Synthesising Observational and Experimental Evidence

Ed Jee

Motivation

- Whilst conducting field experiments researchers often have access to additional, non-experimental, data sources or samples.
  - This can be used to inform pilots or as baseline covariates.

But observational data can also be used to inform our estimates of treatment effects beyond simple covariate adjustment in our experimental sample.

- Combining evidence is attractive for multiple reasons:
  - increase power cheaply.
  - discarding data seems wrong from a scientific/Bayesian angle.

Challenges

Depending on the assumptions we're prepared to make this can be very easy, or a little more difficult:

- Homogeneous treatment effects + selection on observables
  - pool observational + experimental together. LaLonde, 1986 suggests this probably isn't a great idea.
- Rank preserving endogeneity
  - "Bigger causal effects imply bigger bias" Peysakhovich and Lada, 2016
- Primary outcome only measured in observational sample, "un-selected" RCTs
  - Athey, Chetty, and Imbens, 2020
- Internal selection bias observationally, site-selection into RCTs
  - "Mutual Debiasing" Gechter and Meager, 2021

The setting I consider is most similar to Gechter and Meager.

Context

- Suppose we observe the universe of data in a country regarding an outcome Y_i and treatment T_i.
- Due to implementation constraints can only randomise encouragement Z_i for treatment T_i in certain regions.
    - Implementing partner/researcher has preferences over experimental site that may be correlated to treatment effects.
    - i.e. decision $D_i = \{e, o\}$ whether an individual is assigned to experimental or observational study is potentially endogenous.

Akin to Gechter and Meager (GM) but with only one study; full micro-data; and, critically, institutional knowledge of study design assignment.

The Problem

Using the notation of GM, we observe a result R, with switching equation:

$$R = \mathbb{I}\{D = e\} \underbrace{R^e}_{TE} + (1 - \mathbb{I}\{D = e\}) \underbrace{R^o}_{TE+SB}$$

- Fundamental problem of causal inference, we never observe the pair: (R^e, R^o).
- $\hat{R}^o$ is contaminated with "internal" selection bias whilst $\hat{R}^e$ is affected by site selection bias.
- If we can find some way of debiasing one study we can mutually debias the other.

Internal Selection Bias

$$E[Y|T=1] - E[Y|T=0] = E[Y_1 - Y_0|T=1] + \underbrace{E[Y_0|T=1] - E[Y_0|T=0]}_{\text{selection bias}}$$

- Internal selection bias arises in observational studies because individuals taking treatment are fundamentally different to units that decide not to take treatment.

- But an RCT with partial compliance nests a **hypothetical observational study** and we can back out the two bias terms.

- $E[Y_0|T=0] \longrightarrow$ corresponds to the *never takers* in an experiment ($E[Y|\underbrace{T=0, Z=1}_{nt}]$).

- $E[Y_0|T=1] \longrightarrow$ the untreated outcome for a treated unit. This is just the untreated *complier* mean.

- Whilst we don't observe this directly we can back it out since p_c, p_n, $E[Y|nt]$ are all known and makeup the control group.

Internal Selection Bias II

We've uncovered the internal selection bias for a *hypothetical* observational study, given our study is experimental.

- We have $\{R^e|R^o = r^o, D = e\}$
- But we want $\{R^e|R^o = r^o, D = o\}$
    - The hypothetical experimental study, given our study is observational.

This is a classic Heckman sample selection problem - we only observe:

$$(\{R^e|R^o, D = o\}, \{R^o|D = e\})$$

In the GM setting we need to find some instrument that shifts research design choice, D, *but is independent of treatment effects*.

Our Solution

However, in our context we have institutional knowledge of the site selection assignment rule.

- Given rich micro-data we can recast the problem as whether *individuals* are assigned to observational or experimental studies.
- Individuals just ineligible for the experimental study are a valid counterfactual for individuals just eligible for the RCT.
- In neighbourhood of cut-off $\{R^e | R^o = r^o, D = e\} = \{R^e | R^o = r^o, D = o\}$
  - i.e. D as good as randomly assigned.

We can debias observational studies by inferring their hypothetical experimental result.

Next Steps

- Simulations.
- Figuring out how much more we can do with knowledge of assignment mechanism.
- Extrapolate away from cutoff.