Linking Causal Inference Frameworks

Jiding Zhang and GPT-5.5 Pro

The note is mainly based on *Mostly Harmless Econometrics* (MHE, [1]), Judea Pearl's tutorial [2], Oxford APTS notes [4], NBER lecture notes [3], and Imai et al. tutorial [5]. This note is co-created with GPT-5.5 Pro.

Contents

1 Introduction and roadmap **4**

2 Potential outcomes: units, treatments, and estimands **5**

2.1 Estimands 5

2.2 The selection problem 6

2.3 Core assumptions 7

2.4 Identification under unconfoundedness 8

3 DAGs and SCMs: graphical preliminaries **8**

3.1 Graphs, arrows, and causal interpretation 9

3.2 Three elementary path structures 9

3.3 Blocking, open paths, and d-separation 10

3.4 Causal paths, back-door paths, and confounding 12

3.5 Conditioning versus intervening 12

3.6 The back-door criterion and adjustment 13

3.7 The front-door criterion 14

3.8 Structural causal models 15

4 Econometric regression preliminaries **15**

4.1 The conditional expectation function . . . 16

4.2 Population linear projection and its relation to the CEF . . . 17

4.3 Short and long regressions . . . 17

4.4 The omitted-variable-bias formula . . . 18

4.5 Regression exogeneity and endogeneity . . . 19

5 Translation dictionary across frameworks **19**

5.1 ATE, ATT, and intervention distributions . . . 20

5.2 A compact dictionary . . . 20

5.3 What each language contributes to identification . . . 21

6 Regression adjustment, endogeneity, and causal interpretation **21**

6.1 The selection problem in three languages . . . 22

6.2 Adjustment: the same identification result in two notations . . . 23

6.3 What short and long regressions are doing causally . . . 24

6.4 Good controls, bad controls, and the timing of variables . . . 25

6.5 Endogeneity taxonomy across frameworks . . . 26

6.6 Regression, matching, weighting, and propensity scores . . . 27

7 Design-based identification: IV, DiD, and RD **28**

7.1 Instrumental variables and LATE . . . 28

7.2 Difference-in-differences and fixed effects . . . 29

7.3 Regression discontinuity . . . 30

8 Mechanisms, mediation, and moderation **32**

8.1 From total effects to mechanisms . . . 32

8.2 Why randomizing treatment alone is not enough . . . 33

8.3 Sequential ignorability . . . 33

8.4 Experimental designs for mechanisms 33
8.5 Moderation and treatment-effect heterogeneity 34
8.6 Connection to front-door identification 35

9 Final synthesis: comparison, workflow, and pitfalls 36

9.1 What each framework contributes 36
9.2 A recommended empirical workflow 36
9.3 Common pitfalls 37

1 Introduction and roadmap

Most empirical questions in economics, psychology, education, public policy, and the biomedical sciences are not merely predictive. They ask what would happen under a different treatment, assignment rule, policy, stimulus, incentive, or institutional arrangement. Pearl emphasizes that such questions cannot be answered from the observed joint distribution alone; they require assumptions about how the data-generating process would change under interventions [2, Secs. 1–2]. Angrist and Pischke similarly organize empirical work around an ideal experiment, the sources of selection bias that prevent naive comparisons from being causal, and designs that try to recover experiment-like variation [1, Chs. 1–2].

This note links three languages that are often taught separately.

1. **Potential outcomes.** This language defines unit-level counterfactuals and population estimands such as ATE, ATT, CATE, and LATE, and also supports direct and indirect effects in mediation settings [3, 1, 5].

2. **DAGs and structural causal models.** This language encodes causal structure, interventions, confounding, conditioning, mediation, selection, and identification through graphical criteria such as back-door and front-door adjustment [2, 4].

3. **Econometric regression and design.** This language connects causal identification to conditional expectation functions, short and long regressions, omitted-variable bias, endogeneity, IV, fixed effects, difference-in-differences, regression discontinuity, and inference [1, 3].

A useful organizing sequence is

$$\text{Define} \longrightarrow \text{Assume} \longrightarrow \text{Identify} \longrightarrow \text{Estimate}.$$

Pearl uses this four-step workflow explicitly: define the causal target, state assumptions, identify the target as a functional of observed data, and estimate that functional [2, Sec. 4]. Econometric research designs follow the same logic, though the words are different: specify the ideal experiment, describe why the observed variation mimics it, derive the estimating equation, and assess threats to identification.

The structure of the note follows a separation-then-synthesis logic. Sections 2–4 introduce the three languages separately. Section 5 translates between them. Section 6 uses the translation to

explain regression adjustment and endogeneity. Section 7 treats IV, DiD, and RD as design-based identification strategies. Section 8 turns from total effects to mechanisms, mediation, and moderation. Section 9 gives a final workflow and checklist.

Takeaway

Potential outcomes are best for writing the estimand. DAGs/SCMs are best for displaying and checking the causal assumptions. Econometric regressions and quasi-experimental designs are best for implementing credible estimators once the estimand and assumptions are clear.

2 Potential outcomes: units, treatments, and estimands

Let $A \in \{0, 1\}$ denote a binary treatment and Y an outcome. For unit i, define

$$Y_i(1) = \text{unit } i\text{'s outcome under treatment}, \qquad Y_i(0) = \text{unit } i\text{'s outcome under control}.$$

The NBER lecture notes use the same setup with treatment indicator W_i and potential outcomes $Y_i(0), Y_i(1)$, emphasizing that causal effects compare potential outcomes defined on the same unit [3, Lecture 1, Sec. 2]. Only one potential outcome is observed:

$$Y_i = Y_i(A_i) = A_i Y_i(1) + (1 - A_i) Y_i(0).$$

This is a consistency relation: the observed outcome equals the potential outcome under the treatment actually received.

2.1 Estimands

The *population average treatment effect* is

$$\text{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)].$$

The NBER notes call this the population ATE, or PATE, and also define the population average treatment effect for the treated, or PATT [3, Lecture 1, Sec. 2.2]. In the notation of this note,

$$\text{ATT} = \mathbb{E}[Y_i(1) - Y_i(0) \mid A_i = 1].$$

Likewise,

$$\text{ATU} = \mathbb{E}[Y_i(1) - Y_i(0) \mid A_i = 0]$$

is the average effect for currently untreated units. The conditional average treatment effect is

$$\text{CATE}(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x].$$

For a finite sample of N units, one can define sample analogues:

$$\tau_S = \frac{1}{N}\sum_{i=1}^{N}\{Y_i(1) - Y_i(0)\}, \qquad \tau_{S,T} = \frac{1}{N_T}\sum_{i:A_i=1}\{Y_i(1) - Y_i(0)\},$$

where $N_T = \sum_i A_i$. These sample estimands differ conceptually from population estimands: the target is the causal effect for the sample at hand rather than for a superpopulation [3, Lecture 1, Sec. 2.2].

Other estimands will appear later. With an instrument Z for treatment D, the local average treatment effect is

$$\text{LATE} = \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i(1) > D_i(0)],$$

where $D_i(z)$ is treatment receipt under instrument value z. In a two-period DiD design, the target is often a post-treatment ATT. In an RD design, the target is a local effect at the cutoff. In mediation, the target may be a natural direct or indirect effect.

2.2 The selection problem

The fundamental problem is that $Y_i(1) - Y_i(0)$ is not observed for any unit. MHE illustrates the problem through the decomposition

$$\begin{aligned}&\mathbb{E}[Y_i \mid A_i = 1] - \mathbb{E}[Y_i \mid A_i = 0] \\ &= \underbrace{\mathbb{E}[Y_i(1) - Y_i(0) \mid A_i = 1]}_{\text{ATT}} + \underbrace{\Big\{\mathbb{E}[Y_i(0) \mid A_i = 1] - \mathbb{E}[Y_i(0) \mid A_i = 0]\Big\}}_{\text{selection bias}}.\end{aligned}$$

The first term is the causal effect for the treated. The second term compares untreated potential outcomes for treated and untreated units. It is not observed directly. Random assignment solves the selection problem because it makes treatment independent of potential outcomes, removing the selection-bias term [1, Ch. 2].

This decomposition is important because it separates two questions that are often conflated. First, what causal effect do we want? Second, why should treated and untreated units tell us about each other's counterfactual outcomes?

2.3 Core assumptions

The following assumptions are standard in point-treatment analyses. The formulation follows APTS, which separates no unobserved confounding, positivity, consistency, and no interference [4, Ch. 9].

Assumption 2.1 (Consistency and well-defined intervention)**.** If unit i factually receives $A_i = a$, then

$$Y_i = Y_i(a).$$

This requires that the intervention represented by a be well-defined. APTS emphasizes that vague treatments can violate consistency: an intervention on weight could mean a drug, exercise, surgery, or diet, each with different consequences even if the numerical weight change is the same [4, Ch. 9.3].

Assumption 2.2 (No interference)**.** Unit i's potential outcome depends only on unit i's treatment. Thus $Y_i(a_i)$ is sufficient notation rather than $Y_i(a_1, \ldots, a_n)$. This assumption can fail for vaccines, peer effects, classroom interventions, market-wide policies, and social-network treatments [4, Ch. 9.4].

Assumption 2.3 (Conditional exchangeability / unconfoundedness)**.** For binary treatment, the standard joint version is

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp A_i \mid X_i.$$

This implies the marginal statements $Y_i(a) \perp\!\!\!\perp A_i \mid X_i$ for $a \in \{0, 1\}$, which are sufficient for identifying the marginal mean $\mathbb{E}[Y(a)]$. The NBER notes define unconfoundedness in the joint form and note that it is also called the conditional independence assumption, selection on observables, or missing at random in related literatures [3, Lecture 1, Sec. 2.2]. APTS writes the corresponding no-unobserved-confounding condition as $Y(a) \perp\!\!\!\perp A \mid X$ [4, Ch. 9.1].

Assumption 2.4 (Positivity / overlap)**.** For ATE identification,

$$0 < \mathbb{P}(A_i = 1 \mid X_i = x) < 1$$

for covariate values x in the target population. For ATT identification, we only require untreated comparison units in the treated support:

$$\mathbb{P}(A_i = 0 \mid X_i = x) > 0 \quad \text{whenever} \quad \mathbb{P}(X_i = x \mid A_i = 1) > 0.$$

The NBER notes emphasize that overlap should be checked before looking at outcomes and that lack of overlap can make estimates highly sensitive to estimator choice [3, Lecture 1, Secs. 1 and 6]. APTS distinguishes weak positivity from stronger bounds such as $\varepsilon < \pi(X) < 1 - \varepsilon$ for variance control [4, Ch. 9.2].

2.4 Identification under unconfoundedness

Under consistency, no interference, conditional exchangeability, and positivity,

$$\mathbb{E}[Y(a)] = \int \mathbb{E}[Y \mid A = a, X = x] dF_X(x).$$

For a discrete X, replace the integral with a sum. Hence

$$\text{ATE} = \int \{\mathbb{E}[Y \mid A = 1, X = x] - \mathbb{E}[Y \mid A = 0, X = x]\} \, dF_X(x).$$

For ATT, it is enough to identify the missing untreated mean for treated units:

$$\text{ATT} = \mathbb{E}[Y \mid A = 1] - \int \mathbb{E}[Y \mid A = 0, X = x] dF_{X|A=1}(x).$$

This ATT formula weights covariate strata by the treated distribution, not by the population distribution.

The propensity score is

$$e(x) = \mathbb{P}(A = 1 \mid X = x).$$

The propensity-score balancing result stated in the NBER notes implies that, under unconfoundedness given X, adjustment using the propensity score can also suffice [3, Lecture 1, Sec. 2.2]. The NBER notes define $e(x)$ in exactly this way and discuss regression, matching, weighting, and propensity-score estimators as alternative implementations of the same identification idea under unconfoundedness and overlap [3, Lecture 1, Secs. 2–3].

> **Takeaway**
>
> Potential outcomes make the causal target and missing-data problem explicit. Identification then requires assumptions that justify replacing missing counterfactual means with observed conditional means.

3 DAGs and SCMs: graphical preliminaries

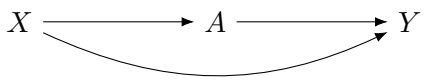
This section introduces the graphical language on its own terms. The purpose is not yet to diagnose omitted-variable bias in a regression. That synthesis is delayed until Section 6. Here we define the graph, the blocking rules, and the intervention notation that will later be translated into potential-outcome and econometric language. The definitions follow Pearl's causal-graph notation and the APTS graphical-model notes [2, 4].

3.1 Graphs, arrows, and causal interpretation

Definition 3.1 (Directed graph and DAG; [4])**.** A directed graph $\mathcal{G} = (V, E)$ consists of a finite set of vertices V and a collection of ordered pairs E called directed edges. If $(v, w) \in E$, write $v \to w$. A directed graph is acyclic if it contains no directed cycle. A directed acyclic graph is called a DAG.

If $v \to w$, then v is a parent of w and w is a child of v. Let $\mathrm{pa}_{\mathcal{G}}(w)$ denote the parents of w and $\mathrm{ch}_{\mathcal{G}}(v)$ the children of v. A node a is an ancestor of v if $a = v$ or there is a directed path $a \to \cdots \to v$. Descendants are defined analogously. A path between two nodes is a sequence of distinct adjacent nodes connecting them, ignoring arrow direction. A path is directed from X to Y if all arrows point away from X and toward Y [4, Ch. 5].

A directed edge $A \to Y$ is usually read as a direct causal claim relative to the variables represented in the graph: if we intervene on A, the distribution of Y may change through that edge. The word "direct" is relative to the graph. If an omitted mediator is later added, the edge $A \to Y$ may be replaced by $A \to M \to Y$. Therefore, the absence of an arrow is often the stronger assumption: it says that one variable is not a direct cause of another, given the other variables in the graph. Pearl emphasizes that causal assumptions in graphs are encoded by the arrows that are absent as well as by the arrows that are present [2, Sec. 3].



The small graph above says that X is a direct cause of both treatment A and outcome Y, and that A is a direct cause of Y. It does not say that X is the only possible common cause unless the graph is intended to be causally complete for the question at hand.

3.2 Three elementary path structures

The blocking rules can be understood from three path fragments. These fragments should be learned before thinking about regression controls.

Name	Graph fragment	Conditioning intuition
Chain	$A \to C \to Y$	Association can flow through C. Conditioning on the noncollider C blocks the path. If C is a mediator, blocking this path also blocks part of the causal effect.
Fork	$A \leftarrow C \to Y$	C is a common cause. Conditioning on C blocks the path between A and Y through the common cause.
Collider	$A \to C \leftarrow Y$	The path is blocked unless we condition on C or on a descendant of C. Conditioning on a collider can create association.

Pearl's tutorial and the APTS graphical notes use the same blocking logic: conditioning blocks paths through noncolliders and opens paths through colliders or their descendants [2, 4]. In a DAG, "conditioning" can mean stratifying, matching, restricting the sample, including a variable in a regression, weighting to balance it, or otherwise working with a conditional distribution. The exact statistical implementation comes later.

Example 3.1 (Conditioning on a collider)**.** Suppose ability U_1 and family resources U_2 each make admission to an elite school S more likely:

$$U_1 \to S \leftarrow U_2.$$

In the full population, U_1 and U_2 may be independent. Among admitted students, however, a lower value of one factor may have to be compensated by a higher value of the other. Conditioning on $S = 1$ can therefore create a negative association between U_1 and U_2. This phenomenon is graphical; it is not specific to any regression model.

3.3 Blocking, open paths, and d-separation

Definition 3.2 (Collider and noncollider)**.** On a path, a non-endpoint node C is a collider if the two edges adjacent to it on the path both point into C, as in $A \to C \leftarrow B$. Otherwise C is a noncollider on that path.

A node can be a collider on one path and a noncollider on another. Collider status is path-specific, not an intrinsic property of the variable.

Definition 3.3 (Blocked path and d-separation; [2, 4])**.** Let S be a set of nodes on which we condition. A path is blocked by S if at least one of the following holds:

1. the path contains a noncollider that is in S;
2. the path contains a collider C such that neither C nor any descendant of C is in S.

If every path between node sets A and B is blocked by S, then A and B are d-separated by S, written $A \perp_{\mathcal{G}} B \mid S$.

D-separation is a statement about a graph. Under the global Markov property, it implies conditional independence in distributions compatible with the graph [4, Ch. 5]. Conversely, if a path is open, the graph generally permits association along that path, although special parameter values can cancel associations in particular distributions.

Definition 3.4 (DAG Markov factorization; [2, 4])**.** A distribution P is Markov with respect to a DAG $\mathcal{G}$ on variables $V_1, \ldots, V_n$ if it factorizes as

$$P(v_1, \ldots, v_n) = \prod_{j=1}^{n} P(v_j \mid \mathrm{pa}_j),$$

where pa_j denotes the realized values of the parents of V_j in $\mathcal{G}$.

This factorization is a bridge between graphs and probability. It builds the joint distribution from local conditional distributions. In causal work, this factorization is interpreted structurally only after the graph and its mechanisms are given a causal interpretation.

Caveat

D-separation is only as credible as the graph. If an important common cause is omitted, the graphical conclusion can be wrong. When important common causes are unobserved, they should be represented explicitly, often as latent nodes or bidirected arcs in more advanced graphical notation.

3.4 Causal paths, back-door paths, and confounding

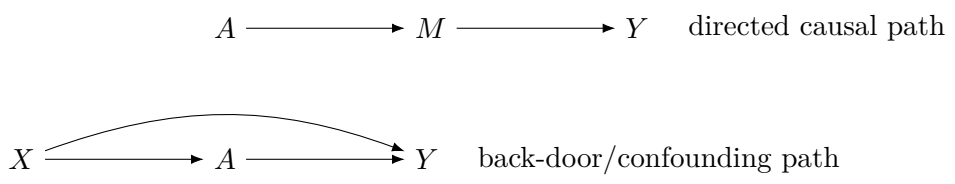
Let A be treatment and Y outcome. A path from A to Y that begins with an arrow out of A is in the causal direction. A directed path such as

$$A \to M \to Y$$

represents a possible mechanism by which treatment affects the outcome. A path from A to Y that begins with an arrow into A is a back-door path. The canonical back-door path is

$$A \leftarrow X \to Y.$$

It carries noncausal association because it connects treatment and outcome through a common cause.



If all back-door paths from A to Y are blocked, then the effect of A on Y is unconfounded relative to the graph. If at least one back-door path is open, a simple comparison of treated and untreated units generally mixes causal association with confounding association.

3.5 Conditioning versus intervening

Pearl's do-operator distinguishes observation from intervention. The observational quantity

$$P(Y = y \mid A = a)$$

asks about units whose treatment happened to equal a. The interventional quantity

$$P(Y = y \mid \mathrm{do}(A = a))$$

asks about a regime that externally sets A to a.

Conditioning on $A = a$ does not alter the mechanism that generated A. Intervening with $\mathrm{do}(A = a)$ replaces the structural assignment rule for A by the constant a. Graphically, incoming arrows into A are deleted. Probabilistically, in a Markovian DAG factorization, the factor for A is removed and the remaining factors are evaluated at $A = a$ [2, 4, Sec. 3.2.1].

Theorem 3.1 (Truncated factorization / intervention distribution; [2, 4])**.** *If* $P(v) = \prod_j P(v_j \mid \mathrm{pa}_j)$ *factorizes according to a Markovian DAG, then the intervention distribution under* $\mathrm{do}(A = a)$ *is*

$$P(v_{-A} \mid \mathrm{do}(A = a)) = \prod_{j:V_j \neq A} P(v_j \mid \mathrm{pa}_j)\big|_{A=a}.$$

The factor that describes how A *is generated from its parents is removed.*

Example 3.2 (Observation is not intervention)**.** In the graph $X \to A \to Y$ and $X \to Y$, $P(Y \mid A = 1)$ compares people who selected or were selected into treatment. It generally depends on the distribution of X among treated people. By contrast, $P(Y \mid \mathrm{do}(A = 1))$ describes the distribution of outcomes if treatment were externally set to one, leaving the distribution of pre-treatment X unchanged.

3.6 The back-door criterion and adjustment

Definition 3.5 (Back-door adjustment set; [2, 4])**.** A set C is a back-door adjustment set for the ordered pair (A, Y) if:

1. C blocks all back-door paths from A to Y;
2. no element of C is a descendant of A.

Pearl calls such a set admissible or sufficient for adjustment [2, Definition 3]. The APTS notes use the same two graphical conditions in their discussion of back-door adjustment sets [4, Ch. 13.2].

Theorem 3.2 (Back-door adjustment; [2, 4])**.** *If* C *is a back-door adjustment set for* (A, Y)*, then*

$$P(Y = y \mid \mathrm{do}(A = a)) = \sum_c P(Y = y \mid A = a, C = c) P(C = c),$$

with integrals replacing sums for continuous C*.*

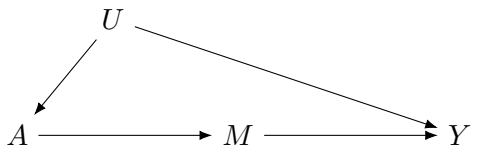
This theorem is the graphical counterpart of conditional exchangeability in potential-outcome notation. Pearl emphasizes that the back-door criterion gives a graphical method for selecting admissible adjustment sets, rather than asking the analyst to assess the counterfactual independence $Y(a) \perp\!\!\!\perp A \mid C$ directly [2, Sec. 3.3.1].

3.7 The front-door criterion

Back-door adjustment is not the only graphical identification strategy. Sometimes no observed set blocks all back-door paths from A to Y, yet the effect is still identifiable through a mediator-like variable.

Definition 3.6 (Front-door adjustment set; [2])**.** A set M satisfies the front-door criterion for the effect of A on Y if:

1. M intercepts all directed paths from A to Y;
2. there is no unblocked back-door path from A to M;
3. all back-door paths from M to Y are blocked by A.



Here U confounds A and Y, so ordinary back-door adjustment fails if U is unobserved. If M satisfies the three front-door conditions, the total effect is identified by

$$P(Y = y \mid \mathrm{do}(A = a)) = \sum_{m} P(M = m \mid A = a) \sum_{a'} P(Y = y \mid M = m, A = a')P(A = a').$$

Pearl's survey explains the logic as a multi-stage adjustment: estimate the effect of A on the mediator, estimate the effect of the mediator on Y after adjusting for A, and combine the pieces into an expression with no do-operator [2, Sec. 3.3.3].

Takeaway

Back-door adjustment removes confounding by blocking noncausal paths. Front-door adjustment can identify an effect even with unobserved treatment-outcome confounding, provided an observed mediator carries the treatment effect and the mediator-outcome relation can itself be deconfounded.

3.8 Structural causal models

A structural causal model specifies how each variable is generated from its parents and exogenous disturbances. The APTS notes define an SCM with respect to a directed graph $\mathcal{G}$ as a collection of functions

$$V_j = f_j(V_{\mathrm{pa}(j)}, U_j), \qquad j = 1, \ldots, n,$$

where U_j are exogenous variables [4, Ch. 7.1]. Pearl's SCM similarly combines structural equations, graphs, interventions, and counterfactuals, and presents it as a unifying framework for graphical, potential-outcome, structural-equation, and interventional approaches [2, Sec. 3].

The SCM semantics of an intervention $\mathrm{do}(A = a)$ is modular: replace the structural equation for A by the constant a and leave the other structural equations intact. Counterfactuals are obtained by applying such modifications while holding the exogenous disturbances for the unit fixed. This is why SCMs can define both intervention distributions, such as $P(Y \mid \mathrm{do}(A = a))$, and unit-level counterfactuals, such as $Y_a(u)$.

> **Caveat**
>
> Ordinary DAGs encode many conditional independences and graphical adjustment rules, but they do not by themselves encode support conditions such as overlap, well-defined treatment versions, interference restrictions, monotonicity in IV, or all cross-world restrictions used in mediation and principal stratification.

4 Econometric regression preliminaries

This section introduces regression as an econometric and statistical object before giving it a causal interpretation. This separation is essential. MHE's Chapter 3 first develops the conditional expectation function, population regression, short and long regressions, and omitted-variable-bias algebra, and only then discusses the causal assumptions under which regression can estimate causal effects [1, Ch. 3]. The causal interpretation is postponed to Section 6.

4.1 The conditional expectation function

Let Y_i be an outcome and X_i a vector of observed variables. The conditional expectation function (CEF) is

$$m(x) = \mathbb{E}[Y_i \mid X_i = x].$$

MHE defines the CEF as the population average of Y_i with X_i held fixed and emphasizes that the CEF summarizes predictive or associational structure, not necessarily causality [1, Sec. 3.1.1]. If Y_i is continuous with conditional density $f_Y(\cdot \mid X_i = x)$, then

$$\mathbb{E}[Y_i \mid X_i = x] = \int t f_Y(t \mid X_i = x)\, dt,$$

with sums replacing integrals for discrete outcomes.

For a binary variable A_i, the CEF $\mathbb{E}[Y_i \mid A_i]$ has two values:

$$\mathbb{E}[Y_i \mid A_i = 1], \qquad \mathbb{E}[Y_i \mid A_i = 0].$$

Their difference is an observed difference in conditional means. It is causal only under assumptions that connect observed outcomes to potential outcomes. This distinction is central throughout the note.

Proposition 4.1 (CEF decomposition)**.** *Let $m(X_i) = \mathbb{E}[Y_i \mid X_i]$ and define the CEF residual $\eta_i = Y_i - m(X_i)$. Then*

$$\mathbb{E}[\eta_i \mid X_i] = 0, \qquad \mathbb{E}[g(X_i)\eta_i] = 0$$

for any integrable function g.

Proof. The first equality follows directly from the definition of conditional expectation:

$$\mathbb{E}[\eta_i \mid X_i] = \mathbb{E}[Y_i - m(X_i) \mid X_i] = m(X_i) - m(X_i) = 0.$$

The second follows from iterated expectations:

$$\mathbb{E}[g(X_i)\eta_i] = \mathbb{E}\{g(X_i)\mathbb{E}[\eta_i \mid X_i]\} = 0.$$

□

This result explains why the CEF is the best predictor of Y in mean-squared-error terms among all functions of X. It also explains the econometric appeal of conditioning: once we condition on X, the remaining CEF residual has no predictable component from X.

4.2 Population linear projection and its relation to the CEF

The population linear projection of Y_i on a vector of regressors R_i is

$$Y_i = R_i'\beta + u_i, \qquad \beta = \arg\min_b \mathbb{E}[(Y_i - R_i'b)^2].$$

If $\mathbb{E}[R_iR_i']$ is nonsingular, then

$$\beta = \mathbb{E}[R_iR_i']^{-1}\mathbb{E}[R_iY_i], \qquad \mathbb{E}[R_iu_i] = 0.$$

The orthogonality condition $\mathbb{E}[R_iu_i] = 0$ is a property of the best linear predictor. It should not be confused with a causal assumption. A causal structural error is a different object.

If the CEF is exactly linear,

$$\mathbb{E}[Y_i \mid R_i] = R_i'\beta,$$

then the linear projection equals the CEF. If the CEF is nonlinear, the population regression coefficient is the best linear approximation to the CEF, with weights determined by the distribution of the regressors. MHE stresses this link between linear regression and the CEF [1, Sec. 3.1.2].

Example 4.1 (Saturated regression)**.** If A is binary and X takes finitely many values, a fully saturated regression with all indicators for X and all treatment-by-X interactions exactly reproduces the cell means $\mathbb{E}[Y \mid A = a, X = x]$. In that special case, regression is just a convenient way to organize conditional means. A non-saturated linear regression imposes additional functional-form restrictions or weighting choices.

4.3 Short and long regressions

Let A_i denote a scalar regressor of interest, such as treatment, schooling, or exposure, and let X_i denote additional controls. The short regression is

$$Y_i = \alpha^S + \beta^S A_i + u_i^S,$$

while the long regression is

$$Y_i = \alpha^L + \beta^L A_i + X_i'\gamma^L + u_i^L.$$

Both equations are population linear projections unless explicitly stated otherwise. The short coefficient compares outcome variation with raw variation in A_i. The long coefficient compares residualized outcome variation with residualized treatment variation.

By the Frisch-Waugh-Lovell theorem,

$$\beta^L = \frac{\mathrm{Cov}(\widetilde{A}_i, \widetilde{Y}_i)}{\mathrm{Var}(\widetilde{A}_i)},$$

where $\widetilde{A}_i$ is the residual from the population projection of A_i on $(1, X_i)$ and $\widetilde{Y}_i$ is the residual from the population projection of Y_i on $(1, X_i)$. Thus, adding controls means asking whether the part of treatment not linearly predicted by X_i is associated with the part of the outcome not linearly predicted by X_i. This statement is algebraic. Whether the remaining variation is as-if random is a causal question.

4.4 The omitted-variable-bias formula

Suppose the long regression is

$$Y_i = \alpha^L + \beta^L A_i + X_i' \gamma^L + u_i^L,$$

and the auxiliary projection of X_i on A_i is

$$X_i = \delta + \pi A_i + r_i.$$

Then the short-regression coefficient satisfies

$$\beta^S = \beta^L + \pi' \gamma^L.$$

For a scalar omitted variable X_i,

$$\beta^S - \beta^L = \gamma^L \frac{\mathrm{Cov}(A_i, X_i)}{\mathrm{Var}(A_i)}.$$

MHE summarizes this formula as: short equals long plus the effect of omitted variables times the regression of omitted variables on included variables [1, Sec. 3.2.2].

The sign logic is useful. If X_i is positively correlated with A_i and positively associated with Y_i conditional on A_i, omitting X_i makes β^S larger than β^L in the population linear-projection sense. If the two associations have opposite signs, the omitted-variable term is negative. But this is only regression algebra. It becomes a causal omitted-variable-bias formula only when the long regression has a causal interpretation and the omitted variable is a genuine confounder.

4.5 Regression exogeneity and endogeneity

Consider the linear equation

$$Y_i = \alpha + \tau A_i + X_i'\gamma + \varepsilon_i.$$

In a purely predictive regression, ε_i can be a projection residual and is automatically orthogonal to included regressors. In a causal or structural equation, ε_i represents omitted determinants of the outcome in the causal model. Then the coefficient τ is causally interpretable only if treatment variation is unrelated to the structural error after conditioning on X_i.

A strong form of exogeneity is

$$\mathbb{E}[\varepsilon_i \mid A_i, X_i] = 0.$$

Endogeneity means that the relevant orthogonality or conditional-mean restriction fails. Classical sources include omitted variables/confounding, bad controls, sample selection, simultaneity or reverse causality, measurement error, and failures of IV, DiD, or RD identifying assumptions. Section 6 translates these failures into potential-outcome and DAG language.

Takeaway

CEF and regression are associational tools. They become causal only after we add assumptions that link observed conditional means or residualized associations to counterfactual comparisons.

5 Translation dictionary across frameworks

The potential-outcome and SCM languages are mathematically compatible when the counterfactuals are generated by an SCM. Pearl writes the potential-outcome variable as $Y_x(u)$, the value of Y for unit u under the intervention $X = x$ [2, Sec. 5]. Thus

$$P(Y(d) = y) = P(Y = y \mid \mathrm{do}(D = d)).$$

The left side names a counterfactual random variable; the right side names an interventional distribution. They should not be confused.

The dictionary below uses A for a generic treatment. In the IV rows, Z is the instrument and D is the treatment received. Thus $D_i(z)$ denotes treatment receipt under instrument value z, while

$Y_i(d, z)$ denotes the outcome under joint intervention $(D = d, Z = z)$. The exclusion restriction is what permits the shorter notation $Y_i(d)$.

5.1 ATE, ATT, and intervention distributions

For a binary treatment A,

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y \mid \text{do}(A = 1)] - \mathbb{E}[Y \mid \text{do}(A = 0)].$$

The ATT is

$$\text{ATT} = \mathbb{E}[Y(1) - Y(0) \mid A = 1].$$

This is a counterfactual query conditional on the factual event $A = 1$. It is not merely the difference between two marginal intervention distributions.

5.2 A compact dictionary

Concept	**Potential-outcome language**	**DAG/SCM/do language**	**Econometric language**
Unit counterfactual	$Y_i(a)$	$Y_a(i)$: value in the submodel where A is set to a	Outcome unit i would have under regime a
Interventional distribution	$P(Y(a) = y)$	$P(Y_a = y) = P(Y = y \mid \text{do}(A = a))$	Outcome distribution under a policy setting $A = a$
Observed outcome	If $A_i = a$, then $Y_i = Y_i(a)$	Consistency follows from intervention semantics	Observed outcome under received treatment
ATE	$\mathbb{E}[Y(1) - Y(0)]$	$\mathbb{E}[Y \mid \text{do}(A = 1)] - \mathbb{E}[Y \mid \text{do}(A = 0)]$	Population effect of assigning treatment
ATT	$\mathbb{E}[Y(1) - Y(0) \mid A = 1]$	$\mathbb{E}[Y_1 - Y_0 \mid A = 1]$, a counterfactual query conditional on a factual event	Effect on participants or treated units
No unobserved confounding	$\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid X$	X blocks all back-door paths from A to Y and contains no descendants of A	Selection on observables; CIA
Back-door adjustment	$\mathbb{E}[Y(a)] = \int \mathbb{E}[Y \mid A = a, X = x] dF_X(x)$	$P(Y \mid \text{do}(A = a)) = \sum_x P(Y \mid A = a, X = x) P(X = x)$	Regression, matching, weighting, or standardization
Overlap	$0 < P(A = a \mid X = x) < 1$ on relevant support	Support condition, not encoded by DAG alone	Comparable treated and controls exist
Front-door	Target remains $P(Y(a))$; primitive PO notation is less compact	Mediator M intercepts directed paths, $A \to M$ unconfounded, and $M \to Y$ back-door paths blocked by A	Identification through mediator when back-door adjustment fails

Continued on next page

Concept	Potential-outcome language	DAG/SCM/do language	Econometric language
Instrument independence	$Z \perp\!\!\!\perp \{D(0), D(1), Y(d,z) : d, z \in \{0,1\}\}$, perhaps conditional on X	No open back-door path from Z to D or Y, given design/covariates	Instrument as-if randomly assigned
Instrument exclusion	$Y(d,1) = Y(d,0) \equiv Y(d)$	No directed causal path from Z to Y bypassing D	Instrument affects outcome only through treatment
Instrument relevance	$\mathbb{E}[D(1) - D(0)] \neq 0$	Directed causal effect $Z \to D$ is nonzero	Nonzero first stage
LATE monotonicity	After orienting Z so that it weakly increases treatment, $D_i(1) \geq D_i(0)$ for all i	Not encoded by an ordinary DAG	No defiers in the chosen orientation
LATE	$\mathbb{E}[Y(1) - Y(0) \mid D(1) > D(0)]$	Complier effect; compliance type is counterfactual	Wald/just-identified IV estimand under LATE assumptions
Bad control	Conditioning can change the estimand or destroy exchangeability	Mediators block causal paths; colliders open noncausal paths	Longer regression can be worse

5.3 What each language contributes to identification

Potential outcomes are strongest at defining the target population and counterfactual contrasts. DAGs are strongest at organizing the assumptions that justify exchangeability or alternative identification strategies. Regression and econometric design are strongest at implementing the observed-data functional and making the identifying variation credible.

Pearl's Section 5 explicitly discusses the formal mapping between SCMs and potential outcomes and argues for a symbiosis between counterfactual and graphical methods [2, Sec. 5]. The NBER notes make a complementary econometric point: under unconfoundedness, several estimators can be used, and the empirical credibility of unconfoundedness and overlap often matters more than the particular estimator chosen [3, Lecture 1, Sec. 1].

6 Regression adjustment, endogeneity, and causal interpretation

Sections 2, 3, and 4 introduced three languages separately. We now synthesize them. The goal is to explain how regression controls, omitted-variable bias, bad controls, and endogeneity map into potential outcomes and DAGs. This section intentionally contains the main cross-framework discussion so that the preliminary sections do not have to repeat it.

6.1 The selection problem in three languages

Start with the observed difference in means:

$$\mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0].$$

Using consistency, $Y = Y(1)$ for treated units and $Y = Y(0)$ for untreated units. Add and subtract $\mathbb{E}[Y(0) \mid A = 1]$:

$$\begin{aligned}\mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0] &= \mathbb{E}[Y(1) \mid A = 1] - \mathbb{E}[Y(0) \mid A = 0] \\ &= \underbrace{\mathbb{E}[Y(1) - Y(0) \mid A = 1]}_{\text{ATT}} + \underbrace{\{\mathbb{E}[Y(0) \mid A = 1] - \mathbb{E}[Y(0) \mid A = 0]\}}_{\text{selection bias}}.\end{aligned}$$

This decomposition is the potential-outcome version of the selection problem emphasized in MHE's discussion of the experimental ideal [1, Ch. 2]. The second term compares the untreated potential outcomes of treated and untreated groups. It is zero in a randomized experiment, but not generally in observational data.

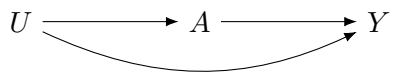
The DAG version is an open back-door path:

$$A \leftarrow U \rightarrow Y.$$

The unobserved common cause U affects both treatment and outcome, so treated and untreated units differ in potential outcomes even before treatment. The regression version is failure of exogeneity in a structural equation

$$Y_i = \alpha + \tau A_i + \varepsilon_i : \qquad \mathbb{E}[\varepsilon_i \mid A_i] \neq 0 \quad \text{or} \quad \operatorname{Cov}(A_i, \varepsilon_i) \neq 0.$$

The short regression coefficient then mixes causal effect and selection.



The three statements are the same diagnosis in different notation:

$$\mathbb{E}[Y(0) \mid A = 1] \neq \mathbb{E}[Y(0) \mid A = 0]$$

$\iff$ open back-door path $\iff$ endogenous treatment in a structural regression.

The equivalence is conceptual rather than automatic: the precise regression statement depends on the structural equation and functional form.

6.2 Adjustment: the same identification result in two notations

Suppose X is a set of pre-treatment covariates. In potential-outcome notation, the key assumption is conditional exchangeability:

$$\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid X.$$

The NBER notes call this unconfoundedness, and also note its close connection to selection on observables and conditional independence assumptions in econometrics [3, Lecture 1, Sec. 2.2]. The APTS notes pair this assumption with positivity, consistency, and no interference as standard assumptions for causal identification [4, Ch. 9].

Under consistency, conditional exchangeability, and positivity,

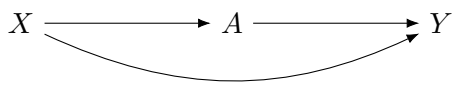
$$\begin{aligned}
\mathbb{E}[Y(a)] &= \mathbb{E}\{\mathbb{E}[Y(a) \mid X]\} \\
&= \mathbb{E}\{\mathbb{E}[Y(a) \mid A = a, X]\} \\
&= \mathbb{E}\{\mathbb{E}[Y \mid A = a, X]\} \\
&= \int \mathbb{E}[Y \mid A = a, X = x] dF_X(x).
\end{aligned}$$

For discrete X, replace the integral by a sum. This is the outcome-regression or g-formula version of covariate adjustment.

In DAG language, a sufficient graphical condition is that X is a valid back-door adjustment set for $A \to Y$. Pearl's back-door theorem gives

$$P(Y \mid \mathrm{do}(A = a)) = \sum_x P(Y \mid A = a, X = x) P(X = x),$$

which is the same identifying formula written as an interventional distribution [2, Sec. 3.3.1].



X blocks the back-door path $A \leftarrow X \to Y$

The ATT uses the same conditional means but a different covariate distribution:

$$\mathrm{ATT} = \mathbb{E}[Y \mid A = 1] - \int \mathbb{E}[Y \mid A = 0, X = x] dF_{X \mid A=1}(x).$$

Thus ATE and ATT may use the same adjustment set but different weights. This distinction is often obscured by a single regression coefficient.

6.3 What short and long regressions are doing causally

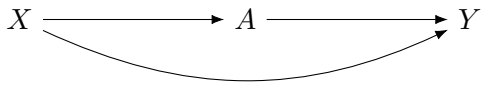
The short and long regressions from Section 4 are

$$Y_i = \alpha^S + \beta^S A_i + u_i^S, \qquad Y_i = \alpha^L + \beta^L A_i + X_i'\gamma^L + u_i^L.$$

The omitted-variable-bias identity says

$$\beta^S = \beta^L + \pi'\gamma^L,$$

where π comes from projecting X_i on A_i. When X is a genuine confounder, this algebra becomes causal intuition.



If X raises both treatment and outcome, then $\mathrm{Cov}(A, X) > 0$ and $\gamma^L > 0$, so the short coefficient is larger than the long coefficient in the scalar OVB formula. In potential-outcome language, treated units have higher untreated potential outcomes. In DAG language, the path $A \leftarrow X \rightarrow Y$ is open. In regression language, the short residual contains an omitted determinant of Y that is correlated with A.

Example 6.1 (Returns to schooling)**.** Let A_i be schooling and Y_i log wages. Ability, family background, and motivation may affect both schooling and wages:

$$A_i \leftarrow X_i \rightarrow Y_i.$$

MHE uses the returns-to-schooling setting to motivate omitted-variable bias: adding family background and ability-related controls changes the schooling coefficient because the controls are correlated with schooling and predict wages [1, Sec. 3.2.2]. The DAG says why such controls may matter for causal interpretation; the OVB formula describes how coefficients move in a linear projection.

Two warnings are important.

First,

$$\text{valid adjustment set} \neq \text{valid linear regression specification.}$$

A valid adjustment set identifies

$$\mathbb{E}[Y(a)] = \int \mathbb{E}[Y \mid A = a, X = x] dF_X(x).$$

A linear long regression estimates this functional exactly only under additional functional-form conditions, or under a saturated specification that correctly represents the relevant conditional means. With treatment-effect heterogeneity, a single OLS coefficient is generally a weighted average of conditional effects, not automatically the ATE or ATT [1, Ch. 3.3].

Second, a variable is not a good control merely because it predicts Y. It must be admissible for the causal estimand. The next subsection makes this point explicit.

6.4 Good controls, bad controls, and the timing of variables

MHE's practical rule is that good controls are variables fixed at the time the regressor of interest is determined, whereas bad controls are themselves outcome variables in the notional experiment [1, Sec. 3.2.3]. DAG language refines this rule by distinguishing confounders, mediators, colliders, descendants, and selection variables.

Variable type	**DAG pattern**	**Regression implication**
Confounder	$A \leftarrow X \rightarrow Y$	Omitting X leaves a back-door path open. Including X can move the long regression toward a causal contrast, subject to consistency, positivity, and no remaining unmeasured confounding.
Precision control	$X \rightarrow Y$, with no open path from X to A	Adjustment is not required for identification, but it may improve precision. Pre-treatment covariates in randomized experiments often play this role.
Mediator	$A \rightarrow M \rightarrow Y$	If the target is the total effect, controlling for M is a bad control because it blocks part of the causal path. The coefficient on A becomes direct-effect-like rather than total-effect-like.
Collider	$A \rightarrow C \leftarrow U \rightarrow Y$	Conditioning on C opens a path from A to Y through U. A long regression can be more biased than the short regression.
Selection variable	$A \rightarrow S \leftarrow Y$ or $A \rightarrow S \leftarrow U \rightarrow Y$	Restricting the sample or controlling for S can induce selection bias. This is the graphical logic behind many sample-selection problems.
Descendant of treatment	$A \rightarrow W$	Post-treatment controls are dangerous for total-effect estimation. They may block mechanisms, induce selection, or condition on variables whose distribution was changed by treatment.

Example 6.2 (College, occupation, and wages)**.** Suppose A is college completion, M is occupation, and Y is wages:

$$A \to M \to Y, \qquad A \to Y.$$

College may raise wages partly by changing occupations. MHE's bad-control discussion warns that comparing college and noncollege workers within occupation no longer estimates the total effect of college, even if college completion were randomized [1, Sec. 3.2.3]. The long regression with occupation controls answers a different question, closer to a controlled direct effect under additional assumptions.

Example 6.3 (Collider control and sample selection)**.** Suppose treatment affects whether a unit appears in the analysis sample S, and an unobserved trait U also affects sample inclusion and the outcome:

$$A \to S \leftarrow U \to Y.$$

Before conditioning on S, the path $A \to S \leftarrow U \to Y$ is blocked at the collider S. Restricting the sample to $S = 1$, or controlling for S, opens the path. In econometric language, this is sample-selection bias. It is distinct from treatment selection. Treatment selection is usually confounding, $A \leftarrow U \to Y$; sample selection is often collider conditioning, $A \to S \leftarrow U \to Y$.

Caveat

Longer regressions are not automatically more credible. Adding a true confounder can reduce bias. Adding a mediator can remove part of the effect. Adding a collider can create bias. Adding a pure outcome predictor can improve precision without changing identification. The decision is causal, not merely statistical.

6.5 Endogeneity taxonomy across frameworks

The word "endogeneity" is broad. It usually means that an econometric regressor is related to the structural error term in the equation being estimated. Potential outcomes and DAGs decompose this broad term into more specific causal failures.

Econometric term	**Potential-outcome expression**	**DAG/SCM expression**	**Common response**
Omitted variables / confounding	$\{Y(0), Y(1)\} \not\perp\!\!\!\perp A$, or $Y(a) \not\perp\!\!\!\perp A \mid X$ after inadequate controls	$A \leftarrow U \rightarrow Y$ or another unblocked back-door path	Measure and adjust for sufficient pre-treatment covariates; use randomized assignment, IV, DiD, RD, or bounds
Bad controls / post-treatment controls	Conditioning on $M(A)$ changes the counterfactual contrast for total effects	$A \rightarrow M \rightarrow Y$; controlling for M blocks part of the causal path	Do not control for mediators when estimating total effects; define a direct-effect or mediation estimand if mechanisms are the target
Collider control / sample selection	Conditioning on selection $S = 1$ can make A dependent on potential outcomes in the selected sample	$A \rightarrow S \leftarrow U \rightarrow Y$; conditioning on S opens a noncausal path	Avoid conditioning on colliders when possible; use selection models, weighting, sensitivity analysis, or redesign
Simultaneity / reverse causality	Treatment is not a well-ordered intervention without timing or a structural model	Feedback is not represented by a simple acyclic graph unless variables are time-indexed, e.g. $Y_t \rightarrow A_{t+1} \rightarrow Y_{t+2}$	Use timing, instruments, structural simultaneous-equation models, or longitudinal causal models
Measurement error	Observed A is not the treatment whose potential outcomes $Y(a)$ are defined for; true treatment may be A^*	$A^* \rightarrow A$ and $A^* \rightarrow Y$; regression uses a noisy proxy	Validation data, repeated measures, IV, latent-variable or errors-in-variables models
Panel unobserved heterogeneity	Treatment histories are associated with unit-specific potential outcome levels or trends	Unit factor α_i affects both A_{it} and Y_{it}; time-varying U_{it} may remain	Fixed effects for time-invariant heterogeneity; DiD/event-study assumptions for trends; longitudinal methods for time-varying confounding

This table is intentionally schematic. Its purpose is to separate regression failures from the causal assumptions that would make a comparison interpretable.

6.6 Regression, matching, weighting, and propensity scores

Under unconfoundedness and overlap, the causal estimand can be written using the observed CEF:

$$\mathbb{E}[Y(a)] = \int \mathbb{E}[Y \mid A = a, X = x] dF_X(x).$$

Different estimators implement this same identified functional in different ways.

Outcome regression estimates $\mathbb{E}[Y \mid A = a, X = x]$ and averages predictions over a target covariate distribution. Matching attempts to compare treated and untreated units with similar X. Inverse probability weighting uses the propensity score $e(X) = \mathbb{P}(A = 1 \mid X)$ to create a pseudo-population in which treatment is balanced with respect to X. Doubly robust methods combine outcome

and propensity-score models. The NBER notes emphasize that, under unconfoundedness, many estimators are possible, but overlap and the credibility of unconfoundedness are often more important than the particular estimator chosen [3, Lecture 1, Sec. 1].

This is also where CEF and potential outcomes meet. Under unconfoundedness,

$$\mathbb{E}[Y \mid A = a, X = x] = \mathbb{E}[Y(a) \mid A = a, X = x] = \mathbb{E}[Y(a) \mid X = x].$$

The observed CEF within treatment cells recovers the counterfactual regression function $\mu_a(x) = \mathbb{E}[Y(a) \mid X = x]$. Without unconfoundedness, the observed CEF remains useful for prediction but no longer equals the causal response function.

Takeaway

Endogeneity is regression language for a failure of the identifying comparison. Potential outcomes say which counterfactual independence failed. DAGs say which path or structural feature explains the failure. Regression provides the algebra and estimators once the causal assumptions are in place.

7 Design-based identification: IV, DiD, and RD

Design-based econometrics asks whether the empirical setting supplies credible variation in treatment. IV, DiD, and RD fit this note because they can all be written in potential-outcome notation, diagnosed with DAGs, and implemented with regressions. They differ in the source of identifying variation and in the estimand they usually identify.

7.1 Instrumental variables and LATE

Consider a binary instrument Z, binary treatment D, and outcome Y. Let $D_i(z)$ be the treatment unit i would receive under instrument value z, and let $Y_i(d, z)$ be the potential outcome under treatment d and instrument z. Observed treatment and outcome are

$$D_i = D_i(Z_i), \qquad Y_i = Y_i(D_i(Z_i), Z_i).$$

The canonical IV DAG is

$$Z \to D \to Y, \qquad U \to D, \qquad U \to Y,$$

where U is unobserved. The instrument shifts treatment but is otherwise unrelated to the outcome.

Assumption 7.1 (LATE assumptions; [1, 3])**.** For binary Z and D:

1. **Independence / as-if random assignment.**

$$Z_i \perp\!\!\!\perp \{D_i(0), D_i(1), Y_i(d,z) : d, z \in \{0,1\}\},$$

   possibly conditional on pre-instrument covariates X_i.

2. **Exclusion.** For all d,

$$Y_i(d,1) = Y_i(d,0) \equiv Y_i(d).$$

3. **Relevance / first stage.**

$$\mathbb{E}[D_i(1) - D_i(0)] \neq 0.$$

4. **Monotonicity.** Either $D_i(1) \geq D_i(0)$ for all i, or $D_i(1) \leq D_i(0)$ for all i. We can usually orient the labels of Z so that $D_i(1) \geq D_i(0)$.

Under the unconditional binary-IV version of these assumptions, with the labels of Z oriented so that $D_i(1) \geq D_i(0)$,

$$\frac{\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0]}{\mathbb{E}[D_i \mid Z_i = 1] - \mathbb{E}[D_i \mid Z_i = 0]} = \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i(1) > D_i(0)].$$

MHE states this as the LATE theorem and emphasizes that monotonicity rules out defiers; without monotonicity, IV need not be a weighted average of individual treatment effects [1, Sec. 4.4.1]. With treatment-effect heterogeneity, this estimand is generally not the ATE or ATT [1, Sec. 4.4.2].

DAGs help represent relevance, exclusion, and independence. They do not by themselves encode monotonicity, which is a restriction on the joint response of $D_i(0)$ and $D_i(1)$.

7.2 Difference-in-differences and fixed effects

DiD is most naturally expressed in potential-outcome notation. Let $G_i = 1$ indicate the group exposed to treatment in the post period, and let $t \in \{0,1\}$ denote pre and post. Suppose treatment occurs only for $G_i = 1$ in period 1:

$$D_{it} = G_i \cdot 1\{t = 1\}.$$

The usual target is the post-period ATT,

$$\mathrm{ATT}_1 = \mathbb{E}[Y_{i1}(1) - Y_{i1}(0) \mid G_i = 1].$$

The missing counterfactual is $\mathbb{E}[Y_{i1}(0) \mid G_i = 1]$: what treated-group units would have experienced in the post period without treatment.

The two-period parallel-trends assumption is

$$\mathbb{E}[Y_{i1}(0) - Y_{i0}(0) \mid G_i = 1] = \mathbb{E}[Y_{i1}(0) - Y_{i0}(0) \mid G_i = 0].$$

Under this assumption,

$$\begin{aligned}\mathrm{ATT}_1 = &\{\mathbb{E}[Y_{i1} \mid G_i = 1] - \mathbb{E}[Y_{i0} \mid G_i = 1]\} \\ &- \{\mathbb{E}[Y_{i1} \mid G_i = 0] - \mathbb{E}[Y_{i0} \mid G_i = 0]\}.\end{aligned}$$

The NBER DiD notes define the basic design exactly as two groups and two periods, with one group treated only in the second period and the control group never treated, and write the regression interaction coefficient as the DiD estimand [3, Lecture 10, Sec. 1]. MHE treats DiD together with fixed effects and panel data, emphasizing permanent group differences and trends [1, Ch. 5].

The corresponding two-way regression is

$$Y_{gt} = \alpha + \eta_g + \lambda_t + \tau(G_g \times Post_t) + u_{gt},$$

where η_g are group effects and λ_t are time effects. In richer panel settings,

$$Y_{it} = \alpha_i + \lambda_t + \tau D_{it} + \varepsilon_{it}$$

controls for time-invariant unit heterogeneity through α_i. But fixed effects do not by themselves remove time-varying confounding.

DAGs are useful for diagnosing DiD threats: anticipation, time-varying confounding, conditioning on post-treatment variables, attrition, and compositional change. However, the parallel-trends assumption is a restriction on counterfactual trends, not merely a static d-separation statement.

7.3 Regression discontinuity

RD exploits a known assignment rule based on a running variable R_i and cutoff c. In sharp RD,

$$D_i = 1\{R_i \geq c\}.$$

The NBER RD notes explicitly frame RD in the potential-outcome framework and state that the running variable may be associated with potential outcomes, but this association is assumed smooth at the cutoff; a discontinuity in observed outcomes at the cutoff is then interpreted as causal [3, Lecture 3, Sec. 2.1].

The sharp RD estimand is

$$\tau_{SRD} = \lim_{r \downarrow c} \mathbb{E}[Y_i \mid R_i = r] - \lim_{r \uparrow c} \mathbb{E}[Y_i \mid R_i = r],$$

where $r \downarrow c$ approaches c from above and $r \uparrow c$ approaches from below. Under continuity of potential-outcome regression functions,

$$\lim_{r \downarrow c} \mathbb{E}[Y_i(d) \mid R_i = r] = \lim_{r \uparrow c} \mathbb{E}[Y_i(d) \mid R_i = r], \qquad d \in \{0, 1\},$$

this contrast identifies the local effect at the cutoff,

$$\mathbb{E}[Y_i(1) - Y_i(0) \mid R_i = c],$$

interpreted through limits when R is continuous.

In fuzzy RD, treatment probability jumps at the cutoff but treatment is not deterministic. The cutoff-crossing indicator acts as an instrument. The estimand is the Wald ratio

$$\tau_{FRD} = \frac{\lim_{r \downarrow c} \mathbb{E}[Y_i \mid R_i = r] - \lim_{r \uparrow c} \mathbb{E}[Y_i \mid R_i = r]}{\lim_{r \downarrow c} \mathbb{E}[D_i \mid R_i = r] - \lim_{r \uparrow c} \mathbb{E}[D_i \mid R_i = r]}.$$

MHE describes fuzzy RD as an IV setup and notes that the resulting Wald estimand is a local average treatment effect for compliers at the cutoff [1, Sec. 6.2].

DAGs can help clarify manipulation or sorting around the cutoff, but RD is primarily a design plus continuity assumption rather than a back-door adjustment argument. MHE emphasizes that sharp RD exploits precise knowledge of treatment rules and that fuzzy RD leads to an IV-type setup [1, Ch. 6].

Takeaway

IV, DiD, and RD are not alternative definitions of causality. They are design-based ways to identify causal estimands when simple adjustment is not credible. Potential outcomes define the estimand and assumptions; DAGs diagnose threats; regressions implement the estimators.

8 Mechanisms, mediation, and moderation

Most of the note so far has focused on total effects: what is the effect of changing A on Y? In psychology, psychometrics, education, laboratory experiments, and mechanism-oriented social science, researchers often ask how or for whom the treatment works. This section links mediation and moderation to the previous frameworks.

8.1 From total effects to mechanisms

Let $T \in \{0, 1\}$ be treatment, M a mediator observed after treatment, and Y an outcome. Imai, Tingley, and Yamamoto define a causal mechanism as the process through which treatment affects the outcome, formalized by natural indirect effects or causal mediation effects [5]. Let

$$M_i(t) = \text{mediator value under } T_i = t,$$

and

$$Y_i(t, m) = \text{outcome under treatment } t \text{ and mediator value } m.$$

The observed mediator and outcome are

$$M_i = M_i(T_i), \qquad Y_i = Y_i(T_i, M_i(T_i)).$$

The unit total effect is

$$\tau_i = Y_i(1, M_i(1)) - Y_i(0, M_i(0)).$$

The unit natural indirect effect at treatment level t is

$$\delta_i(t) = Y_i(t, M_i(1)) - Y_i(t, M_i(0)), \qquad t \in \{0, 1\}.$$

It changes the mediator from the value it would take under control to the value it would take under treatment, while holding treatment fixed at t.

The unit natural direct effect at mediator level induced by t is

$$\zeta_i(t) = Y_i(1, M_i(t)) - Y_i(0, M_i(t)), \qquad t \in \{0, 1\}.$$

It changes treatment while holding the mediator at the value it would naturally take under t. Averaging over units gives $\bar{\tau}$, $\bar{\delta}(t)$, and $\bar{\zeta}(t)$, and the decomposition

$$\bar{\tau} = \bar{\delta}(t) + \bar{\zeta}(1 - t), \qquad t \in \{0, 1\},$$

which follows by adding and subtracting the same nested counterfactual [5, Sec. 2.1].

8.2 Why randomizing treatment alone is not enough

A standard randomized experiment can identify the total effect $\bar{\tau}$. It does not generally identify $\bar{\delta}(t)$ or $\bar{\zeta}(t)$. Imai, Tingley, and Yamamoto emphasize that treatment randomization alone cannot identify causal mechanisms because nested counterfactuals such as $Y_i(t, M_i(1-t))$ are never observed for any unit [5, Sec. 2].

This connects directly to the bad-control discussion in Section 6. A regression of Y on T and M does not automatically identify a direct effect. The mediator M is post-treatment. If there are unobserved common causes of M and Y,

$$M \leftarrow U \rightarrow Y,$$

conditioning on M can produce a biased direct-effect estimate.

8.3 Sequential ignorability

A common identification strategy for the single-experiment mediation design is sequential ignorability. In one standard formulation, for all treatment values t, t' and mediator values m,

$$\{Y_i(t', m), M_i(t)\} \perp\!\!\!\perp T_i \mid X_i = x,$$

and

$$Y_i(t', m) \perp\!\!\!\perp M_i(t) \mid T_i = t, X_i = x,$$

with suitable positivity conditions for treatment and mediator values. Equivalently, because $M_i = M_i(t)$ among units with $T_i = t$, the second condition is often written using the observed mediator M_i. The first condition is treatment ignorability, often satisfied by randomization. The second treats the mediator as if randomized conditional on treatment and pre-treatment covariates. Imai, Tingley, and Yamamoto stress that this mediator ignorability assumption is strong and often difficult for experimentalists to justify because unobserved mediator-outcome confounding is plausible [5, Sec. 2.3].

8.4 Experimental designs for mechanisms

Imai, Tingley, and Yamamoto propose designs that improve on the single-experiment mediation design [5].

1. **Single-experiment design.** Randomize treatment, observe mediator and outcome. This identifies total effects easily but mediation effects only under strong additional assumptions.

2. **Parallel design.** Randomly assign units to two experiments. In one experiment, only treatment is randomized. In the other, both treatment and mediator are manipulated. This can improve identification when the mediator can be directly manipulated.

3. **Crossover design.** Units are sequentially assigned so that each unit can contribute information under multiple treatment/mediator conditions, subject to no carryover or consistency-type assumptions.

4. **Parallel encouragement design.** Rather than perfectly manipulating the mediator, the researcher randomizes an encouragement that shifts the mediator. This is useful when mediators such as emotion, cognition, attention, or beliefs cannot be directly set.

5. **Crossover encouragement design.** This combines repeated exposure with encouragement rather than direct manipulation of the mediator.

These designs are particularly relevant for psychometrics and laboratory experiments. A researcher may randomize a stimulus, feedback, incentive, or information treatment, and then ask whether the effect operates through attention, perceived norms, emotion, memory, self-efficacy, or beliefs. The crucial lesson is the same as for IV and RD: design features must identify the relevant counterfactuals; adding a mediator to a regression is not enough.

8.5 Moderation and treatment-effect heterogeneity

Mediation asks *how* treatment works:

$$T \to M \to Y.$$

Moderation asks *for whom* or *under what conditions* treatment works. In potential-outcome notation, moderation is treatment-effect heterogeneity:

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x].$$

A common regression implementation is

$$Y_i = \alpha + \tau T_i + \theta X_i + \kappa(T_i X_i) + u_i.$$

If T_i is randomized, X_i is pre-treatment, and the linear interaction model is appropriate, κ describes how the treatment effect changes with X_i. If treatment is observational, the interaction coefficient is causal only under the same kind of identifying assumptions required for main effects.

Moderation is not the same as mediation. An interaction between T and X reveals heterogeneity across pre-treatment strata. A mediator is affected by treatment and lies on a causal path. Imai, Tingley, and Yamamoto also caution that statistical interaction is not, by itself, evidence of a causal process through a mediator [5, Sec. 2.2].

8.6 Connection to front-door identification

Front-door identification and mediation both involve a variable on a path from treatment to outcome, but they answer different questions. Mediation analysis decomposes effects into direct and indirect components. Front-door adjustment identifies a total effect by exploiting a mediator-like variable under graphical assumptions. Thus a mediator can be a bad control for the total effect, a target of mechanism analysis, or part of a front-door identification strategy, depending on the estimand and assumptions.

Takeaway

Mechanism analysis requires its own estimands and assumptions. A mediator is not a routine control. It is post-treatment information whose use must be justified by a direct-effect, indirect-effect, or front-door identification argument.

9 Final synthesis: comparison, workflow, and pitfalls

9.1 What each framework contributes

Framework	Primary strength	What it can obscure if used alone
Potential outcomes	Precise estimands, target populations, heterogeneity, missing counterfactuals, LATE, DiD ATT, mediation effects	Which variables to adjust for; collider structure; pathways; some selection mechanisms
DAGs / SCMs	Structural assumptions, interventions, path blocking, back-door and front-door identification, confounders, mediators, colliders	Overlap, treatment versions, target-population weighting, monotonicity, some cross-world restrictions
Econometric regression and design	Conditional means, short/long regressions, OVB algebra, endogeneity diagnostics, IV, DiD, RD, fixed effects, inference	The exact causal estimand and assumptions may remain implicit; estimates can be local or weighted

The three frameworks are complementary. A causal claim is strongest when the estimand is clear in potential-outcome notation, the identifying assumptions are transparent in DAG/SCM language, and the estimator follows from a credible design.

9.2 A recommended empirical workflow

Workflow

1. **State the causal question.** What intervention, treatment version, outcome, timing, unit, and population are under study?
2. **Define the estimand.** Write ATE, ATT, CATE, LATE, DiD ATT, RD local effect, direct effect, indirect effect, or another target explicitly.
3. **Describe the ideal experiment.** What randomized assignment or intervention would identify the estimand directly?

4. **Draw the causal graph or state the assignment rule.** Include confounders, instruments, mediators, colliders, selection variables, timing, and important unobserved variables.
5. **State assumptions in multiple languages.** Examples: $Y(a) \perp\!\!\!\perp A \mid X$, backdoor adjustment, IV exclusion and monotonicity, DiD parallel trends, RD continuity, mediation sequential ignorability.
6. **Derive the observed-data functional.** Examples: adjustment formula, ATT weighting formula, front-door formula, Wald ratio, DiD contrast, RD limit contrast.
7. **Choose an estimator.** Regression, matching, IPW, AIPW, 2SLS, fixed effects, event study, local linear RD, or experimental mediation design.
8. **Report the target population and threats.** Population, treated, untreated, compliers, cutoff units, cohorts, or mechanism-specific effects.

9.3 Common pitfalls

1. **Confusing association with intervention.** $P(Y \mid A = a)$ generally differs from $P(Y \mid \mathrm{do}(A = a))$.
2. **Confusing estimators with estimands.** OLS, matching, 2SLS, DiD, and RD are estimation strategies. ATE, ATT, LATE, and natural indirect effects are estimands.
3. **Assuming longer regressions are always better.** Adding a confounder can reduce bias; adding a mediator can block part of the effect; adding a collider can create bias.
4. **Calling every IV estimate an ATE.** With heterogeneous treatment effects, a valid binary IV generally identifies LATE for compliers.
5. **Ignoring overlap.** A graphical adjustment set does not solve lack of common support.
6. **Treating DiD as valid without parallel trends.** Pre-trends help diagnose plausibility, but the treated group's untreated post-treatment trend is unobserved.
7. **Treating RD as valid without continuity and no manipulation.** The running variable may be related to potential outcomes; RD requires smoothness through the cutoff apart from treatment.

8. **Using mediator regressions as mechanism evidence without assumptions.** Treatment randomization alone identifies total effects, not natural direct and indirect effects.

9. **Confusing moderation and mediation.** Moderation is heterogeneity by pre-treatment variables; mediation concerns post-treatment pathways.

10. **Leaving the intervention vague.** The causal effect of "education," "therapy," or "information" is ill-defined unless the intervention and versions are specified.

Takeaway

The most coherent causal analyses are bilingual or trilingual: they define counterfactual targets with potential outcomes, encode assumptions with graphs/SCMs, and implement estimates with econometric designs and regressions.

References

[1] Angrist, Joshua D., and Joern-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

[2] Pearl, Judea. 2010. "An Introduction to Causal Inference." *The International Journal of Biostatistics* 6(2), Article 7.

[3] Imbens, Guido W., and Jeffrey M. Wooldridge. 2007. *What Is New in Econometrics?* NBER Summer Institute Lecture Notes.

[4] Didelez, Vanessa, and Robin J. Evans. 2025. *Causal Inference*. APTS/StatML/Foundations of AI lecture notes, University of Oxford. https://www.stats.ox.ac.uk/~evans/APTS/.

[5] Imai, Kosuke, Dustin Tingley, and Teppei Yamamoto. 2013. "Experimental Designs for Identifying Causal Mechanisms." *Journal of the Royal Statistical Society: Series A* 176(1): 5–51. https://imai.fas.harvard.edu/research/files/Design.pdf.