Applied Regression and Time Series Analysis: Lab 3

Jeffrey Yau and Devesh Tiwari

October 25, 2016

Instructions:

- **Due Date: 12/16/2016 (Friday)**
- Late submission will not receive any credit.
- Instructions must be followed strictly.
- This lab can be completed in a group of up to 3 people. Each group only needs to make one submission.
- Submit 2 files: (1) a report (in pdf format) detailing your analyses; (2) your R script or juypter notebook supporting all of your answers. Missing one of this files will result in an automatic 50% reduction in score.
- 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 or data series. Detect any anomalies in each of the variables.
- Your report needs to include a comprehensive graphical analysis
- Your analysis needs to be accompanied by detailed narrative. Just printing a bunch of graphs and econometric results will likely receive a very low score.
- Your analysis needs to show that your models are valid (in statistical sense).
- Your rationale of using certian metrics to choose models need to be provided. Explain the validity / pros / cons of the metric you use to choose your "best" model.
- Your rationale of any decisions made in your modeling needs to be explained and supported with empirical evidence.
- All the steps to arrive at your final model need to be shown and explained clearly.
- All of the assumptions of your final model need to be thoroughly tested, explained, and shown to be valid. Don't just write something like, *"the plot looks reasonable", or "the plot looks good*, as different people interpret vague terms like "reasonable" or "good" differently.
- Students are expected to act with regards to UC Berkeley Academic Integrity.

Forecast Inflation-Adjusted Gas Price

During 2013 amid high gas prices, the Associated Press (AP) published an article about the U.S. inflation-adjusted price of gasoline and U.S. oil production. The article claims that there is "*evidence of no statistical correlation*" between oil production and gas prices. The data was not made publicly available, but comparable data was created using data from the Energy Information Administration. The workspace and data frame *gasOil.Rdata* contains the U.S. oil production (in millions of barrels of oil) and the inflation-adjusted average gas prices (in dollars) over the date range the article indicates.

You have three tasks for this exericse, and both tasks need the use of the data set *gasOil.Rdata*.

Task 1: Create a **SARIMA-type** model to forecast the inflation-adjusted gas prices, and use this model to forecast the inflation-adjusted gas price for the next two years.

Task 2: Create a *multivariate* time series model that can be used to predict/forecast inflation-adjusted gas prices, and use this model to forecast the inflation-adjusted gas price for the next two years

Task 3: Compare the accuracy of the two models' forecasting results. Also, compare and contrast the results of these two models. Is one model better than the other? What metric(s) do you use to measure whether one model is "better" than the other? Why or why not? Explain the pros and cons of each of the models in this specific context.