Project 3: Forecasting

Objectives

To use various forecasting algorithms to determine the best model for a time series that will be created using your student id (SI) number. For this analysis you can use any statistical package, such as MatLab, R, SAS, or use Python with all the available statistical functions. You do not need to use one package. You may use different packages/functions to execute different tasks.

Data Set

You will use the data set provided to you that was generated based on your student id (SI) number. The set of all data sets for all the students in the class is posted in the folder entitled "Individual data files" in the "Project 3 - Forecasting" folder. **Use the one that has your ID**.

Your data set is a time series consisting of 2000 observations. Partition your data set into two parts, the *training set* consisting of 1500 observations and the *testing set* consisting of the remaining 500 observations. Use the training set to do tasks 2, 3, and 4. Use the testing set to do task 5.

Task 1. Check for stationarity

Plot the entire time series (i.e. all 2000 observations) and check it visually for stationarity and make necessary appropriate transformations as discussed in section 6.1.2. Comment on your conclusions.

Task 2. Fit a simple moving average model (use the training set)

1.1. Apply the simple moving average model $s_t = (1/k)\sum_{i=t-k}^{t-1} x_i$ to the training data set, for a given k.
1.2. Calculate the error, i.e., the difference between the predicted and original value in the training data set, and compute the root mean squared error (RMSE).
1.3. Repeat the above two steps by varying k and calculate the RMSE.
1.4. Plot RMSE vs k. Select k based on the lowest RMSE value. For the best value of k plot the predicted values against the original values.
1.5. Comment on your results.

Task 3. Fit an exponential smoothing model (use the training set)

2.1 Apply the exponential smoothing model $s_t = ax_{t-1} + (1-a)s_{t-1}$ to the training data set for $\alpha = 0.1$.
2.2 Calculate the error, i.e., the difference between the predicted and original value in the training data set, and compute the root mean squared error (RMSE).
2.3. Repeat steps 2.1 and 2.2 by increasing a each time by 0.1, until $a = 0.9$.
2.4. Plot RMSE vs a. Select a based on the lowest RMSE value.
2.5. For the selected value of a plot the predicted values against the original values, and visually inspect the accuracy of the forecasting model.
2.6. Comment on your results.

Task 4. Fit an AR(p) model (use the training set)

3.1 First select the order p of the AR model by plotting PACF in order to determine the lag k at which PACF cuts off, as discussed in section 6.4.4.

3.2 Estimate the parameters of the AR(p) model. Provide RMSE value and a plot the predicted values against the original values.
3.3 Carry out a residual analysis to verify the validity of the model.
   a. Do a Q-Q plot of the pdf of the residuals against $N(0, s^2)$ In addition, draw the residuals histogram and carry out a χ^2 test that it follows the normal distribution $N(0, s^2)$.
   b. Do a scatter plot of the residuals to see if there are any correlation trends.
3.4. Comment on your results.

Task 5. Comparison of all the models (use the testing set)
Run the the above three trained models on the test data, and chose the best one. Comment on your results.

What submit
Submit all the requested results, your conclusions, and the code that you wrote to obtain the results. It is very important that you provide enough results to support your conclusions. Conclusions without insufficient results will make you lose points. Also, it is important that you develop your own code. Sharing code is not allowed and constitutes cheating, in which case both students (the one that aids and the one that receives) will get a zero for the project and will be reported to the student conduct office.

Grading
The TA will first verify that your code works and produces the results you submitted. The break down of the grades will be as follows:

Task 1: 30 points
Task 2: 30 points
Task 3: 30 points
Task 4: 10 points

Remember that you will be graded mostly on your ability to interpret the results