Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Lab 3

Professor Jeffrey Yau

March 18, 2018

Instructions:

- **Due Date: To be discussed in Live Session 10**
- **Page limit: 12 pages (This is a hard limit; work beyond page 12 will note be graded)**
- Submission:
  - Submit your own assignment via ISVC
  - Submit 2 files:
    1. A pdf file including the summary, the details of your analysis, and all the R codes used to produce the analysis. Please do not suppress the codes in your pdf file.
    2. R markdown file used to produce the pdf file
  - Each group only needs to submit one set of files
  - Use the following file naming convensation; fail to do so will receive 10% reduction in the grade:
    - FirstNameLastName_LabNumber.fileExtension
    - For example, if you have two students named John Smith and Jane Doe, you should name your file the following
      - JohnSmith_JaneDoe_Lab1.Rmd
      - JohnSmith_JaneDoe_Lab1.pdf
  - Although it sounds obvious, please write the name of each members of your group on page 1 of your pdf and Rmd reports.
  - This lab can be completed in a group of up to 3 people. Each group only needs to make one submission. Although you can work by yourself, I encourage you to work in a group.
- Other general guidelines:
  - Try to use only techniques and R libraries that are covered in this course.
  - If you use R libraries and/or functions to conduct hypothesis tests not covered in this course, you will have to explain why the functions you use are appropriate for the hypothesis you are asked to test. Lacking explanations will result in a score of zero for the corresponding question.
  - Thoroughly analyze the given dataset. Detect any anomalies, including missing values, potential of top and/or bottom code, etc, in each of the variables.
  - Your report needs to include a comprehensive Exploratory Data Analysis (EDA) analysis, which includes both graphical and tabular analysis, as taught in this course. Output-dump (that is, graphs and tables that don't come with explanations) will result in a very low, if not zero, score.
  - Your analysis needs to be accompanied by detailed narrative. Remember, make sure your that when your audience (in this case, the professors and your classmates) can easily understand your your main conclusion and follow your the logic of your analysis. Note that just printing a bunch of graphs and model results, which we call "output dump", will likely receive a very low score.
  - Your rationale of any decisions made in your modeling needs to be explained and supported with empirical evidence. Remember to use the insights generated from your EDA step to guide your modeling step, as we discussed in live sessions.

  - All the steps to arrive at your final model need to be shown and explained very clearly.
- Other requirements:
- Groups will be asked to present their lab in the live session following the submission date.
- Students are expected to act with regards to UC Berkeley Academic Integrity.

Question 1

ECOMPCTNSA.csv, contains quarterly data of E-Commerce Retail Sales as a Percent of Total Sales. The data can be found at: https://fred.stlouisfed.org/series/ECOMPCTNSA.

Build a Seasonal ARIMA model and generate quarterly forecast for 2017. Make sure you use all the steps of building a univariate time series model between lecture 6 and 9, such as checking the raw data, conducting a thorough EDA, justifying all modeling decisions (including transformation), testing model assumptions, and clearly articulating why you chose your given model. Measure and discuss your model's performance. Use both in-sample and out-of-sample model performance. When training your model, exclude the series from 2015 and 2016. For the out-of-sample forecast, measure your model's performance in forecasting the quarterly E-Commerce retail sales in 2015 and 2016. Discuss the model performance. Also forecast beyond the observed time-period of the series. Specifically, generate quarterly forecast for 2017.

Question 2

You will use the series contained in *data_2018Spring_MTS.txt* to conduct a multivariate time series analysis. These series could be completely decoupled or interdependent of each other. Your task is to conduct a multivariate time series analysis and build a model to forecast the series in 1993 and 1994. In model estimation, do not use the observations in 1993. All the model building steps covered in lecture 6 - 10 are applicable.

As always, checking the raw data, conducting a thorough EDA, justifying all modeling decisions (including transformation), testing model assumptions, and clearly articulating why you choose your final model. Measure and discuss your model's performance. Use both in-sample and out-of-sample model performance. When training your model, exclude all the observations in 1993. For the out-of-sample forecast, measure your model's performance in forecasting 1993. Discuss the model performance. Also forecast beyond the observed time-period of the series. Specifically, generate a 12-month forecast beyond the last observed month in the given series.