Data Analysis - lab 6

ISEP – December 3rd 2019

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.

Libraries

This lab may require the following libraries : pandas, numpy, matplotlib, sklearn, scipy and statsmodels.

A Stationnarity analysis

1. Load the **USeconomic** dataset from the csv file.
2. We will now create the variables that we are going to study:
   - The *log(GNP)* data are in the 3nd 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 *linspace* from numpy to create a second variable *year* containing all years and trimesters 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 example below to display the correlograms of the logGNP time series. Comment.

```
from statsmodels.graphics.tsaplots import plot_acf
from statsmodels.graphics.tsaplots import plot_pacf

plt.figure(figsize=(20,10))
plt.subplot(211)
plot_acf(logPNB, ax=plt.gca())
plt.subplot(212)
plot_pacf(logPNB, ax=plt.gca())
plt.show()
```

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 *acorr_ ljungbox* function from the *statsmodels.stats.diagnostic* library 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 (*stats.ttest_ ind* in *scipy*).

4. Use the *plot_ acf* and *plot_ 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. To do so, create a function that does the following:

   - Trains an ARMA model with the chosen (p,q) parameters on the first 70% of the data. (classifer ARMA in *statsmodels.tsa.arima_ model*)
   - Applies it on the last 30% of the data.
   - Computes the BIC and AIC of each model. (Remind what these criterions are in your report)
   - Computes the log likelihood and the standard error of the model. (also remind what these are)

6. Based on your previous tests, which model is the best ?

7. Plot the predictions of your models alongside the expected results.

We are now going to study 3 models in particular: ARMA(1,1), ARMA(1,2) 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 (*stats.shapiro()*) on the residuals of all 3 models applied to the logGNP data and display their autocorrelogram. From there, what can you say on the stationarity of the residuals ? How do you justify which model is the best ?

C ARIMA Model

Differentiate the GNP serie a second time and use the analysis of sections B) 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.