Double Machine Learning for Causal Inference

He Li

New York University

Stern School of Business

September 20, 2018

Overview

Partial Linear Regression

Consider the following partial linear regression model:

$$
\begin{aligned}
Y &= D\theta_0 + g_0(X) + U, \quad \mathbb{E}[U|X, D] = 0 \\
D &= m_0(X) + V, \qquad\qquad \mathbb{E}[V|X] = 0
\end{aligned}
$$

Here, Y is the outcome variable, D is the treatment, $X \in \mathbb{R}^p$ is the control variable and U, V are noise term.

We are interested in estimating treatment effect parameter θ_0 and we need to estimate the nuisance parameter $\eta_0 = (m_0, g_0)$ in the same time.

Regularization Bias

A naive way to estimate θ_0 is as follows.

1. split data into two index set, I, I^c
2. Using some sophisticated machine learning algorithm to estimate g_0 as $\hat{g}_0$ on dataset I^c
3. Using $\hat{g}_0$ and dataset I to estimate θ_0 (plug-in regression)

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

Regularization Bias

However, this estimator $\hat{\theta}_0$ has a slower convergence rate, namely,

$$\sqrt{n}\left(\hat{\theta}_0 - \theta_0\right) = \left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} D_i U_i + \left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} D_i \left(g_0\left(X_i\right) - \hat{g}_0\left(X_i\right)\right)$$

where the first part on the RHS converges to $N(0,\overline{\Sigma})$ but the second term diverges in high-dimensional cases.

$$\left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} D_i \left(g_0\left(X_i\right) - \hat{g}_0\left(X_i\right)\right) = \left(E\left[D_i^2\right]\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} m_0\left(X_i\right)\left(g_0\left(X_i\right) - \hat{g}_0\left(X_i\right)\right) + o_P(1)$$

Regularization Bias

We will introduce two technique, Neyman Orthogonality and Cross-fitting from [2] to overcome the problem.

Overview

Definition

For some low dimensional parameter $\theta \in \Theta \subset \mathbb{R}^{d_0}$ with true value θ_0, we first assume θ_0 satisfies the moment conditions.

$$\mathbb{E}_P\left[\psi\left(w; \theta_0, \eta_0\right)\right] = 0 \tag{2.1}$$

where w is some random variables in a measurable space $\mathcal{W}, \mathcal{A}_{\mathcal{W}}$ equipped with a probability P. η_0 is some nuisance parameter and ψ is a score function (i.e. likelihood function, moment condition).

Definition

Gateaux Derivative

For $\widetilde{T} = \{\eta - \eta_0 : \eta \in T\}$ we define the the Gateaux derivative map $\mathrm{D}_r : \widetilde{T} \to \mathbb{R}^{d_\theta}$,

$$\mathrm{D}_r[\eta - \eta_0] := \partial_r \{\mathbb{E}_P[\psi(w; \theta_0, \eta_0 + r(\eta - \eta_0))]\}, \quad \eta \in T$$

for all $r \in [0, 1)$. We also denote

$$\partial_\eta \mathbb{E}_P[\psi(w; \theta_0, \eta_0)][\eta - \eta_0] := \mathrm{D}_0[\eta - \eta_0], \quad \eta \in T$$

Neyman Orthogonality

Neyman Orthogonality

The score function ψ obeys the orthogonality condition at (θ_0, η_0) with respect to the nuisance realization set $\mathcal{T}_N \subset T$ if Equation (2.1) holds and the Gateaux derivative map $\mathrm{D}_r[\eta - \eta_0]$ exists for all $r \in [0, 1)$ and $\eta \in \mathcal{T}_N$ vanishes at $r = 0$; namely,

$$\partial_\eta \mathbb{E}_P [\psi(w; \theta_0, \eta_0)] [\eta - \eta_0] = 0, \qquad \text{for all } \eta \in \mathcal{T}_N$$

Neyman Near-Orthogonality

The score function ψ obeys the λ_N near-orthogonality condition, $\cdots$, and $\eta \in \mathcal{T}_N$ is small at $r = 0$; namely,

$$\partial_\eta \mathbb{E}_P [\psi(w; \theta_0, \eta_0)] [\eta - \eta_0] \leq \lambda_N, \qquad \text{for all } \eta \in \mathcal{T}_N$$

where $0 < \lambda_N = o\left(N^{-1/2}\right)$.

Likelihood with Finite Dimension Nuisance Parameter

Suppose for the maximum likelihood estimation where the true parameter values θ_0 and β_0 solve the optimization problem,

$$\max_{\theta \in \Theta, \beta \in \mathcal{B}} \mathbb{E}_P[\ell(w; \theta, \beta)]$$

With mild condition, we have,

$$\mathbb{E}_P\left[\partial_\theta \ell\left(w; \theta_0, \beta_0\right)\right] = 0, \quad \mathbb{E}_P\left[\partial_\beta \ell\left(w; \theta_0, \beta_0\right)\right] = 0$$

The original choice of score function is

$$\varphi(w; \theta, \beta) = \partial_\theta \ell(w; \theta, \beta)$$

Likelihood with Finite Dimension Nuisance Parameter

In order to achieve Neyman orthogonality, we set

$$\psi(w;\theta,\eta) = \partial_\theta \ell(w;\theta,\beta) - \mu \partial_\beta \ell(w;\theta,\beta)$$

where the nuisance parameter is $\eta = (\beta', \text{vec}(\mu)')' \in T = \mathcal{B} \times \mathbb{R}^{d_\theta d_\beta}$ and and μ is the $d_\theta \times d_\beta$ orthogonalization parameter matrix.
The true value of μ, namely μ_0, solves the equation $J_{\theta\beta} - \mu J_{\beta\beta} = 0$ for

$$J = \begin{pmatrix} J_{\theta\theta} & J_{\theta\beta} \\ J_{\beta\theta} & J_{\beta\beta} \end{pmatrix} = \partial_{(\theta',\beta')} \mathbb{E}_P \left[\partial_{(\theta',\beta')} \ell(w;\theta,\beta) \right] \Big|_{\theta=\theta_0;\beta=\beta_0}$$

We can show that this score function is Neyman orthogonal score when $J_{\beta\beta}$ is invertible.

Likelihood with Infinite Dimension Nuisance Parameter

Still consider the likelihood function $\ell(w; \theta, \beta)$. Now, instead of assuming that $\mathcal{B}$ is a (convex) subset of a finite-dimensional space, we assume that $\mathcal{B}$ is some (convex) set of functions, so that β is the functional nuisance parameter. Let

$$\beta_\theta = \arg\max_{\beta \in \mathcal{B}} \mathbb{E}_P[\ell(w; \theta, \beta)]$$

Now consider the score function using concentrated-out technique

$$\psi(w; \theta, \eta) = \frac{d\ell(w; \theta, \eta(\theta))}{d\theta}$$

The nuisance parameter is $\eta : \Theta \to \mathcal{B}$, and its true value η_0 is given by $\eta_0(\theta) = \beta_\theta, \quad$ for all $\theta \in \Theta$. This score function also satisfies the Neyman orthogonality condition.

Likelihood with Infinite Dimension Nuisance Parameter

Consider our PLR model,

$$\begin{aligned} Y &= D\theta_0 + g_0(X) + U, \ \mathbb{E}[U|X, D] = 0 \\ D &= m_0(X) + V, \qquad \mathbb{E}[V|X] = 0 \end{aligned}$$

We use,

$$\ell(w; \theta, \beta) = -\frac{1}{2}(Y - D\theta - \beta(X))^2$$

and the true values are

$$(\theta_0, \beta_0) = \arg \max_{\theta \in \Theta, \beta \in \mathcal{B}} \mathbb{E}_P[\ell(w; \theta, \beta)]$$

Likelihood with Infinite Dimension Nuisance Parameter

Therefore, the true β can be expressed using θ_0 as,

$$\beta_\theta(X) = \mathbb{E}_P[Y - D\theta | X], \quad \theta \in \Theta$$

Using the concentrated-out technique, our Neyman orthogonal score function is,

$$\psi\left(w; \theta, \beta_\theta\right) = (D - m_0(X)) \times (Y - D\theta - g_0(X))$$

Empirically, this gives the estimator $\hat{\theta}_0$

$$\frac{1}{n} \sum_{i \in I} (D_i - \hat{m}_0(X_i)) \times \left(Y - D_i\hat{\theta}_0 - \hat{g}_0(X)\right) = 0$$

Overview

Double Machine Learning Algorithm

a. Take a K-fold random partition $(I_k)_{k=1}^K$ of observation indices $[N] = \{1, \ldots, N\}$ such that the size of each fold I_k is $n = N/K$. Also, for each $k \in [K] = \{1, \ldots, K\}$, define $I_k^c := \{1, \ldots, N\} \backslash I_k$.
b. For each $k \in [K]$, construct an ML estimator $\hat{\eta}_{0,k} = \hat{\eta}_0 \left((w_i)_{i \in I_k^c} \right)$ of η_0, where $\hat{\eta}_{0,k}$ is a random element in T, and where randomness depends only on the subset of data indexed by I_k^c.
c. For each $k \in [K]$, construct the estimator $\theta_{0,k}$ as the solution of the following equation:

$$\mathbb{E}_{n,k} \left[\psi \left(w; \check{\theta}_{0,k}, \hat{\eta}_{0,k} \right) \right] = 0$$

where ψ is the Neyman orthogonal score, and $E_{n,k}$ is the empirical expectation over the k-th fold of the data.
d. Aggregate the estimators: $\tilde{\theta}_0 = \frac{1}{K} \sum_{k=1}^K \check{\theta}_{0,k}$

Double Machine Learning Algorithm

- In Step (c), if achievement of exact 0 is not possible, we cab define the estimator $\vec{\theta}_{0,k}$ of θ_0 as an approximate ϵ_N-solution:

$$\left\|\mathbb{E}_{n,k}\left[\psi\left(w;\check{\theta}_{0,k},\hat{\eta}_{0,k}\right)\right]\right\| \leq \inf_{\theta\in\Theta}\left\|\mathbb{E}_{n,k}\left[\psi\left(w;\theta,\hat{\eta}_{0,k}\right)\right]\right\| + \epsilon_N,$$

  where $\epsilon_N = o\left(\delta_N N^{-1/2}\right)$ and $(\delta_N)_{N\geq 1}$ is some sequence of positive constants converging to zero.

- We can also aggregate Step (c) and (d) such that

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\psi\left(w;\tilde{\theta}_0,\hat{\eta}_{0,k}\right)\right] = 0$$

Overview

Linear Score Function

We first consider the case of linear score function, where

$$\psi(w;\theta,\eta) = \psi^a(w;\eta)\theta + \psi^b(w;\eta), \quad \text{for all } w \in \mathcal{W}, \theta \in \Theta, \eta \in T \tag{4.1}$$

Assumptions for Linear Score Function

Assumption (4.1)

For all $N \geq 3$ and $P \in \mathcal{P}_N$, the following conditions hold.

- The true parameter value θ_0 obeys Equation (2.1).
- The score ψ is linear in the sense of (4.1).
- The map $\eta \mapsto E_P[\psi(w; \theta, \eta)]$ is twice continuously Gateaux-differentiable on T.
- The score ψ obeys the Neyman orthogonality or, more generally, the Neyman λ_N near-orthogonality condition at (θ_0, η_0) with respect to the nuisance realization set $\mathcal{T}_N \subset T$.
- The identification condition holds; namely, the singular values of the matrix $J_0 := \mathbb{E}_P\left[\psi^a\left(w; \eta_0\right)\right]$ are between c_0 and c_1.

Assumption 4.1 requires scores to be Neyman orthogonal or near-orthogonal and imposes mild smoothness requirements and the canonical identification condition.

Assumptions for Linear Score Function

Assumption (4.2)

For all $N \geq 3$ and $P \in \mathcal{P}_N$, the following conditions hold.

- Given a random subset I of $[M]$ of size $n = N/K$, the nuisance parameter estimator $\hat{\eta}_0 = \hat{\eta}_0\left((w_i)_{i \in It}\right)$ belongs to the realization set $\mathcal{T}_N$ with probability at least $1 - \Delta_N$ where $\mathcal{T}_N$ contains η_0 and is constrained by the next conditions.
- The moment conditions hold:

$$m_N := \sup_{\eta \in \mathcal{T}_N} \left(\mathbb{E}_P\left[\|\psi\left(w; \theta_0, \eta\right)\|^q\right]\right)^{1/q} \leq c_1$$

$$m'_N := \sup_{\eta \in \mathcal{T}_N} \left(E_P\left[\|\psi^a(w; \eta)\|^q\right]\right)^{1/q} \leq c_1$$

Assumptions for Linear Score Function

Assumption (4.2 continued)

- The following conditions on the statistical rates r_N, r'_N, and λ'_N hold:

$$r_N := \sup_{\eta \in \mathcal{T}_N} \left\| \mathbb{E}_P \left[\psi^a (w; \eta) \right] - \mathbb{E}_P \left[\psi^a (w; \eta_0) \right] \right\| \leq \delta_N$$

$$r'_N := \sup_{\eta \in \mathcal{T}_N} \left(\mathbb{E}_P \left[\left\| \psi (w; \theta_0, \eta) - \psi (w; \theta_0, \eta_0) \right\|^2 \right] \right)^{1/2} \leq \delta_N$$

$$\lambda'_N := \sup_{r \in (0,1), \eta \in \mathcal{T}_N} \left\| \partial_r^2 \mathbb{E}_P \left[\psi (w; \theta_0, \eta_0 + r(\eta - \eta_0)) \right] \right\| \leq \delta_N / \sqrt{N}$$

- The variance of the score ψ is non-degenerate: All eigenvalues of the matrix

$$\mathbb{E}_P \left[\psi (w; \theta_0, \eta_0) \, \psi (w; \theta_0, \eta_0)' \right]$$

are bounded from below by c_0.

Theoretical Results for Linear Score Function

Theorem (4.3)

Suppose that Assumptions 4.1 and 4.2 hold. In addition, suppose that $\delta_N \geq N^{-1/2}$ for all $N \geq 1$. Then the DML estimators $\tilde{\theta}_0$ concentrate in a $1/\sqrt{N}$ neighborhood of θ_0 and are approximately linear and centered Gaussian,

$$\sqrt{N}\sigma^{-1}\left(\tilde{\theta}_0 - \theta_0\right) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\overline{\psi}\left(w_i\right) + O_P\left(\rho_N\right) \rightsquigarrow \mathcal{N}\left(0, \mathrm{I}_d\right)$$

uniformly over $P \in \mathcal{P}_N$, where the size of the remainder term obeys $\rho_N := N^{-1/2} + r_N + r'_N + N^{1/2}\lambda_N + N^{1/2}\lambda'_N \lesssim \delta_N$.
Here, $\overline{\psi}(\cdot) := -\sigma^{-1}J_0^{-1}\psi\left(\cdot, \theta_0, \eta_0\right)$ is the influence function, and the approximate variance is

$$\sigma^2 := J_0^{-1}\mathbb{E}_P\left[\psi\left(w; \theta_0, \eta_0\right)\psi\left(w; \theta_0, \eta_0\right)'\right]\left(J_0^{-1}\right)'$$

Theoretical Results for Linear Score Function

Theorem (4.4)

Suppose that Assumptions 4.1 and 4.2 hold. In addition, suppose that $\delta_N \geq N^{-[(1-2/q)\wedge 1/2]}$ for all $N \geq 1$. Consider the following estimator of the asymptotic variance matrix of $\sqrt{N}\left(\tilde{\theta}_0 - \theta_0\right)$:

$$\hat{\sigma}^2 = \widehat{J}_0^{-1} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{n,k}\left[\psi\left(w; \tilde{\theta}_0, \hat{\eta}_{0,k}\right) \psi\left(w; \tilde{\theta}_0, \hat{\eta}_{0,k}\right)'\right] \left(\widehat{J}_0^{-1}\right)'$$

where $\widehat{J}_0 = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{n,k}\left[\psi^a\left(W; \hat{\eta}_{0,k}\right)\right]$. $\hat{\sigma}^2$ satisfies,

$$\hat{\sigma}^2 = \sigma^2 + O_P\left(\varrho_N\right), \quad \varrho_N := N^{-[(1-2/q)\wedge 1/2]} + r_N + r_N' \lesssim \delta_N$$

Non-linear Score Function

1. The assumptions are similar in non-linear score function case.
2. The DML estimator $\tilde{\theta}_0$ also has a optimal $N^{-1/2}$ convergence rate.
3. The variance matrix estimator $\hat{\sigma}^2 \rightarrow_p \sigma^2$ and we can replace σ^2 by $\hat{\sigma}^2$.
4. A confidence interval can be construct using results above.

Intuition

Denote $\mathbb{E}_N[\psi(w; \theta_0, \eta_0)]$ to be the empirical analogue of $\mathbb{E}_P[\psi(w; \theta_0, \eta_0)]$, with Equation (2.1),

$$\mathbb{E}_N\left[\psi\left(w, \hat{\theta}_0, \eta_0\right)\right] = 0$$

Assume the nuisance parameter η_0 is known, then

$$0 = \mathbb{E}_N\left[\psi\left(w, \hat{\theta}_0, \eta_0\right)\right] \approx \mathbb{E}_N[\psi(w, \theta_0, \eta_0)] + \partial_\theta \mathbb{E}_N[\psi(w, \theta_0, \eta_0)]\left(\hat{\theta}_0 - \theta_0\right)$$
$$\Rightarrow \partial_\theta \mathbb{E}_N[\psi(w, \theta_0, \eta_0)]\sqrt{N}\left(\hat{\theta}_0 - \theta_0\right) \approx -\sqrt{N}\mathbb{E}_N[\psi(w, \theta_0, \eta_0)]$$
$$\Rightarrow \sqrt{N}\left(\hat{\theta}_0 - \theta_0\right) \rightarrow_d \mathcal{N}\left(0, J^{-1}\Omega J^{-1\prime}\right)$$

Intuition

Now consider the case where we do not know η_0. Instead, we use $\hat{\eta}_0$ to estimate η_0, and we solve:

$$\mathbb{E}_N\left[\psi\left(w,\hat{\theta}_0,\hat{\eta}_0\right)\right]=0$$

Therefore,

$$0=\mathbb{E}_N\left[\psi\left(w,\hat{\theta}_0,\hat{\eta}_0\right)\right]\approx\mathbb{E}_N\left[\psi\left(w,\theta_0,\hat{\eta}_0\right)\right]+\partial_\theta\mathbb{E}_N\left[\psi\left(w,\theta_0,\hat{\eta}_0\right)\right]\left(\hat{\theta}_0-\theta_0\right)$$
$$\Rightarrow\partial_\theta\mathbb{E}_N\left[\psi\left(w,\theta_0,\hat{\eta}_0\right)\right]\sqrt{N}\left(\hat{\theta}_0-\theta_0\right)\approx-\sqrt{N}\mathbb{E}_N\left[\psi\left(w,\theta_0,\eta_0\right)\right]$$

In order to get a asymptotic result, we need $\mathbb{E}_N\left[\psi\left(w,\theta_0,\eta_0\right)\right]$ to behave well. If the hypothesis space of η has finite VC dimension, we can use a stochastic equicontinuity argument to achieve it. See [1].

Intuition

Further, we expand the equation above and get,

$$\begin{aligned}
&\partial_\theta \mathbb{E}_N \left[\psi\left(w, \theta_0, \hat{\eta}_0\right)\right] \sqrt{N}\left(\hat{\theta}_0 - \theta_0\right) \\
\approx &- \sqrt{N}\mathbb{E}_N \left[\psi\left(w, \theta_0, \eta_0\right)\right] \\
\approx &- \sqrt{N}\mathbb{E}_N \left[\psi\left(w, \theta_0, \eta_0\right) + \partial_\eta \psi\left(w, \theta_0, \eta_0\right)\left[\hat{\eta}_0 - \eta_0\right]\right] \\
&- \sqrt{N}\mathbb{E}_N \left[\frac{1}{2}\partial_{\eta^2}\psi\left(w, \theta_0, \eta_0\right)\left[\hat{\eta}_0 - \eta_0\right]\right]
\end{aligned}$$

1. The first term on the RHS behaves well.
2. The second term on the RHS goes to 0, which is guaranteed by Neyman (near)-orthogonality condition.
3. Cross-fitting and the concentration of $\|\hat{\eta}_0 - \eta_0\|$ guarantees the third term on the RHS goes to 0.

Intuition

- When we plug in an estimate of the nuisance parameter η_0 to estimate θ_0, a small error of $\hat{\eta}_0$ might be undesirable. Neyman (near)-orthogonality condition guarantees that using plug-in estimator won't hurt.
- Estimating η_0 and θ_0 using the same data will cause overfitting problem. Cross-fitting solves this problem.

Overview

Partial Linear Regression

Here we revisit the PLR model.

$$\begin{aligned} Y &= D\theta_0 + g_0(X) + U, \quad \mathbb{E}[U|X, D] = 0 \\ D &= m_0(X) + V, \qquad\qquad \mathbb{E}[V|X] = 0 \end{aligned}$$

We here provide score function

$$\psi(w; \theta, \eta) := \{Y - D\theta - g(X)\}(D - m(X)), \quad \eta = (g, m)$$

which satisfies Neyman orthogonality condition,

$$\begin{aligned} &\mathbb{E}_P \psi\left(w; \theta_0, \eta_0\right) = 0 \\ &\partial_\eta \mathbb{E}_P \psi\left(w; \theta_0, \eta_0\right)\left[\eta - \eta_0\right] = 0 \end{aligned}$$

for $\eta_0 = (g_0, m_0)$.

Partial Linear Regression

Under mild condition, we can show that this score function is a linear one and satisfies Assumption 4.1 and 4.2. Therefore,

1. The DML estimator $\tilde{\theta}_0$ has

$$\sigma^{-1}\sqrt{N}\left(\tilde{\theta}_0 - \theta_0\right) \rightsquigarrow \mathcal{N}(0,1)$$

   where $\sigma^2 = \left(\mathbb{E}_P\left[V^2\right]\right)^{-1}\mathbb{E}_P\left[V^2U^2\right]\left(\mathbb{E}_P\left[V^2\right]\right)^{-1}$.
2. The plug-in estimator $\hat{\sigma}^2$ converges in probability to σ^2.
3. We can construct confidence interval $\tilde{\theta}_0 \pm \Phi^{-1}(1-\alpha/2)\hat{\sigma}/\sqrt{N}$ which has uniform asymptotic validity

$$\lim_{N\to\infty}\sup_{P\in\mathcal{P}}\left|\mathbb{P}_P\left(\theta_0 \in \left[\tilde{\theta}_0 \pm \Phi^{-1}(1-\alpha/2)\hat{\sigma}/\sqrt{N}\right]\right) - (1-\alpha)\right| = 0$$

Inference on Treatment Effect

Consider the following model,

$$
\begin{aligned}
Y &= g_0(D, X) + U, \ \mathbb{E}_P[U|X, D] = 0 \\
D &= m_0(X) + V, \quad \mathbb{E}_P[V|X] = 0
\end{aligned}
$$

Here $D \in \{0, 1\}$ and we are interested in average treatment effect (ATE),

$$\theta_0 = \mathbb{E}_P\left[g_0(1, X) - g_0(0, X)\right]$$

and average treatment effect on the treated (ATTE),

$$\theta_0 = \mathbb{E}_P\left[g_0(1, X) - g_0(0, X) | D = 1\right]$$

Inference on Treatment Effect

We now employ DML method to estimate ATE and ATTE. For the estimation of ATE, we set

$$\psi(w;\theta,\eta) := (g(1,X) - g(0,X)) + \frac{D(Y - g(1,X))}{m(X)} - \frac{(1-D)(Y-g(0,X))}{1-m(X)} - \theta$$

with nuisance parameter $\eta = (g, m)$, and for the estimation of ATTE, we set

$$\psi(w;\theta,\eta) = \frac{D(Y-\bar{g}(X))}{p} - \frac{m(X)(1-D)(Y-\bar{g}(X))}{p(1-m(X))} - \frac{D\theta}{p}$$

with nuisance parameter $\eta = (\bar{g}, m, p)$. The true value is $\bar{g}_0(X) = g_0(0,X)$, $p_0 = \mathbb{E}_P[D]$.

Inference on Treatment Effect

Our score functions above satisfy the moment condition and Neyman orthogonality condition. Under some mild assumptions, we can verify that our model satisfies Assumption 4.1 and 4.2.

1. The DML estimator $\tilde{\theta}_0$ also has a optimal $N^{-1/2}$ convergence rate to the true estimator θ_0 for ATE and ATTE respectively.
2. The variance matrix estimator $\hat{\sigma}^2 \to_p \sigma^2$ and we can replace σ^2 by $\hat{\sigma}^2$.
3. A confidence interval can be construct using results above.

References I

[1] Donald WK Andrews et al. "Asymptotics for semiparametric econometric models via stochastic equicontinuity". In: *ECONOMETRICA-EVANSTON ILL-* 62 (1994), pp. 43–43.

[2] Victor Chernozhukov et al. "Double/debiased machine learning for treatment and structural parameters". In: *The Econometrics Journal* 21 (2018), pp. C1–C68.