Data Analysis - lab 6

ISEP – December 17th 2018

Instructions: Prepare a report including the source code and the results. Deposit your report on Moodle and don't forget your binome's name or to make 2 deposits if you did not work alone.

In this lab, we are going to study the evolution US GNP between 1954 and 1987. In particular, we are going to study the log of this time series.

A Stationnarity analysis

1. Installez and load the **tseries** library, then load the **USeconomic** data using the command line *data(USeconomic)*.
2. We will now create the variables that we are going to study:
   - The *log(GNP)* data are in the 2nd column of the **USeconomic** dataset. Retrieve this column and store it in a variable of your choice.
   - The log(GNP) data were acquired on a trimestrial basis. Use the command *seq* to create a second variable *year* containing all years and semesters from 1954 to 1987.75 .
   - Display the log(GNP) evolution between 1954 and the 3rd trimester of 1987.
3. Remind what "stationnarity" means for a time series. What can you visually say about the stationnarity of the log(GNP) time series.
4. Use the *acf* command to draw this series autocorrelogram. Comment.

We remind that the Box-Pierce test is a satistical test used on time series used to assess whether or not the series is mostly white noise (i.e. $\forall t \quad \epsilon_t \sim i.i.d.(0, \sigma^2)$).

5. Use the *Box.test* function on your series and interpret the result.
6. What can you conclude on this time series stationnarity ?

B Study of diffGNP

1. Create a variable *DiffGNP* so that: $(Y_t = X_t - X_{t-1})_{2 \leq t \leq T}$ where X_t is the logGNP at time t. Explain what this time series represents.
2. Plot the evolution of this series between 1954 and the 3rd semester of 1987.

3. Is this series centered ? You can use the empirical mean value of the series and a Student test to justify your answer.

4. Use the *acf* and *pacf* functions to draw the autocorrelogram and partial autocorrelogram of this time series. From there, deduce the most likely parameter(s) p and q for an ARMA(p,q) model to modelise *DiffGNP*.

5. Test all the couples (p,q) that seemed relevant to you from the previous question:

   - The function *arima(DiffGNP,c(p,0,q))* will help you to evaluate each model.
   - Explain the different values returned by the function *arima*. In particular, you will have te retrieve the fitted parameters and explain the meaning of the two following paramaters: "log likelihood" and "aic" (Akaike's Information Criterion), and how you can you them to rate your model.
   - Which model seems to be the best one if you only use the AIC criterion ?

We are now going to study 3 models in particular: ARMA(0,1), ARMA(0,2) and ARMA(8,2). We remind you that the Shapiro-Wilk test assesses the null hypothesis that a sample follows a normal distribution.

6. Use the Box-Pierce test and the Shapiro-Wilk test (*shapiro.test*) on the residuals of all 3 models applied to the logGNP data and display their autocorrelogram. From there, what can you say on the stationnarity of the residuals ? How do you justify which model is the best ?

C Predictions using ARMA

We are still interested in the 3 models ARMA(0,1), ARMA(0,2) and ARMA(8,2). We will now assess the quality of all 3 models for prediction purposes. To do so, we are going to remove the 10 last values of the data, retrain the model on the first part of the data, and try to predict the last 10 values:

```
n <- 10
T=length(diffGNP)
index <- 1:(T - n - 1)
res01 <- predict(arima(diffGNP[index], c(0, 0, 1)), n)
res02 <- predict(arima(diffGNP[index], c(0, 0, 2)), n)
res82 <- predict(arima(diffGNP[index], c(8, 0, 2)), n)
```

1. For each model, draw on the same graph the 3 following elements (you may want to zoom on the end of the time series):

   - The original GNP series
   - The mean value predicted by each model.

- The 95% confidence interval on your prediction under the hypothesis that the residualsare normally distributed (regardless of question B.6 result). See the code bellow to answer this question (you may have to adapt it to your variables).

2. Using the previous question, which model seem to give the best results ?

3. Translate these models into ARIMA(p,d,q) for the original logGNP time series.

```
par(mfrow = c(1, 3))
plot(annee[(T - 4 * n):T], diffGNP[(T - 4 * n):T - 1], main = "prevision ARMA(0,2)", t = "l", col = "blue", xlab = "temps", ylab = "diff GNP")
lines(annee[(T - n):T], c(diffGNP[T - n - 1], res02$pred))
lines(annee[(T - n):T], c(diffGNP[T - n - 1], res02$pred) + c(0,res02$se) * 1.96, lty = 2)
lines(annee[(T - n):T], c(diffGNP[T - n - 1], res02$pred) - c(0,res02$se) * 1.96, lty = 2)
plot(annee[(T - 4 * n):T], diffGNP[(T - 4 * n):T - 1], main = "prevision ARMA(8,2)",t = "l", col = "blue", xlab = "temps", ylab = "diff GNP")
lines(annee[(T - n):T], c(diffGNP[T - n - 1], res82$pred), col = "red")
lines(annee[(T - n):T], c(diffGNP[T - n - 1], res82$pred) + c(0,res82$se) * 1.96, lty = 2, col = "red")
lines(annee[(T - n):T], c(diffGNP[T - n - 1], res82$pred) - c(0,res82$se) * 1.96, lty = 2, col = "red")
plot(annee[(T - 4 * n):T], diffGNP[(T - 4 * n):T - 1], main = "prevision ARMA(0,1)",t = "l", col = "blue", xlab = "temps", ylab = "diff GNP")
lines(annee[(T - n):T], c(diffGNP[T - n - 1], res01$pred), col = "green")
lines(annee[(T - n):T], c(diffGNP[T - n - 1], res01$pred) + c(0,res01$se) * 1.96, lty = 2, col = "green")
lines(annee[(T - n):T], c(diffGNP[T - n - 1], res01$pred) - c(0,res01$se) * 1.96, lty = 2, col = "green")
```

D ARIMA Model

Differentiate the GNP serie a second time and use the analysis of sections B) and C) to find the best possible ARIMA(p,2,q) possible for the GNP series.

Remark : You may directly use the fonction ARIMA(p,2,q) on the GNP series, it will be more convenient to draw the graphics and make analysis.