Debiased Machine Learning (DML)

Alexander Quispe

May 28, 2024

Citation

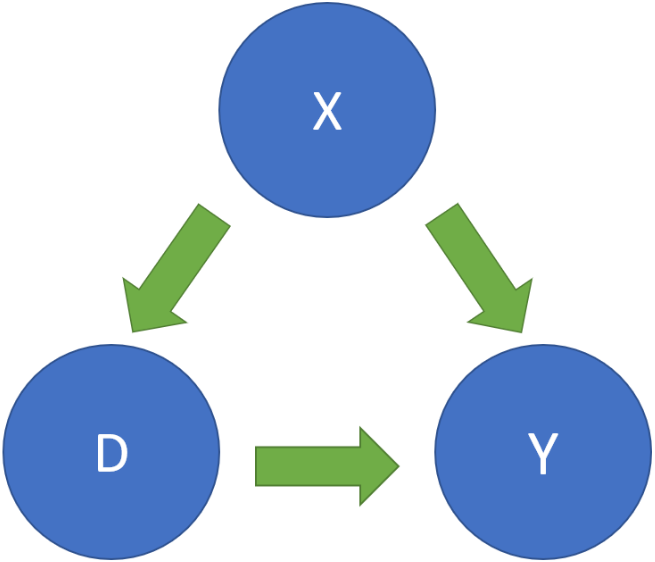
These notes are based on the course 14.38 Inference on Causal and Structural Parameters Using ML and AI at MIT taught by Professor Victor Chernozukhov.

DML - Introduction

In this lecture we discuss debiased machine learning (DML) methods for performing inference on average predictive or causal effects in partially linear regression models.
Main ideas behind the DML method:

1. General framework for estimating causal effects using Machine Learning techniques
2. Confidence intervals for those estimates
3. An estimator that is “root n-consistent”, i.e., an estimator that has good properties in terms of convergence and data-efficiency.

DML - Motivation



PARTIALLY LINEAR REGRESSION

Partially linear regression (PLR) model as in Robinson (1988):

$$Y = \beta D + g(X) + \epsilon, \quad E[\epsilon | D, X] = 0 \tag{1}$$

$$D = h(X) + v, \quad E[v | X] = 0 \tag{2}$$

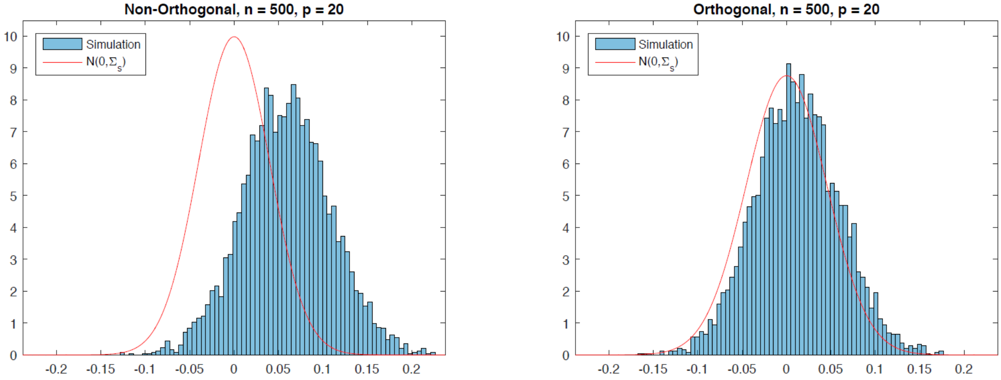
1. Y is the outcome variable
2. D is the regressor of interest
3. X is a high-dimensional vector of other regressors or features, called ”controls.”
4. $g(X)$ is fully nonlinear, but the model is still not fully general because imposes additivity in $g(X)$ and D.

Why not simply use Machine Learning to estimate β?

1. Naive application of machine learning methods directly lead to highly biased estimators.
2. This strategy is not **Neyman-orthogonal**
3. The biases in estimation of g which are unavoidable in high-dimensional estimation, create bias in the estimate of β.
4. This bias is large enough to cause failure of conventional inference.

Lets Experiment!

1. **Left Panel**: We learn g using random forest. g is a very smooth function of a small number of variables.
2. **Right Panel**: We learn g and h using random forest. It shows the behavior of the (Neyman) orthogonal DML estimator.
3. The histogram shows the simulated distribution of the centered estimator $\widetilde{\beta} - \beta$



Lets Experiment!

1. **Left Panel**: Nuisance parameters are estimated with overfitting using the full sample, i.e. without sample splitting.
2. **Right Panel**: the finite-sample distribution of the DML estimator in the partially linear model where nuisance parameters are estimated with sample-splitting using the cross-fitting estimator.

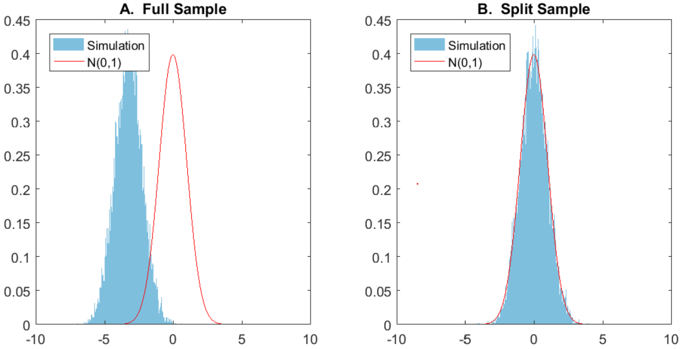


Figure: Left: DML distribution without sample-splitting. Right:DML distribution with crossfitting

DML Inference in PLM

We have our main equation

$$Y = \beta D + g(X) + \epsilon, \quad E[\epsilon | D, X] = 0 \tag{3}$$

we will employ the partialling out X operation, where that inputs a random variable V and outputs residualized form:

$$\tilde{V} := V - E[Y|X] \tag{4}$$

We can apply this to Y and D

$$\tilde{Y} := Y - l(X), \quad \tilde{D} := D - m(X) \tag{5}$$

DML Inference in PLM - FWL

Then by FWL

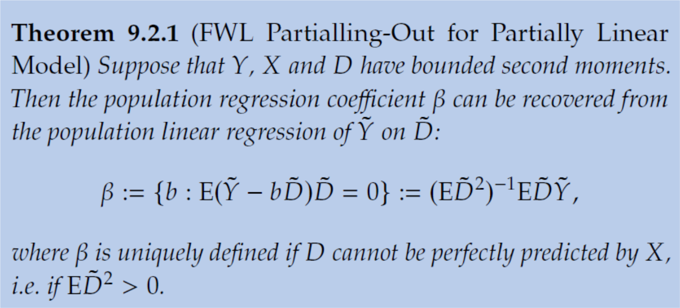
$$\tilde{Y} = \beta\tilde{D} + \epsilon, \;\; E(\epsilon\tilde{D}) = 0 \tag{6}$$

where $l(X)$ and $m(x)$ are defined as conditional expectations of Y and D given X.

$$l(X) := E[Y|X], \;\; m(X) := E[D|X] \tag{7}$$

$E[\epsilon\tilde{D}] = 0$ Is the **Normal Equation** for the population regression of $\tilde{Y}$ on $\tilde{D}$.

DML Inference in PLM - FWL



Theorem 9.2.1 (FWL Partialling-Out for Partially Linear Model) *Suppose that Y, X and D have bounded second moments. Then the population regression coefficient β can be recovered from the population linear regression of $\tilde{Y}$ on $\tilde{D}$:*

$$\beta := \{b : \mathrm{E}(\tilde{Y} - b\tilde{D})\tilde{D} = 0\} := (\mathrm{E}\tilde{D}^2)^{-1}\mathrm{E}\tilde{D}\tilde{Y},$$

where β is uniquely defined if D cannot be perfectly predicted by X, i.e. if $\mathrm{E}\tilde{D}^2 > 0$.

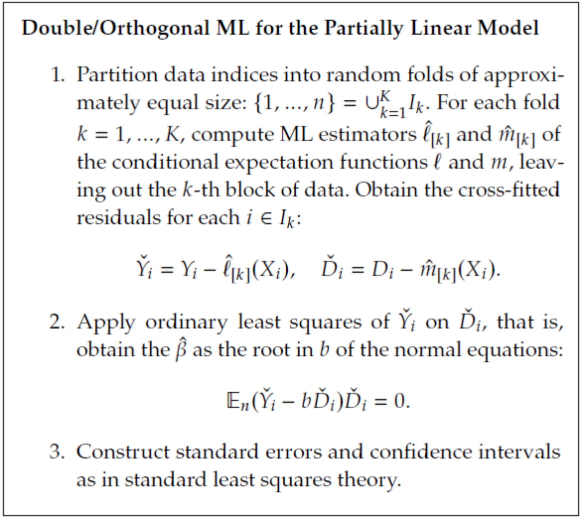
Figure: Applied Causal Inference Powered by ML and AI, Page 186

How to read Beta?

β is a *regression coefficient of residualized* Y *on residualized* D*.*
The residuals are defined as :
1. Y *minus the conditional expectation of* Y *given* X
2. D *minus the conditional expectation of* D *given* X

What about Cross-Fitting?

We need to rely on cross-fitting to make sure the estimated residualized quantities are not overfit.



Double/Orthogonal ML for the Partially Linear Model

1. Partition data indices into random folds of approximately equal size: $\{1, \ldots, n\} = \cup_{k=1}^{K} I_k$. For each fold $k = 1, \ldots, K$, compute ML estimators $\hat{\ell}_{[k]}$ and $\hat{m}_{[k]}$ of the conditional expectation functions ℓ and m, leaving out the k-th block of data. Obtain the cross-fitted residuals for each $i \in I_k$:

$$\check{Y}_i = Y_i - \hat{\ell}_{[k]}(X_i), \quad \check{D}_i = D_i - \hat{m}_{[k]}(X_i).$$

2. Apply ordinary least squares of $\check{Y}_i$ on $\check{D}_i$, that is, obtain the $\hat{\beta}$ as the root in b of the normal equations:

$$\mathbb{E}_n(\check{Y}_i - b\check{D}_i)\check{D}_i = 0.$$

3. Construct standard errors and confidence intervals as in standard least squares theory.

Adaptive Inference on a Target Parameter in PLM

Let´s define

$$||h||_{L^2} = \sqrt{E_X h^2(X)} \tag{8}$$

$$n^{1/4}(||\hat{l}_{[k]} - l||_{L^2} + ||\hat{m}_{[k]} - m||_{L^2}) \approx 0 \tag{9}$$

$$\sqrt{n}(\hat{\beta} - \beta) \approx (\mathbb{E}_n \tilde{D}^2)^{-1} \sqrt{n} \mathbb{E}_n \tilde{D} \epsilon \tag{10}$$

$$\sqrt{n}(\hat{\beta} - \beta) \overset{a}{\sim} N(0, V) \tag{11}$$

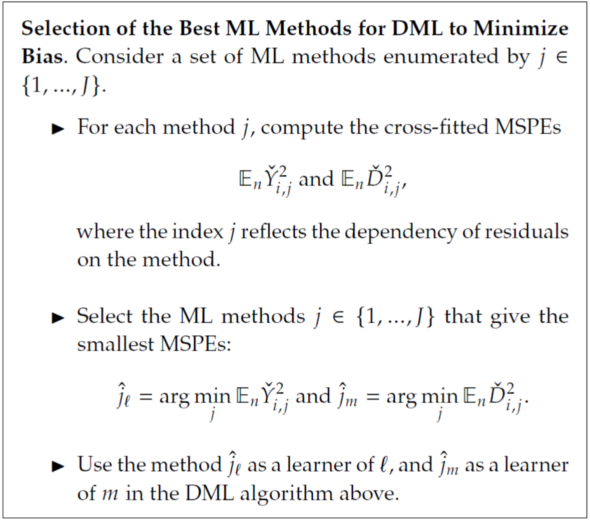
$$V = (E\tilde{D}^2)^{-1} E(\tilde{D}^2 \epsilon^2)(E\tilde{D}^2)^{-1} \tag{12}$$

Confidence interval

$$\left[\hat{\beta} - 2\sqrt{\hat{V}/n}, \hat{\beta} + 2\sqrt{\hat{V}/n}\right] \tag{13}$$

Selection of the Best ML Methods for DML to Minimize Bias.

Therefore, we can select the best ML method for estimating *m* and the best method for estimating *l* to minimize the upper bound on the bias.



Selection of the Best ML Methods for DML to Minimize Bias. Consider a set of ML methods enumerated by $j \in \{1, \ldots, J\}$.

- For each method j, compute the cross-fitted MSPEs

  $$\mathbb{E}_n \check{Y}_{i,j}^2 \text{ and } \mathbb{E}_n \check{D}_{i,j}^2,$$

  where the index j reflects the dependency of residuals on the method.

- Select the ML methods $j \in \{1, \ldots, J\}$ that give the smallest MSPEs:

  $$\hat{j}_\ell = \arg\min_j \mathbb{E}_n \check{Y}_{i,j}^2 \text{ and } \hat{j}_m = \arg\min_j \mathbb{E}_n \check{D}_{i,j}^2.$$

- Use the method $\hat{j}_\ell$ as a learner of ℓ, and $\hat{j}_m$ as a learner of m in the DML algorithm above.