SOC 690S: Machine Learning in Causal Inference

Week 4: Neyman Orthogonality and Causal Inference Basics

Wenhao Jiang
Department of Sociology, Fall 2025

Neyman Orthogonality

Why do we need Neyman Orthogonality

- We want to estimate a causal effect of a treatment D_i on an outcome Y_i
- One problem that is central to our course is that there is a high-dimensional set of controls X_i that confound D_i and Y_i
- Ordinary regression becomes problematic when X_i is large in dimension or is highly nonlinear

Why do we need Neyman Orthogonality

- We want to estimate a causal effect of a treatment D_i on an outcome Y_i
- One problem that is central to our course is that there is a high-dimensional set of controls X_i that confound D_i and Y_i
- Ordinary regression becomes problematic when X_i is large in dimension or is highly nonlinear
- We borrowed the insight from the *Frisch-Waugh-Lovell* (FWL) theorem and produce estimate of Y_i and D_i based on X_i using Machine Learning
- The justification of why such *Double Machine Learning* technique works is the *Neyman Orthogonality*
- There are other methods, such as *Augmented Inverse Propensity Weighting*, that do not rely on the particular form of *double partialling out* but satisfy *Neyman Orthogonality*

The Structural Model

- Assume the following structural equation in the *population*, where the causal effect of D_i on Y_i is well defined:

$$Y_i = \theta_0 D_i + g_0(X_i) + \epsilon_i, \quad E[\epsilon_i | D_i, X_i] = 0$$

- θ_0: parameter of interest (causal effect of D_i)
- $g_0(X_i)$: nuisance function capturing the effect of X_i on Y_i net of D_i
- ϵ_i: error term, mean zero conditional on D_i, X_i
- Note that $g_0(X_i) \neq E[Y_i | X_i]$

$$E[Y_i | X_i] = \theta_0 m_0(X_i) + g_0(X_i), \quad m_0(X_i) = E[D_i | X_i]$$

Normal Equation from the Structural Model

- Define residualized treatment:

$$\tilde{D}_i = D_i - m_0(X_i), \quad m_0(X_i) = E[D_i|X_i]$$

- Define residualized outcome (structural)

$$\tilde{Y}_i = Y_i - g_0(X_i)$$

- *Population* normal equation

$$E\left[(\tilde{Y}_i - \theta_0 \tilde{D}_i)\tilde{D}_i\right] = 0$$

- This identifies θ_0 under exogeneity

Equivalence with FWL Residualization

- By the *Frisch-Waugh-Lovell* (FWL) theorem, we can also residualize using conditional expectation:

$$\tilde{Y}_i = Y_i - E[Y_i|X_i], \quad \tilde{D}_i = D_i - E[D_i|X_i]$$

- The *population* normal equation is

$$\begin{aligned} E[(Y_i - E[Y_i|X_i] - \theta_0(D_i - m_0(X_i))) \cdot (D_i - m_0(X_i)] &= 0 \\ E[(Y_i - g_0(X_i) - \theta_0 m_0(X_i) - \theta_0(D_i - m_0(X_i))) \cdot (D_i - m_0(X_i)] &= 0 \\ E\left[(\tilde{Y}_i - \theta_0 \tilde{D}_i)\tilde{D}_i\right] &= 0 \end{aligned}$$

- FWL residualization and the structural model lead to the *same normal equation*

The Score Function

- Generalize to generic *nuisance functions* $g(\cdot), m(\cdot)$

$$\tilde{Y}_i = Y_i - g(X_i), \quad \tilde{D}_i = D_i - m(X_i)$$

- Define the *score function* that is analogous to the *normal equation* based on the FWL theorem

$$\psi(W_i; \theta, g, m) = (Y_i - g(X_i) - \theta(D_i - m(X_i)))(D_i - m(X_i))$$

- θ_0 can be identified with moment condition satisfying

$$E[\psi(W_i; \theta_0, g, m)] = 0 \quad \text{where } g = E[Y|X],\ m = E[D|X]$$

- We write $\psi(W_i; \theta, g, m)$ as $\psi(W_i; \theta, \eta)$ where $\eta = (g, m)$

Moment Function and Neyman Orthogonality

- Formally, moment condition is defined as

$$M(\theta, \eta) = E[\psi(W; \theta, \eta)]$$

- At the true nuisances $\eta = \eta_0$ (g_0 and m_0 correctly specified), the moment condition has a unique root at θ_0; that is

$$M(\theta, \eta_0) = 0 \text{ if and only if } \theta = \theta_0$$

- Remember in the *structural equation*, $g_0(X_i)$ is defined as the effect of X_i on Y_i *net of* D_i
- In practice, replace $g_0(X_i)$ by $g(X_i) = E[Y_i|X_i]$ produces the same *normal equation* and identify the same θ_0

Moment Function and Neyman Orthogonality

- In reality, we do not know the true *nuisance functions*

$$g_0(X_i) = E[Y_i|X_i], \quad m_0(X_i) = E[D_i|X_i]$$

- We can only approximate them using *finite samples* and *predictive* methods
- The key idea is that we want estimation errors in $\hat{\eta}$ to have minimal impact on $\hat{\theta}$
- *Neyman Orthogonality:* the score ψ is *Neyman orthogonal* if

$$\partial_\eta M(\theta_0, \eta)\Big|_{\eta=\eta_0} = 0$$

- It means that the slope of M in the *nuisance* direction is flat at the truth
- If we plug in $\hat{\eta}$ that is close to η_0, the bias in the estimation of M and the associated *normal equation* is only *second order*, not first order

Neyman Orthogonality via Gateaux Derivative

- Gateaux derivative is the functional derivative

$$\partial_g M(\theta_0, g, m_0)[\Delta]\Big|_{g=g_0} = \lim_{t\to 0} \frac{M(\theta_0, g_0 + t\Delta, m_0) - M(\theta_0, g_0, m_0)}{t}$$

$$M(\theta_0, g_0 + t\Delta, m_0) = E\Big[(Y_i - g_0(X_i) - t\Delta(X_i) - \theta_0(D_i - m_0(X_i)))(D_i - m_0(X_i))\Big]$$

$$= M(\theta_0, g_0, m_0) - tE[\Delta(X_i)(D_i - m_0(X_i))]$$

$$\partial_g M(\theta_0, g, m_0)[\Delta]\Big|_{g=g_0} = -E[\Delta(X_i)(D_i - m_0(X_i))]$$

- Since $E[D_i - m_0(X_i)|X_i] = 0$, this expectation is zero for all directions Δ (*CEF Decomposition Property*)
- By symmetry, the derivative *w.r.t.* m also vanishes at (g_0, m_0)
- The gradient of $M(\theta_0, \eta)$ with respect to $\eta = (g, m)$ is *0 at the truth*

Taylor Expansion of the Moment Function

- Expand around the true nuisances $\eta_0 = (g_0, m_0)$ using Taylor Expansion

$$M(\theta_0, \hat{\eta}) \approx M(\theta_0, \eta)\Big|_{\eta=\eta_0} + \underbrace{\left[\partial_\eta M(\theta_0, \eta)\Big|_{\eta=\eta_0}\right](\hat{\eta} - \eta_0)}_{\text{first order vanishes}}$$

$$+ 1/2(\hat{\eta} - \eta_0)' \left[\partial_\eta^2 M(\theta_0, \eta)\Big|_{\eta=\eta_0}\right](\hat{\eta} - \eta_0) + \textit{ higher order}$$

- *Without orthogonality*, the first-order term drives bias
- *With orthogonality*, the first-order term vanishes, and the remaining error is mainly *second-order* in $(\hat{\eta} - \eta_0)$
- With Neyman orthogonality, it suffices for the nuisance estimates to converge at rate faster than $n^{1/4}$, rather than the much stronger $n^{1/2}$ rate that is generally impossible in high dimensions

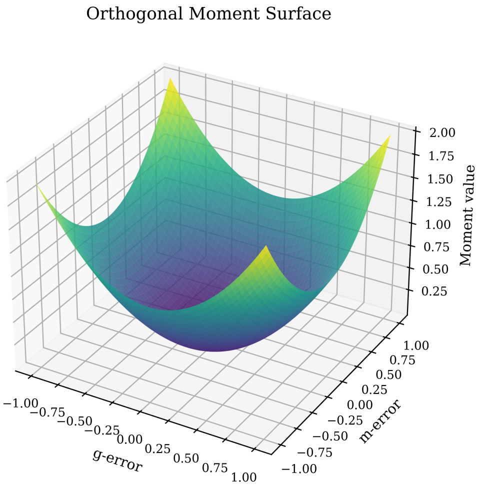
Heuristic Geometric Representation

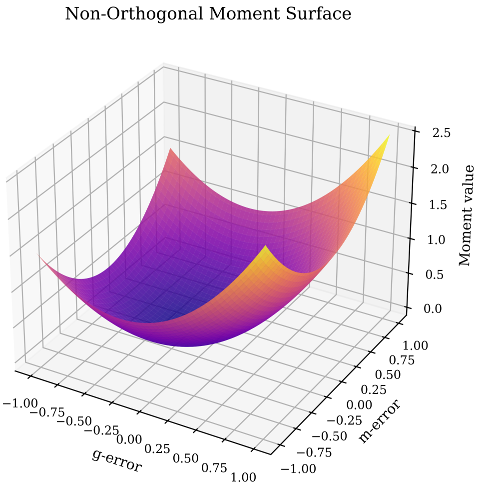
- Think of the moment function

$$M(\theta_0, g, m) = E[\psi(W; \theta_0, g, m)]$$

  as a *surface* over the nuisance directions (g, m)
- If the functional gradient *w.r.t.* (g, m) is nonzero, the surface is *tilted*; small errors in $(\hat{g}, \hat{m})$ shift the zero point and bias the estimation of θ_0
- If the gradient is zero (*orthogonality*), the surface is *flat* in nuisance directions at the truth; θ_0 is robust to small nuisance estimation error

Orthogonal vs. Non-Orthogonal Surfaces





Double Machine Learning

Invalid Single LASSO Estimation (Naive Method)

- We mentioned in Week 2 that an intuitive but incorrect LASSO estimator only does LASSO once (*Neyman Orthogonality* not satisfied)
- One applies LASSO regression of Y_i on D_i and X_i to select relevant covariates X_Y, in addition to the covariate of interest, then refits the model using OLS of Y_i on D_i and X_Y

Why Single LASSO Fails

- The implicit *score function*

$$\psi^{naive}(W_i;\theta,g) = (Y_i - g(X_i) - \theta D_i)D_i$$

  where $g(X_i)$ captures the effect of selected controls X_Y
- *Population* moment is defined as

$$M^{naive}(\theta,g) = E[\psi^{naive}(W_i;\theta,g)]$$

- With Gateaux derivative *w.r.t.* g in direction Δ:

$$\begin{aligned}\partial_g M(\theta_0,g)[\Delta]\Big|_{g=g_0} &= -E[\Delta(X_i)D_i] \\ &= -E[E[\Delta(X_i)D_i|X_i]] \\ &= -E[\Delta(X_i)E[D_i|X_i]] \neq 0\end{aligned}$$

- *Neyman orthogonality* fails; bias in $\hat{g}$ contaminates $\hat{\theta}$ at first order

Double LASSO

- Remember Double LASSO satisfies *Neyman Orthogonality*
- With Gateaux derivative *w.r.t.* g in direction Δ:

$$\partial_g M(\theta_0, g, m_0)[\Delta]\Big|_{g=g_0} = -E[\Delta(X_i)(D_i - m_0(X_i)]$$

- Under *approximate sparsity*, LASSO can consistently approximate $m_0(X_i)$ at rate $\geq n^{1/4}$ in high dimension

$$-E[\Delta(X_i)(D_i - m_0(X_i)] =$$
$$-E[E[\Delta(X_i)(D_i - m_0(X_i)|X_i]] = 0$$

- In actual estimation, we use *plug-in* method to fine-tune *penalty level* λ to find a good approximation to the *nuisance functions* $m_0(X_i)$ (and $g_0(X_i)$)

Double Machine Learning

- Similar to Double LASSO, when we use other Machine Learning methods, we need to *fine-tune hyperparameters* (penalty level, tree depth, or neural network size) to strike the *bias-variance tradeoff* and obtain consistent estimations of the *nuisance functions*

$$m_0(X) = E[D|X], \quad g_0(X) = E[Y|X]$$

- But Double Machine Learning adds an another essential step of *cross-fitting*

Double Machine Learning: Cross-Fitting

- Instead of predicting the *nuisance functions* based on the full *sample*
- We only train nuisance models on $K-1$ folds, and predict the *residualized* Y_i and D_i on the held-out fold k
- We stack predicted *residualized* Y_i and D_i across K folds and for our FWL estimator
- to form the *score function* precisely to *prevent overfitting*—we do not want to use the training data to predict its own nuisance function
- It ensures nuisance errors are *out-of-sample*, so Neyman orthogonality cancels first-order bias

Cross-Fitting and Moment Estimation

- The above intuitive steps can be formally expressed in moment condition
- We take a K-fold random partition $(I_k)_{k=1}^K$ of observation indices $\{1, \ldots, n\}$ such that the size of each fold is about the same
- For each $k \in \{1, \ldots, K\}$, construct a fine-tuned nuisance estimator $\hat{\eta}_{[k]}$ that depends on the subset of data that excludes the k-th fold
- Now let $k(i) = \{k : i \in I_k\}$, the *sample* estimate of the moment equation is then defined as

$$\hat{M}(\theta, \hat{\eta}) = \frac{1}{n} \sum_{i=1}^{n} \psi\left(W_i; \theta, \hat{\eta}_{[k(i)]}\right)$$

- We find $\hat{\theta}$ by solving $\hat{M}(\hat{\theta}, \hat{\eta}) = 0$

Sandwich Variance Estimator

- Sandwich variance estimator is defined in the same fashion as before

$$\hat{V} = \hat{J}^{-1}\hat{\Omega}\hat{J}^{-1}$$

$$\hat{J} = \frac{1}{n}\sum_{i=1}^{n} \partial_\theta \psi(W_i; \hat{\theta}, \hat{\eta})$$

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} \psi(W_i; \hat{\theta}, \hat{\eta})\, \psi(W_i; \hat{\theta}, \hat{\eta})'$$

- This looks scary, but note that for the *score function*

$$\psi(W_i; \theta, \eta) = (\tilde{Y}_i - \theta\tilde{D}_i)\tilde{D}_i$$

$$\hat{J} = -\frac{1}{n}\sum_{i=1}^{n} \tilde{D}_i^2, \quad \hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} (\tilde{Y}_i - \hat{\theta}\tilde{D}_i)^2\tilde{D}_i^2$$

The Use of Cross-Fitting

- Remember the *score function* is defined as

$$\psi(W;\theta,\eta) = (Y_i - g(X_i) - \theta(D_i - m(X_i)))(D_i - m(X_i))$$

- Suppose we have a small estimation error in projecting g_0 and m_0

$$\hat{g}(X_i) = g_0(X_i) + \delta_g(X_i), \quad \hat{m}(X_i) = m_0(X_i) + \delta_m(X_i)$$

$$\partial_g M(\theta_0, g, m)[\Delta]\Big|_{g=g_0} = -E[\delta_g(X_i)(D_i - \hat{m}_0(X_i))] \neq 0$$

- First-order bias terms does not vanish to 0 without *cross-fitting*
- Using the whole sample to train nuisances breaks *orthogonality*

Double LASSO Does not Need Cross-Fit

- In general Double Machine Learning, *nuisance functions* are estimated by flexible ML, and in-sample predictions can *overfit*
- The nuisance errors $\delta_g(X_i), \delta_m(X_i)$ become correlated with residuals $D_i - m_0(X_i)$

Double LASSO Does not Need Cross-Fit

- In general Double Machine Learning, *nuisance functions* are estimated by flexible ML, and in-sample predictions can *overfit*
- The nuisance errors $\delta_g(X_i), \delta_m(X_i)$ become correlated with residuals $D_i - m_0(X_i)$
- *Nuisances* estimated by LASSO regression in a linear, approximately sparse setup
- Shrinkage bias is *analytically controlled* by *approximate sparsity*
- Correlation with residuals does not spoil inference

Potential Outcome Framework

Potential Outcomes Framework

- For each unit i, we fine two *latent* variables

$$\begin{aligned} &Y_i(1) && \text{(outcome if treated)} \\ &Y_i(0) && \text{(outcome if not treated)} \\ &Y_i(d) && d \in \{0,1\} \end{aligned}$$

- Individual treatment effect (ITE) is defined as

$$\tau_i = Y_i(1) - Y_i(0)$$

- The fundamental problem of causal inference is that we cannot observe both $Y_i(1)$ and $Y_i(0)$ for the same unit
- We define Average Treatment Effect (ATE) in the *population* as

$$\delta = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)]$$

Average Predictive Effect and Selection Bias

- Let $D_i \in \{0, 1\}$ denote *actual* treatment assignment
- The observed outcome is defined as

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

- *Population* data directly provide the conditional average

$$\begin{aligned} E[Y_i|D_i = 1] &= E[Y_i(1)|D_i = 1] \\ E[Y_i|D_i = 0] &= E[Y_i(0)|D_i = 0] \\ E[Y_i|D_i = d] &= E[Y_i(d)|D_i = d] \quad d \in \{0, 1\} \end{aligned}$$

Average Predictive Effect and Selection Bias

- The average predictive effect (APE) is defined as the naive difference between Y_i in the treated and control group

$$\pi = E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0]$$

- If there is a *selection bias*, APE π will not agree with the ATE δ
- Using potential outcomes, we want to decompose π

$$\begin{aligned}\pi &= E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] \\ &= E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0] \\ &= \underbrace{\left(E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid D_i = 1]\right)}_{\text{ATET}} + \\ &\quad \underbrace{\left(E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]\right)}_{\text{Selection Bias}}\end{aligned}$$

Randomized Controlled Trials (RCT)

- In a Randomized Controlled Trial (RCT), treatment is randomly assigned:

$$D_i \perp\!\!\!\perp (Y_i(0), Y_i(1)) \quad \text{or } D_i \perp\!\!\!\perp Y_i(d)$$
$$0 \leq P(D_i = 1) \leq 1$$

- The randomization of treatment assignment ensures that

$$E[Y_i \mid D_i = d] = E[Y_i(d) \mid D_i = d] = E[Y_i(d)]$$

- The *selection bias* term

$$E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0] = E[Y_i(0)] - E[Y_i(0)] = 0$$

- APE agrees with ATE

$$\pi = E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = \delta$$

Statistical Inference with Two Sample Means

- The APE is asymptotically normal in distribution
- From an RCT, we collect $\{(Y_i, D_i)\}_{i=1}^n$; we calculate the group means as

$$\hat{\theta}_d = \frac{\sum_{i=1}^n Y_i \cdot \mathbb{1}(D_i = d)}{\sum_{i=1}^n \mathbb{1}(D_i = d)}, \quad d \in \{0, 1\}$$

- APE agrees with ATE

$$\hat{\delta} = \hat{\theta}_1 - \hat{\theta}_0$$

- APE and ATE is asymptotically normal under random assignment

$$\sqrt{n}(\hat{\delta} - \delta) \xrightarrow{d} N(0, \sigma^2)$$

- If the treated and controlled observations are independent, variance is

$$\sigma^2 = \frac{\text{Var}(Y_i \mid D_i = 1)}{P(D_i = 1)} + \frac{\text{Var}(Y_i \mid D_i = 0)}{P(D_i = 0)}$$

Assumptions and Limitations of RCTs

- *Ethical issues:* cannot randomize harmful treatments
- *Practical challenges:* cost of RCTs can be high
- *External validity:* if experiment is localized, results may not generalize

Assumptions and Limitations of RCTs

- *Ethical issues:* cannot randomize harmful treatments
- *Practical challenges:* cost of RCTs can be high
- *External validity:* if experiment is localized, results may not generalize
- *The Stable Unit Treatment Value Assumption (SUTVA)*: Treatment of one unit does not change outcomes of others; no spillover effect and no interference across units

Causal Inference via Conditional Ignorability

Potential Outcome and Ignorability

- In reality, the complete random assignment assumption may be too strong

$$D_i \perp\!\!\!\perp Y_i(d)$$

- The treated and controlled units may differ in some characteristics X_i
- But with the same strata of X_i, treatments are as if randomly assigned

$$D_i \perp\!\!\!\perp Y_i(d) \mid X_i$$

- That is, suppose treatment status D_i is independent of potential outcomes $Y_i(d)$ conditional on a set of covariates X_i—*ignorability* assumption
- We also assume that there is *overlap* or *full support* in the distribution of probability of receiving treatment by X_i

$$p(X_i) := P(D_i = 1|X_i)$$
$$P(0 \leq p(X_i) \leq 1) = 1$$

Potential Outcome and Ignorability

- Conditioning on X_i removes *selection bias*

$$E[Y_i \mid D_i = d, X_i] = E[Y_i(d) \mid D_i = d, X_i] = E[Y_i(d) \mid X_i]$$

- The *selection bias* term

$$E[Y_i(0) \mid D_i = 1, X_i] - E[Y_i(0) \mid D_i = 0, X_i] = E[Y_i(0) \mid X_i] - E[Y_i(0) \mid X_i] = 0$$

- Now the *Conditional APE* (CAPE)

$$\pi(X_i) = E[Y_i \mid D_i = 1, X_i] - E[Y_i \mid D_i = 0, X_i]$$

- Agrees with the *Conditional ATE* (CATE)

$$\delta(X_i) = E[Y_i(1) \mid X_i] - E[Y_i(0) \mid X_i]$$

- Due to *Law of Iterative Expectation*

$$\delta = E[\delta(X_i)] = E[\pi(X_i)] = \pi$$

Regression Adjustment

- We can estimate $E[Y_i \mid D_i, X_i]$ by linear regression if *ignorability* and *linearity* assumptions hold
- We may specify an additive linear model and identify δ by α

$$E[Y_i \mid D_i, X_i] = \alpha D_i + X_i'\beta$$

- Here we also assume the treatment effects are homogeneous; $\delta(x) = \delta$ for all x in the support of X_i
- We can relax the homogeneity assumption by specifying an interactive model

$$E[Y_i \mid D_i, X_i] = \alpha_1 D_i + (D_i X_i)'\alpha_2 + X_i'\beta$$

- Regression adjustment gives unbiased ATE if ignorability holds

Conditioning on Propensity Scores

- Conditioning on only the propensity score also suffices to remove the *selection bias* under *ignorability* assumption
- Balancing property (Rosenbaum–Rubin)

$$D_i \perp\!\!\!\perp X_i \mid p(X_i)$$

- An important consequence is that in scenarios with a known propensity score (*e.g.*, stratified RCT), we can use $p(X_i)$ as a control in place of the high-dimensional set of characteristics X_i
- Bypass a potentially complicated high-dimensional estimation problem
- $p(X_i)$ and controls of X_i and their *transformations* can be combined to (hopefully) improve estimation precision

Horvitz–Thompson Theorem

- Under *conditional ignorability* and *overlap*, the conditional expectation of an appropriately reweighted observed outcome Y_i, given X_i, identifies the conditional average of potential outcome $Y_i(d)$ given X_i

$$
\begin{aligned}
E\left[\frac{Y_i\,\mathbb{1}(D_i = d)}{P(D_i = d \mid X_i)} \,\middle|\, X_i\right] &= E\left[\frac{Y_i(d)\,\mathbb{1}(D_i = d)}{P(D_i = d \mid X_i)} \,\middle|\, X_i\right] \\
&= \frac{E[Y_i(d)\,\mathbb{1}(D_i = d) \mid X_i]}{P(D_i = d \mid X_i)} \\
&= \frac{E[Y_i(d) \mid X_i] \cdot P(D_i = d \mid X_i)}{P(D_i = d \mid X_i)} \\
&= E[Y_i(d) \mid X_i]
\end{aligned}
$$

- Then averaging over X_i identifies average potential outcome

$$E[E[Y_i(d) \mid X_i]] = E[Y_i(d)]$$

Horvitz–Thompson Theorem

- We can therefore define a Horvitz–Thompson transformation

$$H_i = \frac{\mathbb{1}(D_i = 1)}{P(D_i = 1 \mid X_i)} - \frac{\mathbb{1}(D_i = 0)}{1 - P(D_i = 1 \mid X_i)}$$

- And identify CATE by

$$E[Y_i H_i \mid X_i] = E[Y_i(1) - Y_i(0) \mid X_i] = \delta(X_i)$$

Covariate Balance Check

- Given a propensity score $p(X_i)$, we can check if the RCT is valid by performing a *covariate balance check*
- Conditional ignorability implies

$$E[H_i|X_i] = 0$$

- To show this is the case

$$\begin{aligned} E[H_i \mid X_i] &= E\left[\frac{\mathbb{1}(D_i = 1)}{p(X_i)} \middle| X_i\right] - E\left[\frac{\mathbb{1}(D_i = 0)}{1 - p(X_i)} \middle| X_i\right] \\ &= \frac{P(D_i = 1 \mid X_i)}{p(X_i)} - \frac{P(D_i = 0 \mid X_i)}{1 - p(X_i)} \end{aligned}$$

- If we have a reasonable approximation of $p(X_i)$, the two terms above should both be close to 1 and cancel out