Double Machine Learning

PS690 Computational Methods in Social Science

Jiawei Fu
Department of Political Science
Duke University

Sep 23, 2025

Overview

1. Partial Linear with LASSO
2. Partial Linear Model in General
3. Generic Debiased (or Double) Machine Learning
4. Doubly Robust Estimator

Partially Linear Regression Model

- Consider the following partially lenear regression model: (PLM):

$$Y = \beta D + g(X) + \epsilon, \mathbb{E}[\epsilon | D, X] = 0$$
$$D = m(X) + V$$

- Y is the outcome variable, D is the treatment, and X is a high-dimensional vector of controls.
- β is the main regression coefficient that we would like to infer.
- $g(X)$ can be fully nonlinear.
- How to identify β and make valid inference?

Partially Linear Regression Model

- Suppose we learn $g(X)$ using a random forest. And then, we plug in $\hat{g}(X)$, and run least squares to obtain $\hat{\beta}$ from regression model $Y = \beta D + \hat{g}(X) + \epsilon$. (Learn g from half sample)

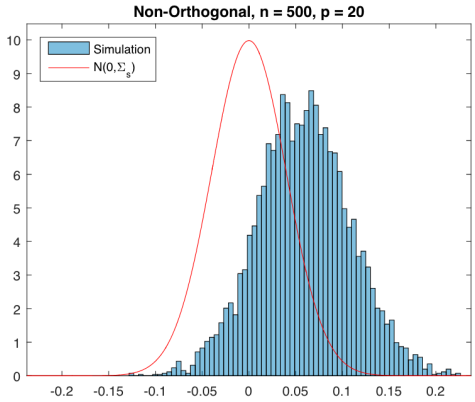

Figure: From [Chernozhukov et al., 2024]: Shows the distribution of $\hat{\beta} - \beta$

Review of FWL theorem

- Consider the simple model:

$$Y = \beta D + \gamma' X$$

  where X has dimension $p < n$.

- β can be estimated partialling-out regression procedure:
  1. Regress Y on X, obtain residuals $\tilde{Y} = Y - \alpha'_{yx} X$, where $\alpha_{yx} \in \arg min \mathbb{E}[(Y - \alpha' X)^2]$
  2. Regress D on X, obtain residuals $\tilde{D} = Y - \alpha'_{dx} X$, where $\alpha_{dx} \in \arg min \mathbb{E}[(D - \alpha' X)^2]$
  3. Regress $\tilde{Y}$ on $\tilde{D}$ to obtain $\beta = \arg min \mathbb{E}[(\tilde{Y} - \beta \tilde{X})^2]$

Double Lasso

- Now consider the linear model with X has high-dimension $p \gg n$:

$$Y = \beta D + \gamma' X$$

.

- How to estimate β and conduct valid inference?
- Note γ the population coefficient defined as mse not lasso type.
- Double Lasso procedure:
  1. Run *Lasso* regression of Y_i on X_i, and obtain residuals $\tilde{Y}_i = Y_i - \hat{\alpha}'_{yx} X_i$, where $\hat{\alpha}_{yx} = \arg min \sum_i (Y_i - \alpha' X_i)^2 + \lambda_1 \sum_j |\alpha_j|$
  2. Run *Lasso* regression of D_i on X_i, and obtain residuals $\tilde{D}_i = D_i - \hat{\alpha}'_{dx} X_i$, where $\hat{\alpha}_{dx} = \arg min \sum_i (D_i - \alpha' X_i)^2 + \lambda_2 \sum_j |\alpha_j|$
  3. Run least squares regression of $\tilde{Y}$ on $\tilde{D}$ to obtain $\hat{\beta}$

We can use standard results from this regression, ignoring that the input variables were previously estimated, to perform inference about the predictive effect, β.

Double LASSO

- Good performance of the Double Lasso procedure relies on
  1. approximate sparsity of the population regression coefficients: $|\gamma_{yx}|_{(j)} \leq Aj^{-a}$ and $|\gamma_{dx}|_{(j)} \leq Aj^{-a}$
  2. careful choice of the Lasso tuning parameters: λ_1 and $\lambda_2 = 2c\hat{\sigma}\sqrt{n}z_{1-a/2p}$,where$\hat{\sigma} \approx \sigma = \sqrt{\mathbb{E}\epsilon^2}$, which can guarantees that we produce high quality prediction of Y and D while simultaneously avoiding overfitting under approximate sparsity.
- Otherwise, we cannot theoretically ensure that first step estimation of $\tilde{Y}$ and $\tilde{D}$ does not have first-order impacts on $\hat{\beta}$.
- Practically, penalty parameter selected via cross-validation can perform poorly in simulations in moderately sized samples.

Double Lasso

- We summarize it into the following theorem.

Theorem (Adaptive Inference with Double Lasso in High-Dimensional Regression [Chernozhukov et al., 2024])

Under approximate sparsity and additional regularity conditions, the estimation error in $\tilde{D}$ and $\tilde{Y}$ has no first order effect on $\hat{\beta}$, and

$$\sqrt{n}(\hat{\beta} - \beta) \approx \sqrt{n}\frac{\mathbb{E}_n \tilde{D}\epsilon}{\mathbb{E}_n \tilde{D}^2} \to N(0, V)$$

where $V = (\mathbb{E}\tilde{D}^2)^{-1}\mathbb{E}[\tilde{D}^2\epsilon^2](\mathbb{E}\tilde{D}^2)^{-1}$

- Just like in the low-dimensional case, we can use these results to construct a confidence interval. The standard error of β is $\sqrt{\frac{\hat{V}}{n}}$, where $\hat{V}$ is a plug-in estimator.

Double Lasso

- In the previous case, β is the target parameter and $\eta = (\alpha_{yx}, \alpha_{dx})$ are called nuisance parameters, with true value $\eta^0 = (\alpha^0_{yx}, \alpha^0_{dx})$.
- Nuisance parameters refer to parameters that must be learned or otherwise adjusted for in order to learn the parameter of interest but are not of direct interest themselves.
- Note that $\hat{\beta}$ depends on the values of nuisance parameters: $\hat{\beta}(\eta)$.
- In high-dimensional settings, we use regularization procedures to estimate the nuisance parameters, which generally results in bias.
- Therefore, $\hat{\eta}$ are close to, but not exactly equal to, the true value η^0.

Why partialling-out works?

- The main idea of the Double Lasso approach is that, in the **population limit**, it corresponds to a procedure for learning the target parameter β that is first-order insensitive to local perturbations of the nuisance parameters around their true values η^0

$$\partial_\eta \beta(\eta^0) = 0$$

- For now, just treat ∂ as traditional derivative. We will see formal definition later.
- We will call the local insensitivity of target parameters to nuisance parameters as Neyman orthogonality of the estimation process.
- Neyman orthogonality, which guarantees that the target parameter is locally insensitive to perturbations of the nuisance parameters around their true values, then ensures that this bias does not transmit to the estimation of the target parameter, at least to the first order.

Why partialling-out works?

- Let us prove $\partial_\eta \beta(\eta^0) = 0$.
- Recall from linear regression class, $\beta(\eta)$ can be defined as the solution to a moment equation:

$$M(\beta, \eta) := \mathbb{E}[(\tilde{Y}(\eta) - \beta \tilde{D}(\eta))\tilde{D}(\eta)] = 0,$$

and

$$\tilde{Y} = Y - \eta_1' X, \quad \tilde{D} = D - \eta_2' X$$

- Then by the implicit function theorem:

$$\partial_\eta \beta(\eta^0) = -\partial_\beta M(\beta, \eta^0)^{-1} \partial_\eta M(\beta, \eta^0)$$

- $\partial_\eta M(\beta, \eta^0)$ has two parts, with respect to η_1 and η_2.

Why partialling-out works?

- Note, we are working on population moment, rather than sample one; in population, those nuisance parameter is defined by BLP: $\mathbb{E}[X(D - \alpha'_{dx}X)] = 0$ and $\mathbb{E}[X(Y - \alpha'_{yx}X)] = 0$.
- With respect to η_1, i.e.α_1, we have

$$\partial_{\eta_1} M(\beta, \eta^0) = \mathbb{E}[X\tilde{D}(\eta^0)] = \mathbb{E}[X(D - \alpha'_{dx}X)] = 0$$

- With respect to η_2, i.e.α_2, we have

$$\begin{aligned}\partial_{\eta_2} M(\beta, \eta^0) &= -\mathbb{E}[X\tilde{Y}(\eta^0)] + 2\mathbb{E}[\beta X\tilde{D}(\eta^0)] \\ &= -\mathbb{E}[X(Y - \alpha'_{yx}X))] + 2\mathbb{E}[\beta X(D - \alpha'_{dx}X)] = 0\end{aligned}$$

Invalid Single LASSO

- Consider the a seemingly sensible approach: Apply Lasso regression of Y on D and X to select covariates X_α, and then fit the model by least squares of Y on D and X_α.
- The inference will be invalid while the prediction is fine.
- Let us check the Neyman condition.
- The moment condition is $M = \mathbb{E}[(Y - \beta D - \gamma' X_\alpha)D] = 0$.
- However, $\partial_\beta M(\alpha, \beta) = \mathbb{E}[DX_\alpha] \neq 0$, unless D is orthogonal to X (for example, in the experiment).
- Then, the bias and the slower than parametric rate of convergence ($\sqrt{s \log(p \vee n)/n}$) of $\gamma' X$ will transmit to bias and slower than $\sqrt{n}$ convergence in estimates of β.

Invalid Single LASSO

Example 4.3.1 In the Notebooks 4.7.1, we compare the performance of the naive and orthogonal methods in a computational experiment where $p = n = 100$, $\beta_j = 1/j^2$, $(\gamma_{DW})_j = 1/j^2$, and

$$Y = 1 \cdot D + \beta' W + \varepsilon_Y, \quad W \sim N(0, I), \quad \varepsilon_Y \sim N(0,1)$$

$$D = \gamma'_{DW} W + \tilde{D}, \quad \tilde{D} \sim N(0,1)/4.$$

From the histograms shown in Figure 4.1, we see that the naive estimator is heavily biased, as expected from the lack of Neyman orthogonality in its estimation strategy. We also see that the Double Lasso estimator, which is based on principled partialling-out such that Neyman orthogonality is satisfied, is approximately unbiased and Gaussian.

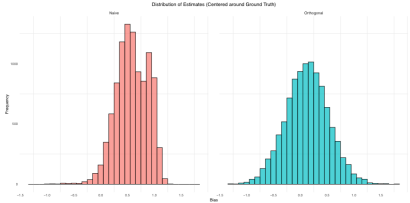

Figure: From [Chernozhukov et al., 2024]

Partial Linear Model in General

- Consider the following partially linear regression model: (PLM):

$$y = \beta D + g(X) + U, \mathbb{E}[U|D, X] = 0$$
$$D = m(X) + V$$

- Double ML for the Partially Linear Model:
  1. Cross-fitting step: Partition data into K folds. For each fold k, compute ML estimators $\hat{l}_{[k]}$ and $\hat{m}_{[k]}$ of $\mathbb{E}[Y|X]$ and $\mathbb{E}[D|X]$, leaving out the kth block of data. Obtain the cross-fitted residuals for each i fold k:

$$\tilde{Y}_i = Y_i - \hat{l}_{[k]}(X_i), \quad \tilde{D}_i = Y_i - \hat{m}_{[k]}(X_i)$$

  2. Apply ordinary least squares of $\tilde{Y}$ on $\tilde{D}$. That is, obtain $\hat{\beta}$ from the normal equations: $\mathbb{E}_n[(\tilde{Y} - \beta\tilde{D})\tilde{D}] = \frac{1}{n}\sum_{i=1}^{n}[(\tilde{Y}_i - \beta\tilde{D}_i)\tilde{D}_i] = 0$
  3. Construct standard errors and confidence intervals as in standard least squares theory.

Partial Linear Model in General

- Let $||h||_{L^2} := \sqrt{E_X[h^2(X)]}$

Theorem (Adaptive Inference on a Target Parameter in PLM)

Suppose that estimators $\hat{l}_{[k]}(X)$ *and* $\hat{m}_{[k]}(X)$ *provide approximations to the best predictors* $l(X)$ *and* $m(X)$ *that are of sufficiently high-quality:*

$$n^{1/4}(||\hat{l}_{[k]} - l||_{L^2} + ||\hat{m}_{[k]} - m||_{L^2}) \approx 0$$

Then, under other regularity conditions,

$$\sqrt{n}(\hat{\beta} - \beta) \approx \sqrt{n}(\mathbb{E}_n \tilde{D}^2)^{-1}\mathbb{E}_n \tilde{D}\epsilon \to N(0, V)$$

, where $V = (\mathbb{E}\tilde{D}^2)^{-1}\mathbb{E}[\tilde{D}^2\epsilon^2](\mathbb{E}\tilde{D}^2)^{-1}$

Remark 1: Neyman Orthogonality by partial out

- Neyman Orthogonality controls regularization bias.
- The lack of Neyman orthogonality means that estimates of the parameter of interest are heavily impacted by estimation of the nuisance parameters.
- Any biases in estimation of $g(X)$, which are essentially unavoidable in high-dimensional estimation, create a non-trivial bias in the estimate of the main effect. This bias is large enough to cause the failure of conventional inference.

Remark 1: Neyman Orthogonality by partial out

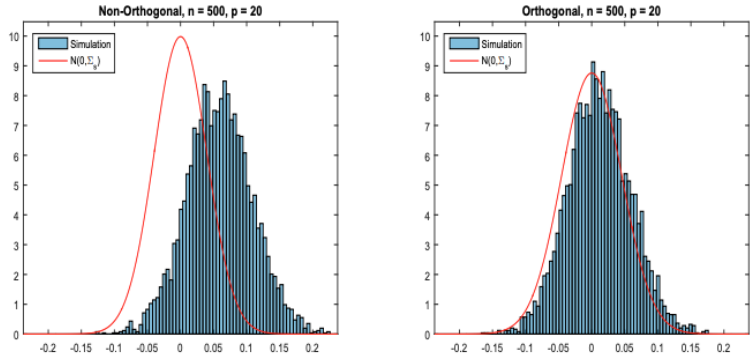

Figure: [Chernozhukov et al., 2018]. The nuisance parameter is estimated by random forest.

Remark 1: Neyman Orthogonality by partial out

- Mathematically,

$$\sqrt{n}(\hat{\beta}-\beta) = \underbrace{(\frac{1}{n}\sum_i D_i^2)^{-1}\frac{1}{\sqrt{n}}\sum_i D_i U_i}_{A}$$
$$+ \underbrace{(\frac{1}{n}\sum_i D_i^2)^{-1}\frac{1}{\sqrt{n}}\sum_i m(X_i)(g_0(X_i)-\hat{g}_0(X_i))}_{B}$$
$$+ \underbrace{(\frac{1}{n}\sum_i D_i^2)^{-1}\frac{1}{\sqrt{n}}\sum_i V_i(g_0(X_i)-\hat{g}_0(X_i))}_{C}$$

Remark 1: Neyman Orthogonality by partial out

- Part A: $(\frac{1}{n}\sum_i D_i^2)^{-1}\frac{1}{\sqrt{n}}\sum_i D_i U_i$
- It is well-behaved, asymptotically normal, under mild conditions.
- Check the expectation is zero: $\mathbb{E}[D_i U_i] = \mathbb{E}[\mathbb{E}[D_i U_i|X, D]] = \mathbb{E}[D_i \mathbb{E}[U_i|X, D]] = 0$

Remark 1: Neyman Orthogonality by partialling out

- Part B: $(\frac{1}{n}\sum_i D_i^2)^{-1}\frac{1}{\sqrt{n}}\sum_i m(X_i)(g_0(X_i) - \hat{g}_0(X_i))$
- This part is problematic: Regularization keeps the variance of the estimator from exploding but also necessarily induces substantive biases in the estimator $\hat{g}_0$ of g.
- B is the sum of n terms that do not have mean zero, $m(X_i)(g_0(X_i) - \hat{g}_0(X_i))$, divided by $\sqrt{n}$.
- It will approach 0, but too slowly for our estimator to be $\sqrt{n}$ consistent.
- Specifically, this results in the low rate of convergence: $g_0(X_i) - \hat{g}_0(X_i) \propto n^{-\phi_g}$, where $\phi_g < \frac{1}{2}$.
- Therefore,

$$\frac{1}{\sqrt{n}}\sum_i m(X_i)(g_0(X_i) - \hat{g}_0(X_i)) \propto \sqrt{n}n^{-\phi_g} \to \infty$$

- To solve this problem, we need Neyman Orthogonality.

Remark 1: Neyman Orthogonality by partialling out

- Let us now look at DML estimator $\hat{\beta} = (\frac{1}{n}\sum_{i=1}^{n} \tilde{D}_i \tilde{D}_i)^{-1} \frac{1}{n}\sum_{i=1}^{n} \tilde{D}_i \tilde{Y}_i$.
- For the ease of exposition, we use a slightly different estimator $\hat{\beta} = (\frac{1}{n}\sum_{i=1}^{n} \tilde{D}_i D_i)^{-1} \frac{1}{n}\sum_{i=1}^{n} \tilde{D}_i \tilde{Y}_i$, which is asymptotically the same as the above DML estimator.
- Similarly, we expand and obtain

$$\sqrt{n}(\hat{\beta} - \beta) = \underbrace{(\frac{1}{n}\sum_i \tilde{D}_i D_i)^{-1} \frac{1}{\sqrt{n}} \sum_i \tilde{D}_i U_i}_{A}$$
$$+ \underbrace{(\frac{1}{n}\sum_i \tilde{D}_i D_i)^{-1} \frac{1}{\sqrt{n}} \sum_i (m(X_i) - \hat{m}(X_i))(g_0(X_i) - \hat{g}_0(X_i))}_{B}$$
$$+ \underbrace{(\frac{1}{n}\sum_i \tilde{D}_i D_i)^{-1} \frac{1}{\sqrt{n}} \sum_i V_i (g_0(X_i) - \hat{g}_0(X_i))}_{C}$$

Remark 1: Neyman Orthogonality by partialling out

- Now, look at the B term.

$$(\frac{1}{n}\sum_i \tilde{D}_i D_i)^{-1}\frac{1}{\sqrt{n}}\sum_i (m(X_i)-\hat{m}(X_i))(g_0(X_i)-\hat{g}_0(X_i))$$

- It is the product of two small error terms, and thus vanishes more qucikly!
- Again, because we are using regularized ML methods, the converge rates are slow

$$g_0(X_i)-\hat{g}_0(X_i) \propto n^{-\phi_g}, where \;\; \phi_g < \frac{1}{2}$$

$$m_0(X_i)-\hat{m}_0(X_i) \propto n^{-\phi_m}, where \;\; \phi_m < \frac{1}{2}$$

- That is, the root mean squared error $\sqrt{\frac{1}{n}(g_0(X_i)-\hat{g}_0(X_i))^2} = ||g_0(X_i)-\hat{g}_0(X_i)||_{L^2} = o_p(n^{-\phi_g})$ etc.
- It will be upper bounded by $\sqrt{n}o_p(n^{-(\phi_m+\phi_g)})$.
- if $\phi_m+\phi_g \geq \frac{1}{2}$, it is $o_p(1)$, then this part can be ignored.

Remarks 2: Cross-fitting

- The DML algorithm also uses a form of sample splitting, called cross-fitting, to guard against a less obvious source of bias that may arise when estimation of nuisance parameters results in overfitting.
- Heuristically, overfitting simply means that an estimator has captured not just generalizable signal, but also noise that is idiosyncratic to each observation.
- Cross-fitting guards against this source of bias, as overfitting resulting from learning nuisance functions in one subsample will not carry over when the nuisance function estimates are applied in a different, separate subsample.

Remarks 2: Cross-fitting

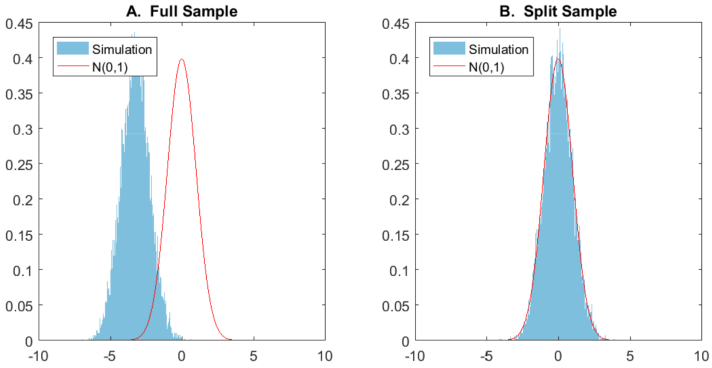

Figure: [Chernozhukov et al., 2018]. This figure illustrates how the bias resulting from overfitting in the estimation of nuisance functions can cause the main estimator to be biased and how sample splitting completely eliminates this problem.

Remarks 2: Cross-fitting

- Note that we did not make use of cross-fitting in previous Double-lasso because the plug-in tuning choice λ theoretically guarantees that overfitting is sufficiently well-controlled that sample splitting is not required.
- Such refined and theoretically rigorous choices of tuning parameters are not yet available for other machine learning methods.
- Indeed, even when using Lasso, crossfitting should be employed if cross-validation, rather than the theoretical plug-in, is used for selecting λ.

Remarks 2: Cross-fitting

- Look at the C term; no matter whether we use DML or naive method, it remains the same (ignore the inverse term, this part converges):

$$\frac{1}{\sqrt{n}} \sum_i V_i (g_0(X_i) - \hat{g}_0(X_i))$$

- We hope this part will vanish. But when will this term do not vanish?
- We will see that this will happen when V_i and $g_0(X_i) - \hat{g}_0(X_i)$ are correlated. For example, when $\hat{g}_0$ is overfit, it will fit parts not explained by X_i in the outcome model.
- For example, we assume $\hat{g}_0(X_i) = g_0(X_i) + \frac{Y_i - g_0(X_i)}{n^{1/2-\epsilon}}$. The second term is the overfit.
- Note that this estimator is not bad; it converges to $g_0(X_i)$ at a nearly parametric rate $n^{1/2-\epsilon}$.

Remarks 2: Cross-fitting

- However, even with this fast rate, the slight overfit will make C term explode.

$$\frac{1}{\sqrt{n}}\sum_i V_i(g_0(X_i) - \hat{g}_0(X_i)) = \frac{1}{\sqrt{n}}\sum_i V_i(\frac{Y_i - g_0(X_i)}{n^{1/2-\epsilon}})$$
$$= \frac{1}{\sqrt{n}}\sum_i V_i(\frac{\beta D_i + U_i}{n^{1/2-\epsilon}})$$

- Note $\mathbb{E}[V_i D_i] = \mathbb{E}[\mathbb{E}[(D_i - m(X_i))D_i|X_i]] = \mathbb{E}[Var(D_i|X_i)] := c$; Then $n^\epsilon \beta c$ explodes.
- The remaining term will explode as well if we do not control enough X to remove endogeneity; but now is fine: $\mathbb{E}[V_i U_i] = O_p(\sqrt{n})$, then $n^{\epsilon-1}O_p(\sqrt{n}) = O_p(n^{\epsilon-\frac{1}{2}})$ will go to zero for small ϵ.

Remarks 2: Cross-fitting

- Now, let us see the magic of cross-fitting.
- Consider the simple case in which we split two sets: I and I^c.
- We use I^C to denote the auxiliary set where we estimate $\hat{g}^{I^C}(X)$, and the remaining evaluation set I to estimate $\hat{\beta}$
- We WTS that conditional expectation of C term for those on the evaluation set I is 0.

$$\begin{aligned}
\mathbb{E}[C^I|I^C] &= \frac{1}{\sqrt{|I|}} \sum_{i \in I} \mathbb{E}[\mathbb{E}[V_i(g_0(X_i) - \hat{g}_0^{I^C}(X_i)|I^C, X_i]|I^C] \\
&= \frac{1}{\sqrt{|I|}} \sum_{i \in I} \mathbb{E}[(g_0(X_i) - \hat{g}_0^{I^C}(X_i)\mathbb{E}[V_i|I^C, X_i]|I^C] \\
&= \frac{1}{\sqrt{|I|}} \sum_{i \in I} \mathbb{E}[(g_0(X_i) - \hat{g}_0^{I^C}(X_i)\mathbb{E}[V_i|X_i]|I^C] \\
&= 0
\end{aligned}$$

- Therefore, unconditionally, it is mean 0 as well.

Remarks 2: Cross-fitting

- That is great! But does it go to zero quickly enough?
- We use the Chebyshev inequality: $\mathbb{P}[|C^I| > t] \leq \frac{\mathbb{E}[(C^I)^2]}{t^2}$, and show that the numerator converges to 0.
- We check the variance. Suppose $\sup_x Var(V_i|X_i) \leq B$. (We ignore some technical details; only sketch the logic.)

$$
\begin{aligned}
Var(C^I|I^C, X_i) &= \frac{1}{|I|}\sum_{i\in I}(g_0(X_i) - \hat{g}_0^{I^C}(X_i))^2 Var(V_i|X_i) \\
&\leq B\frac{1}{|I|}\sum_{i\in I}(g_0(X_i) - \hat{g}_0^{I^C}(X_i))^2 \\
&= Bo_p(n^{-2\phi_g}) \\
&\to 0
\end{aligned}
$$

DML

- We consider estimate and conduct inference based upon a method-of-moments estimator for some low-dimensional parameter θ_0:

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

- η_0 is an infinite-dimensional nuisance parameter.
- Suppose we use data and any (ML) methods get $\hat{\eta}_n$; then plug in and solve $\hat{\theta}$ from sample analogue

$$\frac{1}{n} \sum_{i=1}^{n} \psi(W; \hat{\theta}_n, \hat{\eta}_n) = 0$$

- The problem is that the bias of $\hat{\eta}_n$ will affect $\hat{\theta}$.
- $\hat{\eta}_0$ must be heavily regularized in high-dimensional settings, so these estimators will be biased in general.
- Moreover, $\hat{\theta}$ is not an average of iid data (because of $\hat{\eta}_n$). The distribution is complicated.
- Is it possible that we can conduct valid inference for θ by ignoring $\hat{\eta}_n$?

DML

- Many common problems can be seen as a method of moment problem.
- For example, linear regression model $y_i = x_i'\beta + e_i$,$\mathbb{E}[x_i e_i] = 0$ can be written as

$$\mathbb{E}[x_i(y - x_i'\beta)] = 0$$

- For example, recall the outcome regression estimator, average treatment effects τ can be solved by

$$\mathbb{E}[\mu_1(X_i) - \mu_0(X_i) - \tau] = 0$$

- Also, recall AIPW, we can also use it to solve τ:

$$\mathbb{E}[\mu_1(X_i) - \mu_0(X_i) + Z_i\frac{Y_i - \mu_1(X_i)}{\pi(X_i)} - (1 - Z_i)\frac{Y_i - \mu_0(X_i)}{1 - \pi(X_i)} - \tau] = 0$$

DML

- Is it possible that we can conduct valid inference for θ by ignoring $\hat{\eta}_n$?
- The answer is yes, if using DML.
- There are three key ingredients of DML estimation and inference.
- First ingredient: Neyman orthogonality condition.
- Using a Neyman orthogonal score eliminates the first-order biases arising from the replacement of η_0 with an ML estimator $\hat{\eta}_0$:

$$\frac{d}{dt}\mathbb{E}[\psi(W;\theta_0,\eta_0+t(\eta-\eta_0))|_{t=0}=0$$

- I sometimes use superscript 0 to denote the true value of a parameter if the parameter already has subscript. For example, μ_1^0 is the true parameter of μ_1.

DML

- Let us check moment for partial linear model : $\mathbb{E}[(Y - \beta T - g(X))T] = 0$
- It is easy to see that

$$\frac{d}{dt}\mathbb{E}[(Y - \beta_0 T - g_0(X) - t(g(X) - g_0(X))T]|_{t=0} = \mathbb{E}[(g(X) - g_0(X))T] \neq 0$$

- But residual regression satisfies the Nyeman Orthogonality condition.
- Also, for outcome regression estimator:

$$\frac{d}{dt}\mathbb{E}[\mu_1^0 + t(\mu_1 - \mu_1^0) - \mu_0^0 - t(\mu_0 - \mu_0^0) - \tau]|_{t=0} = \mathbb{E}[\mu_1 - \mu_1^0] + \mathbb{E}[\mu_0 - \mu_0^0]$$

- But we know, in general, outcome regressions with ML are biased.

Neyman orthogonality

- Check that for AIPW, it satisfies the Nyeman Orthogonality condition.
- For AIPW, $\psi = \mu_1^0(X) - \mu_0^0(X) + \frac{Z}{\pi_0(X)}(Y - \mu_1^0(X)) - \frac{1-Z}{1-\pi_0(X)}(Y - \mu_0^0(X)) - \tau$
- We check terms separately. To simplify the notation, we use Δ_* to denote the deviation from the true value $* = \{\mu_1, \mu_2, \pi\}$.
- For the first outcome regression term:

$$\frac{d}{dt}\mathbb{E}[\mu_1^0(X) + t\Delta_{\mu_1} - \mu_0^0(X) - t\Delta_{\mu_0}]t = 0 = \mathbb{E}[\Delta\mu_1 - \Delta\mu_0]$$

- For the second part:

$$\frac{d}{dt}\mathbb{E}[\frac{Z}{\pi_0(X) + t\Delta_\pi}(Y - \mu_1^0(X) - t\Delta_{\mu_1})]|_{t=0}$$
$$= \mathbb{E}[-\frac{Z\Delta_\pi}{\pi_0^2(X)}(Y - \mu_1^0(X)) - \frac{Z}{\pi_0(X)}\Delta_{\mu_1}]$$

Neyman orthogonality

- Similarly, the third part is

$$\mathbb{E}[\frac{(1-Z)\Delta_\pi}{(1-\pi_0(X))^2}(Y-\mu_0^0(X))-\frac{1-Z}{1-\pi_0(X)}\Delta_{\mu_0}]$$

- Combine terms together according to the way we show AIPW is robust, then it is easy to show each part is unconditionally zero by conditional mean.
- For example, $\mathbb{E}[\Delta\mu_1(1-\frac{Z}{\pi_0(X)})|X]=0$.
- I leave you to check $\mathbb{E}[-\frac{Z\Delta_\pi}{\pi_0^2(X)}(Y-\mu_1^0(X))]=0$.

DML

- Secondly, using high-quality machine learning estimators of nuisance parameters. A sufficient condition in the examples given includes the requirement $n^{1/4}||\hat{\eta} - \eta_0||_{L_2} \approx 0$.
- Third, using sample splitting were nuisance functions are estimated on different data than are used in their evaluation when producing the estimator of the main parameter θ_0. This avoids biases arising from overfitting.
- To be efficient, in practice, we are using cross-fitting:
  1. Split data into $K = 5$ or 10 folds: $I_1, ..., I_K$, with equal size N.
  2. For each k, obtain $\hat{\eta}_n^k$ by using the remaining $k - 1$ folds.
  3. Solve $\hat{\theta}$ by

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{N}\sum_{i:I_k}\psi(W_i;\theta,\hat{\eta}_n^k) = 0$$

AIPW

- Now, Let us go back to show the strong double robustness result of AIPW, follow [Wager, 2024].
- Recall, we want to say that $\hat{\tau}^{AIPW}$ is asymptotically normal even if outcome regression and propensity score converges slowly.
- The variance is $V = Var(\tau(X_i)) + \mathbb{E}[\frac{\sigma_0^2(X_i)}{1-\pi(X_i)}] + \mathbb{E}[\frac{\sigma_1^2(X_i)}{\pi(X_i)}]$.
- Consider the oracle AIPW estimator $\tau^*_{AIPW} = \frac{1}{n}\Gamma_i$, where

$$\mu_1(X_i) - \mu_0(X_i) + Z_i \frac{Y_i - \mu_1(X_i)}{\pi(X_i)} - (1 - Z_i)\frac{Y_i - \mu_0(X_i)}{1 - \pi(X_i)}$$

Proposition

Under STUVA, unconfoundedness and strong overlap, the oracle AIPW estimator has distribution

$$\sqrt{n}(\hat{\tau}^*_{AIPW} - \tau) \to_d N(0, V),$$

where $V = Var(\tau(X_i)) + \mathbb{E}[\frac{\sigma_0^2(X_i)}{1-\pi(X_i)}] + \mathbb{E}[\frac{\sigma_1^2(X_i)}{\pi(X_i)}]$.

AIPW

- To establish strong double robustness, we just WTS that $\hat{\tau}^{AIPW}$ is asymptotically equivalent to oracle AIPW estimator $\hat{\tau}^{*}_{AIPW}$.
- Formally, $\sqrt{n}(\hat{\tau}^{AIPW} - \hat{\tau}^{*}_{AIPW}) \to 0$.
- To obtain this property, you may expect we need cross-fitting, as we did before.
- Say we split data into two halves I_1 and I_2.
- Then $\hat{\tau}_{AIPW} = \frac{1}{n}[|I_1|\hat{\tau}^{I_1} + |I_2|\hat{\tau}^{I_2}]$, and

$$\hat{\tau}^{I_1} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_1^{I_2}(X_i) - \hat{\mu}_0^{I_2}(X_i)) + \frac{Z_i(Y_i - \hat{\mu}_1^{I_2}(X_i)}{\hat{\pi}^{I_2}(X_i)} - \frac{(1-Z_i)(Y_i - \hat{\mu}_0^{I_2}(X_i)}{1-\hat{\pi}^{I_2}(X_i)}]$$

- In other words, $\hat{\tau}^{I_1}$ is a treatment effect estimator on I_1 that uses I_2 to estimate its non-parametric components, and vice-versa.

Theorem (Strong Double Robustness)

Under STUVA, unconfoundedness and strong overlap, cross-fitting, and, with the roles of I_1 and I_2 swapped

$$\frac{1}{|I_1|}\sum_{i\in I_1}(\hat{\mu}_z^{I_2}(X_i)-\mu_z(X_i))^2 = o_p(n^{-2\phi_\mu})$$

$$\frac{1}{|I_1|}\sum_{i\in I_1}(\frac{1}{\hat{\pi}^{I_2}(X_i)}-\frac{1}{\pi(X_i)})^2 = o_p(n^{-2\phi_\pi})$$

for some $\phi_\mu + \phi_\mu \geq 1/2$. Then, strong double robustness holds.

AIPW

- Let $\hat{m}_1^{I_1} = \frac{1}{|I_1|}\sum_{i\in I_1}(\hat{\mu}_1^{I_2}(X_i) + \frac{Z_i(Y_i - \hat{\mu}_1^{I_2}(X_i)}{\hat{\pi}^{I_2}(X_i)})$; similarly for $\hat{m}_0^{I_1}$
- See that $\hat{\tau}^{I_1} = \hat{m}_1^{I_1} - \hat{m}_0^{I_1}$.
- For oracle $\hat{\tau}^*_{AIPW} = \frac{1}{n}[|I_1|\hat{\tau}^{I_1,*} + |I_2|\hat{\tau}^{I_2,*}] = \hat{m}_1^{I_1,*} - \hat{m}_0^{I_1,*}$
- Therefore, we only need to show $\sqrt{n}(\hat{m}_1^{I_1} - \hat{m}_1^{I_1,*}) \to_p 0$.
- Check each term of the difference

$$\begin{aligned}\hat{m}_1^{I_1} - \hat{m}_1^{I_1,*} = &\frac{1}{|I_1|}\sum_{i\in I_1}[(\hat{\mu}_1^{I_2}(X_i) - \mu_1(X_i))(1 - \frac{Z_i}{\pi(X_i)})] \\ &+ \frac{1}{|I_1|}\sum_{i\in I_1} Z_i[(Y_1 - \mu_1(X_i))(\frac{1}{\hat{\pi}^{I_2}(X_i)} - \frac{1}{\pi(X_i)})] \\ &- \frac{1}{|I_1|}\sum_{i\in I_1} Z_i[(\hat{\mu}_1^{I_2}(X_i) - \mu_1(X_i))(\frac{1}{\hat{\pi}^{I_2}(X_i)} - \frac{1}{\pi(X_i)})]\end{aligned}$$

AIPW

- See the first term $\frac{1}{|I_1|}\sum_{i\in I_1}[(\hat{\mu}_1^{I_2}(X_i)-\mu_1(X_i))(1-\frac{Z_i}{\pi(X_i)})]$.
- Thanks to cross-fitting, for $i \in I_1$, $\hat{\mu}_1^{I_2}$ is fixed.
- Conditional on I_2, X_i, this term is the average of mean-zero terms.
- And similarly, we check the variance going to zero, and then use Chebyshev's inequality to show the first term is $o_p(1\sqrt{n})$.
- Let us use the same tower property of expectation:

$$
\begin{aligned}
& Var[\frac{1}{|I_1|}\sum_{i\in I_1}[(\hat{\mu}_1^{I_2}(X_i)-\mu_1(X_i))(1-\frac{Z_i}{\pi(X_i)})]|I_2, X_i] \\
&= \frac{1}{|I_1|^2}\sum_{i\in I_1}\mathbb{E}[(\hat{\mu}_1^{I_2}(X_i)-\mu_1(X_i))^2(1-\frac{Z_i}{\pi(X_i)})^2|I_2, X_i] \\
&= \frac{1}{|I_1|^2}\sum_{i\in I_1}\frac{1-\pi(X_i)}{\pi(X_i)}(\hat{\mu}_1^{I_2}(X_i)-\mu_1(X_i))^2 \\
&\leq o_p(\frac{1}{n^{1+2\phi_\mu}})
\end{aligned}
$$

- For the last term, we use Cauchy-Schwarz inequality.

$$\begin{aligned}
&\frac{1}{|I_1|}\sum_{i\in I_1}[(\hat{\mu}_1^{I_2}(X_i)-\mu_1(X_i))(\frac{1}{\hat{\pi}^{I_2}(X_i)}-\frac{1}{\pi(X_i)})] \\
\leq &\sqrt{\frac{1}{|I_1|}\sum_{i\in I_1}(\hat{\mu}_1^{I_2}(X_i)-\mu_1(X_i))^2}\sqrt{\frac{1}{|I_1|}\sum_{i\in I_1}(\frac{1}{\hat{\pi}^{I_2}(X_i)}-\frac{1}{\pi(X_i)})^2} \\
= &\, o_p(\frac{1}{n^{\phi_\mu+\phi_\pi}})
\end{aligned}$$

- That all we need to let this term is also $o_p(1\sqrt{n})$.

References


Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018).
Double/debiased machine learning for treatment and structural parameters.


Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., and Syrgkanis, V. (2024).
Applied causal inference powered by ml and ai.
arXiv preprint arXiv:2403.02467.


Wager, S. (2024).
Causal inference: A statistical learning approach.