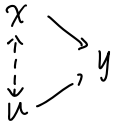
Causality I (C.C. 8.9).

Unit level Causal Model:

y: outcome

(x, u): covariates

g: structural functions.



Suppose $y \in \mathbb{R}$, $x \in \mathcal{X} \subseteq \mathbb{R}$, $u \in \mathcal{U}$

$y \leftarrow g(x, u)$.

implies: $y = g(x, u)$.

Causal Effects:

1. $g(x_1, u_0) - g(x_0, u_0)$.
2. Marginal causal effect:
   $[\nabla_x g](x_0, u_0)$
   where $\nabla_x g = \frac{\partial g}{\partial x}$

Causal Effects for Specific Unit i.

Y_i: outcome variable
(X_i, U_i): covariates } numbers assigned to i

$Y_i \leftarrow g(X_i, U_i)$ for all $i \in \mathcal{X}$.

Fix U_i: $g(x_1, U_i) - g(x_0, U_i)$.

Potential Outcomes:

$Y_i(\cdot): \mathcal{X} \to \mathbb{R}$.

$Y_i(x) \equiv g(x, U_i)$ $\quad y \leftarrow Y_i(x)$.

unit level causal effect: $Y_i(x_1) - Y_i(x_0)$.

learning $Y_i(\cdot)$ or some aspect of it is the main goal of causal analysis.

Counterfactuals:

Realized/Factual outcomes: $Y_i = Y_i(X_i) = g(X_i, U_i)$.

Counterfactual outcomes: $Y_i(x)$ for $x \neq X_i$

Status quo treatment effect: $Y_i(x) - Y_i(X_i)$

Population level Causal Effects: Distribution of Effects.

X: random variable

X_i: value drawn from X

x: fixed, $\in \mathcal{X}$.

for i: $\Delta_i(x_0 \to x_1) = g(x_1, U_i) - g(x_0, U_i) = Y_i(x_1) - Y_i(x_0)$

for population: $\Delta(x_0 \to x_1) = g(x_1, U) - g(x_0, U) = Y(x_1) - Y(x_0)$.

where U is random variable.

Distribution of Treatment Effect:

$$DTE(t, x_0 \to x_1) = F_{Y(x_1) - Y(x_0)}(t)$$
$$= \mathbb{P}(Y(x_1) - Y(x_0) \leq t).$$

* The proportion who benefit (voting criterion):

parameter: $\mathbb{P}[Y(1) > Y(0)]$
$= \mathbb{P}(Y(1) - Y(0) > 0)$
$= 1 - DTE(0, 0 \to 1)$.

* Distribution of Marginal Causal Effect

$\Delta(x_0) = [\nabla_x g](x_0, U)$

$\Delta(x_0) = \frac{\partial Y(x_0)}{\partial x_0}$

Population Level Causal Effects: Summary Statistics

Average Treatment Effect:

$$ATE \equiv \mathbb{E}[\Delta(x_0 \to x_1)]$$
$$= \mathbb{E}_U[g(x_1, U) - g(x_0, U)]$$
$$= \mathbb{E}_U[g(x_1, U)] - \mathbb{E}_U[g(x_0, U)]$$

Average Structual Function:

$$ASF \equiv \mathbb{E}_U[g(x, U)]$$
$$= \mathbb{E}[Y(x)]$$

$\therefore ATE = \mathbb{E}[Y(x_1) - Y(x_0)]$
$= \mathbb{E}[Y(x_1)] - \mathbb{E}[Y(x_0)]$.

Quantile Effects.

τ^{th} quantile of the distribution of u.l.c.e.
$Q_{\Delta(x_0 \to x_1)}(\tau)$.

τ^{th} quantile treatment effect:
$QTE(\tau, x_0 \to x_1) = Q_{g(x_1, U)}(\tau) - Q_{g(x_0, U)}(\tau)$

Usually they are not the same.

Examples of Unit Level Causal Models.

1. Binary:

$(U_i, X_i, Y_i, Y_i(0), Y_i(1))$. X_i: binary.

$Y_i = Y_i(1) X_i + Y_i(0)(1 - X_i)$

2. The Additively Separable Model:

($x \in \mathcal{X} \subseteq \mathbb{R}$.)

$$Y(x) = m(x) + U.$$

$m: \mathcal{X} \to \mathbb{R}$ unknown function. U: unknown scalar r.v.

$Y = m(X) + U$. $m(\cdot)$ not dependent on i.

$Y_i(x) = m(x) + U_i$

$Y_i(x_1) - Y_i(x_0) = m(x_1) - m(x_0)$. homogeneous treatment effects.

(nonlinear in X).

2 (con't). The Linear Constant Coefficient Model:

assume $m(x) = \theta_0 + \theta_1 x$.

$Y(x) = m(x) + U = \theta_0 + \theta_1 x + U$

$Y_i(x) = \theta_0 + \theta_1 x + U_i$ not a "residual", but an unobserved covariate.

$Y_i(x_1) - Y_i(x_0) = \theta_1(x_1 - x_0)$.

Structural Interpretation of unobserved random variable U:

different X_i across i: observed heterogeneity

different U_i across i: unobserved heterogeneity.

The Difference btw Fixing and Conditioning:

Fixing: using marginal distribution of U, define:

potential outcome: $Y(x) = g(x, U)$

then using F_U to define causality $Y(x_1) - Y(x_0)$.

it depends only on: ① $g(\cdot,\cdot)$

② marginal dist. F_U

Here we do not say anything about the joint dist of X & U in the data to define potential outcomes!

Conditioning: implicitly depends on true joint dist of X & U.

e.g. $\mathbb{E}[Y|X=x] = \mathbb{E}[g(X,U)|X=x]$

$= \mathbb{E}[g(x,U)|X=x]$

It is not the same thing as.

$ASF(x) \equiv \mathbb{E}[g(x,U)]$!

Causal Effects for Subpopulations:

treated units: $\{i \in \mathcal{X} : X_i = 1\}$

untreated units: $\{i \in \mathcal{X} : X_i = 0\}$.

$ATT = \mathbb{E}[Y(1) - Y(0) | X = 1]$

$ATU = \mathbb{E}[Y(1) - Y(0) | X = 0]$

by LIE: $ATE = ATT \cdot \mathbb{P}(X=1) + ATU \cdot \mathbb{P}(X=0)$.

Observe another covariate W:

consider units: $\{i \in \mathcal{X} : W_i = w\}$. a subpopulation.

$CATE = \mathbb{E}[Y(1) - Y(0) | W = w]$.

• Common Identification Proof Techniques.

• **Constructive method**: parameter is a known function of the data.

$F_{Y,X,U}$: true population distribution.

$F_{Y,X}$: observed population data.

$\tilde{\mathcal{F}}$: restrictions put on $F_{Y,X,U}$.

$\tilde{F}_{Y,X,U}$: a conjectured value of the population dist.

$\tilde{F}_{Y,X} = MakeData(\tilde{F}_{Y,X,U})$: corresponding population data that would be observed if $\tilde{F}_{Y,X,U}$ were true.

Theorem: suppose there is a known function g such that $\theta(\tilde{F}_{Y,X,U}) = g(\tilde{F}_{Y,X})$. for any $\tilde{F}_{Y,X,U} \in \tilde{\mathcal{F}}$ and suppose the model is not refuted. Then, θ is point identified.

pf: WTS: the identified set:

$\mathcal{H}_I = \{\theta \in \mathcal{H} : \theta = \theta(F_{Y,X,U}) \text{ for some } \tilde{F}_{Y,X,U} \in \tilde{\mathcal{F}} \text{ such that } MakeData(\tilde{F}_{Y,X,U}) = F_{Y,X}\}$.

is a singleton.

∵ model is not refuted

∴ non empty

let $a \in \mathcal{H}_I$. $b \in \mathcal{H}_I$.

by definition of $\mathcal{H}_I$, $\exists\ F^a_{Y,X,U} \in \tilde{\mathcal{F}}$ such that $a = \theta(F^a_{Y,X,U})$

since $F^a_{Y,X,U} \in \tilde{\mathcal{F}}$, $\theta(F^a_{Y,X,U}) = g(F^a_{Y,X})$ by assumption.

therefore $a = g(F^a_{Y,X})$.

by definition of $\mathcal{H}_I$: $\exists\ F^b_{Y,X,U} \in \tilde{\mathcal{F}}$ such that $b = \theta(F^b_{Y,X,U})$

since $F^b_{Y,X,U} \in \tilde{\mathcal{F}}$, $\theta(F^b_{Y,X,U}) = g(F^b_{Y,X})$ by assumption

therefore $b = g(F^b_{Y,X})$.

Finally, notice that

$F^a_{Y,X} = MakeData(F^a_{Y,X,U}) = F_{Y,X}$, $F^b_{Y,X} = MakeData(F^b_{Y,X,U}) = F_{Y,X}$

∴ $a = g(F_{Y,X}) = b$

☆.

A parameter is point identified if it can be expressed as a known function of the distribution of the observable random variables.

Causality 2. Experiments · CC 12.13.

• Identifying Treatment Effects From the Data Alone.

· The fundamental Problem of Causal Inference:

$\mathcal{X}$: the set of logically possible value of treatment.

$X_i \in \mathcal{X}$: realized treatment for i.

Y_i: realized outcome. $Y_i = Y_i(X_i)$.

where $Y_i(x)$ for all $x \in \mathcal{X}$ are potential outcomes.

$Y_i(x)$ for all $x \neq X_i$: counterfactuals.

Unit: $Y_i(x_1) - Y_i(x_0)$.

Problem: we just observe one value of x in the data.

Theorem: Let $\mathcal{Y}$ denote the set of all logically possible values of potential outcomes. For any $i \in \mathcal{X}$, the identified set for $Y_i(x)$ for any $x \in \mathcal{X}$ with $x \neq X_i$ is $\mathcal{Y}$.

Without further assumptions data tells us nothing about counterfactual outcomes.

learning $\uparrow$: main identification problem.

· Relationship btw realized & potential outcomes.

$$Y_i = Y_i(X_i)$$
$$\Leftrightarrow Y_i = \sum_{x \in \mathcal{X}} \mathbb{I}[X_i = x] Y_i(x)$$

In binary Case: $Y_i = Y_i(1) X_i + Y_i(0)(1 - X_i)$

· Marginal Distribution of Potential Outcomes.

Consider some treatment $x \in \mathcal{X}$.

$$\mathbb{P}(Y(x) \le y) = \mathbb{P}(Y(x) \le y \mid X = x)\mathbb{P}(X = x) + \mathbb{P}(Y(x) \le y \mid X \neq x)\mathbb{P}(X \neq x)$$
$$= \underbrace{\mathbb{P}(Y \le y \mid X = x)}_{\text{known}}\mathbb{P}(X = x) + \underbrace{\mathbb{P}(Y \le y \mid X \neq x)}_{\text{unknown}}\mathbb{P}(X \neq x)$$

A distr. of counterfactual outcomes.

Theorem. $\text{supp}(Y(x)) \subseteq S \ \forall x \in \mathcal{X}$.

For fixed $y \in \mathbb{R}$, the no assumptions identified set for $\mathbb{P}(Y(x) \le y)$:

$$[\mathbb{P}(Y \le y \mid X = x)\mathbb{P}(X = x),\ \mathbb{P}(Y \le y \mid X = x)\mathbb{P}(X = x) + \mathbb{P}(X \neq x)]$$

Conditional on $X = x$, $\mathbb{P}(Y(x) \le y \mid X = x) - \mathbb{P}(Y(x) \le y)$

(usually $\mathbb{P}(Y \le y)$ and $\mathbb{P}(Y(x) \le y)$ are not the same.) $= \mathbb{P}(Y \le y \mid X = x) - \mathbb{P}(Y(x) \le y)$

↑ selection bias

· The Joint Identified Set:

We want to know Treatment Effect $Y(x_1) - Y(x_0)$

must know $\mathbb{P}(Y(x_1) \le y_1, Y(x_0) \le y_0)$

Theorem (X). Let $x_1, x_0 \in \mathcal{X}$. The identified set for $F_{Y(x_1), Y(x_0)}(\cdot, \cdot)$ is the set of all joint cdfs whose marginal cdfs are identified.

→ We cannot identify $\mathbb{P}(Y(x) \le y)$ for all $x \in \mathcal{X}$ because we can't observe the counterfactual. That would require $\mathbb{P}(X = x) = 1 \ \forall x \in \mathcal{X}$ which is impossible since we have $\sum_{x \in \mathcal{X}} \mathbb{P}(X = x) = 1$.

· The Average Treatment Effect:

Suppose $\text{supp}[Y(x)] \subseteq \mathcal{Y} \ \forall x \in \mathcal{X}$.

Normalization: $Y^{norm}(x) = \dfrac{Y(x) - y_{min}}{y_{max} - y_{min}}$ $\quad y_{max} = \sup \mathcal{Y}$, $y_{min} = \inf \mathcal{Y}$.

The identified set of mean potential outcomes:

$$\mathbb{E}[Y(x)] \in [LB(x), UB(x)]$$
$$= [\mathbb{E}[Y \mid X = x]\mathbb{P}(X = x),\ \mathbb{E}[Y \mid X = x]\mathbb{P}(X = x) + \mathbb{P}(X \neq x)]$$

The identified set for ATE:

$\mathbb{E}[Y(x_1)] \in [LB(x_1), UB(x_1)]$

$\mathbb{E}[Y(x_0)] \in [LB(x_0), UB(x_0)]$.

By Theorem (X):

Joint identified set $= [LB(x_1), UB(x_1)] \times [LB(x_0), UB(x_0)]$.

$\therefore \underbrace{\mathbb{E}[Y(x_1)] - \mathbb{E}[Y(x_0)]}_{\text{ATE}} \in [LB(x_1) - UB(x_0),\ UB(x_1) - LB(x_0)]$.

WIDTH $= 2 - [\mathbb{P}(X = x_1) + \mathbb{P}(X = x_0)]$

proof: $ATE = \mathbb{E}[Y(x_1) - Y(x_0)]$

$= \mathbb{E}[Y(x_1) \mid X = x_1]\mathbb{P}(X = x_1) + \mathbb{E}[Y(x_1) \mid X \neq x_1]\mathbb{P}(X \neq x_1)$

By LIE. $-[\mathbb{E}[Y(x_0) \mid X = x_0]\mathbb{P}(X = x_0) + \mathbb{E}[Y(x_0) \mid X \neq x_0]\mathbb{P}(X \neq x_0)]$.

$= \mathbb{E}[Y(x_1) \mid X = x_1]\mathbb{P}(X = x_1) - \mathbb{E}[Y(x_0) \mid X = x_0]\mathbb{P}(X = x_0)$.

$+ \underbrace{\mathbb{E}[Y(x_1) \mid X \neq x_1]}_{\text{unknown}}\mathbb{P}(X \neq x_1) - \underbrace{\mathbb{E}[Y(x_0) \mid X \neq x_0]}_{\text{unknown}}\mathbb{P}(X \neq x_0)$.

$\therefore \overline{ATE} = \mathbb{E}[Y(x_1) \mid X = x_1]\mathbb{P}(X = x_1) + \mathbb{E}[Y(x_0) \mid X \neq x_0]\mathbb{P}(X = x_0)$

$+ 1 \cdot \mathbb{P}(X \neq x_1) - 0 \cdot \mathbb{P}(X \neq x_0)$.

$\underline{ATE} = \mathbb{E}[Y(x_1) \mid X = x_1]\mathbb{P}(X = x_1) + \mathbb{E}[Y(x_0) \mid X \neq x_0]\mathbb{P}(X = x_0)$

$+ 0 \cdot \mathbb{P}(X \neq x_1) - 1 \cdot \mathbb{P}(X \neq x_0)$.

$\therefore$ width $= \mathbb{P}(X \neq x_1) + \mathbb{P}(X \neq x_0)$.

$= 1 - \mathbb{P}(X = x_1) + 1 - \mathbb{P}(X = x_0)$

$= 2 - [\mathbb{P}(X = x_1) + \mathbb{P}(X = x_0)]$.

Randomized Experiments.

$x \in \mathcal{X}$: all logically feasible Treatment Values.
$Y(x)$: potential outcomes.
$Y = Y(X)$: realized outcomes.

• Identification of Marginal Distributions of Potential Outcomes:

★ **Theorem:** Suppose $X \perp\!\!\!\perp \{Y(x) : x \in \mathcal{X}\}$ (randomization <=> statistical independence assumption.)

Suppose the joint dist. of (Y, X) is known. Then, $F_{Y(x)}$ is point identified for all $x \in \mathcal{X}$. (marginal distribution of any potential outcome.)

proof: $\forall\, y \in \mathbb{R},\ x \in \text{supp}(X)$:

$\mathbb{P}(Y \le y \mid X = x) = \mathbb{P}(Y(X) \le y \mid X = x)$ (known in data.)
$= \mathbb{P}(Y(x) \le y \mid X = x)$
by randomization: $= \mathbb{P}(Y(x) \le y)$
$= F_{Y(x)}(y)$.

Means: $\mathbb{P}(Y \le \cdot \mid X = x) = \mathbb{P}(Y(x) \le \cdot \mid X = x) = \mathbb{P}(Y(x) \le \cdot \mid X = x')$
$\forall x, x' \in \mathbb{R}$, know one, know the others.

Look at the dist. of observed outcomes among people with treatment values $X = x$. to learn about the counterfactual outcomes of the people with $x' \neq x$.

Likewise, any parameter that depends only on the marginal distributions of $Y(x)$ is point identified.

$ATE(x_0 \to x_1) = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$
$QTE(\tau; x_0 \to x_1) = Q_{Y(x_1)}(\tau) - Q_{Y(x_0)}(\tau)$.

• Identification of the joint distribution of potential outcomes:

partially.
Since we only observe one value of $(Y_i(1), Y_i(0))$ for each i.

• Interpreting Randomization.

randomization <=> statistical independence assumption
$X \perp\!\!\!\perp \{Y(x) : x \in \mathcal{X}\}$

Meaning:
Fix some $x_{fix} \in \mathcal{X}$, look at dist of $Y(x_{fix})$ in diff. groups:
the set of units with $X = x_i$ for $x_i \in \mathcal{X}$, $x_i \neq x_j$
in each group, dist. of $Y(x_{fix})$ is the same.

In binary setting:
$Y(x) \mid X = 1$ has the same dist of $Y(x) \mid X = x_0$.
$\forall\, x \in \{0, 1\}$.

• Covariates in experiments:

W: vector of covariates.
Distinguish: ① W_{pre}. Not causally affected by Treatment
$X \perp\!\!\!\perp (\{Y(x) : x \in \mathcal{X}\}, W_{pre})$
② W_{post}: Might be affected.
$X \not\perp\!\!\!\perp (\{Y(x) : x \in \mathcal{X}\}, W_{post})$.

5 reasons to include:

• To test randomization

Suppose treatment randomly assigned.
$\Rightarrow X \perp\!\!\!\perp (\{Y(x) : x \in \mathcal{X}\}, W_{pre}) \Rightarrow X \perp\!\!\!\perp W_{pre}$ (falsifiable)
Compare $\mathbb{P}(W_{pre} \mid X = 1)$ with $\mathbb{P}(W_{pre} \mid X = 0)$
If different, then: $X \not\perp\!\!\!\perp (\{Y(x) : x \in \mathcal{X}\}, W_{pre})$

• Improve Precision.

Suppose Treatment randomly assigned.
$X \perp\!\!\!\perp (\{Y(x) : x \in \mathcal{X}\}, W_{pre})$.
Suppose X is binary treatment, then:
$ATE = \mathbb{E}[Y \mid X = 1] - \mathbb{E}[Y \mid X = 0]$
Consider Two OLS estimands:
①. Y on $(1, X)$: asymptotic variance of OLS estimator is
$$\frac{\text{var}(Y^{\perp X})}{\text{var}(X)}.$$
② Y on $(1, X, W^{pre})$. $X \perp\!\!\!\perp W_{pre} \to$ coefficient on X same.
Asymptotic covariance of OLS on estimator is:
$$\frac{\text{var}(Y^{\perp X, W_{pre}})}{\text{var}(X^{\perp W_{pre}})}.$$
Since $X \perp\!\!\!\perp W_{pre}$, $X^{\perp W_{pre}} = X$
$\therefore = \dfrac{\text{var}(Y^{\perp X, W_{pre}})}{\text{var}(X)}$

since $\text{Var}(A^{\perp B}) \le \text{Var}(A)$, $(A^{\perp B})^{\perp C} = A^{\perp B, C}$,
$$\frac{\text{var}(Y^{\perp X, W_{pre}})}{\text{var}(X)} \le \frac{\text{var}(Y^{\perp X})}{\text{var}(X)}.$$

• Stratified Randomized Experiments.

researchers assign treatments randomly within groups defined by pre-treatment covariates W.

Theorem: Suppose $X \perp\!\!\!\perp \{Y(x) : x \in \mathcal{X}\} \mid W$
Then $F_{Y(x)|W}(\cdot \mid w)$ is point identified for all $x \in \text{supp}(X \mid W = w)$, $w \in \text{supp}(W)$.

pf: $\forall\, y \in \mathbb{R}$ and $(w, x) \in \text{supp}(W, X)$,

$\mathbb{P}(Y \le y \mid W = w, X = x) = \mathbb{P}(Y(X) \le y \mid W = w, X = x)$ (known in the data.)
$= \mathbb{P}(Y(x) \le y \mid W = w, X = x)$
$= \mathbb{P}(Y(x) \le y \mid W = w)$
$= F_{Y(x)|W}(y \mid w)$

Proposition Suppose $X \perp\!\!\!\perp \{Y(x): x \in \mathcal{X}\} \mid W$.
Then $F_{Y(x)}(\cdot)$ not necessarily $= F_{Y|X}(\cdot|x)$.

• Use Experiments to Learn about Subgroup Treatment Effect

Suppose treatment unconditionally randomly assigned.
$X \perp\!\!\!\perp (\{Y(x): x \in \mathcal{X}\}, W_{pre})$.

Lemma: $A \perp\!\!\!\perp (B, C) \Rightarrow A \perp\!\!\!\perp B \mid C$

Proposition: Suppose $X \perp\!\!\!\perp (\{Y(x): x \in \mathcal{X}\}, W_{pre})$, Then $F_{Y(x)|W_{pre}}(\cdot|w)$ is point identified for any $x \in \mathcal{X}$, $w \in supp(W_{pre})$.
$CATE = \mathbb{E}[Y(1) - Y(0) \mid W_{pre} = w]$

Selection on Observables.

Randomly Assigned Treatments with Observational Data.

not experimental data.

In observational data setting, $X \perp\!\!\!\perp \{Y(x): x \in \mathcal{X}\} \mid W$ is the unconfoundness assumption. (= conditional independence).
→ Binary Treatment:

Assumption ①, unconfoundness: non falsifiable.
$X \perp\!\!\!\perp (Y(1), Y(0)) \mid W$

Assumption ②: overlap: (sufficient variation).
$\mathbb{P}(X=1 \mid W=w) \in (0,1)$ for all $w \in supp(W)$
falsifiable.

Main Identification Results:

$Y = Y(X)$ potential outcomes.

Theorem: Suppose the unconfoundness and overlap hold.
Suppose the distribution of (Y, X, W) is observed.
Then: ① $F_{Y(x)|W}(\cdot|w)$ is point identified for all $x \in \{0,1\}$ and $w \in supp(W)$.
② $F_{Y(x)}$ is point identified for all $x \in \{0,1\}$.

proof: Part 1:

$$\forall\, y \in \mathbb{R}, (w, x) \in supp(W, X):$$
$$\mathbb{P}(Y \le y \mid W=w, X=x) = \mathbb{P}(Y(x) \le y \mid W=w, X=x)$$
$$= \mathbb{P}(Y(x) \le y \mid W=w, X=x)$$
$$= \mathbb{P}(Y(x) \le y \mid W=w)$$
$$= F_{Y(x)|W}(y \mid w)$$

Overlap: $\forall\, x \in \{0,1\}$, LHS is known for all $w \in supp(W)$.
Thus the RHS is known for all $x \in \{0,1\}$ and $w \in supp(W)$.

Part 2:

$$F_{Y(x)}(y) = \mathbb{P}(Y(x) \le y)$$
$$= \mathbb{E}[\mathbb{P}(Y(x) \le y \mid W)]$$
$$= \mathbb{E}[F_{Y(x)|W}(y \mid W)]$$

RHS: ① marginal distribution of W ($\mathbb{E}$). known from data.
② $F_{Y(x)|W}(\cdot|w)$ for all $w \in supp(W)$. known from Part 1.

The theory implies that any parameters that only depends on $F_{Y(x)|W}(\cdot|w)$ and $F_{Y(x)}$ is point identified.

Including:
① CATE: $CATE(w) = \mathbb{E}[Y(1) - Y(0) \mid W = w]$
② ATE: $ATE(w) = \mathbb{E}[Y(1) - Y(0)]$
$= \mathbb{E}[CATE(W)]$
③ CATT: $CATT(w) = \mathbb{E}[Y(1) - Y(0) \mid W = w, X = 1]$
④ ATT: $ATT = \mathbb{E}[Y(1) - Y(0) \mid X = 1]$
⑤: τ-th QTE: $QTE(\tau) = Q_{Y(1)}(\tau) - Q_{Y(0)}(\tau)$.
⑥: τ-th QTT: $QTT(\tau) = Q_{Y(1)|X}(\tau|1) - Q_{Y(0)|X}(\tau|0)$.

Accessing Unconfoundness: What to control for?

Set up. Y: outcome variable of interest.
X: treatment
$W = (W_1 \cdots W_K)$ a vector of covariates.
U: a vector of unobserved variables.

$$(*) \begin{cases} Y \leftarrow g(X, W, U) \\ X \leftarrow h(Y, W, U) \\ W_k \leftarrow m_k(Y, X, W_{-k}, U) \text{ for } k = 1 \cdots K. \end{cases}$$

A graph G: a set of nodes and a set of edges between them.
A directed G: have →. ordered pair of nodes.
A causal graph: nodes are variables.

An walk from v_i to v_j is a sequence of nodes $x_0 \cdots x_k$ such that ① $x_0 = v_i$ ② $x_k = v_j$ ③ there is an edge btw x_t and x_{t+1}, $\forall\, t \in \{0, 1, \cdots, k-1\}$.
A directed walk: all arrows are in some direction.
A path: all the nodes in a walk are different.
+ directed walk = directed path.
A directed cycle: a directed walk that starts and ends at the same node, all others are distinct.
A acyclic graph does not contain any directed cycles.
DAGs: directed acyclic graph.

Baseline Assumptions:

① Initial conditions: U in (*) is external.
no → U, only U →.
② No simultaneity: causal graph of (*) is acyclic.
not possible for two variables to affect each other
①&② → for any joint dist. of U, (*) yields a unique distribution of (Y, X, W).

Simple example:

W: a scalar.

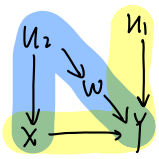
$Y \leftarrow g(X, W, U_1)$

$X \leftarrow h(U_2)$

$W \leftarrow m(U_2)$.

U_1, U_2 independent

Causal Graph:



path: Y collider

path: Y not collider.

X & U d-separated: every path has a collider Y

X & Y d-connected: there is a path with no collider.

Conditional Independence in Causal Graphs.

structure of the causal graph → unconfoundness assumption

d-separation

Definition: Consider a directed graph between V_1 and V_2. A collider on this path is any node on the path with two incoming arrows. (common effect) d: directional.

Definition: A path is d-connected if it contains no collider.

Definition: The nodes V_1 and V_2 is d-connected if there is at least one path between them that is d-connected.

If no path between them is d-connected, then V_1 and V_2 are d-separated.

Including Conditioning Variables.

Definition: Let W be a set of variables. V_1 and V_2 are d-connected given W if there is at least one path from V_1 to V_2 that satisfies all of the following criteria:

1. there are no chains with $W \in \mathcal{W}$ as a middle node.
   that is: no $A \to W \to B$.
2. there are no forks with $W \in \mathcal{W}$ as a middle node.
   that is: no $A \leftarrow W \rightarrow B$.
3. for all colliders C on the path, $(A \to C \leftarrow B)$:
   Either: a. C is in W
   b. at least one of the descendants of C is in W.

e.g. 

X & Y d-separated.

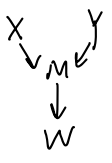
X & Y d-separated given W_1, since $\to W_2 \leftarrow$

X & Y d-connected given W_1, W_2, since $\to W_1 \leftarrow$ & $\to W_2 \leftarrow$

X & Y d-connected (no collider)

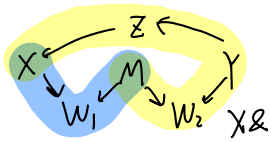
X & Y d-separated given W ($Y \to W \to X$).

X & V d-separated ($U \to Y \leftarrow V$).



X & Y d-separated

X & Y d-connected given W ($M \to W$).



X & M d-separated.

X & M d-connected given W_2.

X & M d-separated given (W_2, Z)

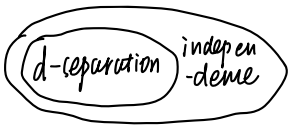
d-separation implies independence.

Theorem: Let G be a causal graph corresponding to (*). Suppose assumption ① & ② holds. Let X, Y, W be subsets of variables (Y, X, W, U). Then for any joint dist of U:

X and Y are d-separated given W

$\Downarrow$

$X \perp\!\!\!\perp Y \mid W$.



However,

$X \perp\!\!\!\perp Y \mid W \not\Rightarrow$ d-separated.

Intuition behind d-separation:

① Mediators:

Structural System: $Y \leftarrow \beta_{Y,M} M + U_Y$, $M \leftarrow \beta_{M,X} X + U_M$, $X \leftarrow U_X$.

all U mutually indep.

$X \to M \to Y$

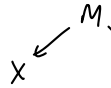
X & Y d-connected $X \not\perp\!\!\!\perp Y$

X & Y d-separated given M $X \perp\!\!\!\perp Y \mid M$

② Common Causes.

Structural System: $Y \leftarrow \beta_{Y,M} M + U_Y$, $X \leftarrow \beta_{X,M} M + U_X$, $M \leftarrow U_M$

all U mutually indep.



X & Y d-connected $X \not\perp\!\!\!\perp Y$

X & Y d-separated given M $X \perp\!\!\!\perp Y \mid M$

③ Colliders

Structural System: $M \leftarrow \beta_{M,X} X + \beta_{M,Y} Y + U_M$, $X \leftarrow U_X$, $Y \leftarrow U_Y$

all U mutually indep.

X & Y d-separated $X \perp\!\!\!\perp Y$

X & Y d-connected given M $X \not\perp\!\!\!\perp Y \mid M$

$W_1 \perp\!\!\!\perp Y$. $W_1 \not\perp\!\!\!\perp Y \mid (M, W_2)$

$W_1 \not\perp\!\!\!\perp Y \mid W_2$.

Deriving Sufficient Conditions for Unconfoundness

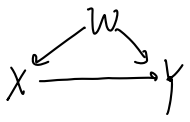
Start from:

$Y \leftarrow g(X, W, U_Y)$

$X \leftarrow h(W, U_X)$

$W \leftarrow U_W$

U's are mutually indep.



Define potential outcome: $Y(x) = g(x, W, U_Y)$

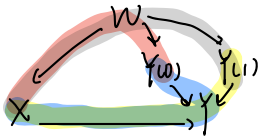
then: $Y \leftarrow g(X, W, U_Y)$

$Y(x) \leftarrow g(x, W, U_Y)$

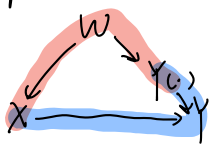
$X \leftarrow h(W, U_X)$

$W \leftarrow U_W$

Then we have the following graph



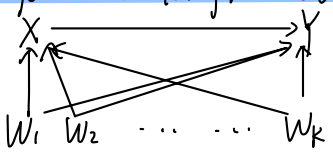
or,



Using the graph theory, X & $Y(\cdot)$ are desseparated condition on W. Thus, $X \perp\!\!\!\perp Y(x) \mid W$. unconfoundness.

In general, condition on post-treatment outcomes is dangerous.

The Unconfoundness Template Causal Graph:



$X \perp\!\!\!\perp Y \mid (W_1, W_2, \cdots, W_K)$

Generally, should control for all observed pre-treatment variables.

Should avoid imprecision from over control:

It has two drawbacks.

① increase standard errors in the estimators.

$X \perp\!\!\!\perp Y(x)$

$X \perp\!\!\!\perp Y(x) \mid W$

But, control for W reduces the remaining variation in X.

② make overlap worse

$p(w) = P(X=1 \mid W=w) \to 0$ or $\to 1$ when W includes too many.

Proposition: There are always at least two different graphs in $\mathcal{G}$, where $\mathcal{G}$ is the set of all causal graphs where $X \perp\!\!\!\perp Y(x) \mid W$.

pf: ① $X \to Y$

② $X \leftarrow W \to Y$

Thus, assumptions on causal graphs provides sufficient, not necessary, conditions for unconfoundness.

Benefits of Causal Graphs: ① arguable more interpretable than unconfoundness alone ② make the model falsifiable.

The Classical Linear Model for Potential Outcomes:

Functional form of potential outcomes are linear in a set of known functions of the treatment with constant coefficients.

Benefits: ① treatment effect parameters equivalent to certain OLS estimands ② more subtle identification results ③ solve extrapolation identifications.

$X \in \mathcal{X}$: a vector of treatment variables.

$\mathcal{X} \subseteq \mathbb{R}^k$: all logical values of these variables.

$Y(x)$: potential outcomes $\forall x \in \mathcal{X}$

Assumption: $\forall x \in \mathcal{X}$, $Y(x) = p(x)'\beta + U$. $p(x)$: J vector.

restrictions: ① linear functional form

② homogeneous treatment effect β.

$Y_i(x_{i1}) - Y_i(x_{i0}) = (p(x_{i1}) - p(x_{i0}))'\beta$.

Exogeneity.

① full independence: $X \perp\!\!\!\perp Y(x)$

$\Downarrow$

under linear functional form: $X \perp\!\!\!\perp U$

$\Downarrow$

② mean independence: $\mathbb{E}[U \mid X = x] = \mathbb{E}[U]$

$\Downarrow$

③ uncorrelatedness: $\text{corr}(U, p_j(X)) = 0$.

Theorem: Identification of the classical linear model:

Suppose $F_{Y,X}$ satisfies.

① Finite moments on observables: $\mathbb{E}[p(X)Y] < \infty$, $\mathbb{E}[p(X)p(X')] < \infty$

② Sufficient Variation: $\mathbb{E}[p(X)p(X)']$ invertible.

③ Linear potential outcomes

④ finite moments on unobservables: $\mathbb{E}[p(X)U] < \infty$, $\mathbb{E}[U] < \infty$.

⑤ Exogeneity: $\mathbb{E}[p(X)U] = \mathbb{E}[p(X)]\mathbb{E}[U]$.

$(\text{corr}(U, p_j(X)) = 0$

⑥ normalization: $\mathbb{E}[U] = 0$.

Then β is point identified.

$\beta = \mathbb{E}[p(X)p(X)']^{-1}\mathbb{E}[p(X)Y]$.

① ④ ⑤ not falsifiable, then the identified set for β is a singleton. $B_I = \{\mathbb{E}[p(X)p(X)']^{-1}\mathbb{E}[p(X)Y]\}$.

proof: $Y = Y(x)$

$= p(x)'\beta + U$

$p(X)Y = p(X)p(X)'\beta + p(X)U$.

$\mathbb{E}[p(X)Y] = \mathbb{E}[p(X)p(X)']\beta + \underbrace{\mathbb{E}[p(X)U]}$.

$= \mathbb{E}[p(X)]\mathbb{E}[U] = 0$.

$\beta = \mathbb{E}[p(X)p(X)']^{-1}\mathbb{E}[p(X)Y]$

Suppose $p(U)$ does not contain a constant.

Let $\tilde{p}(X) = \begin{bmatrix} 1 \\ p(X) \end{bmatrix}$ then: $\mathbb{E}[\tilde{p}(X)\tilde{p}(X)']$ invertible.

define $\tilde{\beta} = \begin{bmatrix} \mathbb{E}[U] \\ \beta \end{bmatrix}$, $\tilde{U} = U - \mathbb{E}[U]$

then $Y = \tilde{p}(x)'\tilde{\beta} + \tilde{U}$ holds.

To prove B_I is a singleton: have to construct $(\tilde{\beta}, \tilde{F}_{U,X})$ consistent with $F_{Y,X}$ such that $\tilde{\beta} = \mathbb{E}[p(X)p(X)']^{-1}\mathbb{E}[p(X)Y]$.

① define $\tilde{\beta} = \begin{bmatrix} \mathbb{E}[U] \\ \beta \end{bmatrix}$.

② define $\tilde{U} = Y - p(X)'\tilde{\beta}$

Thus, $\tilde{F}_{X,U} = F_{X,U}$. $(\tilde{\beta}, \tilde{F}_{X,U})$ consistent with linear ...

$\mathbb{E}[p(X)\tilde{U}] = 0$, $\mathbb{E}[\tilde{U}] = 0$ if $p(X)$ consists a constant.

$\therefore$ this construction is consistent with all model assumptions.

$$P(Y \le y \mid X=x) = P(p(x)'\tilde{\beta} + \tilde{u} \le y \mid X=x)$$
$$= P(\tilde{u} \le y - p(x)'\tilde{\beta} \mid X=x)$$
$$= F_{\tilde{u}|X}(y - p(x)'\tilde{\beta} \mid x).$$

Alternative proof:

Suppose $p(x) = \begin{pmatrix} 1 \\ x \end{pmatrix}$, $Y(x) = \beta_0 + \beta_1 x + U$.

$$cov(Y, x) = cov(\beta_0 + \beta_1 X + U, X)$$
$$= 0 + \beta_1 var(X) + cov(U, X)$$
$$= \beta_1 var(X)$$
$$\therefore \beta_1 = \frac{cov(Y, X)}{var(X)}$$

$cov(U, X) = E[UX] - E[U]E[X]$
$= E[U]E[X] - E[U]E[X]$
$= 0$

Connecting Regression and Causality

The linear potential outcomes model with uncorrelatedness gives a causal interpretation of OLS estimand.

Corollary. Suppose all the assumptions hold.

Let $E = Y - p(X)'\beta^{ols}$

where $\beta^{ols} = E[p(X)p(X)']^{-1}E[p(X)Y]$

Then, $U \equiv E$.

Interpretation: under uncorrelatedness assumption about X & U in linear potential outcome outcomes, U (unobserved heterogeneity) equals E (regression residual, which is by construction uncorrelated to the regressors $p_j(X)$)

Identification without the sufficient variation assumption.

Lemma. Suppose Sufficient Variation does not hold.

① β is not point identified.

② The identified set: $B_I = \{b \in \mathbb{R}^J : E[p(x)p(x)']b = E[p(x)Y]\}$

Identification under mean independence.

Lemma: Suppose $F_{Y,X}$ satisfies: ① sufficient variation ② linear potential outcomes ③ finite moments ④ exogeneity.

(If $p(x)$ contains a constant, then Normalization holds)

Then, β and $F_{X,U}$ are point identified.

Proposition: Suppose all the assumptions hold. If $p(X)$ does not contain a constant,

$$E[Y \mid X=x] = p(x)'\beta + E[U]$$

If $p(X)$ does contain a constant,

$$E[Y \mid X=x] = p_{-1}(x)'\beta_{-1} + (\beta_1 + E[U]).$$

This implies that the model is falsifiable since the conditional mean function must satisfy this specific form.

proof: $E[Y \mid X=x] = E[p(x)'\beta + u \mid X=x]$
$= E[p(x)'\beta \mid X=x] + E[U \mid X=x]$
$= p(x)'\beta + E[U]$

Connecting Regression and Causality.

Step ①. Linear potential outcome model implies

$$ASF(x) = p(x)'\beta + E[U]$$

Step ②. add exogeneity assumption → β point identified.

under mean independence, $E[Y \mid X=x] = p(x)'\beta + E[U]$.

Theorem Suppose all the assumptions satisfy. then

$$ASF(x) = p(x)'\beta + E[U] = E[Y \mid X=x]$$

And it is point identified.

Thus $E[Y \mid X=x]$ has a causal interpretation.

Identification under full independence

Assumption: Exogeneity: $X \perp\!\!\!\perp U$ Implies $Y(x) \perp\!\!\!\perp X$).

Full independence has further falsifying power.

Proposition. Suppose $F_{Y,X}$ satisfies sufficient variation

Let $E[Y \mid X=x] = p(x)'\beta + E[U]$.

Define $\tilde{U} = Y - (p(X)'\beta + E[U])$

Then, linear potential outcomes, finite moments and exogeneity are falsified iff $\tilde{U} \not\perp\!\!\!\perp X$.

$$var(Y \mid X=x) = var(p(X)'\beta + U \mid X=x)$$
$$= var(p(x)'\beta + U \mid X=x)$$
$$= var(U \mid X=x)$$
$$= var(U) \quad (\text{since } X \perp\!\!\!\perp U).$$

→ homoskedasticity

Covariates in the Linear Model.

Reasons to include covariates,

① to test assumptions like unconfoundness.

② to improve the statistical precision of estimators.

③. ★ necessary for causality identification analysis. Here!

④. need them to define other parameter of interest.

⑤ use them to solve identification problems like missing data.

Nonparametric unconfoundness:

$$\{Y(x) : x \in \mathcal{X}\} \perp\!\!\!\perp X \mid W$$

Also assume $Y(x) = p(x)'\beta + U$, then:

$$\{Y(x) : x \in \mathcal{X}\} \perp\!\!\!\perp X \mid W \iff U \perp\!\!\!\perp X \mid W$$

$$E[U \mid X=x, W=w] = E[U \mid W=w]$$

$$E[Y \mid X=x, W=w] = E[Y(x) \mid W]$$

(↑ known)

$$= p(x)'\beta + E[U \mid W=w].$$

(↑ to identify this, need some version of sufficient variation assumption.)

Linear Unconfoundness ⇐ nonparametric unconfoundness #7

$$Y(x) = p(x)'\beta + u.$$

Here, we use "uncorrelatedness" after partialling out W.

Assumption: $corr(p_j(X)^{\perp W}, u^{\perp W}) = 0$ for all $j \in \{1, \cdots, J\}$

Remember: $cov(A^{\perp C}, B^{\perp C}) = cov(A^{\perp C}, B) = cov(A, B^{\perp C})$

then: $corr(p_j(X)^{\perp W}, u) = 0$, $corr(p_j(X), u^{\perp W}) = 0$.

Identification:

Suppose: ① linear potential outcomes ② finite moment, sufficient variation, $var(Y, p(X), W)$ exists, positive definite

③. linear unconfoundness.

Then: 1. β and $\tau_{x,u}$ point identified.

2. $\beta = \mathbb{E}[p(X)^{\perp W}(p(X)^{\perp W})']^{-1}\mathbb{E}[p(X)^{\perp W}Y^{\perp W}]$

by FWL, β is the coefficient on $p(X)$ in OLS of Y on $(p(X), W)$.

3. $\beta_2 = \{\mathbb{E}[p(X)^{\perp W}(p(X)^{\perp W})']^{-1}\mathbb{E}[p(X)^{\perp W}Y^{\perp W}]\}$.

Suppose that $p(x) = \begin{bmatrix}1\\ x\end{bmatrix}$. Let $\beta = \begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix}$

then: $Y(x) = \beta_0 + \beta_1 x + u$. since $(A+B)^{\perp C} = A^{\perp C} + B^{\perp C}$

$$\begin{aligned} cov(Y^{\perp W}, X^{\perp W}) &= cov(\beta_0 + \beta_1 X^{\perp W} + u^{\perp W}, X^{\perp W}) \\ &= 0 + \beta_1 cov(X^{\perp W}, X^{\perp W}) + cov(u^{\perp W}, X^{\perp W}) \\ &= \beta_1 var(X^{\perp W}). \end{aligned}$$

sufficient variation $\Rightarrow$ $\beta_1 = \dfrac{cov(Y^{\perp W}, X^{\perp W})}{var(X^{\perp W})}$

★ With linear unconfoundness assumptions, coefficients on $p(X)$ can be interpreted causally. Coefficient on W cannot!

In practice, most covariates are thought to be endogenous.

The bias from including outcomes as control variables

① endogenous themselves

② destroys the exogeneity of X. ★

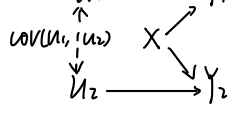
proof 1st.

Set up.

Suppose we have two outcomes. Y_1 Y_2

$Y_1(x) = \alpha_1 + \beta_1 x + u_1$

$Y_2(x) = \alpha_2 + \beta_2 x + u_2$



Suppose $X \perp\!\!\!\perp (u_1, u_2)$,

then, β_1 is point identified.

If include Y_2 as control, do OLS of Y_1 on $(1, X, Y_2)$, then

$$\text{coeff. on } X = \frac{cov(Y_1, X^{\perp Y_2})}{var(X^{\perp Y_2})}$$

Let $(\tilde{\beta_0}, \tilde{\beta_{Y_2}})$ be the coefficient of X on $(1, Y_2)$, residual $= X^{\perp Y_2}$

$$\begin{aligned} cov(Y_1, X^{\perp Y_2}) &= cov(\beta_1 X_1 + u_1, X^{\perp Y_2}) \\ &= \beta_1 cov(X_1, X^{\perp Y_2}) + cov(u_1, X^{\perp Y_2}) \\ &= \beta_1 var(X^{\perp Y_2}) + cov(u_1, X - [\tilde{\beta_0} + \tilde{\beta_{Y_2}} Y_2]) \end{aligned}$$

$$= \beta_1 var(X^{\perp Y_2}) - \tilde{\beta_{Y_2}} cov(u_1, Y_2).$$

Hence $\dfrac{cov(Y_1, X^{\perp Y_2})}{var(X^{\perp Y_2})} = \beta_1 - \tilde{\beta_{Y_2}} \dfrac{cov(u_1, Y_2)}{var(X^{\perp Y_2})}$

Next, $cov(Y_2, u_1) = cov(\alpha_2 + \beta_2 X + u_2, u_1)$

$$= cov(u_2, u_1)$$

Moreover, $\tilde{\beta}_{Y_2} = \dfrac{cov(X, Y_2)}{var(Y_2)}$

$$= \frac{cov(X, \alpha_1 + \beta_2 X + u_2)}{var(Y_2)}$$

$$= \beta_2 \frac{var(X)}{var(Y_2)}$$

$$= \beta_2 \frac{var(X)}{var(\alpha_1 + \beta_2 X + u_2)}$$

$$= \beta_2 \frac{var(X)}{\beta_2^2 var(X) + var(u_2)}$$

Therefore, $\dfrac{cov(Y_1, X^{\perp Y_2})}{var(X^{\perp Y_2})} = \beta_1 - \beta_2 \dfrac{var(X)}{\beta_2^2 var(X) + var(u_2)} \dfrac{cov(u_2, u_1)}{var(X^{\perp Y_2})}$

If either ① $\beta_2 = 0$ (Y_2 actually not a control) or

② $cov(u_2, u_1) = 0$ ⟶ generally will not hold.

then, adding Y_2 to be a covariate will not cause bias.

proof 2nd.

Basic idea in using regression to get causality:

Variation in the conditional mean function yields causal effects.

$$\begin{aligned} \mathbb{E}[Y_1 | X = x, Y_2 = y_2] &= \mathbb{E}[\alpha_1 + \beta_1 x + u_1 | X = x, Y_2 = y_2] \\ &= \alpha_1 + \beta_1 x + \mathbb{E}[u_1 | X = x, Y_2 = y_2] \\ &= \alpha_1 + \beta_1 x + \mathbb{E}[u_1 | X = x, \alpha_2 + \beta_2 X + u_2 = y_2] \\ &= \alpha_1 + \beta_1 x + \mathbb{E}[u_1 | X = x, u_2 = y_2 - \alpha_2 - \beta_2 x] \\ &= \alpha_1 + \beta_1 x + \mathbb{E}[u_1 | u_2 = y_2 - \alpha_2 - \beta_2 x] \end{aligned}$$

$$\frac{\partial \mathbb{E}[Y_1 | X = x, Y_2 = y_2]}{\partial x} = \beta_1 + \underbrace{\frac{\partial \mathbb{E}[u_1 | u_2 = y_2 - \alpha_2 - \beta_2 x]}{\partial x}}_{*}$$

If either ① $\beta_2 = 0$, then $* = 0$ or

② u_1 mean independent of u_2

then, adding Y_2 to be a covariate will not cause bias.

Mean Independence with Endogenous Controls.

$$Y(x, w) = \beta_0 + \beta_1 x + \beta_2 w + u.$$

Mean independence assumption: $\mathbb{E}[u | X = x, W = w] = \mathbb{E}[u]$

★ Rarely possible! Instead, we put:

Assumption: Conditional mean independence (control function).

$$\mathbb{E}[u | X = x, W = w] = \mathbb{E}[u | W = w].$$

Assumption: linearity.

$\mathbb{E}[u | W = w] = \gamma_0 + \gamma_1 W$, allow for $\gamma_1 \neq 0$. ★

do OLS of Y on $(1, X, W)$:

$$\mathbb{E}[Y|X=x,W=w] = \mathbb{E}[\beta_0+\beta_1 X+\beta_2 W+U \mid X=x, W=w]$$
$$= \beta_0+\beta_1 x+\beta_2 w + \mathbb{E}[U|X=x,W=w]$$
$$= \beta_0+\beta_1 x+\beta_2 w + \mathbb{E}[U|W=w]$$
$$= \beta_0+\beta_1 x+\beta_2 w+\gamma_0+\gamma_1 w$$
$$= (\beta_0+\gamma_0)+\beta_1 x+(\beta_2+\gamma_1)w.$$

$\frac{\partial \mathbb{E}[Y|X=x,W=w]}{\partial x} = \beta_1$ even if W is endogenous!
$\therefore \beta_1$ is point identified.

But, $\frac{\partial \mathbb{E}[Y|X=x,W=w]}{\partial w} = \beta_2+\gamma_1 \neq \beta_2$, biased!

Proposition. Under the above assumptions, the coefficient on X from OLS of Y on $(1, X, W)$ is β_1.

proof. by FWL: coeff $= \frac{cov(Y, X^{\perp W})}{var(X^{\perp W})}$.

$$cov(Y, X^{\perp W}) = cov(\beta_0+\beta_1 X+\beta_2 W+U, X^{\perp W})$$
$$= \beta_1 cov(X, X^{\perp W}) + \beta_2 cov(W, X^{\perp W}) + cov(U, X^{\perp W})$$
$$= \beta_1 var(X^{\perp W}) + cov(U, X^{\perp W}).$$
$$cov(U, X^{\perp W}) = \mathbb{E}[X^{\perp W}(U-\mathbb{E}[U])]$$
$$= \mathbb{E}[X^{\perp W}U] \quad \text{assume } \mathbb{E}[U]=0$$
$$= \mathbb{E}_{X,W}[\mathbb{E}[X^{\perp W}U|X,W]]$$
$$= \mathbb{E}_{X,W}[\mathbb{E}(U|X,W)X^{\perp W}]$$
$$= \mathbb{E}_{X,W}[\mathbb{E}[U|W]X^{\perp W}]$$
$$= \mathbb{E}_{X,W}[(\gamma_0+\gamma_1 W)X^{\perp W}]$$
$$= cov(\gamma_0+\gamma_1 W, X^{\perp W}) \quad \text{since } \mathbb{E}[X^{\perp W}]=0$$
$$= 0$$

Non-parametric endogenous controls

without $\mathbb{E}[U|W] = \gamma_0+\gamma_1 W$,

$\mathbb{E}[Y|X=x,W=w] = \beta_0+\beta_1 x+\beta_2 w+\mathbb{E}[U|W=w]$

$\beta_1 = \frac{\partial \mathbb{E}[Y|X=x,W=w]}{\partial x}$ is identified!

Important: $Y(x) \perp\!\!\!\perp X \mid W \not\Leftrightarrow Y(x) \perp\!\!\!\perp W \mid X$

using conditional mean independence $\rightarrow$ allow endogenous control.

Accessing Linearity

The non-parametric additively separable model.

Mean independence.

↳ $m(x) = \mathbb{E}[Y|X=x] = p(x)'\beta + \mathbb{E}[U]$

Suppose that linear model is false.

Theorem. Suppose the joint distribution of (Y, X) is observed. And: Additively Separable Potential Outcomes, Finite Moments, Mean Independence. Then: $g(x)+\mathbb{E}[U]$ is point identified for all $x \in supp(X)$. model is not falsifiable.

proof: $\forall x \in supp(X)$:

$$\mathbb{E}[Y|X=x] = \mathbb{E}[g(x)+U|X=x]$$
$$= \mathbb{E}[g(x)+U|X=x]$$
$$= g(x)+\mathbb{E}[U|X=x]$$
$$= g(x)+\mathbb{E}[U]. \quad \text{identified.}$$

Corollary 7. Suppose all the relevant assumption hold.

1. Let $\tilde{U} = U - \mathbb{E}[U]$. Then $F_{\tilde{U}}$ is point identified.
2. $F_{Y(x)}$ is point identified for all $x \in supp(X)$
3. The joint distribution of any finite set of potential outcomes $(Y(x_1) \cdots Y(x_k))$ is point identified, so long as $x_1 \cdots x_k \in supp(X)$.

proof: 1. we have $U = Y - g(X)$

so $\tilde{U} = Y - (g(X)+\mathbb{E}[U])$.
($g(X)+\mathbb{E}[U]$: known; then known)

That is: $F_{\tilde{U}}(u) = \mathbb{P}(\tilde{U} \le u)$
$= \mathbb{P}(Y-g(X)-\mathbb{E}[U] \le u)$
$= \int_{supp(Y,X)} (y-g(x)-\mathbb{E}[U])\, dF_{Y,X}(y,x)$

It is known.

2. $Y(x) = g(x)+\mathbb{E}[U]+\tilde{U}$

$F_{Y(x)}(y) = \mathbb{P}(Y(x) \le y)$
$= \mathbb{P}(g(x)+\mathbb{E}[U]+\tilde{U} \le y)$
$= \mathbb{P}(\tilde{U} \le y-g(x)-\mathbb{E}[U])$
$= F_{\tilde{U}}(y-g(x)-\mathbb{E}[U])$ ← known. (then known)

3. for example, just consider two values $x_1, x_2 \in supp(X)$

$F_{Y(x_1),Y(x_2)}(y_1,y_2) = \mathbb{P}(Y(x_1) \le y_1, Y(x_2) \le y_2)$
$= \mathbb{P}(\tilde{U} \le y_1-g(x_1)-\mathbb{E}[U], \tilde{U} \le y_2-g(x_2)-\mathbb{E}[U])$
$= \mathbb{P}(\tilde{U} \le \min\{y_1-g(x_1)-\mathbb{E}[U], y_2-g(x_2)-\mathbb{E}[U]\})$
$= F_{\tilde{U}}(\min\{y_1-g(x_1)-\mathbb{E}[U], y_2-g(x_2)-\mathbb{E}[U]\})$
(known; then point identified.)

Cost for allowing a fully non-parametric function form:
cannot extrapolate off the support.

Full independence

Proposition: $m(x) = \mathbb{E}[Y|X=x]$. Define $\tilde{U} = Y - m(X)$

Then. Additively separable + finite moments + full independence are falsified iff $\tilde{U} \not\perp\!\!\!\perp X$.

The non-parametric potential outcomes model

replace $X \perp\!\!\!\perp U$ with: $Y(x) \perp\!\!\!\perp X$ for all $x \in supp(X)$

Thus $F_{Y(x)}(y) = \mathbb{P}(Y \le y \mid X=x)$ is point identified.

Best Linear approximation ASF $= m(x)$.

let $\beta^{OLS} = \mathbb{E}[p(X)p(X)']^{-1}\mathbb{E}[p(X)Y]$. then within a class of functions $f_{p(x)} = p(x)'b$. $p(x)'\beta^{OLS}$ is the best approximation to ASF using weighted L_2 distance.

Accessing Exogeneity

Reasons for exogeneity failure ($E[p(X)U] \neq E[p(X)]E[U]$)

① simultaneity ② Omitted variables.

Interpreting OLS without exogeneity

Omitted Variable Bias X, W scalar

$$Y(x, w) = \beta_0 + \beta_1 x + \beta_2 w + V$$

Let $Y(x) = Y(x, W) = \beta_0 + \beta_1 x + (\beta_2 W + V) = \beta_0 + \beta_1 x + U$; $U \equiv \beta_2 W + V$.

suppose $cov(X, V) = 0$, $cov(W, V) = 0$.

do OLS on $(1, X, W) \rightarrow \beta = (\beta_0, \beta_1, \beta_2)$

do OLS on $(1, X)$: $cov(Y, X) = cov(\beta_0 + \beta_1 X + \beta_2 W + V, X)$

$= \beta_1 var(X) + \beta_2 cov(W, X)$ (true)

then $\frac{cov(Y, X)}{var(X)} = \beta_1 + \beta_2 \frac{cov(W, X)}{var(X)}$

$$Bias = \frac{cov(Y, X)}{var(X)} - \beta_1 = \beta_2 \frac{cov(W, X)}{var(X)}$$

Depends on: ① β_2 ② $\frac{cov(W, X)}{var(X)}$

Different way of seeing: $cov(X, U) = cov(X, \beta_2 W + V)$

$= \beta_2 cov(X, W)$.

For mean independence: $E[V|X=x] = E[V]$.

Then $E[U|X=x] = E[\beta_2 W + V | X = x]$

$= \beta_2 E[W|X=x] + E[V|X=x]$

$= \beta_2 E[W|X=x]$

if $\beta_2 \neq 0$ and mean of W depend on X, then

$E[U|X=x] \neq E[U]$

Consider $E[Y|X=x] = E[\beta_0 + \beta_1 X + U | X = x]$

$= \beta_0 + \beta_1 x + E[U|X=x]$

$$\frac{\partial E[Y|X=x]}{\partial x} = \beta_1 + \frac{\partial E[U|X=x]}{\partial x}$$

suppose $E[W|X=x] = \delta$, then $\frac{\partial E[U|X=x]}{\partial x} = \beta_2 \delta$

$E[Y|X=x] = \beta_0 + (\beta_1 + \beta_2 \delta) x$, $\frac{\partial E[Y|X=x]}{\partial x} = \beta_1 + \beta_2 \delta$

Heterogeneous Treatment Effects: Random Coefficients

Motivation: the unit level causal effects change from person to person.

Non-parametric potential outcomes model

$$Y(x) = g(x, U).$$

$$Y_i(x_2) - Y_i(x_1) = g(x_2, U_i) - g(x_1, U_i)$$

Random Coefficient Models

β random

$$Y(x) = \beta x + U$$

$$Y_i(x) = \beta_i x + U_i$$

The OLS estimand.

$$Y(x) = \beta x + U$$

$$cov(Y, X) = E[(Y - E[Y])(X - E[X])]$$

$$= E[Y(X - E[X])] - E[Y]E[X - E[X]]$$

$$= E[Y(X - E[X])].$$

$$= E[(\beta X + U)(X - E[X])] \quad \text{assume } E[U|X=x] = E[U]$$

$$= E[\beta X (X - E[X])]$$

Also assume $E[\beta | X = x] = E[\beta]$

then $cov(Y, X) = E[\beta X (X - E[X])]$

$= E[E[\beta X (X - E[X]) | X]]$

$= E[E[\beta|X] X (X - E[X])]$

$= E[\beta] E[X(X - E[X])]$

$= E[\beta] var(X)$ then:

Theorem $Y(x) = \beta x + U$, $E[U|X] = E[U]$. $E[\beta|X=x] = E[\beta]$

then: $E[\beta] = \frac{cov(Y, X)}{var(X)}$ = average random effect

$$ATE(x_1 \rightarrow x_2) = E[Y(x_2) - Y(x_1)] = E[(\beta x_2 + U) - (\beta x_1 + U)]$$

$$= E[\beta(x_2 - x_1)] = E[\beta](x_2 - x_1)$$

Heterogeneous Treatment Effects: Quantile Regression

(most common approach).

Theorem (Equivalence). $h: \mathbb{R} \rightarrow \mathbb{R}$ weakly left-continuous.

Y: r.v. then $Q_{h(Y)}(\tau) = h(Q_Y(\tau))$.

proof: suppose F_Y strictly ↑ everywhere, h strictly ↑. h continuous.

$P(Y \leq Q_Y(\tau)) = \tau$.

$P(Y \leq y) = P(h(Y) \leq h(y)) \Rightarrow \tau = P(Y \leq Q_Y(\tau))$

$= P(h(Y) \leq h(Q_Y(\tau)))$

$\therefore h(Q_Y(\tau))$ is the τ^{th} quantile of $h(Y)$, $= Q_{h(Y)}(\tau)$.

Model 1: The additively separable model with quantile independence

Theorem potential outcome: $Y = g(x) + U$.

observed outcome $Y = Y(x)$. (Y, X) observed.

Assume: ① $Q_{U|X}(\tau|x) = Q_U(\tau)$, $Q_U(\tau)$ is known

Then: $g(x)$, $F_{X,U}$ are point identified

proof: $Q_{Y|X}(\tau|x) = Q_{g(x)+U|X}(\tau|x)$

$h(U) = g(x) + U$ $= Q_{g(x)+U|X}(\tau|x)$

$= g(x) + Q_{U|X}(\tau|x)$ (Equivalence).

$= g(x) + Q_U(\tau)$

$$g(x) = \underbrace{Q_{Y|X}(\tau|x)}_{\text{known from the data}} - Q_U(\tau) \quad \text{identified}$$

Corollary: point identification of the classical linear model.

Suppose $Y(x) = p(x)'\beta + U$.

Assume: $Med(U|Y=x) = Med(U)$, $Med(U) = 0$,

then $\theta = (\beta, F_{X,U})$ is point identified.

Model 2: Non-separable Unobservables

if $Y(x) = g(x) + U$, then $\frac{\partial Y(x)}{\partial x} = g'(x)$. not heterogeneous.

To model heterogeneous treatment effect: $Y(x) = g(x, U)$. $\frac{\partial Y(x)}{\partial x} = g'(x, U)$

Theorem: Suppose $Y(x) = g(x, U)$

$Y = Y(X)$, observe (Y, X).

Assume: ① U scalar ② $g(\cdot)$ weakly ↑, left cont. ③ $X \perp\!\!\!\perp U$ ④ $U \sim \text{unif}[0,1]$

Then: $g(x,u)$ is point identified.

pf:
$$\begin{aligned} Q_{Y|X}(\tau|x) &= Q_{g(x,U)|X}(\tau|x) \\ &= Q_{g(x,U)|X}(\tau|x) \quad (\text{known from data}) \\ &= Q_{g(x,U)}(\tau) \\ &= g(x, Q_U(\tau)) \quad \text{since } U \sim \text{unif}[0,1]. \\ &= g(x,\tau) \quad \therefore \text{identified} \end{aligned}$$

Model 3. The linear quantile regression model.

Theorem: Identification of linear quantile regression model.

$$Y(x) = p(x)'\beta(U).$$

$Y = Y(X)$, (Y, X) observed. Assumes:

① U scalar ② $p(x)'\beta(\cdot)$ weakly increasing, left continuous ③ $X \perp\!\!\!\perp U$. ④ $U \sim \text{unif}[0,1]$. ⑤ no multicollinearity.

Then: $\beta: (0,1) \to \mathbb{R}$ is point identified.

pf: let $g(x,u) = p(x)'\beta(u)$, then by previous Theorem, $\forall$.

$$Q_{Y|X}(\tau|x) = p(x)'\beta(\tau).$$

$$\begin{aligned} ASF(x) &= p(x)'\mathbb{E}[\beta(U)] \\ &= p(x)'\int_0^1 \beta(u)\,du \quad \text{since } U \sim \text{unif}[0,1]. \end{aligned}$$

$$\begin{aligned} ATE(x_1 \to x_2) &= ASF(x_2) - ASF(x_1) \\ &= p(x_2)'\mathbb{E}[\beta(U)] - p(x_1)'\mathbb{E}[\beta(U)]. \\ &= (p(x_2) - p(x_1))'\mathbb{E}[\beta(U)]. \end{aligned}$$

$$\begin{aligned} Q_{Y(x)}(\tau) &= Q_{p(x)'\beta(U)}(\tau) \\ &= p(x)'\beta(Q_U(\tau)) \\ &= p(x)'\beta(\tau). \end{aligned}$$

$$\begin{aligned} QTE(x_1 \to x_2) &= Q_{Y(x_2)}(\tau) - Q_{Y(x_1)}(\tau) \\ &= p(x_2)'\beta(\tau) - p(x_1)'\beta(\tau) \\ &= (p(x_2) - p(x_1))'\beta(\tau) \end{aligned}$$

Instrumental Variables.

Simultaneity.

Motivation: to identify supply and demand curve.

Observe in data: (P, Q, Z_1, Z_2).

unknown parameters: $(\alpha_1, \alpha_2, \gamma_1, \gamma_2, \beta_1, \beta_2)$.

When are these parameters point identified?

Model:
$$\begin{cases} P = \alpha_1 + \gamma_1 Q + \beta_1' Z_1 + U_1 \\ Q = \alpha_2 + \gamma_2 P + \beta_2' Z_2 + U_2 \end{cases}$$

Let Z_1 be the supply shifter, Z_2 be the demand shifter.

Then:

Theorem:

Model:
$$\begin{cases} P = \alpha_1 + \gamma_1 Q + \beta_1 Z_1 + \delta_1 Z_2 + U_1 \quad (\text{Supply}) \\ Q = \alpha_2 + \gamma_2 Q + \beta_2 Z_2 + \delta_2 Z_2 + U_2 \end{cases}$$

Assume: ① unique equilibrium: $\gamma_1\gamma_2 \neq 1$

an IV satisfies ②–④:
- ② exclusion: $\delta_1 = 0$, $\delta_2 = 0$
- ③ exogeneity: $\text{cov}(Z_1, U_1) = 0$, $\text{cov}(Z_1, U_2) = 0$; $\text{cov}(Z_2, U_1) = 0$, $\text{cov}(Z_2, U_2) = 0$.
- ④ relevant: $\beta_1 \neq 0$, $\beta_2 \neq 0$.
- ⑤ normalization: $\mathbb{E}[U_1] = 0$, $\mathbb{E}[U_2] = 0$.

Then, (γ_1, γ_2) are point identified.

proof #1: assume $\alpha_1 = 0$, $\alpha_2 = 0$.

$$\begin{cases} P = \gamma_1 Q + \beta_1 Z_1 + \delta_1 Z_2 + U_1 \\ Q = \gamma_2 P + \beta_2 Z_1 + \delta_2 Z_2 + U_2 \end{cases}$$

$$\Leftrightarrow \begin{bmatrix} 1 & -\gamma_1 \\ -\gamma_2 & 1 \end{bmatrix}\begin{bmatrix} P \\ Q \end{bmatrix} = \begin{bmatrix} \beta_1 & 0 \\ 0 & \beta_2 \end{bmatrix}\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} + \begin{bmatrix} U_1 \\ U_2 \end{bmatrix}$$

solve it, we get:
$$P = \frac{\beta_1}{1-\gamma_1\gamma_2}Z_1 + \frac{\gamma_1\beta_2}{1-\gamma_1\gamma_2}Z_2 + \frac{U_1+\gamma_1U_2}{1-\gamma_1\gamma_2}$$
$$Q = \frac{\gamma_2\beta_1}{1-\gamma_1\gamma_2}Z_1 + \frac{\beta_2}{1-\gamma_1\gamma_2}Z_2 + \frac{U_2+\gamma_2U_1}{1-\gamma_1\gamma_2}$$

reduced form: endo… = f(exo…)

let $(\pi_{11}, \pi_{12}) = \left(\frac{\beta_1}{1-\gamma_1\gamma_2}, \frac{\gamma_1\beta_2}{1-\gamma_1\gamma_2}\right)$, $V_1 \equiv \frac{U_1+\gamma_1U_2}{1-\gamma_1\gamma_2}$

$(\pi_{21}, \pi_{22}) = \left(\frac{\gamma_2\beta_1}{1-\gamma_1\gamma_2}, \frac{\beta_2}{1-\gamma_1\gamma_2}\right)$, $V_2 \equiv \frac{U_2+\gamma_2U_1}{1-\gamma_1\gamma_2}$

then
$$\begin{cases} P = \pi_{11}Z_1 + \pi_{12}Z_2 + V_1 \\ Q = \pi_{21}Z_1 + \pi_{22}Z_2 + V_1 \end{cases}$$

$$\begin{aligned} \mathbb{E}[Z_1V_1] &= \mathbb{E}\left[Z_1\frac{U_1+\gamma_2U_2}{1-\gamma_1\gamma_2}\right] \\ &= \frac{\mathbb{E}[Z_1U_1] + \gamma_2\mathbb{E}[Z_1U_2]}{1-\gamma_1\gamma_2} \\ &= 0 \text{ by ③ and ⑤.} \end{aligned}$$

$\mathbb{E}[Z_2V_1] = 0$ also.

Then, π_{11} and π_{12} are point identified.

$\mathbb{E}[Z_1V_2] = \mathbb{E}[Z_2V_2] = 0$ by similar calculation

Then π_{21} and π_{22} are point identified.

Then, $\frac{\pi_{12}}{\pi_{22}} = \gamma_1$, $\frac{\pi_{21}}{\pi_{11}} = \gamma_2$ point identified.

Then, $\beta_1 = \pi_{11}(1-\gamma_1\gamma_2)$, $\beta_2 = \pi_{22}(1-\gamma_1\gamma_2)$ point identified.

proof 2:

$$\begin{aligned} \text{cov}(P, Z_2|Z_1) &= \text{cov}(\alpha_1 + \gamma_1 Q + \beta_1 Z_1 + U_1, Z_2|Z_1) \\ &= \gamma_1\,\text{cov}(Q, Z_2|Z_1) + \underbrace{\beta\,\text{cov}(Z_1, Z_2|Z_1)}_{=0} + \text{cov}(U_1, Z_2|Z_1) \\ &= \gamma_1\,\text{cov}(Q, Z_2|Z_1) + \text{cov}(U_1, Z_2|Z_1) \end{aligned}$$

If strengthen assumption: $\mathbb{E}[U_1|Z_1, Z_2] = 0$, $\mathbb{E}[U_2|Z_1, Z_2] = 0$, then:

$$\begin{aligned} \text{cov}(U_1, Z_2|Z_1) &= \mathbb{E}[U_1Z_2|Z_1] - \mathbb{E}[U_1|Z_1]\mathbb{E}[Z_2|Z_1] \\ &= \mathbb{E}[Z_2\mathbb{E}[U_1|Z_2, Z_1]|Z_1] - \mathbb{E}[\mathbb{E}[U_1|Z_2, Z_1]|Z_1]\mathbb{E}[Z_2|Z_1] \\ &= 0 \text{ by ③ and ⑤} \end{aligned}$$

Then $\gamma_1 = \dfrac{\text{cov}(P, Z_2|Z_1)}{\text{cov}(Q, Z_2|Z_1)}$ point identified.

Next: we show that $\text{cov}(Q, Z_2|Z_1) \neq 0$.

$$\text{cov}(Q, Z_2|Z_1) = \text{cov}(\alpha_2 + \gamma_2 P + \beta_2 Z_1 + U_2, Z_2|Z_1)$$
$$= \gamma_2 \text{cov}(P, Z_2|Z_1) + \beta_2 \text{var}(Z_2|Z_1) + \text{cov}(U_2, Z_2|Z_1)$$
$$= \gamma_2 \text{cov}(P, Z_2|Z_1) + \beta_2 \text{var}(Z_2|Z_1) \text{ by } ③$$

Then: $\text{cov}(P, Z_2|Z_1) = \gamma_1 \text{cov}(Q, Z_2|Z_1)$
$$= \gamma_1\gamma_2 \text{cov}(P, Z_2|Z_1) + \gamma_1\beta_2 \text{var}(Z_2|Z_1)$$

then: $\text{cov}(P, Z_2|Z_1) = \frac{\gamma_1\beta_2 \text{var}(Z_2|Z_1)}{1-\gamma_1\gamma_2}$

then: $\text{cov}(Q, Z_2|Z_1) = \gamma_2 \frac{\gamma_1\beta_2 \text{var}(Z_2|Z_1)}{1-\gamma_1\gamma_2} + \beta_2 \text{var}(Z_2|Z_1)$
$$= \frac{\beta_2}{1-\gamma_1\gamma_2} \text{var}(Z_2|Z_1) \neq 0 \text{ since } \beta_2 \neq 0,\ \text{var}(Z_2|Z) > 0$$

Similarly: $\gamma_2 = \frac{\text{cov}(Q, Z_1|Z_2)}{\text{cov}(P, Z_1|Z_2)}$. point identified.

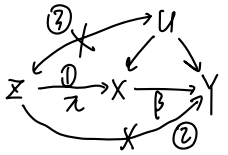
$\beta_2 = (1-\gamma_1\gamma_2)\frac{\text{cov}(Q, Z_2|Z_1)}{\text{var}(Z_2|Z_1)}$ point identified.

similarly, $\beta_1 = (1-\gamma_1\gamma_2)\frac{\text{cov}(P, Z_1|Z_2)}{\text{var}(Z_1|Z_2)}$ point identified.

$\alpha_1 = E[P] - \gamma_1 E[Q] - \beta_1 E[Z_1]$, α_2 similarly.

then: $(\alpha_1, \alpha_2, \beta_1, \beta_2, \gamma_1, \gamma_2)$ all identified

$Y(X, Z) = \beta X + \delta Z + U$. X, Z scalar.
$$\text{cov}(Y, Z) = \text{cov}(\beta X + \delta Z + U, Z)$$
$$= \beta \text{cov}(X, Z) + \delta \text{var}(Z) + \text{cov}(U, Z)$$
$$\therefore \beta = \frac{\text{cov}(Y, Z)}{\text{cov}(X, Z)} - \delta \frac{\text{var}(Z)}{\text{cov}(X, Z)} - \frac{\text{cov}(U, Z)}{\text{cov}(X, Z)}$$



$Z \to Y$: $\beta\pi$
$Z \to X$: π
then: $X \to Y = \frac{\beta\pi}{\pi} = \beta$.

Assumption: ① relevant: $\text{cov}(X, Z) \neq 0$.
② exclusion: $\delta = 0$, Z no direct causal effect on Y.
③ instrument exogeneity: $\text{cov}(U, Z) = 0$.
(like Z randomly assigned)

Then: $\beta = \frac{\text{cov}(Y, Z)}{\text{cov}(X, Z)}$ (2SLS estimand)

$$\beta = \left[\frac{\text{cov}(Y, Z)}{\text{cov}(X, Z)}\right] = \frac{\text{cov}(Y, Z)}{\text{var}(Z)} \Big/ \frac{\text{cov}(X, Z)}{\text{var}(Z)}$$

Second stage (need exclusion) | reduced form: Causal effect of $Z \to Y$. need exogeneity, ITT | first stage: no causal interpretation, falsifiable $\text{cov}(X, Z) = 0$

$Y = \beta X + U$
$= \beta(\underbrace{\pi Z}_{\text{exo.}} + \underbrace{X^{\perp Z}}_{\text{endo.}}) + U$ — first stage, get $X^{\text{pred}} = \pi Z$
$= \underbrace{\beta\pi}_{\text{reduced form}} Z + \beta X^{\perp Z} + U$, $\frac{\text{cov}(Y, Z)}{\text{var}(Z)}$

$Y = \beta X^{\text{pred}} \Rightarrow \beta = \frac{\text{cov}(Y, X^{\text{pred}})}{\text{var}(X^{\text{pred}})} = \frac{\beta\pi}{\pi}$

Generally: $Y(X) = p(X)'\beta + U$
now: $\text{corr}(p_k(X), U) \neq 0$.

Theorem: consider $Y(X, Z) = p(X)'\beta + q(Z)'\delta + U$ (K-vector, L-vector)
Assume: ① exclusion: $\delta = 0$
② exogeneity: $E[q(Z)U] = E[q(Z)]E[U]$
③ relevance: $\text{rank}(E[q(Z)p(X)'] = K$
④ Normalization: $p(X)$ contain a constant and $E[U] = 0$
Then: β is point-identified.
when $L = K$: $\beta = E[q(Z)p(X)']^{-1}E[q(Z)Y]$

proof: we have $0 = E[q(Z)U]$
$= E[q(Z)(Y - p(X)'\beta)]$
$= E[q(Z)Y] - E[q(Z)p(X)']\beta$.

$E[q(Z)p(X)']\beta = E[q(Z)Y]$
$(L \times K)\ (K \times 1) = (L \times 1)$

since $\text{rank}(E[q(Z)p(X)']) = K$, full rank.
And if $L = k$, then:
$\beta = E[q(Z)p(X)']^{-1}E[q(Z)Y]$.
if $L > k$, then: choose a $K \times L$ Π:
$\beta = E[\Pi q(Z)p(X)']^{-1}E[\Pi q(Z)Y]$
(assume invertible. new instrument $\Pi q(Z)$).

Pick optimal Π: same to WLS to minimize variance.
Under homoskedasticity, optimal Π = residual of $\pi Z + X^{\perp Z}$.
If linearity model does not hold, different instrument.

Triangular Models

$$\begin{cases} Y(x) = g(X, U). \\ X(z) = h(z, U). \end{cases}$$ observed outcome: $\begin{cases} Y = g(X, U) \\ X = h(X, U) \end{cases}$

Assume additively separable & linear.
$Y(x) = \gamma_0 + \gamma_1 X + U$
$X(z) = \pi_0 + \pi_1 Z + V$.
Assume Z exogenous: $E[ZV] = E[Z]E[V]$, $\pi_1 \neq 0$,
$E[ZU] = E[Z]E[U]$
then: $(\gamma_0, \gamma_1, \pi_0, \pi_1)$ identified.

Heterogeneous Treatment Effects with IV: Random Coefficients: