STAT 5814: APPLIED TIME SERIES ANALYSIS

# Midterm Exam 2 Report

*Submitted By:*
Sayantan Majumdar
smxnv@mst.edu

*Course Instructor:*
Dr. Akim Adekpedjou
akima@mst.edu

November 21, 2019

# Contents

# List of Figures

# 1 Problem 1

The monthly values of the average hourly for US apparel and textile workers for July 1981 to July 1987 are the file wages of the library TSA.

(a) Plot the series and write a few lines on what you observe.

(b) Fit a linear trend model using least squares. Give the plot of the linear trend and superpose it with that of the data. Give the estimated regression equation.

(c) Plot the standardized residuals from the linear regression versus time. Comments?

(d) Fit a quadratic time trend model using least squares. Give the plot of quadratic trend and superpose it Give the estimated regression equation. Plot the residuals and comment on any patterns.

(e) Perform a diagnostic check of the residuals. Comments.

(f) Plot the autocorrelation function for the standardized residuals from the quadratic regression.

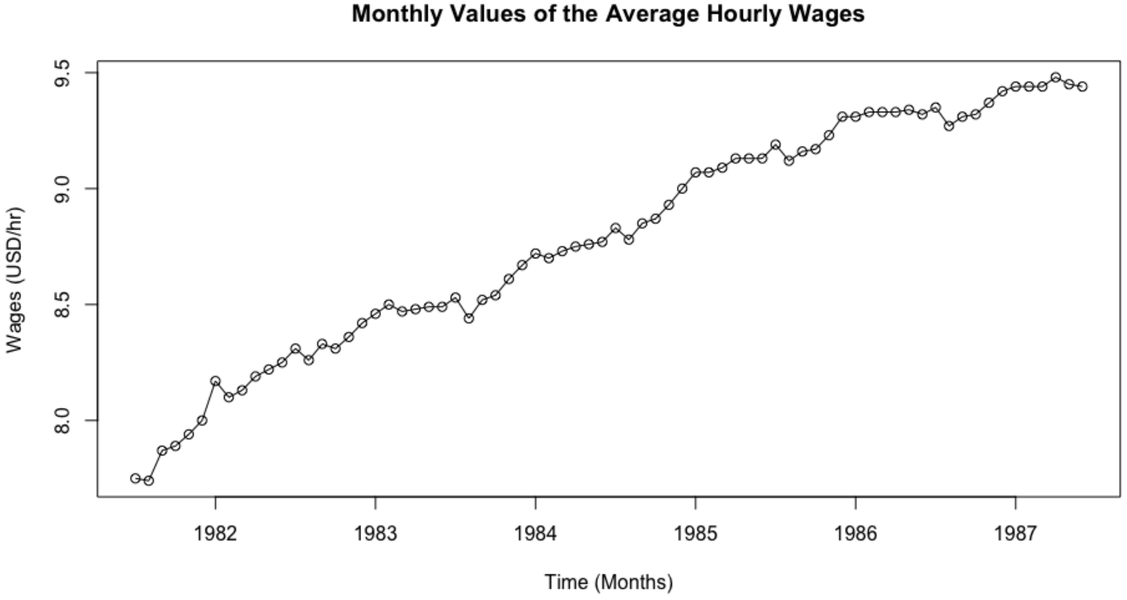
(g) Investigate the normality of the standardized residuals from the quadratic regression. Comment.

## 1.1 R Code

```r
### STAT 5814 MIDTERM 2/PROBLEM 1
### AUTHOR: SAYANTAN MAJUMDAR
### EMAIL: smxnv@mst.edu
### SNO: 12566087

library(TSA)
library(snpar)

#get data
data(wages)
par(mfrow = c(1, 1))
plot(wages, type='o', ylab='Wages (USD/hr)', xlab='Time (Months)', main='Monthly Values of the Average Hourly Wages')

#linear model
wages.lm = lm(wages~time(wages))
summary(wages.lm)
model = ts(wages.lm$fitted.values, frequency=12, start=c(1981, 7), end=c(1987, 6))
lines(model, col='red')
plot(y=rstandard(wages.lm), x=as.vector(time(wages)), type = 'o', ylab = 'Standardized Residuals', xlab = 'Time (Months)', main='Standardized Residual Plot')
par(mfrow = c(2, 2))
plot(wages.lm)
```

```

#Quadratic model trend
wages.qm = lm(wages ~ time(wages) + I(time(wages)^2))
summary(wages.qm)
model = ts(wages.qm$fitted.values, frequency=12, start=c(1981, 7), end=c
    (1987, 6))
par(mfrow=c(2, 2))
lines(model, col='red')
plot(y=rstandard(wages.qm), x=as.vector(time(wages)), type = 'o', ylab = '
    Standardized Residuals', xlab = 'Time (Months)', main='Standardized
    Residual Plot')
par(mfrow = c(2, 2))
plot(wages.qm)

#Residual Diagnostics
par(mfrow=c(1, 3))
hist(wages.qm$residuals, main="Residuals Histogram", xlab='Residuals')
qqnorm(wages.qm$residuals, main="QQ Plot for Residuals")
qqline(wages.qm$residuals, col="Red")
boxplot(wages.qm$residuals, main='Residual Boxplot', ylab='Residuals')

#ACF
acf(wages.qm$residuals, main='ACF of the Residuals')

#Residual Normality
shapiro.test(wages.qm$residuals)
runs.test(wages.qm$residuals, exact=TRUE)
```

## 1.2 Results



(a)

Figure 1.1: Time series plot of the wages data. This shows that the wages are mostly increasing and only deviating from this trend in small proportions.

(b) Linear trend model fit summary:

```
Call:
lm(formula = wages ~ time(wages))

Residuals:
     Min       1Q   Median       3Q      Max
-0.23828 -0.04981  0.01942  0.05845  0.13136

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.490e+02  1.115e+01  -49.24   <2e-16 ***
time(wages)  2.811e-01  5.618e-03   50.03   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08257 on 70 degrees of freedom
Multiple R-squared:  0.9728,Adjusted R-squared:  0.9724
F-statistic:  2503 on 1 and 70 DF,  p-value: < 2.2e-16
```

The $\text{R}^2$ value of this fit is very good ($\approx 0.97$) which signifies that the linear fit is working well. The estimated regression equation is given by Eq (1).

$$\widehat{X_t} = -549 + 0.2811 X_t + \varepsilon \tag{1}$$

where $\varepsilon \sim \mathcal{N}(0, 0.08257)$

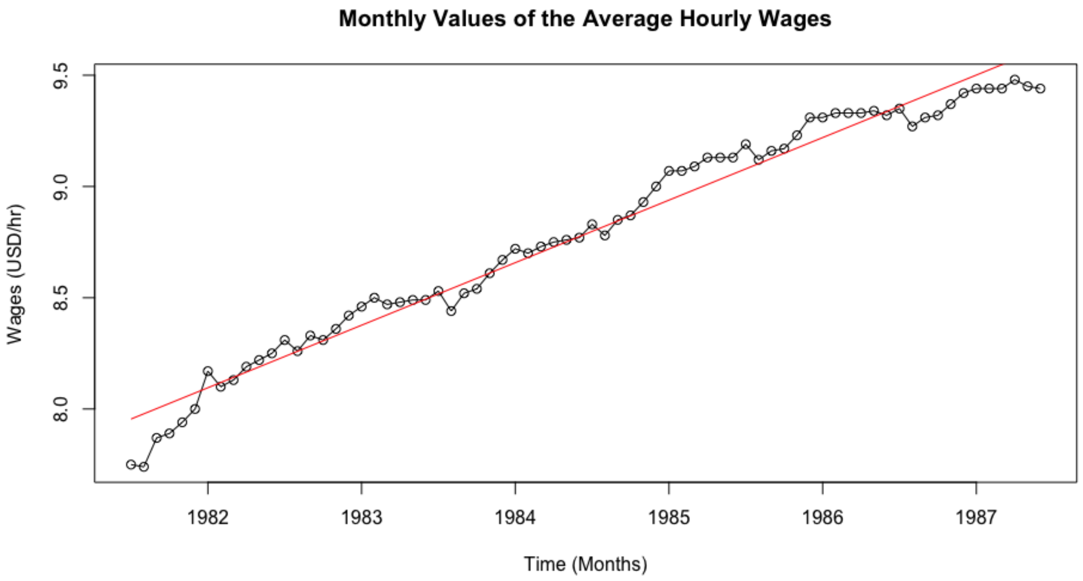


Figure 1.2: Wages plot showing the linear trend in red.

(c)

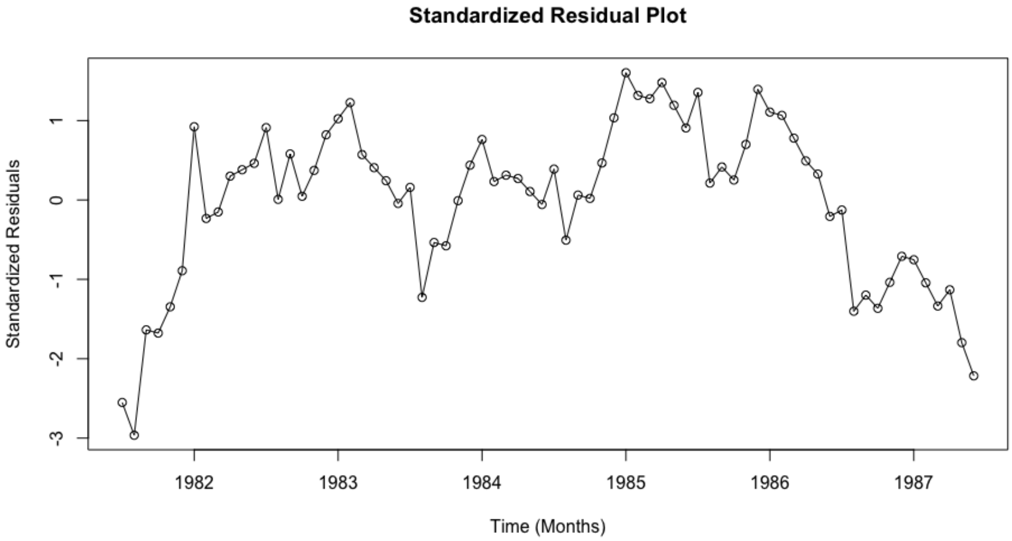


Figure 1.3: Standardized residuals from the linear regression versus time. Since the majority of the standardized residuals are within the [-2, 1] interval, so the model residuals strongly follow a normal distribution. Hence, we have a good fit.

(d) Quadratic time trend model summary:

```
Call:
lm(formula = wages ~ time(wages) + I(time(wages)^2))

Residuals:
      Min        1Q    Median        3Q       Max
-0.148318 -0.041440  0.001563  0.050089  0.139839

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -8.495e+04  1.019e+04  -8.336 4.87e-12 ***
time(wages)       8.534e+01  1.027e+01   8.309 5.44e-12 ***
I(time(wages)^2) -2.143e-02  2.588e-03  -8.282 6.10e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05889 on 69 degrees of freedom
Multiple R-squared:  0.9864,Adjusted R-squared:  0.986
F-statistic:  2494 on 2 and 69 DF,  p-value: < 2.2e-16
```

The R$^2$ value of this fit is even better ($\approx 0.98$) when compared to the linear fit earlier. This implies that the quadratic trend is more appropriate. The estimated regression equation is given by Eq (2).

$$\widehat{X_t} = -84950 + 85.34X_t - 0.02143X_t^2 + \varepsilon \tag{2}$$

where $\varepsilon \sim \mathcal{N}(0, 0.05889)$

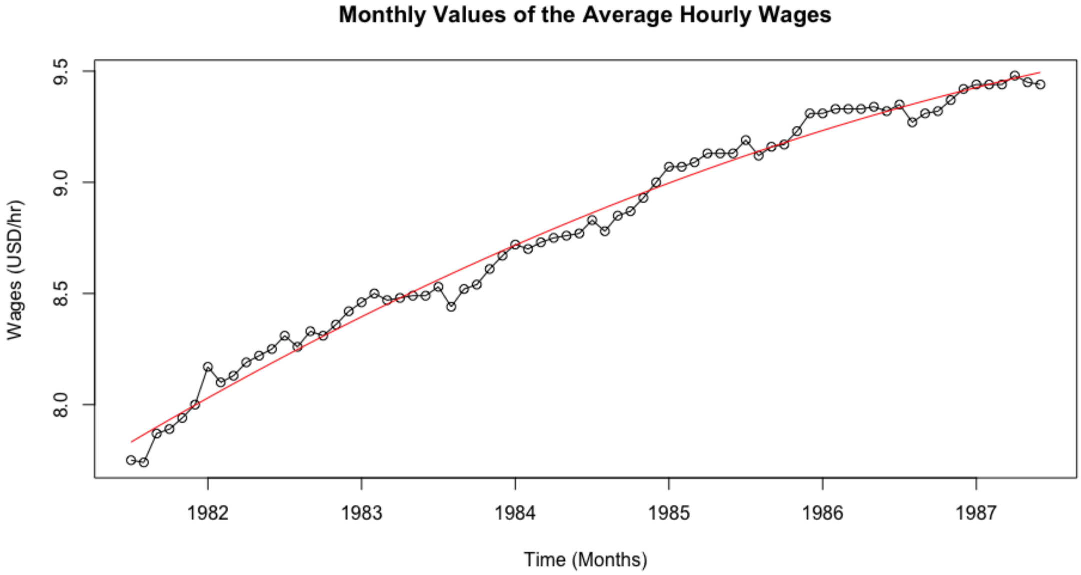


Figure 1.4: Wages plot showing the quadratic fit in red.

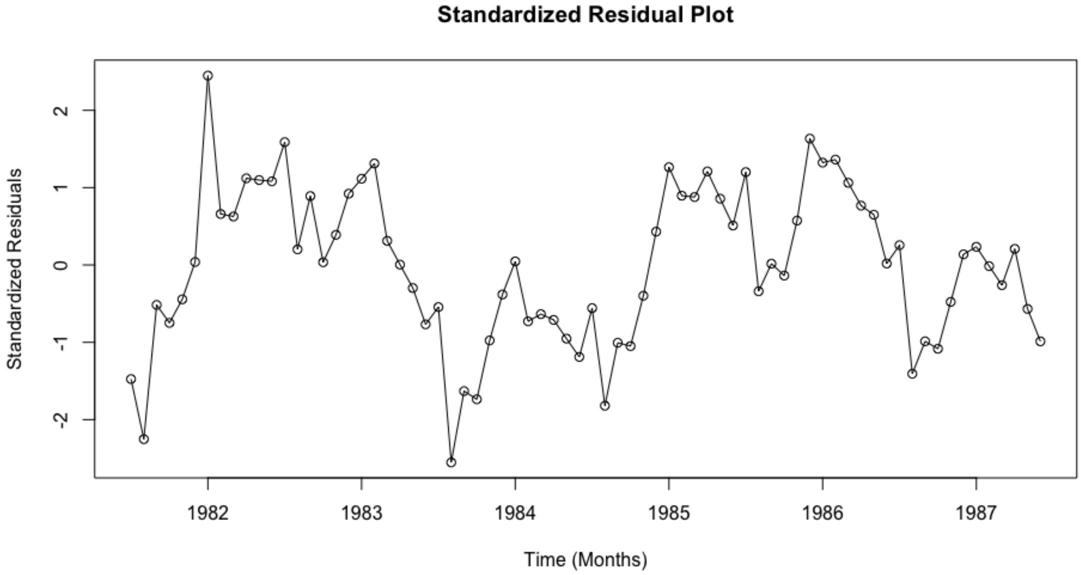


Figure 1.5: Standardized residual plot for the quadratic fit. Here, we observe that these most of these residuals are in the ±2 region. So, the quadratic fit is better than the linear one.

(e)

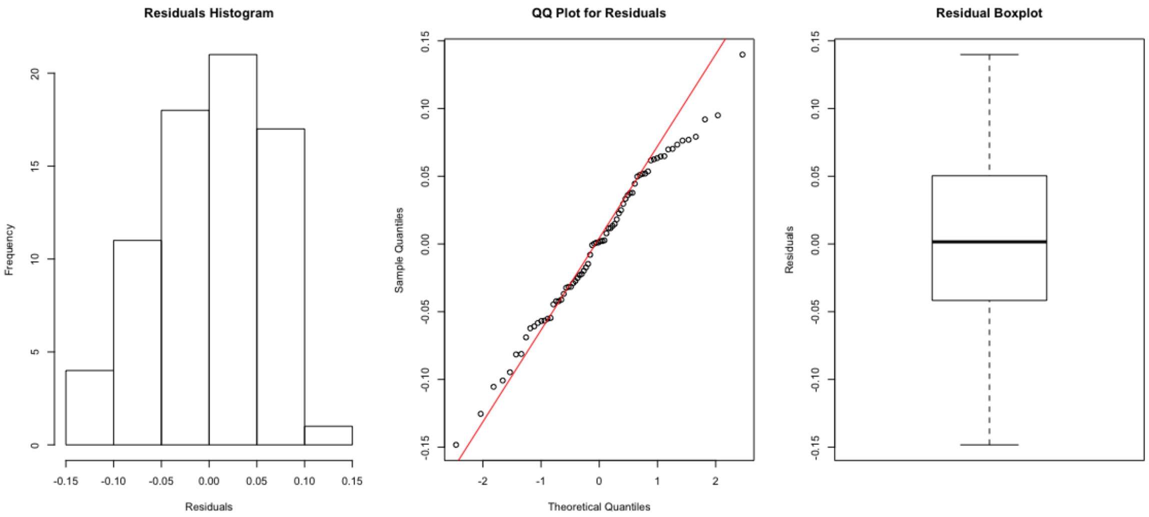


Figure 1.6: Residual diagnostics of the quadratic model showing residual histogram, qq-plot, and boxplot. These plots show that the residuals closely follow a normal distribution.

(f)

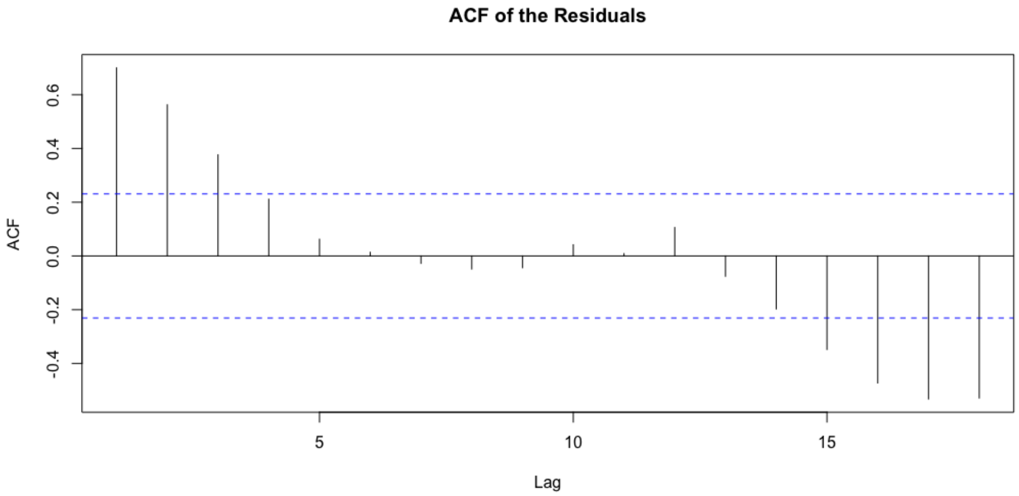


Figure 1.7: Autocorrelation function (ACF) for the standardized residuals from the quadratic regression. This shows that the ACF quickly degrades to zero as the lag increases.

(g) Residual normality checks:

```
Shapiro-Wilk normality test
data:  wages.qm$residuals
W = 0.98856, p-value = 0.7622

Exact runs test
data:  wages.qm$residuals
```

```
Runs = 15, p-value = 1.284e-07
alternative hypothesis: two.sided
```

The Shapiro-Wilk test shows that the residuals are normally distributed (p-value > 0.05 and W is close to 1). However, the runs test suggests that the order of the residuals is not random.

# 2 Problem 2

A data set of 57 consecutive measurements from a machine tool are in the TSA package.

(a) Estimate the parameters of a (mean-centered) AR(1) model for this series. Use the least squares method and maximum likelihood, and report the estimated parameters from each of these methods. Comment on any similarities and differences. Give the confidence intervals of your parameters.

(b) Estimate the parameters of a (mean-centered) AR(2) model for this series. Use the least squares method and maximum likelihood, and report the estimated parameters from each of these methods. Comment on any similarities and differences.

(c) Derive the confidence intervals for the parameters $\phi_1$ and $\phi_2$. Does the confidence interval for $\phi_2$ suggests it should be included in the model?

(d) Compare the results of the maximum likelihood fits from parts (a) and (b). Which model do you believe is preferable? Briefly explain your answer.

(e) Perform a diagnostic check of the residuals. Comments.

(f) Compare the two models using AIC, BIC. Which one do you prefer? Does any of the quantities suggests one model is better than the other? If so, how much do you gain in errors reduction if any. Is that in line with your answer in (d)?

## 2.1 R Code

```r
### STAT 5814 MIDTERM 2/PROBLEM 2
### AUTHOR: SAYANTAN MAJUMDAR
### EMAIL: smxnv@mst.edu
### SNO: 12566087

library(TSA)
library(snpar)
library(stats)
library(forecast)

data("deere3")
par(mfrow = c(1, 1))
plot(deere3, type='o', ylab='Measurements', xlab='Time', main='Deere3 Data Plot')
```

```r

# Conditional least square estimate for AR(1) model
css_model_ar1 = arima(deere3, order=c(1, 0, 0), method='CSS')
css_model_ar1_coef = coef(css_model_ar1)
print(css_model_ar1_coef)
confint(css_model_ar1)


# Maximum Likelihood estimate for AR(1) model
ml_model_ar1 = arima(deere3, order=c(1, 0, 0), method='ML')
ml_model_ar1_coef = coef(ml_model_ar1)
print(ml_model_ar1_coef)
confint(ml_model_ar1)

# Conditional least square estimate for AR(2) model
css_model_ar2 = arima(deere3, order=c(2, 0, 0), method='CSS')
css_model_ar2_coef = coef(css_model_ar2)
print(css_model_ar2_coef)
confint(css_model_ar2)

# Maximum Likelihood estimate for AR(2) model
ml_model_ar2 = arima(deere3, order=c(2, 0, 0), method='ML')
ml_model_ar2_coef = coef(ml_model_ar2)
print(ml_model_ar2_coef)
confint(ml_model_ar2)

# Residual Diagnostics (manually repeated for all the models)
par(mfrow=c(3, 2))
model = ml_model_ar2
st_residuals = rstandard(model)
plot(y=st_residuals, x=as.vector(time(deere3)), type = 'o', ylab = 'Standardized Residuals', xlab = 'Time', main='Standardized Residual Plot')
hist(st_residuals, main="Residuals Histogram", xlab='Residuals')
qqnorm(st_residuals, main="QQ Plot for Residuals")
qqline(st_residuals, col="Red")
boxplot(st_residuals, main='Residual Boxplot', ylab='Residuals')
acf(st_residuals, main='ACF Plot')
#Residual Normality
shapiro.test(st_residuals)
runs.test(st_residuals, exact=TRUE)

#AIC and BIC tests
AIC(model)
BIC(model)

#Confirmation
auto_arima = auto.arima(deere3)
auto_arima
```

## 2.2 Results

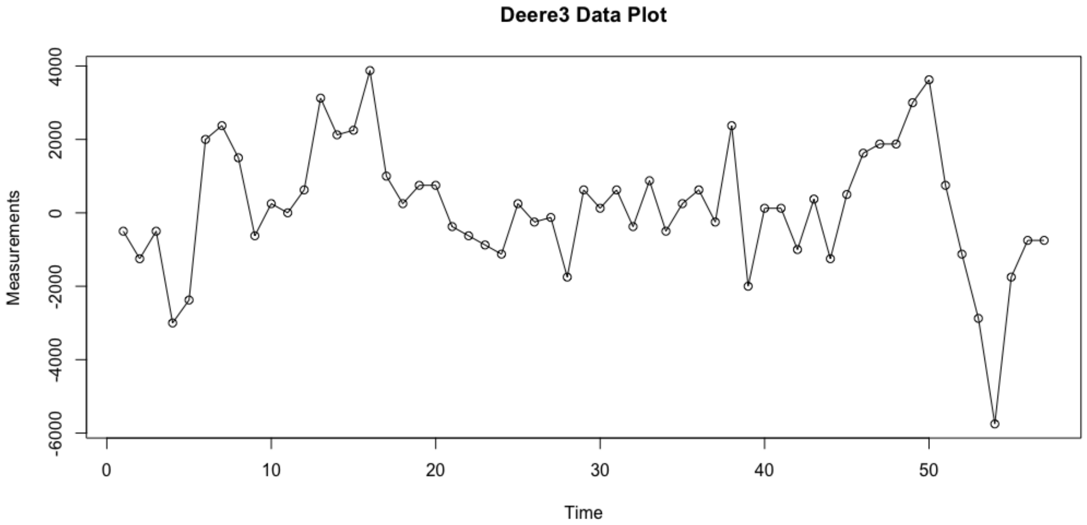


Figure 2.1: Time series plot of the Deere3 data.

(a) Least squares method for AR(1) parameter estimation:

```
> print(css_model_ar1_coef)
        ar1   intercept
  0.5332044 160.0797248
> confint(css_model_ar1)
                 2.5 %      97.5 %
ar1          0.3132122   0.7531966
intercept -646.0111173 966.1705670
```

Maximum Likelihood method for AR(1) parameter estimation:

```
> print(ml_model_ar1_coef)
        ar1   intercept
  0.5255778 124.3524257
> confint(ml_model_ar1)
                 2.5 %      97.5 %
ar1          0.3084104   0.7427452
intercept -648.3281895 897.0330409
```

The estimated $\hat{\phi}$ from both the least squares and maximum likelihood methods are very similar. However, the intercept ($\hat{\mu}$) is significantly different. In addition, the confidence intervals for $\hat{\phi}$ in both cases are similar and $0 \notin \text{CI}(\hat{\phi})$.

(b) Least squares method for AR(2) parameter estimation:

```
> print(css_model_ar2_coef)
         ar1          ar2    intercept
5.245848e-01 7.938728e-03 2.011876e+02
```

Maximum Likelihood method for AR(2) parameter estimation:

```
> print(ml_model_ar2_coef)
         ar1          ar2    intercept
  0.52111096   0.00830321 123.24182209
```

Like before, the estimated $\phi_1$ and $\phi_2$ in both cases are pretty similar, whereas, the intercepts are significantly different.

(c) Confidence intervals for $\phi_1$ and $\phi_2$:

```
Least Squares Method
> confint(css_model_ar2)
                 2.5 %       97.5 %
ar1          0.2660503    0.7831192
ar2         -0.2509164    0.2667939
intercept -607.0745845 1009.4498459

Maximum Likelihood Method
> confint(ml_model_ar2)
                 2.5 %      97.5 %
ar1          0.2642641   0.7779579
ar2         -0.2493401   0.2659465
intercept -656.0388502 902.5224944
```

Since $0 \in \text{CI}(\phi_2)$, therefore, $\phi_2 \approx 0$. So $\phi_2$ should not be included in the model.

(d) The maximum likelihood estimates of the intercepts for the AR(1) and AR(2) models are pretty similar. Also, the confidence interval of the intercept in case of AR(1) is a subset of that of AR(2). But, since our target is develop a parsimonious model and $0 \in \text{CI}(\phi_2)$ for the maximum likelihood method as well, so, AR(1) is the preferable model.

(e) Residual Diagnostics: The residual normality test results are given below.

```
Least Squares Estimation for AR(1):

Shapiro-Wilk normality test
data:  st_residuals
W = 0.98297, p-value = 0.6
```

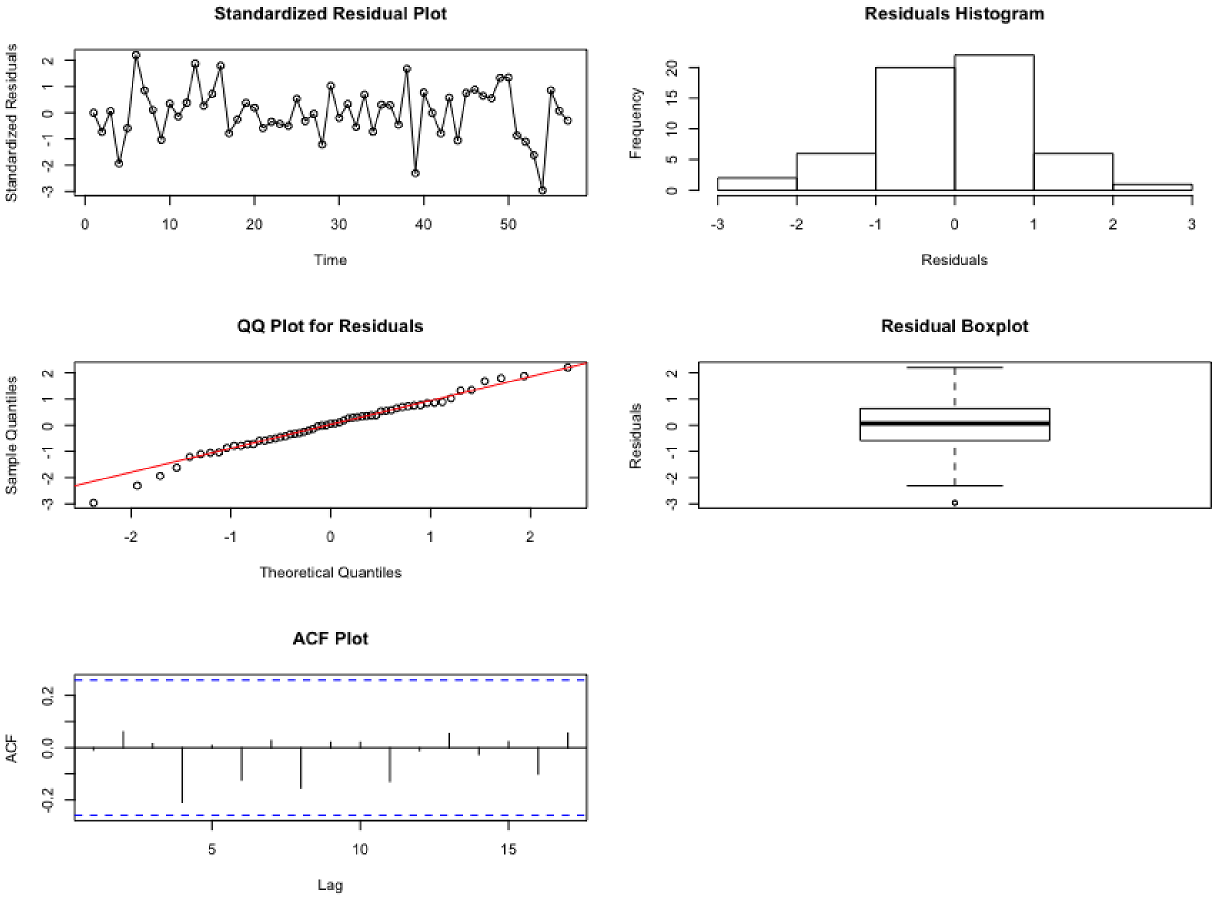


Figure 2.2: Residual diagnostic plots for least squares estimation for AR(1). These plots shows that the residuals tend to closely follow a normal distribution.

```
Exact runs test
data:  st_residuals
Runs = 29, p-value = 1

Maximum Likelihood Estimation for AR(1):

Shapiro-Wilk normality test
data:  st_residuals
W = 0.98261, p-value = 0.5827

Exact runs test
data:  st_residuals
Runs = 29, p-value = 1

Least Squares Estimation for AR(2):

Shapiro-Wilk normality test
```

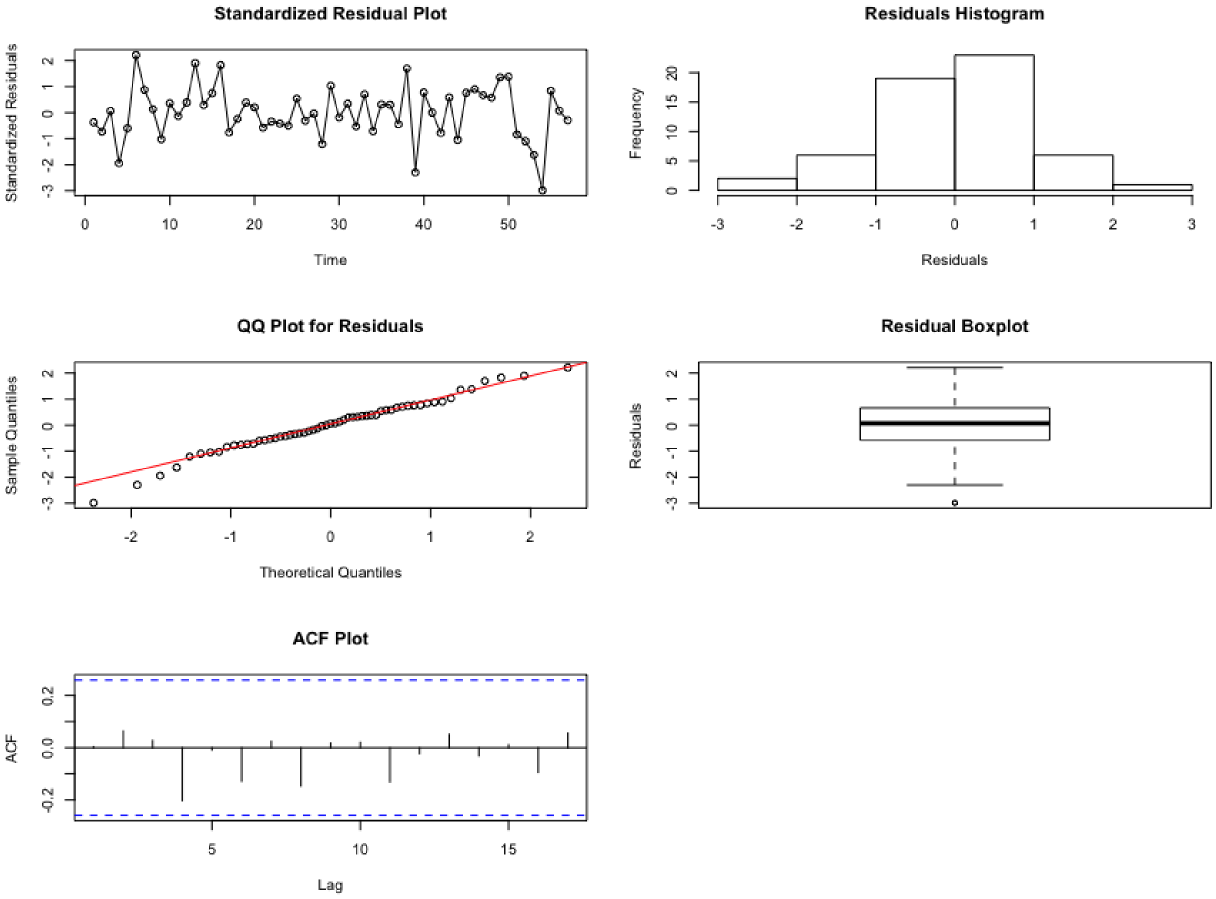


Figure 2.3: Residual diagnostic plots for maximum likelihood estimation for AR(1). Similar observations like that of Fig 2.2 can be made.

```
data:  st_residuals
W = 0.9809, p-value = 0.5028

Exact runs test
data:  st_residuals
Runs = 29, p-value = 1

Maximum Likelihood Estimation for AR(2):

Shapiro-Wilk normality test
data:  st_residuals
W = 0.98294, p-value = 0.599

Exact runs test
data:  st_residuals
Runs = 29, p-value = 1
```

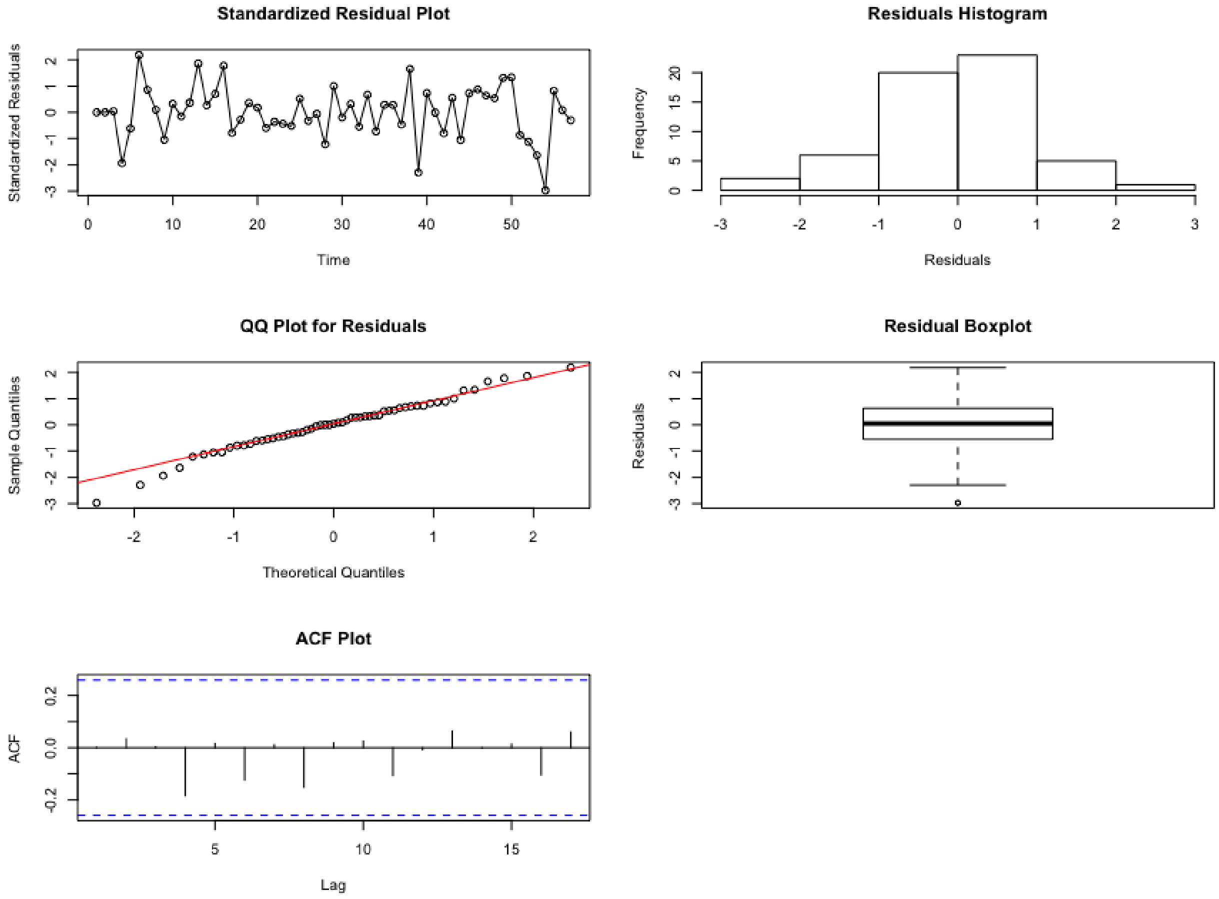


Figure 2.4: Residual diagnostic plots for least squares estimation for AR(2). Like before, these also closely follow a normal distribution but less than AR(1) residuals.

The normality tests show that the residuals from the AR(1) tend to follow the normal distribution more closely than those of AR(2).

(f) AIC and BIC tests for the maximum likelihood method:

```
AR(1):
> AIC(model)
[1] 997.0189
> BIC(model)
[1] 1003.148

AR(2):
> AIC(model)
[1] 999.0148
> BIC(model)
[1] 1007.187
```

As observed, we obtain a reduction of $\approx 2$ and $\approx 4$ in the AIC and BIC values

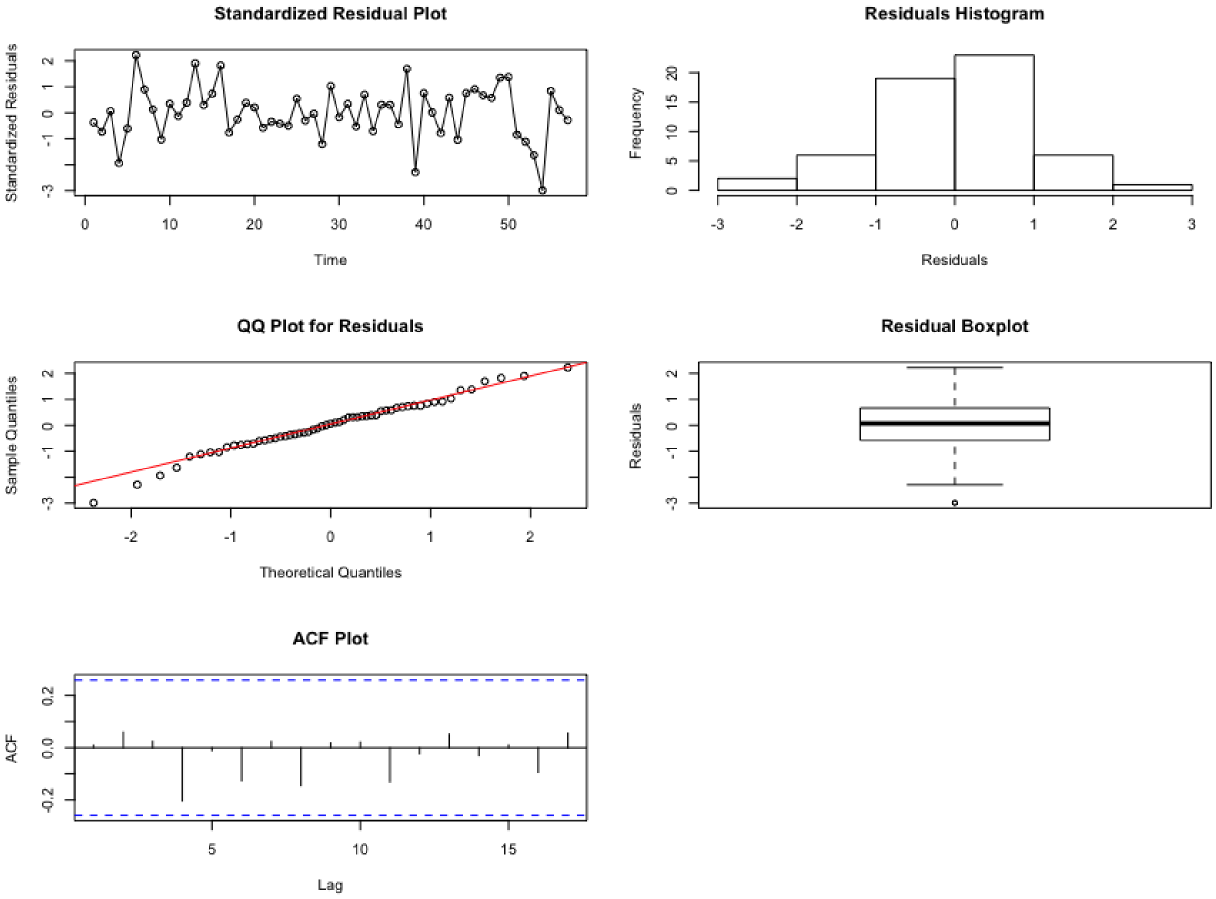


Figure 2.5: Residual diagnostic plots for maximum likelihood estimation for AR(2). Similar observations like that of Fig 2.4 can be made.

respectively for the AR(1) model. Although, these are not significant reductions, our aim is always to develop a parsimonious model which aligns with the results obtained in (d). Therefore, AR(1) is the preferred model which is further confirmed using the auto.arima function in R.

```
> auto_arima
Series: deere3
ARIMA(1,0,0) with zero mean

Coefficients:
         ar1
      0.5291
s.e.  0.1103

sigma^2 estimated as 2109761:  log likelihood=-495.56
AIC=995.12   AICc=995.34   BIC=999.2
```

# 3 Problem 3

Are sales trends for two industrial products related? The two time series objects described below contain Sales of chemicals and allied products; and Sales of motor vehicles and parts, in the U.S. for each month from Jan. 1971 to Dec. 1991. File is petr.txt.

You should analyze the data in the ”chemicals” time series and the ”vehicles” time series and write a report to address such questions as:

(i) What trend model(s) best capture the trends in sales of chemicals over time?

(ii) What trend model(s) best capture the trends in sales of vehicles over time? Once the trend has been accounted for, what can you say about the behavior of the detrended data, for both models? Can either the original time series or the detrended series be described using any common models?

(iii) For the various models you tried, assess their fit. Can any transformations improve the fit? Is there any apparent association between chemical sales and vehicle sales over time? If so, describe the association.

(iv) Make conclusions that relate to how the sales amounts (for both chemicals and vehicles) change, both long-term over the observed period of years, and in terms of patterns of month- to-month variation.

## 3.1 R Code

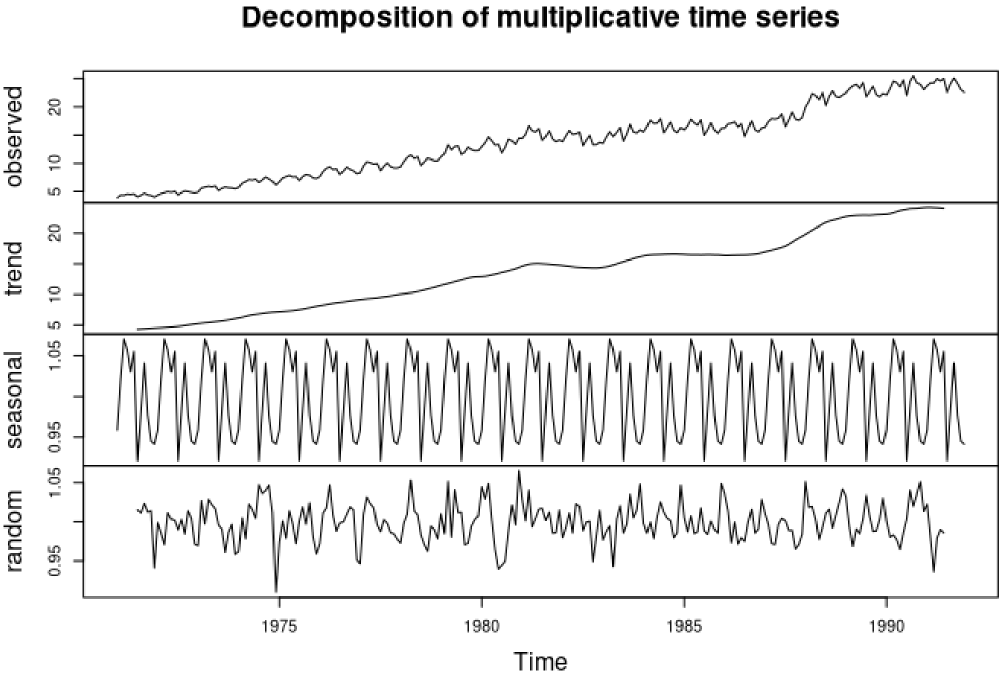
```r
### STAT 5814 MIDTERM 2/PROBLEM 3
### AUTHOR: SAYANTAN MAJUMDAR
### EMAIL: smxnv@mst.edu
### SNO: 12566087

library(TSA)
library(snpar)
library(stats)
library(forecast)
library(tseries)
library(fUnitRoots)

data = read.table('Data/ExamData/petr.txt', header=T)
chemicals.data = ts(data$Chemicals, frequency=12, start=c(1971, 1), end=c
    (1991, 12))
vehicles.data = ts(data$Vehicles, frequency=12, start=c(1971, 1), end=c
    (1991, 12))
plot(chemicals.data, main='Time Series Plot of Chemicals Data')
plot(vehicles.data, main='Time Series Plot of Vehicles Data')
chemicals.decompose = decompose(chemicals.data, type="mult")
plot(chemicals.decompose)
vehicles.decompose = decompose(vehicles.data, type="mult")
plot(vehicles.decompose)

```

```r
# Stationarity Check
adf.test(chemicals.data)
adf.test(vehicles.data)

par(mfrow=c(2, 2))
# Differencing to remove linear trend
chemicals.diff= diff(chemicals.data)
plot(chemicals.diff, main=’Differencing=1: Chemicals Data’)

# Differencing at lag 12 (monthly data) to remove seasonality
chemicals.diff2 = diff(chemicals.diff, lag=12)
plot(chemicals.diff2, main=’Lag=12: Chemicals Data’)
# Stationarity Check
adf.test(chemicals.diff2)

# Differencing to remove linear trend
vehicles.diff= diff(vehicles.data)
plot(vehicles.diff, main=’Differencing=1: Vehicles Data’)

# Differencing at lag 12 (monthly data) to remove seasonality
vehicles.diff2 = diff(vehicles.diff, lag=12)
plot(vehicles.diff2, main=’Lag=12: Vehicles Data’)
# Stationarity Check
adf.test(vehicles.diff2)

par(mfrow=c(2, 1))
# Linear Trend and Harmonic (Seasonal) Trend for Chemicals Data
har = harmonic(chemicals.data, 1)
chemicals.fit = lm(chemicals.data ~ har + time(chemicals.data))
chemicals.residuals = rstudent(chemicals.fit)
summary(chemicals.fit)
chemicals.model = ts(chemicals.fit$fitted.values, frequency=12, start=c(1971, 1), end=c(1991, 12))

plot(chemicals.data, main=’TS Plot: Chemicals Data’)
lines(chemicals.model, col=’red’)
legend(1975, 25, legend=c("Actual", "Fitted"), col=c("Black", "Red"), lty=1:1, cex=0.5)

# Checking Transformation for Chemicals Data
chemicals.bc.lambda = BoxCox.lambda(chemicals.data)
chemicals.bc.tr = BoxCox(chemicals.data, lambda=chemicals.bc.lambda)
har = harmonic(chemicals.bc.tr, 1)
chemicals.bc.fit = lm(chemicals.bc.tr ~ (har + time(chemicals.bc.tr)))
chemicals.bc.residuals = rstudent(chemicals.bc.fit)
summary(chemicals.bc.fit)
chemicals.bc.model = ts(chemicals.bc.fit$fitted.values, frequency=12, start=c(1971, 1), end=c(1991, 12))

plot(chemicals.bc.tr, main=’TS Plot: Transformed Chemicals Data’)
lines(chemicals.bc.model, col=’red’)
legend(1975, 4.2, legend=c("Actual", "Fitted"), col=c("Black", "Red"), lty=1:1, cex=0.5)

```

```r
# Linear Trend and Harmonic (Seasonal) Trend for Vehicles Data
har = harmonic(vehicles.data, 1)
vehicles.fit = lm(vehicles.data ~ har + time(vehicles.data))
vehicles.residuals = rstudent(vehicles.fit)
summary(vehicles.fit)
vehicles.model = ts(vehicles.fit$fitted.values, frequency=12, start=c(1971, 1), end=c(1991, 12))

plot(vehicles.data, main='TS Plot: Vehicles Data')
lines(vehicles.model, col='red')
legend(1975, 20, legend=c("Actual", "Fitted"), col=c("Black", "Red"), lty=1:1, cex=0.5)

# Checking Transformation for Vehicles Data
vehicles.bc.lambda = BoxCox.lambda(vehicles.data)
vehicles.bc.tr = BoxCox(vehicles.data, lambda=vehicles.bc.lambda)
har = harmonic(vehicles.bc.tr, 1)
vehicles.bc.fit = lm(vehicles.bc.tr ~ (har + time(vehicles.bc.tr)))
vehicles.bc.residuals = rstudent(vehicles.bc.fit)
summary(vehicles.bc.fit)
vehicles.bc.model = ts(vehicles.bc.fit$fitted.values, frequency=12, start=c(1971, 1), end=c(1991, 12))

plot(vehicles.bc.tr, main='TS Plot: Transformed Vehicles Data')
lines(vehicles.bc.model, col='red')
legend(1975, 3, legend=c("Actual", "Fitted"), col=c("Black", "Red"), lty=1:1, cex=0.5)

# Differencing after BoxCox Chemicals Data
chemicals.bc.diff= diff(chemicals.bc.tr)
plot(chemicals.bc.diff, main='Differencing=1: Transformed Chemicals Data')

# Differencing at lag 12 (monthly data) to remove seasonality
chemicals.bc.diff2 = diff(chemicals.bc.diff, lag=12)
plot(chemicals.bc.diff2, main='Lag=12: Transformed Chemicals Data')
# Stationarity Check
adf.test(chemicals.bc.diff2)

# Differencing after BoxCox Vehicles Data
vehicles.bc.diff= diff(vehicles.bc.tr)
plot(vehicles.bc.diff, main='Differencing=1: Transformed Vehicles Data')

# Differencing at lag 12 (monthly data) to remove seasonality
vehicles.bc.diff2 = diff(vehicles.bc.diff, lag=12)
plot(vehicles.bc.diff2, main='Lag=12: Transformed Vehicles Data')
# Stationarity Check
adf.test(vehicles.bc.diff2)

par(mfrow=c(2, 3))
# ACF, PACF, and EACF for Detrended Chemicals Data
acf(chemicals.diff2, main="ACF: Detrended Chemicals Data")
pacf(chemicals.diff2, main="PACF: Detrended Chemicals Data")
eacf(chemicals.diff2)
plot(armasubsets(chemicals.diff2, 5, 5))
```

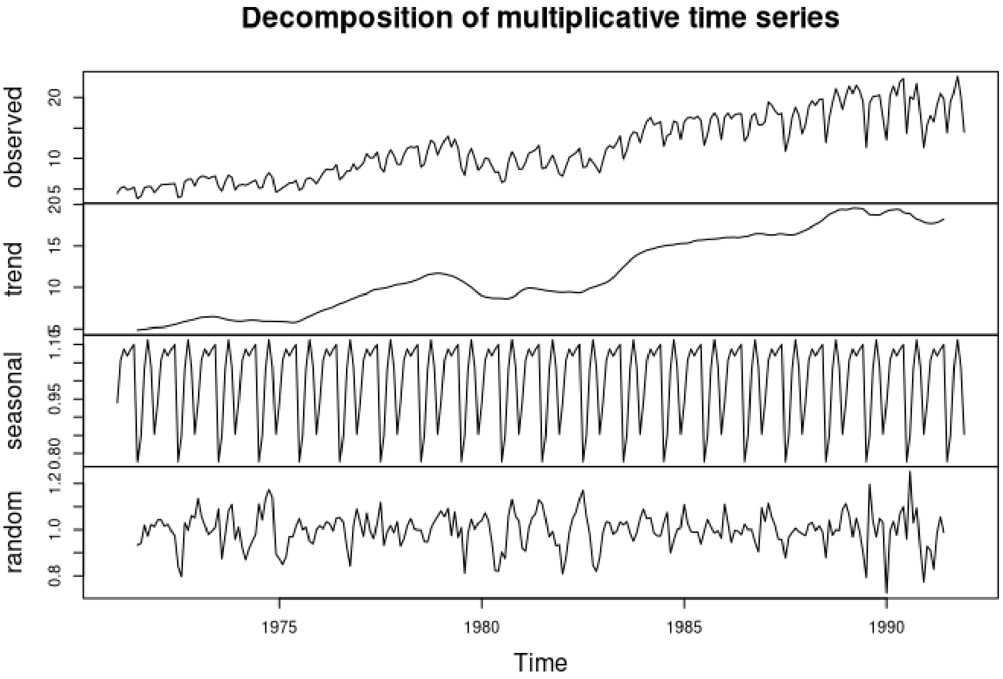
```r

# ACF, PACF, and EACF for Detrended Vehicles Data
acf(vehicles.diff2, main="ACF: Detrended Vehicles Data")
pacf(vehicles.diff2, main="PACF: Detrended Chemicals Data")
eacf(vehicles.diff2)
plot(armasubsets(vehicles.diff2, 5, 5))

# ACF, PACF, and EACF for Stabilized Chemicals Data
acf(chemicals.bc.diff2, main="ACF: Stabilized Chemicals Data")
pacf(chemicals.bc.diff2, main="PACF: Stabilized Chemicals Data")
eacf(chemicals.bc.diff2)
plot(armasubsets(chemicals.bc.diff2, 5, 5))

# ACF, PACF, and EACF for Stabilized Vehicles Data
acf(vehicles.bc.diff2, main="ACF: Stabilized Vehicles Data")
pacf(vehicles.bc.diff2, main="PACF: Stabilized Chemicals Data")
eacf(vehicles.bc.diff2)
plot(armasubsets(vehicles.bc.diff2, 5, 5))

par(mfrow=c(2, 1))
# ARIMA Model Fits for Chemicals Data
chemicals.arima = auto.arima(chemicals.diff2)
chemicals.bc.arima = auto.arima(chemicals.diff2, lambda='auto')
chemicals.bc.diff2.arima = auto.arima(chemicals.bc.diff2)

# ARIMA Model Fits for Vehicles Data
vehicles.arima = auto.arima(vehicles.diff2)
vehicles.bc.arima = auto.arima(vehicles.diff2, lambda='auto')
vehicles.bc.diff2.arima = auto.arima(vehicles.bc.diff2)

# Residual Diagnostics (manually repeated for all the models)
par(mfrow=c(3, 2))
dataset = chemicals.diff2
model = chemicals.arima
# st_residuals = chemicals.residuals
st_residuals = rstandard(model)
plot(y=st_residuals, x=as.vector(time(dataset)), type = 'o', ylab = '
    Standardized Residuals', xlab = 'Time', main='Standardized Residual
    Plot')
hist(st_residuals, main="Residuals Histogram", xlab='Residuals')
qqnorm(st_residuals, main="QQ Plot for Residuals")
qqline(st_residuals, col="Red")
boxplot(st_residuals, main='Residual Boxplot', ylab='Residuals')
acf(st_residuals, main='ACF Plot')
tsdiag(model)
# Residual Normality
shapiro.test(st_residuals)
snpar::runs.test(st_residuals, exact=TRUE)
# AIC and BIC tests
AIC(model)
BIC(model)

# Correlation between Chemicals and Vehicles Data
ccf_cv = ccf(chemicals.data, vehicles.data)
```

## 3.2 Results



(i)

Figure 3.1: Decomposition of the Chemicals data shows that both trend and seasonality are present.



(ii)

Figure 3.2: Decomposition of the Vehicles data.

Here, we again observe that the data exhibit both trend and seasonality. Nevertheless, in both the cases, we check their stationarity.

```
> adf.test(chemicals.data)

Augmented Dickey-Fuller Test
data:  chemicals.data
Dickey-Fuller = -3.0203, Lag order = 6, p-value = 0.1462
alternative hypothesis: stationary

> adf.test(vehicles.data)
Augmented Dickey-Fuller Test
data:  vehicles.data
Dickey-Fuller = -2.9112, Lag order = 6, p-value = 0.1921
alternative hypothesis: stationary
```

The augmented Dickey-Fuller tests for both the datasets fail to reject the null hypothesis, and hence, these time series data are not stationary. Therefore, we detrend (also remove seasonality) these, and the time series plots are shown in Fig 3.3.

The visual observation of the final detrended (Fig 3.3) data show that both the time series have become stationary. We further test this using the augmented Dickey-Fuller test which confirms that we have achieved stationarity.

```
> adf.test(chemicals.diff2)

Augmented Dickey-Fuller Test
data:  chemicals.diff2
Dickey-Fuller = -5.6096, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary

> adf.test(vehicles.diff2)

Augmented Dickey-Fuller Test
data:  vehicles.diff2
Dickey-Fuller = -7.1672, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

Here, the original data can be described as a harmonic model with increasing trend. However, for the detrended data, we can calculate the ACF, PACF, and EACF to get an idea about the order of the ARIMA process.

(iii) Initally, we check the result of the harmonic model fit over the original data.

```
Call:
lm(formula = chemicals.data ~ har + time(chemicals.data))
```

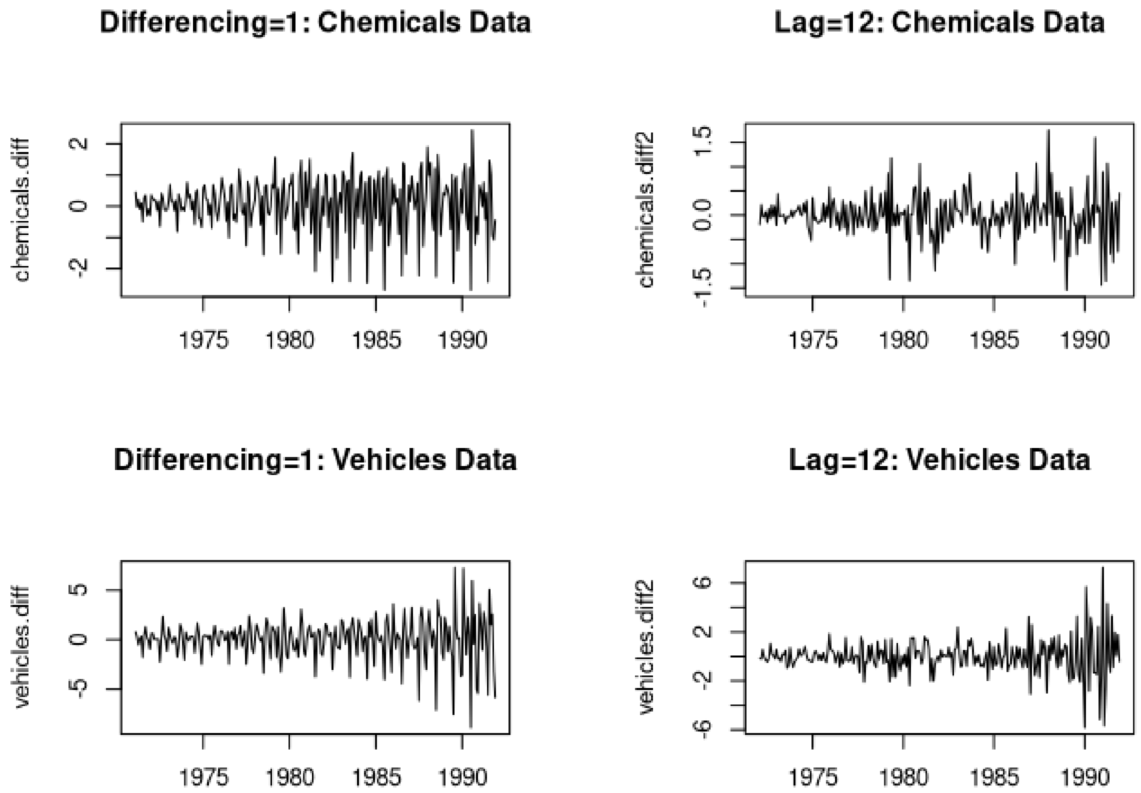


Figure 3.3: Time series plots of the detrended data. At first, the linear trend is removed using the first differences. Then the seasonality of the differenced data is removed by setting lag=12 as we deal with monthly data here.

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.0447 -0.5332  0.0430  0.6677  3.0799

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.951e+03  2.365e+01 -82.478  < 2e-16 ***
harcos(2*pi*t)      -9.564e-02  1.023e-01  -0.935    0.351
harsin(2*pi*t)       5.420e-01  1.023e-01   5.297  2.6e-07 ***
time(chemicals.data)  9.913e-01  1.194e-02  83.058  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.148 on 248 degrees of freedom
Multiple R-squared:  0.9653,Adjusted R-squared:  0.9649
F-statistic:  2302 on 3 and 248 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = vehicles.data ~ har + time(vehicles.data))

Residuals:
    Min      1Q  Median      3Q     Max
-7.0874 -1.3989  0.3886  1.5226  4.5814

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1.499e+03  4.476e+01 -33.497  < 2e-16 ***
harcos(2*pi*t)       2.120e-01  1.935e-01   1.095  0.27449
harsin(2*pi*t)       5.378e-01  1.936e-01   2.777  0.00591 **
time(vehicles.data)  7.626e-01  2.259e-02  33.760  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.172 on 248 degrees of freedom
Multiple R-squared:  0.8217,Adjusted R-squared:  0.8195
F-statistic: 380.9 on 3 and 248 DF,  p-value: < 2.2e-16
```

While the $R^2$ is good for both cases, the residual standard error is quite high along with the F-statistic. The model fits for both the original datasets are depicted in Fig 3.4. We now perform the residual diagnostics of the fit. Furthermore, we check the AIC

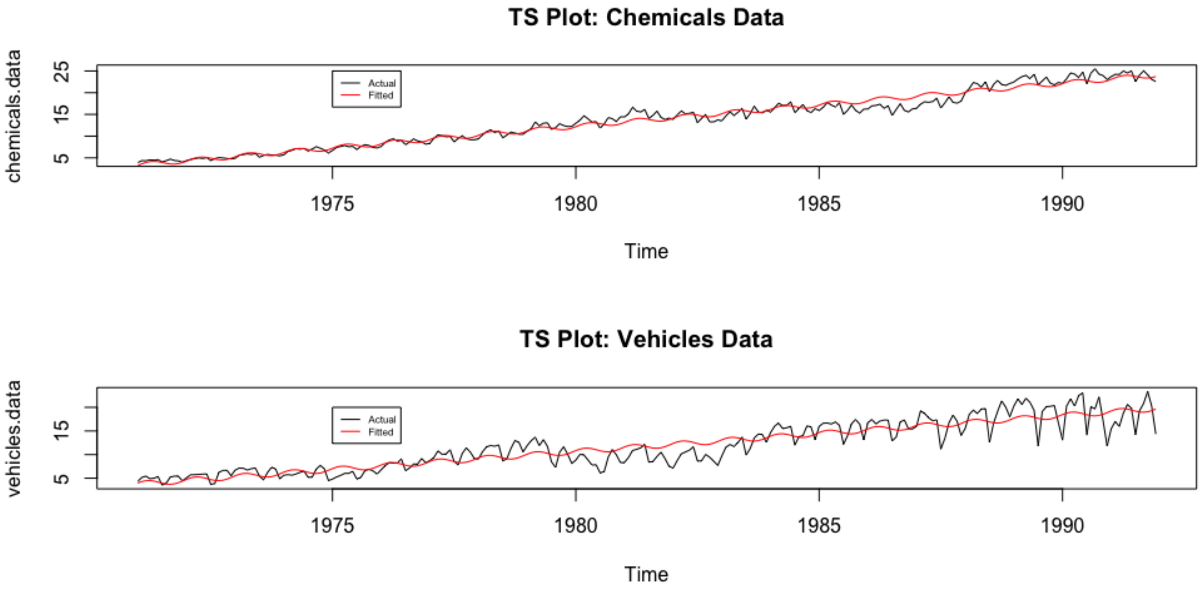


Figure 3.4: Fitting a harmonic model with linear trend over the original data. We see that this model somewhat matches the time series pattern but misses out a lot. Increasing the number of harmonic pairs better fits the model but could lead to possible overfitting and results in a non-parsimonious model.

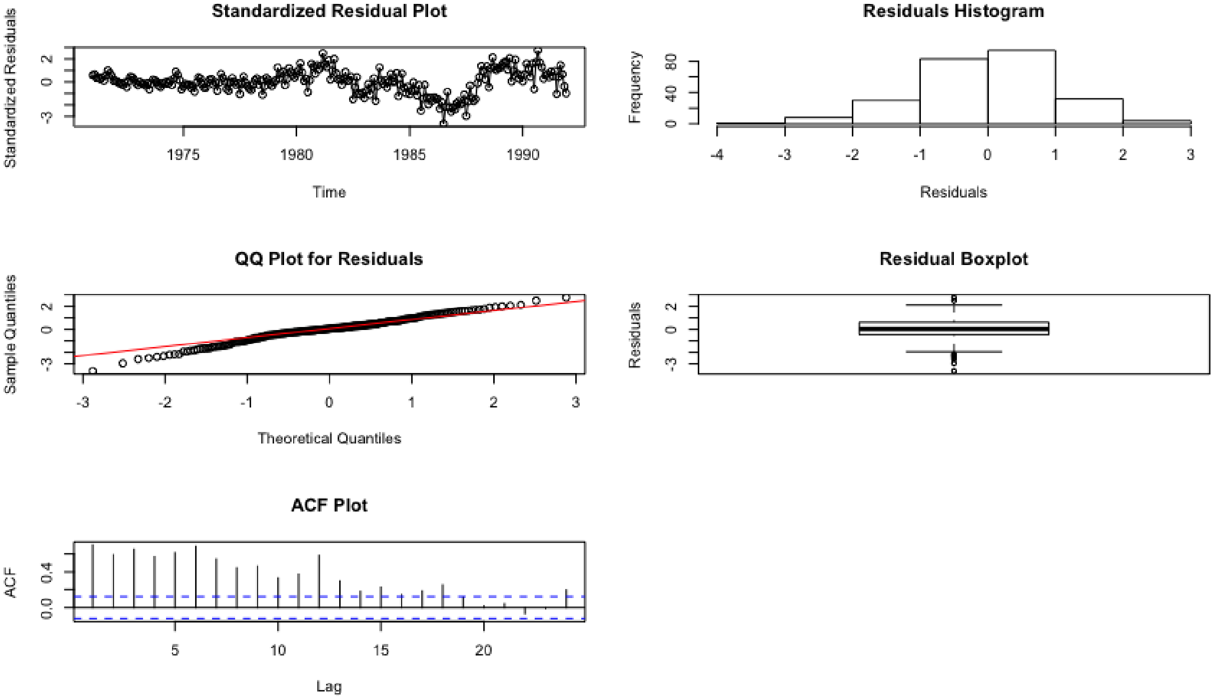


Figure 3.5: Residual diagnostics of the harmonic fit over the original Chemicals data. Here, we see the non-normality of the residuals.

and BIC values along with the p-values of the normality checks. Both the normality checks suggest that the residuals are from a non-normal distribution.

```
Original Chemicals Data

Shapiro-Wilk normality test
W = 0.98358, p-value = 0.005282

Exact runs test
Runs = 56, p-value < 2.2e-16
alternative hypothesis: two.sided

> AIC(chemicals.fit)
[1] 790.5797
> BIC(chemicals.fit)
[1] 808.2269

Original Vehicles Data

Shapiro-Wilk normality test
W = 0.98358, p-value = 0.005282

Exact runs test
```

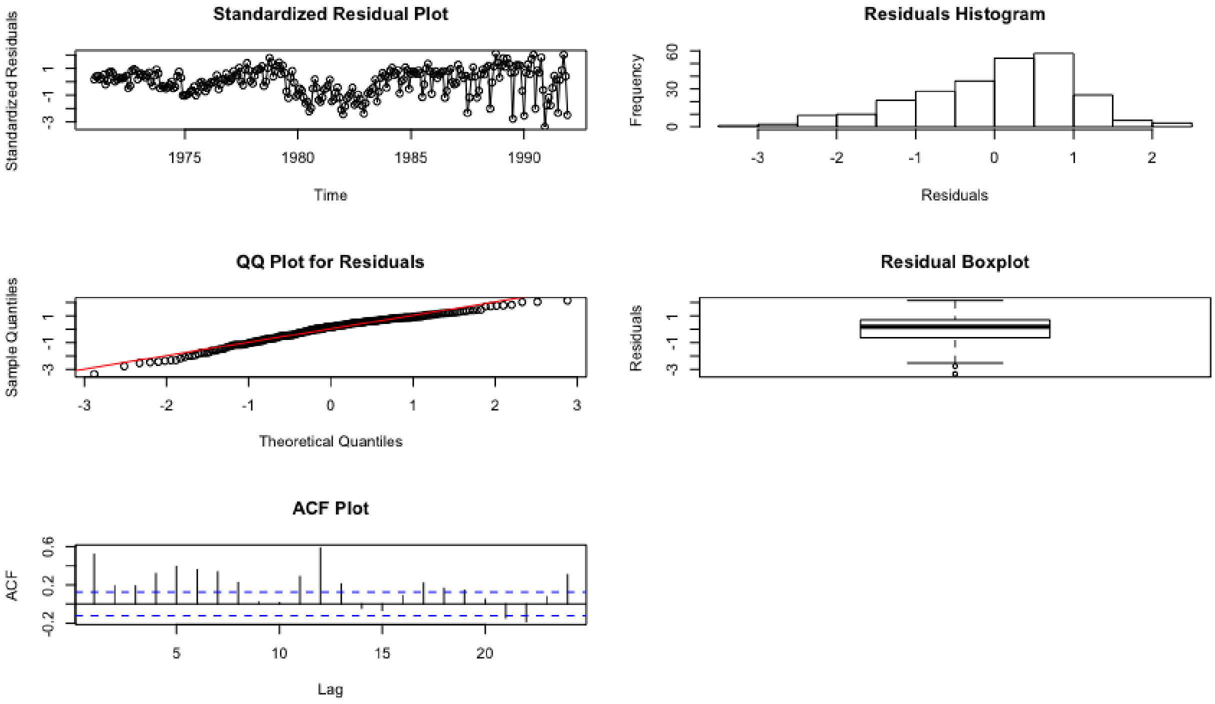


Figure 3.6: Residual diagnostics of the harmonic fit over the original Vehicles data. Here, we again see the non-normality of the residuals.

```
Runs = 56, p-value < 2.2e-16
alternative hypothesis: two.sided

> AIC(vehicles.fit)
[1] 1112.108
> BIC(vehicles.fit)
[1] 1129.755
```

Now, we observe the ACF, PACF, EACF, and ARMA Subset Plots for the detrended data. The EACF matrices for these data are given below.

```
> eacf(chemicals.diff2)
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x o x x o o o x o o o  x  o  o
1 o o x o o o o x o o o  x  o  o
2 x x o o o o o x o o o  x  o  x
3 x x x o o o o o o o o  x  o  x
4 x x o o o o o o o o o  x  o  x
5 x x o o o o o o o o o  x  x  o
6 x x o x x o o o o o o  x  o  o
7 x x x x o o o o o o o  x  o  o
```

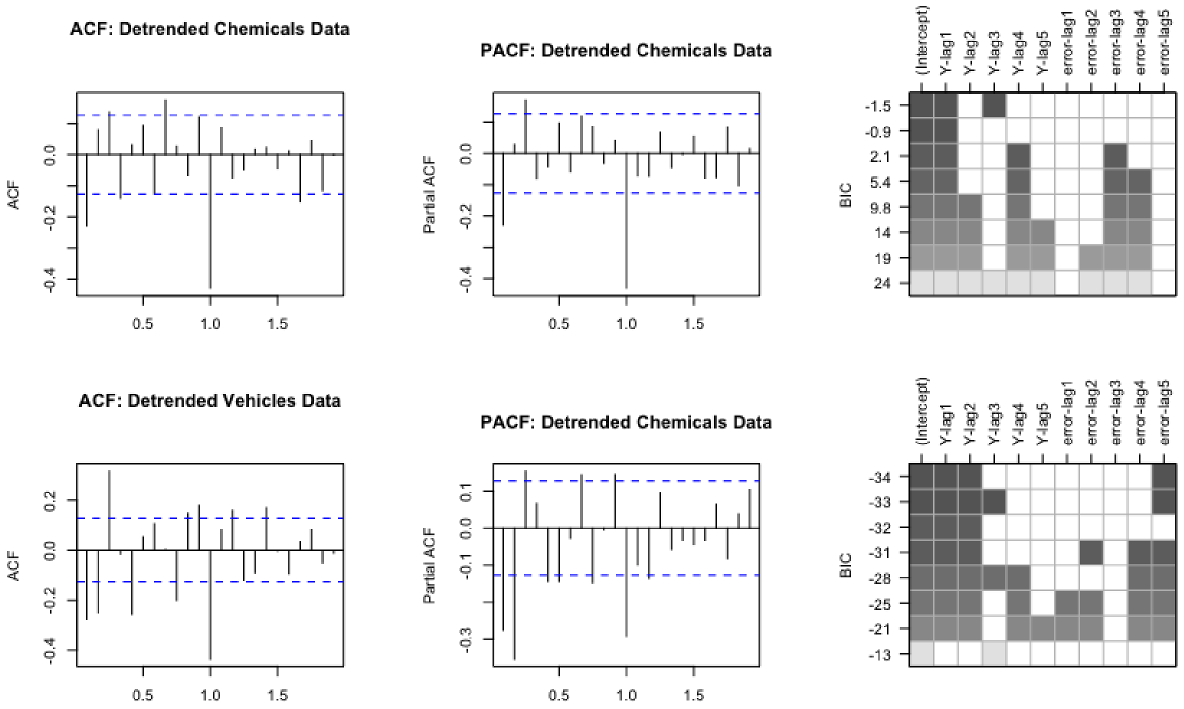


Figure 3.7: ACF, PACF, and ARMA subset plots for the detrended Chemicals and Vehicles data. These show that the detrended data could possible have both AR and MA parts.

```
> eacf(vehicles.diff2)
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x o x o o o x x x  x  o  x
1 x x x o x o o o x o o  x  x  x
2 x x o x x o o o o o o  x  o  x
3 x x x o x o o o o o o  x  x  x
4 x x o o o o o o o o o  x  x  x
5 x x x o x o o o o o o  x  x  o
6 x x o x x o o o o o o  x  x  o
7 x x o o o o o o o o o  x  x  o
```

Next, we fit the best ARIMA model using the auto.arima function in R and perform the necessary residual diagnostics (using tsdiag function). These are shown in Figures 3.8 and 3.9.

```
> chemicals.arima = auto.arima(chemicals.diff2)
> chemicals.arima
Series: chemicals.diff2
ARIMA(2,0,2)(0,0,1)[12] with zero mean

Coefficients:
```

```
          ar1      ar2     ma1     ma2     sma1
      -0.9289  -0.8776  0.7277  0.7503  -0.6089
s.e.   0.0644   0.0712  0.0901  0.0911   0.0583

sigma^2 estimated as 0.1409:  log likelihood=-104.97
AIC=221.95   AICc=222.31   BIC=242.8

> vehicles.arima = auto.arima(vehicles.diff2)
> vehicles.arima
Series: vehicles.diff2
ARIMA(2,0,2)(0,0,1)[12] with zero mean

Coefficients:
          ar1      ar2     ma1     ma2     sma1
      -0.4311  -0.7637  0.1581  0.5965  -0.5006
s.e.   0.0943   0.1128  0.1040  0.1725   0.0623

sigma^2 estimated as 1.259:  log likelihood=-366.3
AIC=744.61   AICc=744.97   BIC=765.47
```

Like before, we carry out the normality tests which align with our graphically interpreted results.

```
Detrended Chemicals Data

Shapiro-Wilk normality test
W = 0.95342, p-value = 5.817e-07

Exact runs test
Runs = 121, p-value = 0.8969
alternative hypothesis: two.sided

Detrended Vehicles Data

Shapiro-Wilk normality test
W = 0.97736, p-value = 0.0007176

Exact runs test
Runs = 109, p-value = 0.1537
alternative hypothesis: two.sided
```

Once these analyses are done, we apply BoxCox transformation (automatically chosen $\lambda_C \approx 0.19$ and $\lambda_V \approx -0.03$ for Chemicals and Vehicles data respectively) to the original data. We also do the same for the detrended data but in that case we first apply the transformation and then detrend.

The model details of the harmonic fit over the transformed data are given below.

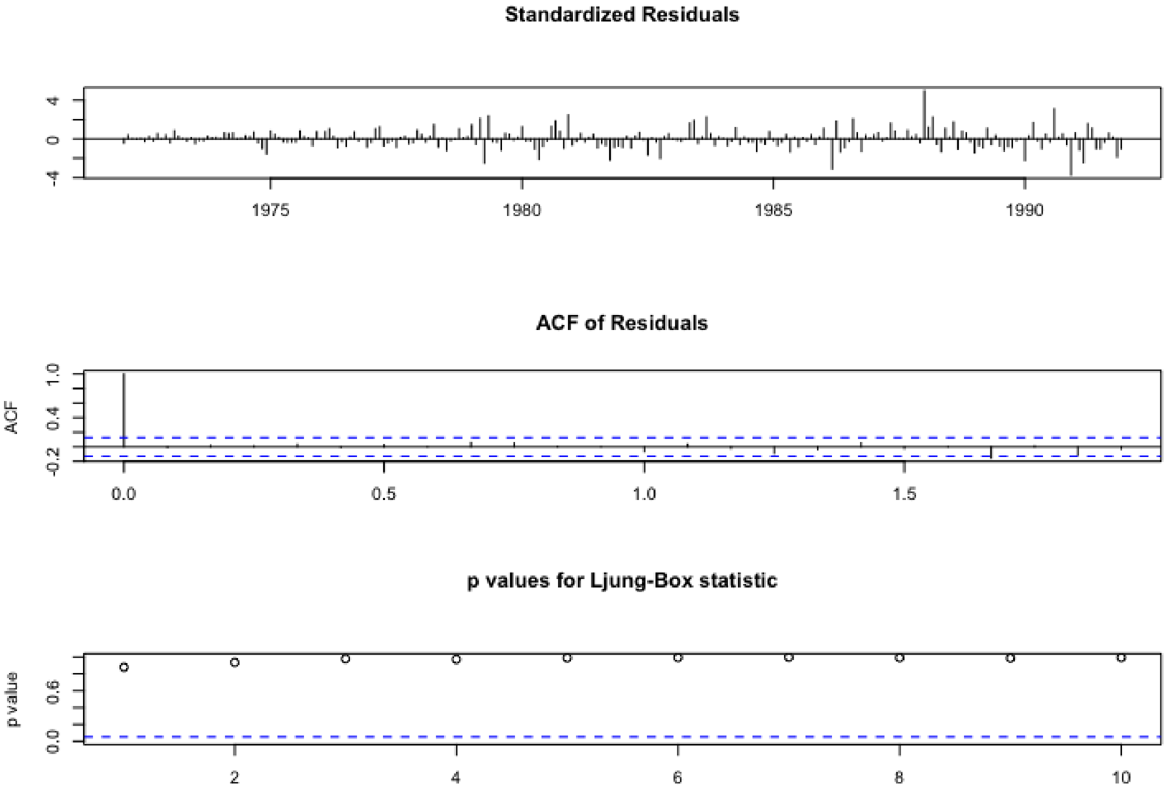


Figure 3.8: Residual diagnostics for the ARIMA fit over the detrended Chemicals data. These show that the residuals closely follow a normal distribution. Also, we observe that the AIC and BIC values have significantly reduced as compared to the harmonic fit over the same data.

```
Transformed Chemicals Data

Call:
lm(formula = chemicals.bc.tr ~ (har + time(chemicals.bc.tr)))

Residuals:
     Min       1Q   Median       3Q      Max
-0.40239 -0.13766 -0.00284  0.12268  0.47168

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           -2.587e+02  3.716e+00 -69.611  < 2e-16 ***
harcos(2*pi*t)        -1.439e-02  1.607e-02  -0.895    0.371
harsin(2*pi*t)         6.795e-02  1.608e-02   4.226 3.34e-05 ***
time(chemicals.bc.tr)  1.322e-01  1.875e-03  70.490  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

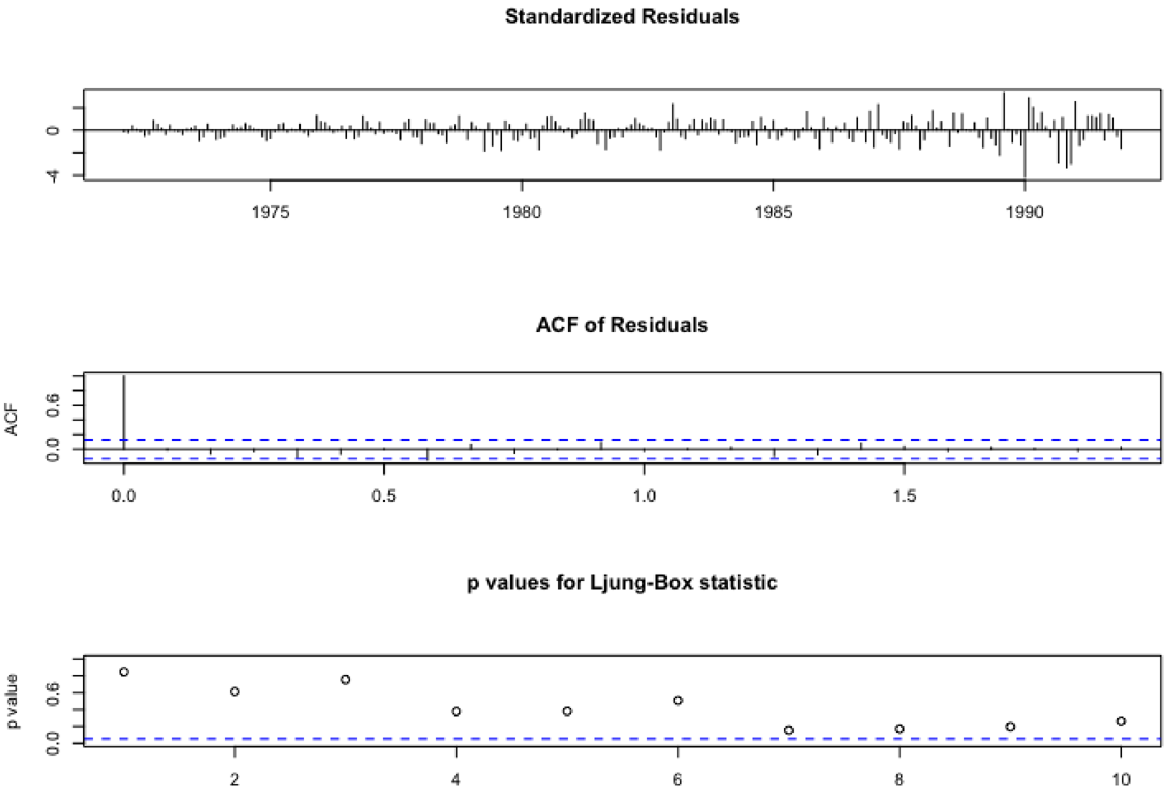


Figure 3.9: Residual diagnostics for the ARIMA fit over the detrended Vehicles data. These show that the residuals do not closely follow a normal distribution. But, we observe that the AIC and BIC values have significantly reduced as compared to the harmonic fit over the same data.

```
Residual standard error: 0.1803 on 248 degrees of freedom
Multiple R-squared:  0.9525,Adjusted R-squared:  0.9519
F-statistic:  1658 on 3 and 248 DF,  p-value: < 2.2e-16

Transformed Vehicles Data
Call:
lm(formula = vehicles.bc.tr ~ (har + time(vehicles.bc.tr)))

Residuals:
     Min       1Q   Median       3Q      Max
-0.51612 -0.12429  0.02448  0.12732  0.41170

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -1.272e+02  3.673e+00 -34.628   <2e-16 ***
harcos(2*pi*t)        2.391e-02  1.588e-02   1.506   0.1334
```

```
harsin(2*pi*t)        4.306e-02  1.589e-02   2.710   0.0072 **
time(vehicles.bc.tr)  6.534e-02  1.854e-03  35.250   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1783 on 248 degrees of freedom
Multiple R-squared:  0.834,Adjusted R-squared:  0.832
F-statistic: 415.3 on 3 and 248 DF,  p-value: < 2.2e-16
```

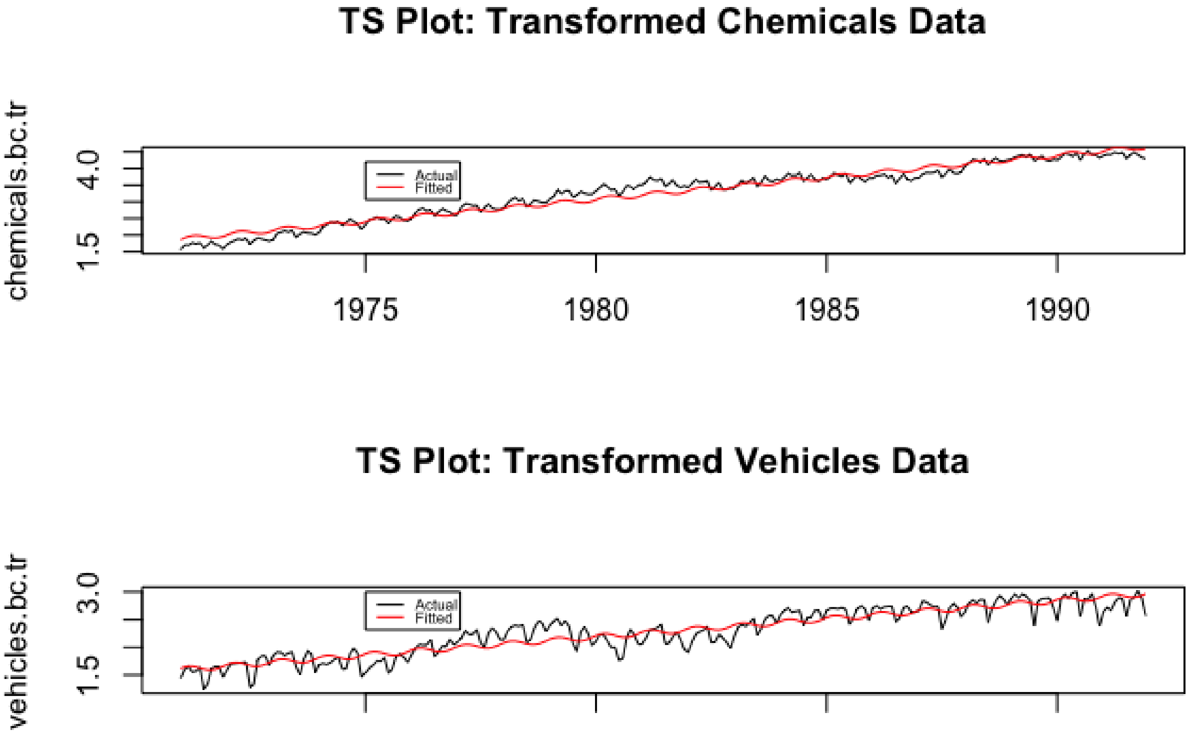


Figure 3.10: Fitting a harmonic model with linear trend over the transformed data. There is not much improvement in the fit as compared to Fig 3.4. In fact, the fit over the Chemicals data is worse. Therefore, we do not perform any other diagnostics for this model and carry on with the detrended and stabilized datasets.

Here, in Fig 3.11 we observe the temporal patterns of the stabilized datasets (transformed, detrended, and seasonality removed).

Now, we check the different ACF plots depicted in Fig 3.12. The EACF matrices for these stabilized datasets are given here.

```
> eacf(chemicals.bc.diff2)
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x o o o o o o o o x  x  o  o
```

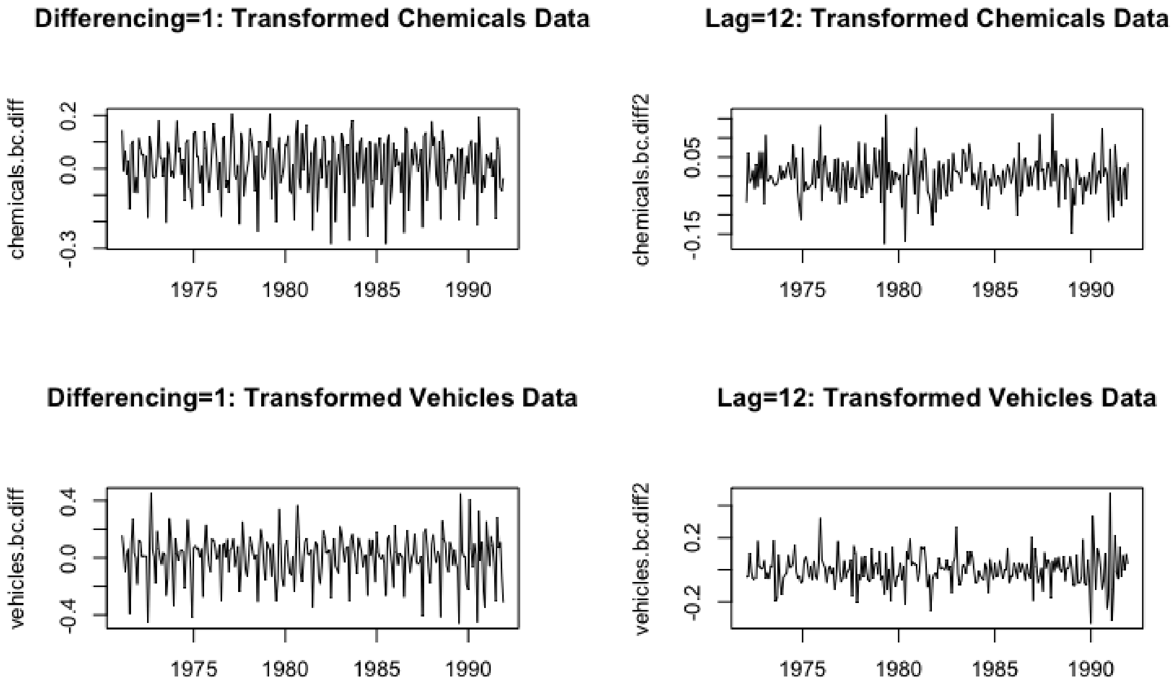


Figure 3.11: Time series plots of the stabilized data. The plots are constructed the same way as in Fig 3.3. However, we see that the stabilized data appears to be more stationary than before.

```
1 x x o o o o o o o o o o  x  x  o
2 x x o o o o o o o o o o  x  x  o
3 o x x o o o o o o o o o  x  x  o
4 x x o o o o o o o o o o  x  x  x
5 x x o o o o o o o o o o  x  o  x
6 x o o o x o o o o o o o  x  o  o
7 x o o x o o o o o o o o  x  x  o

> eacf(vehicles.bc.diff2)
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x o o x o o o o o o x  x  o  o
1 x x o o x o o o o o o o  x  x  o
2 x o o o x o o o o o o o  x  x  x
3 o o x o x o o o o o o o  x  x  x
4 o x x x o o o o o o o o  x  x  o
5 x x x x o o o o o o o o  x  x  o
6 x x o x o o o o o o o o  x  x  o
7 x x o o o o x o o o o o  x  o  o
```

As expected from Fig 3.12, we have greatly reduced the AIC and BIC values using

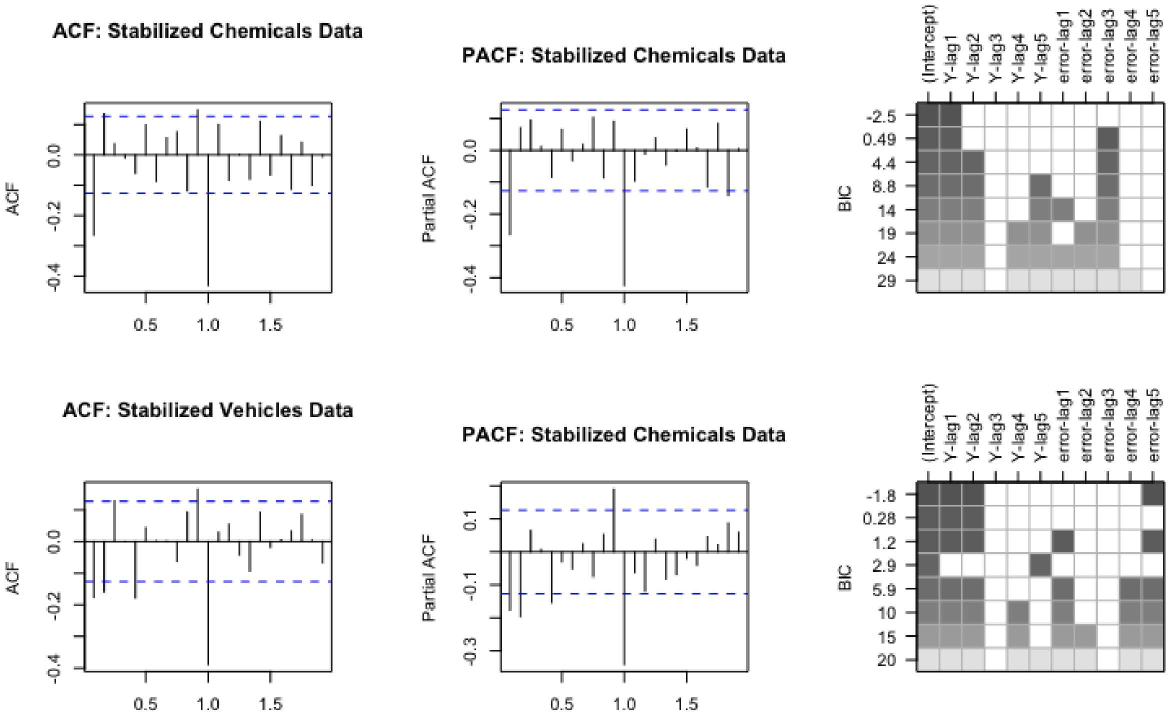


Figure 3.12: ACF, PACF, and ARMA subset plots for the stabilized Chemicals and Vehicles data. These show that the detrended data could possible have both AR and MA parts like before but the AIC and BIC values could be significantly reduced (since $\lambda_C \approx 0.19$ and $\lambda_V \approx -0.03$ for Chemicals and Vehicles data respectively) as compared to Fig 3.7.

transformation. The model details obtained using the auto.arima function are given here. It is noteworthy that, while the number of parameters for the Chemicals data remain the same as before, we have been able to reduce one parameter in case of the Vehicles data. Hence, we do not further check the residuals.

```
> chemicals.bc.diff2.arima = auto.arima(chemicals.bc.diff2)
> chemicals.bc.diff2.arima

Series: chemicals.bc.diff2
ARIMA(2,0,2)(0,0,1)[12] with zero mean

Coefficients:
          ar1      ar2     ma1     ma2     sma1
      -0.9330  -0.9102  0.7469  0.7854  -0.6996
s.e.   0.0548   0.0595  0.0850  0.0741   0.0641

sigma^2 estimated as 0.001787:  log likelihood=415.93
AIC=-819.86   AICc=-819.5   BIC=-799.01

> vehicles.bc.diff2.arima = auto.arima(vehicles.bc.diff2)
> vehicles.bc.diff2.arima

Series: vehicles.bc.diff2
ARIMA(0,0,2)(0,0,2)[12] with zero mean

Coefficients:
          ma1      ma2     sma1     sma2
      -0.2106  -0.1494  -0.6332  -0.1205
s.e.   0.0676   0.0740   0.0678   0.0723

sigma^2 estimated as 0.006902:  log likelihood=252.91
AIC=-495.82   AICc=-495.56   BIC=-478.44
```

Therefore, from the above results, it can be concluded that transformation improves the fit of these models. In order to check the cross-correlation between the Chemicals and Vehicles data, we use the ccf function in R (Fig 3.13). This shows that the sales of chemicals and vehicles are positively correlated.

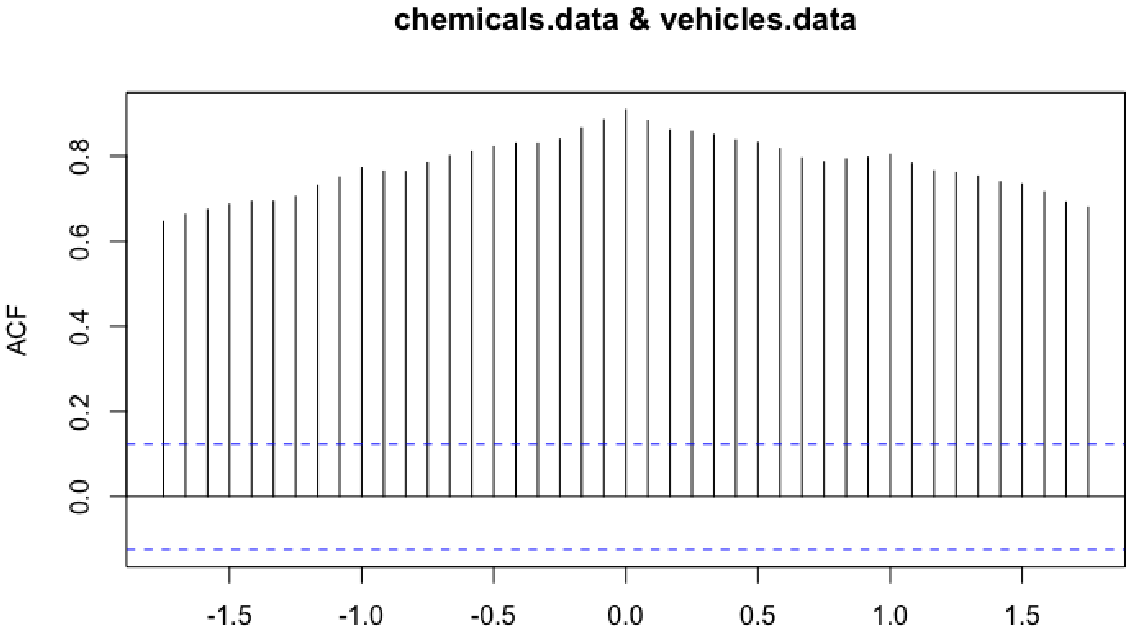


Figure 3.13: Cross-correlation plot of Chemicals and Vehicles data.

(iv) The strong positive correlation between the Chemicals and Vehicles data could be attributed to the increase of vehicle usage over time thereby necessitating more use of chemicals not only for vehicle parts but also for additional utilities like stain removers etc. However, if the month-to-month variation is considered, there are bigger variations in the sales of vehicles. This is because Chemical sales on a monthly basis will not fluctuate much as chemicals are required in a variety of industries. But in case of vehicles, there are additional factors involved such as the economic slowdown which significantly affect the sales.

# 4 Problem 4

France has enjoyed a long and storied history as a world leader in culture, food, literature, film, and revolutions. Over its history, the average life expectancy of people in France has changed notably. The time series object described below contains the annual average life expectancy in France (measured in years lived), measured during the years 1816 to 2019. The data is in the file FLE.txt.

Analyze the data in the time series and write a report addressing the following:

(a) What trend model(s) best capture the trends in French life expectancy over time?

(b) Are there any noticeable patterns, or observations apparent in the series? Can you tell from your plot what could be the causes of your observations?

(c) Regardless of what you observe, suppose you apply some transformation to the data, how does the new data looks like? Any changes in trend?

(d) For the various models you tried, assess the fit of the models using any tools at your disposal. Perform any necessary test to make sure sure your model is appropriate. Write down the equation of your final model. Are any transformations of the data necessary? Perform all diagnostics tests.

Now conclude with brief summary: How the life expectancy changes over time, both long-term over the observed period of years, and in terms of patterns of year-to-year variation. Add any necessary graphs, tests and/or confidence interval etc...

## 4.1 R Code

```
### STAT 5814 MIDTERM 2/PROBLEM 4
### AUTHOR: SAYANTAN MAJUMDAR
### EMAIL: smxnv@mst.edu
### SNO: 12566087

library(TSA)
library(snpar)
library(stats)
library(forecast)
library(tseries)
library(fUnitRoots)

data = read.table('Data/ExamData/FLE.txt')
fle.data = ts(data$V2, frequency=1, start=1816, end=2019)
par(mfrow=c(1, 1))
plot(fle.data, main='Time Series Plot of French Life Expectancy', xlab='Time (Years)', ylab='Annual Average Life Expectancy (years)')
# Stationarity Check
adf.test(fle.data)

```

```
# Quadratic fit over original data
fle.fit = lm(fle.data ~ time(fle.data) + I(time(fle.data)^2))
fle.residuals = rstudent(fle.fit)
summary(fle.fit)
fle.model = ts(fle.fit$fitted.values, frequency=1, start=1816, end=2019)

plot(fle.data, main=’Quadratic Fit: Life Expectancy Data’)
lines(fle.model, col=’red’)
legend(1850, 70, legend=c("Actual", "Fitted"), col=c("Black", "Red"), lty
    =1:1, cex=0.7)

# Transformed Data
fle.bc.lambda = BoxCox.lambda(fle.data)
fle.bc.tr = BoxCox(fle.data, lambda=fle.bc.lambda)
plot(fle.bc.tr, main=’Box Cox Transformed Data’)

fle.bc.fit = lm(fle.bc.tr ~ time(fle.bc.tr) + I(time(fle.bc.tr)^2))
fle.bc.model = ts(fle.bc.fit$fitted.values, frequency=1, start=1816, end
    =2019)
summary(fle.bc.fit)

plot(fle.bc.tr, main=’Box Cox Transformed Data’)
lines(fle.bc.model, col=’red’)
legend(1850, 2500, legend=c("Actual", "Fitted"), col=c("Black", "Red"),
    lty=1:1, cex=0.7)

# Model assessment for the transformed data
fle.bc.residuals = rstudent(fle.bc.fit)

# Detrending of the original data
fle.detrend = diff(fle.data, differences=2)
plot(fle.detrend, main=’Detrended FLE Data’)
# Stationarity Check
adf.test(fle.detrend)

par(mfrow=c(3, 1))
# ACF, PACF, and EACF for Detrended FLE Data
acf(fle.detrend, main="ACF: Detrended FLE Data")
pacf(fle.detrend, main="PACF: Detrended FLE Data")
eacf(fle.detrend)
plot(armasubsets(fle.detrend, 5, 5))

# ARIMA Model Fits for Detrended FLE Data
fle.detrend.arima = arima(fle.detrend, order=c(1, 0, 2))
fle.auto.arima = auto.arima(fle.detrend)
fle.bc.auto.arima = auto.arima(fle.detrend, lambda="auto")

# Residual Diagnostics (manually repeated for all the models)
par(mfrow=c(1, 3))
dataset = fle.detrend
model = fle.auto.arima
# st_residuals = fle.bc.residuals
st_residuals = rstandard(model)
plot(y=st_residuals, x=as.vector(time(dataset)), type = ’o’, ylab = ’
    Standardized Residuals', xlab = 'Time', main='Standardized Residual
    Plot')
hist(st_residuals, main="Residuals Histogram", xlab='Residuals')
qqnorm(st_residuals, main="QQ Plot for Residuals")
qqline(st_residuals, col="Red")
boxplot(st_residuals, main='Residual Boxplot', ylab='Residuals')
acf(st_residuals, main='ACF Plot')
tsdiag(model)
# Residual Normality
shapiro.test(st_residuals)
snpar::runs.test(st_residuals, exact=TRUE)
# AIC and BIC tests
AIC(model)
BIC(model)
confint(fle.bc.auto.arima)
```

## 4.2 Results

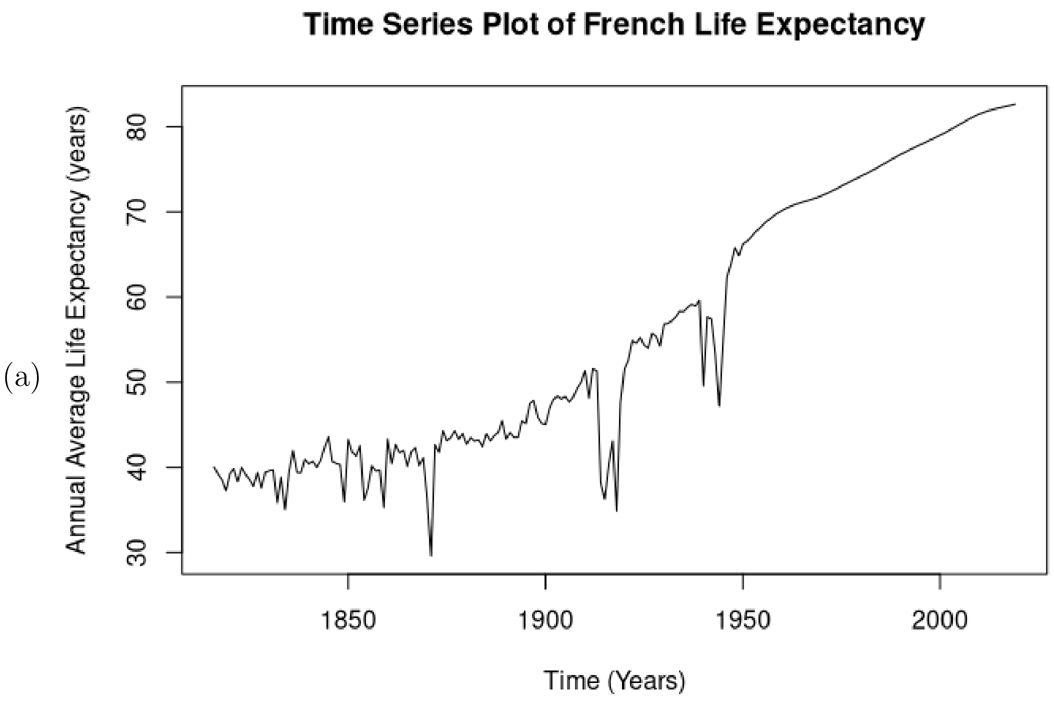


Figure 4.1: Time series plot of the French life expectancy data.

From Fig 4.1 it can be said that a quadratic trend is suitable for this dataset. The model summary is shown here.

```
Call:
lm(formula = fle.data ~ time(fle.data) + I(time(fle.data)^2))

Residuals:
     Min       1Q   Median       3Q      Max
-17.2140  -1.3665   0.4999   1.8464   5.6716
```

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)          3.499e+03  3.042e+02   11.50   <2e-16 ***
time(fle.data)      -3.849e+00  3.175e-01  -12.12   <2e-16 ***
I(time(fle.data)^2)  1.069e-03  8.279e-05   12.92   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.668 on 201 degrees of freedom
Multiple R-squared:  0.946,Adjusted R-squared:  0.9455 
F-statistic:  1762 on 2 and 201 DF,  p-value: < 2.2e-16
```

Here, we observe that the F-statistic and the residual standard are quite high. Moreover, the augmented Dickey-Fuller stationarity test (Dickey-Fuller = -2.9107, Lag order = 5, p-value = 0.1945) confirms that the original data is not stationary. However, by only visually observing the original data (Fig 4.1), we can say that the quadratic fit (Fig 4.2) is the best trend for this dataset.

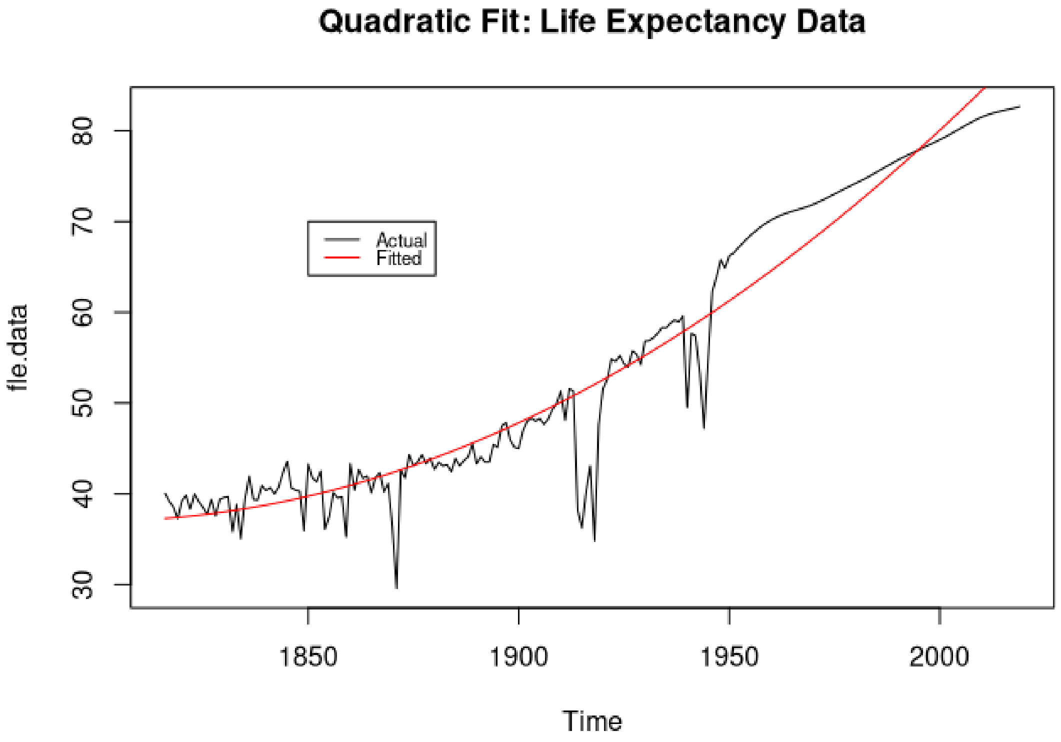


Figure 4.2: Quadratic fit over the original French life expectancy data. The fit performs poorly from 1950 onward as we observe a linear trend from 1950-2019.

(b) The fluctuations in the dataset are mostly present before 1950. The increase in the life expectancy values between 1816-1850 could be attributed to the advancements in medicine. However, the sharp drops in the values, particularly during mid 1915-1920 and mid 1940s could be caused by the two world wars. Interestingly, the life expectancy increases post 1950 which could be primarily attributed to the invention of modern healthcare and economic growth.

(c)

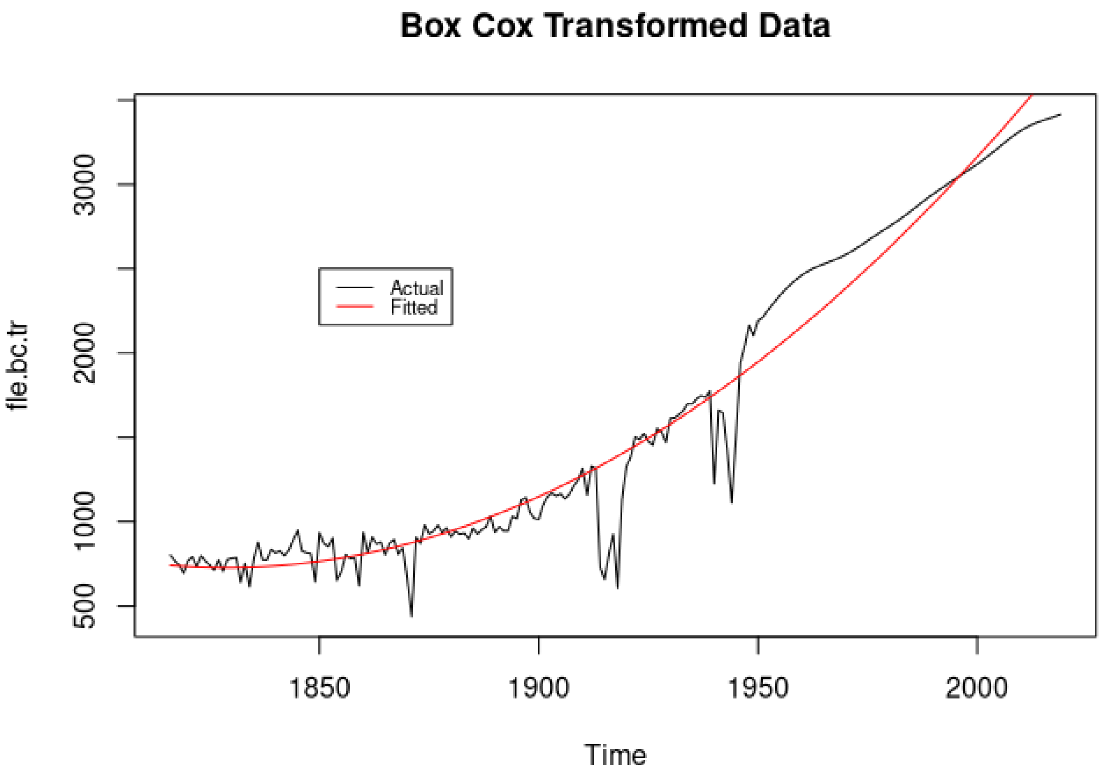


Figure 4.3: Transformed FLE Data using Box Cox method. Here, this transformation is obtained by auto-setting $\lambda \approx 1.99$. The quadratic trend still remains.

The model fit summary for this transformed dataset is given here.

```
Call:
lm(formula = fle.bc.tr ~ time(fle.bc.tr) + I(time(fle.bc.tr)^2))

Residuals:
    Min      1Q  Median      3Q     Max
-782.46  -70.39   23.94   83.28  305.68

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.787e+05  1.494e+04   18.66   <2e-16 ***
time(fle.bc.tr)      -3.040e+02  1.559e+01  -19.50   <2e-16 ***
I(time(fle.bc.tr)^2)  8.311e-02  4.064e-03   20.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 180.1 on 201 degrees of freedom
Multiple R-squared:  0.9625,Adjusted R-squared:  0.9621
F-statistic:  2579 on 2 and 201 DF,  p-value: < 2.2e-16
```

(d) The residual diagnostics for the quadratic model are shown in Fig 4.4. We further verify the non-normality of the residuals using the Shapiro-Wilk and Exact Runs tests.

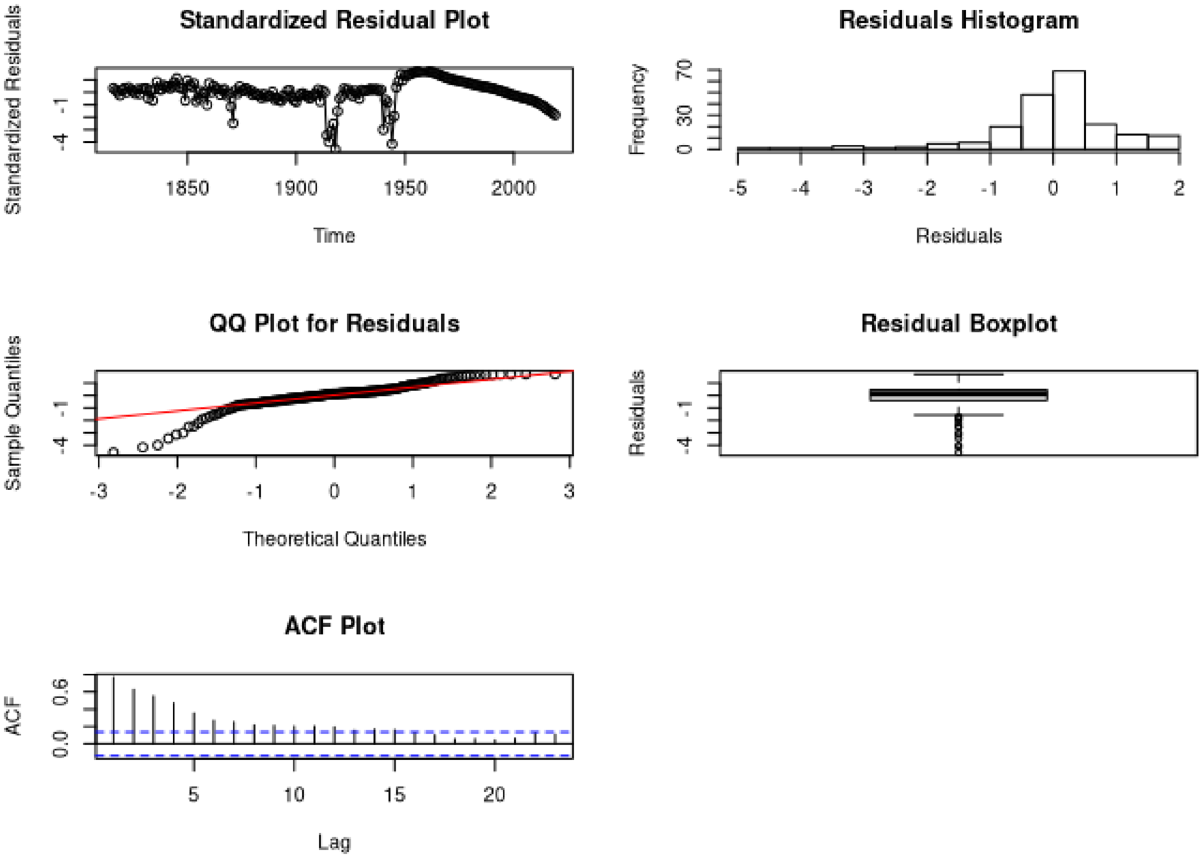


Figure 4.4: Residual diagnostics for the quadratic model. Here, we observe that this model fits poorly as the residuals follow a non-normal distribution.

```
Quadratic Model

Shapiro-Wilk normality test
W = 0.88032, p-value = 1.21e-11

Exact runs test
Runs = 46, p-value = 3.374e-16
alternative hypothesis: two.sided
```

Next, we detrend the transformed dataset (Fig 4.5) using two differences because of the quadratic trend (i.e. the transformed dataset is not stationary). Since no seasonality is present, we set lag=1. Once the detrending is done, we check the ACF, PACF, EACF, and the ARMA subset plots as given in Fig. 4.6.

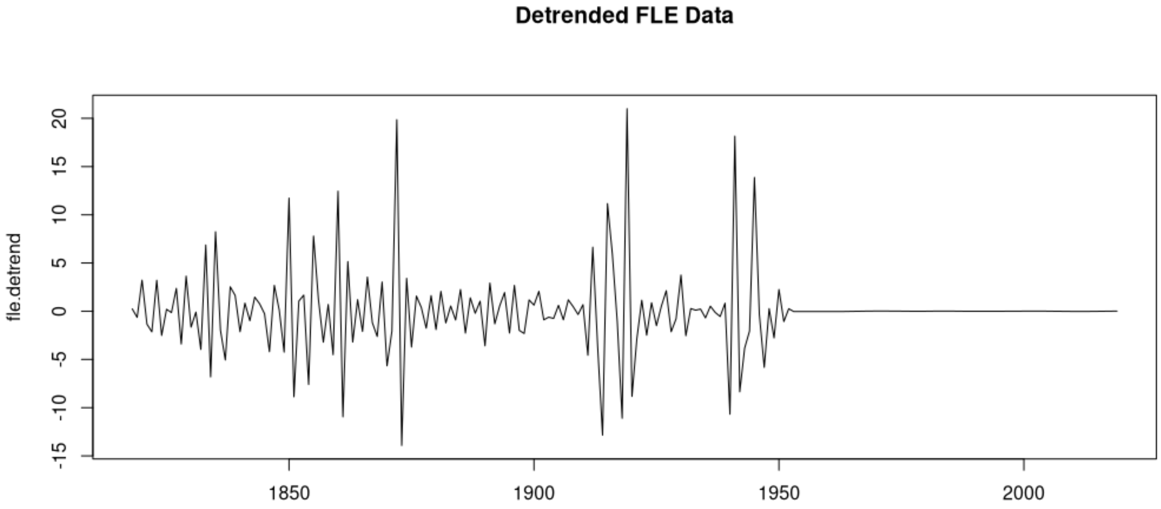


Figure 4.5: Plot of the detrended FLE data. This shows that we have achieved stationarity and can now proceed with various ARIMA modeling. We also verify this using the augmented Dickey-Fuller test (Dickey-Fuller = -10.009, Lag order = 5, p-value = 0.01) where we accept the alternative hypothesis of stationarity.

```
> eacf(fle.detrend)

AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x o o o o o o o o o o  x  x  o
1 x o o o o o o o o o o  o  o  o
2 x x o o o o o o o o o  o  o  o
3 x o x o o o o o o o o  o  o  o
4 x o x x x o o o o o o  o  o  o
5 x x o o x o o o o o o  o  o  o
6 x x x o o o x o o o o  o  o  o
7 x x x o o o o x o o o  o  o  o
```

Initially, we consider ARMA(1, 1) model or ARIMA(1, 0, 1). The model fit summary is given here.

```
>arima(x = fle.detrend, order = c(1, 0, 1))

Coefficients:
```

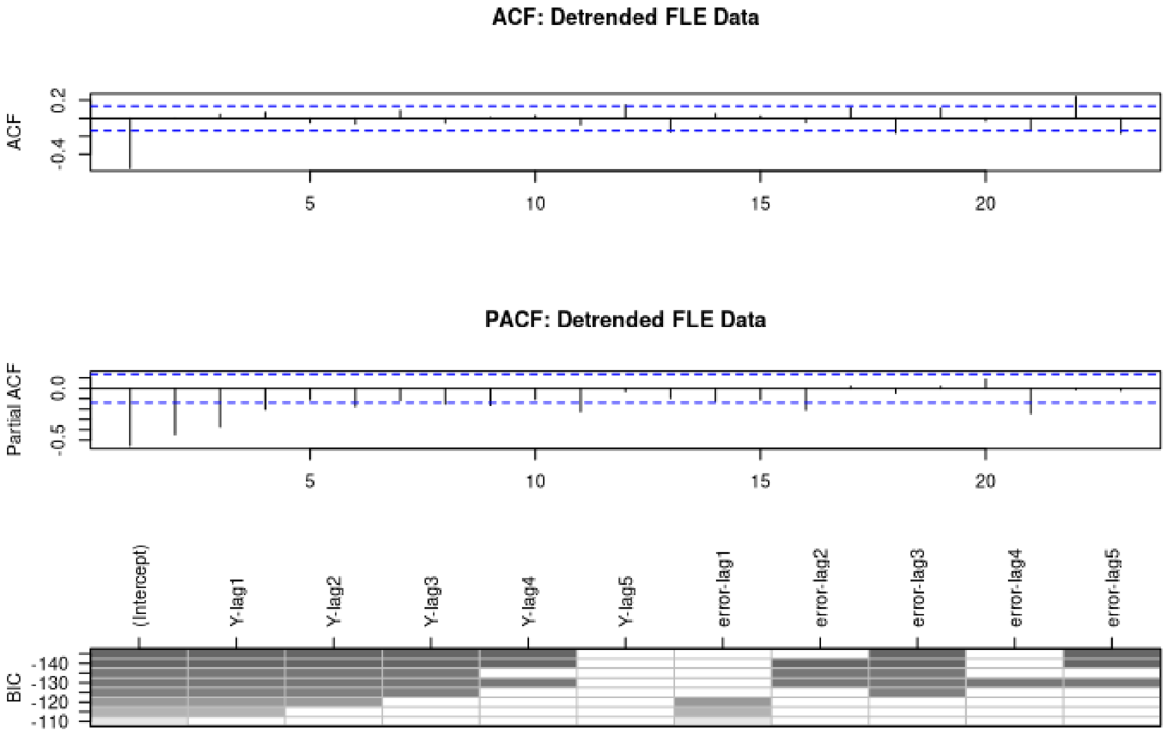


Figure 4.6: ACF, PACF, and ARMA subset plots of the detrended FLE data. From these we can say that ARMA(p, q) is a suitable model.

```
          ar1     ma1  intercept
      -0.2498  -1.000     0.0017
s.e.   0.0680   0.013     0.0025

sigma^2 estimated as 6.765:  log likelihood = -482.63,  aic = 971.26
```

We also perform the residual diagnostics (Figures 4.7 and 4.8)and apply the normality checks. The residual diagnostics show that the ARMA(1, 1) residuals do not closely follow a normal distribution. This is further confirmed by the Shapiro-Wilk and Runs tests.

```
ARMA(1, 1) for the detrended FLE data

Shapiro-Wilk normality test
W = 0.76725, p-value < 2.2e-16

Exact runs test
Runs = 86, p-value = 0.02855
alternative hypothesis: two.sided
```

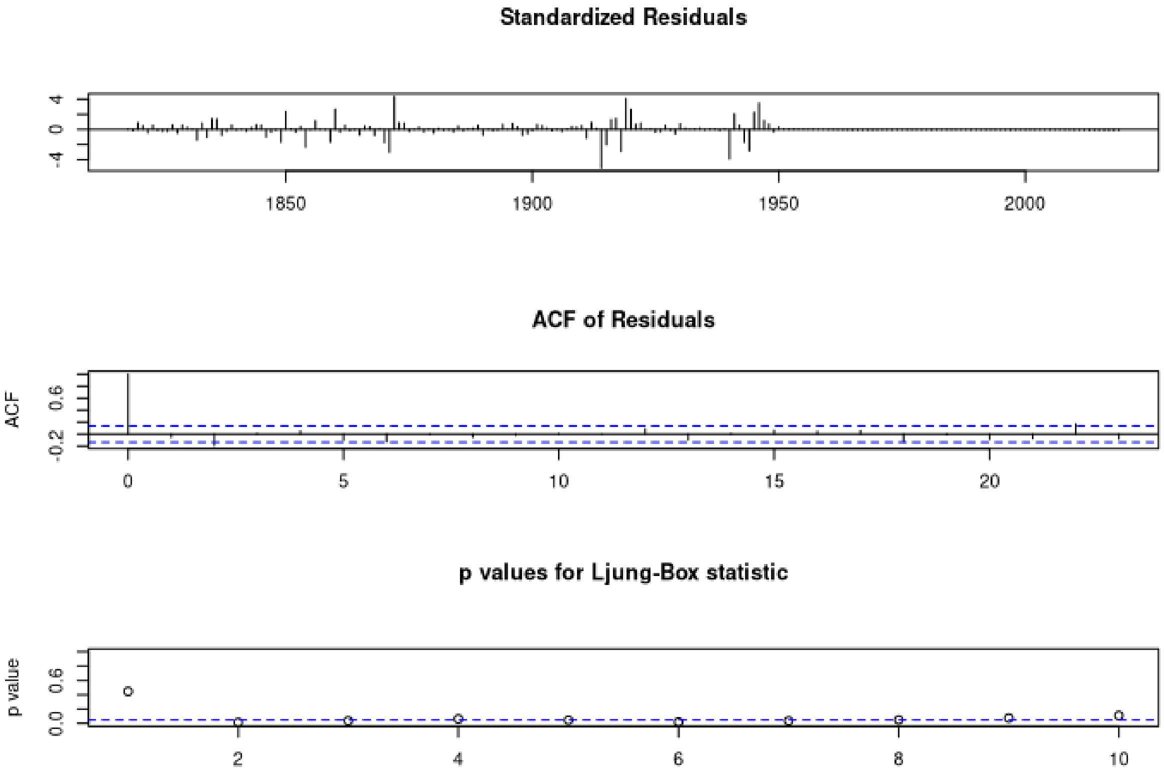


Figure 4.7: Residual diagnostics for the ARMA(1, 1) model for the FLE data.

```
> AIC(model)
[1] 973.2626
> BIC(model)
[1] 986.4957
```

Next, we try with ARMA(1, 2) to see if we have a better fit.

```
Call:
arima(x = fle.detrend, order = c(1, 0, 2))

Coefficients:
         ar1      ma1     ma2  intercept
      0.5658  -1.8831  0.8831     0.0017
s.e.  0.0946   0.0600  0.0594     0.0009

sigma^2 estimated as 6.266:  log likelihood = -476.12,  aic = 960.25
```

The residual diagnostics (Figures 4.9 and 4.10) and normality checks are carried out like before. The residual diagnostics show that the ARMA(1, 1) residuals do not closely follow a normal distribution. This is further confirmed by the Shapiro-Wilk and Runs tests.

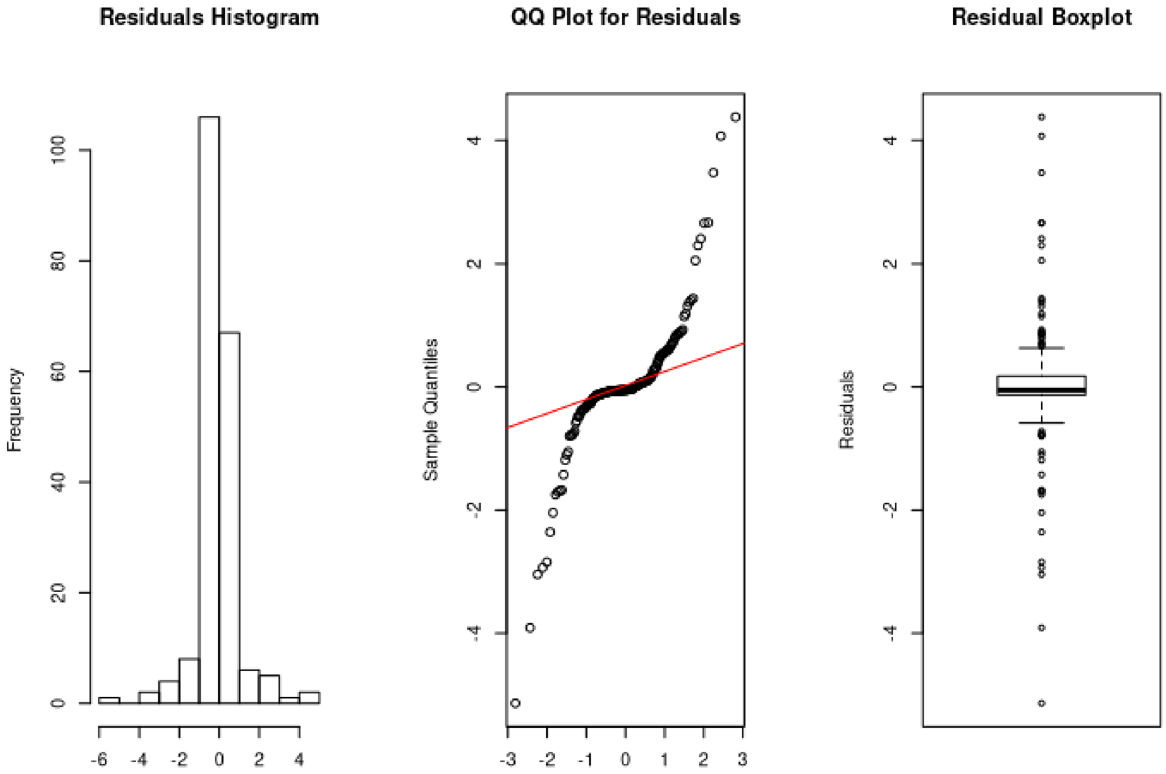


Figure 4.8: Residual plots for the ARMA(1, 1) model for the FLE data.

```
ARMA(1, 2) for the detrended FLE data

Shapiro-Wilk normality test
W = 0.81151, p-value = 6.773e-15

Exact runs test
Runs = 66, p-value = 4.192e-07
alternative hypothesis: two.sided

> AIC(model)
[1] 962.246
> BIC(model)
[1] 978.7874
```

From these we can conclude that ARMA(1, 1) is preferred model. However, to obtain a better fit, we use the auto.arima function and check the effect of transformation. The model summary including the AIC and BIC values and the confidence intervals of the parameters are given here.

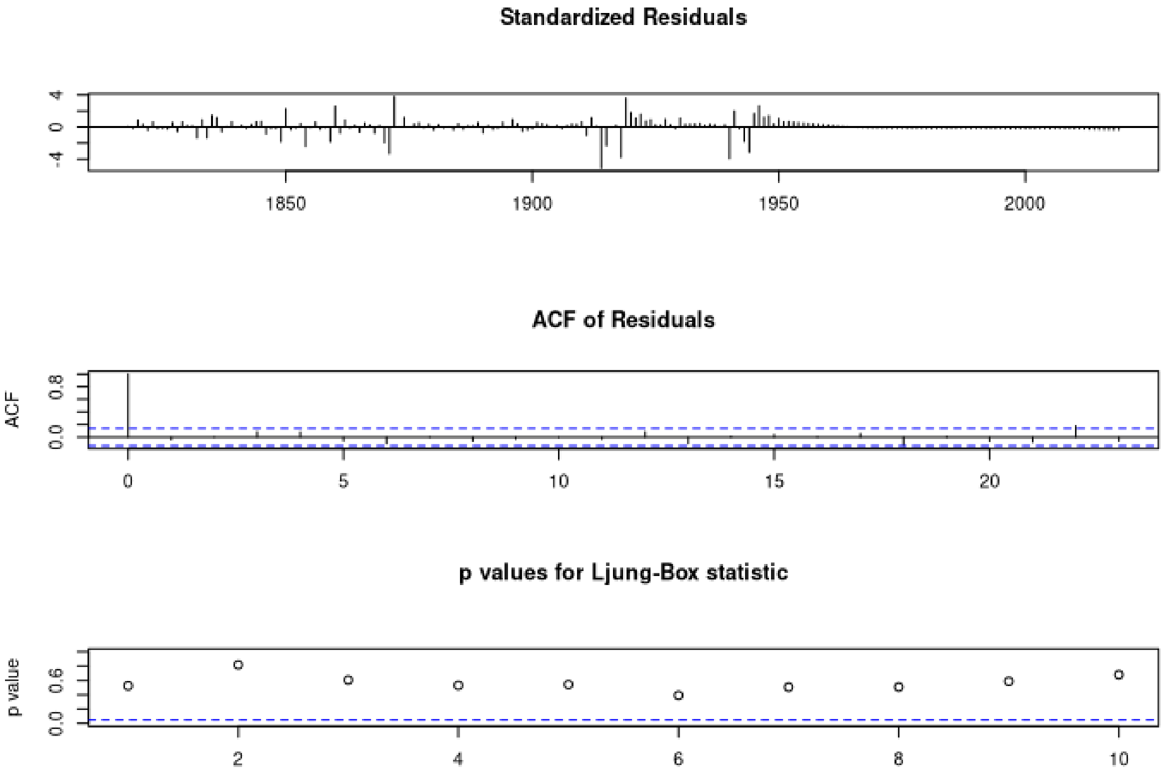
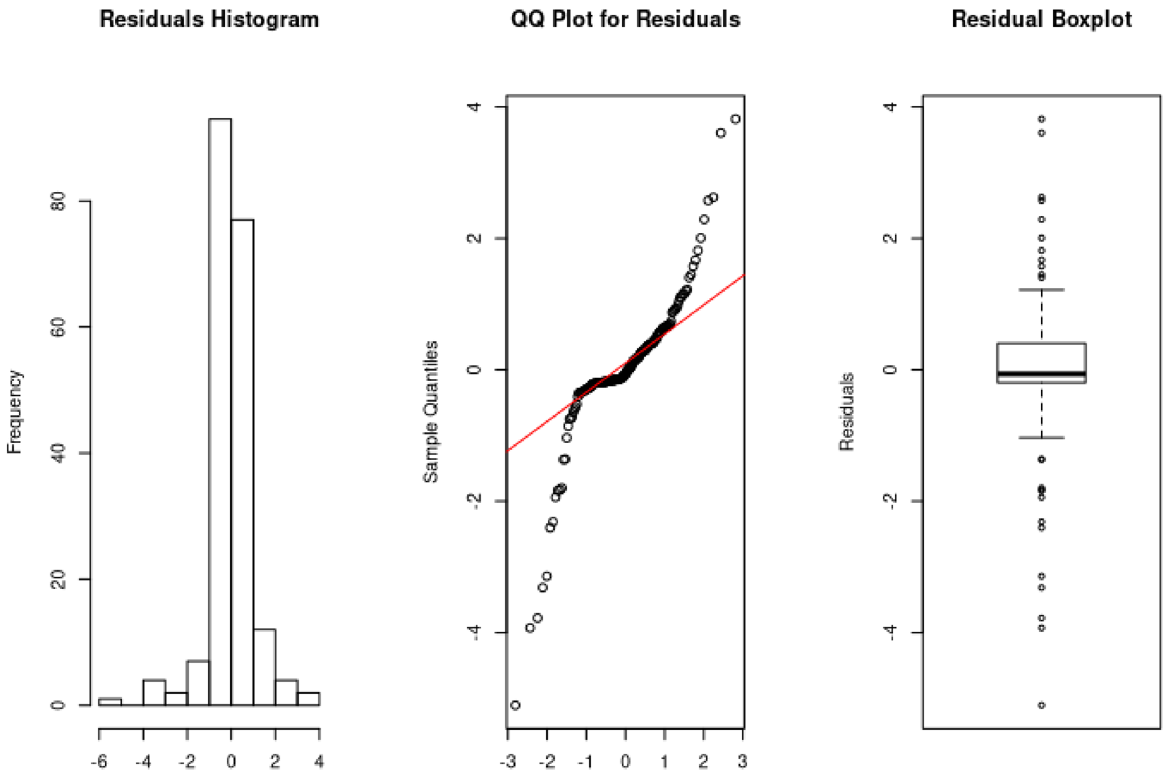


Figure 4.9: Residual diagnostics for the ARMA(1, 2) model for the FLE data.

```
> fle.auto.arima = auto.arima(fle.detrend)
> fle.auto.arima
Series: fle.detrend
ARIMA(1,0,3) with zero mean

Coefficients:
         ar1      ma1     ma2      ma3
      0.7382  -2.0855  1.2361  -0.1467
s.e.  0.0960   0.1212  0.2218   0.1026

sigma^2 estimated as 6.436:  log likelihood=-475.51
AIC=961.01   AICc=961.32   BIC=977.55

> fle.bc.auto.arima = auto.arima(fle.detrend, lambda="auto")
> fle.bc.auto.arima
Series: fle.detrend
ARIMA(1,0,2) with non-zero mean
Box Cox transformation: lambda= 0.7566293

Coefficients:
         ar1      ma1     ma2     mean
```



Residuals Histogram

QQ Plot for Residuals

Residual Boxplot

Figure 4.10: Residual plots for the ARMA(1, 2) model for the FLE data.

```
      0.5021  -1.6531  0.7389  -1.3864
s.e.  0.2453   0.1990  0.1701   0.0253

sigma^2 estimated as 4.409:  log likelihood=-435.52
AIC=881.04   AICc=881.34   BIC=897.58

> confint(fle.bc.auto.arima)
                 2.5 %     97.5 %
ar1        0.02135014  0.9827585
ma1       -2.04318152 -1.2630966
ma2        0.40548839  1.0722244
intercept -1.43600896 -1.3367935
```

Here, we see that the transformed dataset has lower AIC and BIC values along with lesser parameters wherein the confidence intervals do not contain zero. Therefore, data transformation is required.

So from the above analyses, the final model chosen is the ARIMA(1, 2, 2) model (ARMA(1, 2) for the second order differences) where BoxCox transformation ($\lambda \approx 0.75$) is used. The model equation is given by Eq. (3).

$$Z_t = -1.3864 + 0.5021Z_{t-1} + 1.6531\varepsilon_{t-1} - 0.7389\varepsilon_{t-2} + \varepsilon_t \tag{3}$$

where $\varepsilon \sim \mathcal{N}(0, 4.409)$ and $Z_t = \text{BoxCox}(\nabla^2 X_t,\, \lambda)$, $\lambda \approx 0.75$, $t = \{1, 2, 3, \dots\}$, and $X_t$ being the original FLE data.

If the entire duration of the time series is considered, then the life expectancy has greatly varied before 1950 but there is an overall increase in the life expectancy. From 1950 onward, the life expectancy keeps on increasing linearly.
When individual years are taken into account, then the sharp increments and decrements in the values can be attributed to the improved healthcare systems and degrading economy or wars respectively.