Machine Learning and Causal Inference III

PS690 Computational Methods in Social Science

Jiawei Fu
Department of Political Science
Duke University

September 30, 2025

Overview

1. RDD
   1.1 RDD with Many Covariates
2. IV
   2.1 DML IV
   2.2 Optimal IV
   2.3 Proxy Controls
3. DID
   3.1 DML DID

Basic RDD

- Sharp RDD: the binary treatment variable $D_i \in \{0,1\}$ for individual i is assigned based on a running variable X_i in a sharp way: $D_i = 1[X_i \geq c]$.
- The estimand from the continuity framework is $\mathbb{E}[Y_i(1) - Y_i(0)|X_i = c]$.

Assumption

The conditional mean of the potential outcomes $\mathbb{E}[Y(t)|X = x]$ *for* $t \in \{0,1\}$ *are continuous at the cutoff level* c*.*

- $\tau^{RD} = \lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] - lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x]$
- Estimation via local linear regression: $\hat{\tau}^{RD} = \arg min_{\tau} \{\sum_{i=1}^{n} K_h(X_i - c)(Y_i - a - \tau D_i - \beta_0 (X_i - c)_- - \beta_1 (X_i - c)_+)^2\}$, where $K_h(x) = \frac{1}{h} K(\frac{x}{h})$ is the kernel function and h_n is the bandwidth.
- Note that there are multiple ways to express the objective function.

RDD with Many Covariates

- In the lower dimension case, we can simply add covariates $Z_i \in \mathbb{R}^p$ linearly to the objective function to increase the efficiency.

$$\hat{\tau}^{RD} = \arg min_{\tau}\{\sum_{i=1}^{n} K_h(X_i - c)(Y_i - a - \tau D_i - \beta_0(X_i - c) - \beta_1 D_i(X_i - c) - \gamma' Z_i)^2\}$$

- Suppose we transform the outcome variable $\tilde{Y} = Y_i - Z_i^T \hat{\gamma}_h$, where $\hat{\gamma}_h$ is the vector of linear projection coefficients from previous function; then,

$$\hat{\tau}^{lin} = \arg min_{\tau}\{\sum_{i=1}^{n} K_h(X_i - c)(\tilde{Y} - a - \tau D_i - \beta_0(X_i - c) - \beta_1 D_i(X_i - c))^2\}$$

  $\hat{\tau}^{lin}$ is consistent to τ^{RD} if the conditional distribution of the regressors given the running variable varies smoothly around the cutoff.

- The variance of the linear adjustment estimator is asymptotically smaller than unadjusted one.

RDD with Many Covariates

- In the high-dimensional case, a natural choice is to use LASSO select relevant covariates and then run local linear RDD [Kreiss and Rothe, 2023].
- For simplicity, we use V to denote $(1, D_i, X_i - c, D_i(X_i - c))$, and $\theta = (a, \tau, \beta_0, \beta_1)$.
- The post-lasso RDD procedure:
  1. Using a preliminary bandwidth b and a penalty parameter λ, one solves a Lasso version of the local linear regression defining the RDD estimator by adding a penalty term to obtain preliminary estimates:

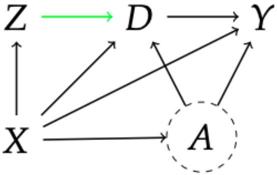
     $$\sum_{i=1}^{n} K_b(X_i - c)(\tilde{Y} - V_i^T \theta - (Z_i - \hat{\mu}_Z)\gamma)^2\} + \lambda \sum_{i=1}^{p} \hat{\omega}_k |\gamma_k|,$$

     where $\hat{\mu}_Z = \frac{1}{n}\sum_{i=1}^{n} Z_i K_i(X_i - c)$, and $\hat{\omega}_k^2 = \frac{b}{n}\sum_{i=1}^{n}(K_b(X_i - c)Z_i^{(k)} - \mu_Z^{(k)})^2$ are the local sample mean and variance, respectively, of the covariates.
  2. Let $\hat{J}$ denote the set of indices of those covariates whose first step Lasso estimates are non-zero. Using it and final bandwidth h to estimate τ.

RDD with Many Covariates

- Recall that, in the lower-dimensional case, $\hat{\tau}^{RD}$ is asymptotically equivalent to running a local linear RDD regression with a modified outcome variable $Y_i - Z_i'\gamma$.
- A natural extension is to consider $Y_i - \eta_0(Z_i')$ for the potentially non-linear function η_0, which can be connected to DML.
- That is, by sample spiting, one estimates any $\hat{\eta}(Z)$ and compute a local linear "no covariates" RDD estimator that uses the adjusted outcome.

Linear IV



- Consider SEM, for example Y is income, D is schooling, A is unobserved ability, Z is IV (distance to college, birth quarter, etc.), X is observed controls,

$$\begin{aligned} Y &= \alpha D + \delta A + f_Y(X) + \epsilon_Y \\ D &= \beta Z + \gamma A + f_D(X) + \epsilon_D \\ Z &= f_Z(X) + \epsilon_Z \\ A &= f_A(X) + \epsilon_A \\ X &= \epsilon_X \end{aligned}$$

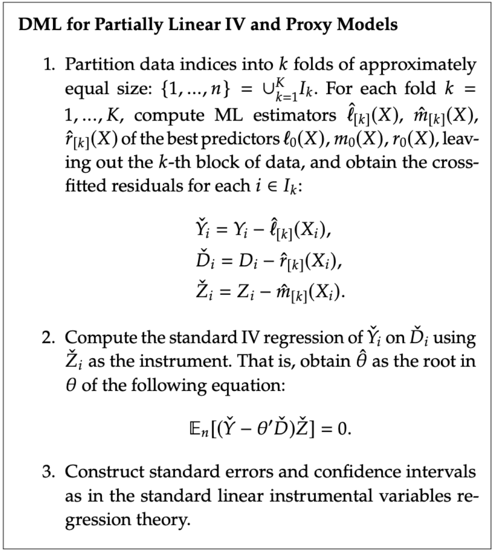
Partial linear IV

- We apply partialling-out: $\tilde{V} = V - \mathbb{E}[V|X]$, and have a simplified system

$$\tilde{Y} = \alpha\tilde{D} + \delta\tilde{A} + \epsilon_Y$$
$$\tilde{D} = \beta Z + \gamma\tilde{A} + \epsilon_D$$
$$\tilde{Z} = \epsilon_Z \; \tilde{A} = \epsilon_A$$

- To identify α, we use moment $\mathbb{E}[(\tilde{Y} - \alpha\tilde{D})\tilde{Z}] = 0$ (exclusion restriction).
- Then, we obtain our old friend $\alpha = \frac{\mathbb{E}[\tilde{Y}\tilde{Z}]}{\mathbb{E}[\tilde{D}\tilde{Z}]}$ (assuming relevance).
- Note that α is Neyman orthogonal to the nuisance parameters $\mathbb{E}[Y|X]$, $\mathbb{E}[D|X]$, $\mathbb{E}[Z|X]$; therefore, we can use DML.

DML Partial linear IV



DML for Partially Linear IV and Proxy Models

1. Partition data indices into k folds of approximately equal size: $\{1, ..., n\} = \cup_{k=1}^{K} I_k$. For each fold $k = 1, ..., K$, compute ML estimators $\hat{\ell}_{[k]}(X)$, $\hat{m}_{[k]}(X)$, $\hat{r}_{[k]}(X)$ of the best predictors $\ell_0(X), m_0(X), r_0(X)$, leaving out the k-th block of data, and obtain the cross-fitted residuals for each $i \in I_k$:

$$\check{Y}_i = Y_i - \hat{\ell}_{[k]}(X_i),$$
$$\check{D}_i = D_i - \hat{r}_{[k]}(X_i),$$
$$\check{Z}_i = Z_i - \hat{m}_{[k]}(X_i).$$

2. Compute the standard IV regression of $\check{Y}_i$ on $\check{D}_i$ using $\check{Z}_i$ as the instrument. That is, obtain $\hat{\theta}$ as the root in θ of the following equation:

$$\mathbb{E}_n[(\check{Y} - \theta'\check{D})\check{Z}] = 0.$$

3. Construct standard errors and confidence intervals as in the standard linear instrumental variables regression theory.

Nonlinear IV: LATE

- Consider the setup of causal inference with binary treatment D and binary instrument Z.
- We know that IV can be used to identify local ATE (Ate for compilers): $\mathbb{E}[Y_i(1) - Y_i(0) | D_i(1) > D_i(0)]$ under assumptions:
  1. Randomization: $Z \perp\!\!\!\perp \{D(1), D(0), Y(1), Y(0)\}$
  2. Monotonicity: $D_i(1) \geq D_i(0)$ for all i; there are no defier.
  3. Exclusion restriction: $Y_i(1) = Y_i(0)$ for always takers and never takers.
- It can be identified by $\tau^{LATE} = \frac{\mathbb{E}[\mathbb{E}[Y|Z=1,X] - \mathbb{E}[Y|Z=0,X]]}{\mathbb{E}[\mathbb{E}[Y|D=1,X] - \mathbb{E}[D|Z=0,X]]}$
- To use DML, we can derive the orthogonal score function. We define three nuisance parameters:$\mu(Z,X) = \mathbb{E}[Y|Z,X]$,$m(Z,X) = \mathbb{E}[D|Z,X]$, and $\pi(X) = \mathbb{E}[Z|X]$.
- The orthogonal score function is similar to two AIPW's:

$$\begin{aligned} \psi = {} & \mu(1,X) - \mu(0,X) + H(\pi)(Y - \mu(Z,X)) \\ & - [m(1,X) - m(0,X) + H(\pi)(Y - \mu(Z,X))]\tau^{LATE} \end{aligned}$$

where $H(\pi) = \frac{Z}{\pi(X)} - \frac{1-Z}{1-\pi(X)}$.

Optimal IV

- Recall that we can use GMM to perform IV regression.
- For a general linear model, $y_i = X_i\beta + u_i$, the moment condition is $\mathbb{E}[Z_i' u_i] = \mathbb{E}[Z_i'(y_i - X_i\beta)] = 0$.
- The sample analogue is $\frac{1}{N}\sum_{i=1}^{N} Z_i'(y_i - X_i\hat{\beta}) = 0$.
- And $\hat{\beta} = (\frac{1}{N}\sum_{i=1}^{N} Z_i'X_i)(\frac{1}{N}\sum_{i=1}^{N} Z_i'y_i)$ for just identification.
- For over-identification, GMM solves

$$\min_b [\sum_{i=1}^{N} Z_i'(y_i - X_i\beta)]' \widehat{W} [\sum_{i=1}^{N} Z_i'(y_i - X_i\beta)]$$

  where $\widehat{W}$ is a positive semidefinite weighting matrix.
- The solution is

$$\hat{\beta}^{GMM} = (X'Z\widehat{W}Z'X)^{-1}(X'Z\widehat{W}Z'X)$$

Optimal IV

- Suppose $\widehat{W} = (\frac{Z'Z}{N})^{-1}$, then $\hat{\beta}^{2sls} = [X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'Y$ is the 2SLS estimator.
- A natural question is how to choose the optimal weighting matrix.
- Look at the asymptotic variance:

$$Avar[\sqrt{N}(\hat{\beta}^{GMM} - \beta)] = (Q'WQ)^{-1}Q'W\Omega WQ(Q'WQ)^{-1}$$

  where $\Omega = \mathbb{E}(Z_i Z_i' u_i^2) = Var(Z_i' u_i)$, and $Q = \mathbb{E}[Z_i X_i']$.
- To minimize it, we can choose $W = \Lambda^{-1}$ so that it becomes $(Q'\Omega Q)^{-1}$.
- To estimate Ω, we first use β^{2sls} to get residual $\hat{u}_i$, and then construct $\hat{\Omega}$ using the sample analog.
- Also note that under heteroskedasticity $\mathbb{E}[u_i^2|Z_i] = \sigma^2$, the weighting matrix for 2SLS and GMM is the same, and thus 2SLS is efficient.

Optimal IV

- If we use the optimal weighting matrix $\widehat{W}$, then adding more instruments will not hurt us.
- Consider a slightly stronger conditional moment restriction: $\mathbb{E}[u_i|Z_i] = 0$.
- If Z is a valid IV, then we can construct infinitely many IVs based on it: any non-linear function of Z.
- Now, we ask what is the optimal set of instrumental variables?
- One solution is to construct an infinite list of potent instruments and then use the first k.
- Another approach is to construct an optimal instrument that minimizes asymptotic variance.
- It turns out that the optimal IV is $Z_i^* = [Var(u_i|Z_i)]^{-1}\mathbb{E}[X_i|Z_i] = [\mathbb{E}(u_i^2|Z_i)]^{-1}\mathbb{E}[X_i|Z_i]$.
- Then, we do not need a weighting matrix to solve $\sum_{i=1}^{N} Z_i^*(y_i - X_i\hat{\beta}) = 0$.

Optimal IV

- Let us assume homeostatic so that we ignore the inverse term in the optimal IV formula.
- To obtain an optimal IV, people may try to estimate $\mathbb{E}[X_i|Z_i]$, then solve the IV estimator.
- It may suffer from overfitting bias.
- The main problem is that if $\mathbb{E}[X_i|Z_i]$ is fitted on the training data, then the estimated optimal IV is a function of X_i, which is correlated with u_i.
- In other words, the optimal IV estimated for each i, is correlated to u_i.
- Therefore, we should use cross-fitting to address this issue.

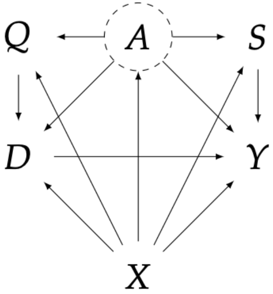
Optimal IV

- We introduce the optimal IV results by LASSO as follows [Belloni et al., 2012].
- We still focus on the homoskedastic case.
- Suppose that there is a very large list of instruments $Z_{i1}, ..., Z_{ik}$, and $\mathbb{E}[X_i|Z_i] = Z_i'\alpha_0$, where β_0 is sparse.
- Then, using LASSO / post-LASSO to estimate the predicted $\hat{X}_i = Z_i'\hat{\alpha}_0{}^{lasso}$, and then obtain the resulting IV estimator.
- Under several lasso-similar conditions, the above estimator achieves the efficiency bound asymptotically (as the optimal IV).
- In the presence of heteroscedasticity, the above IV estimator continues to be $\sqrt{n}$ consistent and asymptotically normal.

Proxy Controls

- Suppose that we observe some proxies Q (test scores or grades in the early period) and S (test scores or grades in later period) of unobserved ability:

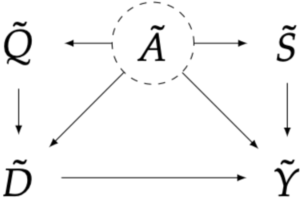
$$
\begin{aligned}
Y &= \alpha D + \delta A + \iota S + f_Y(X) + \epsilon_Y \\
D &= \gamma A + \beta Q + f_D(X) + \epsilon_D \\
Q &= \eta A + f_Q(X) + \epsilon_Q \\
S &= \phi A + f_S(X) + \epsilon_S \\
A &= f_A(X) + \epsilon_A \\
X &= \epsilon_X
\end{aligned}
$$



Proxy Controls

- Similarly, we apply partialling-out,

$$
\begin{aligned}
\tilde{Y} &= \alpha \tilde{D} + \delta \tilde{A} + \iota \tilde{S} + \epsilon_Y \\
\tilde{D} &= \gamma \tilde{A} + \beta \tilde{Q} + \epsilon_D \\
\tilde{Q} &= \eta \tilde{A} + \epsilon_Q \\
\tilde{S} &= \phi \tilde{A} + \epsilon_S \\
\tilde{A} &= \epsilon_A
\end{aligned}
$$



- Via substitution ($\tilde{A} = \frac{\tilde{S} - \epsilon_S}{\phi}$), we obtain $\tilde{Y} = \alpha \tilde{D} + \overline{\delta} \tilde{S} + U$, where $U = \frac{-\delta \epsilon_S}{\phi} + \epsilon_Y$ and $\overline{\delta} = \iota + \frac{\delta}{\phi}$.
- Note that $\mathbb{E}[U\tilde{D}] = \mathbb{E}[U\tilde{Q}] = 0$.
- Therefore, we can use $\tilde{Q}$ as an IV for $\tilde{S}$ and thus identify α.

Canonical DID

- In the simplest two periods DID, we use $Y_t(d)$ to denote potential outcomes for $d = 0, 1$ and $t = 1, 2$.
- To identify ATET $\mathbb{E}[Y_2(1) - Y_2(0)|D = 1]$, we rely on two assumptions.

Assumption (Parallel Trends)

$$\mathbb{E}[Y_2(0) - Y_1(0)|D = 1] = \mathbb{E}[Y_2(0) - Y_1(0)|D = 0]$$

- Parallel trends assumption requires that, in expectation, the change in control potential outcomes among the treatment group is the same as the change in the control potential outcomes among the control group.
- It is important to mention that it is actually functional form dependent (additively separable). It will not hold if you make a non-linear transformation of Y.

Canonical DID

Assumption (No Anticipation)

$\mathbb{E}[Y_1(0)|D = 1] = \mathbb{E}[Y_1(1)|D = 1]$

- No Anticipation imposes that receipt of treatment at $t = 2$ does not impact average period 1 potential outcomes.
- The identification is straightforward:

$$\begin{aligned} ATET &= \mathbb{E}[Y_2(1) - Y_2(0)|D = 1] \\ &= \mathbb{E}[Y_2(1)|D = 1] - \mathbb{E}[Y_2(0)|D = 1] \\ &= \mathbb{E}[Y_2(1)|D = 1] - \{\mathbb{E}[Y_1(1)|D = 1] + \mathbb{E}[Y_2(0) - Y_1(0)|D = 0]\} \\ &= \mathbb{E}[Y_2(1) - Y_1(1)|D = 1] - \mathbb{E}[Y_2(0) - Y_1(0)|D = 1] \end{aligned}$$

Canonical DID

- To estimate, we can estimate conditional means: $\hat{\theta}_s(d) = \frac{\mathbb{E}_n[Y1[D=d, t=s]]}{\mathbb{E}_n[1[D=d, t=s]]}$, and $\hat{\alpha} := \widehat{ATET} = (\hat{\theta}_2(1) - \hat{\theta}_1(1)) - (\hat{\theta}_2(0) - \hat{\theta}_1(0))$.
- A numerically equivalent estimator can be obtained through regression:

$$Y = \beta_0 + \beta_1 D + \beta_2 T + \alpha DT + \epsilon$$

DML DID

- In practice, it is more likely to satisfy conditional parallel trends

Assumption (Conditional DID)

$$\mathbb{E}[Y_2(0) - Y_1(0) | D = 1, X] = \mathbb{E}[Y_2(0) - Y_1(0) | D = 0, X] \textit{ a.s.}$$

and

$$\mathbb{E}[Y_1(0) | D = 1] = \mathbb{E}[Y_1(1) | D = 1] \textit{ a.s.}$$

- Now, we have high-dimensional nuisance parameters to be estimated.

DML DID

- The Neyman orthogonal score function is

$$\psi(W;\alpha,\eta) = \frac{D - m(X)}{p(1 - m(X))}(Y_2 - Y_1 - g(0,X)) - \frac{D}{p}\alpha,$$

  where the true value for nuisance parameters are $p_0 = \mathbb{E}[D]$,$m_0(X) = \mathbb{E}[D|X]$, and $g_0(0,X) = \mathbb{E}[Y_2 - Y_1|D = 0, X]$.
- So, as usual, we use partition data into K folds and estimate nuisance parameters, $\hat{p}$,$\hat{g}$,and $\hat{m}$. And then construct score function $\hat{\psi}$ and estimate $\hat{\alpha}$ using sample analogue of the moment condition $\mathbb{E}_n[\hat{\psi}(W_i,\alpha,\hat{\eta})] = 0$.

References

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012).
Sparse models and methods for optimal instruments with an application to eminent domain.
Econometrica, 80(6):2369–2429.

Kreiss, A. and Rothe, C. (2023).
Inference in regression discontinuity designs with high-dimensional covariates.
The Econometrics Journal, 26(2):105–123.