Empirical Economics Cheat-Sheet

Asymptotics

Modes of Convergence

Let $\{X_n\}_{n\geq 1}$ and X be random variables on $(\Omega, \mathcal{F}, P)$.

Almost Sure Convergence: $X_n \xrightarrow{a.s.} X$ if $P(\{\omega : X_n(\omega) \to X(\omega)\}) = 1$.

Convergence in Probability: $X_n \xrightarrow{p} X$ if $\forall \epsilon > 0$: $P(|X_n - X| > \epsilon) \to 0$.

Convergence in r-th Mean: $X_n \xrightarrow{r} X$ if $\mathbb{E}(|X_n - X|^r) \to 0$.

Convergence in Distribution: $X_n \xrightarrow{d} X$ if $F_{X_n}(x) \to F_X(x)$ for all x where F_X is continuous.

Implications between modes

$$X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$$
$$X_n \xrightarrow{r} X \implies X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$$
$$X_n \xrightarrow{r} X \implies X_n \xrightarrow{s} X \text{ for } s \leq r$$

Note that none of the reverse implications hold in general. The one exception is that $X_n \xrightarrow{d} c$ (a constant) implies $X_n \xrightarrow{p} c$.

Continuous Mapping Theorem

Let $g : \mathbb{R}^k \to \mathbb{R}^n$ be continuous on $S \subset \mathbb{R}^k$ with $P(X \in S) = 1$. Then:

$$X_n \xrightarrow{a.s.} X \implies g(X_n) \xrightarrow{a.s.} g(X)$$
$$X_n \xrightarrow{p} X \implies g(X_n) \xrightarrow{p} g(X)$$
$$X_n \xrightarrow{d} X \implies g(X_n) \xrightarrow{d} g(X)$$

This does not hold for convergence in r-th mean. To see this, take $X_n = n$ w.p. $1/n^2$ and 0 otherwise, with $g(x) = x^2$: then $\mathbb{E}|X_n| \to 0$ but $\mathbb{E}|X_n^2| = 1$.
Notice that we need $P(X \in S) = 1$. For instance, $g(x, y) = x/y$ is continuous on $S = \mathbb{R}^2 \setminus \{(x, 0)\}$; we need $c \neq 0$ for $X_n/Y_n \xrightarrow{d} X/c$.

Slutsky's Theorem

If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ (constant), then:

$$X_n + Y_n \xrightarrow{d} X + c$$
$$X_n Y_n \xrightarrow{d} cX$$
$$X_n/Y_n \xrightarrow{d} X/c \quad (c \neq 0)$$

It is essential that Y_n converges to a *constant*. If $Y_n \xrightarrow{d} Y$ (non-degenerate), Slutsky does not apply.

Weak Law of Large Numbers

If $\{X_i\}_{i\geq 1}$ iid with $\mathbb{E}(X_i) = \mu$, $\mathrm{Var}(X_i) = \sigma^2 < \infty$, then $\bar{X}_n \xrightarrow{p} \mu$. This follows from Chebyshev's inequality: $P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$.

Central Limit Theorem

If $\{X_i\}_{i\geq 1}$ iid with $\mathbb{E}(X_i) = \mu$, $\mathrm{Var}(X_i) = \Sigma$ (finite), then:

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \Sigma).$$

Stochastic Order Notation

$X_n = O_p(1)$: $\{X_n\}$ is bounded in probability, i.e. $\forall \epsilon > 0$, $\exists M$ s.t. $\sup_n P(|X_n| > M) < \epsilon$.
$X_n = o_p(1)$: $X_n \xrightarrow{p} 0$.
The key composition rules are: $O_p(1) \cdot o_p(1) = o_p(1)$; $O_p(1) + O_p(1) = O_p(1)$.
Convergence in distribution implies boundedness: $X_n \xrightarrow{d} X \implies X_n = O_p(1)$.
If $\sqrt{n}(X_n - c) \xrightarrow{d} X$, then $X_n \xrightarrow{p} c$ and $\sqrt{n}(X_n - c) = O_p(1)$.

Delta Method

Delta Method (General)

Let $\{X_n\}_{n\geq 1}$ be $(K \times 1)$ random vectors with $n^r(X_n - c) \xrightarrow{d} X$ for some $r > 0$ and constant c. Let $g : \mathbb{R}^k \to \mathbb{R}^d$ be differentiable at c with Jacobian $Dg(c)$. Then:

$$n^r(g(X_n) - g(c)) \xrightarrow{d} Dg(c)\, X.$$

If $X \sim N(0, \Sigma)$:

$$n^r(g(X_n) - g(c)) \xrightarrow{d} N\big(0,\, Dg(c)\,\Sigma\, Dg(c)'\big)\,.$$

Proof Sketch

By Taylor: $g(x) = g(c) + Dg(c)(x - c) + h_1(x)(x - c)$ with $h_1(c) = 0$. Then:

$$n^r(g(X_n) - g(c)) = Dg(c)\, n^r(X_n - c) + h_1(X_n)\, n^r(X_n - c).$$

Since $X_n \xrightarrow{p} c$, CMT gives $h_1(X_n) \xrightarrow{p} 0$, and since $n^r(X_n - c) = O_p(1)$, the remainder is $o_p(1) \cdot O_p(1) = o_p(1)$.

Second-Order Delta Method

If $g'(c) = 0$ and $g''(c)$ exists (scalar case), then:

$$n^{2r}(g(X_n) - g(c)) \xrightarrow{d} \frac{g''(c)}{2} X^2.$$

Use when first-order term vanishes (degenerate limit).

Application: Sample Variance

Let X_i iid with $\mathbb{E}(X_i) = \mu$, $\mathrm{Var}(X_i) = \sigma^2$, $\mathbb{E}(X_i - \mu)^4 = \kappa$. Let $S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$. Using the Delta Method on $g(\mu, m_2) = m_2 - \mu^2$ applied to the sample moments $(\bar{X}_n, \frac{1}{n}\sum X_i^2)$:

$$\sqrt{n}(S_n^2 - \sigma^2) \xrightarrow{d} N(0, \kappa - \sigma^4).$$

Estimation

Definitions

Given a sample $\{X_i\}_{i=1}^n$ from distribution F, a **statistic** is a function $T_n : (X_1, \ldots, X_n) \to V$. An **estimator** is a statistic used to learn about some feature $\theta(F)$.

Finite Sample Properties

The **bias** of $\hat{\theta}_n$ is $\mathrm{Bias}(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta$. We say $\hat{\theta}_n$ is **unbiased** if $\mathbb{E}(\hat{\theta}_n) = \theta$.
Mean Squared Error:

$$\mathrm{MSE}(\hat{\theta}_n) = \mathbb{E}\Big[(\hat{\theta}_n - \theta)^2\Big] = \mathrm{Var}(\hat{\theta}_n) + \mathrm{Bias}(\hat{\theta}_n)^2.$$

Large Sample Properties

We say $\hat{\theta}_n$ is **consistent** if $\hat{\theta}_n \xrightarrow{p} \theta$ (or $\hat{\theta}_n \xrightarrow{a.s.} \theta$).
We say $\hat{\theta}_n$ is **asymptotically normal** if $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, V)$ for some V.
An estimator is **asymptotically efficient** if it achieves the smallest possible asymptotic variance among regular estimators (e.g. MLE under regularity conditions).

Method of Moments

The **sample analogue principle** replaces population moments with their sample counterparts.
If θ satisfies $\mathbb{E}(m(X, \theta)) = 0$ for moment function m, the MoM estimator solves:

$$\frac{1}{n}\sum_{i=1}^n m(X_i, \hat{\theta}_n) = 0.$$

Consistency follows from SLLN + CMT if identification holds.

Maximum Likelihood Estimation

Setup

Let $\{X_i\}_{i=1}^n$ iid with density f_{θ_0} for some $\theta_0 \in \Theta \subset \mathbb{R}^d$.
Likelihood: $\ell_n(\theta) = \prod_{i=1}^n f_\theta(X_i)$.
Log-likelihood: $L_n(\theta) = \frac{1}{n}\sum_{i=1}^n \ln f_\theta(X_i)$.
MLE: $\hat{\theta}_n \in \arg\max_{\theta\in\Theta} L_n(\theta)$.

Why MLE is Consistent

Under regularity conditions, $L_n(\theta) \xrightarrow{p} L(\theta) := \mathbb{E}(\ln f_\theta(X))$.
Key insight: θ_0 uniquely maximizes $L(\theta)$.
Proof: Let $M(\theta) = L(\theta) - L(\theta_0) = \mathbb{E}\Big[\ln \frac{f_\theta(X)}{f_{\theta_0}(X)}\Big]$. By Jensen's inequality:

$$M(\theta) \leq \ln \mathbb{E}\left[\frac{f_\theta(X)}{f_{\theta_0}(X)}\right] = \ln \int \frac{f_\theta(x)}{f_{\theta_0}(x)} f_{\theta_0}(x)\, dx = \ln 1 = 0,$$

with equality iff $f_\theta(X)/f_{\theta_0}(X) = c$ a.s. Ruled out for $\theta \neq \theta_0$ by assumption.

Asymptotic Distribution

Under regularity conditions:

$$\sqrt{n}(\hat{\theta}_{\mathrm{MLE}} - \theta_0) \xrightarrow{d} N\big(0,\ I(\theta_0)^{-1}\big)\,,$$

where $I(\theta_0) = -\mathbb{E}\Big[\frac{\partial^2}{\partial\theta^2} \ln f_{\theta_0}(Y|X)\Big]$ is the **Fisher Information**.
This variance is optimal: no regular estimator can achieve a smaller asymptotic variance.

Example: Normal

$X_i \sim N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$ unknown.

$$L_n(\theta) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln\sigma^2 - \frac{1}{2n\sigma^2}\sum_{i=1}^n (X_i - \mu)^2.$$

FOCs yield $\hat{\mu} = \bar{X}_n$, $\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \bar{X}_n)^2$.

Example: Bernoulli

$X_i \sim \text{Bernoulli}(\theta)$. Log-likelihood:
$L_n = \bar{X}_n \ln\theta + (1 - \bar{X}_n)\ln(1-\theta)$. FOC: $\hat{\theta} = \bar{X}_n$. Log-likelihood is concave, so FOC suffices.

Conditional MLE

If $Y|X$ has conditional density $f_\theta(y|x)$: $\ell_n(\theta) = \prod_i f_\theta(Y_i|X_i)$. Maximising this is equivalent to OLS when $Y|X \sim N(X'\beta, \sigma^2)$.

Normal Regression as CMLE

If $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$, CMLE of (β_0, β_1) minimizes $\sum(Y_i - \beta_0 - \beta_1 X_i)^2$. These are the OLS estimators.

OLS: Setup & Projections

Linear Model

$$y_i = x_i'\beta + u_i, \quad \mathbb{E}(x_i u_i) = 0.$$

Here $x_i \in \mathbb{R}^{k+1}$ with $x_{i0} = 1$. We say the model is "linear" because it is linear in the parameters β_j. The error u_i captures all unobserved determinants of y_i.

Identification

Assume $\mathbb{E}(xu) = 0$ and no perfect collinearity (no $a \neq 0$ with $P(a'x = 0) = 1$).
$\mathbb{E}(xx')$ invertible $\iff$ no perfect collinearity. Then:

$$\beta = \mathbb{E}(xx')^{-1}\mathbb{E}(xy).$$

Proof ($\mathbb{E}(xx')$ invertible $\iff$ no collinearity):
($\Rightarrow$) If $P(x'a = 0) = 1$ for $a \neq 0$, then $\mathbb{E}(xx')a = \mathbb{E}(x \cdot x'a) = 0$, not invertible.
($\Leftarrow$) No collinearity $\implies c'\mathbb{E}(xx')c = \mathbb{E}[(x'c)^2] > 0\ \forall c \neq 0$, so $\mathbb{E}(xx')$ is positive definite.

OLS Estimator

Given iid sample $\{y_i, x_i\}_{i=1}^n$. Unique OLS estimator (when $X'X$ invertible):

$$\hat{\beta}_n = \left(\frac{1}{n}\sum x_i x_i'\right)^{-1} \frac{1}{n}\sum x_i y_i = (X'X)^{-1}X'Y.$$

Equivalently, $\hat{\beta}_n$ solves $\min_b \|Y - Xb\|^2$. The FOC gives $X'\hat{U} = 0$.

Projection Matrix

$P_X = X(X'X)^{-1}X'$: projects onto column space $\mathcal{S}(X)$.
$M_X = I_n - P_X$: residual maker.
Properties:

- $P_X = P_X'$, $M_X = M_X'$ (symmetric)
- $P_X^2 = P_X$, $M_X^2 = M_X$ (idempotent)
- $P_X M_X = M_X P_X = 0$
- $P_X X = X$, $M_X X = 0$
- For any Y: $Y = P_X Y + M_X Y = \hat{Y} + \hat{U}$

Projection Theorem

Let $\mathcal{S}$ be a nonempty subspace of $\mathbb{R}^n$. There exists a unique $\hat{y} \in \mathcal{S}$ minimizing $\|y - \hat{y}\|$. Necessary and sufficient: $y - \hat{y}$ is orthogonal to every vector in $\mathcal{S}$.
Applying to $\mathcal{S} = \mathcal{S}(X)$: the condition $X'(Y - \hat{Y}) = 0$ yields $\hat{Y} = P_X Y$.

Frisch-Waugh-Lovell

Partition $Y = X_1\beta_1 + X_2\beta_2 + U$. Then:

$$\hat{\beta}_2 = (X_2' M_{X_1} X_2)^{-1} X_2' M_{X_1} Y.$$

That is, $\hat{\beta}_2$ is obtained by regressing the residuals of Y on X_1 onto the residuals of X_2 on X_1.
Proof: $M_{X_1}Y = M_{X_1}X_2\hat{\beta}_2 + \hat{U}$, multiply by X_2':
$X_2' M_{X_1} Y = X_2' M_{X_1} X_2 \hat{\beta}_2$ since $X'\hat{U} = 0$.
Population version: $\beta_2 = \mathbb{E}(\tilde{x}_2\tilde{x}_2')^{-1}\mathbb{E}(\tilde{x}_2 y)$, where $\tilde{x}_2 = x_2 - \tilde{\gamma}x_1$ is the residual from projecting x_2 onto x_1. This holds because $\mathbb{E}(\tilde{x}_2 x_1') = 0$.

Omitted Variables Bias

If $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ but we regress y on x_1 only:

$$b_1 = \beta_1 + \beta_2 \frac{\text{Cov}(x_1, x_2)}{\text{Var}(x_1)}.$$

The bias term $\beta_2 \frac{\text{Cov}(x_1,x_2)}{\text{Var}(x_1)}$ is the effect of the omitted variable $\times$ correlation with included variable.

OLS: Finite Sample Properties

Assumptions

$Y = X\beta + U$ with $\mathbb{E}(U|X) = 0$ (equivalently $\mathbb{E}(Y|X) = X\beta$). Since (u_i, x_i) independent of x_j for $j \neq i$: $\mathbb{E}(u_i|x_1, \ldots, x_n) = 0$.

Unbiasedness

$$\mathbb{E}(\hat{\beta}_n|X) = (X'X)^{-1}X'\mathbb{E}(Y|X) = (X'X)^{-1}X'X\beta = \beta.$$

By LIE: $\mathbb{E}(\hat{\beta}_n) = \mathbb{E}[\mathbb{E}(\hat{\beta}_n|X)] = \beta$.

Variance under Homoskedasticity

Assume $\text{Var}(u_i|x_i) = \sigma^2$ (homoskedastic). Then $\text{Var}(U|X) = \sigma^2 I_n$ and:

$$\text{Var}(\hat{\beta}_n|X) = \sigma^2(X'X)^{-1}.$$

Variance under Heteroskedasticity

If $\mathbb{E}(u_i^2|x_i) = \sigma^2(x_i)$, then $\text{Var}(U|X) = \Omega$ (diagonal, varying entries):

$$\text{Var}(\hat{\beta}_n|X) = (X'X)^{-1}X'\Omega X(X'X)^{-1}.$$

Gauss-Markov Theorem

Under $\mathbb{E}(U|X) = 0$ and homoskedasticity, OLS is BLUE: for any linear unbiased estimator $\tilde{\beta} = AY$ with $AX = I_{k+1}$,

$$\text{Var}(\tilde{\beta}|X) - \text{Var}(\hat{\beta}_n|X) = \sigma^2 CC' \succeq 0,$$

where $C = A - (X'X)^{-1}X'$.
Proof: $\text{Var}(AY|X) = \sigma^2 AA'$. Let $C = A - (X'X)^{-1}X'$. Then $CX = AX - I_{k+1} = 0$, so:

$$AA' - (X'X)^{-1} = CC' + (X'X)^{-1}X'C' + CX(X'X)^{-1} = CC' \succeq 0.$$

Implication: For any $r \in \mathbb{R}^{k+1}$, $r'\hat{\beta}$ is BLUE for $r'\beta$:
$\text{Var}(r'\tilde{\beta}|X) - \text{Var}(r'\hat{\beta}|X) = r'CC'r \geq 0$.

Unbiasedness of $\hat{\sigma}^2$

Under normality, $\hat{\sigma}^2 = \frac{\text{SSR}}{n-k-1}$ is unbiased. **Proof (trace trick):**

$$\begin{aligned}\mathbb{E}[\text{SSR}|X] &= \mathbb{E}[U'M_X U|X] = \mathbb{E}[\text{tr}(U'M_X U)|X] \\ &= \mathbb{E}[\text{tr}(M_X UU')|X] = \text{tr}(M_X\,\mathbb{E}[UU'|X]) \\ &= \sigma^2\,\text{tr}(M_X) = \sigma^2(n-k-1),\end{aligned}$$

since $\text{tr}(M_X) = \text{tr}(I_n) - \text{tr}(P_X) = n - (k+1)$ (idempotent: $\text{tr}(P_X) = \text{tr}(X(X'X)^{-1}X') = \text{tr}(I_{k+1})$).

GLS (Known Heteroskedasticity)

If $\text{Var}(U|X) = \Omega$ with Ω known, pre-multiply by $\Omega^{-1/2}$:
$Y^* = X^*\beta + U^*$, where $\text{Var}(U^*|X) = I_n$.

$$\hat{\beta}_{\text{GLS}} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y.$$

This is OLS applied to the transformed model, hence BLUE by Gauss-Markov in the transformed space. Equivalently, it is BLUE in the original model.

Coefficient of Determination

$R^2 = 1 - \frac{\text{SSR}}{\text{TSS}} = 1 - \frac{\|M_X Y\|^2}{\|M_c Y\|^2} = \frac{\|P_X M_c Y\|^2}{\|M_c Y\|^2}$.

Where $\text{TSS} = \sum(y_i - \bar{y})^2$, $\text{SSR} = \sum \hat{u}_i^2$, $\text{ESS} = \sum(\hat{y}_i - \bar{y})^2$.
Note: TSS = ESS + SSR (so $0 \leq R^2 \leq 1$) holds only when the model includes an intercept; without one, R^2 may be negative.
Adjusted: $\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \cdot \frac{\text{SSR}}{\text{TSS}} \leq R^2$.
Population: $R^2_{\text{pop}} = 1 - \frac{\text{Var}(u)}{\text{Var}(y)}$.

Note that a high R^2 does not imply causality, and a low R^2 does not preclude it.

OLS: Large Sample Properties

Consistency

Under $y = x'\beta + u$, $\mathbb{E}(xu) = 0$, $\mathbb{E}(xx')$ invertible:

$$\hat{\beta}_n = \left(\frac{1}{n}\sum x_i x_i'\right)^{-1} \frac{1}{n}\sum x_i y_i \xrightarrow{a.s.} \mathbb{E}(xx')^{-1}\mathbb{E}(xy) = \beta.$$

by the SLLN and CMT.

Asymptotic Normality

Assume $\text{Var}(xu) = \mathbb{E}(u^2 xx')$ exists. Then:

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \Sigma),$$

where $\Sigma = \mathbb{E}(xx')^{-1}\mathbb{E}(u^2 xx')\mathbb{E}(xx')^{-1}$.

Proof: $\sqrt{n}(\hat{\beta}_n - \beta) = \left(\frac{1}{n}\sum x_i x_i'\right)^{-1} \frac{1}{\sqrt{n}}\sum x_i u_i$. CLT gives $\frac{1}{\sqrt{n}}\sum x_i u_i \xrightarrow{d} N(0, \mathbb{E}(u^2 xx'))$, then apply Slutsky.

Variance Estimation: Homoskedastic Case

Under $\mathbb{E}(u|x)=0$, $\text{Var}(u|x)=\sigma^2$: $\Sigma=\sigma^2\mathbb{E}(xx')^{-1}$. Estimate:

$$\hat\Sigma=\hat\sigma^2\left(\frac{1}{n}\sum x_ix_i'\right)^{-1},\quad \hat\sigma^2=\frac{1}{n}\sum\hat u_i^2.$$

Variance Estimation: Heteroskedastic Case

Without homoskedasticity, use the **Eicker-Huber-White** (robust) estimator:

$$\hat\Sigma=\left(\frac{1}{n}\sum x_ix_i'\right)^{-1}\left(\frac{1}{n}\sum\hat u_i^2x_ix_i'\right)\left(\frac{1}{n}\sum x_ix_i'\right)^{-1}.$$

Consistency of $\hat\Sigma$ (Key Proof Step)

Need $\frac{1}{n}\sum\hat u_i^2x_ix_i'\xrightarrow{p}\mathbb{E}(u^2xx')$. Decompose:

$$\frac{1}{n}\sum\hat u_i^2x_ix_i'=\frac{1}{n}\sum u_i^2x_ix_i'+\frac{1}{n}\sum(\hat u_i^2-u_i^2)x_ix_i'.$$

First term $\xrightarrow{a.s.}\mathbb{E}(u^2xx')$ by SLLN. Second term $=o_p(1)$ because:

$$|\hat u_i^2-u_i^2|=|x_i'(\beta-\hat\beta_n)|\cdot|\hat u_i+u_i|$$
$$\max_{i\le n}|\hat u_i^2-u_i^2|\le\|\beta-\hat\beta_n\|^2\max\|x_i\|^2+2\|\beta-\hat\beta_n\|\max\|x_iu_i\|.$$

Use the lemma: if $\mathbb{E}(\|Z_i\|^r)<\infty$ and Z_i identically distributed, then $\frac{\max_{i\le n}\|Z_i\|}{n^{1/r}}\xrightarrow{p}0$.

Applied: $\frac{\max\|x_i\|^2}{n}=o_p(1)$, $\frac{\max\|x_iu_i\|}{\sqrt n}=o_p(1)$, and $\sqrt n(\hat\beta-\beta)=O_p(1)$.

Hypothesis Testing

Definitions

The **null hypothesis** is $H_0:\theta_0\in\Theta_0$. It is **simple** if Θ_0 is a singleton, and **composite** otherwise.
A **test** is a function $\phi_n(X_1,\dots,X_n)\to\{0,1\}$; we reject H_0 iff $\phi_n=1$.
A **Type I error** occurs when we reject a true H_0; a **Type II error** when we fail to reject a false H_0.
The **power function** is $\beta_n(\theta)=P_\theta(\phi_n=1)$.
The **size** of the test is $\alpha:=\sup_{\theta\in\Theta_0}\beta_n(\theta)$.
A test has **asymptotic size** α if $\lim_{n\to\infty}\sup_{\theta\in\Theta_0}\beta_n(\theta)\le\alpha$.

Confidence Sets

C_n is a $1-\alpha$ confidence set if $P_\theta(\theta\in C_n)\ge1-\alpha$ for all θ.
Pivot: A function of data and unknown parameters whose distribution does not depend on unknown parameters (e.g. $\frac{\sqrt n(\bar X_n-\mu)}{\sigma}\sim N(0,1)$).
Exact (known σ^2): $C_n=\left[\bar X_n\pm\frac{\sigma}{\sqrt n}z_{1-\alpha/2}\right]$.
Exact (unknown σ^2, normal): $\frac{\sqrt n(\bar X_n-\mu)}{\tilde S_n}\sim t_{n-1}$, yielding $C_n=\left[\bar X_n\pm\frac{\tilde S_n}{\sqrt n}t_{n-1,1-\alpha/2}\right]$, where $\tilde S_n^2=\frac{1}{n-1}\sum(X_i-\bar X_n)^2$.

Finite Sample Inference (Normal Regression)

Under $Y|X\sim N(X\beta,\sigma^2I_n)$: $\hat\beta|X\sim N(\beta,\sigma^2(X'X)^{-1})$, $\frac{(n-k-1)\hat\sigma^2}{\sigma^2}\sim\chi^2_{n-k-1}$, and $\hat\beta\perp\hat\sigma^2|X$.
t-statistic: $T_n=\frac{\hat\beta_j-\beta_{j,0}}{se(\hat\beta_j)}\sim t_{n-k-1}$, where $se(\hat\beta_j)=\hat\sigma\sqrt{e_j'(X'X)^{-1}e_j}$.
We reject $H_0:\beta_j=\beta_{j,0}$ if $|T_n|>t_{n-k-1,1-\alpha/2}$.
p-value: $\hat p=2F(-|T_n|)$ where F is the t_{n-k-1} CDF.

Testing Single Linear Restriction (Asymptotic)

$H_0:r'\beta=c$. Under $\sqrt n(\hat\beta_n-\beta)\xrightarrow{d}N(0,V)$ with $\hat V_n\xrightarrow{p}V$:

$$T_n=\frac{\sqrt n(r'\hat\beta_n-c)}{\sqrt{r'\hat V_nr}}\xrightarrow{d}N(0,1)\text{ under }H_0.$$

Reject if $|T_n|>z_{1-\alpha/2}$. CI: $C_n=r'\hat\beta_n\pm z_{1-\alpha/2}\sqrt{r'\hat V_nr/n}$.

Testing Multiple Linear Restrictions

$H_0:R\beta=c$, R is $p\times(k+1)$ full row rank.

$$T_n=n\cdot(R\hat\beta_n-c)'(R\hat V_nR')^{-1}(R\hat\beta_n-c)\xrightarrow{d}\chi^2_p.$$

Reject if $T_n>\chi^2_{p,1-\alpha}$. Confidence set: ellipsoid $\{c:T_n(c)\le\chi^2_{p,1-\alpha}\}$.
RVR' positive definite because: if $a\ne0$, $R'a\ne0$ (full rank), so $(R'a)'V(R'a)>0$.

Testing Non-Linear Restrictions

$H_0:f(\beta)=0$, $f:\mathbb{R}^{k+1}\to\mathbb{R}^p$ continuously differentiable. Delta method:

$$\sqrt n(f(\hat\beta_n)-f(\beta))\xrightarrow{d}N(0,D_\beta f(\beta)\,V\,D_\beta f(\beta)').$$

Construct χ^2_p statistic as before. Note $f(\beta)=R\beta$ yields linear case since $D_\beta f=R$.

Potential Outcomes & Causality

Setup

Individual i has potential outcomes $y_i(1)$ (treated) and $y_i(0)$ (untreated). Treatment $D_i\in\{0,1\}$. Observed outcome:

$$Y_i=y_i(1)D_i+y_i(0)(1-D_i).$$

The **fundamental problem** of causal inference is that we never observe both $y_i(1)$ and $y_i(0)$.

Treatment Effects

ATE: $\mathbb{E}(y(1)-y(0))$.
ATT: $\mathbb{E}(y(1)-y(0)|D=1)$.
ATU: $\mathbb{E}(y(1)-y(0)|D=0)$.
Decomposition:

$$\text{ATE}=\text{ATT}\cdot P(D=1)+\text{ATU}\cdot P(D=0).$$

Naive Comparison and Selection Bias

$$\begin{aligned}&\mathbb{E}(Y|D=1)-\mathbb{E}(Y|D=0)\\&=\underbrace{\mathbb{E}(y(1)-y(0)|D=1)}_{\text{ATT}}+\underbrace{\mathbb{E}(y(0)|D=1)-\mathbb{E}(y(0)|D=0)}_{\text{Selection Bias}}.\end{aligned}$$

The naive comparison equals the ATT only when the selection bias vanishes.

Random Assignment

$D\perp(y(0),y(1))$ implies:

$$\mathbb{E}(y(d)|D)=\mathbb{E}(y(d))\quad\text{for }d\in\{0,1\}.$$

Now, selection bias vanishes, and
$\beta_1=\mathbb{E}(Y|D=1)-\mathbb{E}(Y|D=0)=\text{ATE}$.
OLS of Y on D gives unbiased estimate of ATE.

Conditional Independence (Unconfoundedness)

$y(0),y(1)\perp D|w$ (selection on observables). Then:

$$\text{ATE}=\mathbb{E}[\mathbb{E}(Y|D=1,w)-\mathbb{E}(Y|D=0,w)].$$

Requires **overlap**: $0<P(D=1|w=w')<1$ for all w'.

Homogeneous vs. Heterogeneous Effects

Homogeneous: $y_i(1)-y_i(0)=\beta_1$ for all i. Then $y_i=\beta_0+\beta_1D_i+u_i$ has a causal interpretation: β_1 is the treatment effect.
Heterogeneous: Effects vary across i. Regression coefficient is an average effect, not the individual effect.

Heterogeneous Effects with Interactions

If $x\in\{0,1\}$ and effects vary, the correct specification is:

$$y_i=\beta_0+\beta_1x_i+\beta_2D_i+\beta_3D_ix_i+v_i,\ \ \mathbb{E}(v|D,x)=0.$$

Here $\beta_2=\mathbb{E}(y(1)-y(0)|x=0)$ and $\beta_3=\text{ATE}(x=1)-\text{ATE}(x=0)$.
A common **misspecification trap** arises: if you omit D_ix_i and run $y=b_0+b_1x+b_2D+e$, then b_2 converges to a *variance-weighted* average of conditional ATEs:

$$b_2\xrightarrow{p}\sum_x\frac{\text{Var}(D|x)\,P(x)}{\mathbb{E}(\text{Var}(D|x))}\cdot\text{ATE}(x),$$

which generally $\ne$ ATE unless effects are homogeneous or $P(D=1|x)$ is constant.

Inverse Probability Weighting (IPW)

Under unconfoundedness and overlap, with $p(x):=P(D=1|X=x)$:

$$\text{ATE}=\mathbb{E}\left[\frac{Y(D-p(X))}{p(X)(1-p(X))}\right],$$
$$\text{ATT}=\mathbb{E}\left[\frac{YD}{P(D=1)}\right]-\mathbb{E}\left[\frac{Y(1-D)p(X)}{P(D=1)(1-p(X))}\right].$$

Useful when the propensity score $p(x)$ is easier to model than $\mathbb{E}(Y|D,X)$.

Multiple Treatments

k treatments, $k+1$ potential outcomes. With random assignment:

$$\mathbb{E}(Y|x)=\beta_0+\beta_1x_1+\cdots+\beta_kx_k,$$

where $\beta_0=\mathbb{E}(y(0))$ and $\beta_j=\mathbb{E}(y(j)-y(0))$.

All-Causes (Latent Variable) Framework

Setup

The **all-causes** (or **latent variable**) model specifies:

$$Y = g(D, U),$$

where D denotes observed determinants and U encompasses *all* unobserved determinants of Y. Together, D and U exhaustively cause the outcome. The linear case: $Y = \alpha + \beta D + U$.
Key distinction: U has a *causal* interpretation (it represents unobserved causes of Y), unlike a regression residual ε which is a statistical object minimizing MSE.

From Potential Outcomes to All-Causes

Binary $D \in \{0, 1\}$ with potential outcomes $Y(0), Y(1)$:

$$\begin{aligned} Y &= DY(1) + (1-D)Y(0) \\ &= \underbrace{\mathbb{E}(Y(0))}_{\alpha} + \underbrace{(Y(1) - Y(0))}_{\beta} \cdot D + \underbrace{Y(0) - \mathbb{E}(Y(0))}_{U}. \end{aligned}$$

β is deterministic under homogeneous effects; a random variable under heterogeneous effects.

From All-Causes to Potential Outcomes

Given $Y = \alpha + \beta D + U$, define:

$$Y(0) \equiv g(0, U) = \alpha + U, \quad Y(1) \equiv g(1, U) = \alpha + \beta + U.$$

Both are random through U; β may also be random (heterogeneous effects).

Causal U vs. Regression Residual

In the all-causes model $Y = D'\beta + U$: $\mathbb{E}(DU) = 0$ asserts observed and unobserved *causes* are orthogonal—a substantive causal claim. In contrast, the BLP residual $\varepsilon = Y - D'\beta^*$ satisfies $\mathbb{E}(D\varepsilon) = 0$ by construction (FOC of MSE minimization), with no causal content. $\beta^* = \beta$ iff the causal orthogonality condition $\mathbb{E}(DU) = 0$ holds. When $\mathbb{E}(DU) \neq 0$ (endogeneity), $\beta^* \neq \beta$ and OLS is inconsistent for the causal parameter.

Equivalence Result (Vytlacil, 2002)

The latent variable selection model (Heckman–Vytlacil):

$$Y_d = \mu_d(X, U_d), \quad D^* = \mu_D(Z) - U_D, \quad D = \mathbf{1}[D^* \geq 0],$$

with (A1) $\mu_D(Z)$ nondegenerate $|X$; (A2) $(U_0, U_D), (U_1, U_D) \perp Z|X$; (A3) U_D absolutely continuous; (A4) $\mathbb{E}|Y_d| < \infty$; (A5) $0 < P(D=1|X) < 1$, is **equivalent** to the LATE assumptions of Imbens–Angrist (1994): independence, exclusion, relevance, and monotonicity. The latent variable model generates LATE, and LATE assumptions generate the latent variable model.

Instrumental Variables

The Endogeneity Problem

If $\mathbb{E}(xu) \neq 0$ (endogeneity), OLS is inconsistent for β:

$$\hat{\beta}_n^{\text{OLS}} \xrightarrow{p} \beta + \mathbb{E}(xx')^{-1}\mathbb{E}(xu) \neq \beta.$$

IV Conditions

Use instrument $z \in \mathbb{R}^{l+1}$ ($z_0 = 1$) satisfying:
Validity: $\mathbb{E}(zu) = 0$ (exogeneity + exclusion).
Relevance (Rank Condition): $\mathbb{E}(zx')$ has rank $k+1$.
Order Condition (necessary): $l \geq k$ (at least as many instruments as regressors).
When $l = k$, the model is exactly identified; when $l > k$, it is overidentified.

Identification

From $y = x'\beta + u$ and $\mathbb{E}(zu) = 0$: $\mathbb{E}(zy) = \mathbb{E}(zx')\beta$. If $l = k$:

$$\beta = \mathbb{E}(zx')^{-1}\mathbb{E}(zy).$$

IV Estimator (Exact Identification)

$$\hat{\beta}_{\text{IV}} = \left(\frac{1}{n}\sum z_i x_i'\right)^{-1} \frac{1}{n}\sum z_i y_i = (Z'X)^{-1}Z'Y.$$

Consistent by the LLN and CMT.

Potential Outcomes Framework for IV

Exclusion: $y(d, z') = y(d, z'')$ $\forall d, z', z''$. Write $y(d) \equiv y(d, z')$.
Exogeneity: $y(d, z') \perp z|w$ $\forall d, z'$.
Under constant linear treatment effects: $y(d) = x(w, d)'\beta + u$, $\mathbb{E}(u|w) = 0$.
Exclusion $\implies$ $\mathbb{E}(u|z, w) = 0$, so $\mathbb{E}(zu) = 0$.

Asymptotic Distribution

$$\sqrt{n}(\hat{\beta}_{\text{IV}} - \beta) \xrightarrow{d} N(0, V_{\text{IV}}),$$

where in the scalar case with $\mathbb{E}(u^2|z) = \sigma^2$: $V_{\text{IV}} = \frac{\sigma^2}{\text{Corr}(x,z)^2\text{Var}(x)}$.
IV vs. OLS efficiency: If $\mathbb{E}(u|x) = 0$, OLS more efficient: $V_{\text{OLS}} = \frac{\sigma^2}{\text{Var}(x)}$.

GMM / 2SLS (Overidentification)

When $l > k$: $\mathbb{E}(zx')$ is $l+1 \times k+1$, not square. Use GMM:

$$\hat{\beta}_{\text{2SLS}} = (X'P_Z X)^{-1}X'P_Z Y,$$

where $P_Z = Z(Z'Z)^{-1}Z'$. Equivalently: regress X on Z (first stage), then Y on $\hat{X}$ (second stage).

Overidentification Tests (Sargan/Hansen)

Test $H_0 : \mathbb{E}(zu) = 0$. Under H_0 and homoskedasticity, the test statistic is $n \times R^2$ from regressing $\hat{u}_i$ on all instruments z, distributed χ^2_{L-K} asymptotically, where $L = \#$ instruments and $K = \#$ endogenous regressors. Rejection implies invalid instruments or misspecification, but we cannot tell which instrument is bad.

Must Include Exogenous Regressors in First Stage

With endogenous x_k and instruments $z = (x_{-k}, z_k)$: the first stage *must* regress x_k on *all* of z, not just z_k. Omitting x_{-k} causes OVB in $\tilde{\pi}$, yielding $\hat{\beta}_k \xrightarrow{p} \beta_k \pi/(\pi + \nu) \neq \beta_k$. Intuitively, $\hat{x}_k$ without controls captures variation from x_{-k}, which violates the ceteris paribus interpretation.

LATE Interpretation

With heterogeneous effects and binary D, Z, IV does not identify the ATE. Under **monotonicity** ($D_i(1) \geq D_i(0)$ $\forall i$, no defiers), the Wald estimand is the **LATE**:

$$\frac{\mathbb{E}(Y|Z=1) - \mathbb{E}(Y|Z=0)}{\mathbb{E}(D|Z=1) - \mathbb{E}(D|Z=0)} = \mathbb{E}(y(1) - y(0)|\text{complier}).$$

Derivation: Numerator $= \mathbb{E}(Y(1) - Y(0))$ by exogeneity. Always-takers/never-takers contribute zero (D unchanged by Z); defiers ruled out. So numerator $= P(\text{c}) \cdot \mathbb{E}(y(1) - y(0)|\text{c})$. Denominator $= P(\text{c})$. Ratio is the LATE.

Hausman Test (Exogeneity)

Test $H_0 : \mathbb{E}(xu) = 0$ using both OLS and IV. Under H_0, both are consistent; under H_1, only IV is. Joint distribution:

$$\sqrt{n}\begin{pmatrix} \hat{\beta}_{\text{OLS}} - \beta \\ \hat{\beta}_{\text{IV}} - \beta \end{pmatrix} \xrightarrow{d} N(0, V_{\text{joint}}).$$

Test statistic: $T_n = n(\hat{\beta}_{\text{IV}} - \hat{\beta}_{\text{OLS}})'\hat{W}^{-1}(\hat{\beta}_{\text{IV}} - \hat{\beta}_{\text{OLS}}) \xrightarrow{d} \chi^2_k$ under H_0, where $\hat{W}$ consistently estimates the variance of the difference.

Control Functions

Setup and Equivalence to 2SLS

Model: $y_1 = z_1'\delta_1 + \alpha_1 y_2 + u_1$, $\mathbb{E}(z'u_1) = 0$. Reduced form: $y_2 = z'\pi_2 + v_2$, $\mathbb{E}(z'v_2) = 0$. Endogeneity arises iff $\text{Cov}(u_1, v_2) \neq 0$.
CF idea: Project u_1 on v_2: $u_1 = \rho_1 v_2 + e_1$, where $\mathbb{E}(v_2 e_1) = 0$ and $\mathbb{E}(z'e_1) = 0$. Substituting:

$$y_1 = z_1'\delta_1 + \alpha_1 y_2 + \rho_1 v_2 + e_1.$$

Since $e_1 \perp (z_1, y_2, v_2)$, OLS on this equation is consistent. Replace v_2 with $\hat{v}_2$ (OLS residuals from first stage). In the linear case, CF estimates of δ_1, α_1 are **numerically identical** to 2SLS. Test $H_0 : \rho_1 = 0$ is a test of exogeneity.

Advantage: Nonlinear Models

For $y_1 = z_1'\delta_1 + \alpha_1 y_2 + \gamma_1 y_2^2 + u_1$, standard 2SLS needs extra instruments for y_2^2. CF only adds the scalar $\hat{v}_2$:

$$y_1 \text{ on } z_1, y_2, y_2^2, \hat{v}_2.$$

Requires stronger assumption: $\mathbb{E}(u_1|z, v_2) = \mathbb{E}(u_1|v_2) = \rho_1 v_2$ (independence of v_2 from z, linearity of conditional expectation).

Binary Endogenous Variable

If $y_2 \in \{0, 1\}$ with $y_2 = \mathbf{1}[z'\pi_2 + e_2 \geq 0]$, $e_2 \sim N(0, 1)$: the CF uses the **generalized residual** $\widehat{gr}_{i2} = y_{i2}\lambda(z_i'\hat{\pi}_2) - (1 - y_{i2})\lambda(-z_i'\hat{\pi}_2)$, where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ is the inverse Mills ratio. Regress y_1 on $z_1, y_2, \widehat{gr}_2$. Less robust than IV but more efficient when correct.

CF vs. IV: Tradeoffs

IV (2SLS): More robust — only needs $\mathbb{E}(z'u_1) = 0$ and rank condition. Works regardless of y_2's distribution.
CF: More efficient — solves endogeneity with one scalar control. But requires correct specification of the first-stage distribution and $\mathbb{E}(u_1|v_2)$ linearity. Misspecification $\implies$ inconsistency.

Correlated Random Coefficients

Model: $y_1 = \alpha_0 + z_1'\delta_1 + a_1 y_2 + u_1$ where $a_1 = \alpha_1 + v_1$ is random. Write $e_1 = v_1 y_2 + u_1$. 2SLS is inconsistent if $\text{Cov}(z, v_1 y_2) \neq 0$. CF fix: include $\hat{v}_2$ and $\hat{v}_2 y_2$ as controls $\implies$ consistent $\hat{\alpha}_1$ (Heckman & Vytlacil, 1998).

Marginal Treatment Effects

Framework

Binary treatment $D \in \{0,1\}$, outcome Y, potential outcomes $Y(0), Y(1)$. Selection: $D = \mathbf{1}[U \leq p(Z)]$, where $p(Z) = P(D{=}1|Z)$ is the propensity score and $U \sim \text{Unif}[0,1]$ (normalized). Assumptions: (A1) $p(Z)$ nondegenerate given X; (A2) $(Y(0), Y(1), U) \perp Z|X$; (A3) U absolutely continuous; (A4) $\mathbb{E}|Y_d| < \infty$; (A5) $0 < P(D{=}1|X) < 1$.

MTE Definition

$$\text{MTE}(u) \equiv \mathbb{E}[Y(1) - Y(0)|U = u].$$

The MTE is the average treatment effect for agents at the margin of indifference when $U = u$. Low $U \implies$ high propensity to select into treatment.

Target Parameters as Weighted Averages

All standard treatment parameters are weighted averages of MTE:

$$\text{ATE} = \int_0^1 \text{MTE}(u)\,du, \quad \omega_{\text{ATE}} = 1,$$
$$\text{ATT} = \int_0^1 \text{MTE}(u) \cdot \frac{P(u \leq p(Z))}{P(D{=}1)}\,du,$$
$$\text{ATU} = \int_0^1 \text{MTE}(u) \cdot \frac{P(u > p(Z))}{P(D{=}0)}\,du.$$

ATT overweights MTE at low u (likely treated); ATU overweights high u.

Selection Patterns

Selection on the gain: $\text{MTE}(u)$ decreasing — those who select $D{=}1$ have higher returns. Implies ATT > ATE > ATU.
Selection on the loss: $\text{MTE}(u)$ increasing — those who select $D{=}1$ gain less. ATU > ATE > ATT.
Essential heterogeneity: Agents select based on unobserved idiosyncratic returns. Different instruments identify different weighted averages of MTE $\implies$ different IV estimates are not comparable.

Identification

With continuous Z, the MTE is identified from the derivative of the conditional expectation:

$$\text{MTE}(p) = \frac{\partial}{\partial p}\mathbb{E}[Y|p(Z) = p].$$

Intuition: a marginal increase in p induces the agent at $U = p$ into treatment.

Vytlacil (2002) Equivalence

Assumptions (A1)–(A5) of the nonparametric selection model (generalized Roy) are *equivalent* to the LATE assumptions of Imbens & Angrist (1994) when D is binary. The latent variable model implies their assumptions and vice versa. This equivalence breaks down for multivalued treatments.

Policy Relevant Treatment Effect

For a policy shifting $p(Z)$ from $p_{a'}$ to p_a:
$\text{PRTE} = \int_0^1 \text{MTE}(u) \cdot \omega_{\text{PRTE}}(u)\,du$, where $\omega_{\text{PRTE}}(u)$ depends on the policy change. Unlike LATE, PRTE answers: "what is the effect on people this policy would move into treatment?"

Weak Instruments

Setup

$y = \beta x + u$, $x = \pi z + v$, $\mathbb{E}(zu) = \mathbb{E}(zv) = 0$. Identification requires $\pi \neq 0$.

The Problem with $\pi \approx 0$

$$\hat{\beta}_{\text{IV}} = \beta + \frac{\frac{1}{\sqrt{n}}\sum z_i u_i}{\frac{1}{n}\sum z_i x_i}.$$

If $\pi = 0$: denominator $\xrightarrow{p} 0$ but numerator $\xrightarrow{d}$ normal. We cannot apply Slutsky's theorem. The ratio converges to a **ratio of correlated normals**, not $N(0, V)$.

Finite Sample Bias

By Kinal (1980), $\hat{\beta}_{\text{IV}}$ has finite moments of order up to $(L - K)$, where L is the number of excluded instruments and K the number of endogenous regressors. With $L = K$ (exactly identified): $\hat{\beta}_{\text{IV}}$ has **no finite moments** (not even a finite mean).
$\hat{\beta}_{\text{OLS}} \xrightarrow{p} \frac{\sigma_{uv}}{\sigma_v^2}$ (biased toward OLS probability limit).

Rule of Thumb

First-stage F-statistic ≥ 10 for relative bias $\leq 10\%$ (Stock & Yogo, 2005). Critical values range 9–12 for 3–30 instruments. But this is only an approximation.

Anderson-Rubin Test

Robust to weak instruments. Suppose $y = x'\beta + u$, $\mathbb{E}(zu) = 0$. Under $H_0 : \beta = \beta_0$: $\mathbb{E}(z(y - x'\beta_0)) = 0$, so regress $u(\beta_0) = y - x'\beta_0$ on z:

$$u_i(\beta_0) = z_i'\gamma + \epsilon_i, \quad \mathbb{E}(z\epsilon) = 0.$$

Under H_0: $\gamma = 0$. Test statistic:

$$T_n = n\hat{\gamma}'\hat{V}^{-1}\hat{\gamma} \xrightarrow{d} \chi^2_l.$$

Reject if $T_n > \chi^2_{l,1-\alpha}$.

Why robust: Under H_0, $\sqrt{n}\hat{\gamma} = \left(\frac{1}{n}\sum z_i z_i'\right)^{-1}\frac{1}{\sqrt{n}}\sum z_i u_i$. This uses CLT on $z_i u_i$ directly—no division by a possibly-near-zero first stage.
Power: Under $H_1 : \beta \neq \beta_0$, we have $\gamma = \mathbb{E}(zz')^{-1}\mathbb{E}(zx')(\beta - \beta_0) \neq 0$, so the test has power against any alternative consistent with the maintained model assumptions (exclusion restriction and correct specification).

With Included Instruments

Separate $z = (z_1, z_2)$, $x = (x_1, z_1)$. Under $H_0 : \beta_1 = \beta_{1,0}$, regress $y - x_1'\beta_{1,0}$ on z_1 and z_2. Test whether coefficients on z_2 are zero:

$$T_n = n\hat{\gamma}_{z_2}'\hat{V}^{-1}\hat{\gamma}_{z_2} \xrightarrow{d} \chi^2_{l_2}.$$

Difference-in-Differences

Setup

Two periods: $T \in \{0,1\}$. Two groups: $G \in \{0,1\}$ (treated/control). Observed outcome:

$$Y = \begin{cases} y(1) & \text{if } G = 1, T = 1 \\ y(0) & \text{otherwise.} \end{cases}$$

Target: $\text{ATT} = \mathbb{E}(y(1) - y(0)|G = 1, T = 1)$.

Naive Comparisons Fail

Across time (treated group):

$$\mathbb{E}(Y|G{=}1, T{=}1) - \mathbb{E}(Y|G{=}1, T{=}0) = \text{ATT} + \text{Temporal trend}.$$

Across groups (post-period):

$$\mathbb{E}(Y|G{=}1, T{=}1) - \mathbb{E}(Y|G{=}0, T{=}1) = \text{ATT} + \text{Selection bias in } y(0).$$

Common Trends Assumption

In the absence of treatment, the change in outcomes would be the same for treated and control groups:

$$\mathbb{E}(y(0)|G = 1, T = 1) - \mathbb{E}(y(0)|G = 1, T = 0)$$
$$= \mathbb{E}(y(0)|G = 0, T = 1) - \mathbb{E}(y(0)|G = 0, T = 0).$$

Note that the LHS involves an *unobservable* counterfactual.

DiD Estimand

Under common trends:

$$\text{ATT} = [\mathbb{E}(Y|G{=}1, T{=}1) - \mathbb{E}(Y|G{=}1, T{=}0)]$$
$$- [\mathbb{E}(Y|G{=}0, T{=}1) - \mathbb{E}(Y|G{=}0, T{=}0)].$$

Regression Implementation

$$Y_{it} = \beta_0 + \beta_1 G_i + \beta_2 T_t + \beta_3 (G_i \cdot T_t) + u_{it}.$$

β_3 is the DiD estimator = ATT under common trends.
β_1: group difference at baseline. β_2: common time effect.

Data Requirements

Repeated cross section: Random sample in each period (different units).
Panel data: Same units observed in both periods (stronger). Every panel is a repeated cross section, but not vice versa.

Regression Discontinuity Design

Setup

Running variable X, cutoff c, treatment D. Potential outcomes $y(0), y(1)$.

Sharp RDD

$D = \mathbf{1}(X \geq c)$: treatment is deterministic in X.
Unconfoundedness holds: $y(0), y(1) \perp D|X$ (since D is a function of X).
Overlap fails: $P(D = 1|X = x) = \mathbf{1}(x \geq c) \in \{0,1\}$.

Identification

Under continuity: $\mathbb{E}(y(0)|X=x)$ continuous at c:
$$\mathbb{E}(y(1)-y(0)|X=c)=\mathbb{E}(Y|X=c)-\lim_{x\uparrow c}\mathbb{E}(Y|X=x).$$
This identifies the treatment effect **at the cutoff only**.

Estimation: Local Linear Regression

Choose bandwidth h and solve:
$$\min_{\alpha_0,\beta_0,\gamma,\delta}\sum_{i=1}^{n}\mathbf{1}(|X_i-c|\le h)(y_i-\alpha_0-\beta_0x_i-\gamma d_i-\delta d_ix_i)^2.$$
Equivalent to two separate regressions on $\{i:X_i\in[c-h,c)\}$ and $\{i:X_i\in[c,c+h]\}$.
Recentering: $Y_i=\alpha_0+\beta_0(X_i-c)+\gamma D_i+\delta D_i(X_i-c)+\epsilon_i$ makes γ the discontinuity.

Bandwidth Choice

Optimal: $h=C\cdot n^{-1/5}$ (bias-variance tradeoff). Larger $h\Rightarrow$ lower variance, higher bias.
IK (2012) and CCT (2014) propose data-driven bandwidth selectors. CCT accounts for asymptotic bias.

Threats to Validity

Manipulation: Individuals choosing X values near cutoff. McCrary (2008): test for density discontinuity at c.
Multiple treatments: Cannot identify which treatment caused the jump.
Covariate balance: Pre-determined covariates should be continuous at c; discontinuities suggest violations.

Fuzzy RDD

$P(D=1|X=x)$ is discontinuous at c, but $D\ne\mathbf{1}(X\ge c)$. Then $Z=\mathbf{1}(X\ge c)$ is an instrument for D.
Under monotonicity ($P(D_1\ge D_0)=1$), the estimand is a LATE at the cutoff:
$$\mathbb{E}(y(1)-y(0)|X=c,\text{complier})=\frac{\lim_{x\downarrow c}\mathbb{E}(Y|X=x)-\lim_{x\uparrow c}\mathbb{E}(Y|X=x)}{\lim_{x\downarrow c}\mathbb{E}(D|X=x)-\lim_{x\uparrow c}\mathbb{E}(D|X=x)}.$$

Fuzzy RDD Implementation

2SLS on subsample $\{i:|X_i-c|\le h\}$:
First stage: $D=\pi_0+\pi_1Z+\pi_2(X-c)+\pi_3Z(X-c)+v$.
Second stage: $Y=\beta_0+\beta_1D+\beta_2(X-c)+\beta_3Z(X-c)+u$.

Panel Data

Setup

Consider N individuals observed over T time periods. The linear model is:
$$Y_{it}=X_{it}'\beta+\alpha_i+u_{it},$$
where α_i is an unobserved individual fixed effect.
Asymptotics: Large N, small T.

Problem with Pooled OLS

Pooled OLS treats α_i+u_{it} as a composite error. If $\mathbb{E}(X_{it}\alpha_i)\ne0$, this composite error is correlated with the regressors, making pooled OLS inconsistent.

First Differencing (FD)

Difference across time to eliminate α_i:
$$\Delta Y_{it}=\Delta X_{it}'\beta+\Delta u_{it}.$$
FD estimator: $\hat\beta_{\text{FD}}$ is OLS applied to differenced data.
Consistency requires: $\mathbb{E}(\Delta X_{it}'\Delta u_{it})=0$, that is, the changes in regressors are uncorrelated with the changes in errors. A sufficient (but stronger than necessary) condition is $\mathbb{E}(u_{it}|X_{it},X_{it-1})=0$; FD only requires the moment condition on the differences, not full contemporaneous exogeneity.
Not sufficient if unobservables in *other* time periods are correlated with today's regressors.

Strict Exogeneity

$\mathbb{E}(u_{it}|X_{i1},\dots,X_{iT})=0$ for all t. Stronger than contemporaneous exogeneity. Required for FE consistency.
Violated if, e.g., past outcomes affect future regressors (feedback effects).

Fixed Effects (FE) / Within Estimator

For general $T\ge2$, define within-transformed variables $\ddot Y_{it}=Y_{it}-\bar Y_i$:
$$\ddot Y_{it}=\ddot X_{it}'\beta+\ddot u_{it}.$$
FE estimator is OLS on demeaned data. Eliminates α_i without differencing.
Under strict exogeneity: $\hat\beta_{\text{FE}}$ is consistent as $N\to\infty$ (fixed T).
FE = LSDV equivalence: Regressing Y on X and N individual dummies (LSDV) yields the same $\hat\beta$ as the within estimator. **Proof:** By FWL, $\hat\beta_{\text{LSDV}}=(X'M_DX)^{-1}X'M_DY$ where D is the matrix of individual dummies. M_D demeans within each individual: $(M_DY)_{it}=Y_{it}-\bar Y_i=\ddot Y_{it}$. So $\hat\beta_{\text{LSDV}}=(\ddot X'\ddot X)^{-1}\ddot X'\ddot Y=\hat\beta_{\text{FE}}$.

FD vs. FE

With $T=2$: FD = FE.
With $T>2$: differ in general. FE more efficient under homoskedasticity of u_{it}; FD more robust to serial correlation patterns.

Serial Correlation and Clustered SEs

Standard errors must account for within-individual serial correlation in u_{it}. **Cluster-robust variance:** allows arbitrary within-cluster correlation: $\hat V=(X'X)^{-1}\left(\sum_{j=1}^J X_j'\hat U_j\hat U_j'X_j\right)(X'X)^{-1}$, where j indexes clusters, X_j and $\hat U_j$ are the data and residuals for cluster j.

Tips and Tricks

Proof Strategies for Consistency

1. Write estimator as function of sample averages.
2. Apply SLLN to each sample average.
3. Apply CMT to the composed function.

For extremum estimators (MLE, GMM): show uniform convergence of objective function + identification at θ_0.
Example (OLS): $\hat\beta=(\frac1n\sum x_ix_i')^{-1}(\frac1n\sum x_iy_i)$. SLLN: $\frac1n\sum x_ix_i'\xrightarrow{p}\mathbb{E}(xx')$, $\frac1n\sum x_iy_i\xrightarrow{p}\mathbb{E}(xy)$. CMT: $\hat\beta\xrightarrow{p}\mathbb{E}(xx')^{-1}\mathbb{E}(xy)=\beta$.

Trace Trick for Quadratic Forms

For scalar $U'AU$: $U'AU=\text{tr}(U'AU)=\text{tr}(AUU')$, so $\mathbb{E}[U'AU|X]=\text{tr}(A\,\mathbb{E}[UU'|X])$. Key identity: $\text{tr}(AB)=\text{tr}(BA)$.
Example:
$\mathbb{E}[\text{SSR}|X]=\mathbb{E}[u'M_Xu|X]=\text{tr}\big(M_X\sigma^2I\big)=\sigma^2\,\text{tr}(M_X)=\sigma^2(n-k-1)$, since M_X is idempotent with $\text{tr}(M_X)=n-k-1$.

Proof Strategies for Asymptotic Normality

1. Decompose $\sqrt n(\hat\theta-\theta)$ into a CLT term and remainder.
2. Apply CLT to iid mean-zero term.
3. Show remainder is $o_p(1)$ using Slutsky.

Example (OLS): $\sqrt n(\hat\beta-\beta)=(\frac1n\sum x_ix_i')^{-1}\frac{1}{\sqrt n}\sum x_iu_i$. CLT: $\frac{1}{\sqrt n}\sum x_iu_i\xrightarrow{d}N(0,\Sigma)$. Slutsky: first factor $\xrightarrow{p}Q^{-1}$. Result: $\sqrt n(\hat\beta-\beta)\xrightarrow{d}N(0,Q^{-1}\Sigma Q^{-1})$.

Useful Inequalities

Markov: $P(|X|\ge t)\le\frac{\mathbb{E}|X|}{t}$.
Chebyshev: $P(|X-\mu|\ge t)\le\frac{\text{Var}(X)}{t^2}$.
Jensen: If g convex, $g(\mathbb{E}(X))\le\mathbb{E}(g(X))$. Strict if g strictly convex and X non-degenerate.
Cauchy-Schwarz: $|\mathbb{E}(XY)|^2\le\mathbb{E}(X^2)\mathbb{E}(Y^2)$.

Key o_p/O_p Arguments

To show $\frac1n\sum\hat u_i^2x_ix_i'\xrightarrow{p}\mathbb{E}(u^2xx')$:
$\max_{i\le n}|\hat u_i^2-u_i^2|\le\|\hat\beta-\beta\|^2\max\|x_i\|^2+2\|\hat\beta-\beta\|\max\|x_iu_i\|$.
Use: $\frac{\max\|Z_i\|}{n^{1/r}}=o_p(1)$ when $\mathbb{E}\|Z\|^r<\infty$.

Common Endogeneity Sources

- Omitted variables correlated with both x and y
- Simultaneity / reverse causality
- Measurement error in regressors
- Self-selection into treatment

IV Checklist

1. **Relevance:** First-stage F ≥10 (weak instrument check)
2. **Exclusion:** z affects y only through x (untestable)
3. **Exogeneity:** $\mathbb{E}(zu)=0$ (partially testable via overid)
4. **Monotonicity:** For LATE interpretation with heterogeneous effects

Identification Strategy Summary

- **RCT:** Random assignment $\implies$ ATE from simple regression
- **Selection on observables:** Unconfoundedness + overlap $\implies$ ATE
- **IV:** Exogeneity + relevance $\implies$ causal effect (LATE if heterogeneous)
- **DiD:** Common trends $\implies$ ATT
- **RDD:** Continuity at cutoff $\implies$ treatment effect at cutoff
- **Panel FE/FD:** Eliminates time-invariant unobservables

Bias-Variance Tradeoff (Irrelevant Variables)

Including irrelevant variable ($\beta_2=0$): no bias, but *increases* variance. Omitting relevant variable ($\beta_2\ne0$): introduces OVB, but *decreases* variance. Via FWL: $\text{Var}(\hat\beta_1|X)=\sigma^2/\text{SSR}_1$ where SSR_1 is residual SS from regressing x_1 on other regressors. Adding correlated regressors lowers SSR_1, inflating variance.