Double Machine Learning for DiD Models

Zijian Wang

Research Module in Econometrics and Statistics

January 23, 2025

Basic two-period model

- Treatment $D_i \in \{0, 1\}$ is received between $t = 1$ and $t = 2$
- Potential outcomes framework:

$$Y_{i,t} = D_i Y_{i,t}(1) + (1 - D_i) Y_{i,t}(0) = \begin{cases} Y_{i,t}(1) & \text{if } D_i = 1 \\ Y_{i,t}(0) & \text{if } D_i = 0 \end{cases}$$

- Causal parameter of interest is the ATT in period 2:

$$\theta_0 = \mathbb{E}[Y_{i,2}(1) - Y_{i,2}(0) \,|\, D_i = 1]$$

Basic two-period model (cont.)

- Identifying assumptions:
  - Parallel trends: $\mathbb{E}[Y_{i,2}(0) - Y_{i,1}(0) \,|\, D_i = 1] = \mathbb{E}[Y_{i,2}(0) - Y_{i,1}(0) \,|\, D_i = 0]$
  - No anticipation: $\mathbb{E}[Y_{i,1}(0) \,|\, D_i = 1] = \mathbb{E}[Y_{i,1}(1) \,|\, D_i = 1]$

$\implies \theta_0 = \mathbb{E}[Y_{i,2} - Y_{i,1} \,|\, D_i = 1] - \mathbb{E}[Y_{i,2} - Y_{i,1} \,|\, D_i = 0]$

- DiD estimator:

$$\hat{\theta}_{0,\text{DiD}} = \frac{1}{N_1} \sum_{i:D_i=1} (Y_{i,2} - Y_{i,1}) - \frac{1}{N_0} \sum_{i:D_i=0} (Y_{i,2} - Y_{i,1})$$

Incorporating covariates

- Conditional version of identifying assumptions:
  - Parallel trends:
    $\mathbb{E}[Y_{i,2}(0) - Y_{i,1}(0) \,|\, D_i = 1, X_i] = \mathbb{E}[Y_{i,2}(0) - Y_{i,1}(0) \,|\, D_i = 0, X_i]$
  - No anticipation: $\mathbb{E}[Y_{i,1}(0) \,|\, D_i = 1, X_i] = \mathbb{E}[Y_{i,1}(1) \,|\, D_i = 1, X_i]$
  - Uniform overlap: $\mathbb{P}(D_i = 1 \,|\, X_i) \in (\varepsilon, 1 - \varepsilon)$ for some $\varepsilon > 0$
- Cond. ATT can be written as cond. DiD expression, so that ATT is identified via the LIE

Classical estimation approaches

- Regression adjustment:
  - Moment equation: $\theta_0 = \mathbb{E}[Y_{i,2} - Y_{i,1} \,|\, D_i = 1] - \mathbb{E}[g_0(X_i) \,|\, D_i = 1]$,
    with regression function $g_0(X_i) = \mathbb{E}[Y_{i,2} - Y_{i,1} \,|\, D_i = 0, X_i]$

  $\Longrightarrow \hat{\theta}_{0,\text{reg}} = \frac{1}{N_1} \sum_{i:D_i=1} \big(Y_{i,2} - Y_{i,1} - \hat{g}_0(X_i)\big)$
- Inverse probability weighting:
  - Moment equation: $\theta_0 = \frac{1}{p_0} \cdot \mathbb{E}\big[\frac{D_i - m_0(X_i)}{1 - m_0(X_i)}(Y_{i,2} - Y_{i,1})\big]$,
    with propensity score $m_0(X_i) = \mathbb{P}(D_i = 1 \,|\, X_i)$ and prior prob. $p_0 = \mathbb{P}(D_i = 1)$

  $\Longrightarrow \hat{\theta}_{0,\text{ipw}} = \frac{1}{\bar{D}} \cdot \frac{1}{N} \sum_{i=1}^{N} \big(\frac{D_i - \hat{m}_0(X_i)}{1 - \hat{m}_0(X_i)}(Y_{i,2} - Y_{i,1})\big)$
- Plug in non-parametric estimator $\hat{g}_0$ or $\hat{m}_0$ of nuisance function g_0 or m_0
  $\Longrightarrow$ Estimator of θ_0 won't attain parametric convergence rate $N^{-1/2}$

Double machine learning (DML)

- Proposed by Chernozhukov et al. [2018] and extended to DiD framework by Chang [2020]
- Main ingredients of DML:
  - Moment equation that defines an *infeasible* method-of-moments estimator, as some nuisance parameters need to be estimated beforehand
  - Employ ML methods to learn the nuisance functions
  - Use Neyman orthogonality to "remove" non-parametric convergence rates
  - Use cross-fitting to avoid overfitting bias

Neyman orthogonal moment equation

- Scalar nuisance parameter p with true value $p_0 = \mathbb{P}(D = 1)$
- Infinite-dimensional nuisance parameter $\eta = (g, m)$ with true value $\eta_0 = (g_0, m_0)$
- Consider the score function:

$$\psi(W; \theta, p, \eta) = \frac{D - m(X)}{p(1 - m(X))}(Y_2 - Y_1 - g(X)) - \theta$$

Proposition (Identification and Neyman orthogonality)

The ATT parameter θ_0 is identified by the moment equation $\mathbb{E}[\psi(W; \theta_0, p_0, \eta_0)] = 0$. This moment equation is "insensitive" to small errors in the nuisance parameter η, in the sense that $\partial_\eta \mathbb{E}[\psi(W; \theta_0, p_0, \eta)]\big|_{\eta=\eta_0} = 0$.

DML algorithm

1. Form a K-fold partition of $\{1, \dots, N\}$ into $\{I_k\}_{k=1}^K$ each of the size $n := N/K$, and define $I_k^c := \{1, \dots, N\} \setminus I_k$.
2. For each k, use $\{W_i\}_{i \in I_k^c}$ to compute estimators of η_0 and p_0:
$$\hat{\eta}_0(I_k^c) := \big(\hat{g}_0(\,\cdot\,; I_k^c),\ \hat{m}_0(\,\cdot\,; I_k^c)\big), \quad \bar{D}(I_k^c) := \frac{1}{N-n} \sum\nolimits_{i \in I_k^c} D_i$$
3. For each k, define $\check{\theta}_{0,k}$ as the root θ of:
$$\frac{1}{n} \sum\nolimits_{i \in I_k} \psi(W_i; \theta, \bar{D}(I_k^c), \hat{\eta}_0(I_k^c)) = 0$$
4. Obtain the final estimator of θ_0 by averaging:
$$\hat{\theta}_0 := \frac{1}{K} \sum\nolimits_{k=1}^K \check{\theta}_{0,k}$$

Asymptotics of DML estimator

Existence of moments + High-level assumptions on the ML estimators:

- Conv. rates: $\|\hat{g}_0(X; I_k^c) - g_0(X)\|_{L^2} + \|\hat{m}_0(X; I_k^c) - m_0(X)\|_{L^2} = o_P(N^{-1/4})$
- $\hat{m}_0(X; I_k^c) \in (\varepsilon, 1 - \varepsilon)$ with probability approaching 1

Theorem (Asymptotic normality)

$$\sqrt{N}(\hat{\theta}_0 - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

- Asymptotic variance $\sigma^2 = \mathbb{E}[(\psi(W; \theta_0, p_0, \eta_0) - \theta_0(D - p_0)/p_0)^2]$ can be consistently estimated
- Approximate $(1 - \alpha) \cdot 100\%$ CI: $\left[\hat{\theta}_0 \pm z_{1-\alpha/2} \cdot \hat{\sigma}/\sqrt{N}\right]$

References

N.-C. Chang. Double/debiased machine learning for difference-in-differences models. *The Econometrics Journal*, 23(2):177–191, 02 2020. ISSN 1368-4221. doi: 10.1093/ectj/utaa001. URL https://doi.org/10.1093/ectj/utaa001.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.