477/577 In-class Exercise 3 : Fitting ARMA(p,q)

(due Fri 2/24/2017)

Name:

Use this file as a template for your report. Submit your code and comments together with (selected) output from R console.

- Your comments must be Arial font, and **BOLD FACED**.
- Your code must be Lucida Console font.

You must submit PRINTOUT of this file.

First, copy and paste below command in R console.

```
D  <- read.csv("http://gozips.uakron.edu/~nmimoto/pages/datasets/Steel.csv")
D1 <- ts(D[,2], start=c(1956,1), freq=12)
D2 <- window(D1, start=c(1969, 12))

Mav1 <- aggregate(c(D2), list(month=cycle(D2)), mean)$x
M.av <- ts(Mav1[cycle(D2)], start=start(D2), freq=frequency(D2))
D3   <- D2-M.av
```

Now your "D1" in R contains monthly production of raw steel in Australia (in 1000 tones) from January 1956 to Nov 1993. "D2" is the same data cut from Dec. 1969 to Nov 1993. "Mav1" is the monthly average calculated from "D2". "M.av" is the same as Mav1, except that it is as long as D2. "D3" is the series when the monthly average is subtracted from D2.

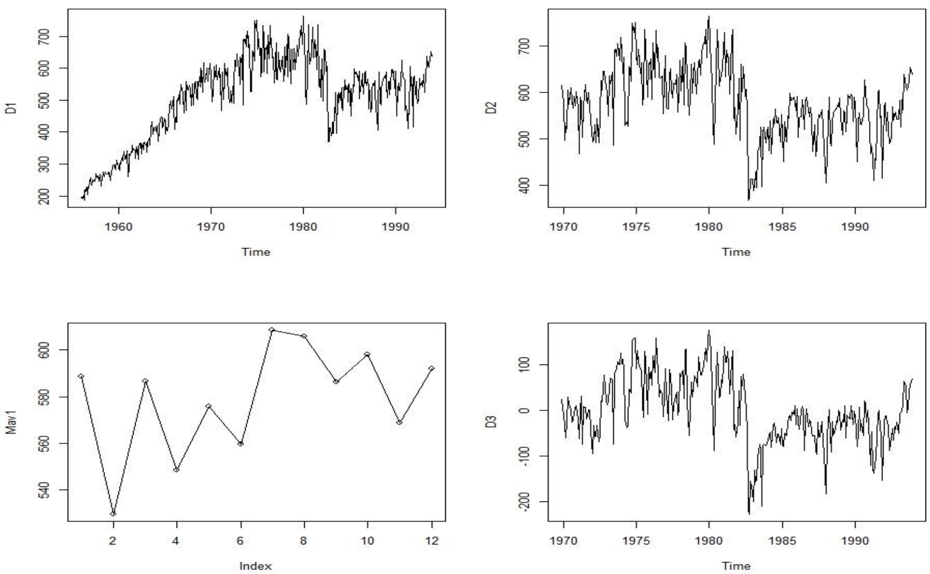
1. Plot the series D1, D2, Mav1, and D3. Comment on the stationary of D1, D2, and D3. Briefly explain what you see in the plot of "D2" regarding the stationarity.

   ```
   layout(matrix(1:4, 2,2, byrow=TRUE))
   plot(D1)
   plot(D2)
   plot(Mav1, type="o")
   plot(D3)
   ```

   D1 does not look stationary, as it has clear upward trend early on.

   D2 and D3 may be stationary, but has big dip around 1993. If you break up D2 before and after 1993, each of them does look like stationary series.

   After the monthly average is subtracted, D3 does not show significant difference in overall appearance from D2. This suggests that subtracting M.av was not effective way to account for monthly differences.



2. Note that D1 was cut after Dec 1969 to produce D2, which was used to produce D3. Does it make sense to do this? If this procedure is to be justified, what is the assumption made?

   It does make sense in this case, because if we speculate that prior to 1969, Australian steel industry was developing with boosting economy, and the industry came to maturity around 1969. By separating, we can model development phase and stable phase differently.

   However, the timing of cut, 1969, is of question, and may influence the result of the analysis. For example, under the same argument we can cut D1 at 1973, instead of 1969. This may change our final model and forecast.

   Therefore, cutting does make sense, but some care/investigation is needed to formalize the process.

3. Note that Monthly average was calculated from D2, in order to create D3. Does it make sense to do this? If this procedure is to be justified, what is the assumption made? Is there a way to verify that assumption?

   It would make sense if the data shows some kind of seasonality that changes from month to month. Looking at M.av, no clear pattern can be seen, except that July and Aug is slightly higher than other months.

   To justify such transformation, data must exhibit similar behavior for each month. For example, if we plot January only for 1970-1995, it must show

stationarity over constant mean. The monthly mean can be different for each month, but for each month, behavior in 70's should be similar to that in 90's.

The fact that D3 looks so similar to D2 tells us that's probably not the case. Looking at any month, data in 70's will be above the mean, and data in 90's will be below the mean. So for this data, splitting at the dip in 1982 seems more sensible than taking out monthly average.

4. Model "D3" with ARMA(p,q) model using auto.arima() function in forecast package. Be sure to use d=0 option inside. What is the best fitting model selected by using minimum AICc? Are all parameter estimates significantly different from zero?

```
library(forecast)
Fit1 <- auto.arima(D3, d=0)
Fit1

ARIMA(1,0,2) with zero mean

Coefficients:
         ar1      ma1      ma2
      0.9444  -0.2876  -0.2613
s.e.  0.0277   0.0681   0.0709

sigma^2 estimated as 2057:  log likelihood=-1506.31
AIC=3020.61   AICc=3020.75   BIC=3035.26
```

ARMA(1,2) without intercept is chosen based on AICc. All parameter looks significantly different from 0.

5. Force to fit ARMA(p, q) model to D3, adding 1 to each p and q you found in #4. E.g if you found ARMA(1,1) to be the best model, for to fit ARMA(2,2) to D3. Use function Arima(D3, order=c(p, 0, q)) . Comment on the parameter significance.

```
Fit2 <- Arima(D3, order=c(2,0,3) )
Fit2

ARIMA(2,0,3) with non-zero mean
```

```
Coefficients:
         ar1      ar2      ma1      ma2     ma3  intercept
      1.6269  -0.6402  -0.9667  -0.0578  0.1446     1.9445
s.e.  0.3043   0.2901   0.3095   0.1260  0.1063    22.4030

sigma^2 estimated as 2074:  log likelihood=-1505.93
AIC=3025.85   AICc=3026.25   BIC=3051.49
```

Estimates of Theta2, Theta3 and intercept are not significantly different from 0. For example for Theta 2, CI is -0.0578 +- 2*(.1260) and includes 0.

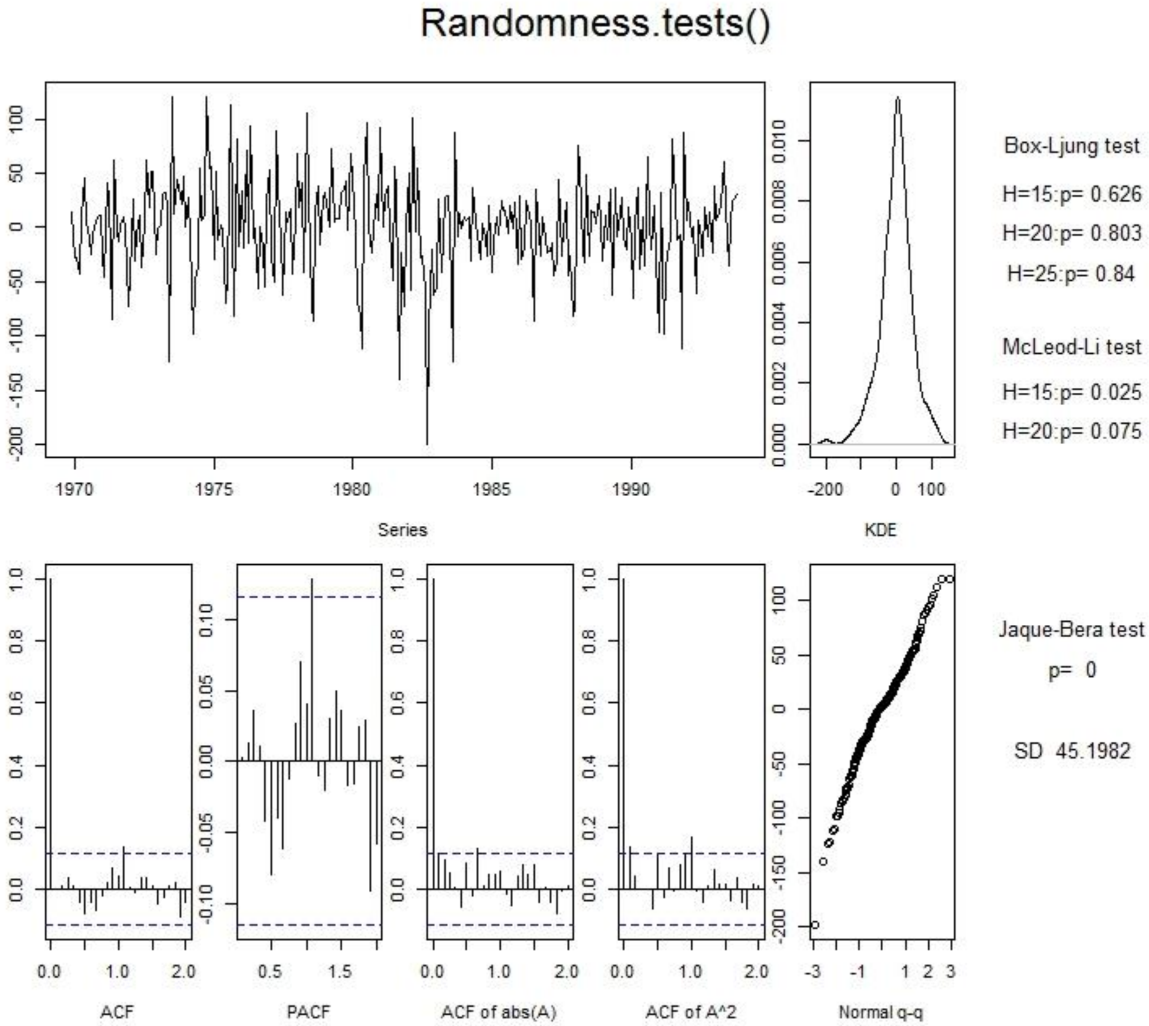
6. Check adequacy of the model you obtained in #4. (i.e. perform residual analysis)

```
source("http://gozips.uakron.edu/~nmimoto/477/TS_R-
90.txt")
Randomness.tests(Fit1$resid)
```

Box-Ljung test shows high p-value, indicating uncorrelated nature of the residuals.

However, one of McLeod-Li test is showing one p-value below .05, indicating some correlation is spotted in squares of the residuals. This is probably the result of big dip still present around 1983, and period of small variance that follows. It is a concern, but not the huge one. Based on the residual plot alone, this is borderline-may-be-ok residual. We should look for better model.

Jaque-Bera test shows very low p-value, indicating non-normality of the residuals. At this point, normality of residuals is not a main concern.

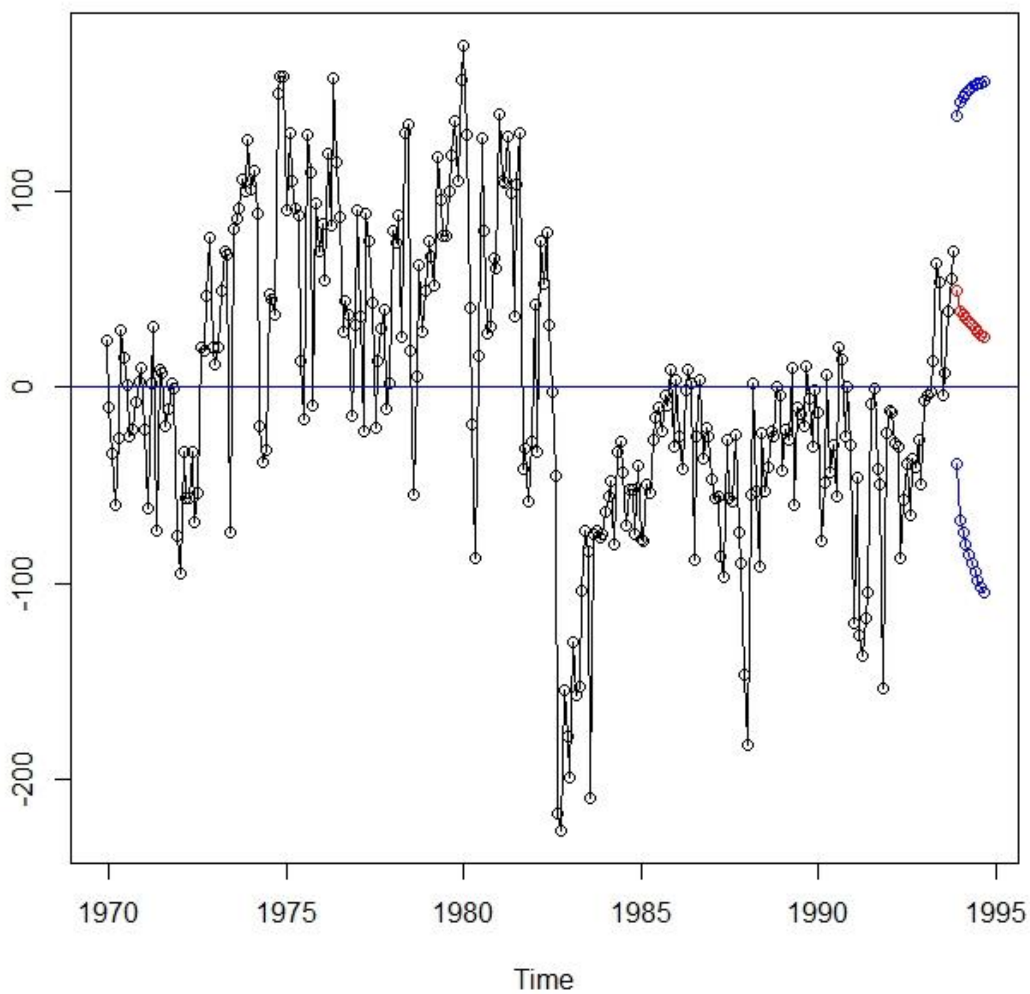


7. Using model obtained in #4, produce 5-step ahead prediction for D3. Then use that to predict December 1993 steel production with 95% prediction interval. (c.f. p60 Chapter 3 slides)

```
Y <- D3

Fit1 <- Arima(Y, order=c(1,0,2) ) #- Force to fit ARMA(1,2)

Y.h <- predict(Fit1, n.ahead=10)
Yhat <- Y.h$pred
Yhat.CIu <- Yhat+1.96*Y.h$se
Yhat.CIl <- Yhat-1.96*Y.h$se
ts.plot(cbind(Y, Yhat, Yhat.CIu, Yhat.CIl ), type="o",
col=c("black","red","blue","blue"))
abline(h=0)
abline(h=mean(Y), col="blue")

Yhat.CIu
Yhat.CIl
```

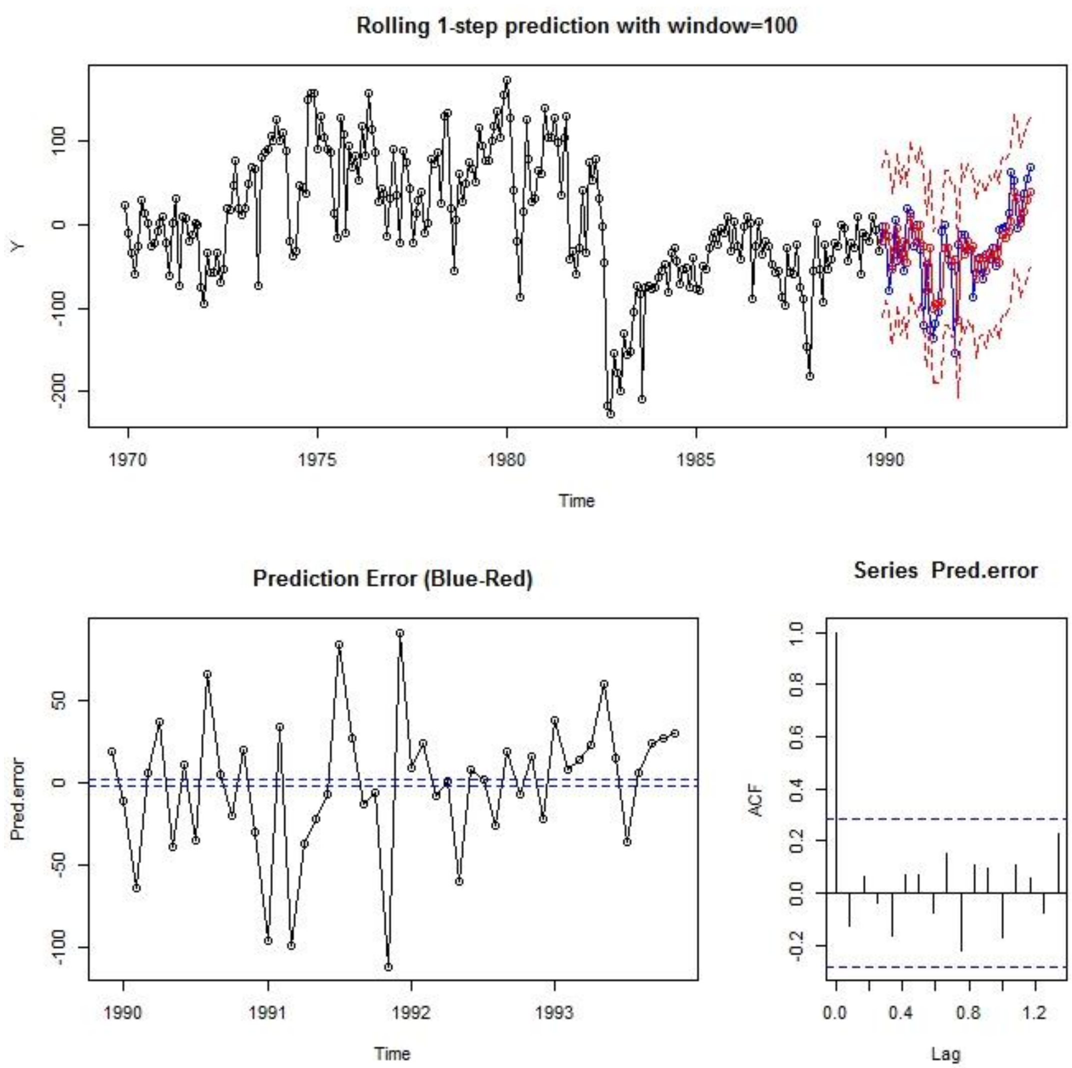
M.av

95% CI for 1-step ahead prediction of D3 : (-39.52, 138.58)
Dec Monthly average: 591.84
95% Prediction Interval for Steel Production in Dec 1993: (552.32 730.42)



8. Perform Rolling 1-step prediction of the last 48 month within the dataset, with window size of 240. Report root mean squared error of the prediction performance. (c.f. p61 and 62 Chapter 3 slides)

   Despite our ignorance over the two-piece nature of D3 data, rolling 1-step prediction of last 48 month seems not too bad. Root Mean Squared Error of the prediction was 41.2, and prediction error seems to be uncorrelated.



```
#--- Rolling 1-step predicton of last 48 month
  Rolling.len = 48
  Window.size = 240

  p = 1
  d = 0
  q = 2

  Y <- D3

  Yhat <- Yhat.CIu <- Yhat.CIl <- Y2<- 0 #- Initialize

  for (i in 1:Rolling.len) {
      window.bgn <- i
      window.end <- i+Window.size-1
      Fit1 <- Arima(Y[window.bgn:window.end], order=c(p,d,q) ) #-
Force to fit AR(p)
      Y.h <- predict(Fit1, n.ahead=1)
      Yhat[i] <- Y.h$pred
      Yhat.CIu[i] <- Yhat[i]+1.96*Y.h$se
      Yhat.CIl[i] <- Yhat[i]-1.96*Y.h$se
  }

  X  <- window(Y, start=time(Y)[1], end=time(Y)[Window.size])
  Y2 <- window(Y, start=time(Y)[Window.size+1],
end=time(Y)[Window.size+Rolling.len])
  Yhat     <- ts(Yhat,     start=c(floor(time(Y)[Window.size+1]),
cycle(Y)[Window.size+1]), freq=frequency(Y) )
  Yhat.CIu <- ts(Yhat.CIu, start=c(floor(time(Y)[Window.size+1]),
cycle(Y)[Window.size+1]), freq=frequency(Y) )
  Yhat.CIl <- ts(Yhat.CIl, start=c(floor(time(Y)[Window.size+1]),
cycle(Y)[Window.size+1]), freq=frequency(Y) )

  Pred.error <- Y2-Yhat
  Pred.rMSE = sqrt( mean( (Pred.error)^2 ) ) #- prediction root Mean
Squared Error
  Pred.rMSE

  mean(Pred.error)
  layout(matrix(c(1,1,1,2,2,3), 2, 3, byrow=TRUE))

  plot(Y, type="o", col="blue", main="Rolling 1-step prediction with
window=100" ) #- Entire dataset
  lines(X, type="o")
  lines(Yhat, type="o", col="red")
  lines(Yhat.CIu, type="l", col="red", lty=2)
  lines(Yhat.CIl, type="l", col="red", lty=2)
  plot(Pred.error, type="o", main="Prediction Error (Blue-Red)")
  abline(h=c(-1.96, 1.96), col="blue", lty=2)
  acf(Pred.error)
```