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

W271 Instructional Team

July 21, 2017

Instructions:

- **Due Date: 8/6/2017 (11:59 p.m. PST, Sunday)**
- 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:
    - SectionNumber_hw01_FirstNameLastNameFirstInitial.fileExtension
    - For example, if you are in Section 1 and have two students named John Smith and Jane Doe, you should name your file the following
      - Section1_hw01_JohnS_JaneD.Rmd
      - Section1_hw01_JohnS_JaneD.pdf
  - Although it sounds obvious, please write the name of each members of your group on page 1 of your report.
  - 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, we encourage you to work in a group.
  - When working in a group, do not use the "division-of-labor" approach to complete the lab. That is, do not divide the lab by having Student 1 completed questions 1 - 3, Student 2 completed questions 4 - 6, etc. Asking your teammate to do the questions for you takes away your own opportunity to learn.
- Other general guidelines:
  - Please read the instructions carefully.
  - Please read the questions carefully.
  - 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 function you use is appropriate for the hypothesis you are asked to test
  - 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.

  - 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.
- Students are expected to act with regards to UC Berkeley Academic Integrity.

Question 1:

During your EDA, you notice that your data exhibits both seasonality (different months have different heights) AND that there is a clear linear trend. How many order of non-seasonal and seasonal differencing would it take to make this time-series stationary in the mean? Why?

Question 2: SARIMA

It is Dec 31, 2016 and you work for a non-partisan think tank focusing on the state of the US economy. You are interested in forecasting the unemployment rate through 2017 (and then 2020) to use it as a benchmark against the incoming administration's economic performance. Use the dataset *UNRATENSA.csv* and answer the following:

(A) Build a SARIMA model using the unemployment data and produce a 1 year forecast and then a 4 year forecast. Because it is Dec 31, 2016, leave out 2016 as your test data.

- How well does your model predict the unemployment rate up until June 2017?
- What does the unemployment rate look like at the end of 2020? How credible is this estimate?

(B) Build a linear time-regressionand incorporate seasonal effects. Be sure to evaluate the residuals and assess this model on the basis of the assumptions of the classical linear model, and then produce a 1 year and a 4 year forecast.

- How well does your model predict the unemployment rate up until June 2017?
- What does the unemployment rate look like at the end of 2020? How credible is this estimate?
- Compare this forecast to the one produced by the SARIMA model. What do you notice?

Qestion 3: VAR

You also have data on automotive car sales. Use a VAR model to produce a 1 year forecast on both the unemployment rate and automotive sales for 2017 in the US.

Compare the 1 year forecast for unemployment produced by the VAR and SARIMA models, examining both the accuracy AND variance of the forecast. Do you think the addition of the automotive sales data helps? Why or why not?