SOC 690S: Machine Learning in Causal Inference

Week 6: Matching, Propensity, Weighting, and Beyond

Wenhao Jiang
Department of Sociology, Fall 2025

Motivation and Roadmap

- In Week 3, we introduced the *potential outcomes framework* and discussed covariate adjustment
- We focused on the *conditional ignorability assumption* (CIA) as the key identifying condition (DAG provides a graphic evaluation of covariate adjustment)
- This week, we will revisit some of those ideas and unpack the anatomy of several adjustment methods
- We will also introduce *machine learning* tools to flexibly estimate the *propensity score* in observational data
- Some of these models are the precursors of Double Machine Learning
    - AIPW will be introduced as an alternative for double partialling out

Matching and Regression

Conditional Ignorability Assumption

- In Week 3, we introduced the *Conditional Ignorability Assumption* (CIA)

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

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

$$e(X_i) := P(D_i = 1|X_i)$$
$$P(0 < e(X_i) < 1) = 1$$

Conditional Ignorability Assumption

- We showed that

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

- Under CIA, we may address the selection bias by *conditioning* on observed covariates and define

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

Conditional Ignorability Assumption

- 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

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

- We are left with

$$\begin{aligned} \pi(X_i) &= E[Y_i(1) \mid D_i = 1, X_i] - E[Y_i(0) \mid D_i = 1, X_i] \\ &= E[Y_i(1) \mid D_i = 1, X_i] - E[Y_i(0) \mid D_i = 0, X_i] \text{ under CIA} \\ &= E[Y_i(1) \mid X_i] - E[Y_i(0) \mid X_i] = \delta(X_i) \end{aligned}$$

- The conditional ATT $\pi(X_i)$ at covariate value $X_i = x$ equals CATE at $X_i = x$

Conditional Ignorability Assumption

- There are many approaches to *condition* on the covariates X_i
- The most straightforward approaches are *matching* and *regression*

Conditional Ignorability Assumption

- There are many approaches to *condition* on the covariates X_i
- The most straightforward approaches are *matching* and *regression*
- *Matching* uses the marginal distribution of X_i in the (treated, controlled, or whole) population
- In the case of the marginal distribution of X_i in the *whole* population, ATE is defined as

$$\delta(X_i) = E[Y_i \mid D_i = 1, X_i] - E[Y_i \mid D_i = 0, X_i]$$
$$\delta_{ATE} = \sum_x \delta(X_i = x) P(X_i = x) = E[Y_i(1) - Y_i(0)]$$

- Averaging the marginal distribution of X_i in the *treated* population

$$\delta_{ATT} = \sum_x \delta(X_i = x) P(X_i = x \mid D_i = 1) = E[Y_i(1) - Y_i(0) \mid D_i = 1]$$

Anatomy of Average Treatment Effect on the Treated

- According to the Bayesian rule

$$P(X_i = x \mid D_i = 1) = \frac{P(D_i = 1 \mid X_i = x) \cdot P(X_i = x)}{P(D_i = 1)}$$

- ATT through *matching* can be written as

$$\delta_{ATT} = E[Y_i(1) - Y_i(0) \mid D_i = 1] = \sum_x \delta(X_i = x) P(X_i = x \mid D_i = 1)$$
$$= \frac{\sum_x \delta(X_i = x) \cdot P(D_i = 1 \mid X_i = x) \cdot P(X_i = x)}{\sum_x P(D_i = 1 \mid X_i = x) \cdot P(X_i = x)}$$

- So the ATT derived through *matching* uses weights to $\delta(X_i = x)$ proportional to the probability of treatment at each value of the covariates

Anatomy of Regression Adjustment

- Another common and straightforward approach *condition* on the covariates is through *regression*
- Instead of assuming full linearity in X_i, we specify a *saturated-in-X_i* model for flexibility

$$Y_i = \sum_x \mathbb{1}(X_i = x)\alpha_x + \delta_R D_i + \epsilon_i$$

- While the *matching*-based ATT uses weights proportional to the probability of treatment at each value of the covariates, regression-based estimand uses the variance of treatment at each value of the covariates

Anatomy of Regression Adjustment

- Instead of assuming full linearity in X_i, we specify a *saturated-in-X_i* model for flexibility

$$
\begin{aligned}
Y_i &= \sum_x \mathbb{1}(X_i = x)\alpha_x + \delta_R D_i + \epsilon_i \\
\delta_R &= \frac{Cov(Y_i, \tilde{D}_i)}{V(\tilde{D}_i)} = \frac{Cov(E[Y_i \mid D_i, X_i], \tilde{D}_i)}{V(\tilde{D}_i)} \\
&= \frac{E\left[\tilde{D}_i E[Y_i \mid D_i, X_i]\right]}{E(\tilde{D}_i^2)} \\
&= \frac{E\left[(D_i - E[D_i \mid X_i])E[Y_i \mid D_i, X_i]\right]}{E[(D_i - E[D_i \mid X_i])^2]}
\end{aligned}
$$

Anatomy of Regression Adjustment

- Expanding $E[Y_i \mid D_i, X_i]$ by

$$\begin{aligned} E[Y_i \mid D_i, X_i] &= E[Y_i \mid D_i = 0, X_i] + \\ &\qquad D_i \left(E[Y_i \mid D_i = 1, X_i] - E[Y_i \mid D_i = 0, X_i]\right) \\ &= E[Y_i \mid D_i = 0, X_i] + D_i \delta(X_i) \end{aligned}$$

- The numerator of the regression-based estimand becomes

$$\begin{aligned} & E\left[(D_i - E[D_i \mid X_i])\left(E[Y_i \mid D_i = 0, X_i] + D_i \delta(X_i)\right)\right] \\ &= E[(D_i - E[D_i \mid X_i])E[Y_i \mid D_i = 0, X_i]] + E[(D_i - E[D_i \mid X_i])D_i \delta(X_i)] \\ &= E[(D_i - E[D_i \mid X_i])^2 \delta(X_i)] \end{aligned}$$

Anatomy of Regression Adjustment

- The regression-based estimand is therefore given by

$$\begin{aligned}\delta_R &= \frac{E[(D_i - E[D_i \mid X_i])^2 \delta(X_i)]}{E[(D_i - E[D_i \mid X_i])^2]} \\ &= \frac{E[E[(D_i - E[D_i \mid X_i])^2 \delta(X_i) \mid X_i]]}{E[(D_i - E[D_i \mid X_i])^2 \mid X_i]]} \\ &= \frac{E[\sigma_D^2(X_i)\delta(X_i)]}{E[\sigma_D^2(X_i)]}\end{aligned}$$

$$\text{where } \sigma_D^2(X_i) \equiv E[(D_i - E[D_i \mid X_i])^2 \mid X_i]$$

- $\sigma_D^2(X_i)$ is the conditional variance of D_i given X_i
- As D_i is binary, $\sigma_D^2(X_i) = P(D_i = 1|X_i)(1 - P(D_i = 1|X_i))$

$$\delta_R = \frac{\sum_x \delta(X_i = x) e(X_i = x)(1 - e(X_i = x))P(X_i = x)}{\sum_x e(X_i = x)(1 - e(X_i = x))P(X_i = x)}$$

Anatomy of Regression Adjustment

- The anatomy of regression adjustment establishes the fact that regression puts the most weight on covariate cells where the conditional variance of treatment status is largest
- It is important to note that neither the regression nor the covariate-matching estimands give any weight to covariate cells that do not contain both treated and control observations
- Indeed, when X_i is continuous, the regression anatomy still holds

$$\delta_R = \frac{\int_x \delta(x)\, e(x)\Big(1 - e(x)\Big)\, f(x)\, dx}{\int_x e(x)\Big(1 - e(x)\Big)\, f(x)\, dx}$$

- With continuous covariates, exact common support at every $X_i = x$ is impossible in finite samples

Anatomy of Regression Adjustment with Continuous Covariates

- Instead, the positivity assumption requires that the propensity score satisfies $0 < e(x) < 1$ for all x in the support of X, so that in neighborhoods of any covariate value, both treated and untreated units occur with positive probability
- Regression with continuous covariates is a *fuzzy* matching method, where neighborhoods with more balanced *overlap* ($e(X_i) = 0.5$) receive larger weights

$$E[Y(d) \mid X_i = x] = \lim_{\epsilon \to 0} E\Big[Y(d) \mid X_i \in (x - \epsilon, x + \epsilon)\Big]$$

- Without common support, regression identifies "treatment effects" only by extending the fitted control line into the treated region
- This linear extrapolation recovers the true effect only if treatment effects are constant across X_i

Effects of Voluntary Military Service on Earnings

Table: Uncontrolled, matching, and regression estimates of the effects of voluntary military service on earnings

	Average earnings 1988–1991	Differences in means by veteran	Matching estimates	Regression estimates	Regression minus matching
Whites	14537	1233.4	-197.2	-88.8	108.4
		(60.3)	(70.5)	(62.5)	(28.5)
Non-whites	11664	2449.1	839.7	1074.4	234.7
		(47.4)	(62.7)	(50.7)	(32.5)

Notes: Adapted from Angrist (1998, Tables II and V). Standard errors are reported in parentheses. The table shows estimates of the effect of voluntary military service on the 1988–1991 Social Security–taxable earnings of men who applied to enter the armed forces between 1979 and 1982. The matching and regression estimates control for applicants' year of birth, education at the time of application, and AFQT score. There are 128,968 whites and 175,262 non-whites in the sample.

The Propensity Score Theorem

- *The propensity score theorem*: suppose the CIA holds such that $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i | X_i$, then $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i | e(X_i)$
- To show this is the case, note the claim is true if

$$P[D_i = 1 \mid Y_i(d), e(X_i)] \text{ does not depend on } Y_i(d).$$

$$\begin{aligned} P[D_i = 1 \mid Y_i(d), e(X_i)] &= E[D_i \mid Y_i(d), e(X_i)] \\ &= E\{E[D_i \mid Y_i(d), e(X_i), X_i] \mid Y_i(d), e(X_i)\} \\ &= E\{E[D_i \mid Y_i(d), X_i] \mid Y_i(d), e(X_i)\} \\ &= E\{E[D_i \mid X_i] \mid Y_i(d), e(X_i)\}, \quad \text{by the CIA} \end{aligned}$$

- Note that

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

- which is just $e(X_i)$

The Propensity Score Theorem

- By the propensity score theorem and CIA

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

Horvitz–Thompson Extension

- We introduced in Week 3 that propensity score can also be used to construct ATE under CIA and *common support*

$$E[Y_i(1) - Y_i(0)] = E\left[\frac{Y_i D_i}{e(X_i)} - \frac{Y_i(1-D_i)}{1-e(X_i)}\right]$$
$$= E\left[\frac{(D_i - e(X_i))Y_i}{e(X_i)(1-e(X_i)}\right]$$

- To see this, iterate over X_i

$$E\left[\frac{Y_i D_i}{e(X_i)}\right] = E\left[E\left[\frac{Y_i D_i}{e(X_i)}\middle| X_i\right]\right]$$
$$= E\left[\frac{E[Y_i \mid D_i = 1, X_i]e(X_i)}{e(X_i)}\right]$$
$$= E\left[E[Y_i \mid D_i = 1, X_i]\right] = E[Y_i(1) \mid X_i]$$

Horvitz–Thompson Extension

- This is the method of *Inverse Propensity Weighting* (IPW), given we have a reliable estimation of $e(X_i)$
- We can similarly define the estimand of ATT using IPW

$$E[Y_i(1) - Y_i(0) \mid D_i = 1] = E\left[\frac{(D_i - e(X_i))Y_i}{(1 - e(X_i))P(D_i = 1)}\right]$$

- Consider again the regression estimand

$$\begin{aligned}\delta_R &= \frac{E[(D_i - e(X_i))Y_i]}{E[e(X_i)(1 - e(X_i))]} \\ &= E\left[\frac{e(X_i)(1 - e(X_i))}{E[e(X_i)(1 - e(X_i))]}\left[\frac{Y_i D_i}{e(X_i)} - \frac{Y_i(1 - D_i)}{1 - e(X_i)}\right]\right]\end{aligned}$$

Horvitz–Thompson Extension

- The IPW estimand for ATE, ATT, and the regression estimand are in the *same* class of weighted average estimands (Hirano, Imbens, and Ridder 2003)

$$E\left[g(X_i)\left[\frac{Y_i D_i}{e(X_i)} - \frac{Y_i(1-D_i)}{1-e(X_i)}\right]\right]$$

- For ATE, set $g(X_i) = 1$
- For ATT, set $g(X_i) = \frac{e(X_i)}{P(D_1=1)}$
- For regression, set

$$g(X_i) = \frac{e(X_i)(1-e(X_i))}{E[e(X_i)(1-e(X_i))]}$$

Stabilized Inverse Probability Weighting (SIPW)

- IPW for ATE is defined as

$$E[Y(1) - Y(0)] = E\left[\frac{D_i Y_i}{e(X_i)} - \frac{(1 - D_i)Y_i}{1 - e(X_i)}\right]$$

- However, if $e(X_i) \approx 0$ or 1, weights explode and the estimand is unstable
- Stabilized IPW *stabilizes* IPW (Hájek estimator) by multiplying weights by marginal treatment probabilities to reduce variance

$$w_i^{SIPW} = \begin{cases} \dfrac{P(D_i = 1)}{e(X_i)} & D_i = 1 \\ \dfrac{P(D_i = 0)}{1 - e(X_i)} & D_i = 0 \end{cases}$$

Stabilized Inverse Probability Weighting (SIPW)

- For $E[Y(1)] = E\left[\frac{D_i Y_i}{e(X_i)}\right]$, multiply the numerator and denominator by $w_i^{SIPW} = \dfrac{P(D_i = 1)}{e(X_i)}$

$$E[Y(1)] = \frac{E\left[\dfrac{P(D_i=1)}{e(X_i)} D_i Y_i\right]}{E\left[\dfrac{P(D_i=1)}{e(X_i)} D_i\right]} = \frac{E\left[\dfrac{1}{e(X_i)} D_i Y_i\right]}{E\left[\dfrac{1}{e(X_i)} D_i\right]} = \frac{E\left[\dfrac{1}{e(X_i)} D_i Y_i\right]}{E\left[E\left[\dfrac{1}{e(X_i)} D_i \middle| X_i\right]\right]}$$

$$= E\left[\frac{D_i Y_i}{e(X_i)}\right]$$

- The sample estimator is therefore given by

$$\hat{\delta}_{SIPW} = \frac{\sum_i w_i^{SIPW} D_i Y_i}{\sum_i w_i^{SIPW} D_i} - \frac{\sum_i w_i^{SIPW} (1 - D_i) Y_i}{\sum_i w_i^{SIPW} (1 - D_i)}$$

Why Stabilized IPW Reduces Variance

- SIPW uses stabilized weights multiply raw IPW weights by the marginal treatment probabilities, ensuring that

$$E[w^{SIPW} \mid D = 1] = E[w^{SIPW} \mid D = 0] = 1$$

- After stabilization, the total weight of each group matches its actual sample size, so the weighted pseudo-population is not artificially inflated
- By centering the weights around 1, the spread of weights (and thus sampling variance of the estimator) is reduced, though extreme weights from poor overlap can still occur

Weighted Linear Regression and IPW

- The IPW has the same ATE estimand as the Weighted Linear Regression of Y_i on D_i with IPW weights (or Horvitz–Thompson transformation)

$$H_i = \frac{D_i}{e(X_i)} - \frac{1 - D_i}{1 - e(X_i)}$$

- See the proof

G-computation (G-formula)

- Besides using propensity score under CIA, G-computation leverages CIA and *common support* to construct counterfactuals for Y_i conditioning on X_i
- Under CIA and *common support*, G-computation models conditional outcomes

$$m_d(x) = E[Y_i \mid D_i = d, X_i = x], \quad d \in \{0, 1\}$$

- The counterfactuals for each unit may be projected by

$$\hat{Y}_i(1) = \hat{m}_1(X_i), \quad \hat{Y}_i(0) = \hat{m}_0(X_i)$$

- Average predictions to estimate causal effects

$$\hat{\delta}_{ATE} = \frac{1}{n} \sum_{i=1}^{n} \Big(\hat{m}_1(X_i) - \hat{m}_0(X_i) \Big)$$

G-computation (G-formula)

- The estimation relies on correct specification of outcome models $m_d(x)$ given D_i and X_i
- Similar to regression-based estimand above, this is sensitive to extrapolation if overlap is weak

An Overview of Doubly Robust Estimation

- The two classical strategies, IPW and G-computation, are respectively consistent only if its *nuisance* models $e(X_i)$ and $m(X_i)$ are correctly specified
- *Doubly Robust* estimand combines both models so that
  - estimator is consistent if *either* $e(X_i)$ or $m(X_i)$ is correct,
  - efficient (minimum variance) if both are correct
- The canonical example estimand is the *Augmented Inverse Probability Weighting* (AIPW)

Augmented Inverse Probability Weighting (AIPW)

- Given the two nuisance functions

$$m_d(x) = E[Y_i \mid D_i = d, X_i], \quad e(X_i) = P(D_i = 1 \mid X_i)$$

- The ATE estimand θ is given by

$$\theta = E\left[m_1(X_i) - m_0(X_i) + \frac{D_i}{e(X_i)}\Big(Y_i - m_1(X_i)\Big) - \frac{1 - D_i}{1 - e(X_i)}\Big(Y_i - m_0(X_i)\Big)\right]$$

- When the outcome regression $m_d(X_i)$ is correctly specified

$$\begin{aligned} E\left[\frac{D_i}{e(X_i)}(Y_i - m_1(X_i))\middle| X_i\right] &= \frac{1}{e(X_i)} P[D_i = 1 \mid X_i]\, E[Y_i - m_1(X_i) \mid D_i = 1, X_i] \\ &= \frac{1}{e(X_i)} E[D_i \mid X_i]\, E[Y_i - m_1(X_i) \mid D_i = 1, X_i] \\ &= 0 \end{aligned}$$

Augmented Inverse Probability Weighting (AIPW)

- Similarly,
$$E\left[\frac{1-D_i}{1-e(X_i)}(Y_i - m_0(X_i)) \mid X_i\right] = 0.$$
- Hence,
$$\theta = E[m_1(X_i) - m_0(X_i)] = E[Y_i(1) - Y_i(0)]$$
- This holds even if the propensity score nuisance function is *incorrectly* specified

Augmented Inverse Probability Weighting (AIPW)

- When the propensity score $e(X_i)$ is correctly specified

$$E\left[\frac{D_i}{e(X_i)}(Y_i - m_1(X_i))\middle| X_i\right] = \frac{1}{e(X_i)}E[D_iY_i \mid X_i] - m_1(X_i)$$

- Since $E[D_iY_i \mid X_i] = e(X_i)E[Y_i(1) \mid X_i]$, this becomes

$$E[Y_i(1) \mid X_i] - m_1(X_i)$$

- Similarly,

$$E\left[-\frac{1-D_i}{1-e(X_i)}(Y_i - m_0(X_i)) \mid X_i\right] = E[Y_i(0) \mid X_i] - m_0(X_i)$$

- Adding up,

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

Neyman Orthogonality of AIPW

- AIPW *score function* (influence function)

$$\psi(W_i;\theta,m,e) = m_1(X_i) - m_0(X_i) + \frac{D_i}{e(X_i)}\Big(Y_i - m_1(X_i)\Big) - \frac{1-D_i}{1-e(X_i)}\Big(Y_i - m_0(X_i)\Big)$$

- The moment condition is defined as

$$M(\theta,m,e) = E[\psi(W_i;\theta,m,e)]$$

- And $\theta_0 = E[Y_i(1) - Y_i(0)]$ is found by the *normal equation*

$$M(\theta_0,m,e) = 0$$

- Neyman orthogonality condition requires

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

- where $\eta = (m,e)$ are nuisance functions

Neyman Orthogonality of AIPW

- To show AIPW satisfies Neyman Orthogonality, define a perturb propensity score

$$e_t(X_i) = e_0(X_i) + t \cdot \Delta(X_i), \quad t \to 0$$

- The Gateaux derivative is defined as

$$\partial_e M(\theta_0, m_0, e)[\Delta] \Big|_{e=e_0} = \lim_{t \to 0} \frac{M(\theta_0, m_0, e_t) - M(\theta_0, m_0, e_0)}{t}$$

Neyman Orthogonality of AIPW

- Focus on $e(X_i)$ terms in the score:

$$\psi_e(W_i; e_t) = \frac{D_i}{e_t(X_i)}(Y_i - m_1(X_i)) - \frac{1 - D_i}{1 - e_t(X_i)}(Y_i - m_0(X_i))$$

- Expand $e_t(X_i) = e_0(X_i) + t\Delta(X_i)$ in the denominator using Taylor expansion

$$\frac{1}{e_0 + t\Delta} = \frac{1}{e_0} - \frac{t\Delta}{e_0^2} + o(t)$$

$$\frac{1}{1 - e_0 - t\Delta} = \frac{1}{1 - e_0} + \frac{t\Delta}{(1 - e_0)^2} + o(t)$$

$$\psi_e(W_i; e_t) = \psi_e(W_i; e_0) - t\frac{D_i\Delta(X_i)}{e_0(X_i)^2}(Y_i - m_1(X_i)) + t\frac{(1 - D_i)\Delta(X_i)}{(1 - e_0(X_i))^2}(Y_i - m_0(X_i)) + o(t)$$

Neyman Orthogonality of AIPW

- The difference in moments defined in the Gateaux derivative is

$$M(\theta_0, m_0, e_t) - M(\theta_0, m_0, e_0) = -tE\left[\frac{D_i\Delta(X_i)}{e_0(X_i)^2}(Y_i - m_1(X_i))\right] + tE\left[\frac{(1-D_i)\Delta(X_i)}{(1-e_0(X_i))^2}(Y_i - m_0(X_i))\right] + o(t)$$

- Divide by t and take limit $t \to 0$

$$\partial_e M(\theta_0, m_0, e)[\Delta]\Big|_{e=e_0} = -E\left[\frac{D_i\Delta(X_i)}{e_0(X_i)^2}(Y_i - m_1(X_i))\right] + E\left[\frac{(1-D_i)\Delta(X_i)}{(1-e_0(X_i))^2}(Y_i - m_0(X_i))\right]$$

Neyman Orthogonality of AIPW

- Taking the iterative expectation

$$E\left[E\left[\frac{D_i\Delta(X_i)}{e_0(X_i)^2}(Y_i - m_1(X_i))\middle|X_i\right]\right] = E\left[\frac{\Delta(X_i)}{e_0(X_i)^2}E\left[D_i(Y_i - m_1(X_i))|X_i\right]\right]$$

$$E\left[D_i(Y_i - m_1(X_i))|X_i\right] = e_0(X_i)(E[Y_i \mid D_i = 1, X_i] - m_1(X_i)) = 0$$

- It is straightforward to show the whole derivative $= 0$
- The DML can be introduced

$$\hat{\theta}_{DML} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}_1^{(-k)}(X_i) - \hat{m}_0^{(-k)}(X_i) + \frac{D_i}{\hat{e}^{(-k)}(X_i)}\left(Y_i - \hat{m}_1^{(-k)}(X_i)\right) - \frac{1 - D_i}{1 - \hat{e}^{(-k)}(X_i)}\left(Y_i - \hat{m}_0^{(-k)}(X_i)\right)\right]$$