Applied Regression and Time Series Analysis: Lab 3

July 30, 2016

Instructions:

- Instructions must be followed stritly.
- Form a group of (ideally) 3 people. Each group only needs to make one submission. Note that unlike what is stated in the syllabus, you have the option to work by yourself and not in a group, although it is not encouraged.
- 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.
- Late submission will not receive any credit.
- Use only techniques and R libraries that are covered in this course.
- Thoroughly analyze the given dataset or data series. Detect any anomalies in each of the variables. Examine if any of the variables that may appear to be top- or bottom-coded.
- 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 and 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.
- We have been told that some members in a group project did not contribute as much as they should. Please respect your teammate and make adequate contribution to the project. If anyone of you feel that any your teammate does not make sufficient contribution to the project, please report to us.
- Most importantly, students are expected to act with regards to UC Berkeley Academic Integrity.

Part 1 (50 points): Forecast the Web Search Activity for global Warming

The file, **google_correlate_flight.csv**, contains the relative web search activity for the phrase "flight prices" over time. Your data science team believes that search activity for this phrase is positively correlated with consumer demand for flying (and possibly prices). Your task is to forecast the relative demand of this phrase for the year 2016. For the purposes of this assignment, ignore the units of the data as they are not relevant here.

Remember to explain and justify each of your modelling decisions. Also, comment on your forecast. Do you notice anything interesting? Do you notice anything worth worrying about?

Part 2 (50 points): 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.

In support of their conclusion, the AP reported a single p-value. You have two tasks for this exericse, and both tasks need the use of the data set *gasOil.Rdata*.

Your first task is to recreate the analysis that the AP likely used to reach their conclusion. Thoroughly discuss all of the errors the AP made in their analysis and conclusion.

Your second task is to create a more statistically-sound model that can be used to predict/forecast inflation-adjusted gas prices. Use your model to forecast the inflation-adjusted gas prices from 2012 to 2016.

Part 3 (50 Points): Forecast two series

The flie *ex3series.csv* contains two monthly time series. You task is to (1) build a forecasting model using techniques covered in lecture 8 - 13, which includes the async. lectures, live sessions, and the assigned readings, and (2) produce a 6-step ahead (monthly) forecas. Please follow the instructions outlined above. To repeat,

- Your analysis 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, if not a score of zero.
- 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 and 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.