Time Series Analysis: 780 (Tampere University)

Assignment 1

Using the Excel file `Case_study.csv` consisting of time series data (Y_t) (in the assigned column), one would like to propose an ARIMA(p, d, q) specification and forecast the time series. In so doing, one proceeds in three steps:

1. Preliminary analysis of the autoregressive order (p), the moving average order (q), and the differencing order (d);
2. Estimation and selection of some ARIMA models;
3. Implementation of some diagnostic tests (especially, for the error terms).

Step 1: Preliminary analysis of orders

1. Plot the times series (Y_t).
2. Analysis of d :
   (a) Display the autocorrelation function of (Y_t) and $\nabla^k Y_t,\ k = 1, 2$ where $\nabla Y_t = Y_t - Y_{t-1}$. Interpret the results:

   (1) What do you observe?

   (2) Using this informal graphical procedure, is there some evidence that the series (Y_t) might be non-stationary? Explain carefully.

   (b) To provide more evidence, implement a unit root test, the so-called Augmented DickeyFuller test (ADF), to assess whether there exists a root on the unit circle in the autoregressive lag polynomial of an AR(p) specification.

   (c) Which order do you suggest for d?
3. Analysis of (p, q):

   Using the result of part 2, plot the ACF and PACF of the relevant series (Y_t or $\nabla^k Y_t$). Provide some upper bounds for p and q, say $p_{\max}$ and $q_{\max}$.

   Hint: Upper bound of the moving average part (respectively, autoregressive part) can be achieved by looking at the ACF (respectively, PACF).

Step 2: Estimation and selection of ARIMA Models

At this stage, the differencing order is assumed to be known (from part 2), and $0 \leq p \leq p_{\max}$ and $0 \leq q \leq q_{\max}$. In the sequel, one assumes $p_{\max} = 4$ and $q_{\max} = 4$. Furthermore, AR(p), MA(q) and weak white noise specifications rules out, i.e. one rules out the following (p, q) combinations $(0, 0), (0, q)$ and $(p, 0)$.

1. Compute the AIC (Akaike's information criterion) and BIC (Bayesian information criterion) for the 16 remaining models.

   Hint: Among other methods, this can be done as follows

   - Define the competing models

- Fit models to the data
- Compute the AIC and BIC for each model

2. Provide a table with the estimates of your best three specifications.

   Hint: Select the models that minimize each information criterion. Note that:

   - The statistical properties of these two information criteria are different (especially, in terms of consistency and efficiency). It turns out that these two criteria might select different models.
   - Instead of selecting a unique model (specification) with each criterion, it is often better to select a model that minimizes the corresponding information criterion but also models that are "adjacent" (i.e., models that have a value of the information criterion close to the minimal one).

Step 3: Diagnostic tests

1. Among the different tests of specifications, one would like to test the absence of correlations of the error(residual) terms. Using the results of part 2 in Step 2, conduct a LjungBox test for the first 10 lags. Also, provide ACF and PACF plots. What can you conclude?

2. Are the residuals normal?

   Hint: Use histograms, QQ plots, and Shapiro-Wilk's test to conclude the normality of the residuals for the selected 3 specifications.

3. All in all, Which model is your preferred specification?

   Hint: Make use of all previous results as well as the principle of parsimony (i.e., among two competing models, one would generally select the one with less parameters, especially with respect to the moving average component).

4. Plot original time series and best-fitted model.

Step 4: Forecast

1. Generate a forecast plot for a time series, incorporating a 95% confidence interval. Plot the original time series and $h-$step prediction. Take $h = 10$ and $h = 25$.

   Hint: Use the preferred specification as in Step 3 part 3.

Additional information

- You are required to use R and the Deadline is: Feb 25, 2024.
- You are required to upload two files: the report file (please see the attached file for format) and the R file which contains the R code for the assignment.
- Please upload your report along with your code by the deadline.
- Your report should be detailed (Please consider the solution to exercise 2 for the format i.e. first the code (perhaps in a different style or color), then the output of the code followed by the precise interpretation). Interpreting results is a key part of the exercise.
- Write your name, ID, and the column assigned at the beginning of the code. The codes should be well commented.