Regression and causal analysis

Harry Ganzeboom

Research Skills, December 4 2008

Lecture #5

Regression analysis is about partial associations

- Note that Berry & Feldman do not give any causal interpretation to regression outcomes.
- In fact, they do not even discuss that coefficients in multiple regression can be interpreted as partial coefficients: they show how X1 varies with Y, while X2 is controlled ('held constant').
- Regression can be used as descriptive tool (also: predictive tool = trend analysis, forecasting), but its power derives from causal interpretation.
- Like in in any kind of causal interpretation, causal conclusions derive from: causal order assumptions + partial associations.

Correlation and causation

- Often heard: “correlation does not imply causation”.
- This is not a productive statement. Better:
  - Causation implies correlation.
  - Correlations are brought about by <u>some</u> causal process:
    - Direct causation: $X \rightarrow Y$
    - Reversed causation: $Y \rightarrow X$
    - Confounding (common cause): $Y \leftarrow Z \rightarrow X$

And what if no correlation?

- The absence of a correlation does not imply the absence of causation.
- Because (positive) correlation brought about by causation X→Y can be suppressed by a confounder Z that has a positive effect on X and a positive effect on Y.

Causation

- X is regarded to be a cause of Y, if:
  - X is causally prior to Y [causal order]
  - All variables Z that are causally prior to both X and Y are controlled (‘held constant’)
  - And there is still a correlation between X and Y.
- Note that causal order needs to brought into the analysis by research design and/or as a (plausible) assumptions.

Causation in experiments

- X is causally prior to Y by design
- Confounders are ruled out by design (randomization)
- Correlation ➔ causation

- Control variables in experiments may be used:
  - To boost statistical power
  - For post-stratification (post-hoc matching)

Causation in observational designs

- Step 1: Identify possible confounders
- Step 2: Make sure that you can safely assume that these confounders are indeed causally prior:
  - Design: retrospective or panel measurement
  - Assumptions: e.g. life-course, generic-specific
- Step 3: Measure all confounders without error.
- Step 4: Control all confounders:
  - Table elaboration
  - Regression analysis
  - Propensity score matching.

Frequent problems in causal modeling

- Controlling intervening variables
- Controlling variables with ambiguous status with respect to causal order.
- Incomplete or badly measured confounders.

Controlling (1): elaboration

- In elaboration, you split up the data by values of the confounders and find the average effect across these partial tables.
- This is all problematic when you have many and multi-valued control variables.
- The technique is often done with cross-tabs, but it works better with conditional means.

Controlling (2): regression

- Regression coefficients have a partial interpretation: b(Y,X1) shows the effect of X1, while controlling other X.
- This can be easily be done with a large number of (continuous) X-variables.

Controlling (3): matching

- Increasingly regaining popularity is the use of propensity score matching: you compare each case in the treatment group to a matching case in the control group.
- This works well for discrete (0/1) X (=treatment) variables.
- Is an appealing technique for the laymen, but there is also a claim that it is superior to regression analysis.