Selection on Observables

Causal Inference using Machine Learning
Master in Economics, UNT

Andres Mena

Spring 2024

Content

Confounding in Observational Studies

Causal Inference for Observational Studies:

- Experimental studies are often not feasible due to ethical, practical, or financial constraints.
- Instead, we rely on observational data, where treatment assignment is not controlled by the researcher.

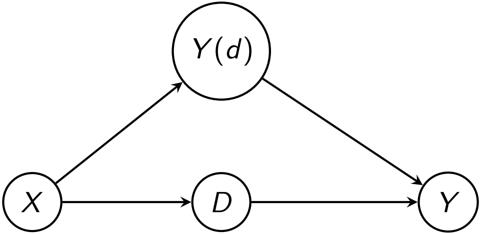
Notation:

- Data: $\{(Y_i, D_i, X_i) : i = 1, \ldots, N\}$ are i.i.d. from an infinite super-population.
- $X_i \in \mathbb{R}^K$: vector of pre-treatment covariates.

Key Concern: Confounding Factors

- Confounders are variables related to both the treatment assignment and the outcome.
- If not properly accounted for, confounding leads to biased estimates of causal effects.

Causal Graph: Potential Outcomes and Confounding



- X influences both the treatment D and the potential outcomes $Y(d)$.
- Potential outcomes $Y(d)$ determine the realized outcome Y.
- Note: D does not affect $Y(d)$ directly; $Y(d)$ is defined as the outcome *if* D were set to d.

Key Assumptions

Unconfoundedness (Conditional Ignorability)

Assumption: $(Y(1), Y(0)) \perp D \mid X$.

Key Assumptions

Unconfoundedness (Conditional Ignorability)

Assumption: $(Y(1), Y(0)) \perp D \mid X$.

Overlap

Assumption: For all $x \in \mathcal{X}$, $0 < p(x) < 1$, where $p(x) = P(D = 1 \mid X = x)$.

Conditioning Removes Selection Bias

Theorem 1(Conditioning on X Removes Selection Bias)

Under Unconfoundedness and Overlap,

$$E[Y \mid D = d, X] = E[Y(d) \mid X].$$

Proof:

Identification of the Average Treatment Effect (ATE)

Theorem 2 (Identification of ATE)

Statement: Under Unconfoundedness and Overlap,

$$\text{ATE} \;=\; \int_{\mathcal{X}} \big(E[Y \mid D=1, X=x] - E[Y \mid D=0, X=x]\big)\, dF_X(x).$$

Proof:

Content

Definition of ATE and ATT

Average Treatment Effect (ATE):

$$\text{ATE} = E[Y(1)-Y(0)] = \int_{\mathcal{X}} \big(E[Y \mid D=1, X=x]-E[Y \mid D=0, X=x]\big)\, dF$$

Definition of ATE and ATT

Average Treatment Effect (ATE):

$$\text{ATE} = E[Y(1) - Y(0)] = \int_{\mathcal{X}} \big(E[Y \mid D = 1, X = x] - E[Y \mid D = 0, X = x] \big) \, dF$$

Average Treatment Effect on the Treated (ATT):

$$\text{ATT} = E[Y(1) - Y(0) \mid D = 1] = \int_{\mathcal{X}} \big(E[Y \mid D = 1, X = x] - E[Y \mid D = 0, X$$

Definition of ATE and ATT

Average Treatment Effect (ATE):

$$\text{ATE} = E[Y(1) - Y(0)] = \int_{\mathcal{X}} \big(E[Y \mid D = 1, X = x] - E[Y \mid D = 0, X = x]\big)\, d$$

Average Treatment Effect on the Treated (ATT):

$$\text{ATT} = E[Y(1) - Y(0) \mid D = 1] = \int_{\mathcal{X}} \big(E[Y \mid D = 1, X = x] - E[Y \mid D = 0, X$$

- Both ATE and ATT rely on conditional expectations $E[Y \mid D, X]$.
- Once we identify these conditional expectations, we integrate over the appropriate distribution of X.

Estimating ATE and ATT by Parts (Separate Regressions)

Step 1: Estimate $E[Y|D = 0, X]$ by OLS using observations with $D = 0$ only.

Estimating ATE and ATT by Parts (Separate Regressions)

Step 1: Estimate $E[Y|D = 0, X]$ by OLS using observations with $D = 0$ only.
Step 2: Estimate $E[Y|D = 1, X]$ by OLS using observations with $D = 1$ only.

Estimating ATE and ATT by Parts (Separate Regressions)

Step 1: Estimate $E[Y|D=0,X]$ by OLS using observations with $D=0$ only.
Step 2: Estimate $E[Y|D=1,X]$ by OLS using observations with $D=1$ only.
Estimate ATE:

$$\widehat{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N}\big(\widehat{E}(Y|D=1,X_i) - \widehat{E}(Y|D=0,X_i)\big)$$

Estimating ATE and ATT by Parts (Separate Regressions)

Step 1: Estimate $E[Y|D=0,X]$ by OLS using observations with $D=0$ only.
Step 2: Estimate $E[Y|D=1,X]$ by OLS using observations with $D=1$ only.
Estimate ATE:

$$\widehat{\text{ATE}} = \frac{1}{N}\sum_{i=1}^{N}\big(\widehat{E}(Y|D=1,X_i) - \widehat{E}(Y|D=0,X_i)\big)$$

Estimate ATT:

$$\widehat{\text{ATT}} = \frac{1}{N_1}\sum_{i:D_i=1}\big(\widehat{E}(Y|D=1,X_i) - \widehat{E}(Y|D=0,X_i)\big)$$

where $N_1 = \sum_{i=1}^{N} 1\{D_i = 1\}$.

Estimating ATE and CATE with a Pooled Regression

Pooled Model:

$$E[Y|D,X] = \alpha_1 D + \alpha_2' WD + \beta_1 + \beta_2' W,$$

where W includes X and its transformations (centered $E[W] = 0$).

Estimating ATE and CATE with a Pooled Regression

Pooled Model:

$$E[Y|D,X] = \alpha_1 D + \alpha_2' WD + \beta_1 + \beta_2' W,$$

where W includes X and its transformations (centered $E[W] = 0$).

Interpreting Parameters:

- ATE: $\widehat{\alpha_1}$ recovers the ATE when W is centered.
- CATE: $\delta(X) = \alpha_1 + \alpha_2' W$ captures treatment heterogeneity.

Estimating ATE and CATE with a Pooled Regression

Pooled Model:

$$E[Y|D,X] = \alpha_1 D + \alpha_2' WD + \beta_1 + \beta_2' W,$$

where W includes X and its transformations (centered $E[W] = 0$).

Interpreting Parameters:

- ATE: $\widehat{\alpha_1}$ recovers the ATE when W is centered.
- CATE: $\delta(X) = \alpha_1 + \alpha_2' W$ captures treatment heterogeneity.

Estimating ATT: If interested in ATT, you can take the estimated conditional means from the model and average them over the treated sample distribution of X, analogous to the separate regressions approach:

$$\widehat{\text{ATT}} = \frac{1}{N_1} \sum_{i:D_i=1} \left(\widehat{\alpha_1} + \widehat{\alpha_2}' W_i \right).$$

Estimating ATE and CATE with a Pooled Regression

Pooled Model:

$$E[Y|D,X] = \alpha_1 D + \alpha_2' WD + \beta_1 + \beta_2' W,$$

where W includes X and its transformations (centered $E[W] = 0$).

Interpreting Parameters:

- ATE: $\widehat{\alpha_1}$ recovers the ATE when W is centered.
- CATE: $\delta(X) = \alpha_1 + \alpha_2' W$ captures treatment heterogeneity.

Estimating ATT: If interested in ATT, you can take the estimated conditional means from the model and average them over the treated sample distribution of X, analogous to the separate regressions approach:

$$\widehat{\text{ATT}} = \frac{1}{N_1} \sum_{i:D_i=1} \left(\widehat{\alpha_1} + \widehat{\alpha_2}' W_i \right).$$

Note: In high-dimensional settings, partialling out and machine learning methods (e.g., Double Lasso) can be employed to improve flexibility and inference.

Regression on the Propensity Score

Theorem 3: Rosenbaum & Rubin (1983)

Under unconfoundedness:

$$(Y(1), Y(0)) \perp D \mid p(X),$$

where $p(X) = P(D = 1 \mid X)$ is the propensity score.

Regression on the Propensity Score

Theorem 3: Rosenbaum & Rubin (1983)

Under unconfoundedness:

$$(Y(1), Y(0)) \perp D \mid p(X),$$

where $p(X) = P(D = 1 \mid X)$ is the propensity score.

Implication:

- Under unconfoundedness, conditioning on $p(X)$ (the propensity score) suffices to remove confounding.
- This allows for more parsimonious models by reducing the dimensionality of X.

Regression on the Propensity Score

Theorem 3: Rosenbaum & Rubin (1983)

Under unconfoundedness:

$$(Y(1), Y(0)) \perp D \mid p(X),$$

where $p(X) = P(D = 1 \mid X)$ is the propensity score.

Implication:

- Under unconfoundedness, conditioning on $p(X)$ (the propensity score) suffices to remove confounding.
- This allows for more parsimonious models by reducing the dimensionality of X.

Steps for Propensity Score Regression:

1. *Estimate* $p(X)$ using a flexible binary regression model (e.g., logistic regression or machine learning methods).

Steps for Propensity Score Regression:

1. *Estimate* $p(X)$ using a flexible binary regression model (e.g., logistic regression or machine learning methods).
2. *Run regressions* to estimate $E(Y \mid D = d, p(X))$ using the estimated propensity scores $\widehat{p}(X)$.

Steps for Propensity Score Regression:

1. *Estimate* $p(X)$ using a flexible binary regression model (e.g., logistic regression or machine learning methods).
2. *Run regressions* to estimate $E(Y \mid D = d, p(X))$ using the estimated propensity scores $\widehat{p}(X)$.
3. *Compute:*

$$\widehat{\delta}(x_i) = \widehat{E}(Y \mid D = 1, p = \widehat{p}(X_i)) - \widehat{E}(Y \mid D = 0, p = \widehat{p}(X_i)).$$

Steps for Propensity Score Regression:

1. *Estimate* $p(X)$ using a flexible binary regression model (e.g., logistic regression or machine learning methods).
2. *Run regressions* to estimate $E(Y \mid D = d, p(X))$ using the estimated propensity scores $\widehat{p}(X)$.
3. *Compute:*

$$\widehat{\delta}(x_i) = \widehat{E}(Y \mid D = 1, p = \widehat{p}(X_i)) - \widehat{E}(Y \mid D = 0, p = \widehat{p}(X_i)).$$

4. *Take the sample average of* $\widehat{\delta}(x_i)$ *to estimate ATE or ATT.*

Propensity Score Blocking

Alternative Approach: Propensity Score Blocking

- Divide the range of $\widehat{p}(X)$ into blocks or strata (e.g., deciles).

Propensity Score Blocking

Alternative Approach: Propensity Score Blocking

- Divide the range of $\widehat{p}(X)$ into blocks or strata (e.g., deciles).
- Within each block, assume $E(Y \mid D, p(X))$ is approximately constant.

Propensity Score Blocking

Alternative Approach: Propensity Score Blocking

- Divide the range of $\widehat{p}(X)$ into blocks or strata (e.g., deciles).
- Within each block, assume $E(Y \mid D, p(X))$ is approximately constant.
- Estimate $E(Y \mid D = d, p(X))$ within each block as the average outcome for treated $(D = 1)$ and control $(D = 0)$ units.

Propensity Score Blocking

Alternative Approach: Propensity Score Blocking

- Divide the range of $\widehat{p}(X)$ into blocks or strata (e.g., deciles).
- Within each block, assume $E(Y \mid D, p(X))$ is approximately constant.
- Estimate $E(Y \mid D = d, p(X))$ within each block as the average outcome for treated $(D = 1)$ and control $(D = 0)$ units.
- Aggregate across blocks to compute:

$$\widehat{\text{ATE}} = \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{E}(Y \mid D = 1, p = \widehat{p}(X_i)) - \widehat{E}(Y \mid D = 0, p = \widehat{p}(X_i)) \right).$$

Propensity Score Blocking

Alternative Approach: Propensity Score Blocking

- Divide the range of $\widehat{p}(X)$ into blocks or strata (e.g., deciles).
- Within each block, assume $E(Y \mid D, p(X))$ is approximately constant.
- Estimate $E(Y \mid D = d, p(X))$ within each block as the average outcome for treated $(D = 1)$ and control $(D = 0)$ units.
- Aggregate across blocks to compute:

$$\widehat{\text{ATE}} = \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{E}(Y \mid D = 1, p = \widehat{p}(X_i)) - \widehat{E}(Y \mid D = 0, p = \widehat{p}(X_i)) \right).$$

Pros:

- Nonparametric approach that avoids imposing a functional form on $E(Y \mid D, p(X))$.

Cons:

- Requires sufficient sample size within each block to ensure reliable estimates.
- Sensitivity to the choice of the number and width of blocks.

Content

Proving Identification of $E[Y(0)]$ and $E[Y(1)]$ Using IPW

Theorem 4 (Horvitz-Thompson: Propensity Score Reweighting Removes Bias)

$$E\left[\frac{Y \cdot 1(D=d)}{P(D=d|X)} \mid X\right] = E[Y(d) \mid X]$$

Proof Outline:

Estimating ATE and ATT Using IPW

ATE Estimation Using the Horvitz-Thompson Formula:

$$\widehat{\text{ATE}}_{\text{IPW}} = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{D_i Y_i}{\widehat{p}(X_i)} - \frac{(1 - D_i) Y_i}{1 - \widehat{p}(X_i)} \right].$$

Estimating ATE and ATT Using IPW

ATE Estimation Using the Horvitz-Thompson Formula:

$$\widehat{\text{ATE}}_{\text{IPW}} = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{D_i Y_i}{\widehat{p}(X_i)} - \frac{(1 - D_i) Y_i}{1 - \widehat{p}(X_i)} \right].$$

Normalized Propensity Score Weighting (Recommended):

- Adjust weights to sum to 1 within treated and control groups for improved stability.

Estimating ATE and ATT Using IPW

ATE Estimation Using the Horvitz-Thompson Formula:

$$\widehat{\text{ATE}}_{\text{IPW}} = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{D_i Y_i}{\widehat{p}(X_i)} - \frac{(1 - D_i) Y_i}{1 - \widehat{p}(X_i)} \right].$$

Normalized Propensity Score Weighting (Recommended):

- Adjust weights to sum to 1 within treated and control groups for improved stability.

ATT Estimation Using IPW:

$$\widehat{\text{ATT}}_{\text{IPW}} = \frac{1}{N_t} \sum_{i:D_i=1} Y_i - \frac{1}{N_c} \sum_{i:D_i=0} \frac{P(\widehat{D = 0})}{P(\widehat{D = 1})} \frac{Y_i}{1 - \widehat{p}(X_i)}.$$

Estimating ATE and ATT Using IPW

ATE Estimation Using the Horvitz-Thompson Formula:

$$\widehat{\text{ATE}}_{\text{IPW}} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{D_i Y_i}{\widehat{p}(X_i)} - \frac{(1-D_i)Y_i}{1-\widehat{p}(X_i)}\right].$$

Normalized Propensity Score Weighting (Recommended):

- Adjust weights to sum to 1 within treated and control groups for improved stability.

ATT Estimation Using IPW:

$$\widehat{\text{ATT}}_{\text{IPW}} = \frac{1}{N_t}\sum_{i:D_i=1} Y_i - \frac{1}{N_c}\sum_{i:D_i=0}\frac{\widehat{P(D=0)}}{\widehat{P(D=1)}}\frac{Y_i}{1-\widehat{p}(X_i)}.$$

Role of Propensity Score Weighting:

- Equalizes the distribution of X across treatment and control groups.
- For ATE: $X \mid D=1$ weighted by $P(D=1)/p(X)$ matches the marginal distribution of X.
- For ATT: $X \mid D=0$ weighted by $\frac{P(D=0)p(X)}{P(D=1)(1-p(X))}$ matches $X \mid D=1$.

Content

Doubly Robust Methods

Setup:

- Let $E(Y \mid D = d, X = x) = E(Y(d) \mid X = x) = \mu_d(x; \beta)$, the outcome regression model.
- Let $P(D = 1 \mid X = x) = p(x; \gamma)$, the propensity score model.

Doubly Robust Methods

Setup:

- Let $E(Y \mid D = d, X = x) = E(Y(d) \mid X = x) = \mu_d(x; \beta)$, the outcome regression model.
- Let $P(D = 1 \mid X = x) = p(x; \gamma)$, the propensity score model.

Moment Condition for $\theta_{\textbf{ATE}}$:

$$\theta_{\text{ATE}} = E\left[\mu_1(X;\beta) - \mu_0(X;\beta) + \frac{D(Y - \mu_1(X;\beta))}{p(X;\gamma)} - \frac{(1-D)(Y - \mu_0(X;\beta))}{1 - p(X;\gamma)}\right].$$

Doubly Robust Methods

Setup:

- Let $E(Y \mid D = d, X = x) = E(Y(d) \mid X = x) = \mu_d(x; \beta)$, the outcome regression model.
- Let $P(D = 1 \mid X = x) = p(x; \gamma)$, the propensity score model.

Moment Condition for θ_{ATE}:

$$\theta_{\text{ATE}} = E\left[\mu_1(X;\beta) - \mu_0(X;\beta) + \frac{D(Y - \mu_1(X;\beta))}{p(X;\gamma)} - \frac{(1-D)(Y - \mu_0(X;\beta))}{1 - p(X;\gamma)}\right].$$

Doubly Robust (DR) Estimator for θ_{ATE}:

$$\widehat{\theta}_{\text{DR, ATE}} = \frac{1}{N}\sum_{i=1}^{N}\left[\mu_1(X_i;\widehat{\beta}) - \mu_0(X_i;\widehat{\beta}) + \frac{D_i(Y_i - \mu_1(X_i;\widehat{\beta}))}{\widehat{p}(X_i;\widehat{\gamma})} - \frac{(1-D_i)(Y_i - \mu_0(X_i;\widehat{\beta}))}{1 - \widehat{p}(X_i;\widehat{\gamma})}\right].$$

Doubly Robust Methods

Setup:

- Let $E(Y \mid D = d, X = x) = E(Y(d) \mid X = x) = \mu_d(x;\beta)$, the outcome regression model.
- Let $P(D = 1 \mid X = x) = p(x;\gamma)$, the propensity score model.

Moment Condition for $\theta_{\textbf{ATE}}$:

$$\theta_{\text{ATE}} = E\left[\mu_1(X;\beta) - \mu_0(X;\beta) + \frac{D(Y - \mu_1(X;\beta))}{p(X;\gamma)} - \frac{(1-D)(Y - \mu_0(X;\beta))}{1 - p(X;\gamma)}\right].$$

Doubly Robust (DR) Estimator for $\theta_{\textbf{ATE}}$:

$$\widehat{\theta}_{\text{DR, ATE}} = \frac{1}{N}\sum_{i=1}^{N}\left[\mu_1(X_i;\widehat{\beta}) - \mu_0(X_i;\widehat{\beta}) + \frac{D_i(Y_i - \mu_1(X_i;\widehat{\beta}))}{\widehat{p}(X_i;\widehat{\gamma})} - \frac{(1-D_i)(Y_i - \mu_0(X_i;\widehat{\beta}))}{1 - \widehat{p}(X_i;\widehat{\gamma})}\right].$$

Key Properties:

- **Doubly Robust:** The estimator is consistent if either:
  - The outcome regression model $\mu_d(X;\beta)$ is correctly specified, OR
  - The propensity score model $p(X;\gamma)$ is correctly specified.

Doubly Robust Methods

Setup:

- Let $E(Y \mid D = d, X = x) = E(Y(d) \mid X = x) = \mu_d(x; \beta)$, the outcome regression model.
- Let $P(D = 1 \mid X = x) = p(x; \gamma)$, the propensity score model.

Moment Condition for $\theta_{\textbf{ATE}}$:

$$\theta_{\text{ATE}} = E\left[\mu_1(X;\beta) - \mu_0(X;\beta) + \frac{D(Y - \mu_1(X;\beta))}{p(X;\gamma)} - \frac{(1-D)(Y - \mu_0(X;\beta))}{1 - p(X;\gamma)}\right].$$

Doubly Robust (DR) Estimator for $\theta_{\textbf{ATE}}$:

$$\widehat{\theta}_{\text{DR, ATE}} = \frac{1}{N}\sum_{i=1}^{N}\left[\mu_1(X_i;\widehat{\beta}) - \mu_0(X_i;\widehat{\beta}) + \frac{D_i(Y_i - \mu_1(X_i;\widehat{\beta}))}{\widehat{p}(X_i;\widehat{\gamma})} - \frac{(1-D_i)(Y_i - \mu_0(X_i;\widehat{\beta}))}{1 - \widehat{p}(X_i;\widehat{\gamma})}\right].$$

Key Properties:

- **Doubly Robust:** The estimator is consistent if either:
  - The outcome regression model $\mu_d(X; \beta)$ is correctly specified, OR
  - The propensity score model $p(X; \gamma)$ is correctly specified.
- **Efficiency:** If both models are correctly specified, the estimator is more efficient than using either model alone.

Content

Neyman Orthogonality

Motivation: In modern econometrics, we often estimate causal parameters θ while also estimating high-dimensional nuisance functions β and/or γ. Examples include:

- Outcome regression functions $\mu_d(X; \beta)$
- Propensity scores $p(X; \gamma)$

Neyman Orthogonality

Motivation: In modern econometrics, we often estimate causal parameters θ while also estimating high-dimensional nuisance functions β and/or γ. Examples include:

- Outcome regression functions $\mu_d(X; \beta)$
- Propensity scores $p(X; \gamma)$

Problem: Naive estimators are sensitive to estimation errors in these nuisance parameters. If $\widehat{\beta}$ or $\widehat{\gamma}$ converge slowly, such errors can cause large biases in the causal parameter estimates.

Neyman Orthogonality

Motivation: In modern econometrics, we often estimate causal parameters θ while also estimating high-dimensional nuisance functions β and/or γ. Examples include:

- Outcome regression functions $\mu_d(X;\beta)$
- Propensity scores $p(X;\gamma)$

Problem: Naive estimators are sensitive to estimation errors in these nuisance parameters. If $\widehat{\beta}$ or $\widehat{\gamma}$ converge slowly, such errors can cause large biases in the causal parameter estimates.

Neyman Orthogonality: A property of a moment equation (or estimator) that makes it *insensitive* to small perturbations in the nuisance parameter estimates. Formally, the first-order derivative of the moment condition with respect to the nuisance parameters at the true value is zero. This ensures that small estimation errors in β or γ do not induce first-order bias in the estimator of θ.

Naive estimatior

Comparison: Regression-based Estimator for ATE

Consider a regression-based ATE estimator:

$$\widehat{\theta} = \frac{1}{N}\sum_{i=1}^{N}[\mu_1(X_i;\widehat{\beta}) - \mu_0(X_i;\widehat{\beta})] = \frac{1}{N}\sum_{i=1}^{N}\Delta\mu(X_i,\hat{\beta}).$$

Naive estimatior

Comparison: Regression-based Estimator for ATE

Consider a regression-based ATE estimator:

$$\widehat{\theta} = \frac{1}{N}\sum_{i=1}^{N}[\mu_1(X_i;\widehat{\beta}) - \mu_0(X_i;\widehat{\beta})] = \frac{1}{N}\sum_{i=1}^{N}\Delta\mu(X_i,\hat{\beta}).$$

Taylor Expansion: Let θ_0 be the true ATE and β_0 the true parameter. Using a Taylor expansion around β_0:

$$\sqrt{N}(\widehat{\theta} - \theta_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} m_{1i}(\beta_0) + G_\beta \cdot \sqrt{N}(\widehat{\beta} - \beta_0) + o_p(1),$$

Comparison: Regression-based Estimator for ATE

Consider a regression-based ATE estimator:

$$\widehat{\theta} = \frac{1}{N}\sum_{i=1}^{N}[\mu_1(X_i;\widehat{\beta}) - \mu_0(X_i;\widehat{\beta})] = \frac{1}{N}\sum_{i=1}^{N}\Delta\mu(X_i,\hat{\beta}).$$

Taylor Expansion: Let θ_0 be the true ATE and β_0 the true parameter. Using a Taylor expansion around β_0:

$$\sqrt{N}(\widehat{\theta} - \theta_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} m_{1i}(\beta_0) + G_\beta \cdot \sqrt{N}(\widehat{\beta} - \beta_0) + o_p(1),$$

where:

- G_β is the **derivative (gradient)** of the estimator's moment condition with respect to β at β_0.

Naive estimatior

Comparison: Regression-based Estimator for ATE

Consider a regression-based ATE estimator:

$$\widehat{\theta} = \frac{1}{N}\sum_{i=1}^{N}[\mu_1(X_i;\widehat{\beta}) - \mu_0(X_i;\widehat{\beta})] = \frac{1}{N}\sum_{i=1}^{N}\Delta\mu(X_i,\hat{\beta}).$$

Taylor Expansion: Let θ_0 be the true ATE and β_0 the true parameter. Using a Taylor expansion around β_0:

$$\sqrt{N}(\widehat{\theta} - \theta_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} m_{1i}(\beta_0) + G_\beta \cdot \sqrt{N}(\widehat{\beta} - \beta_0) + o_p(1),$$

where:

- G_β is the **derivative (gradient)** of the estimator's moment condition with respect to β at β_0.
- $m_{1i}(\beta_0) = \Delta\mu(X_i,\hat{\beta}) - E[\Delta\mu(X_i,\hat{\beta})]$

Comparison: Regression-based Estimator for ATE

Consider a regression-based ATE estimator:

$$\widehat{\theta} = \frac{1}{N}\sum_{i=1}^{N}[\mu_1(X_i;\widehat{\beta}) - \mu_0(X_i;\widehat{\beta})] = \frac{1}{N}\sum_{i=1}^{N}\Delta\mu(X_i,\hat{\beta}).$$

Taylor Expansion: Let θ_0 be the true ATE and β_0 the true parameter. Using a Taylor expansion around β_0:

$$\sqrt{N}(\widehat{\theta} - \theta_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} m_{1i}(\beta_0) + G_\beta \cdot \sqrt{N}(\widehat{\beta} - \beta_0) + o_p(1),$$

where:

- G_β is the **derivative (gradient)** of the estimator's moment condition with respect to β at β_0.
- $m_{1i}(\beta_0) = \Delta\mu(X_i,\hat{\beta}) - E[\Delta\mu(X_i,\hat{\beta})]$

Consequence: If $\widehat{\beta} - \beta_0$ converges more slowly than $N^{-1/2}$, the term $G_\beta \cdot \sqrt{N}(\widehat{\beta} - \beta_0)$ introduces a first-order bias. Hence, the accuracy of $\widehat{\theta}$ critically depends on the rate at which $\widehat{\beta}$ converges.

Neyman Orthogonality and Double Robustness

Double Robust (DR) Estimator for ATE:

$$\widehat{\theta}_{\text{DR}} = \frac{1}{N}\sum_{i=1}^{N}\left[(\mu_1(X_i;\widehat{\beta}) - \mu_0(X_i;\widehat{\beta})) + \frac{D_i(Y_i - \mu_1(X_i;\widehat{\beta}))}{\widehat{p}(X_i;\widehat{\gamma})} - \frac{(1-D_i)(Y_i - \mu_0(X_i;\widehat{\beta}))}{1-\widehat{p}(X_i;\widehat{\gamma})}\right].$$

Neyman Orthogonality and Double Robustness

Double Robust (DR) Estimator for ATE:

$$\widehat{\theta}_{\text{DR}} = \frac{1}{N}\sum_{i=1}^{N}\left[(\mu_1(X_i;\widehat{\beta}) - \mu_0(X_i;\widehat{\beta})) + \frac{D_i(Y_i - \mu_1(X_i;\widehat{\beta}))}{\widehat{p}(X_i;\widehat{\gamma})} - \frac{(1-D_i)(Y_i - \mu_0(X_i;\widehat{\beta}))}{1-\widehat{p}(X_i;\widehat{\gamma})}\right].$$

Taylor Expansion of the DR Estimator: Let $\psi(X_i;\beta,\gamma)$ represent the influence function in the above bracketed term. Expanding around (β_0,γ_0):

$$\widehat{\theta}_{\text{DR}} - \theta_0 = \frac{1}{N}\sum_{i=1}^{N}[\psi(X_i;\beta_0,\gamma_0) - E\{\psi(X;\beta_0,\gamma_0)\}] + \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial\psi}{\partial\beta},\frac{\partial\psi}{\partial\gamma}\right)\Big|_{(\beta_0,\gamma_0)}(\widehat{\beta}-\beta_0,\widehat{\gamma}-\gamma_0)' + o_p(1)$$

Neyman Orthogonality and Double Robustness

Double Robust (DR) Estimator for ATE:

$$\widehat{\theta}_{\mathrm{DR}} = \frac{1}{N}\sum_{i=1}^{N}\left[(\mu_1(X_i;\widehat{\beta}) - \mu_0(X_i;\widehat{\beta})) + \frac{D_i(Y_i - \mu_1(X_i;\widehat{\beta}))}{\widehat{p}(X_i;\widehat{\gamma})} - \frac{(1-D_i)(Y_i - \mu_0(X_i;\widehat{\beta}))}{1-\widehat{p}(X_i;\widehat{\gamma})}\right].$$

Taylor Expansion of the DR Estimator: Let $\psi(X_i;\beta,\gamma)$ represent the influence function in the above bracketed term. Expanding around (β_0,γ_0):

$$\widehat{\theta}_{\mathrm{DR}} - \theta_0 = \frac{1}{N}\sum_{i=1}^{N}[\psi(X_i;\beta_0,\gamma_0) - E\{\psi(X;\beta_0,\gamma_0)\}] + \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial\psi}{\partial\beta},\frac{\partial\psi}{\partial\gamma}\right)\Big|_{(\beta_0,\gamma_0)}(\widehat{\beta}-\beta_0,\widehat{\gamma}-\gamma_0)' + o_p(1)$$

Neyman Orthogonality: For the DR estimator, the partial derivatives $\frac{\partial\psi}{\partial\beta}$ and $\frac{\partial\psi}{\partial\gamma}$ at (β_0,γ_0) are zero. Hence:

$$G_\beta = \frac{\partial E[\psi]}{\partial\beta}\Big|_{\beta_0,\gamma_0} = 0, \quad G_\gamma = \frac{\partial E[\psi]}{\partial\gamma}\Big|_{\beta_0,\gamma_0} = 0.$$

Neyman Orthogonality and Double Robustness

Double Robust (DR) Estimator for ATE:

$$\widehat{\theta}_{\text{DR}} = \frac{1}{N}\sum_{i=1}^{N}\left[(\mu_1(X_i;\widehat{\beta}) - \mu_0(X_i;\widehat{\beta})) + \frac{D_i(Y_i - \mu_1(X_i;\widehat{\beta}))}{\widehat{p}(X_i;\widehat{\gamma})} - \frac{(1-D_i)(Y_i - \mu_0(X_i;\widehat{\beta}))}{1-\widehat{p}(X_i;\widehat{\gamma})}\right].$$

Taylor Expansion of the DR Estimator: Let $\psi(X_i;\beta,\gamma)$ represent the influence function in the above bracketed term. Expanding around (β_0,γ_0):

$$\widehat{\theta}_{\text{DR}} - \theta_0 = \frac{1}{N}\sum_{i=1}^{N}[\psi(X_i;\beta_0,\gamma_0) - E\{\psi(X;\beta_0,\gamma_0)\}] + \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial\psi}{\partial\beta},\frac{\partial\psi}{\partial\gamma}\right)\Bigg|_{(\beta_0,\gamma_0)}(\widehat{\beta}-\beta_0,\widehat{\gamma}-\gamma_0)' + o_p(1)$$

Neyman Orthogonality: For the DR estimator, the partial derivatives $\frac{\partial\psi}{\partial\beta}$ and $\frac{\partial\psi}{\partial\gamma}$ at (β_0,γ_0) are zero. Hence:

$$G_\beta = \frac{\partial E[\psi]}{\partial\beta}\bigg|_{\beta_0,\gamma_0} = 0, \quad G_\gamma = \frac{\partial E[\psi]}{\partial\gamma}\bigg|_{\beta_0,\gamma_0} = 0.$$

Implications:

- With $G_\beta = G_\gamma = 0$, the leading bias terms vanish.
- The DR estimator is robust to first-order estimation errors in β and γ.
- If either the outcome model or the propensity score model is correctly specified, the DR estimator remains consistent, making it highly valuable in high-dimensional or complex modeling scenarios.

Content

Defining the Moment Condition

Key Ingredients

- DML is based on the **method-of-moments** framework, targeting a low-dimensional parameter of interest, θ_0, defined via the moment condition:

$$E[\psi(W; \theta_0, \eta_0)] = 0,$$

  where:

  - ψ: Score function.
  - W: Data vector.
  - θ_0: Parameter of interest.
  - η_0: Nuisance parameters (unknown high-dimensional functions).

Defining the Moment Condition

Key Ingredients

- DML is based on the **method-of-moments** framework, targeting a low-dimensional parameter of interest, θ_0, defined via the moment condition:

  $$E[\psi(W;\theta_0,\eta_0)] = 0,$$

  where:
  - ψ: Score function.
  - W: Data vector.
  - θ_0: Parameter of interest.
  - η_0: Nuisance parameters (unknown high-dimensional functions).
- **Interpretation:** θ_0 is identified when the above equation holds.

Neyman Orthogonality

Key Concept

- A score function $\psi(W;\theta,\eta)$ satisfies **Neyman orthogonality** if:

$$\frac{\partial}{\partial\eta}E[\psi(W;\theta_0,\eta)]\bigg|_{\eta=\eta_0} = 0.$$

Neyman Orthogonality

Key Concept

- A score function $\psi(W;\theta,\eta)$ satisfies **Neyman orthogonality** if:

$$\frac{\partial}{\partial\eta} E[\psi(W;\theta_0,\eta)]\Big|_{\eta=\eta_0} = 0.$$

- **Importance:** Eliminates first-order bias from errors in the nuisance parameter estimates, $\hat{\eta}$.

Neyman Orthogonality

Key Concept

- A score function $\psi(W;\theta,\eta)$ satisfies **Neyman orthogonality** if:

$$\frac{\partial}{\partial \eta} E[\psi(W;\theta_0,\eta)]\bigg|_{\eta=\eta_0} = 0.$$

- **Importance:** Eliminates first-order bias from errors in the nuisance parameter estimates, $\hat{\eta}$.

Remark

- Named after Jerzy Neyman.
- Ensures robustness in high-dimensional settings, where $\hat{\eta}$ is regularized and inherently biased.

Gateaux Derivative

Definition

- The **Gateaux derivative** formalizes sensitivity to small perturbations:

$$\frac{\partial}{\partial \eta} E[\psi(W;\theta,\eta)][\Delta] := \frac{\partial}{\partial t} E[\psi(W;\theta,\eta + t\Delta)]\Big|_{t=0}.$$

Gateaux Derivative

Definition

- The **Gateaux derivative** formalizes sensitivity to small perturbations:

$$\frac{\partial}{\partial \eta} E[\psi(W;\theta,\eta)][\Delta] := \frac{\partial}{\partial t} E[\psi(W;\theta,\eta + t\Delta)]\bigg|_{t=0}.$$

- **Implication:** Neyman orthogonality implies:

$$\frac{\partial}{\partial \eta} E[\psi(W;\theta_0,\eta_0)][\Delta] = 0, \quad \forall \Delta.$$

Gateaux Derivative

Definition

- The **Gateaux derivative** formalizes sensitivity to small perturbations:

$$\frac{\partial}{\partial \eta} E[\psi(W;\theta,\eta)][\Delta] := \frac{\partial}{\partial t} E[\psi(W;\theta,\eta + t\Delta)]\Big|_{t=0}.$$

- **Implication:** Neyman orthogonality implies:

$$\frac{\partial}{\partial \eta} E[\psi(W;\theta_0,\eta_0)][\Delta] = 0, \quad \forall \Delta.$$

- **Admissible Directions:** Δ is admissible if $\eta_0 + t\Delta$ stays in the parameter space for small t.

Good Learners for Nuisance Functions

Requirements for High-Quality Learners

- Learners must approximate the true nuisance parameters η_0 well:

$$n^{1/4}\|\hat{\eta} - \eta_0\|_{L^2} \approx 0.$$

Good Learners for Nuisance Functions

Requirements for High-Quality Learners

- Learners must approximate the true nuisance parameters η_0 well:

$$n^{1/4}\|\hat{\eta} - \eta_0\|_{L^2} \approx 0.$$

- Examples of Machine Learning Methods:
  1. **LASSO:** For sparsely parameterized η_0.
  2. **Random Forests:** For tree-like structures in η_0.
  3. **Deep Neural Networks:** For η_0 approximable by sparse deep nets.
  4. **Ensemble Models:** Combining methods to leverage strengths of each.

Good Learners for Nuisance Functions

Requirements for High-Quality Learners

- Learners must approximate the true nuisance parameters η_0 well:

$$n^{1/4}\|\hat{\eta} - \eta_0\|_{L^2} \approx 0.$$

- Examples of Machine Learning Methods:
  1. **LASSO:** For sparsely parameterized η_0.
  2. **Random Forests:** For tree-like structures in η_0.
  3. **Deep Neural Networks:** For η_0 approximable by sparse deep nets.
  4. **Ensemble Models:** Combining methods to leverage strengths of each.
- Cross-validation and careful tuning are critical for robust performance.

Cross-Fitting

Why Cross-Fitting?

- Prevents **overfitting**, which occurs when nuisance parameter estimates are correlated with the same data used for inference.

Cross-Fitting

Why Cross-Fitting?

- Prevents **overfitting**, which occurs when nuisance parameter estimates are correlated with the same data used for inference.
- Mechanism:
  1. Split data into K folds.
  2. Use $K - 1$ folds to estimate nuisance parameters ($\hat{\eta}$).
  3. Use the left-out fold to compute residuals and estimate the target parameter.

Cross-Fitting

Why Cross-Fitting?

- Prevents **overfitting**, which occurs when nuisance parameter estimates are correlated with the same data used for inference.
- Mechanism:
  1. Split data into K folds.
  2. Use $K-1$ folds to estimate nuisance parameters ($\hat{\eta}$).
  3. Use the left-out fold to compute residuals and estimate the target parameter.
- **Outcome:** Avoids biases arising from overfitting complex machine learning methods.

Example 1: Partially Linear Model (PLM)

Moment Condition for PLM

$$\psi(W;\theta,\eta) = (Y - \ell(X) - \theta(D - m(X)))\,(D - m(X)).$$

- $W = (Y, D, X)$: Observable variables.
- $\eta = (\ell, m)$: Nuisance parameters.
    - $\ell(X) = E[Y|X]$, $m(X) = E[D|X]$.

Example 1: Partially Linear Model (PLM)

Moment Condition for PLM

$$\psi(W;\theta,\eta) = (Y - \ell(X) - \theta(D - m(X)))\,(D - m(X)).$$

- $W = (Y, D, X)$: Observable variables.
- $\eta = (\ell, m)$: Nuisance parameters.
    - $\ell(X) = E[Y|X]$, $m(X) = E[D|X]$.

Neyman Orthogonality

- Using elementary calculations:

$$\frac{\partial}{\partial \eta} E[\psi(W;\theta,\eta)]\Big|_{\eta=\eta_0} = 0.$$

Example 1: Partially Linear Model (PLM)

Moment Condition for PLM

$$\psi(W;\theta,\eta) = (Y - \ell(X) - \theta(D - m(X)))\,(D - m(X)).$$

- $W = (Y, D, X)$: Observable variables.
- $\eta = (\ell, m)$: Nuisance parameters.
  - $\ell(X) = E[Y|X]$, $m(X) = E[D|X]$.

Neyman Orthogonality

- Using elementary calculations:

$$\frac{\partial}{\partial\eta} E[\psi(W;\theta,\eta)]\Big|_{\eta=\eta_0} = 0.$$

Interpretation: $\psi(W;\theta,\eta)$ generalizes residualization in linear models, enabling robust inference.

Example 2: Doubly Robust IPW

Score for ATE

$$\psi(W;\theta,\eta) = (g(1,X) - g(0,X)) + H(D,X)(Y - g(D,X)) - \theta,$$

where:

$$H(D,X) = \frac{D}{m(X)} - \frac{(1-D)}{1-m(X)}.$$

Example 2: Doubly Robust IPW

Score for ATE

$$\psi(W;\theta,\eta) = (g(1,X) - g(0,X)) + H(D,X)(Y - g(D,X)) - \theta,$$

where:

$$H(D,X) = \frac{D}{m(X)} - \frac{(1-D)}{1-m(X)}.$$

- $g(D,X) = E[Y|D,X]$, $m(X) = P[D=1|X]$.
- Neyman Orthogonality:

$$\frac{\partial}{\partial\eta}E[\psi(W;\theta,\eta)] = 0.$$

Generic DML Algorithm

Steps

1. **Input:** Data $\{W_i\}_{i=1}^n$, Neyman orthogonal score $\psi(W;\theta,\eta)$, and machine learning methods for η.
2. **Cross-Fitting:**
   - Split data into K folds.
   - Train $\hat{\eta}[k]$ on $K-1$ folds and compute residuals on the left-out fold.
3. **Moment Estimation:**
$$\hat{M}(\theta,\hat{\eta}) = \frac{1}{n}\sum_{i=1}^{n}\psi(W_i;\theta,\hat{\eta}[k(i)]).$$
4. **Solve for θ:**
$$\hat{M}(\hat{\theta},\hat{\eta}) = 0.$$
5. **Variance and Confidence Intervals:**
   - Estimate variance $\hat{V}$:
$$\hat{V} = \frac{1}{n}\sum_{i=1}^{n}\phi(W_i)\phi(W_i)'.$$