Topics in Applied Econometrics

Regression, IV in the potential outcomes framework

Attila Gyetvai

Duke Economics
Summer 2020

How to tackle an empirical project

1. What causal effects are we interested in?
2. What ideal experiment would capture this effect?
3. What is our identification strategy?
4. What is our mode of statistical inference?

Regression primer

- Linear relationship between outcome and confounding factors

$$Y_i = \alpha + \beta X_i + u_i$$

- "Best" linear relationship fits the data best
- Ordinary least squares (OLS)
- Assumptions
  1. Correct specification: $\mathbb{E}(u_i \mid X_i) = 0$
  2. Random sample: (X_i, Y_i) are i.i.d.
  3. No extreme outliers: $0 < \mathbb{E}(X_i^4) < +\infty$, $0 < \mathbb{E}(u_i^4) < +\infty$
- Under these assumptions, OLS estimates are unbiased, consistent, and asymptotically normally distributed

Regression and potential outcomes

- Treatment status as a confounder
- No compliance issues

$$
\begin{aligned}
Y_i &= \alpha + \beta D_i + u_i \\
&= \begin{cases} \alpha + u_i & \text{if } D_i = 0 \\ \alpha + \beta + u_i & \text{if } D_i = 1 \end{cases}
\end{aligned}
$$

Therefore

$$\mathbb{E}(Y_i \mid D_i = 1) = \alpha + \beta \qquad \mathbb{E}(Y_i \mid D_i = 0) = \alpha$$

- If treatment status is random, ATE is β

Noncompliance

- If treatment status is not random but assignment is, we still can identify LATE
- Recall: $LATE = \mathbb{E}(Y_{1i} - Y_{0i} \,|\, D_{1i} = 1, D_{0i} = 0$
- Let's look at compliance types
  - π_C compliers (C): $D_{1i} = 1, D_{0i} = 0$
  - π_A always-takers (A): $D_{1i} = 1, D_{0i} = 1$
  - π_N never-takers(N): $D_{1i} = 0, D_{0i} = 0$
  - No defiers
- Consider the regression

$$Y_i = \alpha + \beta Z_i + u_i \quad \rightsquigarrow \quad \beta = \mathbb{E}(Y_i \,|\, Z_i = 1) - \mathbb{E}(Y_i \,|\, Z_i = 0)$$

Noncompliance (cont'd)

- Consider the regression

$$
\begin{aligned}
Y_i = \alpha + \beta Z_i + u_i \quad \rightsquigarrow \quad \beta &= \mathbb{E}(Y_i \,|\, Z_i = 1) - \mathbb{E}(Y_i \,|\, Z_i = 0) \\
\mathbb{E}(Y_i \,|\, Z_i = 1) &= \Pr(C) \cdot \mathbb{E}(Y_i \,|\, Z_i = 1, C) \\
&\quad + \Pr(A) \cdot \mathbb{E}(Y_i \,|\, Z_i = 1, A) \\
&\quad\quad + \Pr(N) \cdot \mathbb{E}(Y_i \,|\, Z_i = 1, N) \\
&= \pi_C \, \mathbb{E}(Y_{1i} \,|\, C) + \pi_A \, \mathbb{E}(Y_{1i} \,|\, A) + \pi_N \, \mathbb{E}(Y_{0i} \,|\, N) \\
\mathbb{E}(Y_i \,|\, Z_i = 0) &= \Pr(C) \cdot \mathbb{E}(Y_i \,|\, Z_i = 0, C) \\
&\quad + \Pr(A) \cdot \mathbb{E}(Y_i \,|\, Z_i = 0, A) \\
&\quad\quad + \Pr(N) \cdot \mathbb{E}(Y_i \,|\, Z_i = 0, N) \\
&= \pi_C \, \mathbb{E}(Y_{0i} \,|\, C) + \pi_A \, \mathbb{E}(Y_{1i} \,|\, A) + \pi_N \, \mathbb{E}(Y_{0i} \,|\, N) \\
\beta &= \pi_C \, \mathbb{E}(Y_{1i} - Y_{0i} \,|\, C)
\end{aligned}
$$

Noncompliance (cont'd)

- How can we get π_C?
- Consider the regression

$$D_i = \gamma + \delta Z_i + u_i \quad \rightsquigarrow \quad \delta = \mathbb{E}(D_i \,|\, Z_i = 1) - \mathbb{E}(D_i \,|\, Z_i = 0) = \pi_C$$

- Therefore we get LATE as

$$LATE = \mathbb{E}(Y_{1i} - Y_{0i} \,|\, C) = \frac{\beta}{\delta} = \frac{\mathbb{E}(Y_i \,|\, Z_i = 1) - \mathbb{E}(Y_i \,|\, Z_i = 0)}{\mathbb{E}(D_i \,|\, Z_i = 1) - \mathbb{E}(D_i \,|\, Z_i = 0)}$$

- This is the Wald estimator, i.e., IV with binary instrument

LATE as IV

- IV: instrumental variables regression
- $Y_i = \alpha + \beta D_i + u_i$ with endogenous regressor D ($\mathbb{E}(u_i \mid D_i) \neq 0$)
- $D_i = \gamma + \delta Z_i + v_i$ with exogenous instrument Z ($\mathbb{E}(v_i \mid Z_i) = 0$)
- Two-stage estimation (2SLS)
  1. Regress D on Z, obtain predicted $\hat{D}$
  2. Regress Y on $\hat{D}$
- IV identifies LATE

Regression, IV in practice

- Canned procedures in statistical software
- Stata:
  - `reg Y D, r`
  - `ivregress 2sls Y (D=Z), vce(r)`
- R:
  - `lm(Y ~ D, data)` (plus extra stuff for robust errors)
  - `ivreg(Y ~ D | Z, data)`