acf plot
acf(WomenHeartDeath, lag.max = 40)
acf(MenHeartDeath, lag.max = 40)

#pacf plot
pacf(WomenHeartDeath, lag.max = 40)
pacf(MenHeartDeath, lag.max = 40)

#Compute square root for removing trends
WomenHeartDeath.sqrt <- sqrt(WomenHeartDeath)
tsplot(Year, WomenHeartDeath.sqrt,type="l")
MenHeartDeath.sqrt <- sqrt(MenHeartDeath)
tsplot(Year, MenHeartDeath.sqrt,type="l")

#Compute log death
WomenHeartDeath.log <- log(WomenHeartDeath)
tsplot(Year, WomenHeartDeath.log)
MenHeartDeath.log <- log(MenHeartDeath)
tsplot(Year, MenHeartDeath.log)

#Estimate the trend of the logarithm transformed time series.
ma.WomenHeartDeath.log <- filter(WomenHeartDeath.log, sides=2, filter=rep(1/2,2))
detrended.WomenHeartDeath.log <- WomenHeartDeath.log - ma.WomenHeartDeath.log
plot(Year, ma.WomenHeartDeath.log,type="l", main="Trend (log(WomenHeartDeath))")
```

```
ma.MenHeartDeath.log <- filter(MenHeartDeath.log, sides=2, filter=rep(1/2,2))
detrended.MenHeartDeath.log <- MenHeartDeath.log - ma.MenHeartDeath.log
plot(Year, ma.MenHeartDeath.log,type="l", main="Trend (log(MenHeartDeath))")

# This series has a nonstationary trend.
# Let's compute the first-order difference to remove the trend.
WomenHeartDeath.diff.log <- diff(WomenHeartDeath.log)
tsplot( Year[1:51], WomenHeartDeath.diff.log)
MenHeartDeath.diff.log <- diff(MenHeartDeath.log)
tsplot( Year[1:51], MenHeartDeath.diff.log)

# Let us take a look at the sample ACF and sample PACF
acf(WomenHeartDeath.diff.log, lag.max = 40)
pacf(WomenHeartDeath.diff.log, lag.max = 40)
acf(MenHeartDeath.diff.log, lag.max = 40)
pacf(MenHeartDeath.diff.log, lag.max = 40)

# Compute AIC
ord.max = 15
WomenHeartDeath.diff.log.model <-
ar(WomenHeartDeath.diff.log,order.max=ord.max,aic=TRUE,method="mle")
WomenHeartDeath.diff.log.model$aic
ord.max = 13
MenHeartDeath.diff.log.model <-
ar(MenHeartDeath.diff.log,order.max=ord.max,aic=TRUE,method="mle")
MenHeartDeath.diff.log.model$aic

# Compute BIC
n = length(WomenHeartDeath.diff.log)
WomenHeartDeath.diff.log.BIC <- WomenHeartDeath.diff.log.model$aic - 2*((0:ord.max)+1) + log(n) *
((0:ord.max)+1)
WomenHeartDeath.diff.log.BIC       # Choose model with smallest BIC
# BIC = Bayesian Information Criterion
n = length(MenHeartDeath.diff.log)
MenHeartDeath.diff.log.BIC <- MenHeartDeath.diff.log.model$aic - 2*((0:ord.max)+1) + log(n) *
((0:ord.max)+1)
MenHeartDeath.diff.log.BIC       # Choose model with smallest BIC
# BIC = Bayesian Information Criterion

# Fit an AR(0) model with maximum likelihood
WomenHeartDeath.diff.log.model0 <-
ar(WomenHeartDeath.diff.log,order.max=0,aic=FALSE,method="mle")
WomenHeartDeath.diff.log.model0
MenHeartDeath.diff.log.model0 <- ar(MenHeartDeath.diff.log,order.max=0,aic=FALSE,method="mle")
MenHeartDeath.diff.log.model0

order = 0
MLE.fit = ar.mle(MenHeartDeath.diff.log, order=order, aic=FALSE)
MLE.fit$x.mean
MLE.fit$ar
sqrt(diag(MLE.fit$asy.var.coef))


# Now let's loop through several possible models to search for the best model
n = length(MenHeartDeath.log)
max.p = 1
max.d = 1
max.q = 0
max.P = 2
max.D = 1
max.Q = 2
BIC.array =array(NA,dim=c(max.p+1,max.d+1,max.q+1,max.P+1,max.D+1,max.Q+1))
AIC.array =array(NA,dim=c(max.p+1,max.d+1,max.q+1,max.P+1,max.D+1,max.Q+1))
best.bic <- 1e8
x.ts = MenHeartDeath.log
for (p in 0:max.p) for(d in 0:max.d) for(q in 0:max.q)
 for (P in 0:max.P) for(D in 0:max.D) for(Q in 0:max.Q)
 {
  # This is a modification of a function originally from the book:
  # Cowpertwait, P.S.P., Metcalfe, A.V. (2009), Introductory Time
  # Series with R, Springer.
  # Modified by M.A.R. Ferreira (2016, 2020).
  cat("p=",p,", d=",d,", q=",q,", P=",P,", D=",D,", Q=",Q,"\n")
  fit <- tryCatch(
   {  arima(x.ts, order = c(p,d,q),
         seas = list(order = c(P,D,Q),
               frequency(x.ts)),method="CSS-ML")
   },
   error = function(cond){
    message("Original error message:")
    message(cond)
    # Choose a return value in case of error
    return(NA)
   }
  )
  condition = !is.na(fit)
  if(length(condition)>1)condition = all(condition)
  if(condition) {
      number.parameters <- length(fit$coef) + 1
      BIC.array[p+1,d+1,q+1,P+1,D+1,Q+1] = -2*fit$loglik + log(n)*number.parameters
      AIC.array[p+1,d+1,q+1,P+1,D+1,Q+1] = -2*fit$loglik + 2*number.parameters
      if (BIC.array[p+1,d+1,q+1,P+1,D+1,Q+1] < best.bic)
      {
        best.bic <- BIC.array[p+1,d+1,q+1,P+1,D+1,Q+1]
        best.fit <- fit
        best.model <- c(p,d,q,P,D,Q)
      }
    }
  }
best.bic
best.fit
best.model


# Let's perform model diagnostics for the ARIMA
# with the sarima function:
sarima(WomenHeartDeath.log,0,0,0,1,1,1,12)
sarima(MenHeartDeath.log,0,0,0,1,1,1,12)


# Let's perform forecasting for 5 year using the ARIMA model:
WomenHeartDeath.log.for <- sarima.for(WomenHeartDeath.log,5,0,0,0,1,1,1,12)
MenHeartDeath.log.for <- sarima.for(MenHeartDeath.log,5,0,0,0,1,1,1,12)

# Here are predictions
exp(WomenHeartDeath.log.for$pred)
exp(MenHeartDeath.log.for$pred)

# Here are 95% prediction intervals
# Lower bounds
exp(WomenHeartDeath.log.for$pred - 1.96*WomenHeartDeath.log.for$se)
exp(MenHeartDeath.log.for$pred - 1.96*MenHeartDeath.log.for$se)

# Upper bounds
exp(WomenHeartDeath.log.for$pred + 1.96*WomenHeartDeath.log.for$se)
exp(MenHeartDeath.log.for$pred + 1.96*MenHeartDeath.log.for$se)
```