Module 6(b): Two Stage Least Squares

Fall 2021

Matthew Blackwell

Gov 2003 (Harvard)

1/ Basic two-stage least squares

TSLS

- **Two stage least squares** (TSLS) is the classical approach to IV.

TSLS

- **Two stage least squares** (TSLS) is the classical approach to IV.
- Basic idea is to assume two constant effects linear models:

$$
\begin{aligned}
Y_i &= \alpha + \tau D_i + \varepsilon_i \\
D_i &= \delta + \gamma Z_i + \eta_i
\end{aligned}
$$

TSLS

- **Two stage least squares** (TSLS) is the classical approach to IV.
- Basic idea is to assume two constant effects linear models:

$$Y_i = \alpha + \tau D_i + \varepsilon_i$$
$$D_i = \delta + \gamma Z_i + \eta_i$$

- Here the treatment D_i is **endogenous** so $\mathbb{E}[\varepsilon_i \mid D_i] \neq 0$

TSLS

- **Two stage least squares** (TSLS) is the classical approach to IV.
- Basic idea is to assume two constant effects linear models:

$$
\begin{aligned}
Y_i &= \alpha + \tau D_i + \varepsilon_i \\
D_i &= \delta + \gamma Z_i + \eta_i
\end{aligned}
$$

- Here the treatment D_i is **endogenous** so $\mathbb{E}[\varepsilon_i \mid D_i] \neq 0$
- But we have an **instrument** Z_i that is exogenous $\mathbb{E}[\varepsilon_i \mid Z_i] = 0$

TSLS

- **Two stage least squares** (TSLS) is the classical approach to IV.
- Basic idea is to assume two constant effects linear models:

$$Y_i = \alpha + \tau D_i + \varepsilon_i$$
$$D_i = \delta + \gamma Z_i + \eta_i$$

- Here the treatment D_i is **endogenous** so $\mathbb{E}[\varepsilon_i \mid D_i] \neq 0$
- But we have an **instrument** Z_i that is exogenous $\mathbb{E}[\varepsilon_i \mid Z_i] = 0$
    - It also is exogenous for treatment, so $\mathbb{E}[\eta_i \mid Z_i] = 0$.

TSLS

- **Two stage least squares** (TSLS) is the classical approach to IV.
- Basic idea is to assume two constant effects linear models:

$$Y_i = \alpha + \tau D_i + \varepsilon_i$$
$$D_i = \delta + \gamma Z_i + \eta_i$$

- Here the treatment D_i is **endogenous** so $\mathbb{E}[\varepsilon_i \mid D_i] \neq 0$
- But we have an **instrument** Z_i that is exogenous $\mathbb{E}[\varepsilon_i \mid Z_i] = 0$
    - It also is exogenous for treatment, so $\mathbb{E}[\eta_i \mid Z_i] = 0$.
- This implies the following CEF form for Y_i conditional on Z_i:

$$\mathbb{E}[Y_i \mid Z_i] = \alpha + \tau\mathbb{E}[D_i \mid Z_i] = \alpha + \tau \cdot (\gamma Z_i)$$

TSLS estimands

- Under the model, we have the following CEF: $\mathbb{E}[Y_i \mid Z_i] = \alpha + \tau \cdot (\gamma Z_i)$

TSLS estimands

- Under the model, we have the following CEF: $\mathbb{E}[Y_i \mid Z_i] = \alpha + \tau \cdot (\gamma Z_i)$
  - $\rightsquigarrow$ a regression of Y_i on γZ_i would have τ as the slope.

TSLS estimands

- Under the model, we have the following CEF: $\mathbb{E}[Y_i \mid Z_i] = \alpha + \tau \cdot (\gamma Z_i)$
    - $\rightsquigarrow$ a regression of Y_i on γZ_i would have τ as the slope.
- If the CEF is linear, we have this simple relationship slopes:

$$\mathbb{E}[D_i \mid Z_i] = \delta + \gamma Z_i \quad \rightsquigarrow \quad \gamma = \frac{\text{cov}(D_i, Z_i)}{\mathbb{V}[Z_i]}$$

TSLS estimands

- Under the model, we have the following CEF: $\mathbb{E}[Y_i \mid Z_i] = \alpha + \tau \cdot (\gamma Z_i)$
  - $\rightsquigarrow$ a regression of Y_i on γZ_i would have τ as the slope.
- If the CEF is linear, we have this simple relationship slopes:

$$\mathbb{E}[D_i \mid Z_i] = \delta + \gamma Z_i \quad \rightsquigarrow \quad \gamma = \frac{\text{cov}(D_i, Z_i)}{\mathbb{V}[Z_i]}$$

- Applying this to above CEF we have:

$$\tau = \frac{\text{cov}(Y_i, \gamma Z_i)}{\mathbb{V}[\gamma Z_i]} = \frac{\text{cov}(Y_i, Z_i)}{\gamma \mathbb{V}[Z_i]} = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(D_i, Z_i)}$$

TSLS estimands

- Under the model, we have the following CEF: $\mathbb{E}[Y_i \mid Z_i] = \alpha + \tau \cdot (\gamma Z_i)$
  - $\rightsquigarrow$ a regression of Y_i on γZ_i would have τ as the slope.
- If the CEF is linear, we have this simple relationship slopes:

$$\mathbb{E}[D_i \mid Z_i] = \delta + \gamma Z_i \quad \rightsquigarrow \quad \gamma = \frac{\text{cov}(D_i, Z_i)}{\mathbb{V}[Z_i]}$$

- Applying this to above CEF we have:

$$\tau = \frac{\text{cov}(Y_i, \gamma Z_i)}{\mathbb{V}[\gamma Z_i]} = \frac{\text{cov}(Y_i, Z_i)}{\gamma \mathbb{V}[Z_i]} = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(D_i, Z_i)}$$

- TSLS estimator:

TSLS estimands

- Under the model, we have the following CEF: $\mathbb{E}[Y_i \mid Z_i] = \alpha + \tau \cdot (\gamma Z_i)$
  - $\rightsquigarrow$ a regression of Y_i on γZ_i would have τ as the slope.
- If the CEF is linear, we have this simple relationship slopes:

$$\mathbb{E}[D_i \mid Z_i] = \delta + \gamma Z_i \quad \rightsquigarrow \quad \gamma = \frac{\text{cov}(D_i, Z_i)}{\mathbb{V}[Z_i]}$$

- Applying this to above CEF we have:

$$\tau = \frac{\text{cov}(Y_i, \gamma Z_i)}{\mathbb{V}[\gamma Z_i]} = \frac{\text{cov}(Y_i, Z_i)}{\gamma \mathbb{V}[Z_i]} = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(D_i, Z_i)}$$

- TSLS estimator:
  - Estimate $\widehat{\gamma}$ from regression of treatment D_i on instrument Z_i

TSLS estimands

- Under the model, we have the following CEF: $\mathbb{E}[Y_i \mid Z_i] = \alpha + \tau \cdot (\gamma Z_i)$
  - $\rightsquigarrow$ a regression of Y_i on γZ_i would have τ as the slope.
- If the CEF is linear, we have this simple relationship slopes:

$$\mathbb{E}[D_i \mid Z_i] = \delta + \gamma Z_i \quad \rightsquigarrow \quad \gamma = \frac{\text{cov}(D_i, Z_i)}{\mathbb{V}[Z_i]}$$

- Applying this to above CEF we have:

$$\tau = \frac{\text{cov}(Y_i, \gamma Z_i)}{\mathbb{V}[\gamma Z_i]} = \frac{\text{cov}(Y_i, Z_i)}{\gamma \mathbb{V}[Z_i]} = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(D_i, Z_i)}$$

- TSLS estimator:
  - Estimate $\widehat{\gamma}$ from regression of treatment D_i on instrument Z_i
  - Estimate $\widehat{\tau}_{2SLS}$ as the slope of a regression of Y_i on $\widehat{\gamma} Z_i$

TSLS estimands

- Under the model, we have the following CEF: $\mathbb{E}[Y_i \mid Z_i] = \alpha + \tau \cdot (\gamma Z_i)$
  - $\rightsquigarrow$ a regression of Y_i on γZ_i would have τ as the slope.
- If the CEF is linear, we have this simple relationship slopes:

$$\mathbb{E}[D_i \mid Z_i] = \delta + \gamma Z_i \quad \rightsquigarrow \quad \gamma = \frac{\text{cov}(D_i, Z_i)}{\mathbb{V}[Z_i]}$$

- Applying this to above CEF we have:

$$\tau = \frac{\text{cov}(Y_i, \gamma Z_i)}{\mathbb{V}[\gamma Z_i]} = \frac{\text{cov}(Y_i, Z_i)}{\gamma \mathbb{V}[Z_i]} = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(D_i, Z_i)}$$

- TSLS estimator:
  - Estimate $\widehat{\gamma}$ from regression of treatment D_i on instrument Z_i
  - Estimate $\widehat{\tau}_{2SLS}$ as the slope of a regression of Y_i on $\widehat{\gamma} Z_i$
  - Under this model, $\widehat{\tau}_{2SLS} \xrightarrow{p} \tau$ (but don't use SEs from second stage)

Binary treatment and instrument

- Under binary treatment/instrument, TSLS estimand is the LATE:

$$\tau = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(D_i, Z_i)} = \frac{\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0]}{\mathbb{E}[D_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]} = \frac{\text{ITT}_Y}{\text{ITT}_D} = \tau_{\text{LATE}}$$

Binary treatment and instrument

- Under binary treatment/instrument, TSLS estimand is the LATE:

$$\tau = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(D_i, Z_i)} = \frac{\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0]}{\mathbb{E}[D_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]} = \frac{\text{ITT}_Y}{\text{ITT}_D} = \tau_{\text{LATE}}$$

- And the TSLS estimator is the Wald estimator:

$$\widehat{\tau}_{2SLS} = \frac{\widehat{\text{cov}}(Y_i, Z_i)}{\widehat{\text{cov}}(D_i, Z_i)} = \frac{\overline{Y}_1 - \overline{Y}_0}{\overline{D}_1 - \overline{D}_0} = \frac{\widehat{\text{ITT}}_Y}{\widehat{\text{ITT}}_D} = \widehat{\tau}_{iv}$$

Binary treatment and instrument

- Under binary treatment/instrument, TSLS estimand is the LATE:

$$\tau = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(D_i, Z_i)} = \frac{\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0]}{\mathbb{E}[D_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]} = \frac{\text{ITT}_Y}{\text{ITT}_D} = \tau_{\text{LATE}}$$

- And the TSLS estimator is the Wald estimator:

$$\widehat{\tau}_{2SLS} = \frac{\widehat{\text{cov}}(Y_i, Z_i)}{\widehat{\text{cov}}(D_i, Z_i)} = \frac{\overline{Y}_1 - \overline{Y}_0}{\overline{D}_1 - \overline{D}_0} = \frac{\widehat{\text{ITT}}_Y}{\widehat{\text{ITT}}_D} = \widehat{\tau}_{iv}$$

- $\rightsquigarrow$ constant effects model not required for TSLS in this setting.

Binary treatment and instrument

- Under binary treatment/instrument, TSLS estimand is the LATE:

$$\tau = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(D_i, Z_i)} = \frac{\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0]}{\mathbb{E}[D_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]} = \frac{\text{ITT}_Y}{\text{ITT}_D} = \tau_{\text{LATE}}$$

- And the TSLS estimator is the Wald estimator:

$$\widehat{\tau}_{2SLS} = \frac{\widehat{\text{cov}}(Y_i, Z_i)}{\widehat{\text{cov}}(D_i, Z_i)} = \frac{\overline{Y}_1 - \overline{Y}_0}{\overline{D}_1 - \overline{D}_0} = \frac{\widehat{\text{ITT}}_Y}{\widehat{\text{ITT}}_D} = \widehat{\tau}_{iv}$$

- $\rightsquigarrow$ constant effects model not required for TSLS in this setting.
- But we need constant effects when we add covariates:

$$Y_i = \alpha + \tau D_i + \mathbf{X}_i'\beta_y + \varepsilon_i$$
$$D_i = \delta + \gamma Z_i + \mathbf{X}_i'\beta_d + \eta_i$$

Binary treatment and instrument

- Under binary treatment/instrument, TSLS estimand is the LATE:

$$\tau = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(D_i, Z_i)} = \frac{\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0]}{\mathbb{E}[D_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]} = \frac{\text{ITT}_Y}{\text{ITT}_D} = \tau_{\text{LATE}}$$

- And the TSLS estimator is the Wald estimator:

$$\widehat{\tau}_{2SLS} = \frac{\widehat{\text{cov}}(Y_i, Z_i)}{\widehat{\text{cov}}(D_i, Z_i)} = \frac{\overline{Y}_1 - \overline{Y}_0}{\overline{D}_1 - \overline{D}_0} = \frac{\widehat{\text{ITT}}_Y}{\widehat{\text{ITT}}_D} = \widehat{\tau}_{iv}$$

- $\rightsquigarrow$ constant effects model not required for TSLS in this setting.
- But we need constant effects when we add covariates:

$$Y_i = \alpha + \tau D_i + \mathbf{X}_i'\beta_y + \varepsilon_i$$
$$D_i = \delta + \gamma Z_i + \mathbf{X}_i'\beta_d + \eta_i$$

  - Otherwise, τ is an odd weighted function of causal effects and $\tau \neq \tau_{\text{LATE}}$

Weak instruments

- IV is unstable when instrument weakly affects treatment $\text{cov}(D_i, Z_i) \approx 0$.

Weak instruments

- IV is unstable when instrument weakly affects treatment $\text{cov}(D_i, Z_i) \approx 0$.
- **Example** completely irrelevant instrument:

$$\begin{aligned} Y_i &= \tau D_i + \varepsilon_i & \mathbb{E}[\varepsilon_i \mid D_i] &\neq 0 \\ D_i &= 0 \times Z_i + \eta_i & \mathbb{E}[\varepsilon_i \mid Z_i] = \mathbb{E}[\eta_i \mid Z_i] &= 0 \end{aligned}$$

Weak instruments

- IV is unstable when instrument weakly affects treatment $\text{cov}(D_i, Z_i) \approx 0$.
- **Example** completely irrelevant instrument:

$$\begin{aligned} Y_i &= \tau D_i + \varepsilon_i & \mathbb{E}[\varepsilon_i \mid D_i] &\neq 0 \\ D_i &= 0 \times Z_i + \eta_i & \mathbb{E}[\varepsilon_i \mid Z_i] = \mathbb{E}[\eta_i \mid Z_i] &= 0 \end{aligned}$$

  - Note that we only assume mean independence, so $\text{cov}(D_i, Z_i)$ could be nonzero.

Weak instruments

- IV is unstable when instrument weakly affects treatment $\text{cov}(D_i, Z_i) \approx 0$.
- **Example** completely irrelevant instrument:

$$
\begin{aligned}
Y_i &= \tau D_i + \varepsilon_i & \mathbb{E}[\varepsilon_i \mid D_i] \neq 0 \\
D_i &= 0 \times Z_i + \eta_i & \mathbb{E}[\varepsilon_i \mid Z_i] = \mathbb{E}[\eta_i \mid Z_i] = 0
\end{aligned}
$$

  - Note that we only assume mean independence, so $\text{cov}(D_i, Z_i)$ could be nonzero.
- We can write the bias of the Wald estimator as:

$$
\widehat{\tau}_{iv} - \tau = \frac{\widehat{\text{cov}}(\tau D_i + \varepsilon_i, Z_i)}{\widehat{\text{cov}}(D_i, Z_i)} - \tau
$$

Weak instruments

- IV is unstable when instrument weakly affects treatment $\text{cov}(D_i, Z_i) \approx 0$.
- **Example** completely irrelevant instrument:

$$\begin{aligned} Y_i &= \tau D_i + \varepsilon_i & \mathbb{E}[\varepsilon_i \mid D_i] \neq 0 \\ D_i &= 0 \times Z_i + \eta_i & \mathbb{E}[\varepsilon_i \mid Z_i] = \mathbb{E}[\eta_i \mid Z_i] = 0 \end{aligned}$$

  - Note that we only assume mean independence, so $\text{cov}(D_i, Z_i)$ could be nonzero.
- We can write the bias of the Wald estimator as:

$$\widehat{\tau}_{iv} - \tau = \frac{\widehat{\text{cov}}(\tau D_i + \varepsilon_i, Z_i)}{\widehat{\text{cov}}(D_i, Z_i)} - \tau = \frac{\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i Z_i}{\frac{1}{n}\sum_{i=1}^{n} \eta_i Z_i}$$

Weak instruments

- IV is unstable when instrument weakly affects treatment $\text{cov}(D_i, Z_i) \approx 0$.
- **Example** completely irrelevant instrument:

$$\begin{aligned} Y_i &= \tau D_i + \varepsilon_i & \mathbb{E}[\varepsilon_i \mid D_i] &\neq 0 \\ D_i &= 0 \times Z_i + \eta_i & \mathbb{E}[\varepsilon_i \mid Z_i] = \mathbb{E}[\eta_i \mid Z_i] &= 0 \end{aligned}$$

  - Note that we only assume mean independence, so $\text{cov}(D_i, Z_i)$ could be nonzero.

- We can write the bias of the Wald estimator as:

$$\widehat{\tau}_{iv} - \tau = \frac{\widehat{\text{cov}}(\tau D_i + \varepsilon_i, Z_i)}{\widehat{\text{cov}}(D_i, Z_i)} - \tau = \frac{\frac{1}{n}\sum_{i=1}^n \varepsilon_i Z_i}{\frac{1}{n}\sum_{i=1}^n \eta_i Z_i} \xrightarrow{d} \underbrace{\frac{\text{cov}(\varepsilon_i, \eta_i)}{\mathbb{V}[\varepsilon_i]}}_{\text{bias}} + \underbrace{W_i}_{\text{Cauchy r.v.}}$$

Weak instruments

- IV is unstable when instrument weakly affects treatment $\text{cov}(D_i, Z_i) \approx 0$.
- **Example** completely irrelevant instrument:

$$\begin{aligned} Y_i &= \tau D_i + \varepsilon_i & \mathbb{E}[\varepsilon_i \mid D_i] \neq 0 \\ D_i &= 0 \times Z_i + \eta_i & \mathbb{E}[\varepsilon_i \mid Z_i] = \mathbb{E}[\eta_i \mid Z_i] = 0 \end{aligned}$$

  - Note that we only assume mean independence, so $\text{cov}(D_i, Z_i)$ could be nonzero.
- We can write the bias of the Wald estimator as:

$$\widehat{\tau}_{iv} - \tau = \frac{\widehat{\text{cov}}(\tau D_i + \varepsilon_i, Z_i)}{\widehat{\text{cov}}(D_i, Z_i)} - \tau = \frac{\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i Z_i}{\frac{1}{n}\sum_{i=1}^{n} \eta_i Z_i} \xrightarrow{d} \underbrace{\frac{\text{cov}(\varepsilon_i, \eta_i)}{\mathbb{V}[\varepsilon_i]}}_{\text{bias}} + \underbrace{W_i}_{\text{Cauchy r.v.}}$$

- Inconsistent and asymptotically heavy tails (bc of Cauchy)

Weak instruments

- IV is unstable when instrument weakly affects treatment $\text{cov}(D_i, Z_i) \approx 0$.
- **Example** completely irrelevant instrument:

$$\begin{aligned} Y_i &= \tau D_i + \varepsilon_i & \mathbb{E}[\varepsilon_i \mid D_i] \neq 0 \\ D_i &= 0 \times Z_i + \eta_i & \mathbb{E}[\varepsilon_i \mid Z_i] = \mathbb{E}[\eta_i \mid Z_i] = 0 \end{aligned}$$

  - Note that we only assume mean independence, so $\text{cov}(D_i, Z_i)$ could be nonzero.
- We can write the bias of the Wald estimator as:

$$\widehat{\tau}_{iv} - \tau = \frac{\widehat{\text{cov}}(\tau D_i + \varepsilon_i, Z_i)}{\widehat{\text{cov}}(D_i, Z_i)} - \tau = \frac{\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i Z_i}{\frac{1}{n}\sum_{i=1}^{n} \eta_i Z_i} \xrightarrow{d} \underbrace{\frac{\text{cov}(\varepsilon_i, \eta_i)}{\mathbb{V}[\varepsilon_i]}}_{\text{bias}} + \underbrace{W_i}_{\text{Cauchy r.v.}}$$

- Inconsistent and asymptotically heavy tails (bc of Cauchy)
  - When $Z \rightarrow D$ effect is small but non-zero we see similar behavior.

What to do about weak instruments?

- Detecting weak instruments:

What to do about weak instruments?

- Detecting weak instruments:
    - F-test on instruments (excluded from second stage): $H_0 : \gamma = 0$.

What to do about weak instruments?

- Detecting weak instruments:
  - F-test on instruments (excluded from second stage): $H_0 : \gamma = 0$.
  - Rule of thumb: bias is small when F-stat ≥ 10 (Stock & Yogo, 2005)

What to do about weak instruments?

- Detecting weak instruments:
  - F-test on instruments (excluded from second stage): $H_0 : \gamma = 0$.
  - Rule of thumb: bias is small when F-stat ≥ 10 (Stock & Yogo, 2005)
  - Correct coverage may require cutoff $F \geq 104.7$ (Lee et al, 2020)

What to do about weak instruments?

- Detecting weak instruments:
  - F-test on instruments (excluded from second stage): $H_0 : \gamma = 0$.
  - Rule of thumb: bias is small when F-stat ≥ 10 (Stock & Yogo, 2005)
  - Correct coverage may require cutoff $F \geq 104.7$ (Lee et al, 2020)
  - The latter is a worst-case, typical data maybe ok with 10 cutoff

What to do about weak instruments?

- Detecting weak instruments:
  - F-test on instruments (excluded from second stage): $H_0 : \gamma = 0$.
  - Rule of thumb: bias is small when F-stat ≥ 10 (Stock & Yogo, 2005)
  - Correct coverage may require cutoff $F \geq 104.7$ (Lee et al, 2020)
  - The latter is a worst-case, typical data maybe ok with 10 cutoff
- Anderson-Rubin (1949) test (simplified setting, binary Z/D)

What to do about weak instruments?

- Detecting weak instruments:
  - F-test on instruments (excluded from second stage): $H_0 : \gamma = 0$.
  - Rule of thumb: bias is small when F-stat ≥ 10 (Stock & Yogo, 2005)
  - Correct coverage may require cutoff $F \geq 104.7$ (Lee et al, 2020)
  - The latter is a worst-case, typical data maybe ok with 10 cutoff
- Anderson-Rubin (1949) test (simplified setting, binary Z/D)
  - $H_0 : \tau = \tau_0$ equivalent to $H_0 : \text{ITT}_Y - \text{ITT}_D \cdot \tau_0 = 0$

What to do about weak instruments?

- Detecting weak instruments:
  - F-test on instruments (excluded from second stage): $H_0 : \gamma = 0$.
  - Rule of thumb: bias is small when F-stat ≥ 10 (Stock & Yogo, 2005)
  - Correct coverage may require cutoff $F \geq 104.7$ (Lee et al, 2020)
  - The latter is a worst-case, typical data maybe ok with 10 cutoff
- Anderson-Rubin (1949) test (simplified setting, binary Z/D)
  - $H_0 : \tau = \tau_0$ equivalent to $H_0 : \text{ITT}_Y - \text{ITT}_D \cdot \tau_0 = 0$
  - Under the null, asymptotically we have

$$g(\tau_0) = \widehat{\text{ITT}}_Y - \widehat{\text{ITT}}_D \tau_0 \sim N(0, \Omega(\tau_0))$$
$$\Omega(\tau_0) = \mathbb{V}[\widehat{\text{ITT}}_Y] + \tau_0^2 \mathbb{V}[\widehat{\text{ITT}}_D] - 2\tau_0 \text{cov}(\widehat{\text{ITT}}_Y, \widehat{\text{ITT}}_D)$$

What to do about weak instruments?

- Detecting weak instruments:
    - F-test on instruments (excluded from second stage): $H_0 : \gamma = 0$.
    - Rule of thumb: bias is small when F-stat ≥ 10 (Stock & Yogo, 2005)
    - Correct coverage may require cutoff $F \geq 104.7$ (Lee et al, 2020)
    - The latter is a worst-case, typical data maybe ok with 10 cutoff
- Anderson-Rubin (1949) test (simplified setting, binary Z/D)
    - $H_0 : \tau = \tau_0$ equivalent to $H_0 : \text{ITT}_Y - \text{ITT}_D \cdot \tau_0 = 0$
    - Under the null, asymptotically we have

$$g(\tau_0) = \widehat{\text{ITT}}_Y - \widehat{\text{ITT}}_D \tau_0 \sim N(0, \Omega(\tau_0))$$
$$\Omega(\tau_0) = \mathbb{V}[\widehat{\text{ITT}}_Y] + \tau_0^2 \mathbb{V}[\widehat{\text{ITT}}_D] - 2\tau_0 \text{cov}(\widehat{\text{ITT}}_Y, \widehat{\text{ITT}}_D)$$

    - AR test statistic: $g(\tau_0)^2 / \Omega(\tau_0) \sim \chi^2$ no matter first-stage effect.

What to do about weak instruments?

- Detecting weak instruments:
  - F-test on instruments (excluded from second stage): $H_0 : \gamma = 0$.
  - Rule of thumb: bias is small when F-stat ≥ 10 (Stock & Yogo, 2005)
  - Correct coverage may require cutoff $F \geq 104.7$ (Lee et al, 2020)
  - The latter is a worst-case, typical data maybe ok with 10 cutoff
- Anderson-Rubin (1949) test (simplified setting, binary Z/D)
  - $H_0 : \tau = \tau_0$ equivalent to $H_0 : \text{ITT}_Y - \text{ITT}_D \cdot \tau_0 = 0$
  - Under the null, asymptotically we have

$$g(\tau_0) = \widehat{\text{ITT}}_Y - \widehat{\text{ITT}}_D \tau_0 \sim N(0, \Omega(\tau_0))$$
$$\Omega(\tau_0) = \mathbb{V}[\widehat{\text{ITT}}_Y] + \tau_0^2 \mathbb{V}[\widehat{\text{ITT}}_D] - 2\tau_0 \text{cov}(\widehat{\text{ITT}}_Y, \widehat{\text{ITT}}_D)$$

  - AR test statistic: $g(\tau_0)^2 / \Omega(\tau_0) \sim \chi^2$ no matter first-stage effect.
  - Can invert (analytically!) to get confidence intervals

Multi-valued treatments

- Generalization of these ideas:

Multi-valued treatments

- Generalization of these ideas:
    - Multi-valued treatment: $D_i \in \{0, 1, \dots, K-1\}$

Multi-valued treatments

- Generalization of these ideas:
  - Multi-valued treatment: $D_i \in \{0, 1, \dots, K-1\}$
  - Binary instrument: $Z_i \in \{0, 1\}$

Multi-valued treatments

- Generalization of these ideas:
    - Multi-valued treatment: $D_i \in \{0, 1, \dots, K-1\}$
    - Binary instrument: $Z_i \in \{0, 1\}$
- Assumptions:

Multi-valued treatments

- Generalization of these ideas:
  - Multi-valued treatment: $D_i \in \{0, 1, \dots, K-1\}$
  - Binary instrument: $Z_i \in \{0, 1\}$
- Assumptions:
  - Randomization: $[\{Y_i(d, z), \forall d, z\}, D_i(1), D_i(0)] \perp\!\!\!\perp Z_i$

Multi-valued treatments

- Generalization of these ideas:
  - Multi-valued treatment: $D_i \in \{0, 1, \dots, K-1\}$
  - Binary instrument: $Z_i \in \{0, 1\}$
- Assumptions:
  - Randomization: $[\{Y_i(d, z), \forall d, z\}, D_i(1), D_i(0)] \perp\!\!\!\perp Z_i$
  - Monotonicity: $D_i(1) \geq D_i(0)$ (instrument only increases treatment)

Multi-valued treatments

- Generalization of these ideas:
  - Multi-valued treatment: $D_i \in \{0, 1, \dots, K-1\}$
  - Binary instrument: $Z_i \in \{0, 1\}$
- Assumptions:
  - Randomization: $[\{Y_i(d, z), \forall d, z\}, D_i(1), D_i(0)] \perp\!\!\!\perp Z_i$
  - Monotonicity: $D_i(1) \geq D_i(0)$ (instrument only increases treatment)
  - Exclusion restriction: $Y_i(1, d) = Y_i(0, d)$ for all $d = 0, 1, \dots, K-1$

Multi-valued treatments

- Generalization of these ideas:
  - Multi-valued treatment: $D_i \in \{0, 1, \dots, K-1\}$
  - Binary instrument: $Z_i \in \{0, 1\}$
- Assumptions:
  - Randomization: $[\{Y_i(d,z), \forall d, z\}, D_i(1), D_i(0)] \perp\!\!\!\perp Z_i$
  - Monotonicity: $D_i(1) \geq D_i(0)$ (instrument only increases treatment)
  - Exclusion restriction: $Y_i(1,d) = Y_i(0,d)$ for all $d = 0, 1, \dots, K-1$
- Can't identify the proportion of all compliance types here.

Multi-valued treatments

- Generalization of these ideas:
  - Multi-valued treatment: $D_i \in \{0, 1, \dots, K-1\}$
  - Binary instrument: $Z_i \in \{0, 1\}$
- Assumptions:
  - Randomization: $[\{Y_i(d, z), \forall d, z\}, D_i(1), D_i(0)] \perp\!\!\!\perp Z_i$
  - Monotonicity: $D_i(1) \geq D_i(0)$ (instrument only increases treatment)
  - Exclusion restriction: $Y_i(1, d) = Y_i(0, d)$ for all $d = 0, 1, \dots, K-1$
- Can't identify the proportion of all compliance types here.
- Example: $K = 3 \rightsquigarrow 9$ principal strata

Multi-valued treatments

- Generalization of these ideas:
    - Multi-valued treatment: $D_i \in \{0, 1, \ldots, K-1\}$
    - Binary instrument: $Z_i \in \{0, 1\}$
- Assumptions:
    - Randomization: $[\{Y_i(d, z), \forall d, z\}, D_i(1), D_i(0)] \perp\!\!\!\perp Z_i$
    - Monotonicity: $D_i(1) \geq D_i(0)$ (instrument only increases treatment)
    - Exclusion restriction: $Y_i(1, d) = Y_i(0, d)$ for all $d = 0, 1, \ldots, K-1$
- Can't identify the proportion of all compliance types here.
- Example: $K = 3 \rightsquigarrow 9$ principal strata
    - Affected: $(D_i(0), D_i(1)) \in \{(0, 1), (0, 2), (1, 2)\}$

Multi-valued treatments

- Generalization of these ideas:
  - Multi-valued treatment: $D_i \in \{0, 1, \dots, K-1\}$
  - Binary instrument: $Z_i \in \{0, 1\}$
- Assumptions:
  - Randomization: $[\{Y_i(d, z), \forall d, z\}, D_i(1), D_i(0)] \perp\!\!\!\perp Z_i$
  - Monotonicity: $D_i(1) \geq D_i(0)$ (instrument only increases treatment)
  - Exclusion restriction: $Y_i(1, d) = Y_i(0, d)$ for all $d = 0, 1, \dots, K-1$
- Can't identify the proportion of all compliance types here.
- Example: $K = 3 \rightsquigarrow 9$ principal strata
  - Affected: $(D_i(0), D_i(1)) \in \{(0, 1), (0, 2), (1, 2)\}$
  - Unaffected: $(D_i(0), D_i(1)) \in \{(0, 0), (1, 1), (2, 2)\}$

Multi-valued treatments

- Generalization of these ideas:
  - Multi-valued treatment: $D_i \in \{0, 1, \dots, K-1\}$
  - Binary instrument: $Z_i \in \{0, 1\}$
- Assumptions:
  - Randomization: $[\{Y_i(d, z), \forall d, z\}, D_i(1), D_i(0)] \perp\!\!\!\perp Z_i$
  - Monotonicity: $D_i(1) \geq D_i(0)$ (instrument only increases treatment)
  - Exclusion restriction: $Y_i(1, d) = Y_i(0, d)$ for all $d = 0, 1, \dots, K-1$
- Can't identify the proportion of all compliance types here.
- Example: $K = 3 \rightsquigarrow 9$ principal strata
  - Affected: $(D_i(0), D_i(1)) \in \{(0, 1), (0, 2), (1, 2)\}$
  - Unaffected: $(D_i(0), D_i(1)) \in \{(0, 0), (1, 1), (2, 2)\}$
  - Negatively affected: $(D_i(0), D_i(1)) \in \{(1, 0), (2, 0), (2, 1)\}$

Multi-valued treatments

- Generalization of these ideas:
    - Multi-valued treatment: $D_i \in \{0, 1, \dots, K-1\}$
    - Binary instrument: $Z_i \in \{0, 1\}$
- Assumptions:
    - Randomization: $[\{Y_i(d,z), \forall d, z\}, D_i(1), D_i(0)] \perp\!\!\!\perp Z_i$
    - Monotonicity: $D_i(1) \geq D_i(0)$ (instrument only increases treatment)
    - Exclusion restriction: $Y_i(1, d) = Y_i(0, d)$ for all $d = 0, 1, \dots, K-1$
- Can't identify the proportion of all compliance types here.
- Example: $K = 3 \rightsquigarrow 9$ principal strata
    - Affected: $(D_i(0), D_i(1)) \in \{(0,1), (0,2), (1,2)\}$
    - Unaffected: $(D_i(0), D_i(1)) \in \{(0,0), (1,1), (2,2)\}$
    - Negatively affected: $(D_i(0), D_i(1)) \in \{(1,0), (2,0), (2,1)\}$
    - Last ruled out by monotonicity.

Multi-valued treatments

- Generalization of these ideas:
    - Multi-valued treatment: $D_i \in \{0, 1, \ldots, K-1\}$
    - Binary instrument: $Z_i \in \{0, 1\}$
- Assumptions:
    - Randomization: $[\{Y_i(d,z), \forall d, z\}, D_i(1), D_i(0)] \perp\!\!\!\perp Z_i$
    - Monotonicity: $D_i(1) \geq D_i(0)$ (instrument only increases treatment)
    - Exclusion restriction: $Y_i(1,d) = Y_i(0,d)$ for all $d = 0, 1, \ldots, K-1$
- Can't identify the proportion of all compliance types here.
- Example: $K = 3 \rightsquigarrow 9$ principal strata
    - Affected: $(D_i(0), D_i(1)) \in \{(0,1), (0,2), (1,2)\}$
    - Unaffected: $(D_i(0), D_i(1)) \in \{(0,0), (1,1), (2,2)\}$
    - Negatively affected: $(D_i(0), D_i(1)) \in \{(1,0), (2,0), (2,1)\}$
    - Last ruled out by monotonicity.
    - 5 unknowns and 4 knowns under monotonicity.

TSLS with multivalued treatments

- Let $C_i = jk$ be an indicator for compliance type $D_i(1) = j$ and $D_i(0) = k$.

TSLS with multivalued treatments

- Let $C_i = jk$ be an indicator for compliance type $D_i(1) = j$ and $D_i(0) = k$.
    - People that are moved from k to j by the instrument.

TSLS with multivalued treatments

- Let $C_i = jk$ be an indicator for compliance type $D_i(1) = j$ and $D_i(0) = k$.
    - People that are moved from k to j by the instrument.
    - Let $\rho_{jk} = \mathbb{P}(D_i(1) = j, D_i(0) = k)$ be the strata size.

TSLS with multivalued treatments

- Let $C_i = jk$ be an indicator for compliance type $D_i(1) = j$ and $D_i(0) = k$.
    - People that are moved from k to j by the instrument.
    - Let $\rho_{jk} = \mathbb{P}(D_i(1) = j, D_i(0) = k)$ be the strata size.
- We can show that the 2SLS estimator converges to:

$$\widehat{\tau}_{2SLS} \xrightarrow{p} \sum_{k=0}^{K-1} \sum_{j=k+1}^{K-1} \omega_{jk} \mathbb{E}\left(\frac{Y_i(1) - Y_i(0)}{j - k} \mid C_i = jk \right)$$

$$\omega_{jk} = \frac{(j-k)\rho_{jk}}{\sum_{s=0}^{K-1} \sum_{t=s+1}^{K-1} (s-t)\rho_{st}}$$

TSLS with multivalued treatments

- Let $C_i = jk$ be an indicator for compliance type $D_i(1) = j$ and $D_i(0) = k$.
    - People that are moved from k to j by the instrument.
    - Let $\rho_{jk} = \mathbb{P}(D_i(1) = j, D_i(0) = k)$ be the strata size.
- We can show that the 2SLS estimator converges to:

$$\widehat{\tau}_{2SLS} \xrightarrow{p} \sum_{k=0}^{K-1} \sum_{j=k+1}^{K-1} \omega_{jk} \mathbb{E}\left(\frac{Y_i(1) - Y_i(0)}{j-k} \mid C_i = jk \right)$$

$$\omega_{jk} = \frac{(j-k)\rho_{jk}}{\sum_{s=0}^{K-1} \sum_{t=s+1}^{K-1} (s-t)\rho_{st}}$$

- Intuition: a weighted average of effects per dose for each affected type.

TSLS with multivalued treatments

- Let $C_i = jk$ be an indicator for compliance type $D_i(1) = j$ and $D_i(0) = k$.
  - People that are moved from k to j by the instrument.
  - Let $\rho_{jk} = \mathbb{P}(D_i(1) = j, D_i(0) = k)$ be the strata size.
- We can show that the 2SLS estimator converges to:

$$\widehat{\tau}_{2SLS} \xrightarrow{p} \sum_{k=0}^{K-1} \sum_{j=k+1}^{K-1} \omega_{jk} \mathbb{E}\left(\frac{Y_i(1) - Y_i(0)}{j-k} \mid C_i = jk \right)$$

$$\omega_{jk} = \frac{(j-k)\rho_{jk}}{\sum_{s=0}^{K-1} \sum_{t=s+1}^{K-1} (s-t)\rho_{st}}$$

- Intuition: a weighted average of effects per dose for each affected type.
  - Weights are proportional to size of the strata and how big the effect of the instrument is for that strata.

TSLS with multivalued treatments

- Let $C_i = jk$ be an indicator for compliance type $D_i(1) = j$ and $D_i(0) = k$.
  - People that are moved from k to j by the instrument.
  - Let $\rho_{jk} = \mathbb{P}(D_i(1) = j, D_i(0) = k)$ be the strata size.
- We can show that the 2SLS estimator converges to:

$$\widehat{\tau}_{2SLS} \xrightarrow{p} \sum_{k=0}^{K-1} \sum_{j=k+1}^{K-1} \omega_{jk} \mathbb{E}\left(\frac{Y_i(1) - Y_i(0)}{j-k} \mid C_i = jk \right)$$

$$\omega_{jk} = \frac{(j-k)\rho_{jk}}{\sum_{s=0}^{K-1} \sum_{t=s+1}^{K-1} (s-t)\rho_{st}}$$

- Intuition: a weighted average of effects per dose for each affected type.
  - Weights are proportional to size of the strata and how big the effect of the instrument is for that strata.
  - If instrument can only increase by 1 dose, then simplifies to weighted average of principal strata effects.

2/ General two-stage least squares

General 2SLS

- Linear model for each i:

$$Y_i = \mathbf{X}_i'\beta + \varepsilon_i$$

General 2SLS

- Linear model for each i:

$$Y_i = \mathbf{X}_i'\beta + \varepsilon_i$$

  - $\mathbf{X}_i$ is $k \times 1$ now includes D_i and any pretreatment covariates.

General 2SLS

- Linear model for each i:

$$Y_i = \mathbf{X}_i'\beta + \varepsilon_i$$

  - $\mathbf{X}_i$ is $k \times 1$ now includes D_i and any pretreatment covariates.
  - Parts of $\mathbf{X}_i$ are endogenous so that $\mathbb{E}[\varepsilon_i \mid \mathbf{X}_i] \neq 0$

General 2SLS

- Linear model for each i:

$$Y_i = \mathbf{X}_i'\beta + \varepsilon_i$$

  - $\mathbf{X}_i$ is $k \times 1$ now includes D_i and any pretreatment covariates.
  - Parts of $\mathbf{X}_i$ are endogenous so that $\mathbb{E}[\varepsilon_i \mid \mathbf{X}_i] \neq 0$

- Instruments $\mathbf{Z}_i$ that is $\ell \times 1$ vector such that $\mathbb{E}[\varepsilon_i \mid \mathbf{Z}_i] = 0$.

General 2SLS

- Linear model for each i:

$$Y_i = \mathbf{X}_i'\beta + \varepsilon_i$$

  - $\mathbf{X}_i$ is $k \times 1$ now includes D_i and any pretreatment covariates.
  - Parts of $\mathbf{X}_i$ are endogenous so that $\mathbb{E}[\varepsilon_i \mid \mathbf{X}_i] \neq 0$

- Instruments $\mathbf{Z}_i$ that is $\ell \times 1$ vector such that $\mathbb{E}[\varepsilon_i \mid \mathbf{Z}_i] = 0$.

  - $\mathbf{Z}_i$ might include exogenous/pretreatment variables from $\mathbf{X}_i$ as well.

General 2SLS

- Linear model for each i:

$$Y_i = \mathbf{X}_i'\beta + \varepsilon_i$$

  - $\mathbf{X}_i$ is $k \times 1$ now includes D_i and any pretreatment covariates.
  - Parts of $\mathbf{X}_i$ are endogenous so that $\mathbb{E}[\varepsilon_i \mid \mathbf{X}_i] \neq 0$
- Instruments $\mathbf{Z}_i$ that is $\ell \times 1$ vector such that $\mathbb{E}[\varepsilon_i \mid \mathbf{Z}_i] = 0$.
  - $\mathbf{Z}_i$ might include exogenous/pretreatment variables from $\mathbf{X}_i$ as well.
  - Rank condition: $\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i']$ and $\mathbb{E}[\mathbf{X}_i\mathbf{Z}_i']$ have full rank.

General 2SLS

- Linear model for each i:

$$Y_i = \mathbf{X}_i'\beta + \varepsilon_i$$

  - $\mathbf{X}_i$ is $k \times 1$ now includes D_i and any pretreatment covariates.
  - Parts of $\mathbf{X}_i$ are endogenous so that $\mathbb{E}[\varepsilon_i \mid \mathbf{X}_i] \neq 0$
- Instruments $\mathbf{Z}_i$ that is $\ell \times 1$ vector such that $\mathbb{E}[\varepsilon_i \mid \mathbf{Z}_i] = 0$.
  - $\mathbf{Z}_i$ might include exogenous/pretreatment variables from $\mathbf{X}_i$ as well.
  - Rank condition: $\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i']$ and $\mathbb{E}[\mathbf{X}_i\mathbf{Z}_i']$ have full rank.
- Identification:

General 2SLS

- Linear model for each i:
$$Y_i = \mathbf{X}_i'\beta + \varepsilon_i$$
    - $\mathbf{X}_i$ is $k \times 1$ now includes D_i and any pretreatment covariates.
    - Parts of $\mathbf{X}_i$ are endogenous so that $\mathbb{E}[\varepsilon_i \mid \mathbf{X}_i] \neq 0$
- Instruments $\mathbf{Z}_i$ that is $\ell \times 1$ vector such that $\mathbb{E}[\varepsilon_i \mid \mathbf{Z}_i] = 0$.
    - $\mathbf{Z}_i$ might include exogenous/pretreatment variables from $\mathbf{X}_i$ as well.
    - Rank condition: $\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i']$ and $\mathbb{E}[\mathbf{X}_i\mathbf{Z}_i']$ have full rank.
- Identification:
    - $k = \ell$: just-identified.

General 2SLS

- Linear model for each i:

$$Y_i = \mathbf{X}_i'\beta + \varepsilon_i$$

  - $\mathbf{X}_i$ is $k \times 1$ now includes D_i and any pretreatment covariates.
  - Parts of $\mathbf{X}_i$ are endogenous so that $\mathbb{E}[\varepsilon_i \mid \mathbf{X}_i] \neq 0$
- Instruments $\mathbf{Z}_i$ that is $\ell \times 1$ vector such that $\mathbb{E}[\varepsilon_i \mid \mathbf{Z}_i] = 0$.
  - $\mathbf{Z}_i$ might include exogenous/pretreatment variables from $\mathbf{X}_i$ as well.
  - Rank condition: $\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i']$ and $\mathbb{E}[\mathbf{X}_i\mathbf{Z}_i']$ have full rank.
- Identification:
  - $k = \ell$: just-identified.
  - $k < \ell$: over-identified (can test the exclusion restriction, kinda)

General 2SLS

- Linear model for each i:

$$Y_i = \mathbf{X}_i'\beta + \varepsilon_i$$

  - $\mathbf{X}_i$ is $k \times 1$ now includes D_i and any pretreatment covariates.
  - Parts of $\mathbf{X}_i$ are endogenous so that $\mathbb{E}[\varepsilon_i \mid \mathbf{X}_i] \neq 0$

- Instruments $\mathbf{Z}_i$ that is $\ell \times 1$ vector such that $\mathbb{E}[\varepsilon_i \mid \mathbf{Z}_i] = 0$.

  - $\mathbf{Z}_i$ might include exogenous/pretreatment variables from $\mathbf{X}_i$ as well.
  - Rank condition: $\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i']$ and $\mathbb{E}[\mathbf{X}_i\mathbf{Z}_i']$ have full rank.

- Identification:

  - $k = \ell$: just-identified.
  - $k < \ell$: over-identified (can test the exclusion restriction, kinda)
  - $k > \ell$: unidentified (fails rank condition)

Nasty Matrix Algebra

- Projection matrix projects values of $\mathbf{X}_i$ onto $\mathbf{Z}_i$:

$$\boldsymbol{\Pi} = (\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i'])^{-1}\mathbb{E}[\mathbf{Z}_i\mathbf{X}_i'] \qquad \text{(projection matrix)}$$

$$\tilde{\mathbf{X}}_i = \boldsymbol{\Pi}'\mathbf{Z}_i \qquad \text{(projected values)}$$

Nasty Matrix Algebra

- Projection matrix projects values of $\mathbf{X}_i$ onto $\mathbf{Z}_i$:

$$\mathbf{\Pi} = (\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i'])^{-1}\mathbb{E}[\mathbf{Z}_i\mathbf{X}_i'] \qquad \text{(projection matrix)}$$
$$\tilde{\mathbf{X}}_i = \mathbf{\Pi}'\mathbf{Z}_i \qquad \text{(projected values)}$$

- To derive the 2SLS estimator, take the fitted values, $\mathbf{\Pi}'\mathbf{Z_i}$ and multiply both sides of the outcome equation by them:

$$Y_i = \mathbf{X}_i'\beta + \varepsilon_i$$

Nasty Matrix Algebra

- Projection matrix projects values of $\mathbf{X}_i$ onto $\mathbf{Z}_i$:

$$\mathbf{\Pi} = (\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i'])^{-1}\mathbb{E}[\mathbf{Z}_i\mathbf{X}_i'] \qquad \text{(projection matrix)}$$
$$\tilde{\mathbf{X}}_i = \mathbf{\Pi}'\mathbf{Z}_i \qquad \text{(projected values)}$$

- To derive the 2SLS estimator, take the fitted values, $\mathbf{\Pi}'\mathbf{Z_i}$ and multiply both sides of the outcome equation by them:

$$Y_i = \mathbf{X}_i'\beta + \varepsilon_i$$
$$\mathbf{\Pi}'\mathbf{Z}_i Y_i = \mathbf{\Pi}'\mathbf{Z}_i\mathbf{X}_i'\beta + \mathbf{\Pi}'\mathbf{Z}_i\varepsilon_i$$

Nasty Matrix Algebra

- Projection matrix projects values of $\mathbf{X}_i$ onto $\mathbf{Z}_i$:

$$\mathbf{\Pi} = (\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i'])^{-1}\mathbb{E}[\mathbf{Z}_i\mathbf{X}_i'] \qquad \text{(projection matrix)}$$
$$\tilde{\mathbf{X}}_i = \mathbf{\Pi}'\mathbf{Z}_i \qquad \text{(projected values)}$$

- To derive the 2SLS estimator, take the fitted values, $\mathbf{\Pi}'\mathbf{Z_i}$ and multiply both sides of the outcome equation by them:

$$Y_i = \mathbf{X}_i'\beta + \varepsilon_i$$
$$\mathbf{\Pi}'\mathbf{Z}_i Y_i = \mathbf{\Pi}'\mathbf{Z}_i\mathbf{X}_i'\beta + \mathbf{\Pi}'\mathbf{Z}_i\varepsilon_i$$
$$\mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_i Y_i] = \mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_i\mathbf{X}_i']\beta + \mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_i\varepsilon_i]$$

Nasty Matrix Algebra

- Projection matrix projects values of $\mathbf{X}_i$ onto $\mathbf{Z}_i$:

$$\mathbf{\Pi} = (\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i'])^{-1}\mathbb{E}[\mathbf{Z}_i\mathbf{X}_i'] \qquad \text{(projection matrix)}$$

$$\tilde{\mathbf{X}}_i = \mathbf{\Pi}'\mathbf{Z}_i \qquad \text{(projected values)}$$

- To derive the 2SLS estimator, take the fitted values, $\mathbf{\Pi}'\mathbf{Z_i}$ and multiply both sides of the outcome equation by them:

$$Y_i = \mathbf{X}_i'\beta + \varepsilon_i$$

$$\mathbf{\Pi}'\mathbf{Z}_i Y_i = \mathbf{\Pi}'\mathbf{Z}_i\mathbf{X}_i'\beta + \mathbf{\Pi}'\mathbf{Z}_i\varepsilon_i$$

$$\mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_i Y_i] = \mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_i\mathbf{X}_i']\beta + \mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_i\varepsilon_i]$$

$$\mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_i Y_i] = \mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_i\mathbf{X}_i']\beta + \mathbf{\Pi}'\mathbb{E}[\mathbf{Z}_i\varepsilon_i]$$

Nasty Matrix Algebra

- Projection matrix projects values of $\mathbf{X}_i$ onto $\mathbf{Z}_i$:

$$\boldsymbol{\Pi} = (\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i'])^{-1}\mathbb{E}[\mathbf{Z}_i\mathbf{X}_i'] \qquad \text{(projection matrix)}$$
$$\tilde{\mathbf{X}}_i = \boldsymbol{\Pi}'\mathbf{Z}_i \qquad \text{(projected values)}$$

- To derive the 2SLS estimator, take the fitted values, $\boldsymbol{\Pi}'\mathbf{Z_i}$ and multiply both sides of the outcome equation by them:

$$\begin{aligned}
Y_i &= \mathbf{X}_i'\beta + \varepsilon_i \\
\boldsymbol{\Pi}'\mathbf{Z}_i Y_i &= \boldsymbol{\Pi}'\mathbf{Z}_i\mathbf{X}_i'\beta + \boldsymbol{\Pi}'\mathbf{Z}_i\varepsilon_i \\
\mathbb{E}[\boldsymbol{\Pi}'\mathbf{Z}_i Y_i] &= \mathbb{E}[\boldsymbol{\Pi}'\mathbf{Z}_i\mathbf{X}_i']\beta + \mathbb{E}[\boldsymbol{\Pi}'\mathbf{Z}_i\varepsilon_i] \\
\mathbb{E}[\boldsymbol{\Pi}'\mathbf{Z}_i Y_i] &= \mathbb{E}[\boldsymbol{\Pi}'\mathbf{Z}_i\mathbf{X}_i']\beta + \boldsymbol{\Pi}'\mathbb{E}[\mathbf{Z}_i\varepsilon_i] \\
\mathbb{E}[\boldsymbol{\Pi}'\mathbf{Z}_i Y_i] &= \mathbb{E}[\boldsymbol{\Pi}'\mathbf{Z}_i\mathbf{X}_i']\beta
\end{aligned}$$

Nasty Matrix Algebra

- Projection matrix projects values of $\mathbf{X}_i$ onto $\mathbf{Z}_i$:

$$\mathbf{\Pi} = (\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i'])^{-1}\mathbb{E}[\mathbf{Z}_i\mathbf{X}_i'] \qquad \text{(projection matrix)}$$

$$\tilde{\mathbf{X}}_i = \mathbf{\Pi}'\mathbf{Z}_i \qquad \text{(projected values)}$$

- To derive the 2SLS estimator, take the fitted values, $\mathbf{\Pi}'\mathbf{Z_i}$ and multiply both sides of the outcome equation by them:

$$\begin{aligned}
Y_i &= \mathbf{X}_i'\beta + \varepsilon_i \\
\mathbf{\Pi}'\mathbf{Z}_i Y_i &= \mathbf{\Pi}'\mathbf{Z}_i\mathbf{X}_i'\beta + \mathbf{\Pi}'\mathbf{Z}_i\varepsilon_i \\
\mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_i Y_i] &= \mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_i\mathbf{X}_i']\beta + \mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_i\varepsilon_i] \\
\mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_i Y_i] &= \mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_i\mathbf{X}_i']\beta + \mathbf{\Pi}'\mathbb{E}[\mathbf{Z}_i\varepsilon_i] \\
\mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_i Y_i] &= \mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_i\mathbf{X}_i']\beta \\
\mathbb{E}[\tilde{\mathbf{X}}_i Y_i] &= \mathbb{E}[\tilde{\mathbf{X}}_i\mathbf{X}_i']\beta
\end{aligned}$$

Nasty Matrix Algebra

- Projection matrix projects values of $\mathbf{X}_i$ onto $\mathbf{Z}_i$:

$$\begin{aligned} \mathbf{\Pi} &= (\mathbb{E}[\mathbf{Z}_i\mathbf{Z}_i'])^{-1}\mathbb{E}[\mathbf{Z}_i\mathbf{X}_i'] && \text{(projection matrix)} \\ \tilde{\mathbf{X}}_i &= \mathbf{\Pi}'\mathbf{Z}_i && \text{(projected values)} \end{aligned}$$

- To derive the 2SLS estimator, take the fitted values, $\mathbf{\Pi}'\mathbf{Z_i}$ and multiply both sides of the outcome equation by them:

$$\begin{aligned} Y_i &= \mathbf{X}_i'\beta + \varepsilon_i \\ \mathbf{\Pi}'\mathbf{Z}_iY_i &= \mathbf{\Pi}'\mathbf{Z}_i\mathbf{X}_i'\beta + \mathbf{\Pi}'\mathbf{Z}_i\varepsilon_i \\ \mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_iY_i] &= \mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_i\mathbf{X}_i']\beta + \mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_i\varepsilon_i] \\ \mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_iY_i] &= \mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_i\mathbf{X}_i']\beta + \mathbf{\Pi}'\mathbb{E}[\mathbf{Z}_i\varepsilon_i] \\ \mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_iY_i] &= \mathbb{E}[\mathbf{\Pi}'\mathbf{Z}_i\mathbf{X}_i']\beta \\ \mathbb{E}[\tilde{\mathbf{X}}_iY_i] &= \mathbb{E}[\tilde{\mathbf{X}}_i\mathbf{X}_i']\beta \\ \beta &= (\mathbb{E}[\tilde{\mathbf{X}}_i\mathbf{X}_i'])^{-1}\mathbb{E}[\tilde{\mathbf{X}}_iY_i] \end{aligned}$$

How to estimate the parameters

- Collect $\mathbf{X}_i$ into a $n \times k$ matrix $\mathbb{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_n')$

How to estimate the parameters

- Collect $\mathbf{X}_i$ into a $n \times k$ matrix $\mathbb{X} = (\mathbf{X}_1', \dots, \mathbf{X}_n')$
- Collect $\mathbf{Z}_i$ into a $n \times \ell$ matrix $\mathbb{Z} = (\mathbf{Z}_1', \dots, \mathbf{Z}_n')$

How to estimate the parameters

- Collect $\mathbf{X}_i$ into a $n \times k$ matrix $\mathbb{X} = (\mathbf{X}_1', \dots, \mathbf{X}_n')$
- Collect $\mathbf{Z}_i$ into a $n \times \ell$ matrix $\mathbb{Z} = (\mathbf{Z}_1', \dots, \mathbf{Z}_n')$
- In-sample projection matrix produces fitted values: $\widehat{\mathbb{X}} = \mathbb{Z}(\mathbb{Z}'\mathbb{Z})^{-1}\mathbb{Z}'\mathbb{X}$

How to estimate the parameters

- Collect $\mathbf{X}_i$ into a $n \times k$ matrix $\mathbb{X} = (\mathbf{X}_1', \dots, \mathbf{X}_n')$
- Collect $\mathbf{Z}_i$ into a $n \times \ell$ matrix $\mathbb{Z} = (\mathbf{Z}_1', \dots, \mathbf{Z}_n')$
- In-sample projection matrix produces fitted values: $\widehat{\mathbb{X}} = \mathbb{Z}(\mathbb{Z}'\mathbb{Z})^{-1}\mathbb{Z}'\mathbb{X}$
  - Fitted values of regression of $\mathbb{X}$ on $\mathbb{Z}$.

How to estimate the parameters

- Collect $\mathbf{X}_i$ into a $n \times k$ matrix $\mathbb{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_n')$
- Collect $\mathbf{Z}_i$ into a $n \times \ell$ matrix $\mathbb{Z} = (\mathbf{Z}_1', \ldots, \mathbf{Z}_n')$
- In-sample projection matrix produces fitted values: $\widehat{\mathbb{X}} = \mathbb{Z}(\mathbb{Z}'\mathbb{Z})^{-1}\mathbb{Z}'\mathbb{X}$
  - Fitted values of regression of $\mathbb{X}$ on $\mathbb{Z}$.
  - Matrix party trick: $\mathbb{X}'\mathbb{Z}/n = (1/n)\sum_i^n \mathbf{X}_i\mathbf{Z}_i' \overset{p}{\to} \mathbb{E}[\mathbf{X}_i\mathbf{Z}_i']$.

How to estimate the parameters

- Collect $\mathbf{X}_i$ into a $n \times k$ matrix $\mathbb{X} = (\mathbf{X}_1', \dots, \mathbf{X}_n')$
- Collect $\mathbf{Z}_i$ into a $n \times \ell$ matrix $\mathbb{Z} = (\mathbf{Z}_1', \dots, \mathbf{Z}_n')$
- In-sample projection matrix produces fitted values: $\widehat{\mathbb{X}} = \mathbb{Z}(\mathbb{Z}'\mathbb{Z})^{-1}\mathbb{Z}'\mathbb{X}$
  - Fitted values of regression of $\mathbb{X}$ on $\mathbb{Z}$.
  - Matrix party trick: $\mathbb{X}'\mathbb{Z}/n = (1/n)\sum_i^n \mathbf{X}_i\mathbf{Z}_i' \xrightarrow{p} \mathbb{E}[\mathbf{X}_i\mathbf{Z}_i']$.
- Take the population formula for the parameters:

$$\beta = (\mathbb{E}[\tilde{\mathbf{X}}_i\mathbf{X}_i'])^{-1}\mathbb{E}[\tilde{\mathbf{X}}_i Y_i]$$

How to estimate the parameters

- Collect $\mathbf{X}_i$ into a $n \times k$ matrix $\mathbb{X} = (\mathbf{X}_1', \dots, \mathbf{X}_n')$
- Collect $\mathbf{Z}_i$ into a $n \times \ell$ matrix $\mathbb{Z} = (\mathbf{Z}_1', \dots, \mathbf{Z}_n')$
- In-sample projection matrix produces fitted values: $\widehat{\mathbb{X}} = \mathbb{Z}(\mathbb{Z}'\mathbb{Z})^{-1}\mathbb{Z}'\mathbb{X}$
    - Fitted values of regression of $\mathbb{X}$ on $\mathbb{Z}$.
    - Matrix party trick: $\mathbb{X}'\mathbb{Z}/n = (1/n)\sum_i^n \mathbf{X}_i\mathbf{Z}_i' \overset{p}{\to} \mathbb{E}[\mathbf{X}_i\mathbf{Z}_i']$.
- Take the population formula for the parameters:

$$\beta = (\mathbb{E}[\tilde{\mathbf{X}}_i\mathbf{X}_i'])^{-1}\mathbb{E}[\tilde{\mathbf{X}}_i Y_i]$$

- And plug in the sample values (the n cancels out):

$$\hat{\beta}_{2SLS} = (\widehat{\mathbb{X}}'\mathbb{X})^{-1}\widehat{\mathbb{X}}'\mathbf{y} \overset{p}{\to} \beta$$

How to estimate the parameters

- Collect $\mathbf{X}_i$ into a $n \times k$ matrix $\mathbb{X} = (\mathbf{X}_1', \dots, \mathbf{X}_n')$
- Collect $\mathbf{Z}_i$ into a $n \times \ell$ matrix $\mathbb{Z} = (\mathbf{Z}_1', \dots, \mathbf{Z}_n')$
- In-sample projection matrix produces fitted values: $\widehat{\mathbb{X}} = \mathbb{Z}(\mathbb{Z}'\mathbb{Z})^{-1}\mathbb{Z}'\mathbb{X}$
    - Fitted values of regression of $\mathbb{X}$ on $\mathbb{Z}$.
    - Matrix party trick: $\mathbb{X}'\mathbb{Z}/n = (1/n)\sum_i^n \mathbf{X}_i\mathbf{Z}_i' \xrightarrow{p} \mathbb{E}[\mathbf{X}_i\mathbf{Z}_i']$.
- Take the population formula for the parameters:

$$\beta = (\mathbb{E}[\tilde{\mathbf{X}}_i\mathbf{X}_i'])^{-1}\mathbb{E}[\tilde{\mathbf{X}}_i Y_i]$$

- And plug in the sample values (the n cancels out):

$$\hat{\beta}_{2SLS} = (\widehat{\mathbb{X}}'\mathbb{X})^{-1}\widehat{\mathbb{X}}'\mathbf{y} \xrightarrow{p} \beta$$

- This is how R/Stata estimates the 2SLS parameters

Asymptotic variance for 2SLS

- We can write the centered, normalized TSLS estimator as:

$$\sqrt{n}(\widehat{\beta}_{2SLS} - \beta) = \underbrace{\left(n^{-1}\sum_i \widehat{\mathbf{X}}_i \widehat{\mathbf{X}}_i'\right)^{-1}}_{\xrightarrow{p}(\mathbb{E}[\widehat{\mathbf{X}}_i \widehat{\mathbf{X}}_i'])^{-1}} \underbrace{\left(n^{-1/2}\sum_i \widehat{\mathbf{X}}_i \varepsilon_i\right)}_{\xrightarrow{d} N(0, \mathbb{E}[\widehat{\mathbf{X}}_i' \varepsilon_i' \varepsilon_i \widehat{\mathbf{X}}_i])}$$

Asymptotic variance for 2SLS

- We can write the centered, normalized TSLS estimator as:

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) = \underbrace{\left(n^{-1}\sum_i \widehat{\mathbf{X}}_i \widehat{\mathbf{X}}_i'\right)^{-1}}_{\xrightarrow{p}(\mathbb{E}[\widehat{\mathbf{X}}_i \widehat{\mathbf{X}}_i'])^{-1}} \underbrace{\left(n^{-1/2}\sum_i \widehat{\mathbf{X}}_i \varepsilon_i\right)}_{\xrightarrow{d} N(0, \mathbb{E}[\widehat{\mathbf{X}}_i' \varepsilon_i' \varepsilon_i \widehat{\mathbf{X}}_i])}$$

- Thus, we have that $\sqrt{n}(\hat{\beta}_{2SLS} - \beta)$ has asymptotic variance:

$$(\mathbb{E}[\widehat{\mathbf{X}}_i \widehat{\mathbf{X}}_i'])^{-1} \mathbb{E}[\widehat{\mathbf{X}}_i' \varepsilon_i' \varepsilon_i \widehat{\mathbf{X}}_i] (\mathbb{E}[\widehat{\mathbf{X}}_i \widehat{\mathbf{X}}_i'])^{-1}$$

Asymptotic variance for 2SLS

- We can write the centered, normalized TSLS estimator as:

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) = \underbrace{\left(n^{-1}\sum_i \widehat{\mathbf{X}}_i \widehat{\mathbf{X}}_i'\right)^{-1}}_{\xrightarrow{p} (\mathbb{E}[\widehat{\mathbf{X}}_i \widehat{\mathbf{X}}_i'])^{-1}} \underbrace{\left(n^{-1/2}\sum_i \widehat{\mathbf{X}}_i \varepsilon_i\right)}_{\xrightarrow{d} N(0, \mathbb{E}[\widehat{\mathbf{X}}_i' \varepsilon_i' \varepsilon_i \widehat{\mathbf{X}}_i])}$$

- Thus, we have that $\sqrt{n}(\hat{\beta}_{2SLS} - \beta)$ has asymptotic variance:

$$(\mathbb{E}[\widehat{\mathbf{X}}_i \widehat{\mathbf{X}}_i'])^{-1} \mathbb{E}[\widehat{\mathbf{X}}_i' \varepsilon_i' \varepsilon_i \widehat{\mathbf{X}}_i] (\mathbb{E}[\widehat{\mathbf{X}}_i \widehat{\mathbf{X}}_i'])^{-1}$$

- **Robust 2SLS variance estimator** with residuals $\hat{u}_i = Y_i - \mathbf{X}_i'\hat{\beta}$:

$$\widehat{\text{var}}(\hat{\beta}_{2SLS}) = (\widehat{\mathbb{X}}'\widehat{\mathbb{X}})^{-1} \left(\sum_i \hat{u}_i^2 \widehat{\mathbf{X}}_i \widehat{\mathbf{X}}_i'\right) (\widehat{\mathbb{X}}'\widehat{\mathbb{X}})^{-1}$$

Asymptotic variance for 2SLS

- We can write the centered, normalized TSLS estimator as:

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) = \underbrace{\left(n^{-1}\sum_i \widehat{\mathbf{X}}_i\widehat{\mathbf{X}}_i'\right)^{-1}}_{\xrightarrow{p}(\mathbb{E}[\widehat{\mathbf{X}}_i\widehat{\mathbf{X}}_i'])^{-1}} \underbrace{\left(n^{-1/2}\sum_i \widehat{\mathbf{X}}_i\varepsilon_i\right)}_{\xrightarrow{d}N(0,\mathbb{E}[\widehat{\mathbf{X}}_i'\varepsilon_i'\varepsilon_i\widehat{\mathbf{X}}_i])}$$

- Thus, we have that $\sqrt{n}(\hat{\beta}_{2SLS} - \beta)$ has asymptotic variance:

$$(\mathbb{E}[\widehat{\mathbf{X}}_i\widehat{\mathbf{X}}_i'])^{-1}\mathbb{E}[\widehat{\mathbf{X}}_i'\varepsilon_i'\varepsilon_i\widehat{\mathbf{X}}_i](\mathbb{E}[\widehat{\mathbf{X}}_i\widehat{\mathbf{X}}_i'])^{-1}$$

- **Robust 2SLS variance estimator** with residuals $\hat{u}_i = Y_i - \mathbf{X}_i'\hat{\beta}$:

$$\widehat{\text{var}}(\hat{\beta}_{2SLS}) = (\widehat{\mathbb{X}}'\widehat{\mathbb{X}})^{-1}\left(\sum_i \hat{u}_i^2\widehat{\mathbf{X}}_i\widehat{\mathbf{X}}_i'\right)(\widehat{\mathbb{X}}'\widehat{\mathbb{X}})^{-1}$$

- HC2, clutering, and autocorrelation versions exist