Foundations of machine learning

Debiased Machine Learning

Maximilian Kasy

Department of Economics, University of Oxford

Hilary term 2022

Outline

- Supervised machine learning as a first stage estimator in econometrics.
- Two problems that arise using a plugin approach.
- Two solutions - orthogonalized scores and sample splitting.
- How to derive orthogonalized scores.
- Examples.
- Asymptotics.

Takeaways for this part of class

- Supervised learning can be useful as a first-stage
  in econometric estimation problems.
- But simple plug-in estimators are often poorly behaved.
- Well-behaved estimators can be constructed using
  1. Orthogonal scores, and
  2. Sample splitting and averaging.
- Examples:
  1. Partial linear regression.
  2. Average treatment effect und unconfoundedness.
  3. Local average treatment effect under conditional instrument exogeneity.

Setup

- Many settings in econometrics:
  - The object of interest is low-dimensional (or real-valued),
  - but high-dimensional parameters are of intermediate relevance.
- General two stage structure:
  1. The high-dimensional g_0 is given by the solution to some supervised learning problem.
  2. The low-dimensional parameter of interest θ_0 then solves

$$E[\phi(W, \theta_0, g_0)] = 0.$$

- Can we estimate g_0 using supervised machine learning, and plug it in?

Plugin estimation

- Most obvious estimator of θ_0:
  1. First estimate g_0 using some supervised ML method.
  2. Then plug in the estimate and solve for $\hat{\theta}$ in

$$E_n\left[\phi(W_i, \hat{\theta}, \hat{g})\right] = 0.$$

- This causes two **problems**, however:
  1. Bias of $\hat{g}$ might distort $\hat{\theta}_0$.
  2. The statistical dependence of $\hat{g}$ and W_i might distort $\hat{\theta}_0$.
- Both of these issues might cause large biases.
- Let us consider some examples, before solving these problems.

Example 1: Partially linear regression

- Model:

$$Y = D \cdot \theta_0 + g_0(X) + U, \qquad\qquad E[U|X,D] = 0.$$

- Plugin estimator:
  1. Estimate g_0, using some supervised ML method.
  2. Then solve $E_n[\phi(W_i, \theta_0, \hat{g})] = 0$, where E_n is the sample average across observations W_i, and

$$\phi(W, \theta, g) = (Y - D \cdot \theta - g(X)) \cdot D,$$

- Thus

$$\hat{\theta} = E_n\left[D_i^2\right]^{-1} \cdot E_n[D_i \cdot (Y_i - g(X_i))]$$

Example 2: Average treatment effect

- Model:

$$Y = g_0(D, X) + U \qquad\qquad E[U|X, D] = 0$$
$$\theta_0 = E[g_0(1, X) - g_0(0, X)].$$

- Under unconfoundedness, θ_0 is the average treatment effect.
- Plugin estimator:
  1. Estimate g_0, using some supervised ML method.
  2. Then solve $E_n[\phi(W_i, \theta_0, \hat{g})] = 0$, where

$$\phi(W, \theta, g) = g(1, X) - g(0, X) - \theta.$$

Example 3: Local average treatment effect

- Model:

$$Y = g_0^y(Z,X) + U, \quad D = g_0^d(Z,X) + V, \quad E[(U,V)|X,D] = 0,$$
$$\theta_0 = \frac{E[g_0^y(1,X) - g_0^y(0,X)]}{E[g_0^d(1,X) - g_0^d(0,X)]}.$$

- Under conditional instrument exogeneity, exclusion restriction, θ_0 is the local average treatment effect.
- Plugin estimator:
  1. Estimate g_0, using some supervised ML method.
  2. Then solve $E_n[\phi(W_i, \theta_0, \hat{g})] = 0$, where

$$\phi(W,\theta,g) = g^y(1,X) - g^y(0,X) - \left(g^d(1,X) - g^d(0,X)\right)\cdot\theta.$$

Approximating $\hat{\theta}$

- Telescope sum; Taylor approximation;
  approximating sample averages by expectations:

$$
\begin{aligned}
0 = E_n\left[\phi(W_i,\hat{\theta},\hat{g})\right] &= E_n\left[\phi(W_i,\hat{\theta},\hat{g}) - \phi(W_i,\hat{\theta},g_0)\right] \\
&\quad + E_n\left[\phi(W_i,\hat{\theta},g_0) - \phi(W_i,\theta_0,g_0)\right] + E_n\left[\phi(W_i,\theta_0,g_0)\right] \\
&\approx E\left[\partial_g \phi(W_i,\theta_0,g_0)\cdot(\hat{g}-g_0)\right] \\
&\quad + E\left[\partial_\theta \phi(W_i,\theta_0,g_0)\right]\cdot(\hat{\theta}-\theta_0) + E_n\left[\phi(W_i,\theta_0,g_0)\right].
\end{aligned}
$$

- Solving for $\hat{\theta}-\theta_0$:

$$
\begin{aligned}
(\hat{\theta}-\theta_0) \approx E\left[\partial_\theta \phi(W_i,\theta_0,g_0)\right]^{-1}\cdot \big[E_n\left[\phi(W_i,\theta_0,g_0)\right] + \\
+ E\left[\partial_g \phi(W_i,\theta_0,g_0)\cdot(\hat{g}-g_0)\right]\big]
\end{aligned}
$$

- We can further decompose the last term, which is the cause of bias:

$$
\begin{aligned}
&E\left[\partial_g \phi(W_i,\theta_0,g_0)\cdot(\hat{g}-g_0)\right] \\
=&E\left[\partial_g \phi(W_i,\theta_0,g_0)\right]\cdot(E[\hat{g}]-g_0) + E\left[\partial_g \phi(W_i,\theta_0,g_0)\cdot(\hat{g}-E[\hat{g}])\right]
\end{aligned}
$$

Practice problem

Write out this decomposition for average treatment effect estimation and the plugin estimator.

1. Recall what is ϕ and $\boldsymbol{g}$ here.
2. What is $\partial_{\boldsymbol{\theta}}\phi$, what is $\partial_{\boldsymbol{g}}\phi$?
3. What do we get for the red and magenta terms?

Problem 1: Bias in the first stage

- As we discussed previously, ML estimators use regularization, and therefore are **biased**: $E[\hat{g}] \neq g_0$.
- Suppose however that we had a score function which satisfies "**Neyman orthogonality**:"
$$E\left[\partial_g \phi(W_i, \theta_0, g_0)\right] = 0.$$
- Then
$$E\left[\partial_g \phi(W_i, \theta_0, g_0)\right] \cdot \left(E[\hat{g}] - g_0\right) = 0.$$
- $\Rightarrow$ Bias of $\hat{g}$ does not matter to first order.

Problem 2: Statistical dependence of first stage and data

- In general, W_i and $\hat{g}$ are not statistically independent, and $\hat{g}$ has non-negligible **variance**.
- Therefore $E\left[\partial_g \phi(W_i, \theta_0, g_0) \cdot (\hat{g} - E[\hat{g}])\right] \neq 0$.
- Suppose however we used sample splitting:
  1. Estimate $\hat{g}$ on one part of the data.
  2. Average $\phi(W_i, \hat{\theta}, \hat{g})$ over the remaining data.
- Then this term automatically vanishes!

Debiased Machine Learning

Combining these two ideas: (*Definition 3.2 in the paper.*)

1. Start with an estimation problem of the form $E[\phi(W,\theta_0,g_0)] = 0$.

2. Derive an orthogonal Neyman score ψ, which satisfies

$$
\begin{aligned}
E[\psi(W,\theta_0,\eta_0)] &= 0, \\
E\left[\partial_\eta \psi(W_i,\theta_0,\eta_0)\right] &= 0.
\end{aligned}
$$

   We will discuss next how to do this.

3. Split the sample into K subsamples I_k.
   Estimate $\hat{\eta}_k$ based on I_k^c. Denote $E_{n,k}$ the sample average over I_k.

4. Estimate θ by solving

$$
\sum_{k=1}^{k} E_{n,k}\left[\psi(W,\hat{\theta},\hat{\eta}_k)\right] = 0.
$$

How to derive orthogonal scores

- Suppose that
$$(\theta_0, \beta_0) = \underset{\theta,\beta}{\operatorname{argmax}}\ E[L(W, \theta, \beta)].$$
- β takes the role of g here.
  We focus on the parametric case for ease of exposition.
- Two approaches to deriving an orthogonal score:
  1. Construction from moment functions.
  2. Concentrating out.

Construction from moment functions

- Suppose that
$$(\theta_0, \beta_0) = \underset{\theta,\beta}{\operatorname{argmax}}\ E[L(W, \theta, \beta)],$$
and thus
$$E[\partial_\theta L(W, \theta_0, \beta_0)] = 0, \qquad E[\partial_\eta L(W, \theta_0, \beta_0)] = 0.$$
- Define
$$\psi(W, \theta, \eta) = \partial_\theta L(W, \theta, \beta) - \mu \cdot \partial_\beta L(W, \theta, \beta),$$
where $\eta = (\mu, \beta)$, and μ_0 solves
$$\partial_\beta E[\partial_\theta L(W, \theta_0, \beta_0)] - \mu_0 \cdot \partial_\beta E[\partial_\beta L(W, \theta_0, \beta_0)] = 0.$$
- Then
$$E[\psi(W, \theta_0, \eta_0)] = 0,$$
$$E\,[\partial_\eta \psi(W_i, \theta_0, \eta_0)] = 0.$$

Construction by concentrating out

- Suppose again that

$$(\theta_0, \eta_0) = \underset{\theta,\beta}{\operatorname{argmax}}\ E[L(W, \theta, \beta)].$$

- Define

$$\begin{aligned}
\beta(\theta) &= \underset{\beta}{\operatorname{argmax}}\ E[L(W, \theta, \beta)], \\
\psi(W, \theta, \eta) &= \partial_\theta \left(L(W, \theta, \beta(\theta)) \right) \\
&= \partial_\theta L(W, \theta, \beta) + \partial_\theta \beta(\theta) \cdot \partial_\beta L(W, \theta, \beta),
\end{aligned}$$

where $\eta = (\beta, \partial_\theta \beta(\theta))$.

- Then, again

$$\begin{aligned}
E[\psi(W, \theta_0, \eta_0)] &= 0, \\
E\left[\partial_\eta \psi(W_i, \theta_0, \eta_0)\right] &= 0.
\end{aligned}$$

Example 1: Partially linear regression

- Recall the model

$$Y = D \cdot \theta_0 + g_0(X) + U, \qquad\qquad E[U|X,D] = 0.$$

- Define

$$m_0(X) = E[D|X].$$

- Then

$$\psi(W, \theta, \eta) = (Y - D \cdot \theta + g(X)) \cdot (D - m(X))$$

  satisfies the orthogonality condition.

- In the first stage, we need to estimate $g_0(X)$ and $m(X)$.

Example 2: Average treatment effect

- Recall the model

$$Y = g_0(D,X) + U \qquad\qquad E[U|X,D] = 0$$
$$\theta_0 = E[g_0(1,X) - g_0(0,X)].$$

- Define

$$m_0(X) = E[D|X].$$

- Then

$$\psi(W,\theta,\eta) = (g(1,X) - g(0,X)) + \left(\frac{DY}{m(X)} - \frac{(1-D)Y}{1-m(X)}\right)$$
$$- \left(\frac{Dg(1,X)}{m(X)} - \frac{(1-D)g(0,X)}{1-m(X)}\right) - \theta$$

satisfies the orthogonality condition.

- This is the famous “doubly robust” estimation approach.

Asymptotics for debiased ML estimators

Theorem 3.3.

- Assume a number of regularity conditions.
- Consider a Debiased Machine Learning estimator.
- Then

$$\sqrt{n}(\hat{\theta} - \theta) \sim^{A} N(0, \sigma^2),$$

- where

$$\sigma^2 = J^{-1} \cdot \mathrm{Var}(\psi(W, \theta_0, \eta_0)) \cdot J^{-1},$$

  for

$$J = \partial_\theta E[\psi(W, \theta_0, \eta_0)].$$

Intuition of proof

- Recall our earlier expansion

$$(\hat{\theta} - \theta_0) \approx E\left[\partial_\theta \psi(W_i, \theta_0, \eta_0)\right]^{-1} \cdot \Big[E_n\left[\psi(W_i, \theta_0, \eta_0)\right] + \\ + E\left[\partial_\eta \psi(W_i, \theta_0, \eta_0) \cdot (\hat{\eta} - \eta_0)\right]\Big].$$

- Using the Debiased Machine Learning approach, we have killed the blue term.
- The other terms give asymptotic normality and the variance by standard arguments.

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. The Econometrics Journal, *21(1):C1–C68.*