Recitation 5: Data Science

Eugene Tan (TA)

10/11/2020

Recitation today

- Go through problem set
- Data Science as a process (mostly from r4ds)
- Oil and gasoline

Datacamp

- 6 of you haven't finished 'Data Manipulation with data.table in R'
- Essential for finishing the course, I said that you should finish it before the homework.
- One datacamp exercise/mini-exercise due each week - most of you will have already finished it.
- Strong correlation in datacamp progress and HW1 grades.

Go through problem set

- Questions at the end
- Causality

Causality

Does the previous regression capture the causal effect of coal capacity on carbon emissions? Why?

- Does coal capacity affect carbon emissions?
- Q: If we changed coal capacity, would it affect emissions?
- It limits generation (this is the *mechanism*)
- So the regression captures some of the causal effect *but does not identify it*.

Identification

If there is bias, do you expected to go in a particular direction? Explain.

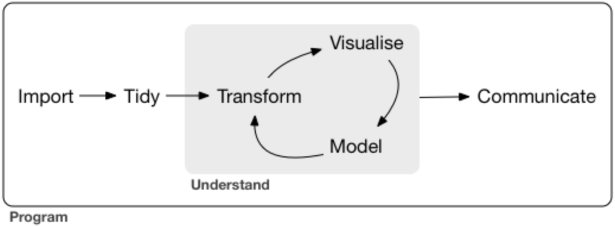
- We identify or establish causality when we have eliminated or ruled out sources of bias.
- In linear regression, then we can say if we change x, y should change by beta.
- This is why you need controls.
- Was there an upward bias or downward bias?
- Consider that natural gas is correlated with coal.

Data Science vs Econometrics

- Data science is the study of how to apply scientific tools, methods, and mindset to extract knowledge from data.
- Econometrics is the branch of economics concerned with the use of mathematical methods (especially statistics) in describing economic systems.

Data Science can be described as a process

- Econometricians don't consider getting/cleaning the data econometrics
- Data Science is a whole process



Flip the classroom

Work on importing, and visualizing this dataset.

Can be useful to look at the report alongside the dataset

Import

take data stored in a file, database, or web application programming interface (API), and load it into a data frame in R. * If you can't get your data into R, you can't do anything with it! * read excel, csv, apis

Tidy $\in$ Wrangle

- store it in a consistent form that matches the semantics of the dataset with the way it is stored.
- each column is a variable, and each row is an observation.
- consistent structure lets you focus your struggle on questions about the data

Transform $\in$ Wrangle

- narrowing in on observations of interest (like all people in one city, or all data from the last year),
- creating new variables that are functions of existing variables (like computing speed from distance and time),
- calculating a set of summary statistics (like counts or means).

Visualize

- show you things that you did not expect
- raise new questions about the data
- statistics is about the shape of data

Modelling ("Econometrics")

- DS will normally use simulations and Machine Learning
- Econ just uses regressions

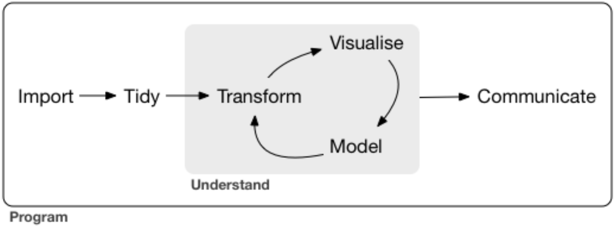
ML vs Econometrics

- Prediction vs causality
- iid data vs panel, time-series, cross-sectional data
- Cross-validation vs theory

Communicate

- Write words to communicate!
- Present work well!

Back to the framework



Exploratory Data analysis

Develop an understanding of your data * What type of variation occurs within my variables? * What type of covariation occurs between my variables?