Notes on Double Machine Learning (for Applied Research)

Jiding Zhang

1 The Omitted Variable Bias

The true model ("long regression"):

$$Y = X_1\beta_1 + X_2\beta_2 + u.$$

Consider estimating the following "short regression" model (omitting X_1) instead:

$$Y = X_2\beta_2 + v \quad \text{where we do the projection: } X_2 = \delta X_1 + \varepsilon.$$

then $\hat{\beta}_2$ is biased if $\delta \neq 0$, i.e., here the result of the "short regression" is not right!

2 The FWL Theorem: The "Correct" Short Regression

Consider the model:

$$Y = X_1\beta_1 + X_2\beta_2 + u.$$

We define the residual maker matrix M_X as (projecting on X and taking the residuals):

$$M_X = I - X(X'X)^{-1}X'$$

and applying the matrix to both sides of the equation gives us:

$$M_{X_1}Y = M_{X_1}X_1\beta_1 + M_{X_1}X_2\beta_2 + M_{X_1}u$$

which simplifies to the short regression using the residuals:

$$\begin{aligned}\underbrace{M_{X_1}Y}_{\text{projecting } Y \text{ on } X_1 \text{ and taking the residuals}} &= (X_1\beta_1 - X_1\beta_1) + M_{X_1}X_2\beta_2 + M_{X_1}u \\ &= \underbrace{M_{X_1}X_2}_{\text{projecting } X_2 \text{ on } X_1 \text{ and taking the residuals}}\beta_2 + M_{X_1}u\end{aligned}$$

i.e., the modified "short regression" is correct.

<u>Note</u>, if we extend the linear form to allow for the more flexible "taking the residual" approach, we have:

$$Y - \mathbb{E}[Y|X_1] = \beta_2(X_2 - \mathbb{E}[X_2|X_1]) + \varepsilon.$$

3 The Partial Linear Case in Chernozhukov et al. [2018]

Consider outcome Y is generated from treatment variable D (*exact!*) and some high-dimensional confounder X. The goal is to estimate θ_0 (effect of treatment D). With high-dimensional X (potentially many confounders), we cannot include them all linearly (omitting them: OVB if X correlated with D). We assume confounder X affects outcome variable (and treatment variable) through nuisance functions (for variable selection, can estimate nuisance function through Lasso, see Belloni et al. [2014]).

The model is given by:

$$Y = D\theta_0 + g_0(X) + U \quad \text{with} \quad \mathbb{E}[U|X, D] = 0.$$

Note, there is no endogeneity here!

The naive approach: estimate g_0 from a subsample and obtain $\hat{g}_0$, then do the regression on the plug-in estimator $\hat{g}_0(X_i)$:

$$\hat{\theta}_0 = \left(\frac{1}{n}\sum D_i^2\right)^{-1}\left(\frac{1}{n}\sum D_i(Y_i - \hat{g}_0(X_i))\right).$$

This is our OLS estimator $(X'X)^{-1}X'y$ but here in y we substract from the estimated nuisance prarmeter.

The convergence of the estimator is given by:

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = \left(\frac{1}{n}\sum D_i^2\right)^{-1}\frac{1}{\sqrt{n}}\sum D_iU_i + \left(\frac{1}{n}\sum D_i^2\right)^{-1}\frac{1}{\sqrt{n}}\sum D_i(g_0(X_i) - \hat{g}_0(X_i))$$

where the second term on the right converges slower than $n^{-\frac{1}{2}}$ (see, e.g., Farrell et al. [2021] for the convergence rate if g_0 is estimated through neural network).

The problem can be solved by orthogonalization. Now, consider how confounder X affects the treatment variable D. We assume in the true model (recall, D should be correlated with X for the OVB to matter):

$$D = m_0(X) + V \quad \text{with} \quad \mathbb{E}[V|X] = 0.$$

We partial out the effect of X from D by taking

$$\hat{V} = D - \hat{m}_0(X)$$

where $\hat{m}_0(X)$ is a machine learning estimator of m_0. The DML estimator θ_0 is given by (using the main sample):

$$\hat{\theta}_0 = \left(\frac{1}{n}\sum \hat{V}_iD_i\right)^{-1}\left(\frac{1}{n}\sum \hat{V}_i(Y_i - \hat{g}_0(X_i))\right).$$

Note, this is an analog of FWL but not exactly the same (e.g., $g_0 \neq \mathbb{E}(Y|X)$ and $Y_i - \hat{g}_0(X_i)$ is not exactly the residual of Y projected on X). A closer analog is considered in Chernozhukov et al. [2018]:

$$\hat{\theta}_0 = \left(\frac{1}{n}\sum \hat{V}_i\hat{V}_i\right)^{-1}\left(\frac{1}{n}\sum \hat{V}_i(Y_i - \widehat{\mathbb{E}(Y_i|X_i)})\right).$$

To estimate $\hat{m}_0$ and $\hat{g}_0$, we split the sample and use the auxiliary sample for estimation.

4 Constructing Neyman Orthogonality

4.1 Notations (General Case)

- The score function:
$$\psi(W;\theta,\eta) \quad \text{s.t.} \quad \mathbb{E}_P[\psi(W;\theta_0,\eta_0)] = 0 \tag{1}$$

- The Gateaux (pathwise) derivative:
$$D_r[\eta-\eta_0] := \partial_r \left\{\mathbb{E}_P[\psi(W;\theta_0,\eta_0+r(\eta-\eta_0))]\right\}$$
for all $r \in [0,1)$ and denote
$$\partial_\eta \mathbb{E}_P \psi(W;\theta_0,\eta_0)[\eta-\eta_0] := D_0[\eta-\eta_0]$$

- Neyman orthogonality (score function should be robust to small perturbations in the nuisance function):
$$\partial_\eta \mathbb{E}_P \psi(W;\theta_0,\eta_0)[\eta-\eta_0] = 0 \quad \forall \eta.$$

Rough idea on why Neyman orthorgonal score matters (Theorem 3.1): If you estimate the target parameter from a score function that satisfies Neyman orthogonality, you get the correct convergence rate!

In the GMM case:

- moment condition:
$$\mathbb{E}_P\left[m(W;\theta_0,h_0(Z))|R\right] = 0$$

- W: (all) data/observation

- R: conditions in moments (subvector of W)

- Z: nuisance vectors (subvector of R, e.g., high-dim confounders) with true nuisance function h_0

- A: arbitrary moment selection function

- Ω: weighting function on moments

- μ: a functional parameter with the true value $\mu_0(R)$ is given by:
$$\mu_0(R) = A(R)'\Omega(R)^{-1} - G(Z)\Gamma(R)'\Omega(R)^{-1}$$
where
$$\Gamma(R) = \partial_{v'} \mathbb{E}_P[m(W;\theta_0,v)|R]|_{v=h_0(Z)}$$
$$G(Z) = \mathbb{E}_P[A(R)'\Omega(R)^{-1}\Gamma(R)|Z] \times \left(\mathbb{E}_P[\Gamma(R)'\Omega(R)^{-1}\Gamma(R)|Z]\right)^{-1}$$

Constructing the Neyman orthogonal score (Lemma 2.6): In this case, the Neyman orthogonal score is:
$$\psi(W;\theta,\eta) = \mu(R)m(W;\theta,h(Z))$$

4.2 The Partial Linear Case (Corollary of Lemma 2.6)

The partial linear model moment condition is:

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

It's a special case of GMM where we pick: $W = (Y, D, X), R = (D, X), Z = X, h(Z) = g(X)$, $A(R) = -D$, $\Omega(R) = 1$, and the moment function is

$$m(W; \theta, v) = Y - D\theta - v.$$

So we can derive the score function:

$$\begin{aligned}
\Gamma(R) &= \partial_{v'}\mathbb{E}_P[m(W; \theta_0, v)|R]|_{v=h_0(Z)} = \partial_{v'}\mathbb{E}_P[Y - D\theta - v|D, X]|_{v=g_0(X)} = -1 \\
G(Z) &= \mathbb{E}_P[A(R)'\Omega(R)^{-1}\Gamma(R)|Z] \times \left(\mathbb{E}_P[\Gamma(R)'\Omega(R)^{-1}\Gamma(R)|Z]\right)^{-1} \\
&= \mathbb{E}_P[(-D)' \times 1 \times (-1)|X] \times (\mathbb{E}_P[(-1) \times 1 \times (-1)|Z])^{-1} = \mathbb{E}_P(D|X) \\
\mu(R) &= A(R)'\Omega(R)^{-1} - G(Z)\Gamma(R)'\Omega(R)^{-1} = (-D) \times 1 - \mathbb{E}_P(D|X) \times (-1) \times 1 = -D + \mathbb{E}_P(D|X)
\end{aligned}$$

Hence,

$$\psi(W; \theta, \eta) = \mu(R)m(W; \theta, h(Z)) = (-D + \underbrace{\mathbb{E}_P(D|X)}_{m_0(X)})(Y - D\theta - g_0(X)).$$

Flipping the sign, we have

$$(D - m_0(X))(Y - D\theta - g_0(X)).$$

Note, this looks like the moment condition for IV!

Next, we prove the score function satisfies (A) condition (1); and (B) the Neyman orthogonality condition.

To show (A), we need to show

$$\begin{aligned}
&\mathbb{E}_P\left[\left(D - \mathbb{E}_P[D|X]\right)\left(Y - D\theta_0 - g_0(X)\right)\right] = 0 \\
\Leftrightarrow &\mathbb{E}_P\left\{\Big(D - \mathbb{E}_P[D|X]\Big)\underbrace{\mathbb{E}_P^{D,X}\Big[Y - D\theta_0 - g_0(X)|D, X\Big]}_{=0}\right\} = 0.
\end{aligned}$$

Note we use the law of iterated expectation (similar to the IV case).

To show (B), note that $\eta = (\mu, h)$, i.e., two nuisance functions:

$$\mathbb{E}_P[\psi(W; \theta_0, \eta_0) + r(\eta - \eta_0)] = \mathbb{E}_P\left\{\Big[\mu_0(R) + r(\mu(R) - \mu_0(R))\Big]m\Big(W; \theta_0, h_0(Z) + r(h(Z) - h_0(Z))\Big)\right\}.$$

Define

$$\begin{aligned}
I_1 &= \mathbb{E}_P\left[(\mu(R) - \mu_0(R))m(W, \theta_0, h_0(Z))\right], \\
I_2 &= \mathbb{E}_P\left[\mu_0(R)\partial_{v'}m(W, \theta_0, v)|_{v=h_0(Z)}(h(Z) - h_0(Z))\right]
\end{aligned}$$

and

$$\partial_\eta\mathbb{E}_P\psi(W, \theta_0, \eta_0)[\eta - \eta_0] = I_1 + I_2$$

where

- I_1 corresponds to the derivative with respect to μ at $r = 0$,
- I_2 corresponds to the derivative with respect to h at $r = 0$.

We note that $I_1 = 0$ due to iterative expectation, and (see p. 55 of the paper for details on the last equality)

$$\begin{aligned} I_2 &= \mathbb{E}_P\left[\mu_0(R)\mathbb{E}_P^X\left[\partial_{v'}m(W,\theta_0,v)|_{v=h_0(Z)}|X\right](h(Z)-h_0(Z))\right] \\ &= \mathbb{E}_P\left[\mu_0(R)\Gamma(R)(h(Z)-h_0(Z))\right] \\ &= \mathbb{E}_P\left[\mathbb{E}_P^Z\left[\mu_0(R)\Gamma(R)|Z\right](h(Z)-h_0(Z))\right] \\ &= 0. \end{aligned}$$

5 Other Remarks

5.1 Intuition: The Estimator

How to get the estimator from $\mathbb{E}[(D - m_0(X))(Y - D\theta - g_0(X))] = 0$? Consider $\hat{\beta}_{IV} = (Z'X)^{-1}Z'y$. Here $Z := D - m_0(X)$, and $y = Y - g_0(X)$.

5.2 Endogeneity

Consider the model

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

We set: $W = (Y, D, X, Z), R = (X, Z), Z = X,\ A(R) = -Z,\ \Omega(R) = 1$, and $m(W; \theta_0, v) = Y - D\theta_0 - v$.

$$\begin{aligned} \Gamma(R) &= \partial_{v'}\mathbb{E}_P[m(W;\theta_0,v)|R]|_{v=h_0(Z)} = \partial_{v'}\mathbb{E}_P[Y - D\theta - v|X,Z,D]|_{v=g_0(X)} = -1 \\ G(Z) &= \mathbb{E}_P[A(R)'\Omega(R)^{-1}\Gamma(R)|Z] \times \left(\mathbb{E}_P[\Gamma(R)'\Omega(R)^{-1}\Gamma(R)|Z]\right)^{-1} \\ &= \mathbb{E}_P[(-Z)' \times 1 \times (-1)|X] \times \left(\mathbb{E}_P[(-1) \times 1 \times (-1)|X]\right)^{-1} = \mathbb{E}_P(Z|X) \\ \mu(R) &= A(R)'\Omega(R)^{-1} - G(Z)\Gamma(R)'\Omega(R)^{-1} = (-Z) \times 1 - \mathbb{E}_P(Z|X) \times (-1) \times 1 = -Z + \mathbb{E}_P(Z|X) \end{aligned}$$

Hence we have the condition:

$$\mathbb{E}_P[(Z - m_0(X))(Y - D\theta_0 - g_0(X))] = 0.$$

References

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650, 2014.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.