Observational Studies and Treatment Effect Estimation

Statistics · Module 11 — Causality

Node 11.5

Position in Knowledge Graph

System: Statistics. Structural Domain: Causality (Module 11). Node 11.5.

When randomization is infeasible, observational studies must rely on identification assumptions to estimate causal effects. This node develops three major classes of methods: (1) propensity score methods under ignorability (IPW, matching), (2) instrumental variables when ignorability fails, and (3) quasi-experimental designs (difference-in-differences, regression discontinuity) that exploit specific structural features of the data-generating process. Each method identifies a specific estimand for a specific subpopulation under distinct assumptions. The doubly robust estimator unifies propensity score and outcome modeling approaches.

Formal Definition

In observational studies, treatment T_i may depend on covariates X_i and unobserved factors. Under **strong ignorability** $\{Y(0), Y(1)\} \perp\!\!\!\perp T \mid X$ and **overlap** $0 < e(x) < 1$, where

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

is the **propensity score**, the ATE is identified via the backdoor formula (Node 11.3).

When ignorability fails (unmeasured confounding), additional structure is needed: instruments, panel variation, or threshold rules.

Intuition

Observational studies face the core challenge that treated and control groups differ systematically. Propensity score methods attempt to recreate the conditions of a randomized experiment by reweighting or matching observations so that covariate distributions are balanced. Instrumental variables circumvent confounding by finding an external source of variation that affects treatment but not the outcome directly. Regression discontinuity exploits a known threshold rule to create local randomization. Each approach trades off generality (what estimand, what subpopulation) against the strength of required assumptions.

Structural Role

Dependencies: Requires identification theory (Node 11.3) as a logical prerequisite, randomized experiments (Node 11.4) as the benchmark, large-sample theory (Node 7.5), and regression (Node 8.6). **Enables:** Policy evaluation (Node 11.6), which extends these methods to decision rules and sequential settings.

All methods in this node require assumptions that are fundamentally untestable from observational data alone. The choice between methods depends on which assumptions are most credible in the specific application.

Mathematical Development

Propensity Score Methods

Theorem (Rosenbaum–Rubin, 1983). Under ignorability given X, $\{Y(0), Y(1)\} \perp\!\!\!\perp T \mid e(X)$.

This reduces conditioning from the potentially high-dimensional X to the scalar $e(X)$, enabling practical estimation.

Inverse Probability Weighting (IPW)

The **Horvitz–Thompson IPW estimator** of the ATE is

$$\hat{\tau}^{\mathrm{IPW}} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{T_i Y_i}{\hat{e}(X_i)} - \frac{(1 - T_i) Y_i}{1 - \hat{e}(X_i)} \right).$$

Derivation. Under ignorability:

$$\mathbb{E}\left[\frac{TY}{e(X)}\right] = \mathbb{E}\left[\frac{TY(1)}{e(X)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{T}{e(X)} \mid X\right] Y(1)\right] = \mathbb{E}[Y(1)],$$

since $\mathbb{E}[T/e(X) \mid X] = e(X)/e(X) = 1$. Similarly for $\mathbb{E}[(1-T)Y/(1-e(X))] = \mathbb{E}[Y(0)]$.

Hajek (normalized) estimator: $\hat{\tau}^{\mathrm{H}} = \frac{\sum_i T_i Y_i / \hat{e}(X_i)}{\sum_i T_i / \hat{e}(X_i)} - \frac{\sum_i (1-T_i) Y_i / (1-\hat{e}(X_i))}{\sum_i (1-T_i)/(1-\hat{e}(X_i))}$. This is more stable (lower variance) than the Horvitz–Thompson version.

Sensitivity to extreme propensity scores. When $e(x)$ is near 0 or 1, the weights $1/e(x)$ or $1/(1-e(x))$ explode, inflating variance. **Trimming** (excluding units with $e(x) < \epsilon$ or $e(x) > 1 - \epsilon$) or **truncation** (capping weights) mitigates this at the cost of changing the estimand.

Outcome Regression

The **outcome regression (OR) estimator** models $\mu_t(x) = \mathbb{E}[Y \mid T = t, X = x]$ and computes

$$\hat{\tau}^{\mathrm{OR}} = \frac{1}{n} \sum_{i=1}^{n} [\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)].$$

This is consistent under correct specification of $\mu_t(x)$ and does not require modeling the propensity score. However, it is sensitive to extrapolation in regions where treatment and control groups do not overlap in covariate space.

Doubly Robust (DR) Estimation

The **augmented IPW (AIPW)** estimator combines both approaches:

$$\hat{\tau}^{\mathrm{DR}} = \frac{1}{n} \sum_{i=1}^{n} \left[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{T_i (Y_i - \hat{\mu}_1(X_i))}{\hat{e}(X_i)} - \frac{(1 - T_i)(Y_i - \hat{\mu}_0(X_i))}{1 - \hat{e}(X_i)} \right].$$

Double robustness. $\hat{\tau}^{\mathrm{DR}}$ is consistent if either $\hat{e}$ or $\hat{\mu}_t$ is correctly specified (but not necessarily both).

Proof sketch. Write $\hat{\tau}^{\mathrm{DR}} = \hat{\tau}^{\mathrm{OR}} +$ correction. If $\hat{\mu}_t$ is correct, the residuals $Y_i - \hat{\mu}_t(X_i)$ have conditional mean zero and the correction vanishes. If $\hat{e}$ is correct, the IPW component identifies the ATE independently.

The correction term is the product of estimation errors in both models, which is $o_p(1/\sqrt{n})$ when either model is correct.

Semiparametric efficiency. The DR estimator achieves the semiparametric efficiency bound for the ATE under the nonparametric model. Its influence function is

$$\phi(Y,T,X) = \frac{T(Y-\mu_1(X))}{e(X)} - \frac{(1-T)(Y-\mu_0(X))}{1-e(X)} + \mu_1(X) - \mu_0(X) - \tau,$$

and the efficiency bound is $\mathrm{Var}(\phi)/n$.

Matching

Nearest-neighbor matching constructs a matched control for each treated unit:

$$\hat{\tau}^{\text{match}} = \frac{1}{n_1}\sum_{i:T_i=1}\big(Y_i - Y_{j(i)}\big), \quad j(i) = \arg\min_{j:T_j=0}\|X_i - X_j\|.$$

This estimates the ATT. Matching on the propensity score $e(X)$ is valid under ignorability and avoids the curse of dimensionality in X-space.

Bias of matching. With a fixed number of matches M, the bias is $O(n^{-1/d})$ in d dimensions. Bias correction (Abadie–Imbens, 2011) adjusts for the remaining difference between matched and target covariate values.

Instrumental Variables (IV)

When ignorability fails, an **instrument** Z provides identification if:

1. Z affects T (**relevance**).
2. Z affects Y only through T (**exclusion restriction**).
3. Z is independent of confounders (**independence**).

Wald estimator (binary Z, binary T):

$$\hat{\tau}^{\text{IV}} = \frac{\bar{Y}_{Z=1} - \bar{Y}_{Z=0}}{\bar{T}_{Z=1} - \bar{T}_{Z=0}} = \frac{\text{Reduced form}}{\text{First stage}}.$$

Two-stage least squares (2SLS). Stage 1: regress T on Z and covariates to get $\hat{T}$. Stage 2: regress Y on $\hat{T}$ and covariates. The coefficient on $\hat{T}$ is $\hat{\tau}^{\text{2SLS}}$.

Under monotonicity ($T_i(1) \geq T_i(0)$ for all i), IV identifies the **LATE** (Local Average Treatment Effect) for compliers:

$$\tau^{\text{LATE}} = \mathbb{E}[Y(1) - Y(0) \mid T(1) > T(0)].$$

Difference-in-Differences (DiD)

With panel data (units observed before and after treatment):

$$\hat{\tau}^{\text{DiD}} = (\bar{Y}_{1,\text{post}} - \bar{Y}_{1,\text{pre}}) - (\bar{Y}_{0,\text{post}} - \bar{Y}_{0,\text{pre}}).$$

Parallel trends assumption: absent treatment, treated and control units would have followed the same time trend: $\mathbb{E}[Y_i(0)_{\text{post}} - Y_i(0)_{\text{pre}} \mid T=1] = \mathbb{E}[Y_i(0)_{\text{post}} - Y_i(0)_{\text{pre}} \mid T=0]$.

Under parallel trends, DiD identifies the ATT. The assumption is untestable but can be assessed by checking pre-treatment trend similarity.

Regression Discontinuity (RD)

Treatment is assigned by a running variable R_i crossing a threshold c: $T_i = \mathbf{1}(R_i \geq c)$ (sharp RD).

$$\tau^{\mathrm{RD}} = \lim_{r \downarrow c} \mathbb{E}[Y \mid R = r] - \lim_{r \uparrow c} \mathbb{E}[Y \mid R = r].$$

This identifies the ATE at the threshold $R = c$, under continuity of potential outcome means.

Fuzzy RD. When T is not a deterministic function of R but exhibits a discontinuity in $e(R)$ at c, the fuzzy RD estimand is

$$\tau^{\mathrm{FRD}} = \frac{\lim_{r \downarrow c} \mathbb{E}[Y \mid R = r] - \lim_{r \uparrow c} \mathbb{E}[Y \mid R = r]}{\lim_{r \downarrow c} \mathbb{E}[T \mid R = r] - \lim_{r \uparrow c} \mathbb{E}[T \mid R = r]},$$

which is a local Wald (IV) estimator using the threshold as the instrument.

Worked Examples

Example 1: IPW Estimation

Binary treatment, 3 covariate strata. Data:

Stratum	n	n_1	$\bar{Y}_1$	$\bar{Y}_0$	e
A	100	70	15	10	0.7
B	200	100	12	9	0.5
C	100	20	18	14	0.2

IPW estimate of $\mathbb{E}[Y(1)]$: Weight treated observations by $1/e$. Within each stratum, $\sum T_i Y_i / e = n_1 \bar{Y}_1 / e$. Normalized:

$\hat{\mathbb{E}}[Y(1)] = \frac{70 \cdot 15/0.7 + 100 \cdot 12/0.5 + 20 \cdot 18/0.2}{70/0.7 + 100/0.5 + 20/0.2} = \frac{1500+2400+1800}{100+200+100} = \frac{5700}{400} = 14.25.$

$\hat{\mathbb{E}}[Y(0)] = \frac{30 \cdot 10/0.3 + 100 \cdot 9/0.5 + 80 \cdot 14/0.8}{30/0.3 + 100/0.5 + 80/0.8} = \frac{1000+1800+1400}{100+200+100} = \frac{4200}{400} = 10.5.$

$\hat{\tau}^{\mathrm{IPW}} = 14.25 - 10.5 = 3.75.$

Naive (unadjusted): $\hat{\tau}^{\mathrm{naive}} = \frac{70 \cdot 15 + 100 \cdot 12 + 20 \cdot 18}{190} - \frac{30 \cdot 10 + 100 \cdot 9 + 80 \cdot 14}{210} = 13.05 - 11.33 = 1.72$. The naive estimate is biased because treatment is more likely in stratum A (high e) which has lower outcome.

Example 2: Regression Discontinuity

Test score threshold for a scholarship: $c = 75$. Running variable: test score R. Treatment: scholarship ($T = \mathbf{1}(R \geq 75)$). Outcome: college GPA.

Fit local linear regressions on each side of the cutoff using a bandwidth $h = 5$:

Left of cutoff ($R \in [70, 75)$): $\hat{\mathbb{E}}[Y \mid R = 75^-] = 2.80$. Right of cutoff ($R \in [75, 80]$): $\hat{\mathbb{E}}[Y \mid R = 75^+] = 3.15$.

$\hat{\tau}^{\mathrm{RD}} = 3.15 - 2.80 = 0.35$ GPA points.

This is the causal effect of the scholarship for students at the threshold. Validity requires that students cannot precisely manipulate their test scores to cross the threshold (no sorting), so that the density of R is continuous at c. A McCrary density test checks this.

Cross-Links

AI. Off-policy evaluation in reinforcement learning uses IPW and doubly robust estimators to estimate the value of a target policy from data collected under a different behavior policy. This is the sequential generalization of observational causal inference, where the "treatment" is a sequence of actions.

Economics. IV estimation is the workhorse of empirical economics (Angrist–Imbens Nobel, 2021). DiD is standard in policy evaluation (Card–Krueger minimum wage study). RD is used in education (test score thresholds), healthcare (age cutoffs for eligibility), and political science (close elections).

Linear Algebra. 2SLS can be written as a projection: $\hat{\beta}^{2\text{SLS}} = (X^\top P_Z X)^{-1} X^\top P_Z Y$, where $P_Z = Z(Z^\top Z)^{-1} Z^\top$ is the projection onto the instrument space. The first stage is the projection of T onto the column space of Z.

Quantum Mechanics. In quantum causal inference, unmeasured confounders correspond to entangled quantum states shared between parties. Instrumental variable-type arguments (using random measurement settings as instruments) are used to bound causal effects in the presence of quantum confounding, with Bell inequalities as the key identification constraints.

Structural Compression

Invariant

Under ignorability and overlap, the ATE is identified and the DR estimator achieves the semiparametric efficiency bound. When ignorability fails, IV identifies the LATE for compliers, DiD identifies the ATT under parallel trends, and RD identifies the threshold ATE under continuity.

Mechanism

IPW reweights the observed distribution to remove confounding; outcome regression models the conditional expectation surface; the DR estimator combines both and is consistent under misspecification of either nuisance model. IV exploits exogenous variation; DiD exploits time-series variation; RD exploits threshold rules. Each design replaces ignorability with a different structural assumption.

Cross-System Echo

Doubly robust estimation is a special case of semiparametric efficiency theory (Bickel–Klaassen–Ritov–Wellner). In AI, off-policy evaluation uses these estimators for distribution shift correction. In economics, the hierarchy IV $>$ DiD $>$ RD $>$ selection-on-observables reflects decreasing reliance on untestable assumptions, and the credibility revolution is built on matching research designs to credible identification strategies.

Exercises

1. Derive the IPW estimator by showing that $\mathbb{E}[TY/e(X)] = \mathbb{E}[Y(1)]$ under ignorability and overlap. State where each assumption is used.

2. Show that the DR estimator is consistent when the propensity score model is correctly specified but the outcome model is misspecified. Then show the converse.

3. For the Wald estimator $\hat{\tau}^{\text{IV}} = (\bar{Y}_{Z=1} - \bar{Y}_{Z=0})/(\bar{T}_{Z=1} - \bar{T}_{Z=0})$, derive the asymptotic distribution using the delta method and compute the standard error.

4. State the parallel trends assumption for DiD formally using potential outcomes notation. Construct a numerical example where parallel trends holds and DiD recovers the ATT, and another where it fails.

5. Show that the sharp RD estimand $\tau^{\text{RD}} = \lim_{r \downarrow c} \mathbb{E}[Y(1) \mid R = r] - \lim_{r \uparrow c} \mathbb{E}[Y(0) \mid R = r]$ equals the ATE at the threshold under continuity of $\mathbb{E}[Y(t) \mid R = r]$ at $r = c$.

6. Compare the assumptions and estimands of IPW, IV, DiD, and RD in a table. For each method, give one example where it is the appropriate design and explain why the alternatives are not credible.
7. Argue, structurally, why no observational method can replicate the unconditional guarantees of a randomized experiment, and identify the specific untestable assumption that each method introduces as the price for working without randomization.

Statistics · Module 11 — Causality · Node 11.5
Concepts: conditional-probability, maximum-likelihood-estimation, convergence
Prerequisites: 11.3, 11.4, 7.5, 8.6
Enables: 11.6