W271 Live Session
Week 7
Jeffrey Yau and Devesh Tiwari

Agenda:

1. Comments and discussions on Lab3 proposal (45 minutes)
2. Comments on homework (and the bonus exercise) in general (10 minutes)
3. A few comments Causal impact analysis (10 minutes)
4. An overview of the 2nd half of the course (i.e. Time Series Statistical Analysis) (25 minutes)

1. Proposal comments

There were several types of proposals we have seen.

Overall / General Comments:

- Most of them propose to answer interesting (and in some cases, very important) questions.
- However, many of them, as it stated, are too broad and general to be appropriately addressed in a data science project that can be addressed using techniques taught in this course.
- Below are three types of proposals, loosely categorized, we have seen in all three sections.

1. Kaggle predictions.

- Those who use examples from Kaggle competitions have to be very careful. The focus on Kaggle competition is purely on precision; regardless of the algorithm proposed, that achieving the best "metric measure" wins. The techniques taught in this class go way beyond that. Prediction is important, but in many problems, prediction accuracy is not the only objective.
- When building statistical models, be that classical linear regression models or time series models we will learn starting next week, you will need to pay attention to the validity of the models by conducting diagnostics and formal hypothesis testing. Your lab will be graded on **the entire model building process and not just the prediction results**, the codes of many of which are already available on Kaggle.
- Again, we want to emphasize the need of building statistical models, which thus requires you to assess your model in light of the underlying statistical assumptions. Even if you are going down the prediction route, ten you still need to evaluate your model in this vein!
- More importantly, if you do use other analysts' codes (and model specification), you must cite them. I know that Kaggle posts a lot of codes to the public.

2. Understanding the drivers of Y

- I read proposals that wanted to understand what variables had an impact on some dependent variable. One problem with this approach is that it does not specify which variables the researcher should focus on, or why.
- But a deeper problem is that we do not know what it means for a variable to be a "driver." Are we asking which variables — from a set of variables — are statistically significant? Or do we care about practical significance?
- Also, remember that the statistical AND practical significance for variables change depending on which variables are included.
- So, the key is to find a valid model that can address the underlying questions you'd like to answer, which means you need to evaluate it in light of the CLM assumptions — and then discuss which variables are important and why.

3. Understanding the relationship between X and Y

- These proposals clearly identified a dependent and a set of independent variables and an independent variable of interest.
- These proposals areappropriate for the methods we have discussed in class thus far, but these projects still need to conduct thorough analyses in order to find the best model, they should take the time to explain the results of each variable in the model, and they should explain whether their results are practically significant.
- Finally, they should be able to answer the question they posed upfront.

On using Panel data for your lab:

- With the exception of differences-in-differences, we do not cover techniques (both EDA and statistical models) in this course.
- Some of the questions, however, indeed require panel data

2. **Homework**

- In HW1, we graded very generously.
- Many do not follow file naming format, making it very difficult for the instructors to organize the file.
- Although I have not specified in the instruction, I thought I do not need to explicitly spell the that you should put your name or names of the people in your group in your work.
- Formatting is important. Please answer questions clearly in the main body of your documents and not in your R-code. If you do the latter, then it is difficult for us to find your answer. I don't think the rest of the homework will involve hand-written exercise, but some of the hand-written solutions were not legible at all.
- Do not just create a large set of tables / charts without providing any context or narrative. In general, think of each step of the model building process as answering a series of questions, write out you answers in clean prose, and then refer to tables and charts below.
- When you are testing hypothesis, be sure to formally state your null hypothesis and say which test you are going to use and why.
- For residual diagnostics, don't just use the plot command, as it does not provide all the residual plots. Also, after all, these are only diagnostics, formal testing of the assumptions are needed.

3. A discussion on quantifying causal impact of some "treatment" of interest

4. An overall of time series analysis

- Our approach to learn time series analysis
  - 1. Learn about the mathematical formulation of a model or an important concept (such as autocorrelation function)
  - 2. Derive some of the most important properties of the model
  - 3. Simulate (using R or Python) a number of realizations from the model
  - 4. Examine the empirical properties exhibited in the simulated realizations
  - 5. Apply the model or concept to a real world data sets.
- A (very general) Approach to ARIMA modeling
  - 1. Based on the interaction of theory, subject-matter expertise, and practice, consider a useful class of models
  - 2. Collect and cleanse the data
  - 3. Conduct with exploratory time series data analysis (ETSDA) by plotting the series and examine the main patterns and atypical observations of the graph, after collecting and "cleaning" the data,
    - Trend
    - The fluctuation around a trend
    - – Seasonal variation (or Seasonality)
    - – Cyclical variation (that does not appear to be seasonal variation)
    - Sharp change in behaviour (i.e. structural change or jumps)
    - Outliers
  - 4. Examine and (statistically) test whether the series is stationary (when applying a stationary model)
  - 5. If the series is not stationary, transform the series to a stationary series, because the time series models covered in this course apply only to stationary or integrated times series. Common transformation techniques include trend removal (i.e. detrending), seasonality removal, logarithmic, and difference transformation.
  - 6. Model the transformed series with a stationary or integrated time series model
  - 7. Examine the validity of the model's underlying assumptions.
    - This is an important step, because if the model's underlying assumptions are not satisfied, one should not proceed to conducting statistical inference and forecasting.
  - 8. Among the valid models, choose the one that perform "best" according to some pre-specified metrics.
  - 9. Once a (statistically) valid model is chosen, Conduct statistical inference and/or forecasting, if the underlying statistical assumptions are all satisfied.