Brief Introduction to Double/debiased Machine Learning

Xingche Guo

Department of Statistics
University of Connecticut

AI Study Group, November 2024

Introduction

- This presentation is based on paper: *Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1):C1–C68.* by **Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, James Robins (2018).**
- The ML methods seem remarkably effective in prediction contexts.
- However, good performance in prediction does not necessarily translate into good performance for estimation or inference about 'causal' parameters.
- Example: deep learning models use big data to predict diseases with high accuracy. However, they don't reveal causal relationships, like which lifestyle or treatment changes can improve outcomes.

Main points via a partially linear model

Consider the following partially linear regression (PLR) model as in Robinson (1988) [1]:

$$Y = D\theta_0 + g_0(X) + U, \quad E[U \mid X, D] = 0 \tag{1}$$

$$D = m_0(X) + V, \quad E[V \mid X] = 0 \tag{2}$$

- Y - outcome variable
- D - policy/treatment variable
- θ_0 is the target parameter of interest
- X is a high-dimensional vector of other covariates, called 'controls' or 'confounders'.
- $m_0 \neq 0$ is typically the case in observational studies.

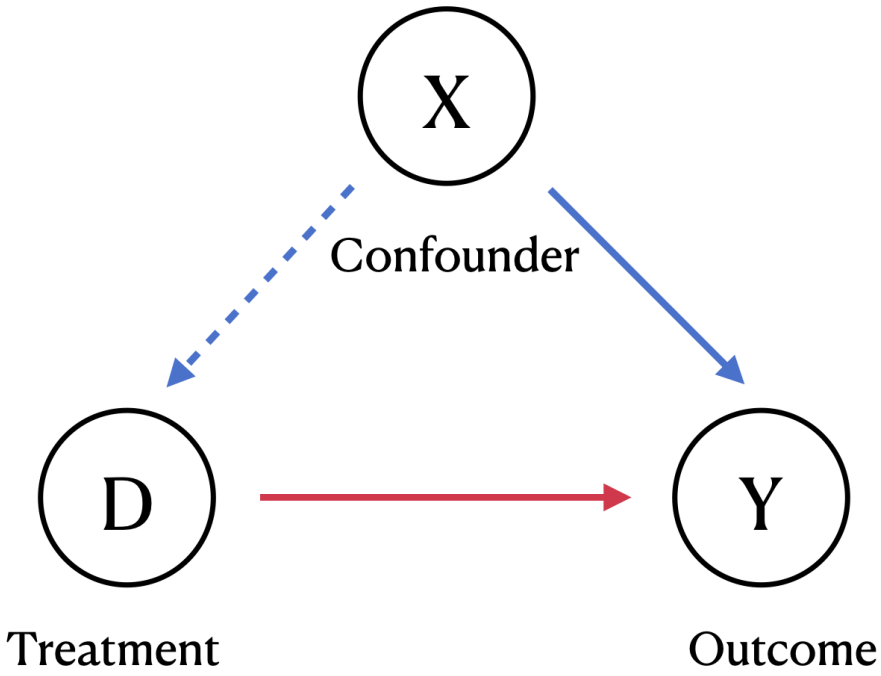
[1] Robinson, P. M. (1988). Root-N-consistent semi-parametric regression. Econometrica 56, 931–54.

Main points via a partially linear model

Consider the following partially linear regression (PLR) model as in Robinson (1988):

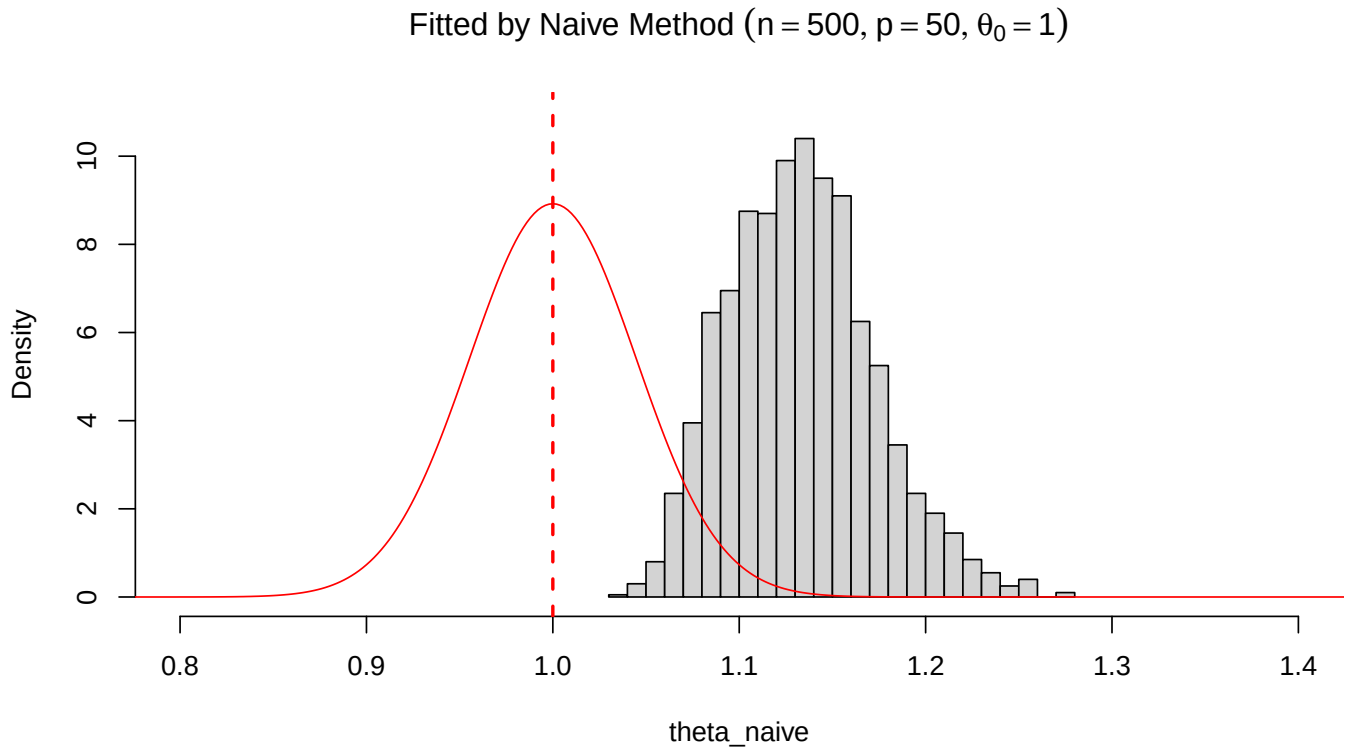
$$Y = D\theta_0 + g_0(X) + U, \quad E[U \mid X, D] = 0 \tag{3}$$

$$D = m_0(X) + V, \quad E[V \mid X] = 0 \tag{4}$$



‘Naive’ or prediction-based ML approach is bad

- Predict Y using D and X and obtain $D\hat{\theta}_0 + \hat{g}_0(X)$
- Excellent prediction performance. BUT the distribution of $\hat{\theta}_0$ looks like this ($\theta_0 = 1$):



- $D = \sum_{j=1}^{p} X_j \gamma_{0j} + N(0, \sigma_v^2)$, $Y = D\theta_0 + \sum_{j=1}^{p} X_j \beta_{0j} + N(0, \sigma_u^2)$.

Why 'naive' approach is bad

Suppose we use (1) with an estimator $\hat{g}_0(X)$ to estimate θ_0:

$$\widehat{\theta}_0 = \left(\frac{1}{n} \sum_{i=1}^{n} D_i^2 \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} D_i \left(Y_i - \widehat{g}_0 \left(X_i \right) \right).$$

$$\sqrt{n} \left(\widehat{\theta}_0 - \theta_0 \right) = \left(\frac{1}{n} \sum_{i=1}^{n} D_i^2 \right)^{-1} \Bigg[\underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_i U_i}_{:=a} + \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} m_0(X_i) \left(g_0 \left(X_i \right) - \widehat{g}_0 \left(X_i \right) \right)}_{:=b}$$
$$+ \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i \left(g_0 \left(X_i \right) - \widehat{g}_0 \left(X_i \right) \right)}_{:=c} \Bigg].$$

Question: $\sqrt{n} \left(\widehat{\theta}_0 - \theta_0 \right) \stackrel{?}{\to} N(0, \sigma^2)$.

Suffer from regularization bias

Have a closer look at b:

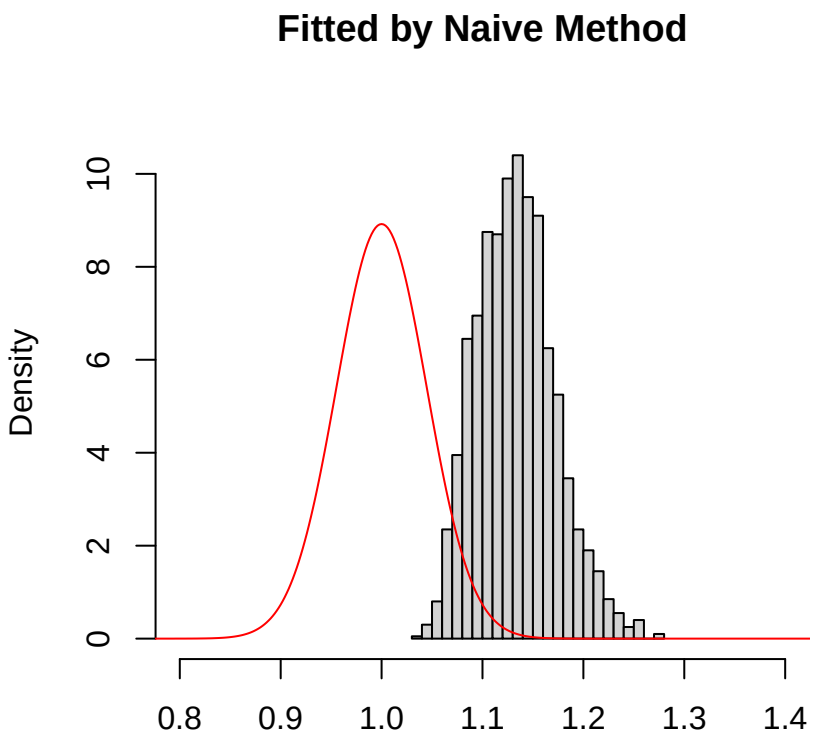
$$b = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m_0(X_i) \left(g_0(X_i) - \widehat{g}_0(X_i) \right).$$

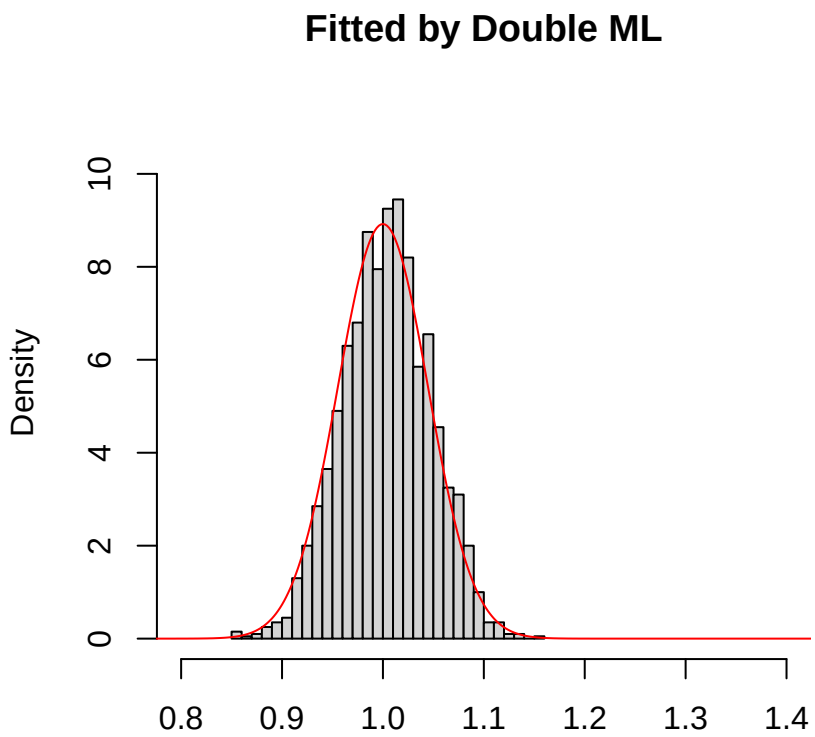
In the (nonparametric/high-dimensional/machine learning) setting, there is typically regularization bias to avoid overfitting (e.g. smoothing in splines, L_1 penalty for LASSO, drop-out in NN):

- The error $g_0(X_i) - \widehat{g}_0(X_i)$ is of order $n^{-\tau}$ for $0 < \tau < \frac{1}{2}$.
- b will then look like $\sqrt{n}\, n^{-\tau} \to \infty$.
- The 'naive' estimator $\hat{\theta}_0$ is not root-n consistent.

‘Double’ ML approach is good

1. Predict Y using X by $\widehat{E[Y|X]}$ by the ‘best performing ML’ tools, then residualize $\widehat{U} = Y - \widehat{E[Y|X]}$.
2. Predict D using X by $\widehat{E[D|X]}$ by the ‘best performing ML’ tools, then residualize $\widehat{V} = D - \widehat{E[D|X]}$.
3. Regress $\widehat{U}$ on $\widehat{V}$ to get $\check{\theta}_0$.





Why 'double' ML approach is good

Define $l_0(X) = E[Y \mid X] = m_0(X)\theta_0 + g_0(X)$:

$$\check{\theta}_0 = \left(\frac{1}{n} \sum_{i=1}^{n} \widehat{V}_i^2 \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \widehat{V}_i \widehat{U}_i,$$

where $\widehat{V} = D - \hat{m}_0(X)$, $\widehat{U} = Y - \widehat{l}_0(X)$.

$$\sqrt{n}\left(\check{\theta}_0 - \theta_0\right) = \left(\frac{1}{n} \sum_{i=1}^{n} \widehat{V}_i^2 \right)^{-1} \Bigg[\underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i U_i}_{:=a^*} + \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(m_0(X_i) - \hat{m}_0(X_i)\right) \left(l_0\left(X_i\right) - \widehat{l}_0\left(X_i\right)\right)}_{:=b^*}$$
$$+ \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ V_i \left(l_0\left(X_i\right) - \widehat{l}_0\left(X_i\right)\right) + U_i \left(m_0\left(X_i\right) - \widehat{m}_0\left(X_i\right)\right) \right\}}_{:=c^*} \Bigg] + o_p(1).$$

Question: $\sqrt{n}\left(\check{\theta}_0 - \theta_0\right) \stackrel{?}{\to} N(0, \sigma^2)$.

Why 'double' ML approach is good

Have a closer look at a^*:

$$a^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i U_i \to N(0, \sigma_a^2).$$

Have a closer look at b^*:

$$b^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (m_0(X_i) - \hat{m}_0(X_i)) \left(l_0(X_i) - \hat{l}_0(X_i) \right).$$

- The error $l_0(X_i) - \hat{l}_0(X_i)$ is of order $n^{-\tau}$ for $0 < \tau < \frac{1}{2}$.
- The error $m_0(X_i) - \hat{m}_0(X_i)$ is of order $n^{-\phi}$ for $0 < \phi < \frac{1}{2}$.
- b^* will then look like $n^{\frac{1}{2}-\tau-\phi}$.
- $o(n^{-1/4})$ is often attainable rate for estimating $l_0(X)$ and $m_0(X)$, in this case, $b^* = o_p(1)$.

Key difference between b and b^* is Neyman Orthogonality

Consider the score function for the naive method and Double ML:

$$\Phi(\theta, g) = D\big(Y - D\theta - g(X)\big),$$
$$\Psi(\theta, \eta) = \big(D - m(X)\big)\Big(Y - l(X) - \big(D - m(X)\big)\theta\Big), \quad \eta = (l, m).$$

The moment condition holds for both methods:

$$E\Phi(\theta_0, g_0) = 0, \quad E\Psi(\theta_0, \eta_0) = 0.$$

The Neyman Orthogonality condition only holds for the Double ML:

$$\lim_{\rho \to 0} \frac{\partial}{\partial \rho} E[\Phi(\theta_0, g_0 + \rho(g - g_0))] = -E(m_0(X)(g(X) - g_0(X))) \neq 0,$$
$$\lim_{\rho \to 0} \frac{\partial}{\partial \rho} E[\Psi(\theta_0, \eta_0 + \rho(\eta - \eta_0))] = 0.$$

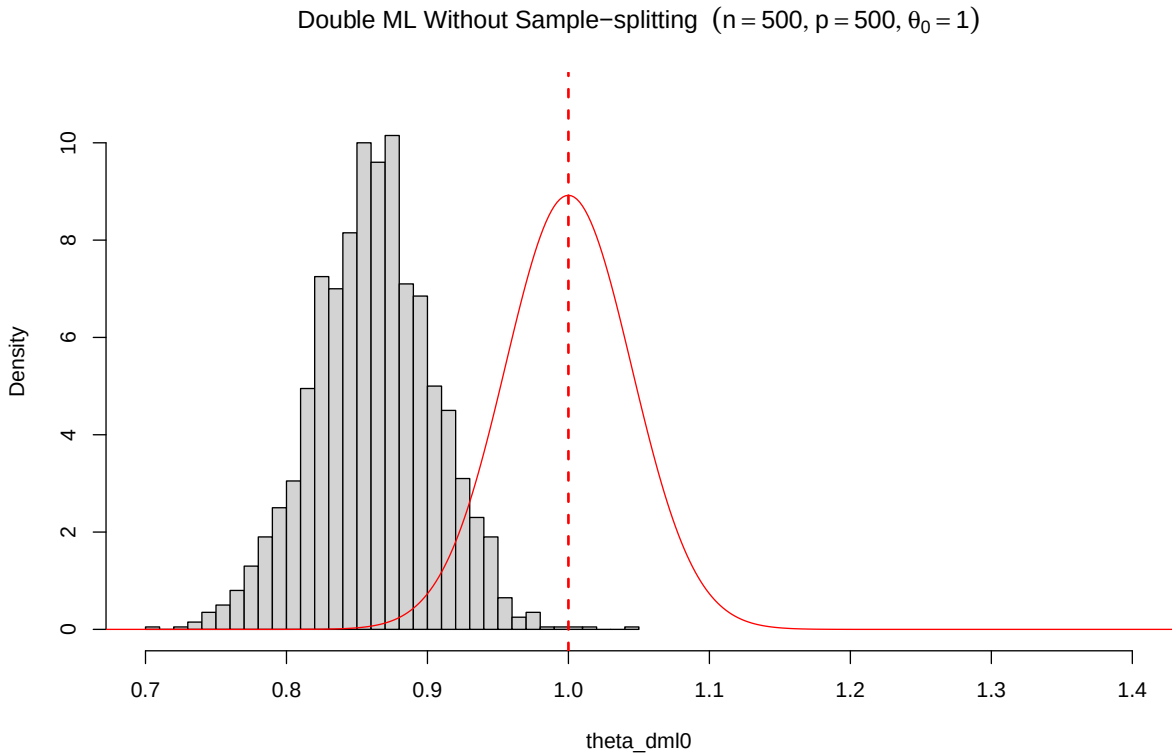
For DML: small deviations in nuisance functions do not invalidate moment conditions.

Suffer from overfitting

Have a closer look at c^*:

$$c^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \left(m_0(X_i) - \hat{m}_0(X_i) \right) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i \left(l_0 \left(X_i \right) - \widehat{l_0} \left(X_i \right) \right).$$

When $\hat{m}_0(X)$ or $\widehat{l_0}(X)$ overfits (i.e., incorrectly models the noise in the data), then c^* may not be bounded.



Suffer from overfitting

Have a closer look at c^*:

$$c^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \left(m_0(X_i) - \hat{m}_0(X_i) \right) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i \left(l_0(X_i) - \widehat{l}_0(X_i) \right).$$

Suppose $\widehat{l}_0(X_i)$ overfits (i.e., model the noise):

$$\widehat{l}_0(X_i) = l_0(X_i) + \frac{Y_i - l_0(X_i)}{n^{\frac{1}{2}-\epsilon}} = l_0(X_i) + \frac{U_i + \theta_0 V_i}{n^{\frac{1}{2}-\epsilon}}.$$

Then

$$\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i \left(l_0(X_i) - \widehat{l}_0(X_i) \right) \right| \propto n^{\epsilon} \to \infty.$$

Suffer from overfitting

Have a closer look at c^*:

$$c^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i \left(m_0(X_i) - \hat{m}_0(X_i) \right) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i \left(l_0 \left(X_i \right) - \widehat{l_0} \left(X_i \right) \right).$$

The second term of c^* is bounded by

$$C_2^* = \sup_{l \in \mathcal{F}_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i \left(l_0 \left(X_i \right) - l \left(X_i \right) \right) \right|,$$

where $\mathcal{F}_n$ is the class of functions that contains estimators of l_0.

- In conventional semi-parametric statistical setup, the complexity of $\mathcal{F}_n$ is controlled by Donsker conditions. In this case $C_2^* = o_p(1)$.
- In modern nonparametric/high-dimensional/ML setup, the Donsker conditions are inappropriate. In this case C_2^* may diverge.

Why sample splitting?

Apply sample splitting to mitigate bias induced by overfitting:

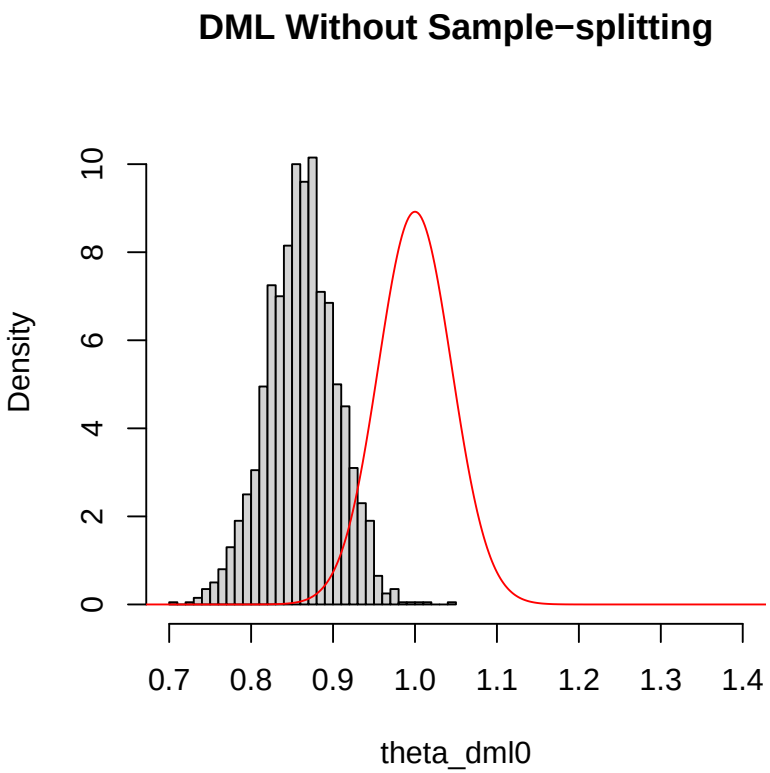
1. Split the data into two index sets, I and I^c, with size n_1 and n_2, $n = n_1 + n_2$.
2. Fit the ML estimator $\widehat{l}_0^{(-I)}(X)$ and $\widehat{m}_0^{(-I)}(X)$ using the data in I^c.
3. Fit the parameter of interest $\check{\theta}_0$ using the data in I.

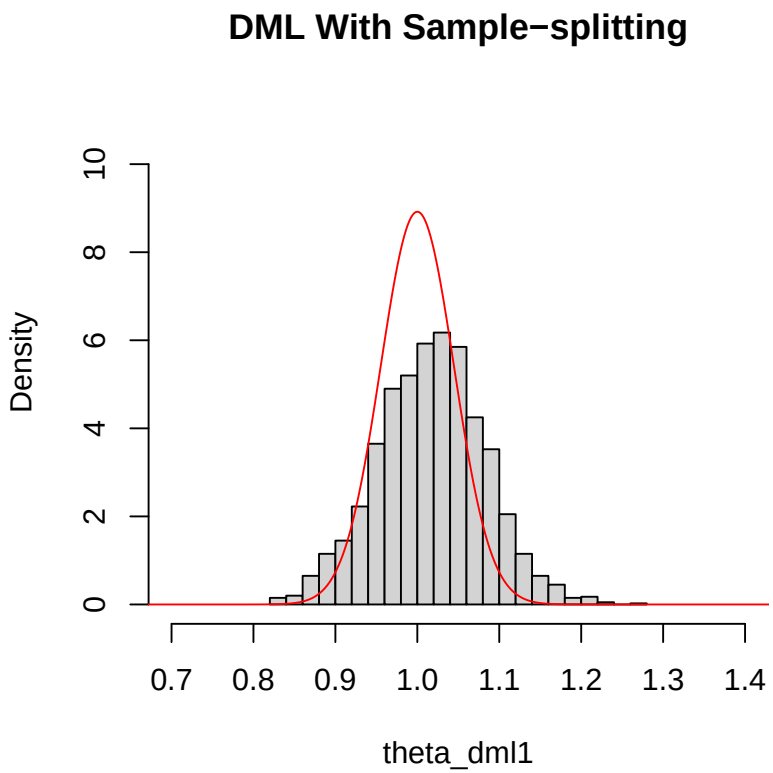
Then we can rewrite c^*:

$$c^* = \frac{1}{\sqrt{n_1}} \sum_{i \in I} U_i \left(m_0(X_i) - \widehat{m}_0^{(-I)}(X_i) \right) + \frac{1}{\sqrt{n_1}} \sum_{i \in I} V_i \left(l_0(X_i) - \widehat{l}_0^{(-I)}(X_i) \right).$$

We can tell that $U_i \perp \widehat{m}_0^{(-I)}(X_i)$, $V_i \perp \widehat{l}_0^{(-I)}(X_i)$. As a result, $c^* = o_p(1)$. By sampling splitting, we do not require Donsker condition any more for the ML model classes.

Why sample splitting?





Attaining full efficiency

- 2-way split: we evenly split data into (I, I^c) and obtain estimates $\check{\theta}_0(I, I^c)$ and $\check{\theta}_0(I^c, I)$, then we average them

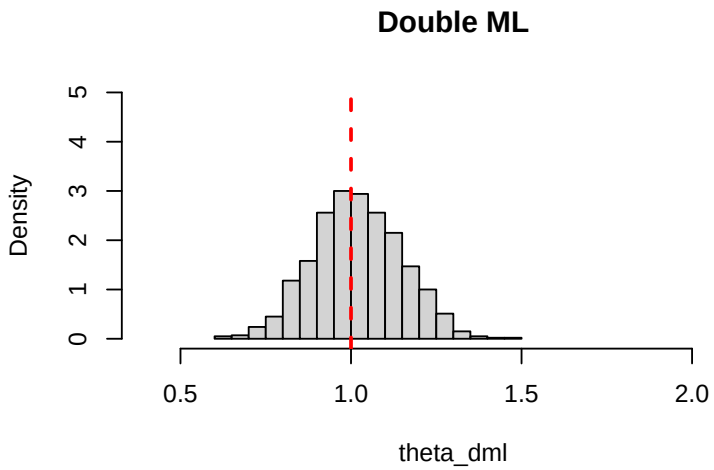
$$\check{\theta}_0 = \frac{1}{2}\left\{\check{\theta}_0(I, I^c) + \check{\theta}_0(I, I^c)\right\}.$$

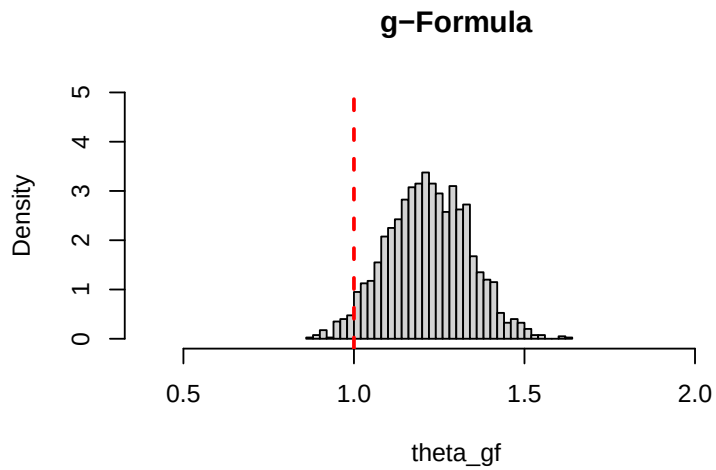
- K-way split: we evenly split data into $(I_1, \ldots, I_K)$ and obtain estimates $\check{\theta}_0(I_k, I_k^c)$ $(k = 1, \ldots, K)$, then we average them

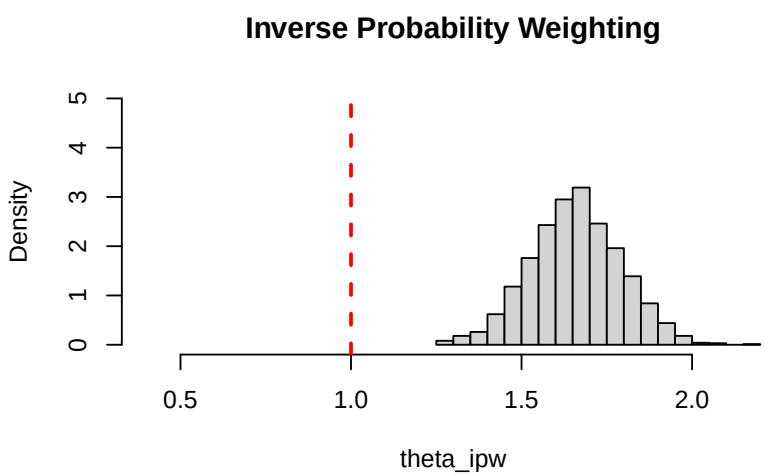
$$\check{\theta}_0 = \frac{1}{K}\sum_{k=1}^{K} \check{\theta}_0(I_k, I_k^c).$$

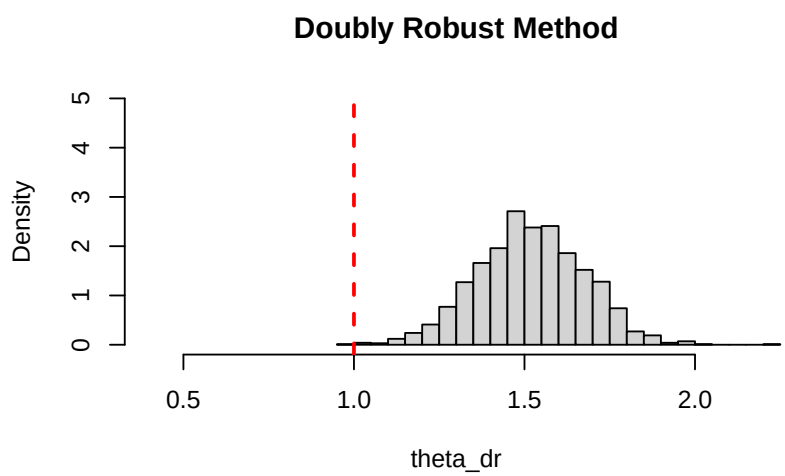
Double ML for estimating ATE

- $D = \text{Bern}(\text{logistic}(\sum_{j=1}^{p} X_j \gamma_{0j}))$, $Y = D\theta_0 + \sum_{j=1}^{p} X_j \beta_{0j} + N(0, \sigma_u^2)$.









Constructing score functions satisfy neyman orthogonality

Suppose loss function is given by $\ell(W, \theta, \beta)$.

- θ: d-dimensional parameter of interest.
- β: p_0-dimensional nuisance parameter.

Under regularity, true parameter (θ_0, β_0) values satisfy:

$$\mathrm{E}\left[\partial_\theta \ell\left(W, \theta_0, \beta_0\right)\right] = 0, \quad \mathrm{E}\left[\partial_\beta \ell\left(W, \theta_0, \beta_0\right)\right] = 0$$

The original score $\partial_\theta \ell\left(W, \theta, \beta\right)$ in general does not possess the Neyman Orthogonality property.

Solution: construct new estimating equation with desired orthogonality property:

$$\psi(W, \theta, \eta) = \partial_\theta \ell(W, \theta, \beta) - \mu \partial_\beta \ell(W, \theta, \beta), \quad \eta = (\beta^\top, \mathrm{vec}(\mu)^\top)^\top,$$

where $\mu_0 = J_{\theta\beta}^{-1} J_{\beta\beta}$ with

$$J_{\theta\beta} = \partial_\beta \mathrm{E}\left[\partial_\theta \ell\left(W, \theta, \beta\right)\right]\Big|_{\theta=\theta_0, \beta=\beta_0}, \quad J_{\beta\beta} = \partial_\beta \mathrm{E}\left[\partial_\beta \ell\left(W, \theta, \beta\right)\right]\Big|_{\beta=\beta_0}.$$

Main Theoretical Result

Under conditions for

1. Parameter not on boundary.
2. Identiability.
3. Twice continuously differentiability
4. Mild smoothness.
5. Neyman orthogonality.
6. Sample splitting.
7. Rate for ML estimator.

$$\sqrt{n}\left(\check{\theta}_0 - \theta_0\right) \xrightarrow{d} N(0, \sigma^2).$$

Take home message

1. Naive ML methods fail to provide reliable inference about causal parameters. Similarly, causal models may struggle when dealing with high-dimensional confounders.
2. By applying Double ML, we can construct high-quality point and interval estimates for causal parameters.
3. **Key Ingredient I:** Neyman Orthogonality Condition — Debias from regularization bias.
4. **Key Ingredient II:** Sample Splitting — Prevent overfitting.