DSC 275/475: Time Series Analysis and Forecasting (Fall 2022)

Project-1 (Total points: 60 for undergraduate students; 70 for Graduate students and including extra credit for undergraduate students)

Overview

This project is designed to provide you hands-on experience working on an end-to-end time series analysis and forecasting solution using AR/ARMA/ARIMA/SARIMA modeling. You are welcome to use any external libraries/packages for this project. A few recommendations are provided for each problem below.

This is a guided exercise in that you are expected to answer each of the questions below. *For some of the questions, you will appreciate there is no single correct answer. In such cases, you have flexibility to decide the approach. Any conclusions you make or decisions you take must be stated and accompanied by a reasonable justification.*

Your submission should be a PDF document with responses (including figures/plots) to each question along with the code either included inline [e.g. Notebook] or as a separate file.

Problem:1 (60 pts: Required for all students)

The data for this project (Problem1_DataSet.csv) represents 7 years of monthly data on airline miles flown in the United Kingdom. You are tasked with the goal of developing a forecasting model that can accurately predict the trend for future years. To achieve the final goal, answer each of the questions below.

1. Create a time series of the plot of the data provided. **(5 pts)**

2. Plot the autocorrelation function (ACF). From the ACF, what is the seasonal period? **(5 pts)**

3. Compute a moving average for the data to determine the trend in the data and overlay on the original time-series plot. What is a suitable choice for the moving average window length? **(5 pts)**

4. Observing the moving average plot in Q3, is the trend line increasing or decreasing? **(5 pts)**

5. Compute the first difference of the data and plot the ACF and PACF for the differenced data. What are the significant lags based on the ACF and PACF? **(5 pts)**

6. Using the output from Q5 above, perform a first seasonal difference with the seasonal period you identified in Q2, and plot the ACF and PACF again. What are the significant lags based on the ACF and PACF? **(5 pts)**

7. Develop a suitable SARIMA model that can be applied on the time series. Use the first 6 years of data only to develop the model. **(20 pts)**

a. To develop the model, vary the model parameters for the non-seasonal *(p,d,q)* and seasonal components *(P,D,Q)* and calculate the output for each combination of parameters.
b. Use an evaluation criteria such as AIC, BIC or sum squared error or mean squared error to determine the ***best choice of parameters*** *(p,d,q,P,D,Q)*. *Note:* AIC and BIC are metrics that is readily output by the ARIMA model.

Suggestions for Problem 1, Q7:

- *For Q7, in Python, we suggest using the package/function SARIMAX in the "statsmodels.tsa.statespace" library*
- *You can choose the range of values to search for the model parameters. We suggest varying p, q and P, Q each over the range 0 to 3 to constrain the search range.*
- *The SARIMA estimation procedure internally uses numerical optimization procedures to find a set of coefficients for the model. These procedures can fail for some combination of model parameter values which in turn can throw Python errors. We must catch these exceptions and skip those configurations that cause a problem. To solve this problem, include "try/except" blocks in your code when iterating through the parameter values (pseudocode below):*

```
try:
        ##Your code here with the SARIMAX function
except:
        continue
```

8. Use the model parameters determined in Q7 above to forecast for the 7th year. Compare the forecast with actual values. Comment on your observations. **(10 pts)**

Problem 2 (10 pts): (Required for Graduate students; Extra credit opportunity for Undergraduate students)

In this problem, you will develop a time-series model to analyze Wine consumption from the data file "TotalWine.csv".

a) Plot the time series for TotalWine. What is the seasonal period for this time-series? **(1 pt)**

b) Apply seasonal differencing to the original time-series. Vary the difference lag from 1, 2, 4, 6. Plot the result for each of these lags. Which of these differences is most suitable to remove the seasonality? **(2 pts)**

c) Compute and plot the Auto-correlation (ACF) function for the original time-series. What is the seasonal period you estimate from the ACF? **(1 pt)**

d) Define an AR model using *tsa.AR* available in statsmodels.api. Determine the optimal order using the "select_order" function. You will need to specify a maximum order *p* (recommend p=10) to consider and a criterion for deciding which model order is "best". [e.g. You can use AIC as the model selection criteria] **(2 pts)**

e) Now, evaluate an AR(p) model for the time-series generated after seasonal differencing (using the best lag you found in part b above) **(4 pts)**

   i. use the fit method specifying the optimal lag found above

ii. use the predict method to generate values starting at the optimal lag

iii. plot the predicted results and the corresponding seasonally differenced time-series

iv. Calculate the Mean Absolute Error (MAE) by comparing the predicted results with the seasonally differenced data.