Causal Inference (6.S059/15.C08/17.C08)

Recitation, Week 12.

Topic: Instrumental Variables

Benjamín Muñoz

May 14, 2024

MIT

Massachusetts Institute of Technology

Two motivations for IV

1. Randomized experiment with non-compliance.
    - Random treatment assignment Z is instrument for actual treatment take-up D.

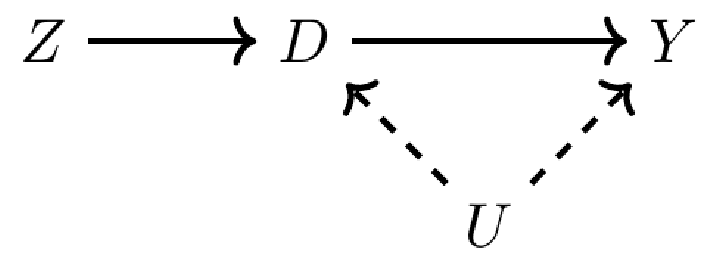
    ★ **Noncompliance:** Participants do not follow the treatment regimen prescribed by the experiment (One-Sided: Treatment Group, Two-Sided: Both Groups).

2. Observational study with potential confounder.
    - Exogeneous variable Z is instrument for endogeneous variable D

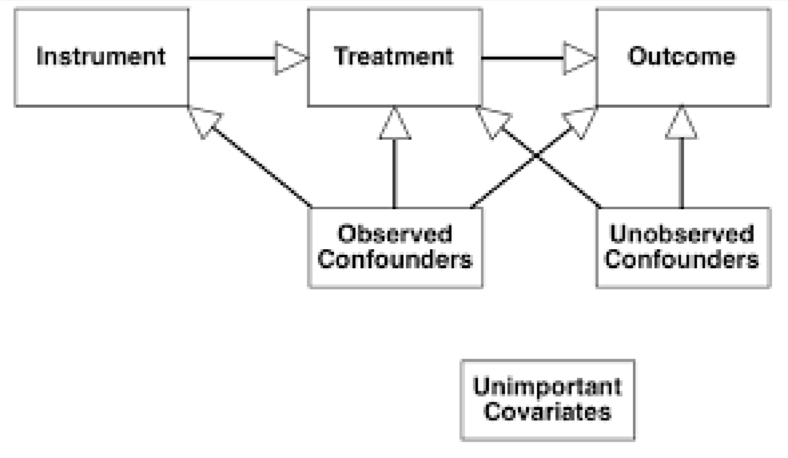
    ★ **Endogeneity:** an explanatory variable is correlated with the error term in a regression model.

    - Simultaneity
    - Omitted Variable Bias
    - Measurement Error
    - Reverse Causality

Basic DAG IV



More Advanced DAG IV



Formalization of Estimands

1. **Experiments with Non-Compliance**
    - Intent-to-treat (ITT) effect: $\mathbb{E}[Y_i^1 - Y_i^0]$
    - "The least we could do" $\longrightarrow$ useful fallback for experiments.
2. **Observational setting with Endogeneity**
    - Local Average Treatment Effect (LATE) ("local" = "compliers").
    - $\mathbb{E}[Y_i^1 - Y_i^0, D_i^1 = 1, D_i^0 = 0]$
    - More important for your theory (we really care about D, not Z).
    - Requires additional assumptions to identify and estimate.
    - Probably "the best we could do."

Intuition of Assumptions

1. **SUTVA**: potential outcomes and treatments for each unit i are unrelated to the treatment status of all other units.
2. **Exogeneity** of the instrument: instrument is ignorable (independent of the treatment and outcome).
3. **Relevance** of the instrument: the effect of instrument on treatment is on average nonzero
4. **Monotonicity**: no defiers
5. **Exclusion restriction**: all the causal effects of the instrument on the outcome must be via the change in the treatment that is induced by the instrument.

Formalization of Assumptions

- ITT requires no further assumptions from a randomized experiment.
- LATE requires several additional assumptions:
  1. **SUTVA** $Y_i(z,d)$ sufficient for ITT, also need $D_i(z)$ for LATE.
  2. **Exogeneity** of the instrument: $\{Y_i^1, Y_i^0, D_i^1, D_i^0\} \perp\!\!\!\perp Z_i$
     - $\{Y_i^z\} \perp\!\!\!\perp Z_i$ (sufficient for ITT, required for LATE)
     - $\{D_i^z\} \perp\!\!\!\perp Z_i$ (required for LATE)
  3. **Relevance** of the instrument (LATE):
     - $0 < Pr(Z_i = 1) < 1$ and $Pr(D_i^1 = 1) \neq Pr(D_i^0 = 1)$
  4. **Monotonicity** (LATE): $D_i^1 \geq D_i^0 \; \forall \; i$ (no defiers).
  5. **Exclusion restriction** (LATE):
     - $Yi(1,d) = Yi(0,d)$ for d = 0,1.

Estimation

- For ITT, just do what you do with a randomized experiment:
  - Difference-in-means
  - Regression with covariates

- For LATE, two common estimators exist: IV/plug-in (Wald) and 2SLS:
  1. **IV/plug-in** $\widehat{LATE} = \frac{\widehat{Cov}(Y_i, Z_i)}{\widehat{Cov}(D_i, Z_i)}$

     The IV/plug-in estimator can also be expressed as:

$$\frac{\widehat{Cov}(Y_i, Z_i)}{\widehat{Cov}(D_i, Z_i)} = \frac{\frac{\widehat{Cov}(Y_i, Z_i)}{\hat{V}(Z_i)}}{\frac{\widehat{Cov}(D_i, Z_i)}{\hat{V}(Z_i)}} = \frac{\text{Reduced Form (RF)}}{\text{1st Stage (1st)}}$$

     Reduced Form: $Y_i = \pi_0 + \pi_1 Z + \nu_i$, First Stage: $D_i = \gamma_0 + \gamma_1 Z_i + \omega_i$
     Second Stage $Y_i = \alpha_0 + \beta_1 \hat{D} + \epsilon_i$

     $\longrightarrow$ the expression becomes 0 if RF goes to 0.

     $\longrightarrow$ "there is no IV result if there is no reduced form."

     $\longrightarrow$ try a reduced-form regression before you do anything fancy.

Estimation

2. **Two Stage Least Squares (2SLS/TSLS) Estimator:**
    - Stage 1: Regress D_i on Z_i (and X_i), obtain $\hat{D}_i$.
    - Stage 2: Regress Y_i on $\hat{D}_i$. (and X_i) (don't drop the intercept!)

    - When n is small (and/or Z is weak), Problems with IV design $\longrightarrow$ try randomization inference
    - Don't do 2SLS "manually" (see next slide).
    - 2SLS variance-covariance matrix can be made robust to heteroskedasticity and clustering using the sandwich approach.

Instrumental Variables Regression in R and Python

In R:

```
# Load necessary package
library(AER)

# Running IV regression
ivmodel <- ivreg(formula = log(packs) ~ log(rprice) + log(income) |
log(income) + tax, data=CigarettesSW)
summary(ivmodel)
```

In Python:

```
# Import necessary libraries
from linearmodels.iv import IV2SLS

# Define variables
dependent = df['y']
exog = df[['const', 'x1']]
endog = df['x2']
instruments = df['z']

# Running IV regression
ivmodel = IV2SLS(dependent, exog, endog, instruments).fit()
print(ivmodel.summary())
```

Some interesting (optional) equivalences

- When there is one-sided non-compliance: LATE = ATT.
- When there are multiple instruments: 2SLS = Generalized Least Squares/GLS
- When first-stage is irrelevant: 2SLS = OLS
- When there are heterogeneous treatment effects: 2SLS/IV = weighted averages of individual LATEs

.	Group/Occupation	None	One	Between 2 and
1	Manager or Chairman	1872 (64.2%)	426 (14.6%)	389 (13.3%)
2	Street Vendor	1401 (48.0%)	390 (13.4%)	685 (23.5%)
3	Secretary	1276 (43.7%)	530 (18.2%)	722 (24.8%)
4	Car Mechanic	1018 (34.9%)	730 (25.0%)	861 (29.5%)
5	Store Clerk	957 (32.8%)	549 (18.8%)	882 (30.2%)
6	Lawyer	1628 (55.8%)	598 (20.5%)	503 (17.2%)
7	Office Cleaner	1707 (58.5%)	391 (13.4%)	568 (19.5%)
8	Doctor	1500 (51.4%)	521 (17.9%)	597 (20.5%)
9	Kindergarten Teacher	1222 (41.9%)	656 (22.5%)	720 (24.7%)
10	Taxi Driver	1199 (41.1%)	568 (19.5%)	741 (25.4%)
11	Waiter	2090 (71.6%)	293 (10.0%)	366 (12.5%)
12	Accountant	1496 (51.3%)	646 (22.1%)	592 (20.3%)
13	University Professor	1839 (63.0%)	383 (13.1%)	397 (13.6%)
14	Catholic Priest	1906 (65.1%)	655 (22.4%)	279 (9.5%)
15	Mapuche	1565 (53.6%)	412 (14.1%)	512 (17.5%)
16	Member of the UDI	2427 (84.4%)	160 (5.6%)	162 (5.6%)
17	Immigrant	2035 (69.6%)	291 (10.0%)	371 (12.7%)
18	Member of the PC	2316 (80.6%)	182 (6.3%)	204 (7.1%)