Econometrics Notes

Jiahui Shui

April 1, 2025

Contents

Remark

Sections 1–4 are covered in ECON 220A and ECON 220B (the propensity score section has been removed, and the M-estimation part is combined with the corresponding section in ECON 220C).
Sections 5 onward are covered in ECON 220C.

1 Probability and Statistics Review

1.1 Probability Space

1.2 Conditional Probability and Independence

1.3 Random Variable

1.4 Convergence

Definition 1.1 (Convergence in Probability). Let $\{X_n\}, X$ be random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. We say X_n converges to X in probability if, $\forall \varepsilon > 0$, $\mathbb{P}(|X_n - X| \geq \varepsilon) \to 0$ as $n \to \infty$. Or equivalently,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| < \varepsilon) = 1, \quad \forall \varepsilon > 0 \tag{1}$$

We denote it as $X_n \xrightarrow{p} X$.

Definition 1.2 (Almost Surely Convergence). Let $\{X_n\}, X$ be random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. We say X_n converges to X almost surely, if

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1 \tag{2}$$

This can be written as $X_n \xrightarrow{a.s.} X$

It is easy to say that $X_n \xrightarrow{a.s} X \Rightarrow X_n \xrightarrow{p} X$. But the opposite direction is not true. Counterexample: consider $((0,1], \mathcal{B}_{(0,1]}, \lambda)$, where λ is Lebesgue measure. Let

$$\xi_n = \mathbf{1}_{(n/2^k - 1, (n+1)/2^k - 1)}, \quad 2^k \leq n < 2^{k+1}$$

Then $\mathbb{P}(|\xi_n| > \varepsilon) \leq 1/2^k \to 0$ but $\lim \xi_n(\omega)$ does not exist for any $\omega \in (0,1]$.

Definition 1.3 (Bounded in Probability). $\{X_n\}$ is said to be bounded in probability if $\forall \varepsilon > 0$, $\exists M > 0$ such that

$$\inf_n \mathbb{P}(|X_n| \leq M) \geq 1 - \varepsilon \tag{3}$$

or equivalently, $\inf_n \mathbb{P}(|X_n| > M) < \varepsilon$

If $X_n \xrightarrow{p} 0$, the we denote it as $X_n = o_p(1)$.
If X_n is bounded in probability, then we denote it as $X_n = O_p(1)$. Moreover:

$$X_n = o_p(a_n) \Leftrightarrow \frac{X_n}{a_n} = o_p(1)$$

and

$$X_n = O_p(a_n) \Leftrightarrow \frac{X_n}{a_n} = O_p(1)$$

Exercise. Prove each of the followings:

(i) $o_p(1) + o_p(1) = o_p(1)$
(ii) $o_p(1) + O_p(1) = O_p(1)$
(iii) $o_p(1)O_p(1) = o_p(1)$
(iv) $(1 + o_p(1))^{-1} = O_p(1)$
(v) $o_p(a_n) = a_n o_p(1)$
(vi) $O_p(a_n) = a_n O_p(1)$
(vii) $o_p(O_p(1)) = o_p(1)$

Remark. (iii) is an implication of Slutsky theorem. Since $o_p(1)O_p(1) \xrightarrow{d} 0$, then it must converge to 0 in probability by proposition 1.6

A powerful theorem to prove convergence in probability:

Theorem 1.4 (Continuous Mapping Theorem). Suppose that a measurable function g is (a.s.) continuous, then

$$X_n \xrightarrow{p} X_\infty \Rightarrow g(X_n) \xrightarrow{p} g(X_\infty) \tag{4}$$

Moreover, it applies to $a.s.$ convergence and convergence in distribution.

Definition 1.5 (Convergence in Distribution). We say $X_n \xrightarrow{d} X$ if the distribution $P_n := \mathbb{P}\{X_n \in \cdot\}$ converges to $P := \mathbb{P}\{X \in \cdot\}$. Or, equivalently

$$F_{X_n}(x) \to F(x), \quad \text{for any point } x \text{ that } F(x) \text{ is continuous} \tag{5}$$

Another important result is that if for any $t \in \mathbb{R}$, the characteristic function $\phi_{X_n}(t) \to \phi_X(t)$, then $X_n \xrightarrow{d} X$.
We can prove that

$$X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X \tag{6}$$

Another important result for convergence in distribution is:

Proposition 1.6. If $X_n \xrightarrow{d} c$ where c is a constant, then $X_n \xrightarrow{p} c$.

Proof. Let $X = c$, then $F_X(x) = \mathbf{1}_{\{x \geq c\}}$. Hence, $\forall \varepsilon > 0$

$$\begin{aligned} \lim_{n\to\infty} \mathbb{P}(|X_n - c| < \varepsilon) &= \lim_{n\to\infty} \mathbb{P}(c - \varepsilon < X_n < c + \varepsilon) \\ &= \lim_{n\to\infty} (F_{X_n}(c+\varepsilon) - F_{X_n}(c-\varepsilon)) \\ &= 1 - 0 = 1 \end{aligned} \tag{7}$$

Then we know that $X_n \xrightarrow{p} c$. □

Lemma 1.7 (Marginal Convergence and Joint Convergence).

- If $X_n \xrightarrow{p} X, Y_n \xrightarrow{p} Y$, then $(X_n, Y_n) \xrightarrow{p} (X, Y)$
- If $X_n \xrightarrow{d} X, Y_n \xrightarrow{d} Y$ and X_n, Y_n are independent for all n, X, Y are independent, then $(X_n, Y_n) \xrightarrow{d} (X, Y)$
- If $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} c$, then $(X_n, Y_n) \xrightarrow{d} (X, c)$
- If $X_n \xrightarrow{a.s.} X, Y_n \xrightarrow{a.s.} Y$, then $(X_n, Y_n) \xrightarrow{a.s.} (X, Y)$

Theorem 1.8 (Slutsky's Theorem). Let $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, where c is constant.

- $X_n + Y_n \xrightarrow{d} X + c$
- $X_n Y_n \xrightarrow{d} cX$
- $X_n / Y_n \xrightarrow{d} X/c$ if $c \neq 0$

Exercise. Let $\{X_n\}$ be independent with $X_n \sim \text{Gamma}(\alpha_n, \beta_n)$. $\alpha_n \to \alpha, \beta_n \to \beta$ for some positive real number α, β. Now, let $\hat{\beta}_n$ be a consistent estimator for β. Prove that $X_n / \hat{\beta}_n \xrightarrow{d} \text{Gamma}(\alpha, 1)$

Theorem 1.9 (Delta Method). First order expansion: Suppose that g is differentiable at $\boldsymbol{c}$, for any sequence $0 < a_n \to \infty$, we have

$$a_n(\boldsymbol{X}_n - \boldsymbol{c}) \xrightarrow{d} \boldsymbol{X} \Rightarrow a_n[g(\boldsymbol{X}_n) - g(\boldsymbol{c})] \xrightarrow{d} [\nabla g(\boldsymbol{c})]^\top \boldsymbol{X} \tag{8}$$

If $\nabla g(\boldsymbol{c}) = 0$, then we have similar expansion: To be completed

Example. Suppose that $\{X_i\}$ i.i.d with mean μ and variance σ^2. Consider the following estimator:

$$\hat{\theta} = \bar{x}^2 \tag{9}$$

Let $\theta := \mu^2$. (1) Find the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta)$ provided $\mu \neq 0$. (2) If $\mu = 0$, find the convergence rate of $\hat{\theta}$ and its limit distribution.

Theorem 1.10 (Prohorov's Theorem). If $X_n \xrightarrow{d} X$, then $X_n = O_p(1)$

Proof. $\forall \varepsilon > 0$, we can choose M_0 sufficiently large such that

$$\mathbb{P}(|X| > M_0) < \varepsilon) \tag{10}$$

Then, since $\mathbb{P}(|X_n| > M_0) \to \mathbb{P}(|X| > M_0)$, then we can choose n_0 such that for all $n \geq n_0$, $\mathbb{P}(|X_n| > M_0) < \varepsilon$. Now, we can select M_1 such that

$$\mathbb{P}(|X_i| > M_1) < \varepsilon, \quad \forall i = 1, \cdots, n_0 - 1 \tag{11}$$

Then let $M = \max(M_0, M_1)$ we have $\mathbb{P}(|X_n| > M) < \varepsilon$ for all n. □

A natural question is: will bounded in probability imply convergence in distribution? The answer is **No**. Consider $X_n = 2 + 1/n$ for even n and $X_n = 1 + 1/(n+1)$ for odd n. Then the sequence (X_{2k}) converges in distribution to $Y = 2$. And (X_{2k-1}) converges in distribution to $W = 1$. Since $Y \neq W$ then the sequence does not converge in distribution. Since all X_n lie in the interval $[1, 5/2]$,then we can easily show that $X_n = O_p(1)$.

1.5 Law of Large Numbers

Theorem 1.11 (WLLN, Khintchin)**.** If $\{X_n\}$ are i.i.d with $\mathbb{E}[X_1] = \mu < \infty$, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu \tag{12}$$

Theorem 1.12 (SLLN, Kolmogorov)**.** If $\{X_n\}$ are i.i.d with $\mathbb{E}[X_1] = \mu < \infty$, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu \tag{13}$$

1.6 Central Limit Theorem

Theorem 1.13 (Levy CLT)**.** Suppose that $\{X_n\}$ i.i.d with mean μ and variance σ^2, then

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma^2) \tag{14}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

1.7 Normal Distribution

Consider multivariate normal distribution: $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} \in \mathbb{R}^d, \boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$. The moment generating function is

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = \mathbb{E}[e^{\boldsymbol{t'X}}] = e^{\boldsymbol{t'\mu} + \frac{1}{2}\boldsymbol{t'\Sigma t}}, \quad \boldsymbol{t} \in \mathbb{R}^d \tag{15}$$

For any $\boldsymbol{A} \in \mathbb{R}^{m\times d}$, we have $\boldsymbol{AX} + \boldsymbol{b} \sim N(\boldsymbol{A\mu} + \boldsymbol{b}, \boldsymbol{A\Sigma A'})$. The probability density function of $\boldsymbol{X}$ is given by

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\det(\boldsymbol{\Sigma})^{1/2}} \exp\left\{-\frac{1}{2}\boldsymbol{(x-\mu)'\Sigma^{-1}(x-\mu)}\right\}, \quad \boldsymbol{x} \in \mathbb{R}^d \tag{16}$$

- If $X_1, \cdots, X_n \sim N(0,1)$ i.i.d, then $X_1^2 + \cdots + X_n^2 \sim \chi_n^2$.
- If $X \sim N(0,1)$ and $Q \sim \chi_n^2$ are independent, then $\frac{X}{\sqrt{Q/n}} \sim t_n$
- If $Q_1 \sim \chi_m^2, Q_2 \sim \chi_n^2$ are independent, then $\frac{Q_1/m}{Q_2/n} \sim F_{m,n}$

1.8 Hypothesis Testing

Consider $H_0 : \theta \in \Theta$, this is called the null hypothesis. The alternative hypothesis is $H_1 : \theta \in \Theta_1$, where $\Theta_1 = \Theta \backslash \Theta_0$. We have to decide between H_0 and H_1. Let R denotes the reject region. A Test might have two types of mistake.

- **Type I Error**: Reject H_0 when $\theta \in \Theta_0$. $\mathbb{P}_\theta(\boldsymbol{X} \in R)$ for $\theta \in \Theta_0$
- **Type II Error**: Accept H_0 when $\theta \in \Theta_1$. $\mathbb{P}_\theta(\boldsymbol{X} \in R^c)$ for $\theta \in \Theta_1$

In this notes, we will use φ to denote *power function* for a hypothesis test, i.e.

$$\beta(\theta) = \mathbb{P}_\theta(\boldsymbol{X} \in R) \tag{17}$$

When $\theta \in \Theta_0$, then $\beta(\theta) = \mathbb{P}(\text{Type I Error})$. If $\theta \in \Theta_1$, then $\beta(\theta) = 1 - \mathbb{P}(\text{Type II Error})$

Definition 1.14. For $\alpha \in [0,1]$, a test with power function $\beta(\theta)$ is a *size* α test if

$$\sup_{\theta\in\Theta_0} \beta(\theta) = \alpha,$$

is a *level* α test if

$$\sup_{\theta\in\Theta_0} \beta(\theta) \le \alpha$$

Definition 1.15. A test is *unbiased* if $\beta(\theta') \ge \beta(\theta'')$ for all $\theta' \in \Theta_1$ and $\theta'' \in \Theta_0$. A test is *consistent* if

$$\lim_{n\to\infty} \beta(\theta) = 1, \quad \forall \theta \in \Theta_1$$

Definition 1.16. Let $\mathcal{C}$ be a class of tests. A test in class $\mathcal{C}$ with power function $\beta(\theta)$ is a *uniformly most powerful* (UMP) class $\mathcal{C}$ test if $\beta(\theta) \ge g(\theta)$ for every $\theta \in \Theta_1$ and every $g(\theta)$ that is a power function of a test in class $\mathcal{C}$. Generally we take $\mathcal{C}$ as the class of all level α test.

Theorem 1.17 (Neyman-Pearson Lemma)**.** Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, where the pdf or pmf corresponding to θ_i is $f(x|\theta_i), i = 0, 1$. Then

$$R = \{x : f(x|\theta_1) > kf(x|\theta)\}$$

for some $k \geq 0$ and $\alpha = \mathbb{P}_{\theta_0}(X \in R)$ is a UMP level α test.

Definition 1.18 (p-value)**.** A p-value $p(X)$ is a test statistic satisfying $0 \leq p(x) \leq 1$ for every sample point x. Small values of $p(X)$ give evidence that H_1 is true. A p-value is valid if, for every $\theta \in \Theta_0$ and every $0 \leq \alpha \leq 1$,

$$\mathbb{P}_\theta(p(X) \leq \alpha) \leq \alpha \tag{18}$$

If $p(X)$ is a valid p-value, then it is easy to construct a level α test based on this statistics. We rejects H_0 if and only if $p(X) \leq \alpha$.

Theorem 1.19. Suppose that $T(X)$ is a test statistic such that large values of T give evidence that H_1 is true. For each sample point x, define

$$p(x) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(X) \geq T(x)) \tag{19}$$

Then $p(X)$ is a valid p-value.

Now, suppose that β is a parameter, and $\hat{\beta}$ is an estimator of β. Moreover, we assume that $\hat{\beta}$ is consistent and asymptotically normal, i.e.

$$\hat{\beta} \xrightarrow{p} \beta, \quad \sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2)$$

Also, suppose that $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$ is an estimator for asymptotical variance. Now we want to test the hypothesis: $H_0 : \beta = c$, where c is a constant. To this end, we can employ t-test:

$$T = \frac{\hat{\beta} - c}{\text{se}(\hat{\beta})} = \frac{\hat{\beta} - c}{\hat{\sigma}/\sqrt{n}} \tag{20}$$

Then under the null, T is asymptotically normal since by Slutsky's theorem, we have

$$\frac{\hat{\beta} - c}{\hat{\sigma}/\sqrt{n}} = \frac{1}{\hat{\sigma}}\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, 1) \tag{21}$$

Then let the rejection region be

$$R := \{|T| \geq z_{\alpha/2}\}, \quad z_{\alpha/2} := \Phi^{-1}(1 - \alpha/2)$$

where $\Phi(x)$ is standard normal cdf. Also, we can construct confidence interval by test inversion:

$$\text{CI}_{1-\alpha} = \left[\hat{\beta} - z_{\alpha/2}\text{se}(\hat{\beta}), \hat{\beta} + z_{\alpha/2}\text{se}(\hat{\beta})\right] \tag{22}$$

Exercise. To be completed.

1.9 Wald's Test

1.10 Introduction to Bayesian Statistics

1.11 Some Exercise

2 Causal Inference and Potential Outcomes

2.1 Potential Outcomes Model

We denote x_i the treatment status, i.e. $x_i = 1$ implies the individual is treated and $x_i = 0$ indicates the individual is not treated. Also, the out come $y_i(1)$ is "the outcome of the i^{th} individual had she received the treatment. Therefore, for each i, we have three random variables, $(x_i, y_i(1), y_i(1))$.

Definition 2.1. Individual treatment effect (ITE):

$$\tau_i = y_i(1) - y_i(0)$$

Average treatment effect (ATE):

$$\tau_{\text{ATE}} = \mathbb{E}[\tau_i] = \mathbb{E}[y_i(1) - y_i(0)]$$

Treatment effect on the treated (ATT):

$$\tau_{\text{ATT}} = \mathbb{E}[\tau_i | x_i = 1] = \mathbb{E}[y_i(1) - y_i(0) | x_i = 1]$$

Treatment effect on the untreated (ATU):

$$\tau_{\text{ATU}} = \mathbb{E}[\tau_i | x_i = 0] = \mathbb{E}[y_i(1) - y_i(0) | x_i = 0]$$

We can always write y_i as

$$y_i = y_i(1)\mathbb{1}_{\{x_i=1\}} + y_i(0)\mathbb{1}_{\{x_i=0\}} = x_i y_i(1) + (1 - x_i)y_i(0) \tag{23}$$

In words, we will never be able to observe the two potential outcomes simultaneously for any individual.

2.2 Randomized Controlled Trials

Definition 2.2 (Missing completely at random; independent treatment)**.** The potential outcomes are said to be missing completely at random, if

$$x_i \perp\!\!\!\perp (y_i(1), y_i(0)) \tag{24}$$

It means the treatment status is independent of the potential outcomes. And therefore two individuals with different treatment status should not differ systematically. This is a very strong assumption. But there is a special case that we believe this assumption holds: randomized controlled trials (RCT).

In an RCT, a sample of units are randomized into two groups, the treatment group and the control group. Then individuals who are in the treatment group will be exposed to the treatment, while those in the control group are not. Due to the randomization of treatment, it is plausible to believe that treatment assignment is independent of the potential outcomes, and hence this assumption holds.

How this assumption will help to identify the treatment effects? Consider the ATE:

$$\tau_{ATE} = \mathbb{E}[y_i(1) - y_i(0)] = \mathbb{E}[y_i(1)] - \mathbb{E}[y_i(0)] \tag{25}$$

Since the treatment is independent of the potential outcomes, then

$$\mathbb{E}[y_i(1)] = \mathbb{E}[y_i(1)|x_i = 1] = \mathbb{E}[y_i|x_i = 1] \tag{26}$$

Similarly, $\mathbb{E}[y_i(0)] = \mathbb{E}[y_i|x_i = 0]$. Hence

Proposition 2.3. Under the independent treatment assumption, the ATE is identified by

$$\tau_{ATE} = \mathbb{E}[y_i(1) - y_i(0)] = \mathbb{E}[y_i|x_i = 1] - \mathbb{E}[y_i|x_i = 0] \tag{27}$$

Now we re-write the potential outcomes into an expectation component and an error term:

$$y_i(1) = \mathbb{E}[y_i(1)] + \underbrace{y_i(1) - \mathbb{E}[y_i(1)]}_{u_i(1)}, \quad y_i(0) = \mathbb{E}[y_i(0)] + \underbrace{y_i(0) - \mathbb{E}[y_i(0)]}_{u_i(0)} \tag{28}$$

Then

$$\begin{aligned} y_i &= x_i y_i(1) + (1 - x_i) y_i(0) \\ &= x_i(\mathbb{E}[y_i(1)] + u_i(1)) + (1 - x_i)(\mathbb{E}[y_i(0)] + u_i(0)) \\ &= \mathbb{E}[y_i(0)] + x_i(\mathbb{E}[y_i(1)] - \mathbb{E}[y_i(0)]) + x_i u_i(1) + (1 - x_i) u_i(0) \\ &= \beta_0 + \beta_1 x_i + u_i \end{aligned} \tag{29}$$

Then under the independent treatment assumption, we have the regression expression, where $\beta_0 = \mathbb{E}[y_i(0)], \beta_1 = \tau_{ATE}$ and $\mathbb{E}[u_i|x_i] = 0$. The OLS estimators are

$$\hat{\beta}_0 = \frac{1}{n_0} \sum_{i=1}^{n} y_i \mathbb{1}_{\{x_i=0\}}, \quad \hat{\beta}_1 = \frac{1}{n_1} \sum_{i=1}^{n} y_i \mathbb{1}_{\{x_i=1\}} - \frac{1}{n_0} \sum_{i=1}^{n} y_i \mathbb{1}_{\{x_i=0\}} \tag{30}$$

where $n_1 = \sum_{i=1}^{n} x_i$ is the size of the treatment group, and $n_0 = \sum_{i=1}^{n} (1 - x_i)$ is the size of the control group. Also we can re-write them as

$$\hat{\beta}_0 = \frac{n}{n_0} \frac{1}{n} \sum_{i=1}^{n} y_i(1 - x_i) = \frac{n}{n_0} \frac{1}{n} \sum_{i=1}^{n} y_i(0)(1 - x_i)$$

Note that by LLN

$$\frac{n}{n_0} \xrightarrow{p} \frac{1}{\mathbb{P}(x_i = 0)}$$

and

$$\frac{1}{n} \sum_{i=1}^{n} y_i(0)(1 - x_i) \xrightarrow{p} \mathbb{E}[y_i(0)(1 - x_i)] = \mathbb{E}[y_i(0)]\mathbb{E}[1 - x_i] = \mathbb{E}[y_i(0)]\mathbb{P}(x_i = 0)$$

Hence

$$\hat{\beta}_0 \xrightarrow{p} \beta_0 = \mathbb{E}[y_i(0)]$$

Also

$$\frac{n}{n_1} \frac{1}{n} \sum_{i=1}^{n} y_i \mathbb{1}_{\{x_i=1\}} \xrightarrow{p} \frac{1}{\mathbb{P}(x_i = 1)} \mathbb{E}[y_i(1)]\mathbb{P}(x_i = 1) = \mathbb{E}[y_i(1)]$$

Hence

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 = \mathbb{E}[y_i(1)] - \mathbb{E}[y_i(0)] = \tau_{ATE}$$

2.3 Selection Bias

What if we don't assume the independent assumption? Then $\hat{\beta}_1$ is consistent for

$$\hat{\beta}_1 \xrightarrow{p} \mathbb{E}[y_i(1)|x_i = 1] - \mathbb{E}[y_i(0)|x_i = 0]$$

Without the independent assumption, we can not pull out ATE from the conditional expectation, but we have

$$\mathbb{E}[y_i(1)|x_i = 1] - \mathbb{E}[y_i(0)|x_i = 1] = \underbrace{\mathbb{E}[y_i(1)|x_i = 1] - \mathbb{E}[y_i(0)|x_i = 1]}_{\tau_{ATT}} + \underbrace{\mathbb{E}[y_i(0)|x_i = 1] - \mathbb{E}[y_i(0)|x_i = 0]}_{\text{selection bias}}$$

3 Linear Regression

Definition 3.1 (Missing at random; conditional independence; selection on observables)**.** The potential outcomes are said to be missing at random (or, alternatively, the treatment assignment satisfies the selection on observables assumption), if

$$x_i \perp\!\!\!\perp (y_i(1), y_i(0))|w_i \tag{31}$$

where w_i is some characteristics of each individual that can be observed by researcher.

Then we have

$$y_i(1) = \underbrace{\mathbb{E}[y_i(1)|w_i]}_{g_1(w_i)} + u_i(1), \quad y_i(0) = \underbrace{\mathbb{E}[y_i(0)|w_i]}_{g_0(w_i)} + u_i(0) \tag{32}$$

where $u_i(1) = y_i(1) - g_1(w_i)$ and $u_i(0) = y_i(0) - g_0(w_i)$. Then

$$y_i = g_0(w_i) + x_i(g_1(w_i) - g_0(w_i)) + \underbrace{x_i u_i(1) + (1 - x_i)u_i(0)}_{u_i} \tag{33}$$

We have to make another very strong assumption:

$$g_1(w_i) = \mathbb{E}[y_i(1)] + w_i^\top \delta, \quad g_0(w_i) = \mathbb{E}[y_i(0)] + w_i^\top \delta \tag{34}$$

Then the outcome variable is

$$y_i = \mathbb{E}[y_i(0)] + w_i^\top \delta + x_i \left(\mathbb{E}[y_i(1)] + w_i^\top \delta - \mathbb{E}[y_i(0)] - w_i^\top \delta\right) + u_i = \mathbb{E}[y_i(0)] + x_i \tau_{ATE} + w_i^\top \delta + u_i \tag{35}$$

3.1 Algebraic Properties

Model:

$$y_i = x_i^\top \beta + u_i \tag{36}$$

we assume that $\mathbb{E}[u_i] = 0$ and $\mathbb{E}[u_i x_i] = 0$. The moment condition is

$$\frac{1}{n}\sum_{i=1}^n x_i(y_i - x_i^\top \hat{\beta}) = 0 \tag{37}$$

Standard algebra leads to

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_i y_i\right) \tag{38}$$

Matrix form: $\hat{\beta} = (X^\top X)^{-1} X^\top y$.

Definition 3.2 (Sum of squares, R-squared)**.** Let $\hat{y}_i = x_i^\top \hat{\beta}$. Define

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2, \quad \text{ESS} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2, \quad \text{SSR} = \sum_{i=1}^n (y_i - \hat{y})^2$$

We have TSS = ESS + SSR. A goodness-of-fit measure is the R-squared, which is defined as

$$R^2 = \frac{\text{ESS}}{\text{TSS}} \tag{39}$$

We often define projection matrix

$$P_x = X(X'X)^{-1}X'y, \quad \hat{y} = P_x y \tag{40}$$

It has following properties: (1) P_x is symmetric. (2) P_x is idempotent, i.e. $P_x P_x = P_x$. (3) $P_x X = X$.

Similarly, we can define $M_x = I - P_x$, is called the annihilator or the residual maker. Then $\hat{u} = M_x y$. It has following properties: (1) M_x is symmetric (2) M_x is idempotent, i.e. $M_x M_x = M_x$ (3) $M_x \hat{u} = \hat{u}$ (4) $M_x P_x = P_x M_x = 0$.

Also, for any vector $a \in \mathbb{R}^n$, $\|P_x a\| \le \|a\|, \|M_x a\| \le \|a\|$.

Theorem 3.3 (Frisch-Waugh-Lovell)**.** Suppose that $x_{i1} \in \mathbb{R}^{d_1}$, $x_{i2} \in \mathbb{R}^{d_2}$ for every i. And

$$y_i = x_{i1}^\top \beta_1 + x_{i2}^\top \beta_2 + u_i \tag{41}$$

The estimated coefficient of $\mathbf{x}_{i2}$ in the regression of y_i on $\mathbf{x}_{i1}$ and $\mathbf{x}_{i2}$ is given by

$$\hat{\beta}_2 = (X_2^\top M_{X_1} X_2)^{-1}(X_2^\top M_{X_1} y) = (\breve{X}_2^\top \breve{X}_2)^{-1}(\breve{X}_2^\top \breve{y}) \tag{42}$$

where $\breve{X}_2$ and $\breve{y}$ are the residuals obtained from regression x_{i2} and y_i on x_{i1}, respectively.

Proof. Consider

$$y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{u}$$

Then pre-multiply M_{X_1} at BHS,

$$M_{X_1} y = 0 + M_{X_1} X_2 \hat{\beta}_2 + M_{X_1}\hat{u} \Rightarrow M_{X_1} y - M_{X_1} X_2 \hat{\beta}_2 = \hat{u}$$

Then, multiply $X_2^\top$ at BHS

$$X_2^\top M_{X_1} y - X_2^\top M_{X_1} X_2 \hat{\beta}_2 = X_2^\top \hat{u} = 0 \Rightarrow \hat{\beta}_2 = (X_2^\top M_{X_1} X_2)^{-1}(X_2^\top M_{X_1} y)$$

□

A simple, but useful application of this is, consider $x_i = (z_i, 1)$. Then run the regression, what will we get?

3.2 Large Sample Properties

Note that

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_i y_i\right) = \beta + \left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_i u_i\right) \tag{43}$$

We first make an assumption on the data generating process.

(i) x_i has a finite second moments, $\mathbb{E}[|x_i|^2] < \infty$. And the matrix $\mathbb{E}[x_i x_i^\top]$ is invertible.

(ii) The error term u_i is mean-zero, $\mathbb{E}[u_i] = 0$. And it has finite variance, also uncorrelated with x: $\mathbb{E}[x_i u_i] = 0$

Proposition 3.4. Under those assumptions:

$$\frac{1}{n}\sum_{i=1}^n x_i x_i^\top \xrightarrow{p} \mathbb{E}[x_i x_i^\top], \quad \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top \text{ is non-singular}\right) \to 1 \tag{44}$$

Also,

$$\frac{1}{n}\sum_{i=1}^n x_i u_i \xrightarrow{p} 0$$

Hence

$$\hat{\beta} \xrightarrow{p} \beta, \quad \hat{\beta} = \beta + o_p(1) \tag{45}$$

If we further assume x_i and u_i has finite fourth moments, and $\mathbb{E}[x_i x_i^\top u_i^2]$ is non-singular, then

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V), \quad V = (\mathbb{E}[x_i x_i])^{-1}(\mathbb{E}[x_i x_i^\top u_i^2])(\mathbb{E}[x_i x_i])^{-1} \tag{46}$$

3.3 Standard Error

The challenge is to estimate $\mathbb{E}[x_i x_i^\top u_i^2]$, since it involves unknown u_i. Now consider

$$\frac{1}{n}\sum_{i=1}^n x_i x_i^\top \hat{u}_i^2, \quad \hat{u}_i = y_i - x_i^\top \beta$$

We can not apply LLN here since $\hat{u}_i$ are not i.i.d. Hence we have to decompose it first.

$$\hat{u}_i^2 - u_i^2 = (\hat{u}_i - u_i)(\hat{u}_i + u_i), \quad \hat{u}_i - u_i = x_i^\top(\beta - \hat{\beta}) \Rightarrow \hat{u}_i^2 = u_i^2 + x_i^\top(\beta - \hat{\beta})(y_i - x_i^\top\hat{\beta} + u_i)$$

Hence

$$\frac{1}{n}\sum_{i=1}^n x_i x_i^\top \hat{u}_i^2 = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top u_i^2 + \frac{1}{n}\sum_{i=1}^n x_i x_i^\top (y_i + u_i) x_i^\top(\beta - \hat{\beta}) - \frac{1}{n}\sum_{i=1}^n x_i x_i^\top x_i^\top \hat{\beta} x_i^\top(\beta - \hat{\beta}) \tag{47}$$

Then we know that

$$\frac{1}{n}\sum_{i=1}^n x_i x_i^\top u_i^2 \xrightarrow{p} \mathbb{E}[x_i x_i^\top u_i^2], \quad \left\|\frac{1}{n}\sum_{i=1}^n x_i x_i^\top (y_i + u_i) x_i^\top(\beta - \hat{\beta})\right\| \le \|\beta - \hat{\beta}\| \left\|\frac{1}{n}\sum_{i=1}^n x_i x_i^\top (y_i + u_i)\right\| \tag{48}$$

Since $\beta - \hat{\beta} \xrightarrow{p} 0$, and $\frac{1}{n}\sum_{i=1}^n x_i x_i^\top (y_i + u_i) \xrightarrow{p} \mathbb{E}[x_i x_i^\top x_i^\top (y_i + u_i)]$. Hence, by $o_p(1)O_p(1) = o_p(1)$, we know that

$$\left\|\frac{1}{n}\sum_{i=1}^n x_i x_i^\top (y_i + u_i) x_i^\top(\beta - \hat{\beta})\right\| \xrightarrow{p} 0 \tag{49}$$

Moreover, impose $\beta - \hat{\beta} = O_p(\frac{1}{\sqrt{n}})$, we have

$$\left\|\frac{1}{n}\sum_{i=1}^n x_i x_i^\top (y_i + u_i) x_i^\top(\beta - \hat{\beta})\right\| = O_p(\frac{1}{\sqrt{n}}) = o_p(1) \tag{50}$$

Similarly,

$$\left\|\frac{1}{n}\sum_{i=1}^n x_i x_i^\top x_i^\top \hat{\beta} x_i^\top(\beta - \hat{\beta})\right\| \le \|\beta - \hat{\beta}\|\|\hat{\beta}\|\frac{1}{n}\sum_{i=1}^n \|x_i\|^4 = O_p(\frac{1}{\sqrt{n}}) = o_p(1)$$

Hence

$$\frac{1}{n}\sum_{i=1}^n x_i x_i^\top \hat{u}_i^2 \xrightarrow{p} \mathbb{E}[x_i x_i^\top u_i^2]$$

Then

$$\hat{V} = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)^{-1}\left[\frac{1}{n}\sum_{i=1}^n x_i x_i^\top \hat{u}_i^2\right]\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top\right)^{-1} \xrightarrow{p} V \tag{51}$$

Sometimes we also write

$$\hat{\beta} \overset{a}{\sim} N(\beta, \frac{\hat{V}}{n}) \tag{52}$$

The standard error (or, more precisely, the variance estimator) we discussed above is widely known as the Huber-Eicker-White standard error. We also call it HC0, where HC stands for "heteroskedasticity consistent. `robust` in `Stata` corresponds to the original HC0 standard error. Compared to HC0, HC1 incorporates a degrees of freedom adjustment of $n/(n-d)$.

$$\begin{aligned}
\hat{V}_{HC1} &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)^{-1} \left[\frac{1}{n-d}\sum_{i=1}^{n} x_i x_i^\top \hat{u}_i^2\right] \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)^{-1} = \frac{n}{n-d}\hat{V} \\
\hat{V}_{HC2} &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top \frac{\hat{u}_i^2}{1-p_{ii}}\right] \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)^{-1}, \quad p_{ii} = x_i^\top (X'X)^{-1} x_i, \\
\hat{V}_{HC3} &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top \frac{\hat{u}_i^2}{(1-p_{ii})^2}\right] \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)^{-1} \\
\hat{V}_{HC4} &= \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top \frac{\hat{u}_i^2}{(1-p_{ii})^{\delta_i}}\right] \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top\right)^{-1}, \quad \delta_i = \min\left\{4, \frac{np_{ii}}{d}\right\}
\end{aligned} \tag{53}$$

There are a lot of excercises to be done here

Exercise. Prove the Gauss-Markov Theorem

3.4 Hypothesis Testing

Consider general problem

$$H_0 : R\beta = r \tag{54}$$

Note that $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V)$. Then under the null

$$\sqrt{n}(R\hat{\beta} - R\beta) = \sqrt{n}(R\hat{\beta} - r) \xrightarrow{d} N(0, RVR')$$

We consider the Wald's statistic for this test:

$$n(R\hat{\beta} - r)'[RVR']^{-1}(R\hat{\beta} - r) \xrightarrow{d} \chi_k^2$$

4 Instrumental Variables

Previously, we assume that the covariates are uncorrelated with the error term. In some cases, however, it is not possible to control for certain information, such as intellectual ability, and hence the correlation $\text{Cov}(x_i, u_i) \neq 0$ cannot be removed. Instead, we assume there is another variable z_i, known as the instrument, such that it is uncorrelated with the error term: $\text{Cov}(z_i, u_i) = 0$.

Assumption (Linear IV model). Let z_i be an instrumental variable. Then (i) $\text{Cov}(z_i, u_i) = 0$ and (ii) $\text{Cov}(z_i, x_i) \neq 0$.

Still, we motivate estimation by moment conditions:

$$\begin{aligned}
0 &= \mathbb{E}[z_i u_i] = \mathbb{E}[z_i(y_i - \beta_0 - \beta_1 x_i)] \\
0 &= \mathbb{E}[u_i] = \mathbb{E}[y_i - \beta_0 - \beta_1 x_i]
\end{aligned}$$

Then, for estimation, we consider their sample analogues

$$\begin{aligned}
0 &= \frac{1}{n}\sum_{i=1}^{n} z_i(y_i - \beta_0 - \beta_1 x_i) \\
0 &= \frac{1}{n}\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)
\end{aligned}$$

The solution to those systems is

$$\hat{\beta}_1 = \frac{\overline{zy} - \bar{z}\cdot\bar{y}}{\overline{zx} - \bar{z}\cdot\bar{x}} \tag{55}$$

where $\bar{x}$ is sample averages. Actually it can be written as a ratio of two regression estimates:

$$\hat{\beta}_1 = \left(\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(z_i - \bar{z})}{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2}\right) \Big/ \left(\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})}{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2}\right) \tag{56}$$

That is, to obtain $\hat{\beta}_1$, we can first regress y_i on the instrument (and an intercept), then regress the endogenous variable x_i on the instrument (and an intercept), and finally take the ratio of the two slope estimates. Consider the linear projection:

$$x_i = \pi_0 + z_i\pi_1 + v_i$$

By construction, the error term v_i will be uncorrelated with the instrument. Then

$$y_i = \beta_0 + \beta_1\pi_0 + z_i\beta_1\pi_1 + u_i + \beta_1 v_i$$

Note that the product $\beta_1\pi_1$ can be consistently estimated by a regression of y_i on the instrument z_i. This regression is known as the reduced-form regression for y_i. π_1 can also be consistently estimated by regression x_i on the instrument,

this regression is known as the reduced-form regression for x_i, or the first-stage regression.

From the first regression, we can save the predicted $\hat{x}_i = \hat{\pi}_0 + z_i\hat{\pi}_1$. Then we can run a second stage regression of y_i on predicted values $\hat{x}_i$, and we will obtain exactly the same estimate $\hat{\beta}_1$.

$$y_i = \beta_0 + x_i\beta_1 + u_i = \beta_0 + (\pi_0 + z_i\pi_1)\beta_1 + u_i + \beta_1 v_i$$

which means a consistent estimate of β_1 can be obtained by regressing y_i on $\pi_0 + z_i\pi_1$. Then one can write done the second stage regression

$$\tilde{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})(\hat{x}_i - \bar{\hat{x}})}{\frac{1}{n}\sum_{i=1}^n (\hat{x}_i - \bar{\hat{x}})^2}$$

Note that $\hat{x}_i - \bar{\hat{x}} = (z_i - \bar{z})\hat{\pi}_1$. Then one can prove that $\tilde{\beta}_1 = \hat{\beta}_1$.

Loosely speaking, a control function is a random quantity, which can be added to the model to mitigate the endogeneity issue. To discuss the control function perspective, we start from the structural equation and the first stage:

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad x_i = \pi_0 + \pi_1 z_i + v_i$$

Then we consider a linear projection of the error terms, and in particular, we write

$$u_i = \theta_1 v_i + \varepsilon_i$$

By construction, ε_i is uncorrelated with v_i. Moreover, $\varepsilon_i = u_i - \theta_1 v_i$ is uncorrelated with z_i. Hence ε_i is uncorrelated with x_i. Intuitively, $\theta_1 v_i$ is the part of u_i that is correlated to x_i, and ε_i is the part that uncorrelate with x_i. If we write

$$y_i = \beta_0 + \beta_1 x_i + \theta_1 v_i + \varepsilon_i$$

then v_i will be a valid control function. Including this extra regressor will solve the endogeneity issue. For estimation, the control function approach suggests to estimate the first stage, and save the residuals $\hat{v}_i = x_i - \hat{\pi}_0 - \hat{\pi}_1 z_i$. Then in the next step, we run a regression of y_i on x_i and the first stage residual $\hat{v}_i$.

Exercise. Show that the control function approach leads to the same estimate. That is, assume we obtained the following fitted regression equation:

$$\breve{\beta}_0 + \breve{\beta}_1 x_i + \breve{\theta}_1 \hat{v}_1$$

Show that $\breve{\beta}_1 = \hat{\beta}_1$.

Proof. Note that by partition regression formula, $\breve{\beta}_1$ is obtained by regressing y_i on $\tilde{x}_i$, where $\tilde{x}_i$ is the residual obtained from regressing x_i on $\hat{v}_i$. Note that

$$x_i = \hat{\pi}_0 + \hat{\pi}_1 z_i + \hat{v}_i$$

Also

$$x_i = \hat{\gamma}_0 + \hat{\gamma}_1 \hat{v}_i + \tilde{x}_i$$

Then it must be $\hat{\gamma}_0 = \hat{\pi}_0 + \hat{\pi}_1\bar{z}$ and $\hat{\gamma}_1 = 1$. As a result, $\tilde{x}_i = \hat{\pi}_1(z_i - \bar{z})$. Since those sample mean are all 0, then regress y_i on $\tilde{x}_i$ will give us

$$\breve{\beta}_1 = \frac{\sum_{i=1}^n y_i(z_i - \bar{z})\hat{\pi}_1}{\sum_{i=1}^n (z_i - \bar{z})^2\hat{\pi}_1^2} = \frac{\sum_{i=1}^n y_i(z_i - \bar{z})}{\sum_{i=1}^n (z_i - \bar{z})^2} \cdot \frac{1}{\hat{\pi}_1} = \hat{\beta}_1$$

□

4.1 Local Average Treatment Effect

We can re-write $y_i = x_i y_i(1) + (1 - x_i)y_i(0)$ as

$$y_i = y_i(0) + \tau_i x_i, \quad \tau_i = y_i(1) - y_i(0)$$

where τ_i is the individual treatment effect. Also $y_i(1) = \mathbb{E}[y_i(1)] + u_i(1)$ and $y_i(0) = \mathbb{E}[y_i(0)] + u_i(0)$, then

$$y_i = \underbrace{\mathbb{E}[y_i(0)]}_{\beta_0} + x_i \underbrace{\tau_{ATE}}_{\beta_1} + \underbrace{u_i(0) + (u_i(1) - u_i(0))x_i}_{u_i}$$

Let z_i be some instrument. If x_i is independent of $(u_i(1), u_i(0))^\top$, then x_i will be uncorrelated with u_i, and hence the ATE can be identified. But in the second case, x_i is correlated with $(u_i(0), u_i(1))$.

Note that even when an instrument is exogenously determined, the 2SLS may not identify the ATE.

(Assumption) Potential outcomes IV model: Let z_i be the binary instrument, $x_i(1)$ and $x_i(0)$ be the two potential treatments, and $y_i(1)$ and $y_i(0)$ be the two potential outcomes.

1. z_i is independent of the potential treatments and the potential outcomets:
$$z_i \perp\!\!\!\perp (x_i(1), x_i(0), y_i(1), y_i(0))^\top$$
2. The instrument is relevant
$$\mathbb{P}(x_i(1) = 1) \neq \mathbb{P}(x_i(0) = 1)$$
3. Either $x_i(1) \geq x_i(0)$ for all individuals or $x_i(1) \leq x_i(0)$ for all individuals (this is known as the monotonicity assumption).

The second assumption is actually

$$\mathrm{Cov}(x_i, z_i) = \mathbb{P}(x_i z_i = 1) - \mathbb{P}(x_i = 1)\mathbb{P}(z_i = 1) = (\mathbb{P}[x_i = 1|z_i = 1] - \mathbb{P}(z_i = 1))\mathbb{P}(z_i = 1)$$

Note that $\mathbb{P}(x_i = 1|z_i = 1) = \mathbb{P}(x_i(1) = 1|z_i = 1) = \mathbb{P}(x_i(1) = 1)$. Similarly,

$$\mathbb{P}(x_i = 1) = \mathbb{P}(x_i(1) = 1)\mathbb{P}(z_i = 1) + \mathbb{P}(x_i(0) = 1)\mathbb{P}(z_i = 0)$$

Hence

$$\text{Cov}(x_i, z_i) = (\mathbb{P}[x_i(1) = 1] - \mathbb{P}[x_i(1) = 1]\mathbb{P}[z_i = 1] - \mathbb{P}[x_i(0) = 1]\mathbb{P}[z_i = 0])\,\mathbb{P}[z_i = 1]$$

This is not zero as long as $\mathbb{P}(x_i(1) = 1) \neq \mathbb{P}(x_i(0) = 1)$ and $0 < \mathbb{P}(z_i = 1) < 1$.

In the following we will discuss what the 2SLS identifies when both the endogenous variable and the instrument are binary:

$$x_i = x_i(1)z_i + x_i(0)(1 - z_i)$$

Without loss of generality, we will assume that $x_i(1) \geq x_i(0)$, which means the instrument will encourage treatment take-up. Recall that the slope estimate from the 2SLS takes the form. We have

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})(z_i - \bar{z})}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(z_i - \bar{z})} \xrightarrow{p} \frac{\text{Cov}(y_i, z_i)}{\text{Cov}(x_i, z_i)}$$

Theorem 4.1 (2SLS identification; local average treatment effect)**.** Consider the setting of assumptions above with $x_i(1) \geq x_i(0)$. Then the 2SLS slope estimator is consistent for

$$\hat{\beta}_1 \xrightarrow{p} \mathbb{E}[y_i(1) - y_i(0)|x_i(1) > x_i(0)]$$

which is the treatment effect for the subpopulation with $x_i(1) > x_i(0)$. This is also known as the local average treatment effect (LATE).

4.2 Large-Sample Properties of the 2SLS and Weak Instruments

We start from two equations:

$$y_i = \beta_0 + x_i\beta_1 + u_i, \quad x_i = \pi_0 + z_i\pi_1 + v_i$$

Hence

$$y_i = \underbrace{\beta_0 + \beta_1\pi_0}_{\gamma_0} + z_i \underbrace{\beta_1\pi_1}_{\gamma_1} + \underbrace{u_i + \beta_1 v_i}_{\eta_i}$$

We can re-write the 2SLS estimator as

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_1 - \beta_1) &= \sqrt{n}\left(\frac{\hat{\gamma}_1}{\hat{\pi}_1} - \frac{\gamma_1}{\pi_1}\right) = \frac{1}{\hat{\pi}_1}\sqrt{n}(\hat{\gamma}_1 - \gamma_1) - \frac{\gamma_1}{\pi_1\hat{\pi}_1}\sqrt{n}(\hat{\pi}_1 - \pi_1) \\
&= \frac{1}{\hat{\pi}_1}\sqrt{n}(\hat{\gamma}_1 - \gamma_1 - \beta_1(\hat{\pi}_1 - \pi_1)) \\
&= \frac{1}{\pi_1}\sqrt{n}(\hat{\gamma}_1 - \gamma_1 - \beta_1(\hat{\pi}_1 - \pi_1)) + o_p(1)
\end{aligned}$$

where the last step follows from $\pi_1 \neq 0, |\hat{\pi}_1 - \pi_1| = o_p(1)$ and that $\sqrt{n}(\hat{\gamma}_1 - \gamma_1 - \beta_1(\hat{\pi}_1 - \pi_1) = O_p(1)$

To be completed in the future

5 Causality

5.1 Model