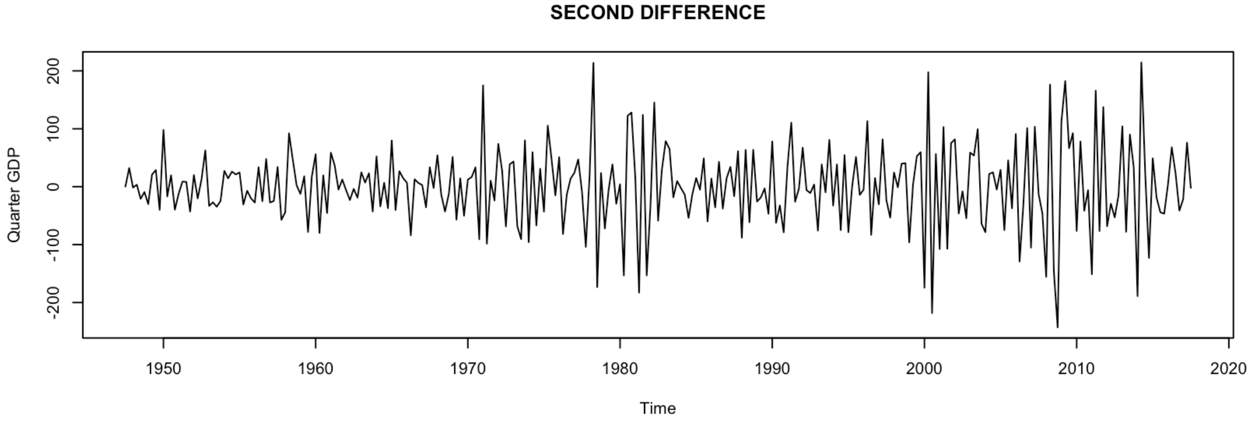
Time Series Analysis and Forecasting for US GDP

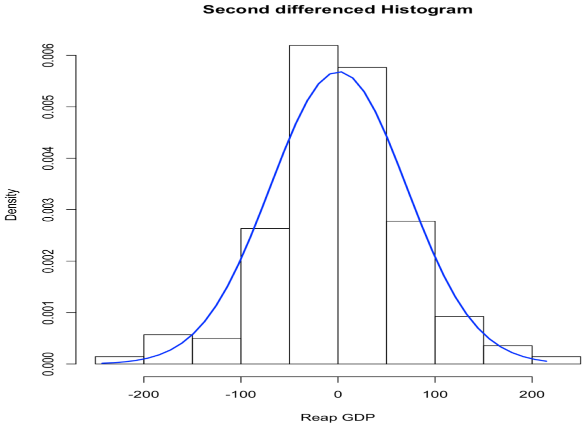
EXPLORATORY DATA ANAYSIS

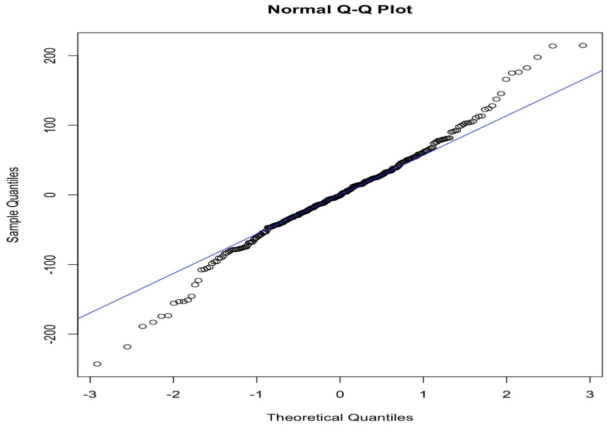
The dataset used for our analysis is US's Real Gross Domestic Product as reported by the Federal Reserve bank of St. Louis. Real GDP is seasonally adjusted. (More detail about the data is found at https://alfred.stlouisfed.org/series?seid=GDPC1). The dataset includes quarterly date from first quarter in 1947 to third quarter in 2017 and Quarterly GDP in billions of 2009 dollars. There are 283 observations on both data (*Figure 1*).

By running a plot of values over time, the data shows strong evidence of a linear up-ward trend gradually with a crash in 2008 (*Figure 2*). Also, the histogram shows that the dataset is not normally distributed(*Figure 3*). Since the linear trend is detected and the dataset has no normal distribution, it gives us to decide to conducting differencing. After the differencing, the time plot shows fluctuations over time with constant mean as seen in *Figure 2*. There appears to be a slight positive trend over time as well as non-consistent variance. Additionally, its histogram shows that the first differenced data is left-skewed distributed. Therefore, we conducted second differencing of raw dataset. The time plot shows ideal time plot with a constant mean and variance over time, and the linear trend is removed. Thus, it has a stationarity.



According to normality tests, the histogram shows that the data is normally distributed and qq-plot shows that the dataset is close to the normal distribution.





Based on the basic statistics function (*Figure 3*), skewness is -0.073 which indicates the data is skewed to left, but very close to zero. Also, kurtosis is 1.155 which is less than 3 of the normal distribution. However, the skewness and kurtosis provide provide all evidence to be close to the normal distribution. Furthermore, Jarque-Bera nomality test in *Figure 5*, we accept the null hypothesis of the data is normally distributed. Therefore, we conclude that the appropriate transformed data is identified since second differenced gdp data is normally distributed and the time plot is stationary

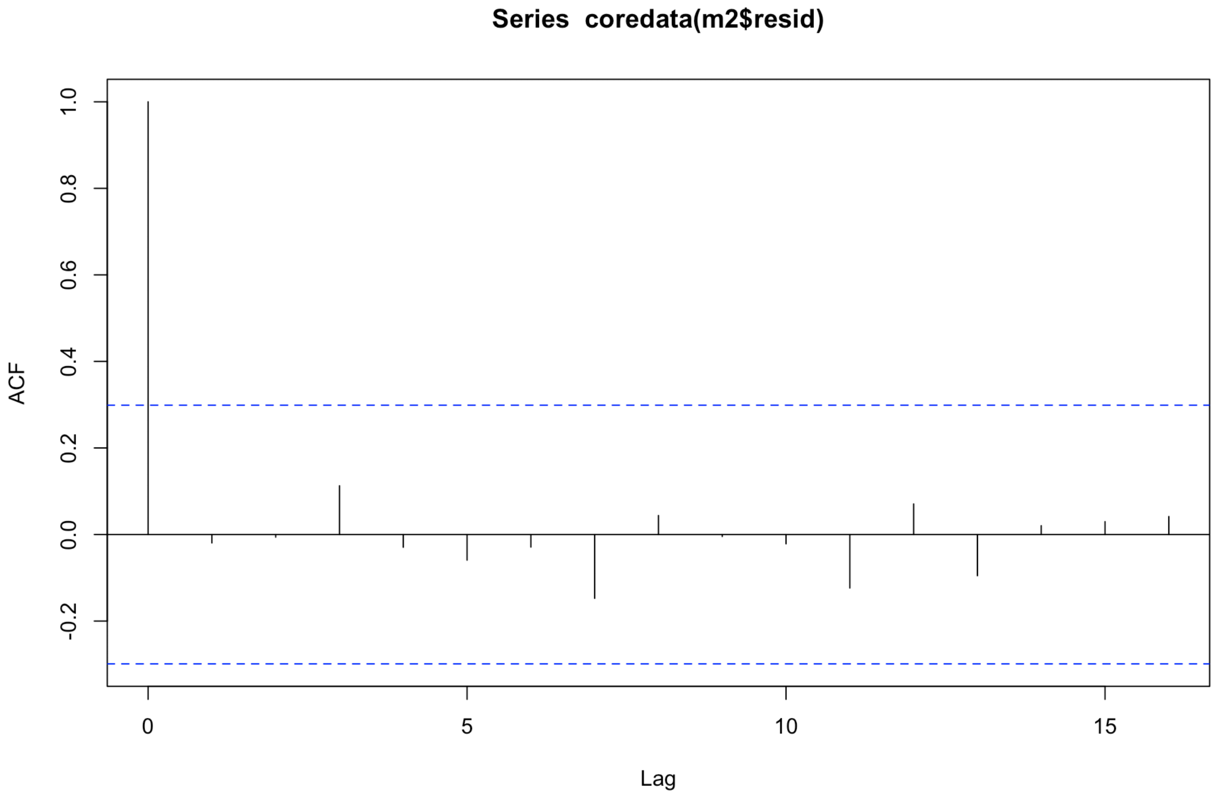
MODEL APPROACHING AND DIAGNOSTICS

Given the data, autoregressive function (ACF) of second differenced data shows zero-autocorrelations after lag 1 while Partial ACF (PACF) shows the correlograms are slowly decaying. Thus, the first attempt to fit the data is ARIMA(0,2,1) with BIC score of 3141.81. However, this model is concluded as an inadequate model by diagnostics from both residuals ACF test (*Figure 10*) and the Ljung-box(LB) test (*Figure 11*). The ACF test shows that there exists few spikes out of the significance level at lag2, 9, and 10. This is a strong evidence that the residuals are not white noise. Additionally, we accept the null hypothesis of zero-serial correlations. Therefore, both residuals on ACF and residuals on LB-test conclude that this model is not an adequate model for our analysis.

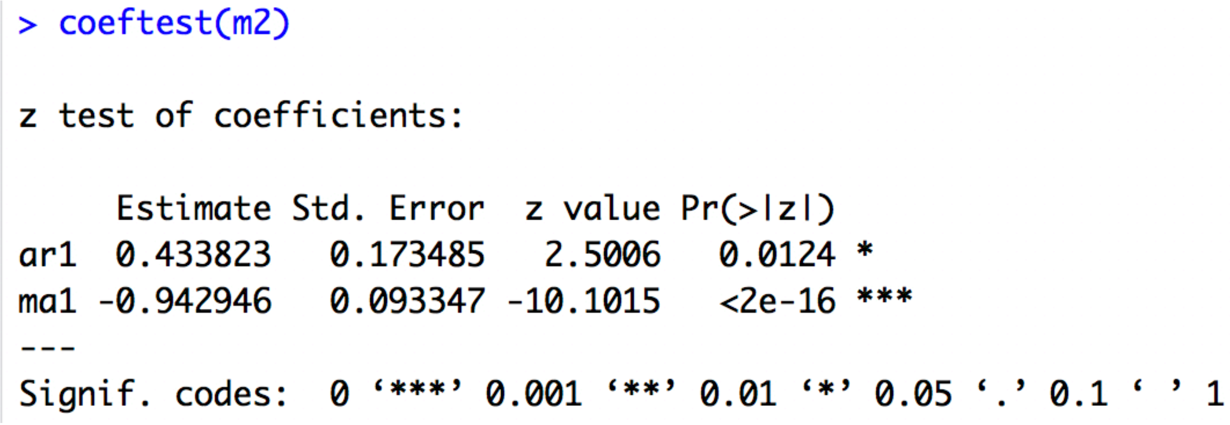
Moreover, to select a better model, we used auto ARIMA function, and it selects ARIMA(1,2,1) with BIC score of 3104.28. Compared to first model, BIC score gets lower by 37.53. The residuals on ACF gives zero autocorrelations, except at lag 2 (*Figure 13*) so that this model is not white noise perfectly. By testing residuals on Ljung box (*Figure 14*), we reject the null hypothesis of white noise with lag 3. However, the p-values of the tests gets higher than 0.05 after lag 6. Hence, LB-test shows that the residuals are not perfect white noise. We conclude that the model is not an adequate model for our analysis too. However, we consider ARIMA(1,2,1) as our model since only one flaw is detected from each residual analyses.

Lastly, we decided to windowed time frame from 2007 to 2017 and used auto arima function. It gives us with ARIMA(1,1,0) with drift (*Figure 16)* and BIC score of 497.49. Since BIC score here is the lowest, we think this model with shorten time frame would be a good model. Residuals on

ACF finally shows a perfect white noise.



Also, Ljung-box test on residuals shows statistically great p-values with this model (*Figure 17*). Coefficient test with significance level of 5% shows very significant p-value.

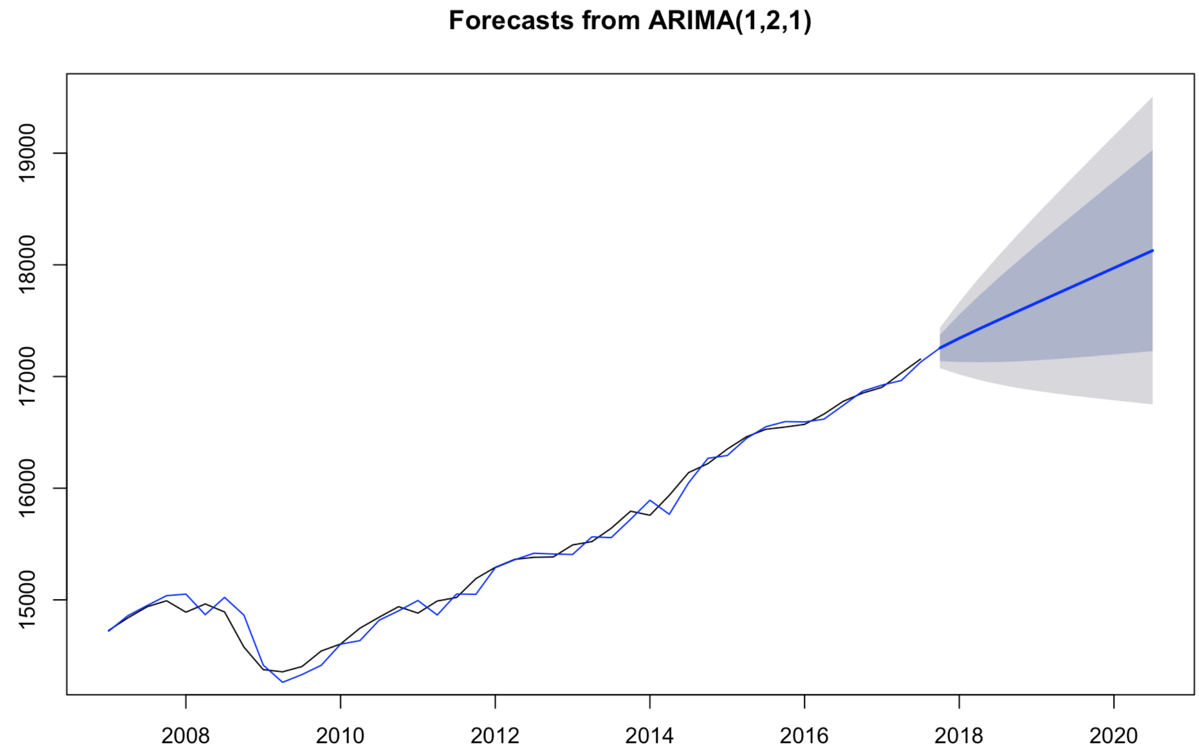


```
> coeftest(m2)

z test of coefficients:

     Estimate Std. Error  z value Pr(>|z|)
ar1  0.433823   0.173485   2.5006   0.0124 *
ma1 -0.942946   0.093347 -10.1015   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Therefore, we can conclude that the model ARIMA(1,2,1) with last decade gives us a best model to fit the data, and the expression of the model is $\mathrm{X}_{\mathrm{t}} = 0.4334\,\mathrm{X}_{\mathrm{t-1}} + e_t - 0.943e_{t-1}$.

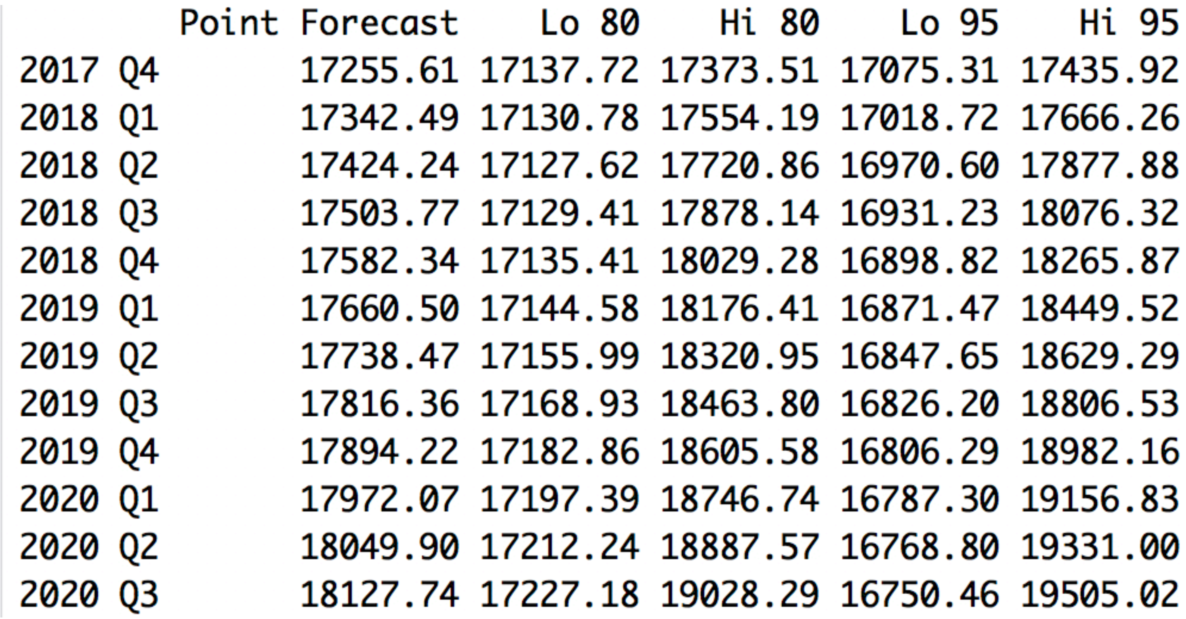
FORECASTING

Our model appears to forecast the trend of the actual values well.



According to Backtesting, the difference between our prediction value and true value are 0.659 percentage far away.

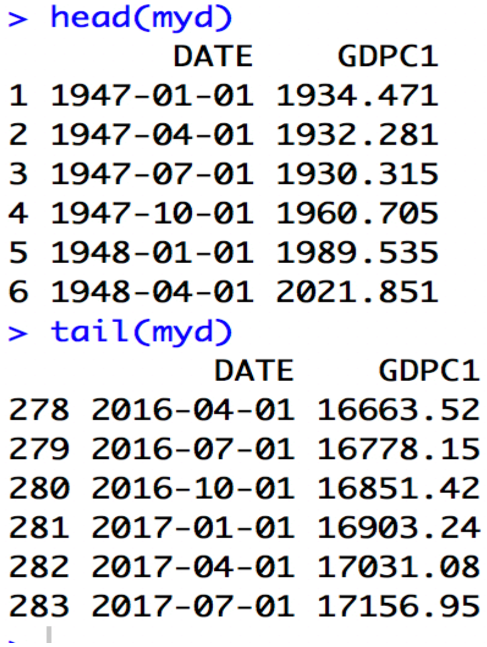
Projecting ahead fourth quarter of 2017 through fourth quarter of 2019, the results shows below



```
        Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2017 Q4       17255.61 17137.72 17373.51 17075.31 17435.92
2018 Q1       17342.49 17130.78 17554.19 17018.72 17666.26
2018 Q2       17424.24 17127.62 17720.86 16970.60 17877.88
2018 Q3       17503.77 17129.41 17878.14 16931.23 18076.32
2018 Q4       17582.34 17135.41 18029.28 16898.82 18265.87
2019 Q1       17660.50 17144.58 18176.41 16871.47 18449.52
2019 Q2       17738.47 17155.99 18320.95 16847.65 18629.29
2019 Q3       17816.36 17168.93 18463.80 16826.20 18806.53
2019 Q4       17894.22 17182.86 18605.58 16806.29 18982.16
2020 Q1       17972.07 17197.39 18746.74 16787.30 19156.83
2020 Q2       18049.90 17212.24 18887.57 16768.80 19331.00
2020 Q3       18127.74 17227.18 19028.29 16750.46 19505.02
```

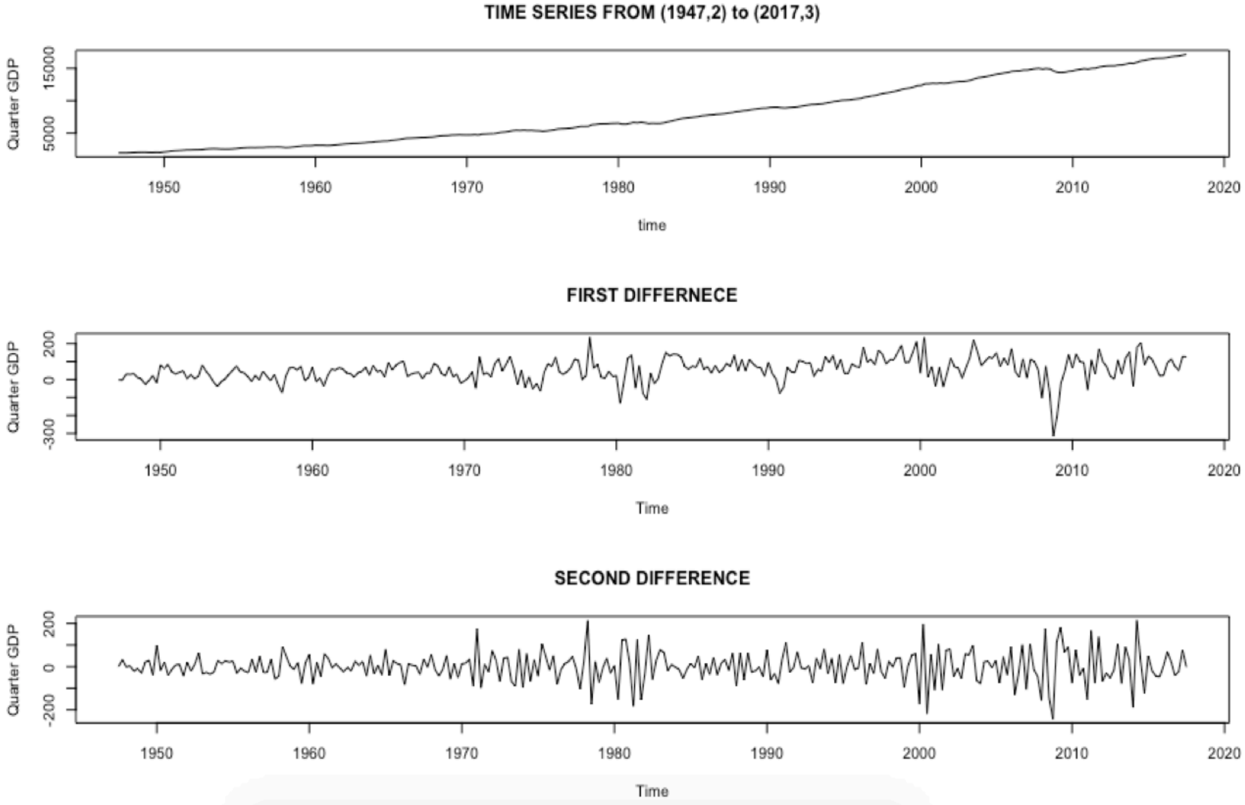
APPENDEX

Figure 1. GDP DATASETS



```
> head(myd)
        DATE    GDPC1
1 1947-01-01 1934.471
2 1947-04-01 1932.281
3 1947-07-01 1930.315
4 1947-10-01 1960.705
5 1948-01-01 1989.535
6 1948-04-01 2021.851
> tail(myd)
          DATE    GDPC1
278 2016-04-01 16663.52
279 2016-07-01 16778.15
280 2016-10-01 16851.42
281 2017-01-01 16903.24
282 2017-04-01 17031.08
283 2017-07-01 17156.95
```

Figure 2. Time plot of raw data, first differenced data, second differenced data



TIME SERIES FROM (1947,2) to (2017,3)

FIRST DIFFERNECE

SECOND DIFFERENCE

Figure 3. HISTOGRAM FOR RAW DATA and FIRST DIFFERENCED DATA

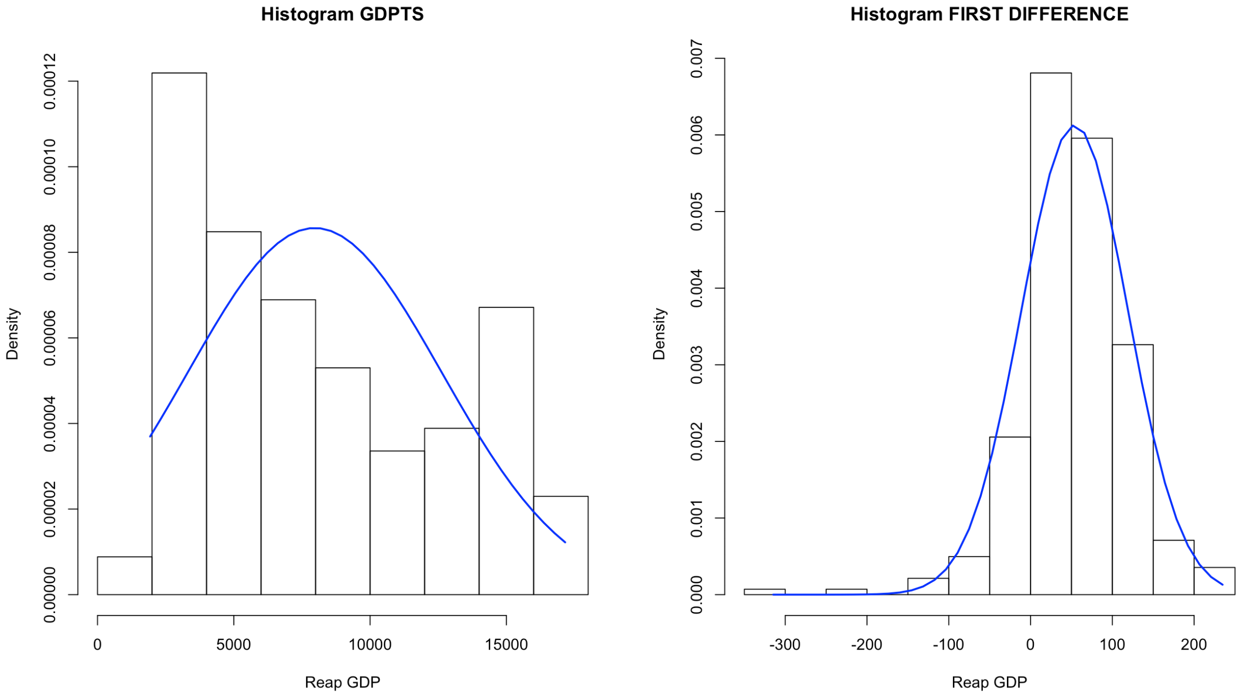
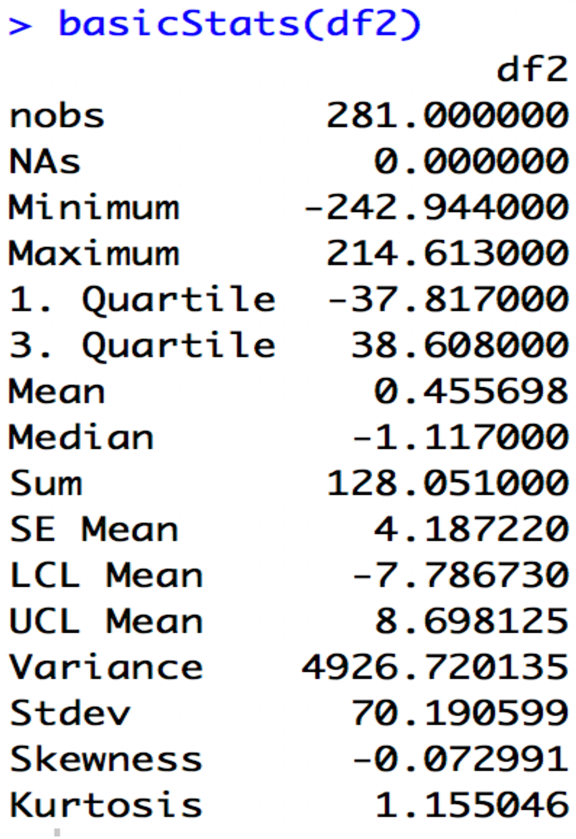
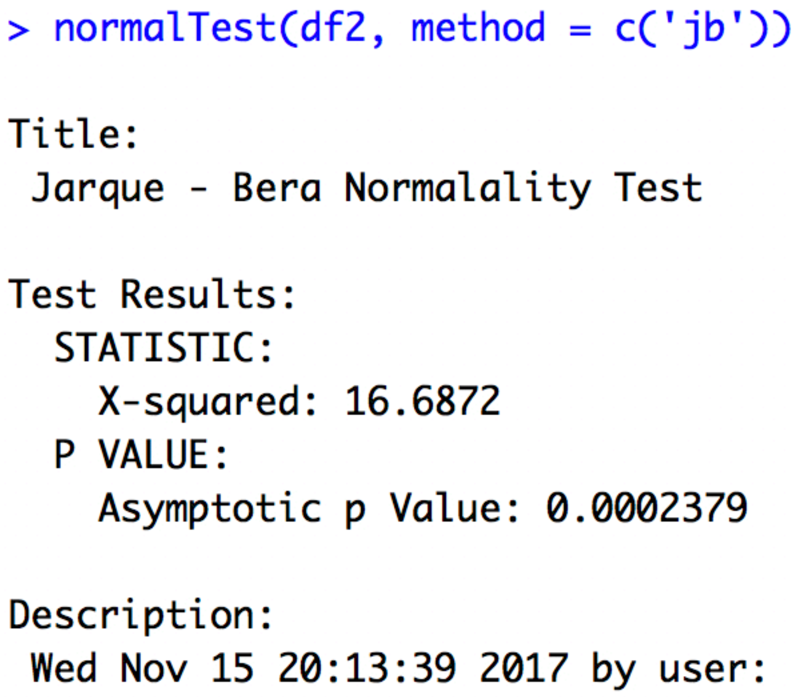


Figure 4. BASIC STATISTICS for SECOND DIFFERENCED DATA



```
> basicStats(df2)
                    df2
nobs         281.000000
NAs            0.000000
Minimum     -242.944000
Maximum      214.613000
1. Quartile  -37.817000
3. Quartile   38.608000
Mean           0.455698
Median        -1.117000
Sum          128.051000
SE Mean        4.187220
LCL Mean      -7.786730
UCL Mean       8.698125
Variance    4926.720135
Stdev         70.190599
Skewness      -0.072991
Kurtosis       1.155046
```

Figure 5. JARQUE BERA NORMALITY TEST for SECONED DIFFERENCED DATA



```
> normalTest(df2, method = c('jb'))

Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 16.6872
  P VALUE:
    Asymptotic p Value: 0.0002379

Description:
 Wed Nov 15 20:13:39 2017 by user:
```

Figure 6. ACF and PACF of RAW DATASET

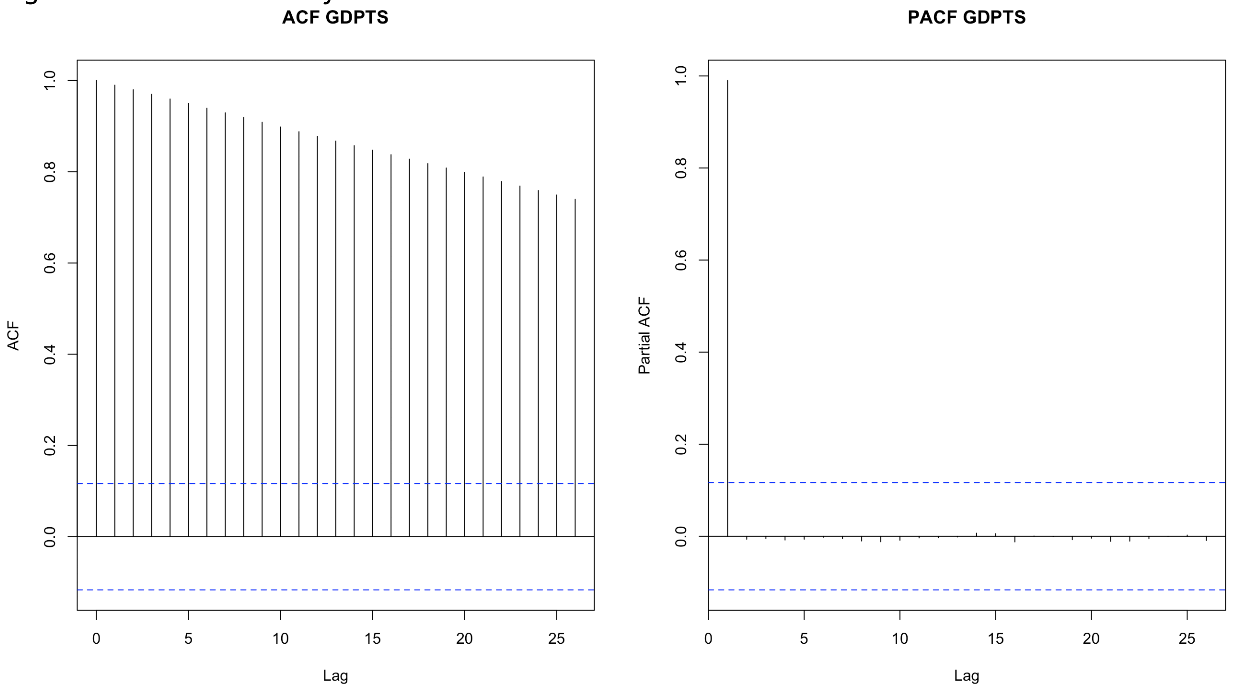


Figure 7. ACF and PACF of FIRST DIFFERENCED DATASET

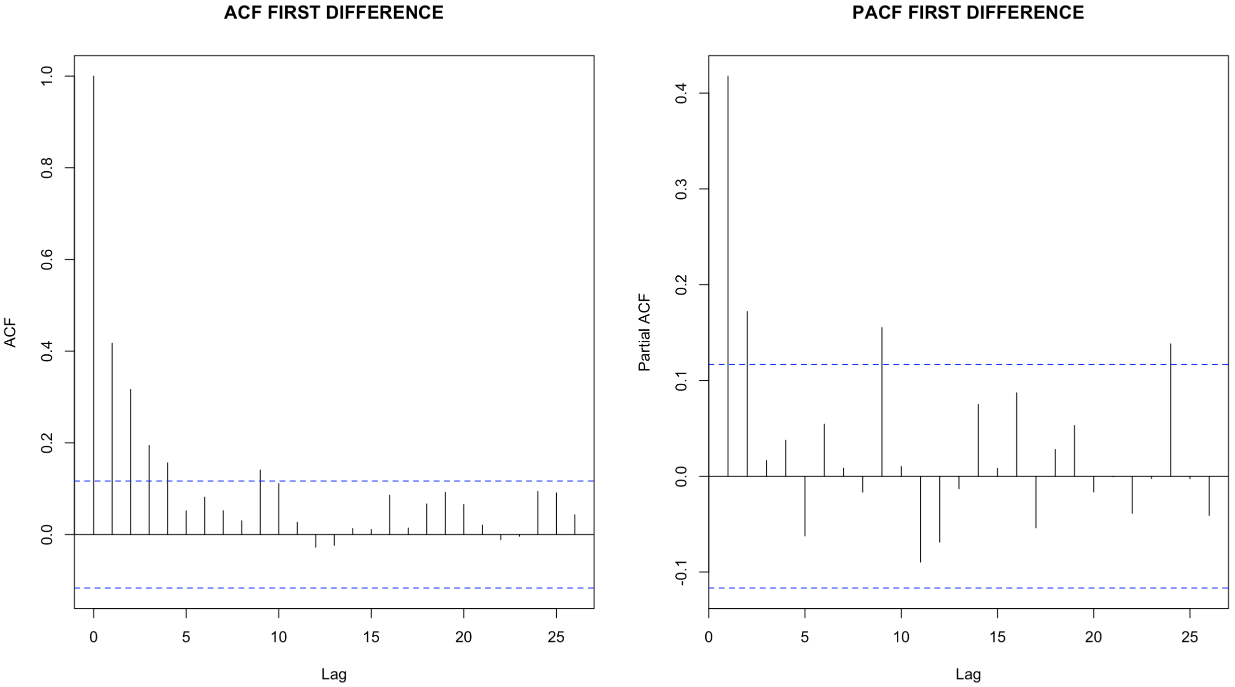


Figure 8. ACF and PACF of SECOND DIFFERENCED DATASET

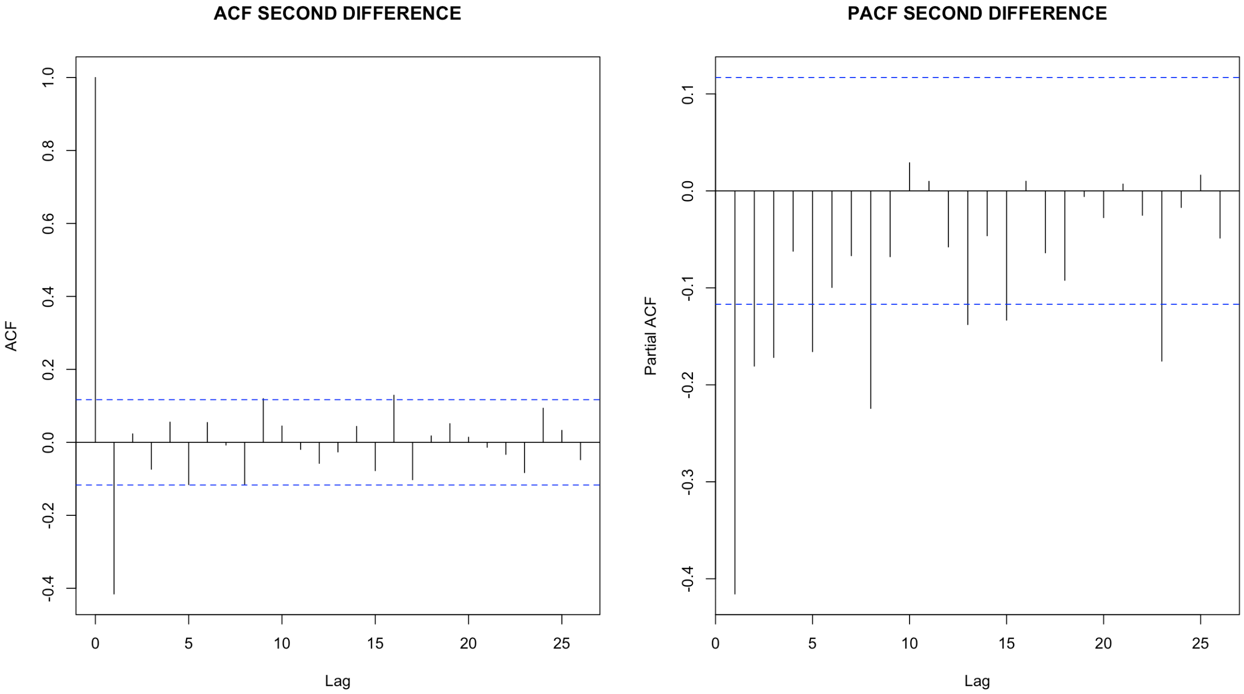
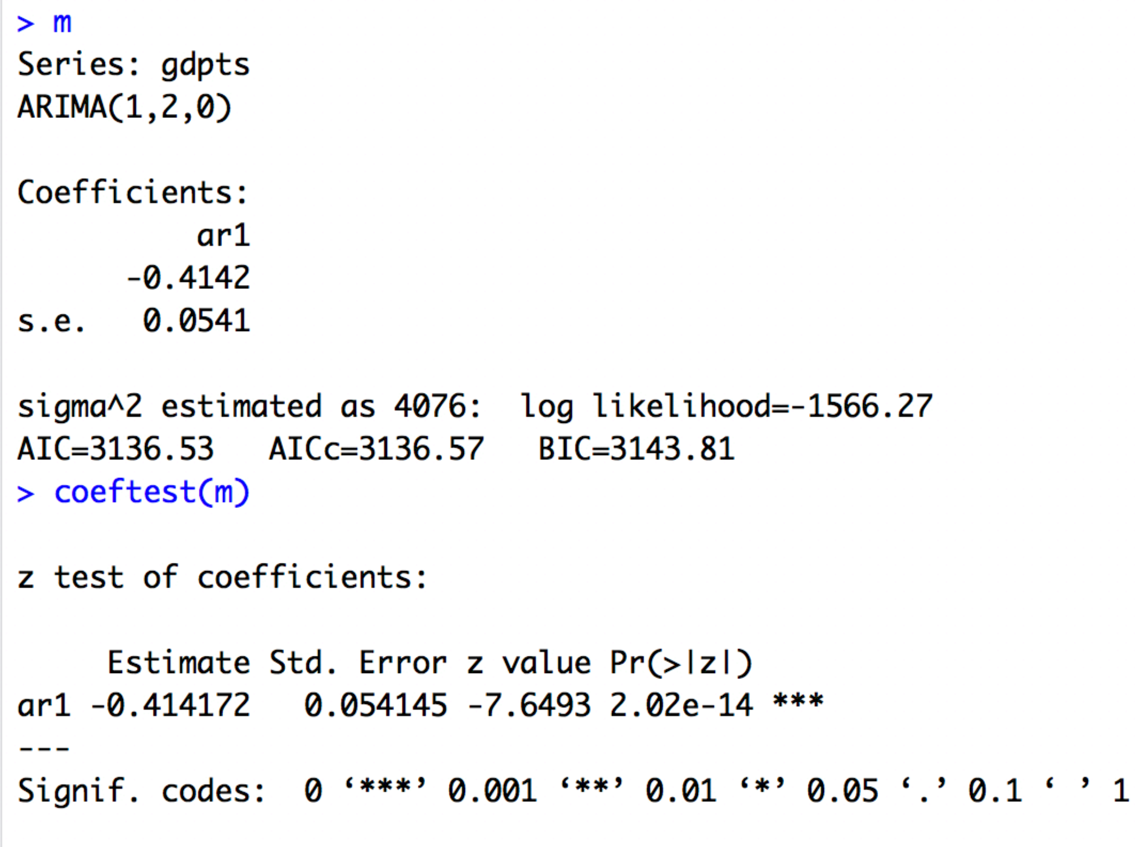


Figure 9. FIRST MODEL ARIMA(1,2,0) and COEFFICIENT TEST



```
> m
Series: gdpts
ARIMA(1,2,0)

Coefficients:
          ar1
      -0.4142
s.e.   0.0541

sigma^2 estimated as 4076:  log likelihood=-1566.27
AIC=3136.53   AICc=3136.57   BIC=3143.81
> coeftest(m)

z test of coefficients:

     Estimate Std. Error z value Pr(>|z|)
ar1 -0.414172   0.054145 -7.6493 2.02e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10. FIRST MODEL ARIMA(1,2,0) RESIDUALS on ACF

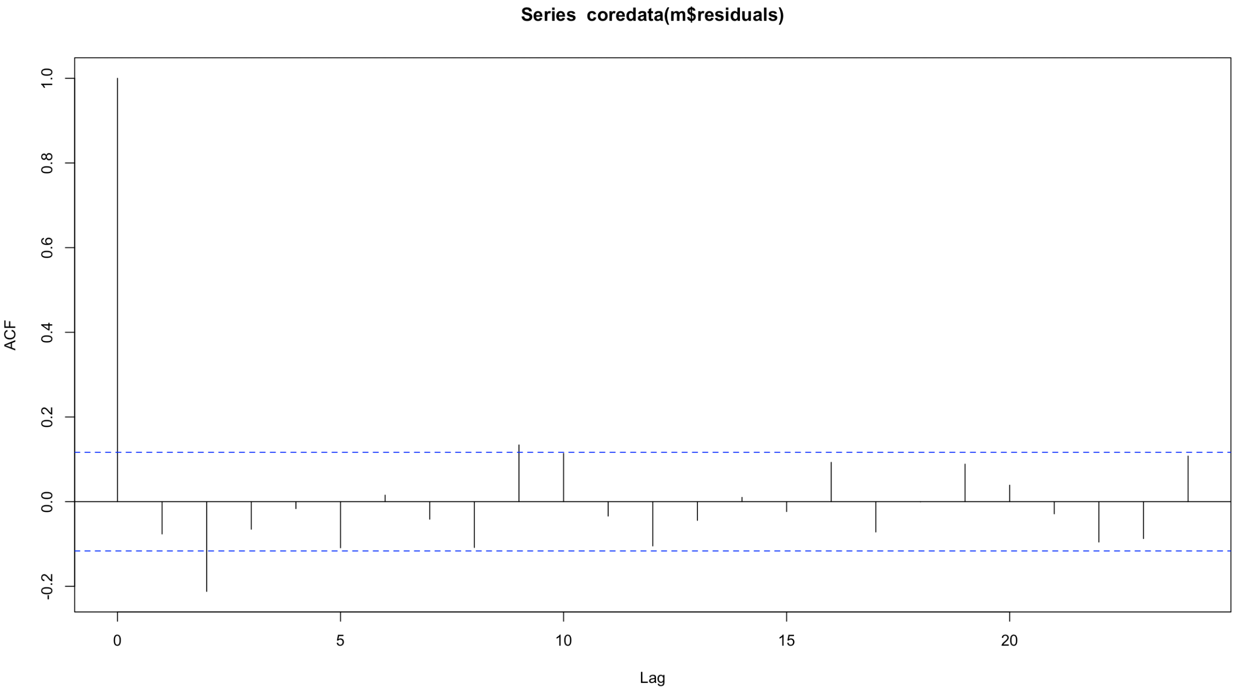
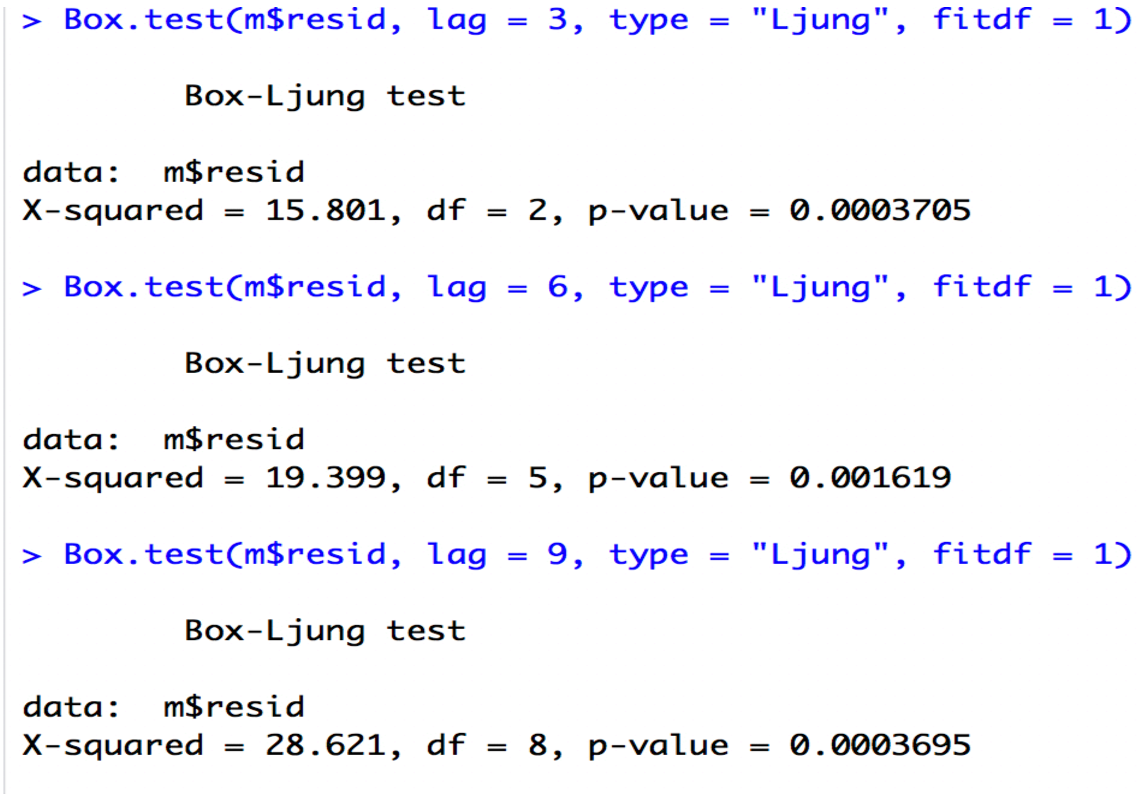
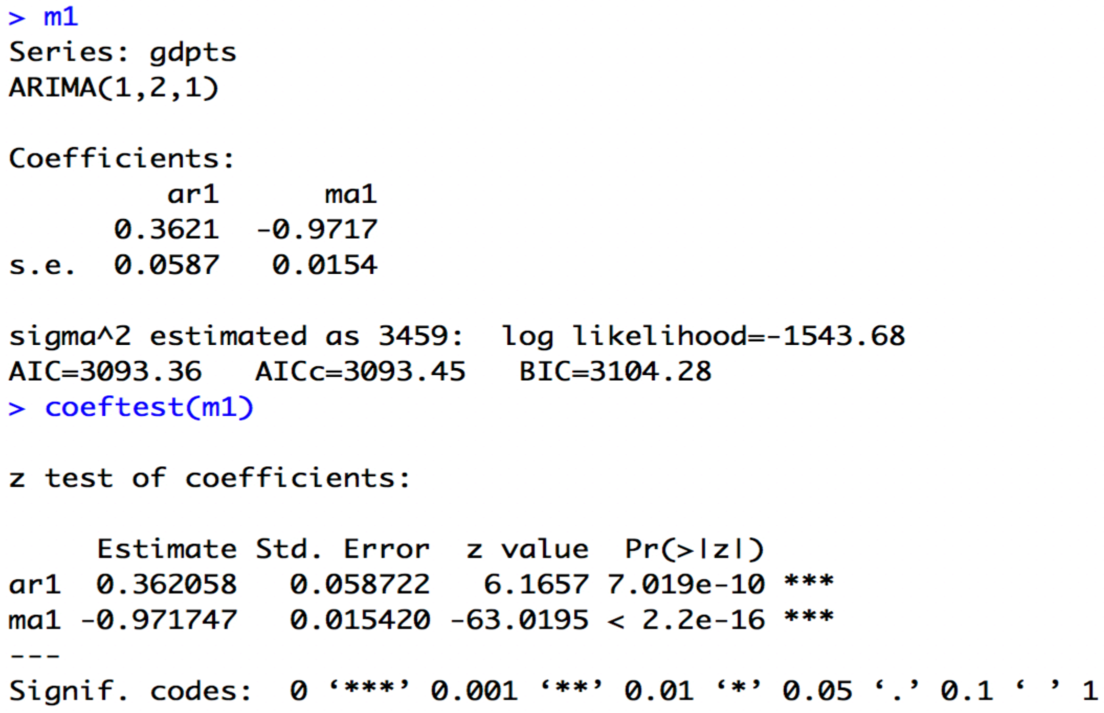


Figure 11. LJUNG BOX TEST ON FIRST MODEL ARIMA(1,2,0)



```
> Box.test(m$resid, lag = 3, type = "Ljung", fitdf = 1)

        Box-Ljung test

data:  m$resid
X-squared = 15.801, df = 2, p-value = 0.0003705

> Box.test(m$resid, lag = 6, type = "Ljung", fitdf = 1)

        Box-Ljung test

data:  m$resid
X-squared = 19.399, df = 5, p-value = 0.001619

> Box.test(m$resid, lag = 9, type = "Ljung", fitdf = 1)

        Box-Ljung test

data:  m$resid
X-squared = 28.621, df = 8, p-value = 0.0003695
```

Figure 12. SECOND MODEL AUTO-ARIMA(1,2,1) and Its COEFFICIENT TEST



```
> m1
Series: gdpts
ARIMA(1,2,1)

Coefficients:
         ar1      ma1
      0.3621  -0.9717
s.e.  0.0587   0.0154

sigma^2 estimated as 3459:  log likelihood=-1543.68
AIC=3093.36   AICc=3093.45   BIC=3104.28
> coeftest(m1)

z test of coefficients:

     Estimate Std. Error  z value  Pr(>|z|)    
ar1  0.362058   0.058722   6.1657 7.019e-10 ***
ma1 -0.971747   0.015420 -63.0195 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 13. SECOND MODEL AUTO-ARIMA(1,2,1)RESIDUALS on ACF

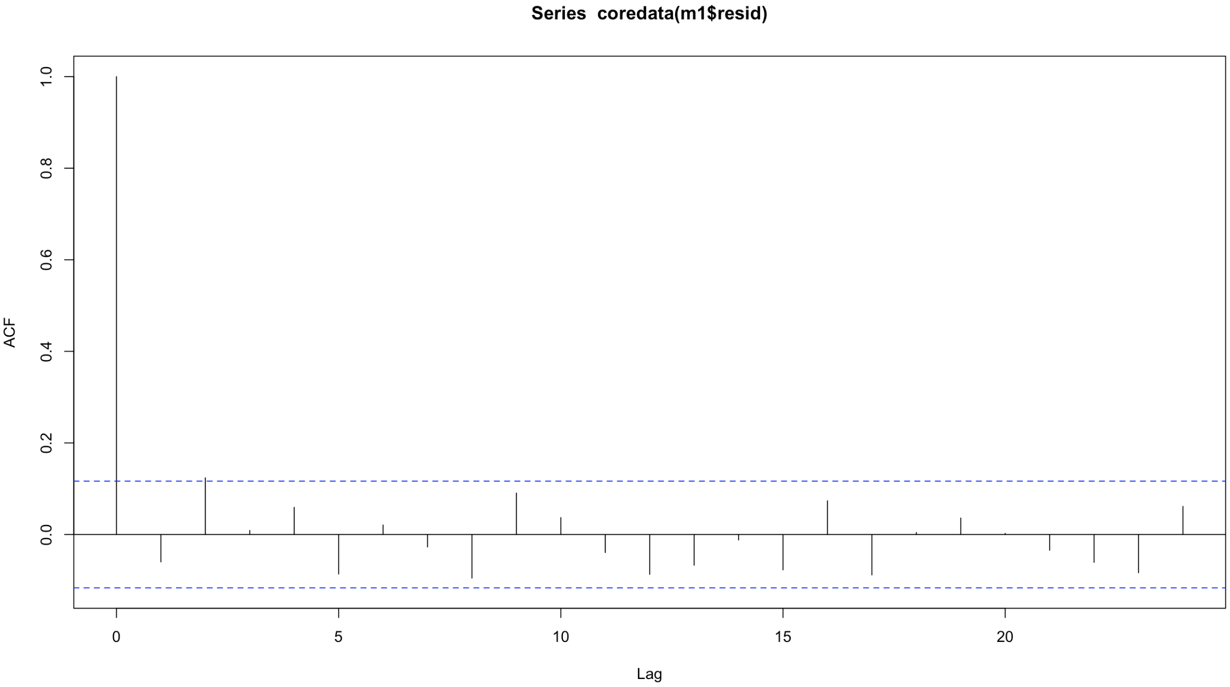
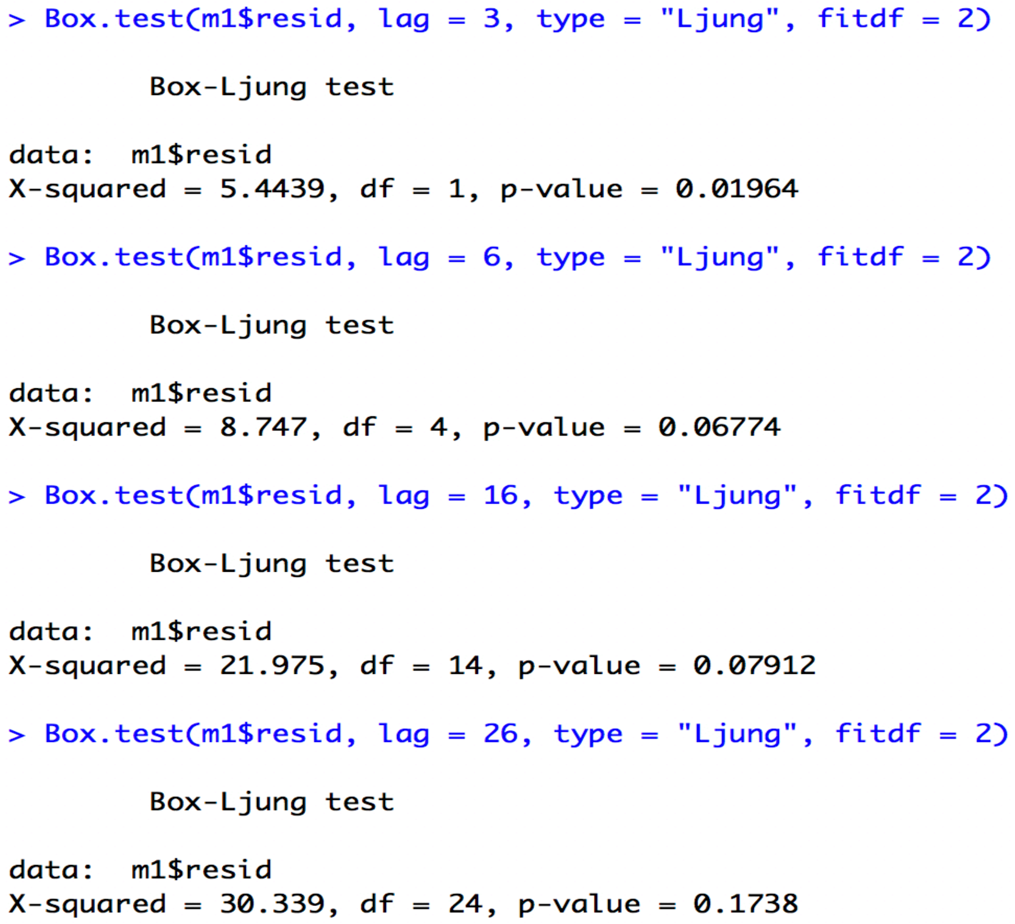


Figure 14. LJUNG BOX TEST ON SECOND MODEL AUTO-ARIMA(1,2,1)



```
> Box.test(m1$resid, lag = 3, type = "Ljung", fitdf = 2)

        Box-Ljung test

data:  m1$resid
X-squared = 5.4439, df = 1, p-value = 0.01964

> Box.test(m1$resid, lag = 6, type = "Ljung", fitdf = 2)

        Box-Ljung test

data:  m1$resid
X-squared = 8.747, df = 4, p-value = 0.06774

> Box.test(m1$resid, lag = 16, type = "Ljung", fitdf = 2)

        Box-Ljung test

data:  m1$resid
X-squared = 21.975, df = 14, p-value = 0.07912

> Box.test(m1$resid, lag = 26, type = "Ljung", fitdf = 2)

        Box-Ljung test

data:  m1$resid
X-squared = 30.339, df = 24, p-value = 0.1738
```

Figure 15. HISTOGRAM and QQ-PLOT of SECOND MODEL RESIDUALS

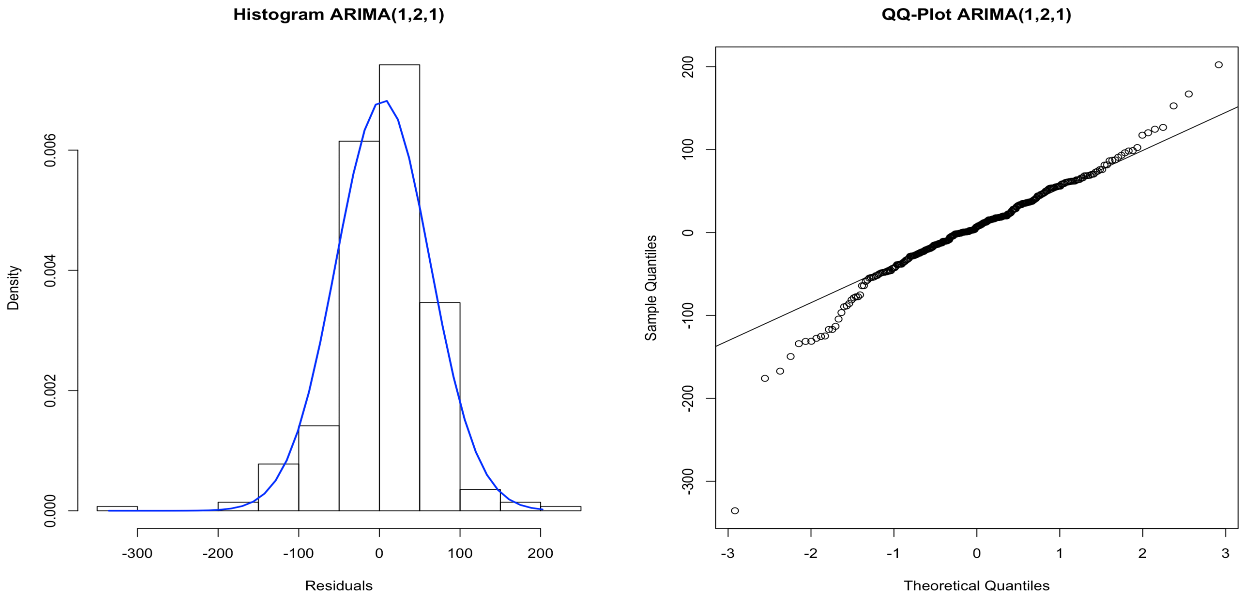
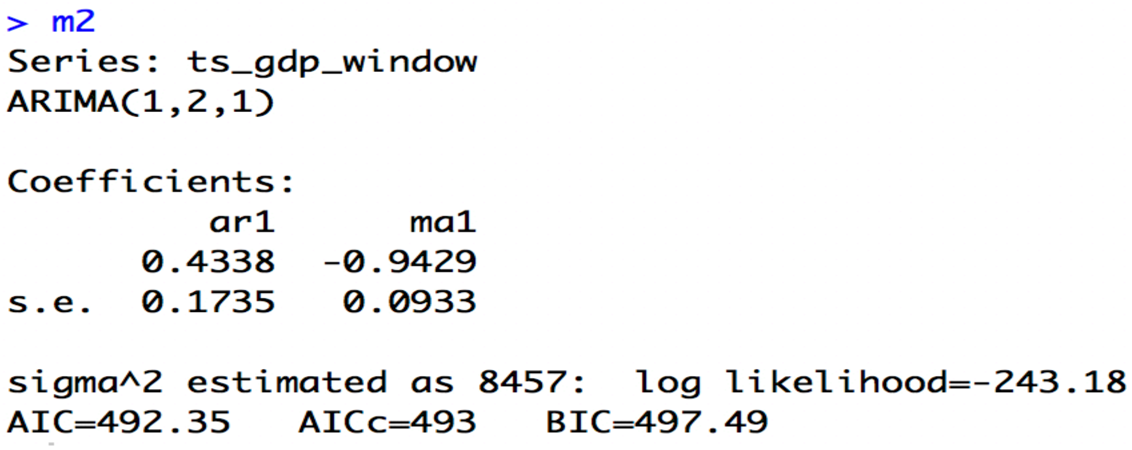
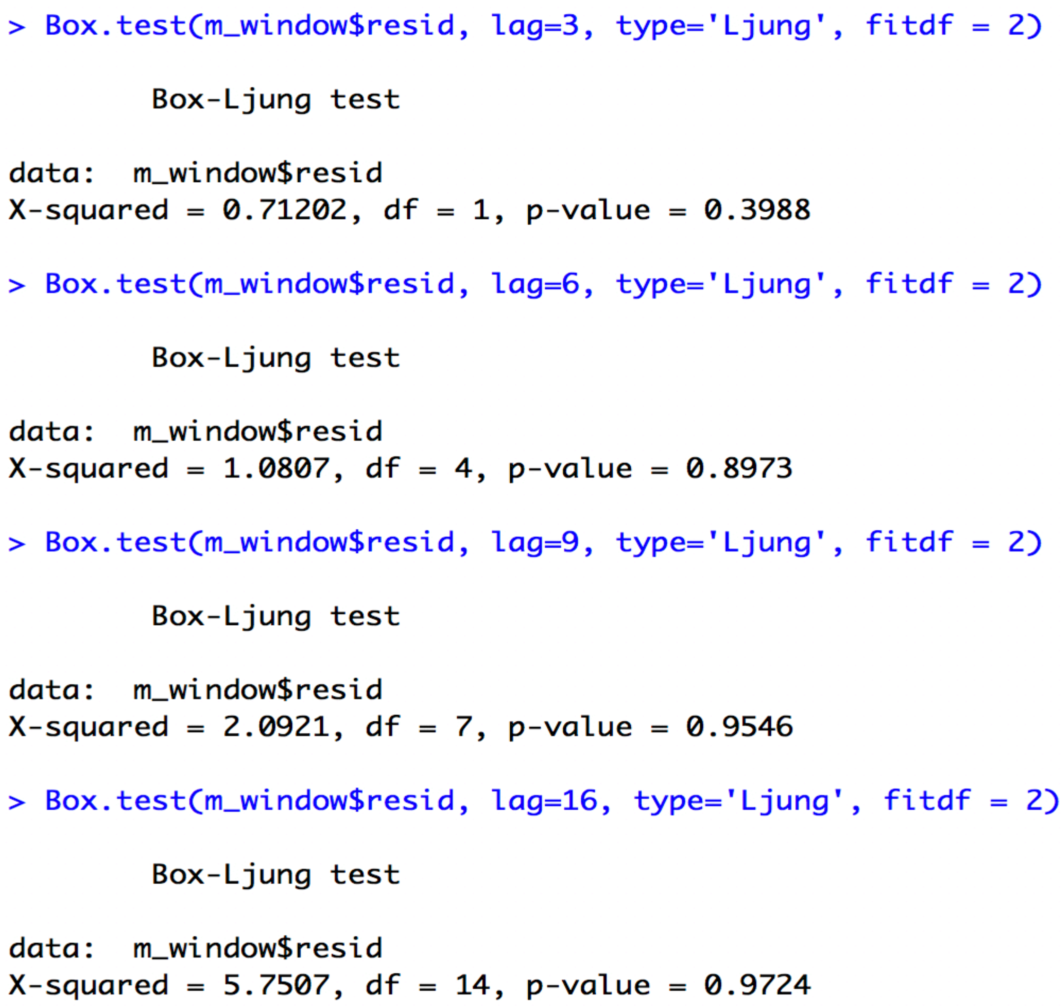


Figure 16. THE THIRD MODEL WINDOWED ARIMA(1,2,1) FROM 2007 to 2017 and COEFFICIENT TEST



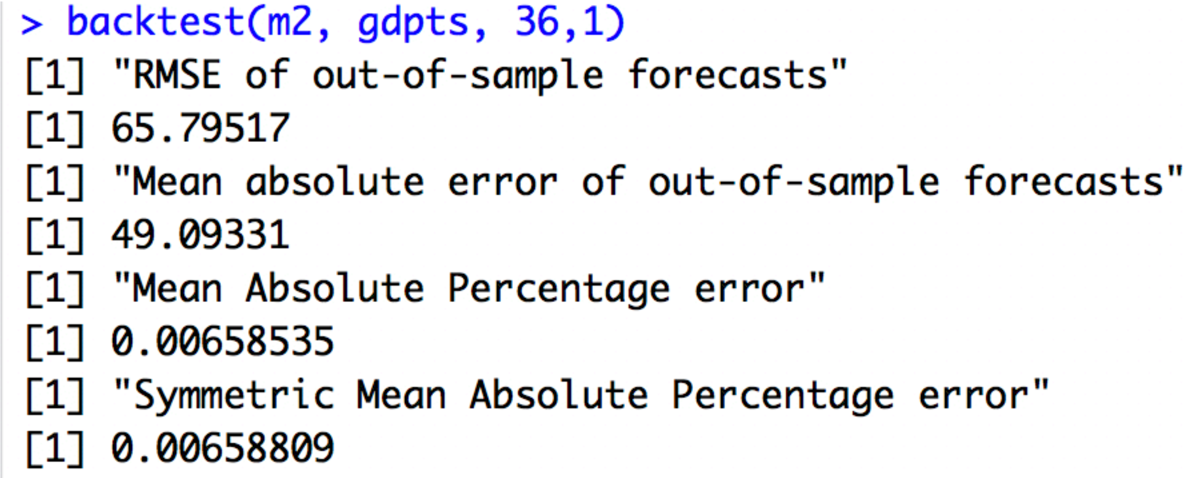
```
> m2
Series: ts_gdp_window
ARIMA(1,2,1)

Coefficients:
         ar1      ma1
      0.4338  -0.9429
s.e.  0.1735   0.0933

sigma^2 estimated as 8457:  log likelihood=-243.18
AIC=492.35   AICc=493   BIC=497.49
```

Figure 17. LJUNG-BOX TEST ON THE THIRD MODEL WINDOWED ARIMA(1,2,1)



```
> Box.test(m_window$resid, lag=3, type='Ljung', fitdf = 2)

	Box-Ljung test

data:  m_window$resid
X-squared = 0.71202, df = 1, p-value = 0.3988

> Box.test(m_window$resid, lag=6, type='Ljung', fitdf = 2)

	Box-Ljung test

data:  m_window$resid
X-squared = 1.0807, df = 4, p-value = 0.8973

> Box.test(m_window$resid, lag=9, type='Ljung', fitdf = 2)

	Box-Ljung test

data:  m_window$resid
X-squared = 2.0921, df = 7, p-value = 0.9546

> Box.test(m_window$resid, lag=16, type='Ljung', fitdf = 2)

	Box-Ljung test

data:  m_window$resid
X-squared = 5.7507, df = 14, p-value = 0.9724
```

Figure 18. BACKTESTING THE THIRD MODEL WINDOWED ARIMA(1,2,1)



```
> backtest(m2, gdpts, 36,1)
[1] "RMSE of out-of-sample forecasts"
[1] 65.79517
[1] "Mean absolute error of out-of-sample forecasts"
[1] 49.09331
[1] "Mean Absolute Percentage error"
[1] 0.00658535
[1] "Symmetric Mean Absolute Percentage error"
[1] 0.00658809
```

R CODE

```
library(tseries)
library(fBasics)
library(zoo)
library(lmtest)
library(forecast)

## 1.Exploratory Data Analysis
setwd("~/Desktop/CSC425/")
#quarter gdp
myd = read.table("GDPC1.csv", header = T, sep = ',')
head(myd)
tail(myd)
x = myd[,2]
head(x)
tail(x)
#((x/x-1)^4)-1)*100
par(mfcol = c(1,1))

## CREATE TIME SERIES OBJECT
gdpts = ts(x, start = c(1947,1), freq = 4)
gdpts
plot(gdpts, type = 'l', xlab = 'time', ylab = 'Quarter GDP', main = 'TIME SERIES FROM (1947,2) to (2017,3)')
### It has a non-statinarity and strict upward trend

## FIRST DIFFERENCE TO REMOVE THE TREND
df1 = diff(gdpts)
plot(df1, main = "FIRST DIFFERNECE", ylab = 'Quarter GDP')
### the time plot gets improved but does not have a consistent variance over time

## SECONDE DIFFERENCE
df2 = diff(diff(gdpts))
plot(df2, main="SECOND DIFFERENCE", ylab ='Quarter GDP')
### Now we have a good time series over time as desired.
Box.test(df2, type="Ljung")

basicStats(df2)
#### The skewness here is -0.072991. This value implies that the distribution of the data is slightly skewed to the left, but is vlose to zero.
#### The kurtosis is 1.155046. It has a less kurtosis than normal distribution (3).
```

```r
#### Therefore, our second differenced time plot is not a perfect normal, but we could say it is close to the normal though.

par(mfcol = c(1,2))
hist(gdpts, xlab = 'Reap GDP', main = "Histogram GDPTS", prob = T)
xfit<-seq(min(gdpts), max(gdpts), length = 40)
yfit<-dnorm(xfit, mean = mean(gdpts), sd = sd(gdpts))
lines(xfit, yfit, col = "blue", lwd = 2)

hist(df1, xlab = 'Reap GDP', main = "Histogram FIRST DIFFERENCE", prob = T)
xfit<-seq(min(df1), max(df1), length = 40)
yfit<-dnorm(xfit, mean = mean(df1), sd = sd(df1))
lines(xfit, yfit, col = "blue", lwd = 2)

## HISTOGRAM FOR SECOND DIFFERNCE
par(mfcol = c(1,2))
hist(df2, xlab = 'Reap GDP', main = "Second differenced Histogram", prob = T)
xfit<-seq(min(df2), max(df2), length = 40)
yfit<-dnorm(xfit, mean = mean(df2), sd = sd(df2))
lines(xfit, yfit, col = "blue", lwd = 2)
#### The histogram shows that the distribution is normal.

## QQ-PLOT FOR SECOND DIFFERNCE
qqnorm(df2)
qqline(df2, col = "blue")
#### It looks like it is close to the normal distribution

normalTest(df2, method = c('jb'))
# 2. MODELING APPROACH
## ACF AND PACF FOR RAW DATSET
acf(coredata(gdpts), plot = T, lag = 26, main = 'ACF GDPTS')
pacf(coredata(gdpts), lag = 26, main = 'PACF GDPTS')
# acf and pacf suggests AR(1)
par(mfcol = c(2,3))

acf(coredata(df1), plot = T, lag = 26, main = 'ACF FIRST DIFFERENCE')
pacf(coredata(df1), plot = T, lag = 26, main = 'PACF FIRST DIFFERENCE')

acf(coredata(df2), plot = T, lag = 26, main = 'ACF SECOND DIFFERENCE')
pacf(coredata(df2), plot = T, lag = 26, main = 'PACF SECOND DIFFERENCE')

## FIRST MODEL
m = Arima(gdpts, c(1,2,0))
m
```

```
##
par(mfcol = c(1,1))
acf(coredata(m$residuals))
coeftest(m)
#### few spikes detected
Box.test(m$resid, lag = 3, type = "Ljung", fitdf = 1)
Box.test(m$resid, lag = 6, type = "Ljung", fitdf = 1)
Box.test(m$resid, lag = 9, type = "Ljung", fitdf = 1)
#### failed the test
coeftest(m)

## SECOND MODEL AUTO ARIMA
m1 = auto.arima(gdpts, ic = "bic")
m1
# suggests ARIMA(1,2,1) = m1
# White Noise
coeftest(m1)
acf(coredata(m1$resid))

# stationary
Box.test(m1$resid, lag = 3, type = "Ljung", fitdf = 2)
Box.test(m1$resid, lag = 6, type = "Ljung", fitdf = 2)
#Box.test(m1$resid, lag = , type = "Ljung", fitdf = 2)
#Box.test(m1$resid, lag = 12, type = "Ljung", fitdf = 2)
Box.test(m1$resid, lag = 16, type = "Ljung", fitdf = 2)
Box.test(m1$resid, lag = 26, type = "Ljung", fitdf = 2)

#passed the test
hist(m1$resid, xlab = "Residuals", prob = T, main = "Histogram ARIMA(1,2,1)")
xfit <- seq(min(m1$resid, na.rm = T), max(m1$resid, na.rm = T), length = 40)
yfit <- dnorm(xfit, mean = mean(m1$resid, na.rm = T), sd = sd(m1$resid, na.rm = T))
lines(xfit, yfit, col = "blue", lwd = 2)

qqnorm(m1$resid, main = 'QQ-Plot ARIMA(1,2,1)')
qqline(m1$resid)



# THIRD MODEL windowed data
ts_gdp_window = tail(gdpts,43)
head(ts_gdp_window)
plot(ts_gdp_window)
library(fBasics)
basicStats(ts_gdp_window)
```

```
plot(diff(diff(ts_gdp_window)))

acf(coredata(ts_gdp_window), plot = 'T', lag = 26)
pacf(coredata(ts_gdp_window), plot = 'T', lag = 26)

m2 = auto.arima(ts_gdp_window)
m2
coeftest(m2)

par(mfcol = c(1,1))
acf(coredata(m2$resid))
Box.test(m2$resid, lag=3, type='Ljung', fitdf = 2)
Box.test(m2$resid, lag=6, type='Ljung', fitdf = 2)
Box.test(m2$resid, lag=9, type='Ljung', fitdf = 2)
Box.test(m2$resid, lag=16, type='Ljung', fitdf = 2)
Box.test(m2$resid, lag=26, type='Ljung', fitdf = 2)

par(mfcol = c(1,2))
hist(m2$resid, xlab = "Residuals", prob = T, main = "Histogram ARIMA(1,2,1)")
xfit <- seq(min(m2$resid, na.rm = T), max(m2$resid, na.rm = T), length = 40)
yfit <- dnorm(xfit, mean = mean(m2$resid, na.rm = T), sd = sd(m2$resid, na.rm = T))
lines(xfit, yfit, col = "blue", lwd = 2)

qqnorm(m2$resid, main = 'QQ-Plot ARIMA(1,2,1)')
qqline(m2$resid)

## FORECASTING SECOND MODEL
f1 = forecast(m1, h=12)
f1
f1_mean = ts(c(f1$fitted, f$mean), start = c(1947,1), freq = 4)
plot(f1, include = 50)
lines(f1_mean,col = 'blue')

source("backtest.R")
#### apply the backtesting procedure using 85% (=239 values) of the data for training and 15%
for testing to evaluate the forecasting power of both models.
backtest(m1, gdpts, 50, 1)


## FORECASTING THIRD MODEL
par(mfcol = c(1,1))

f2 = forecast(m2, h=9)
f2
```

```
f2_mean = ts(c(f2$fitted, f2$mean), start = c(2007,1), freq = 4)
plot(f2, include = 50)
lines(f2_mean,col = 'blue')
source("backtest.R")
#### apply the backtesting procedure using 85% (=239 values) of the data for training and 15% for testing to evaluate the forecasting power of both models.
backtest(m2, gdpts, 36,1)
basicStats(ts_gdp_window)
```

```
#Before the crash
ts_gdp_crash = head(gdpts,244)
tail(ts_gdp_crash)
plot(ts_gdp_crash)
library(fBasics)
basicStats(ts_gdp_crash)
plot(diff(ts_gdp_crash))

acf(coredata(ts_gdp_crash), plot = 'T', lag = 26)
pacf(coredata(ts_gdp_crash), plot = 'T', lag = 26)

m_past = auto.arima(ts_gdp_crash)
m_past

coeftest(m_past)

par(mfcol = c(1,1))
acf(coredata(m_past$resid))
Box.test(m_past$resid, lag=3, type='Ljung', fitdf = 2)
Box.test(m_past$resid, lag=6, type='Ljung', fitdf = 2)
Box.test(m_past$resid, lag=9, type='Ljung', fitdf = 2)
Box.test(m_past$resid, lag=12, type='Ljung', fitdf = 2)

f = forecast(m_past, h=12)
f
f_mean = ts(c(f$fitted, f$mean), start = c(1947,1), freq = 4)
plot(f, include = 40, ylab="Real GDP (Billions)", xlab="Year (Quarterly)")
lines(f_mean,col = 'blue')
lines(tail(head(gdpts,256),52))
```

```
source("backtest.R")
#### apply the backtesting procedure using 85% (=239 values) of the data for training and 15% for testing to evaluate the forecasting power of both models.
backtest(m1, gdpts, 239,1)
```