PSTAT174, FINAL PROJECT

Zhenhui Jiang, perm:8155962, section: Friday 8:00—8:50

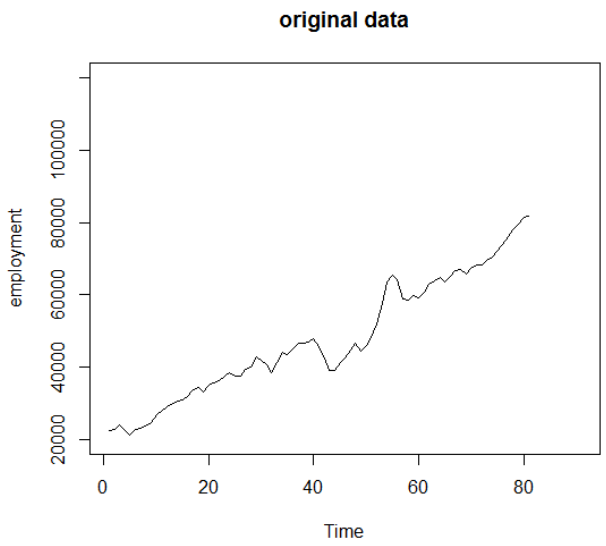
1. Abstract.

The project studies the annual employment of U.S. from 1890 to 1970. For individuals, a job means the live quality of their family and for a country, the employment means the performance of a government. The analysis will answer the following questions:

Did the number of employment in U.S. increase from 1890 to 1970? Were there any sharp changes over that time-period? What's the number of employment in the next 10 years?

In the analysis, I used box-cox method to transform and then difference with lag 1 to remove the linear trend. Furthermore, I use the AIC method to choose models and diagnose check to decide one model. Finally, I find the employment number would increase in the next 10 years.

2. Introduction



I choose this data set since employment relates the strength of a country and the live fundamentals of people. If the number decreased, it reflected the disability of the government and it had to change. If the number increased with stationary rate, it means that the government worked well and need to keep the performance.

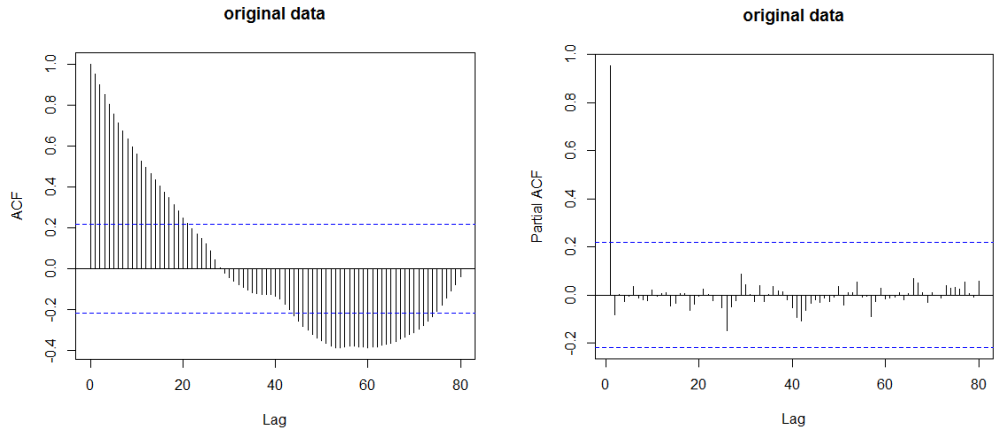
Did the number of employment in U.S. increase from 1890 to 1970? Were there any sharp changes over that time-period? By looking at the plot, We can see there is an obvious increasing trend.

Starting from the time 39 in the plot, which is 1929, the time of the Great Depression. It was a severe worldwide economic depression that took place during the 1930s. The depression originated in the United States, after a major fall in stock prices that began around September 4, 1929, and became worldwide news with the stock market crash of October 29, 1929 (known as Black Tuesday). Not only the U.S, between 1929 and 1932, worldwide GDP fell by an estimated 15%.

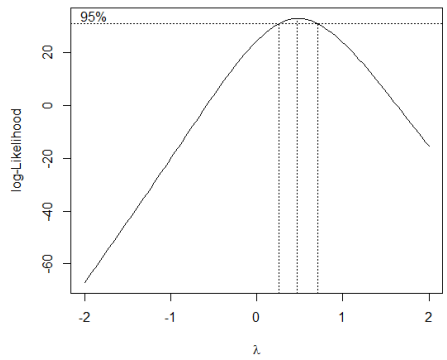
After the Great Depression period, the number of U.S. employment increased in a fast rate until the period around 1948. After the short-term financial crisis, the number of employment increased smoothly.

3. Analysis of data

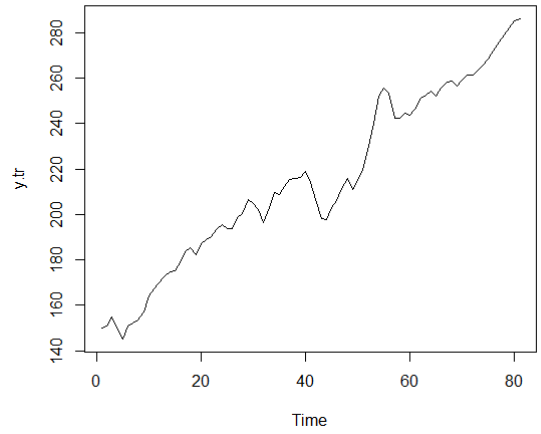
First, I plot the ACF and PACF of the data.



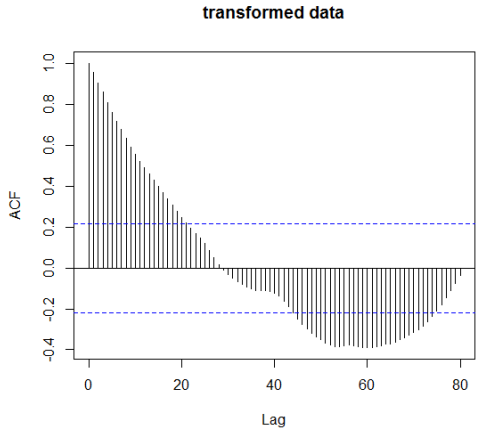
By the graph of ACF, we know that it needs to be transformed and differenced. By using the box-cox transformation, we can see that the best lamda is 0.4646465, and 1/2 is in the confident intervel and is very near to the best lambda.

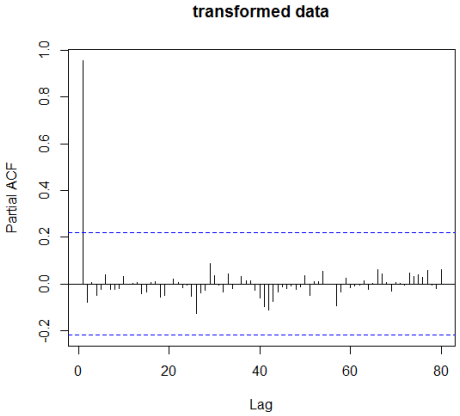


Then I tried lambda equals 1/2, the plot looks like

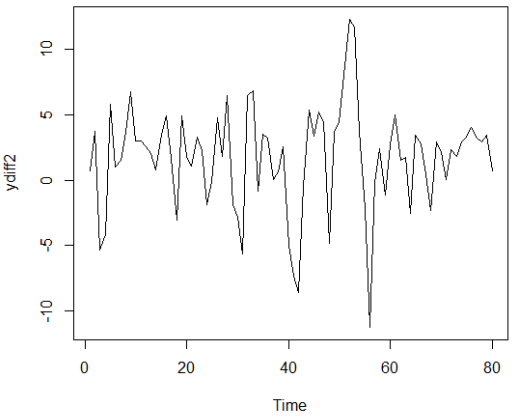


The ACF and PACF are:

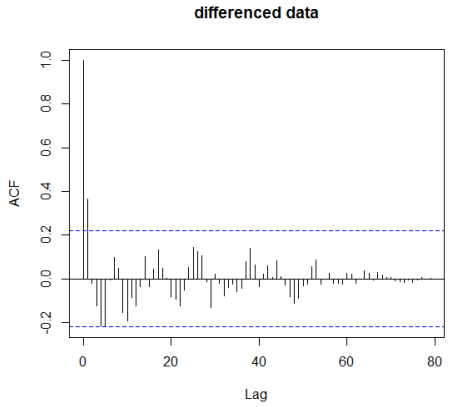


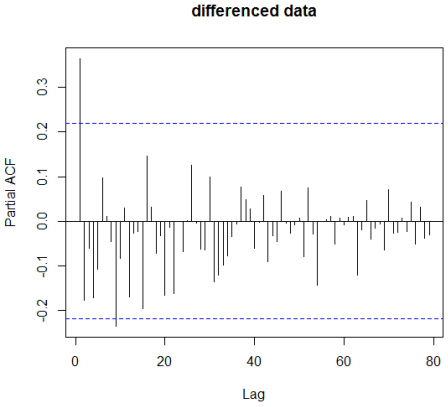


By the ACF and the plot, we can see there still exist trend.
Then I difference the transformed data with lag 1, the plot is:

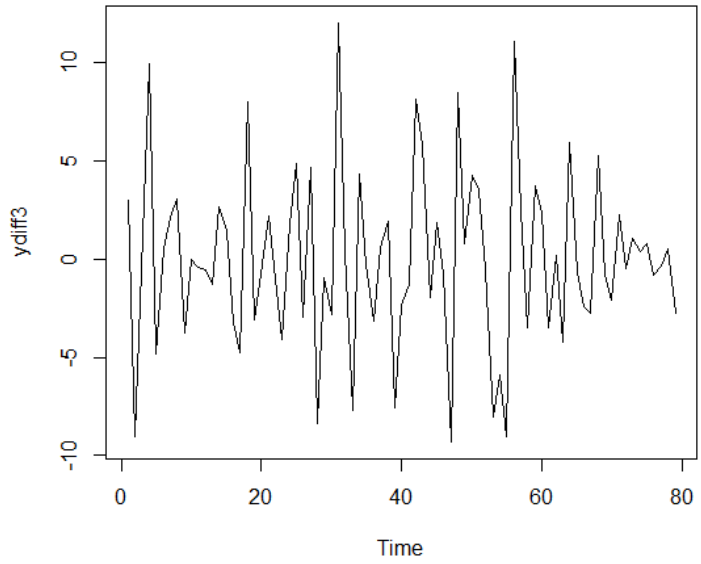


The variance is 15.80926.
The ACF and PACF are:

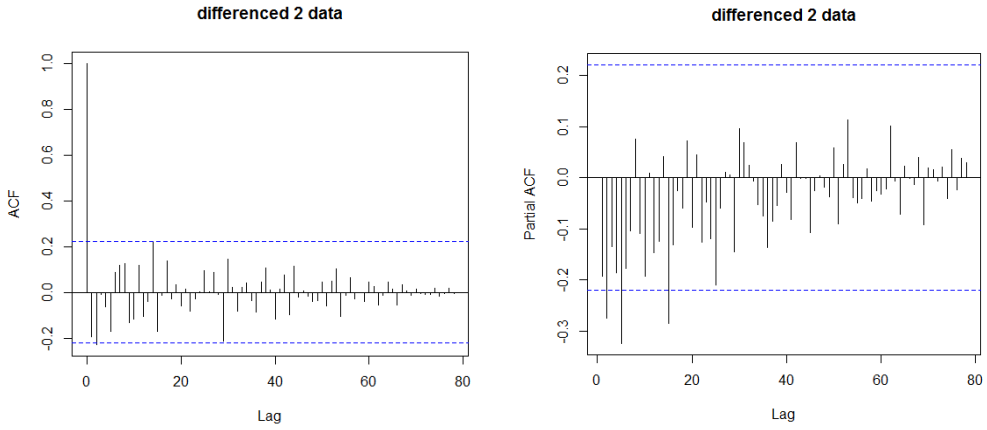




By the plots, we can see that it's good. Then I tried to difference with lag 2. The plot is:



The ACF and PACF are:



The variance is 20.32411, which is larger than the first one, so it's over-differencing.

Then I went to model choice. By the ACF of differenced data, I prefer to choose MA(1) since it cuts off after lag 1 and by the PACF and differenced data, I prefer AR(9), since it cuts off after lag 9. Then I did the estimation of ARIMA(9,1,1), it gives me the result:

```
Call:
arima(x = y.tr, order = c(9, 1, 1), xreg = 1:length(y.tr),
method = "ML")

Coefficients:
          ar1      ar2     ar3      ar4      ar5     ar6      ar7
      ar8      ar9      ma1  1:length(y.tr)
```

```
     1.2648  -0.5148  0.1609  -0.1271  -0.0866  0.2420  -0.
0839  -0.0204  -0.0786  -1.0000         1.5997
s.e.  0.1113   0.1811  0.1911   0.1888   0.1924  0.1945   0.
1957   0.1857   0.1142   0.0418         0.0673

sigma^2 estimated as 10.66:  log likelihood = -210.14,  aic
 = 444.27
```

By the estimation, the model failed because the absolute value of the coef ficient of MA(1) is equal to 1, which is non-invertible.
Then I try to compare AICc to get the best model.

```
> aiccs
   q
p         0        1        2        3        4        5
  0 450.9222 439.7086 441.7951 443.7176 445.7035 438.2725
  1 441.7380 441.8096 443.5030 439.0964 441.0016 438.7831
  2 441.3345 436.5595 438.5420 441.3874 442.8270 441.0333
  3 443.3014 438.6244 440.7335 443.2054 443.9309 443.4787
  4 443.1033 444.9702 443.5341 439.7522 445.4241 444.8575
  5 444.4101 446.0344 448.4699 448.1623 446.7510 445.3028
```

There are too many values to compare, so I let the lowest AIC be TRUE.

```
> (aiccs==min(aiccs))
   q
p      0     1     2     3     4     5
  0 FALSE FALSE FALSE FALSE FALSE FALSE
  1 FALSE FALSE FALSE FALSE FALSE FALSE
  2 FALSE  TRUE FALSE FALSE FALSE FALSE
  3 FALSE FALSE FALSE FALSE FALSE FALSE
  4 FALSE FALSE FALSE FALSE FALSE FALSE
  5 FALSE FALSE FALSE FALSE FALSE FALSE
```

The model ARIMA(2,1,1) has the lowest AIC value.
Then I went to model estimate:

```
Call:
arima(x = y.tr, order = c(2, 1, 1), xreg = 1:length(y.tr),
method = "ML")

Coefficients:
         ar1      ar2     ma1  1:length(y.tr)
      1.2689  -0.4435  -1.000          1.6212
s.e.  0.0983   0.0986   0.036          0.0859
```

```
sigma^2 estimated as 11.61:  log likelihood = -213.02,  aic
 = 436.03
```

It's not good since the absolute value of coefficient of MA(1) part is not less than 1. Also, the roots for AR parts are: -0.6433983 and 3.5045032, one absolute value of the roots is smaller than 1, which means it's non-stationary too. Then, I tried ARIMA(5,1,1) and ARIMA(5,1,0), which both failed at diagnosis check.

Then I tend to try ARIMA(0,1,1) model.

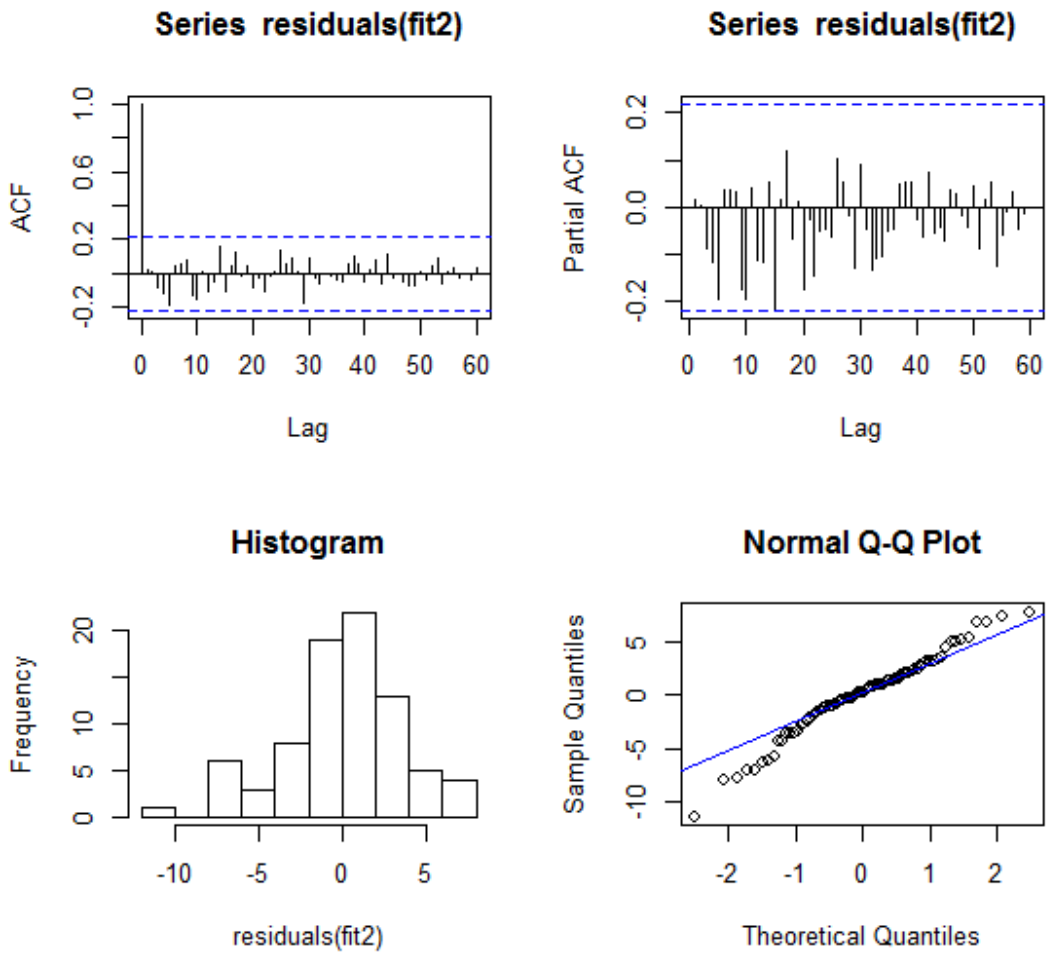
```
Call:
arima(x = y.tr, order = c(0, 1, 1), xreg = 1:length(y.tr),
method = "ML")

Coefficients:
         ma1  1:length(y.tr)
      0.4133         1.6766
s.e.  0.0967         0.5717

sigma^2 estimated as 13.19:  log likelihood = -216.78,  aic
 = 439.55
```

All of them looks good. Then I tried to do diagnosis check:

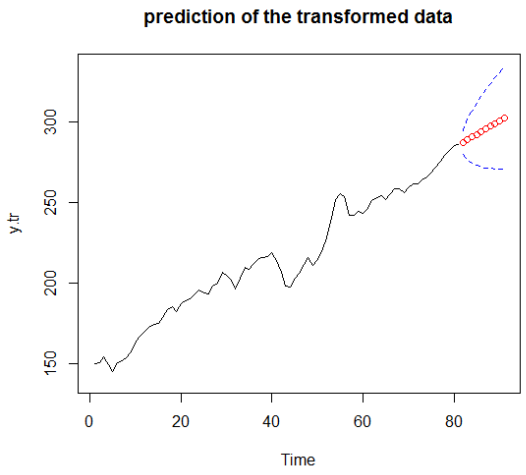
```
> Box.test(residuals(fit2),type="Ljung-Box")

        Box-Ljung test

data:  residuals(fit2)
X-squared = 0.019953, df = 1, p-value = 0.8877

> Box.test(residuals(fit2),type="Box-Pierce")

        Box-Pierce test

data:  residuals(fit2)
X-squared = 0.019231, df = 1, p-value = 0.8897

> Box.test((residuals(fit2))^2,type="Ljung-Box")

        Box-Ljung test

data:  (residuals(fit2))^2
X-squared = 2.5197, df = 1, p-value = 0.1124

> shapiro.test(residuals(fit2))

        Shapiro-Wilk normality test
```

```
data:  residuals(fit2)
W = 0.97296, p-value = 0.08408
```

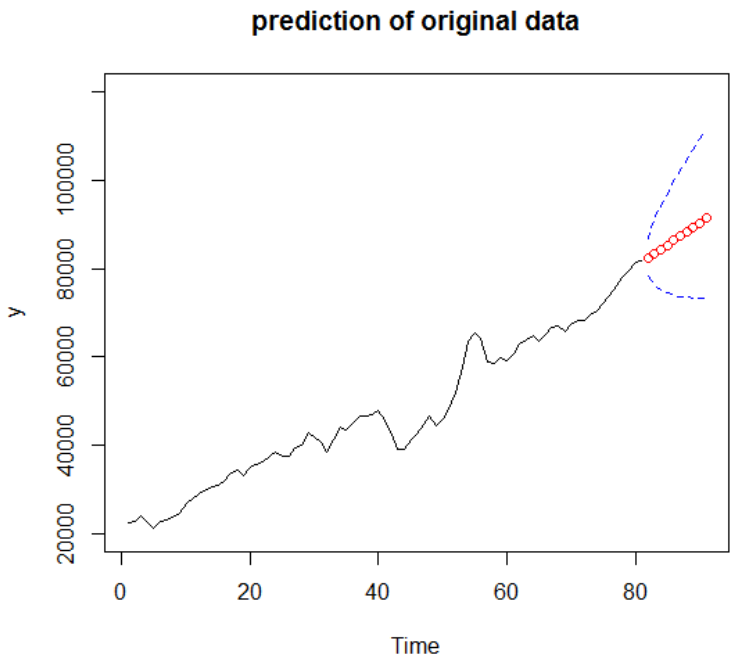
It passed all the tests since the p-values are all larger than 0.05.



The normality of residuals looks good. So I tend to choose ARIMA(0,1,1).
The prediction of the transformed data in the next 10 years is:



To get the original data, I squared it and get the plot:



The Final model is:

$\nabla X_t=1.6766+Z_t+0.4133Z_{t-1}$ where $X_t=y.tr=Y_t^{0.5,}$

It's equivalent to $\nabla Y_t^{0.5}=1.6766+Z_t+0.4133Z_{t-1}$

4. Conclusion

After analysis of the data, I could conclude that the U.S. government did well about the employment issue except two periods of economical crisis. By the prediction plot, we can predict the number of employment of US would still increase.

5. Reference

http://www.history.com/topics/great-depression

https://datamarket.com/data/set/22v2/annual-employment-us-1890-to-1970#!ds=22v2&display=line

6. Appendix

```
y<-scan("C:/Users/11514/Downloads/employ.txt")
head(y)
ts.plot(y,xlim=c(1,length(y)+10),ylim = c(20000,120000),main="original data",ylab="employment")
op <- par(mfrow=c(1,1))
acf(y,lag=80,main="original data")
pacf(y,lag=80,main="original data")
require(MASS)
bcTransform <- boxcox(y ~ as.numeric(1:length(y)))
bcTransform$x[which(bcTransform$y == max(bcTransform$y))]
y.tr<-y^0.5
ts.plot(y.tr)
var(y.tr)
acf(y.tr,lag=80,main="transformed data")
pacf(y.tr,lag=80,main="transformed data")
ydiff2<-diff(y.tr,differences = 1)
ts.plot(ydiff2)
var(ydiff2)
acf(ydiff2,lag=80,main="differenced data")
pacf(ydiff2,lag=80,main="differenced data")
ydiff3<-diff(y.tr,differences = 2)
ts.plot(ydiff3)
var(ydiff3)
acf(ydiff3,lag=80,main="differenced 2 data")
pacf(ydiff3,lag=80,main="differenced 2 data")
library(qpcR)
# Calculate AICc for ARMA models with p and q running from 0 to 5
aiccs <- matrix(NA, nr = 6, nc = 6)
dimnames(aiccs) = list(p=0:5, q=0:5)
for(p in 0:5)
{
  for(q in 0:5)
  {
    aiccs[p+1,q+1] = AICc(arima(y.tr, order = c(p,1,q), method="ML",xreg=1 : length(y.tr)))
  }
}
aiccs
(aiccs==min(aiccs))
fit3=arima(y.tr,order=c(9,1,1),method='ML',xreg=1 : length(y.tr))
fit3
fit1=arima(y.tr,order=c(2,1,1),method='ML',xreg=1 : length(y.tr))
fit1
```

```
fit4=arima(y.tr,order=c(5,1,0),method='ML',xreg=1 : length(y.tr))
fit4
fit5=arima(y.tr,order=c(5,1,1),method='ML',xreg=1 : length(y.tr))
fit5
fit2=arima(y.tr,order=c(0,1,1),method='ML',xreg=1 : length(y.tr))
fit2
Box.test(residuals(fit2),type="Ljung-Box")
Box.test(residuals(fit2),type="Box-Pierce")
Box.test((residuals(fit2))^2,type="Ljung-Box")
shapiro.test(residuals(fit2))
ts.plot(residuals(fit2),main='fitted value')
op <- par(mfrow=c(2,2))
acf(residuals(fit2),lag=60)
pacf(residuals(fit2),lag=60)
# Histogram
hist(residuals(fit2),main = "Histogram")
# q-q plot
qqnorm(residuals(fit2))
qqline(residuals(fit2),col ="blue")
op <- par(mfrow=c(1,1))
pred.tr <- predict(fit2, n.ahead = 10, newxreg=(length(y.tr)+1) : (length(y.tr)+10))
U.tr= pred.tr$pred + 2*pred.tr$se # upper bound for the C.I. for transformed data
L.tr= pred.tr$pred - 2*pred.tr$se # lower bound
ts.plot(y.tr, xlim=c(1,length(y.tr)+10), ylim = c(140,max(U.tr)),main="prediction of the transformed data") #plot y.tr and forecast
lines(U.tr, col="blue", lty="dashed")
lines(L.tr, col="blue", lty="dashed")
points((length(y.tr)+1):(length(y.tr)+10), pred.tr$pred, col="red")
pred.orig <- pred.tr$pred^2 # back-transform to get predictions of original time series
 U= U.tr^2 # bounds of the confidence intervals
 L=L.tr^2
# Plot forecasts with original data
 ts.plot(y, xlim=c(1,length(y)+10), ylim = c(20000,120000),main="prediction of original data")
 lines(U, col="blue", lty="dashed")
lines(L, col="blue", lty="dashed")
 points((length(y)+1):(length(y)+10), pred.orig, col="red")
```