17.802 – Recitation 8
Instrumental Variables

Yiqing Xu

MIT

April 18, 2014
1 2SLS Calculators

2 The LATE Framework
2SLS Lingo

- **Y** - dependent variables: the outcome variables
- **S** - variables of interest: the endogenous variables to be instrumented
- **Z** - instruments: exogenous
- **X** - covariates: exogenous

In any IV analysis, variables are either dependent, (other) endogenous variables, instruments, or covariates.
\[Y_i = \alpha + \rho S_i + A_i' \gamma + \nu_i\]
\[= \alpha + \rho S_i + \eta_i\]

- Now we have an instrument \(Z_i\) for \(S_i\), and \(\text{Cov}(\eta_i, Z_i) = 0\)
- **Exclusion restriction** is implicitly assumed (no \(Z_i\) in the formula)
- **IV estimator**

\[
\rho = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(S_i, Z_i)}
\]
Many Faces of 2SLS

\[ Y_i = \alpha + \rho S_i + X_i' \pi + (A_i' \gamma + \upsilon_i) \]

- **2SLS estimator**
  1. Regress \( S_i \) on \( Z_i \) and \( X_i \); obtain \( \hat{S}_i \)
  2. Regress \( Y_i \) on \( \hat{S}_i \) and \( X_i \), obtain \( \rho \)

- **Indirect Least Squares (ILS) estimator**

  \[
  \begin{align*}
  Y_i &= X_i' \pi_{10} + \pi_{11} Z_i + \zeta_{1i} \\
  S_i &= X_i' \pi_{20} + \pi_{21} Z_i + \zeta_{2i} \\
  \hat{\rho} &= \frac{\hat{\pi}_{21}}{\hat{\pi}_{21}}
  \end{align*}
  \]

- **IV estimator (\( \hat{S}_i^* \) as the instrument)**

  \[
  \rho = \frac{\text{Cov}(Y_i, \hat{S}_i^*)}{\text{Cov}(S_i, \hat{S}_i^*)} = \frac{\text{Cov}(Y_i, \tilde{Z}_i)}{\text{Cov}(S_i, \tilde{Z}_i)} = \frac{\pi_{21}}{\pi_{21}}
  \]

\( \hat{S}_i^* \) residual from Reg: \( \hat{S}_i \) on \( X_i \); \( \tilde{Z}_i \) residual from Reg: \( Z_i \) on \( X_i \)
Why one’s birthday is connected with his/her schooling, and future earnings?

- Age at school entry: kids born early in a calendar year are typically older when entering school
- Compulsory schooling laws
A. Average Education by Quarter of Birth (first stage)
B. Average Weekly Wage by Quarter of Birth (reduced form)
### Table 4.1.1
2SLS estimates of the economic returns to schooling

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>2SLS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Years of education</td>
<td>.071</td>
<td>.067</td>
</tr>
<tr>
<td></td>
<td>(.0004)</td>
<td>(.0004)</td>
</tr>
<tr>
<td><strong>Exogenous Covariates</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (in quarters)</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Age (in quarters) squared</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 year-of-birth dummies</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>50 state-of-birth dummies</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Instruments</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dummy for QOB = 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dummy for QOB = 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dummy for QOB = 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QOB dummies interacted with year-of-birth dummies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(30 instruments total)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:** The table reports OLS and 2SLS estimates of the returns to schooling using the Angrist and Krueger (1991) 1980 census sample. This sample includes native-born men, born 1930–39, with positive earnings and nonallocated values for key variables. The sample size is 329,509. Robust standard errors are reported in parentheses. QOB denotes quarter of birth.
Other Stuff You Should Know

- **Bias of 2SLS**
  - 2SLS estimates are consistent for causal effects but biased towards OLS estimates
  - The bias comes from weak instruments. The bigger the first-stage F, the smaller the bias
  - You don’t have much of the problem in the just-identified case
  - The reduced form is unbiased. Always check it first.
  - Limited maximum likelihood (LIML) is median-unbiased

- **Forbidden regressions**
  - Non-linear first stage
    \[ Y = X'\pi + \rho\hat{S} + \zeta + \rho(S - \hat{S}) \]
  - Non-linear second stage
1 2SLS Calculators

2 The LATE Framework
The LATE Framework

\[ Y_i = Y_{0i} + D_i(Y_{1i} - Y_{0i}) = \alpha_i + \rho_i D_i + \eta_i \]

- The Wald estimator
  \[ \rho = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(D_i, Z_i)} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} = E[Y_{1i} - Y_{0i}|D_{1i} > D_{0i}] \]

- Assumptions
  - Independence
    (Sufficient for identifying the causal effects of the instrument)
  - Exclusion
    (No direct effect; should be more controversial)
  - Monotonicity
    (Not always true)
The Story of Draft Lottery

Lottery number (RSN) $\rightarrow$ Veteran status $\rightarrow$ Earnings

- Angrist and Krueger (1990)
- Angrist and Krueger (1999)
- Angrist and Chen (2007)
- Angrist, Chen, and Song (2011)
### IV Estimates of the Effects of Military Service on the Earnings of White Men born in 1950

<table>
<thead>
<tr>
<th>Earnings year</th>
<th>Earnings (Mean)</th>
<th>Eligibility Effect (Mean)</th>
<th>Veteran Status (Mean)</th>
<th>Eligibility Effect (Mean)</th>
<th>Wald Estimate of Veteran Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
</tr>
<tr>
<td>1981</td>
<td>16,461</td>
<td>-435.8</td>
<td>.267</td>
<td>.159</td>
<td>-2,741</td>
</tr>
<tr>
<td></td>
<td>(210.5)</td>
<td>(0.040)</td>
<td></td>
<td></td>
<td>(1,324)</td>
</tr>
<tr>
<td>1971</td>
<td>3,338</td>
<td>-325.9</td>
<td></td>
<td></td>
<td>-2050</td>
</tr>
<tr>
<td></td>
<td>(46.6)</td>
<td></td>
<td></td>
<td></td>
<td>(293)</td>
</tr>
<tr>
<td>1969</td>
<td>2,299</td>
<td>-2.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(34.5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Adapted from Table 5 in Angrist and Krueger (1999) and author tabulations. Standard errors are shown in parentheses. Earnings data are from Social Security administrative records. Figures are in nominal dollars. Veteran status data are from the Survey of Program Participation. There are about 13,500 individuals in the sample.
Figure 1. Social Security Earnings Profiles by Draft-Eligibility Status
Figure 3. Earnings and the Probability of Veteran Status by Lottery Number
Figure 1. Draft-lottery Estimates of Vietnam-era Service Effects on ln(Earnings) for White Men Born 1950-52
Other Issues

- You cannot identify compliers, but you can count them (first stage) and characterize them (LARF)
- Two-sample IV is feasible, as long as the two samples are representative of the same population
- External validity comes from many experiments (instruments) or the same experiment in many occasions/places
In Summary

- IVs are powerful (w/o RCT, w/ non-complier)
- But they are to find
  - more and more from experiments
  - unanticipated shocks (e.g., natural disaster)
  - binding institutional rules
- The most demanding assumption is exclusion restriction
- Transparency is always the most important (show plots)
- The reduced form gives you a rough idea of whether it’s gonna work (The S.E.s are pretty much the same)
- External validity is an issue, but let’s worry about it later (counting and characterizing compliers help)