Double Machine Learning

Vira Semenova

The partially linear model

$$Y = D\theta_0 + g_0(x) + U, \quad \mathbb{E}[U|D,Z] = 0, \quad \dim(\theta_0) = 1$$

The partially linear model

$$Y = D\theta_0 + g_0(x) + U, \quad \mathbb{E}[U|D, Z] = 0, \quad \dim(\theta_0) = 1$$

- Y: outcome
- $D \in \mathrm{R}$: treatment/policy variable
- Z: controls
- θ_0: the target parameter - true treatment effect

The partially linear model

$$Y = D\theta_0 + g_0(x) + U, \quad \mathbb{E}[U|D, Z] = 0, \quad \dim(\theta_0) = 1$$

- Y: outcome
- $D \in \mathrm{R}$: treatment/policy variable
- Z: controls
- θ_0: the target parameter - true treatment effect

Major difficulty: vector Z is high-dimensional: $\dim(x) \gg n$

The partially linear model

$$Y = D\theta_0 + g_0(x) + U, \quad \mathbb{E}[U|D,Z] = 0, \quad \dim(\theta_0) = 1$$

- Y: outcome
- $D \in \mathrm{R}$: treatment/policy variable
- Z: controls
- θ_0: the target parameter - true treatment effect

Major difficulty: vector Z is high-dimensional: $\dim(x) \gg n$

Youtube link to Victor Chernozhukov presentation (2016)

Code for Double ML is here

The partially linear model, intro

- data - a collection of random variables observed in the data set. Example: data $= (D, Z, Y)$. Our sample is $(D_i, X_i, Y_i)_{i=1}^n$.
- The target parameter θ (i.e., a number/vector) whose true value is θ_0. Goal is to derive $\check{\theta}$ so that
$$\sqrt{n}(\check{\theta} - \theta_0) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(\text{data}_i) + o_P(1) \Rightarrow^d N(0, \sigma_\psi^2)$$
Example: θ_0 - the true treatment effect in the partially linear model
- the nuisance parameter - unknown parameters/functions in the models. Example: $g_0(x)$
- auxiliary sample $\mathcal{A}_n$
- main sample $\{1, 2, \ldots, n\}$

Quiz

$$Y = D\theta_0 + Z'\gamma_0 + U, \quad \mathbb{E}[U|D,Z] = 0$$

Quiz

$$Y = D\theta_0 + Z'\gamma_0 + U, \quad \mathbb{E}[U|D,Z] = 0$$

What is an appropriate notion of oracle in the partially linear model? In the partially linear model, oracle knows

(a) θ_0 but not γ_0

(b) γ_0 but not θ_0

(c) the set of relevant controls $T = \{j : \gamma_{0,j} \neq 0\}$ in Z but neither γ_0 nor θ_0

(d) θ_0 and γ_0 but not the distribution of U

(e) θ_0, γ_0, and the distribution of U

Quiz, discussion

What is an appropriate notion of oracle in the partially linear model? In the partially linear model, oracle knows

(a) θ_0 but not γ_0

(b) $\boxed{\gamma_0 \text{ but not } \theta_0.}$

(c) $\boxed{\text{the set of relevant controls } T = \{j : \gamma_{0,j} \neq 0\} \text{ in } Z \text{ but neither } \gamma_0 \text{ nor } \theta_0}$

(d) θ_0 and γ_0 but not the distribution of U

(e) θ_0, γ_0, and the distribution of U

Outline

- Frish-Waugh-Lowell
- Double Lasso. Double selection
- Double Machine Learning
- Newey (1994) rule
- Double robustness

Frish-Waugh-Lowell theorem

Long regression coef. $\widetilde{\theta}$ on

$$Y_i = D_i\theta_0 + (X_i)'_T(\gamma_0)_T + U_i$$

is equivalent to residual-on-residual regression

Frish-Waugh-Lowell theorem

Long regression coef. $\widetilde{\theta}$ on

$$Y_i = D_i\theta_0 + (X_i)'_T(\gamma_0)_T + U_i$$

is equivalent to residual-on-residual regression

1. Treatment least squares regression

$$\widehat{\delta} = \arg\min_{\delta} \frac{1}{n}\sum_{i=1}^{n}(D_i - X_i'\delta)^2$$

First-stage residual is $\widehat{V}_i = D_i - X_i'\widehat{\delta}$

Frish-Waugh-Lowell theorem

Long regression coef. $\widetilde{\theta}$ on

$$Y_i = D_i\theta_0 + (X_i)'_T(\gamma_0)_T + U_i$$

is equivalent to residual-on-residual regression

1. Treatment least squares regression

$$\widehat{\delta} = \arg\min_\delta \frac{1}{n}\sum_{i=1}^n (D_i - X_i'\delta)^2$$

First-stage residual is $\widehat{V}_i = D_i - X_i'\widehat{\delta}$

2. Outcome least squares regression

$$\widehat{\rho} = \arg\min_\rho \frac{1}{n}\sum_{i=1}^n (Y_i - X_i'\rho)^2$$

Second-stage residual is $\widehat{W_i} = Y_i - X_i'\widehat{\rho}$

Frish-Waugh-Lowell theorem

Long regression coef. $\widetilde{\theta}$ on

$$Y_i = D_i\theta_0 + (X_i)'_T(\gamma_0)_T + U_i$$

is equivalent to residual-on-residual regression

1. Treatment least squares regression
$$\widehat{\delta} = \arg\min_{\delta} \frac{1}{n}\sum_{i=1}^{n}(D_i - X_i'\delta)^2$$
First-stage residual is $\widehat{V}_i = D_i - X_i'\widehat{\delta}$
2. Outcome least squares regression
$$\widehat{\rho} = \arg\min_{\rho} \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\rho)^2$$
Second-stage residual is $\widehat{W}_i = Y_i - X_i'\widehat{\rho}$
3. Residual-on-residual least squares regression
$\check{\theta} = \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n}(\widehat{W}_i - \widehat{V}_i\theta)^2$

Frisch-Waugh-Lowell says

$$\check{\theta} = \widetilde{\theta} \text{ a.s.}$$

Frish-Waugh-Lowell theorem, cont.

Suppose oracle knows δ_0 and ρ_0. The oracle estimator $\widetilde{\theta}_{\text{oracle}}$ is

$$\widetilde{\theta}_{\text{oracle}} = \arg\min_\theta \frac{1}{n} \sum_{i=1}^n (W_i - V_i\theta)^2$$

The Frisch-Waugh-Lowell estimator is

$$\check{\theta} = \arg\min_\theta \frac{1}{n} \sum_{i=1}^n (\widehat{W_i} - \widehat{V}_i\theta)^2$$

Frisch-Waugh-Lowell says:

$\check{\theta}$ has the asymptotic distribution as the oracle $\widetilde{\theta}_{\text{oracle}}$!

$$\sqrt{n}(\check{\theta} - \theta_0) \approx \sqrt{n}(\widetilde{\theta}_{\text{oracle}} - \theta_0) \approx \frac{1}{\sqrt{n}} (\mathbb{E} V_i^2)^{-1} \sum_{i=1}^n V_i \cdot U_i + o_P(1)$$

Frish-Waugh-Lowell theorem, cont.

Suppose oracle knows δ_0 and ρ_0. The oracle estimator $\widetilde{\theta}_{\text{oracle}}$ is

$$\widetilde{\theta}_{\text{oracle}} = \arg\min_\theta \frac{1}{n} \sum_{i=1}^n (W_i - V_i\theta)^2$$

The Frisch-Waugh-Lowell estimator is

$$\check{\theta} = \arg\min_\theta \frac{1}{n} \sum_{i=1}^n (\widehat{W_i} - \widehat{V}_i\theta)^2$$

Frisch-Waugh-Lowell says:

$\check{\theta}$ has the asymptotic distribution as the oracle $\widetilde{\theta}_{\text{oracle}}$!

$$\sqrt{n}(\check{\theta} - \theta_0) \approx \sqrt{n}(\widetilde{\theta}_{\text{oracle}} - \theta_0) \approx \frac{1}{\sqrt{n}}(\mathbb{E} V_i^2)^{-1} \sum_{i=1}^n V_i \cdot U_i + o_P(1)$$

Frish-Waugh-Lowell theorem, orthogonality

The partially linear moment equation is

$$m(\text{data}, \theta_0, \gamma_0) = \mathbb{E}[Y - D\theta_0 - Z'\gamma_0]D = 0$$

Frish-Waugh-Lowell theorem, orthogonality

The partially linear moment equation is

$$m(\text{data}, \theta_0, \gamma_0) = \mathbb{E}[Y - D\theta_0 - Z'\gamma_0]D = 0$$

The derivative w.r.t γ_0 is

$$\partial_\gamma \mathbb{E}[m(\text{data}, \theta_0, \gamma_0)] = \mathbb{E}[DZ] \neq 0.$$

Frish-Waugh-Lowell theorem, orthogonality

The partially linear moment equation is

$$m(\text{data}, \theta_0, \gamma_0) = \mathbb{E}[Y - D\theta_0 - Z'\gamma_0]D = 0$$

The derivative w.r.t γ_0 is

$$\partial_\gamma \mathbb{E}[m(\text{data}, \theta_0, \gamma_0)] = \mathbb{E}[DZ] \neq 0.$$

Taylor expansion of $m(\text{data}, \theta_0, \gamma_0)$ around θ_0

$$\mathbb{E}[m(\text{data}, \theta_0, \widehat{\gamma}) - m(\text{data}, \theta_0, \gamma_0)] \approx \partial_\gamma \mathbb{E}[m(\text{data}, \theta_0, \gamma_0)][\widehat{\gamma} - \gamma_0]$$

Frish-Waugh-Lowell theorem, orthogonality

The partially linear moment equation is

$$m(\text{data}, \theta_0, \gamma_0) = \mathbb{E}[Y - D\theta_0 - Z'\gamma_0]D = 0$$

The derivative w.r.t γ_0 is

$$\partial_\gamma \mathbb{E}[m(\text{data}, \theta_0, \gamma_0)] = \mathbb{E}[DZ] \neq 0.$$

Taylor expansion of $m(\text{data}, \theta_0, \gamma_0)$ around θ_0

$$\mathbb{E}[m(\text{data}, \theta_0, \widehat{\gamma}) - m(\text{data}, \theta_0, \gamma_0)] \approx \partial_\gamma \mathbb{E}[m(\text{data}, \theta_0, \gamma_0)][\widehat{\gamma} - \gamma_0]$$

First-order effect of $\widehat{\gamma} - \gamma_0$ on $\widehat{\theta} - \theta_0$ is non-zero

$$\mathbb{E}[DZ](\widehat{\gamma} - \gamma_0) \neq 0$$

(except when $\widehat{\gamma}$ is estimated by OLS or series)

- $\check{\theta} = \widetilde{\theta}$ only holds when $(\widehat{\gamma}, \widetilde{\theta})$ is estimated by OLS

Frish-Waugh-Lowell theorem, orthogonality, cont.

The Frisch-Waugh-Lowell moment equation is

$$g(\text{data}, \theta_0, \{\delta_0, \rho_0\}) = \mathbb{E}(Y - Z'\rho_0 - (D - Z'\delta_0)\theta_0) \cdot (D - Z'\delta_0)$$

Frish-Waugh-Lowell theorem, orthogonality, cont.

The Frisch-Waugh-Lowell moment equation is

$$g(\mathsf{data}, \theta_0, \{\delta_0, \rho_0\}) = \mathbb{E}(Y - Z'\rho_0 - (D - Z'\delta_0)\theta_0) \cdot (D - Z'\delta_0)$$

The derivative w.r.t ρ_0 is

$$\partial_\rho \mathbb{E}[g(\mathsf{data}, \theta_0, \{\delta_0, \rho_0\})] = -\mathbb{E}(D - Z'\delta_0)Z = 0.$$

Frish-Waugh-Lowell theorem, orthogonality, cont.

The Frisch-Waugh-Lowell moment equation is

$$g(\text{data}, \theta_0, \{\delta_0, \rho_0\}) = \mathbb{E}(Y - Z'\rho_0 - (D - Z'\delta_0)\theta_0) \cdot (D - Z'\delta_0)$$

The derivative w.r.t ρ_0 is

$$\partial_\rho \mathbb{E}[g(\text{data}, \theta_0, \{\delta_0, \rho_0\})] = -\mathbb{E}(D - Z'\delta_0)Z = 0.$$

The derivative w.r.t δ_0 is

$$\begin{aligned}\partial_\delta \mathbb{E}[g(\text{data}, \theta_0, \{\delta_0, \rho_0\})] = & -\mathbb{E}Z \cdot U + \\ & - \mathbb{E}Z \cdot (D - Z'\rho_0)\theta_0 = 0.\end{aligned}$$

First-order effect of $\widehat{\delta} - \delta_0$ and $\widehat{\rho} - \rho_0$ on $\widehat{\theta} - \theta_0$ is zero

Outline

- Frish-Waugh-Lowell
- Double Lasso. Double selection
- Double Machine Learning
- Newey (1994) rule
- Double robustness

Point I. "Naive" or prediction-based Lasso is bad

Run Lasso with outcome Y and covariates D and Z. Obtain

$$D\widehat{\theta}_0 + Z'\widehat{\gamma}_0$$

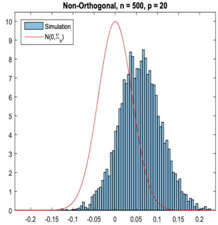

Figure: Figure 1 (a) from Chernozhukov et al. (2018)

Point II. The Double Lasso is good

1. Treatment ℓ_1-penalized least squares regression

$$\widehat{\delta} = \arg\min_{\delta} \frac{1}{n} \sum_{i=1}^{n} (D_i - X_i'\delta)^2 + \lambda_{\delta} \|\delta\|_1$$

First-stage residual is $\widehat{V}_i = D_i - X_i'\widehat{\delta}$

Point II. The Double Lasso is good

1. Treatment ℓ_1-penalized least squares regression

$$\widehat{\delta} = \arg\min_{\delta} \frac{1}{n} \sum_{i=1}^{n} (D_i - X_i'\delta)^2 + \lambda_{\delta} \|\delta\|_1$$

First-stage residual is $\widehat{V}_i = D_i - X_i'\widehat{\delta}$

2. Outcome ℓ_1-penalized least squares regression

$$\widehat{\rho} = \arg\min_{\rho} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i'\rho)^2 + \lambda_{\rho} \|\rho\|_1$$

Second-stage residual is $\widehat{W}_i = Y_i - X_i'\widehat{\rho}$

Point II. The Double Lasso is good

1. Treatment ℓ_1-penalized least squares regression
$$\widehat{\delta} = \arg\min_{\delta} \frac{1}{n} \sum_{i=1}^{n} (D_i - X_i'\delta)^2 + \lambda_{\delta}\|\delta\|_1$$
First-stage residual is $\widehat{V}_i = D_i - X_i'\widehat{\delta}$
2. Outcome ℓ_1-penalized least squares regression
$$\widehat{\rho} = \arg\min_{\rho} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i'\rho)^2 + \lambda_{\rho}\|\rho\|_1$$
Second-stage residual is $\widehat{W}_i = Y_i - X_i'\widehat{\rho}$
3. The double Lasso estimator is $\check{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} (\widehat{W}_i - \widehat{V}_i\theta)^2$

Point II. The Double Lasso is good

1. Treatment ℓ_1-penalized least squares regression

$$\widehat{\delta} = \arg\min_{\delta} \frac{1}{n} \sum_{i=1}^{n} (D_i - X_i'\delta)^2 + \lambda_{\delta} \|\delta\|_1$$

   First-stage residual is $\widehat{V}_i = D_i - X_i'\widehat{\delta}$

2. Outcome ℓ_1-penalized least squares regression

$$\widehat{\rho} = \arg\min_{\rho} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i'\rho)^2 + \lambda_{\rho} \|\rho\|_1$$

   Second-stage residual is $\widehat{W}_i = Y_i - X_i'\widehat{\rho}$

3. The double Lasso estimator is $\check{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} (\widehat{W}_i - \widehat{V}_i\theta)^2$

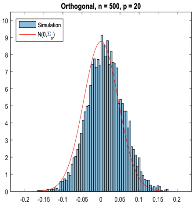

Figure: Figure 1 (b) from Chernozhukov et al. (2018)

Why is “naive” or prediction-based ML bad?

First-stage estimate is

$$\widehat{g}_0(X_i) = X_i'\widehat{\gamma}$$

Why is "naive" or prediction-based ML bad?

First-stage estimate is

$$\widehat{g}_0(X_i) = X_i'\widehat{\gamma}$$

Naive OLS estimator $\widehat{\theta}$ and its estimation error $\widehat{\theta} - \theta_0$ are

$$\widehat{\theta} = \left(\frac{1}{n}\sum_{i=1}^{n} D_i^2\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} D_i(Y_i - \widehat{g}_0(X_i))$$

$$\widehat{\theta} - \theta_0 = \left(\frac{1}{n}\sum_{i=1}^{n} D_i^2\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} D_i(Y_i - D_i\theta_0 - \widehat{g}_0(X_i))$$

$Y_i - D_i\theta_0 - \widehat{g}_0(X_i)$ is the sum of sampling and estimation error

$$\begin{aligned} Y_i - D_i\theta_0 - \widehat{g}_0(X_i) &= Y_i - D_i\theta_0 - g_0(X_i) + g_0(X_i) - \widehat{g}_0(X_i) \\ &= U_i + g_0(X_i) - \widehat{g}_0(X_i) \end{aligned}$$

Why is "naive" or prediction-based ML bad?, cont.

Numerator of $\widehat{\theta} - \theta_0$ is

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_i(Y_i - D_i\theta_0 - \widehat{g}_0(X_i))$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_i U_i + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_i(g_0(X_i) - \widehat{g}_0(X_i)) = a + b$$

Central Limit Theorem implies

$$a = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_i U_i \Rightarrow_d N(0, \sigma^2)$$

Why is "naive" or prediction-based ML bad?, cont.

Denote $m_0(X_i) = X_i'\delta_0$ as the true first stage.

$$b \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m_0(X_i)(g_0(X_i) - \widehat{g}_0(X_i))$$

Why is "naive" or prediction-based ML bad?, cont.

Denote $m_0(X_i) = X_i'\delta_0$ as the true first stage.

$$b \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m_0(X_i)(g_0(X_i) - \widehat{g}_0(X_i))$$

Cauchy-Schwartz implies an upper bound on b

$$|b| \leq \sqrt{n}\left(\frac{1}{n} \sum_{i=1}^{n} m_0^2(X_i)\right)^{1/2}\left(\frac{1}{n} \sum_{i=1}^{n}(g_0(X_i) - \widehat{g}_0(X_i))^2\right)^{1/2}$$

Why is "naive" or prediction-based ML bad?, cont.

Denote $m_0(X_i) = X_i'\delta_0$ as the true first stage.

$$b \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} m_0(X_i)(g_0(X_i) - \widehat{g}_0(X_i))$$

Cauchy-Schwartz implies an upper bound on b

$$|b| \leq \sqrt{n}\left(\frac{1}{n} \sum_{i=1}^{n} m_0^2(X_i)\right)^{1/2}\left(\frac{1}{n} \sum_{i=1}^{n}(g_0(X_i) - \widehat{g}_0(X_i))^2\right)^{1/2}$$

The first term converges to $\mathbb{E}m_0(x)^2$

$$\frac{1}{n} \sum_{i=1}^{n} m_0^2(X_i) \rightarrow \mathbb{E}m_0^2(X_i)$$

Why is "naive" or prediction-based ML bad?, cont.

Denote $m_0(X_i) = X_i'\delta_0$ as the true first stage.

$$b \approx \frac{1}{\sqrt{n}} \sum_{i=1}^n m_0(X_i)(g_0(X_i) - \widehat{g}_0(X_i))$$

Cauchy-Schwartz implies an upper bound on b

$$|b| \leq \sqrt{n}\left(\frac{1}{n} \sum_{i=1}^n m_0^2(X_i)\right)^{1/2}\left(\frac{1}{n} \sum_{i=1}^n (g_0(X_i) - \widehat{g}_0(X_i))^2\right)^{1/2}$$

The first term converges to $\mathbb{E} m_0(x)^2$

$$\frac{1}{n} \sum_{i=1}^n m_0^2(X_i) \rightarrow \mathbb{E} m_0^2(X_i)$$

The second term is bounded as

$$\frac{1}{n} \sum_{i=1}^n (g_0(X_i) - \widehat{g}_0(X_i))^2 = \frac{1}{n} \sum_{i=1}^n (X_i'(\widehat{\gamma}_L - \gamma_0))^2 \lesssim \|\widehat{\gamma}_L - \gamma_0\|_2^2 \lesssim_P C \frac{s_\gamma \log p}{n}$$

$$\boxed{\text{The naive or prediction-focused ML is not root-}n\text{ consistent}}$$

Why is double Lasso good?

Double Lasso estimator $\check{\theta}$ and its estimation error $\check{\theta} - \theta_0$ are

$$\check{\theta} = \left(\frac{1}{n} \sum_{i=1}^{n} \widehat{V}_i^2 \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \widehat{V}_i \widehat{W}_i$$

$$\check{\theta} - \theta_0 = \left(\frac{1}{n} \sum_{i=1}^{n} \widehat{V}_i^2 \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \widehat{V}_i (\widehat{W}_i - \widehat{V}_i \theta_0)$$

$\widehat{W}_i - \widehat{V}_i \theta_0$ is the sum of sampling and estimation error

$$\begin{aligned} \widehat{W}_i - \widehat{V}_i \theta_0 &= Y_i - l_0(X_i) - V_i \theta_0 + (\widehat{m}_0(X_i) - m_0(X_i))\theta_0 - (\widehat{l}_0(X_i) - l_0(X_i)) \\ &= U_i + (\widehat{m}_0(X_i) - m_0(X_i))\theta_0 - (\widehat{l}_0(X_i) - l_0(X_i)) \end{aligned}$$

Why is double Lasso good?, cont.

The numerator of $\check{\theta} - \theta_0$ is the sum of sampling and estimation error

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i U_i$$

$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i \left((\widehat{m}_0(X_i) - m_0(X_i))\theta_0 - (\widehat{l}_0(X_i) - l_0(X_i)) \right)$$

$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\widehat{V}_i - V_i) U_i$$

$$+ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\widehat{V}_i - V_i) \left((\widehat{m}_0(X_i) - m_0(X_i))\theta_0 - (\widehat{l}_0(X_i) - l_0(X_i)) \right) = a^* + c^* + d^* + b^*$$

Central Limit Theorem implies

$$a^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i U_i \Rightarrow_d N(0, \sigma_{uv}^2)$$

Why is double Lasso good?, cont.

By Law of Iterated Expectations,

$$\mathbb{E}[c^*] = \mathbb{E}(D - m_0(x))\left((\widehat{m}_0(x) - m_0(x))\theta_0 - (\widehat{l}_0(x) - l_0(x))\right) = 0$$

$$\mathbb{E}[d^*] = \mathbb{E}(\widehat{m}_0(x) - m_0(x))\theta_0 U = 0$$

We show later that d^*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i(m_0(X_i) - \widehat{m}(X_i))'\theta_0 = o_P(1)$$

and c^* likewise

Why is double Lasso good?, cont.

$$b \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (m_0(X_i) - \widehat{m}_0(X_i)) \cdot (l_0(X_i) - \widehat{l}_0(X_i))$$

Why is double Lasso good?, cont.

$$\color{red}{b} \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (m_0(X_i) - \widehat{m}_0(X_i)) \cdot (l_0(X_i) - \widehat{l}_0(X_i))$$

Cauchy-Schwartz implies an upper bound on b

$$|\color{red}{b}| \leq \sqrt{n}(\frac{1}{n} \sum_{i=1}^{n} (m_0(X_i) - \widehat{m}_0(X_i))^2 \frac{1}{n} \sum_{i=1}^{n} (l_0(X_i) - \widehat{l}_0(X_i))^2$$

Why is double Lasso good?, cont.

$$\color{red}{b} \approx \frac{1}{\sqrt{n}} \textstyle\sum_{i=1}^{n} (m_0(X_i) - \widehat{m}_0(X_i)) \cdot (l_0(X_i) - \widehat{l}_0(X_i))$$

Cauchy-Schwartz implies an upper bound on b

$$|\color{red}{b}| \leq \sqrt{n} (\frac{1}{n} \textstyle\sum_{i=1}^{n} (m_0(X_i) - \widehat{m}_0(X_i))^2 \frac{1}{n} \textstyle\sum_{i=1}^{n} (l_0(X_i) - \widehat{l}_0(X_i))^2$$

The first term is bounded as

$$\frac{1}{n} \sum_{i=1}^{n} (m_0(X_i) - \widehat{m}_0(X_i))^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i'(\widehat{\delta} - \delta_0))^2 \lesssim \|\widehat{\delta} - \delta_0\|_2^2 \lesssim_P C \frac{s_\delta \log p}{n}$$

The second term is bounded as

$$\frac{1}{n} \sum_{i=1}^{n} (l_0(X_i) - \widehat{l}_0(X_i))^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i'(\widehat{\rho} - \rho_0))^2 \lesssim \|\widehat{\rho} - \rho_0\|_2^2 \lesssim_P C \frac{s_\rho \log p}{n}$$

Why is double Lasso good?, summary

- a^* is $N(0, \sigma^2)$
- b^* is $\|\widehat{\rho} - \rho\|_2 \|\widehat{\delta} - \delta\|_2$
- $s_\rho = \|\rho_0\|_0$ and $s_\delta = \|\delta_0\|_0$ are the sparsity indices of ρ_0 and δ_0
- b^* is bounded by $\sqrt{n}\sqrt{\dfrac{s_\rho \log p}{n}}\sqrt{\dfrac{s_\delta \log p}{n}}$ where s_ρ and s_δ are sparsity indices of δ and ρ
- if $s_\rho \cdot s_\delta$ is sufficiently small, $\sqrt{n}\sqrt{\dfrac{s_\rho \log p}{n}}\sqrt{\dfrac{s_\delta \log p}{n}} = o(1)$

The Double ML estimator $\check{\theta}_0$ is a $\sqrt{n}$ consistent and approximately centered.

Outline

- Frish-Waugh-Lowell
- Double Lasso. Double selection
- Double Machine Learning
- Newey (1994) rule
- Double robustness

From double lasso to double ML

Double Lasso relies on two key primitive assumptions:

$$m_0(x) = \mathbb{E}[D|X = x] = x'\delta_0$$

and

$$l_0(x) = \mathbb{E}[Y|X = x] = x'\rho_0$$

are linear sparse functions of z with s_δ and s_ρ.

From double lasso to double ML

Double Lasso relies on two key primitive assumptions:

$$m_0(x) = \mathbb{E}[D|X = x] = x'\delta_0$$

and

$$l_0(x) = \mathbb{E}[Y|X = x] = x'\rho_0$$

are linear sparse functions of z with s_δ and s_ρ.

In Double ML, $m_0(x)$ and $l_0(x)$ need not be linear in z. One can use an arbitrary ML technique to compute

$$\widehat{m}(x) \text{ and } \widehat{l}(x)$$

as long as $\left(\frac{1}{n}\sum_{i=1}^{n}(\widehat{m}(X_i) - m_0(X_i))^2\right)^{1/2} = o_P(n^{-1/4})$ and $\left(\frac{1}{n}\sum_{i=1}^{n}(\widehat{l}(X_i) - l_0(X_i))^2\right)^{1/2} = o_P(n^{-1/4})$

Double Machine Learning

1. Estimate

$$m_0(x) \text{ and } l_0(x)$$

   by some ML technique (lasso, random forest, NN) on $\mathcal{A}_n$

2. On the main sample $\cup_{i=1}^n (D_i, X_i, Y_i)$, compute

$$\widehat{V}_i = D_i - \widehat{m}(X_i) \text{ and } \widehat{W}_i = Y_i - \widehat{l}(X_i)$$

3. The Double ML estimator is

$$\check{\theta} = \left(\sum_{i=1}^n \widehat{V}_i^2\right)^{-1} \sum_{i=1}^n \widehat{V}_i \widehat{W}_i$$

Why Sample Splitting?

Sample splitting means using two independent samples

- auxiliary sample $\mathcal{A}_n$ to estimate the first stage parameters
- main sample $\{1, 2, \ldots, n\}$ to compute $(\widehat{V}_i, \widehat{W}_i)_{i=1}^n$ and $\check{\theta}$

In the expansion

$$\sqrt{n}(\check{\theta}_0 - \theta_0) = a^* + b^* + c^*$$

the numerator of c^* contains terms like

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i(m_0(X_i) - \widehat{m}(X_i))$$

- with sample splitting, easy to control and claim $o_P(1)$
- without sample splitting, hard to control and claim $o_P(1)$

Why Sample Splitting?

- With sample splitting, conditional on the auxiliary data $\mathcal{A}_n$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i(m_0(X_i) - \widehat{m}(X_i))$$

is the sum of i.i.d terms.

Why Sample Splitting?

- With sample splitting, conditional on the auxiliary data $\mathcal{A}_n$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i(m_0(X_i) - \widehat{m}(X_i))$$

  is the sum of i.i.d terms.
- Let $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is an i.i.d sample average.
- Markov inequality: For any $\delta > 0$,

$$\Pr\left(\sqrt{n}(\bar{X} - \mathbb{E}[X]) > \delta\right) \leq \frac{\text{Var}(X)}{\delta^2}$$

Why Sample Splitting?

- With sample splitting, conditional on the auxiliary data $\mathcal{A}_n$
$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_i(m_0(X_i) - \widehat{m}(X_i))$$
is the sum of i.i.d terms.
- Let $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is an i.i.d sample average.
- Markov inequality: For any $\delta > 0$,
$$\Pr\left(\sqrt{n}(\bar{X} - \mathbb{E}[X]) > \delta\right) \leq \frac{\text{Var}(X)}{\delta^2}$$

Why Sample Splitting?, cont.

Conditional on $\mathcal{A}_n$, Markov inequality implies

$$\Pr\left(\left|\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} U_i(m_0(X_i) - \widehat{m}(X_i)) > \delta\right|\mathcal{A}_n\right)$$
$$\leq \delta^{-2}\mathbb{E}[U_i \cdot (m_0(X_i) - \widehat{m}(X_i))|\mathcal{A}_n]^2 \sim \delta^{-2} n^{-2\cdot\phi_m}$$

and one can take $\delta = \delta_n = o(1)$, where

$$U_i(m_0(X_i) - \widehat{m}(X_i)) = U_i(m_0(X_i) - \widehat{m}(X_i)) - \mathbb{E}[U_i(m_0(X_i) - \widehat{m}(X_i))|\mathcal{A}_n]$$

Why Sample Splitting?, cont.

Conditional on $\mathcal{A}_n$, Markov inequality implies

$$\Pr\left(\left|\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} U_i(m_0(X_i) - \widehat{m}(X_i)) > \delta\right| \mathcal{A}_n\right)$$
$$\leq \delta^{-2}\mathbb{E}[U_i \cdot (m_0(X_i) - \widehat{m}(X_i))|\mathcal{A}_n]^2 \sim \delta^{-2} n^{-2\cdot\phi_m}$$

and one can take $\delta = \delta_n = o(1)$, where

$$U_i(m_0(X_i) - \widehat{m}(X_i)) = U_i(m_0(X_i) - \widehat{m}(X_i)) - \mathbb{E}[U_i(m_0(X_i) - \widehat{m}(X_i))|\mathcal{A}_n]$$

is the demeaned sample average.

Why Sample Splitting?, cont.

Conditional on $\mathcal{A}_n$, Markov inequality implies

$$\Pr\left(\left|\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} U_i(m_0(X_i) - \widehat{m}(X_i)) > \delta\right|\mathcal{A}_n\right)$$

$$\leq \delta^{-2}\mathbb{E}[U_i \cdot (m_0(X_i) - \widehat{m}(X_i))|\mathcal{A}_n]^2 \sim \delta^{-2} n^{-2\cdot\phi_m}$$

and one can take $\delta = \delta_n = o(1)$, where

$$U_i(m_0(X_i) - \widehat{m}(X_i)) = U_i(m_0(X_i) - \widehat{m}(X_i)) - \mathbb{E}[U_i(m_0(X_i) - \widehat{m}(X_i))|\mathcal{A}_n]$$

is the demeaned sample average.

Conditional convergence to zero implies unconditional convergence

(Chernozhukov et al. (2018), Lemma 6.1).

Why Sample Splitting? (cont.)

Without sample splitting, the terms in c^* are not i.i.d. .

Why Sample Splitting? (cont.)

Without sample splitting, the terms in c^* are not i.i.d. .

The bound on

$$\sup_{m\in\mathcal{M}_n} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i(m) := \sup_{m\in\mathcal{M}_n} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (U_i(m(X_i) - m_0(X_i)),$$

depends on the complexity of the function class $\mathcal{M}_n$. Large (uncontrollable) complexity results in overfitting bias:

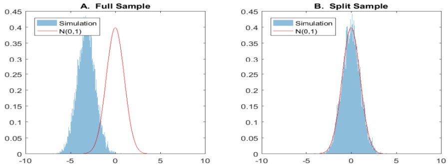

Figure: Figure 2 from Chernozhukov et al. (2018)

Definition (Cross-Fitting)

1. For a random sample of size N, denote a K-fold random partition of the sample indices $[N] = \{1, 2, ..., N\}$ by $(J_k)_{k=1}^{K}$, where K is the number of partitions and the sample size of each fold is $n = N/K$. For each $k \in [K] = \{1, 2, ..., K\}$ define $J_k^c = \{1, 2, ..., N\} \setminus J_k$.
2. For each $k \in [K]$, construct an estimator $\widehat{m}_k = \widehat{(}V_{i \in J_k^c})$ of the nuisance parameter m_0 using only the data $\{V_j : j \in J_k^c\}$. For any observation $i \in J_k$, define $\widehat{m}(X_i) = \widehat{m}_k(X_i)$.

Outline

- Frish-Waugh-Lowell
- Double Lasso. Double selection
- Double Machine Learning
- Newey (1994) rule
- Double robustness

General method for orthogonal moment (Newey (1994))

Start with an arbitrary moment equation

$$\mathbb{E}[\bar{m}(\text{data}, \theta_0, m_0)] = 0$$

where θ_0 is the target parameter and $m_0(x)$ is the nuisance function.

General method for orthogonal moment (Newey (1994))

Start with an arbitrary moment equation

$$\mathbb{E}[\bar{m}(\text{data}, \theta_0, m_0)] = 0$$

where θ_0 is the target parameter and $m_0(x)$ is the nuisance function.

Suppose the nuisance function $m_0(x)$

$$m_0(x) = \mathbb{E}[D|Z]$$

is a conditional expectation function (CEF)

General method for orthogonal moment (Newey (1994))

Start with an arbitrary moment equation

$$\mathbb{E}[\bar{m}(\text{data}, \theta_0, m_0)] = 0$$

where θ_0 is the target parameter and $m_0(x)$ is the nuisance function.

Suppose the nuisance function $m_0(x)$

$$m_0(x) = \mathbb{E}[D|Z]$$

is a conditional expectation function (CEF)

Goal is to obtain a new moment equation

$$\mathbb{E}[\bar{g}(\text{data}, \theta_0, \{m_0, l_0\})] = 0$$

that is orthogonal with respect to m_0 and l_0

Newey (1994) rule for partially linear model, step 1

Start with partially linear model

$$Y = D\theta_0 + g_0(x) + U, \quad \mathbb{E}[U|D, Z] = 0$$

Newey (1994) rule for partially linear model, step 1

Start with partially linear model

$$Y = D\theta_0 + g_0(x) + U, \quad \mathbb{E}[U|D,Z] = 0$$

Treatment CEF is

$$m_0(x) = \mathbb{E}[D|X = x]$$

Plugging in $D = (D - m_0(x)) + m_0(x)$ in (1) gives

$$Y = (D - m_0(x))\theta_0 + \underbrace{(g_0(x) + \theta_0 m_0(x))}_{l_0(x)} + U, \quad \mathbb{E}[U|D,Z] = 0$$

Newey (1994) rule for partially linear model, step 1

Start with partially linear model

$$Y = D\theta_0 + g_0(x) + U, \quad \mathbb{E}[U|D, Z] = 0$$

Treatment CEF is

$$m_0(x) = \mathbb{E}[D|X = x]$$

Plugging in $D = (D - m_0(x)) + m_0(x)$ in (1) gives

$$Y = (D - m_0(x))\theta_0 + \underbrace{(g_0(x) + \theta_0 m_0(x))}_{l_0(x)} + U, \quad \mathbb{E}[U|D, Z] = 0$$

Therefore,

$$Y - (D - m_0(x))\theta_0 = l_0(x) + U \text{ is independent of } D - m_0(x)$$

Newey (1994) rule for partially linear model, step 1

Start with partially linear model

$$Y = D\theta_0 + g_0(x) + U, \quad \mathbb{E}[U|D,Z] = 0$$

Treatment CEF is

$$m_0(x) = \mathbb{E}[D|X = x]$$

Plugging in $D = (D - m_0(x)) + m_0(x)$ in (1) gives

$$Y = (D - m_0(x))\theta_0 + \underbrace{(g_0(x) + \theta_0 m_0(x))}_{l_{\mathbf{0}}(x)} + U, \quad \mathbb{E}[U|D,Z] = 0$$

Therefore,

$$Y - (D - m_0(x))\theta_0 = l_0(x) + U \text{ is independent of } D - m_0(x)$$

The moment function

$$\bar{m}(\text{data}, \theta_0, m_0) = (Y - (D - m_0(x))\theta_0)(D - m_0(x))$$

obeys

$$\mathbb{E}\bar{m}(\text{data}, \theta_0, m_0) = 0$$

Newey (1994) rule for partially linear model, step 2

Fix $X = x$ at a specific value of z. The derivative of $\bar{m}(\text{data}, \theta_0, m_0)$ with respect to $m_0(x)$ is

$$\partial_{m_{\mathbf{0}}(x)} \bar{m}(\text{data}, \theta_0, m_0) = (D - m_0(x))\theta_0 - (Y - (D - m_0(x))\theta_0)$$

Newey (1994) rule for partially linear model, step 2

Fix $X = x$ at a specific value of z. The derivative of $\bar{m}(\text{data}, \theta_0, m_0)$ with respect to $m_0(x)$ is

$$\partial_{m_{\mathbf{0}}(x)} \bar{m}(\text{data}, \theta_0, m_0) = (D - m_0(x))\theta_0 - (Y - (D - m_0(x))\theta_0)$$

The first-order effect of $m(x) - m_0(x)$ on $\bar{m}(\text{data}, \theta_0, m_0)$ is

$$\begin{aligned}
&\partial_{m_{\mathbf{0}}(x)} \bar{m}(\text{data}, \theta_0, m_0) \cdot (m(x) - m_0(x)) \\
&= ((D - m_0(x))\theta_0 - (Y - (D - m_0(x))\theta_0)) (m(x) - m_0(x)) = I + II.
\end{aligned}$$

In expectation, the first-order effect is

$$\begin{aligned}
\mathbb{E}[I] &= \mathbb{E}[(D - m_0(x))\theta_0(m(x) - m_0(x))] = 0 \\
\mathbb{E}[II] &= \mathbb{E}[(Y - (D - m_0(x))\theta_0)(m(x) - m_0(x))] \\
&= \mathbb{E}(l_0(x) + U) \cdot (m(x) - m_0(x)) = \mathbb{E} l_0(x)(m(x) - m_0(x)) \neq 0
\end{aligned}$$

Newey (1994) rule for partially linear model, step 3

Newey (1994) proposes a new moment equation

$$g(\text{data}, \theta_0, \{m_0, l_0\}) = \bar{m}(\text{data}, \theta_0, m_0) - \cdot l_0(x)(D - m_0(x))$$

Newey (1994) rule for partially linear model, step 3

Newey (1994) proposes a new moment equation

$$g(\text{data}, \theta_0, \{m_0, l_0\}) = \bar{m}(\text{data}, \theta_0, m_0) - \cdot l_0(x)(D - m_0(x))$$

The summand $l_0(x)(D - m_0(x))$ is called the correction term

The proposed moment equation holds

$$\mathbb{E}g(\text{data}, \theta_0, \{m_0, l_0\}) = \mathbb{E}\bar{m}(\text{data}, \theta_0, m_0) - \mathbb{E}l_0(x) \cdot (\mathbb{E}[D|Z] - m_0(x)) = 0$$

at $\theta, \{m_0, l_0\}$.

Newey (1994) rule for partially linear model, step 3

Newey (1994) proposes a new moment equation

$$g(\text{data}, \theta_0, \{m_0, l_0\}) = \bar{m}(\text{data}, \theta_0, m_0) - \cdot l_0(x)(D - m_0(x))$$

The summand $l_0(x)(D - m_0(x))$ is called the correction term

The proposed moment equation holds

$$\mathbb{E}g(\text{data}, \theta_0, \{m_0, l_0\}) = \mathbb{E}\bar{m}(\text{data}, \theta_0, m_0) - \mathbb{E}l_0(x) \cdot (\mathbb{E}[D|Z] - m_0(x)) = 0$$

at $\theta, \{m_0, l_0\}$.

The moment equation obeys orthogonality condition

$$\partial_l \mathbb{E}[g(\text{data}, \theta_0, \{m_0, l_0\})] = \mathbb{E}0 \cdot (l(x) - l_0(x)) = 0$$

$$\partial_m \mathbb{E}[g(\text{data}, \theta_0, \{m_0, l_0\})] = \mathbb{E}0 \cdot (m(x) - m_0(x)) = 0$$

Newey (1994) rule for partially linear model, step 3

Newey (1994) proposes a new moment equation

$$g(\text{data}, \theta_0, \{m_0, l_0\}) = \bar{m}(\text{data}, \theta_0, m_0) - \cdot l_0(x)(D - m_0(x))$$

The summand $l_0(x)(D - m_0(x))$ is called the correction term

The proposed moment equation holds

$$\mathbb{E}g(\text{data}, \theta_0, \{m_0, l_0\}) = \mathbb{E}\bar{m}(\text{data}, \theta_0, m_0) - \mathbb{E}l_0(x) \cdot (\mathbb{E}[D|Z] - m_0(x)) = 0$$

at $\theta, \{m_0, l_0\}$.

The moment equation obeys orthogonality condition

$$\partial_l \mathbb{E}[g(\text{data}, \theta_0, \{m_0, l_0\})] = \mathbb{E}0 \cdot (l(x) - l_0(x)) = 0$$

$$\partial_m \mathbb{E}[g(\text{data}, \theta_0, \{m_0, l_0\})] = \mathbb{E}0 \cdot (m(x) - m_0(x)) = 0$$

Newey (1994) rule, summary

Take-aways

- Goal is to obtain an orthogonal moment equation

$$\mathbb{E}g(\text{data}, \theta_0) = 0$$

  starting from an arbitrary non-orthogonal one $\mathbb{E}[m(\text{data}, \theta_0, m_0)] = 0$

- Newey (1994): If $m_0(x) = \mathbb{E}[D|X = x]$ is a CEF, add the correction term

$$\partial_{m_0}\mathbb{E}[m(\text{data}, \theta_0, m_0)|X = x] \cdot (D - m_0(x))$$

- The nuisance parameter of g becomes

$$\{m_0(x), \partial_{m_0}\mathbb{E}[m(\text{data}, \theta_0, m_0)|X = x]\}$$

Newey (1994) rule, examples

Examples in the literature

- Average Treatment Effect (Robins and Rotnitzky (1995))
- Partially linear IV model

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. *Econometrics Journal*, 21:C1–C68.

Newey, W. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):245–271.

Robins, J. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of American Statistical Association*, 90(429):122–129.