Final Exam

Jeffrey Yau

November 25, 2015

Instructions:

- Instructions must be followed stritly.
- Form a group of 3 or 4 people. Each group only need to make one submission.
- Send me the names of your group by **12/9/2015**. I will post them to the ISVC Wall as a confirmation.
- 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.
- ** Due Date: 11:59pm Pacific Standard Time on 12/18/15. Late submission will not be accepted.**
- Use only techniques and R libraries that have been 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.
- 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.

- 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".

Part 1 (40 points): Classical Linear Regression Model

In Part 1, you will use the data set **houseValue.csv** to build a linear regression model, which includes the possible use of the instrumental variable approach, to answer a set of questions interested by a philanthropist group. You will also need to test hypotheses using these questions.

The philanthropist group hires a think tank to examine the relationship between the house values and neighborhood characteristics. For instance, they are interested in the extent to which houses in neighbhorhood with desirable features command higher values. They are specifically interested in environmental features, such as proximity to water body (i.e. lake, river, or ocean) or air quality.

The think tank has collected information from tens of thousands of neighborhoods throughout the United States. The think tank hires your group as contractors, and you are given a sample and selected variables of the original data collected to conduct an initial analysis. Many variables, in their original form or transfomed forms, that can explain the house values are included in the dataset. Analyze each of these variables (as well as a combination of them) very carefully and use them (or a subset of them) to build a model and test hypotheses to address the questions. Also address potential (statistical) issues that may be casued by omitted variables.

Part 2 (20 points): Time Series Modeling

Build a time-series model for the series in **series02.txt** and use it to perform a 24-step ahead forecast. Possible models include AR, MA, ARMA, ARIMA, Seasonal ARIMA, GARCH, ARIMA-GARCH, or Seasonal ARIMA-GARCH models. Note that the original series may need to be transformed before it be modelled.

Part 3 (20 points): Time Series Modeling

Build a time-series model for the series in **series03.csv** and use it to perform a 24-step ahead forecast. Possible models include AR, MA, ARMA, ARIMA, Seasonal ARIMA, GARCH, ARIMA-GARCH, or Seasonal ARIMA-GARCH models. Note that the original series may need to be transformed before it be modelled.

Part 4 (20 points): Time Series Modeling

Build a time-series model for the series in **series04.csv** and use it to perform a 24-step ahead forecast. Possible models include AR, MA, ARMA, ARIMA, Seasonal ARIMA, GARCH, ARIMA-GARCH, or Seasonal ARIMA-GARCH models. Note that the original series may need to be transformed before it be modelled.