Double/Debiased Machine Learning

Tom Ben-Shahar

Basics

While traditional machine learning (ML) methods perform well in prediction, they tend to struggle with causal inference.

Consider the following partially linear model:

$$Y = \alpha D + g(X) + \epsilon$$

$$D = m(X) + v$$

where

Y **:** Outcome of interest

D **:** Treatment or policy variable of interest

X **:** High-dimensional vector of control variables

α **:** Parameter targeted for estimation (treatment effect)

ϵ, v **:** Error terms

m, g **:** Unknown "nuisance" functions

Naive Approach

To estimate the causal parameter α, the given sample of data can be split into two smaller samples, which we can call the **main sample** and the **auxiliary sample**. We then use the auxiliary sample to train a simple ML model (e.g. random forest) to estimate the nuisance function $g(X)$. Given the estimated function $\hat{g}(X)$, we can use the main sample to estimate the parameter α such that

$$\hat{\alpha} = \left(\frac{1}{n} \sum_{i \in I_{main}} D_i^2 \right)^{-1} \frac{1}{n} \sum_{i \in I_{main}} D_i (Y_i - \hat{g}(X_i)) \tag{1}$$

However, this naive approach introduces **regularization bias**. Put simply, the regressor D is correlated with the control variables X. This causes the the estimated effect of D (which is $\hat{\alpha}$) to be biased by the correlated effect of X. To solve this, we can decorrelate D from X . We call this step **orthogonalization**.

Regularization Bias

Bias arises due to the slow convergence of $\hat{\alpha}$ relative to $\sqrt{n}$. This means that as n approaches infinity, $\hat{\alpha}$ diverges rather than converging to the true value α.

$$|\sqrt{n}(\hat{\alpha} - \alpha)| \to_p \infty$$

A way to conceptualize how this bias affects our estimate is to consider a **ridge regression**, in which large coefficients are penalized to prevent overfitting. Such a method is excellent for prediction, but the downward bias presented by the penalty on large coefficients causes biased estimates that are generally invalid for causal inference. So, if our true parameter value is, say, $\alpha = 3$, then while OLS may predict $\hat{\alpha} = 2.99$, a ridge regression may predict $\hat{\alpha} = 2.5$ regularization biases the estimate downward.

Orthogonalization

Similarly to the naive approach, we can begin by splitting our data into main and auxiliary samples. We can then use the auxiliary sample to estimate both $\hat{g}(X)$ and $\hat{m}(X)$. This is again done using a simple ML model. Given the estimated function $\hat{m}(X)$, we can use the main sample to estimate the **orthogonalized component** of D by

$$\hat{v} = D - \hat{m}(X)$$

This is the component of D that is not biased by correlation with X. We can then use the main sample given the estimated $\hat{v}$ to better estimate the parameter α such that

$$\hat{\alpha} = \left(\frac{1}{n} \sum_{i \in I_{main}} \hat{v}_i D_i \right)^{-1} \frac{1}{n} \sum_{i \in I_{main}} \hat{v}_i (Y_i - \hat{g}(X_i))$$

The estimator is now **root-N consistent**, meaning the estimate $\hat{\alpha}$ converges to the true parameter α as n grows towards infinity.

Put more simply, we can use our auxiliary sample to predict $\hat{Y}$ and $\hat{D}$ over X with an ML method like random forests. We can then take the residuals

$$\hat{\epsilon} = Y - \hat{Y}$$

$$\hat{v} = D - \hat{D}$$

and regress $\hat{\epsilon}$ on $\hat{v}$ to get an **unbiased estimate of the causal parameter of interest** $\hat{\alpha}$.