Machine Learning
for Economists

High-Dimensional Confounding

Anthony Strittmatter

Outline

1. Double Selection Procedure
2. General Thoughts
3. Modified Outcome Method
    - Selection-on-Observables
    - Instrumental Variable Approach
    - Difference-in-Differences
4. Practical Considerations

Literature

- Belloni, Chernozhukov, and Hansen (2014): "High-Dimensional Methods and Inference on Structural and Treatment Effects", Journal of Economic Perspectives, 28 (2), pp. 29-50, download.
- Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey (2017): "Double/Debiased/Neyman Machine Learning of Treatment Effects", American Economic Review, P&P, 107 (5), pp. 261-265, download.
- Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2017): "Double/Debiased Machine Learning for Treatment and Structural Parameters", Econometrics Journal, 21 (1), pp. C1-C68, download.
- Zimmert (2019): "Efficient Difference-in-Differences Estimation with High-Dimensional Common Trend Confounding", arXiv:1809.01643, download.

1. Double Selection Procedure

- **Partial Linear Model:**

$$\begin{aligned} Y_i &= D_i\delta + g(X_i) + U_i \\ D_i &= m(X_i) + V_i \end{aligned}$$

  with $E[U_i|D_i, X_i] = 0$ and $E[V_i|X_i] = 0$

- **Approximation with Linear Model:**

$$\begin{aligned} Y_i &= D_i\delta + X_i\beta_g + r_{gi} + U_i \\ D_i &= X_i\beta_m + r_{mi} + V_i \end{aligned}$$

  where X_i can include interactions and non-linear terms.

- r_{gi} and r_{mi} are approximation errors of functions $g(\cdot)$ and $m(\cdot)$, respectively

Reference: Belloni, Chernozhukov, and Hansen (2014)

Types of Covariates

Relation between covariates and outcome (for some $s_g > 0$):

- $|\beta_{gj}| > s_g$: covariate X_j has a **strong association** with Y_i
- $0 < |\beta_{gj}| \leq s_g$: covariate X_j has a **weak association** with Y_i
- $\beta_{gj} = 0$: covariate X_j has a **no association** with Y_i

Relation between covariates and treatment (for some $s_m > 0$):

- $|\beta_{mj}| > s_m$: covariate X_j has a **strong association** with D_i
- $0 < |\beta_{mj}| \leq s_m$: covariate X_j has a **weak association** with D_i
- $\beta_{mj} = 0$: covariate X_j has a **no association** with D_i

$\rightarrow$ All covariates are standardised

Types of Covariates (cont.)

	$\beta_{gj} = 0$	$0 < \lvert\beta_{gj}\rvert \leq s_g$	$\lvert\beta_{gj}\rvert > s_g$
$\beta_{mj} = 0$	Irrelevant	Irrelevant	Irrelevant
$0 < \lvert\beta_{mj}\rvert \leq s_m$	Irrelevant	Unclear?	Weak Confounder
$\lvert\beta_{mj}\rvert > s_m$	Irrelevant	Weak Confounder	Strong Confounder

- $|\beta_{gj}| > s_g$ and $0 < |\beta_{mj}| \leq s_m$: "Weak Outcome Confounder"
- $|\beta_{mj}| > s_m$ and $0 < |\beta_{gj}| \leq s_g$: "Weak Treatment Confounder"

$\rightarrow$ Approximate sparsity means (roughly) that covariates with $|\beta_{gj}| \leq s_g$ and $|\beta_{mj}| \leq s_m$ are not important

Naive Approaches

- Apply LASSO to structural model

$$\min_{\beta_g} E[(Y_i - D_i\delta - X_i\beta_g)^2] + \lambda \|\beta_g\|_1$$

  without a penalty on δ

  - Covariates that are highly correlated with D_i are probably not selected, even though they could be "strong confounders"
  - "Weak treatment confounders" are less likely selected

- Apply LASSO to selection model

$$\min_{\beta_m} E[(D_i - X_i\beta_m)^2] + \lambda \|\beta_m\|_1$$

  - "Weak outcome confounders" are less likely selected

Double Selection Procedure

1. Apply LASSO to the reduced form models

$$\min_{\tilde{\beta}_g} E[(Y_i - X_i\tilde{\beta}_g)^2] + \lambda \|\tilde{\beta}_g\|_1 \quad (1)$$

$$\min_{\beta_m} E[(D_i - X_i\beta_m)^2] + \lambda \|\beta_m\|_1 \quad (2)$$

   with $\tilde{\beta}_g = \delta\beta_m + \beta_g$

   - "Strong confounders" and "weak treatment confounders" are likely selected in (2)
   - $\tilde{\beta}_{gj} \approx \beta_g$ when $\beta_{mj} \approx 0$, such that "weak outcome confounders" are likely selected in (1)
   - Possibly, we additionally select less important variables in (1)

2. Take the union of all covariates $\tilde{X}_i$ with estimated LASSO coefficients of either $\hat{\beta}_{gj} \neq 0$ **or** $\hat{\beta}_{mj} \neq 0$ and estimate the OLS model

$$Y_i = D_i\delta + \tilde{X}_i\beta_g^* + u_i$$

Asymptotic Results

(Main) regularity conditions:

- Approximate sparsity
- Sparse eigenvalues $\rightarrow$ restriction on the correlation structure between covariates

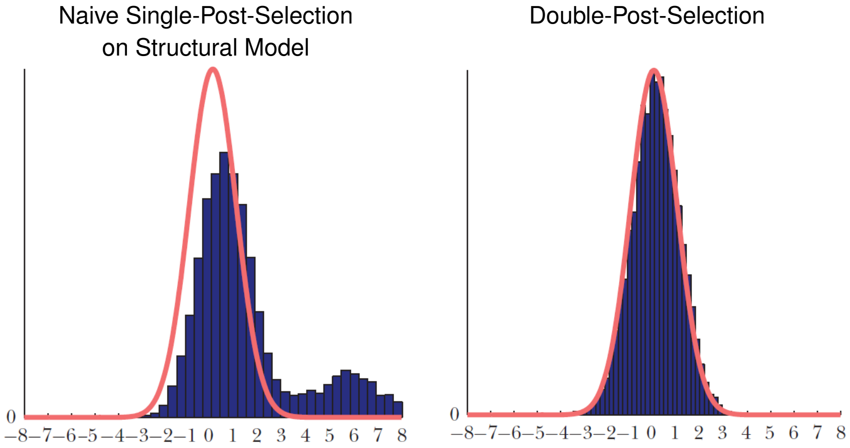
Asymptotic results of the double selection procedure:

- Asymptotic normality
$$\sqrt{N}(\hat{\delta} - \delta) \xrightarrow{d} N(0, \sigma)$$
- Model selection step is asymptotically negligible for building confidence intervals
- Optimal penalty parameter $\lambda^* = 2c \cdot \Phi^{-1}(1 - \gamma/2p)/\sqrt{N}$ (e.g., $c = 1.1$ and $\gamma \leq 0.05$) for "Feasible LASSO"

$$\min_{\beta} E[(Y_i - X_i\beta)^2] + \lambda^* \|\beta\|_1$$

Simulation Exercise

Distribution of Estimators



Source: Belloni, Chernozhukov, and Hansen (2014)

Example: Effect of Abortion on Crime

	Crime Type					
	Violent		Property		Murder	
	Effect	Std. err.	Effect	Std. err.	Effect	Std. err.
Donohue and Levitt (2001)	-.157***	0.034	-.106***	0.021	-.218***	0.068
284 controls	0.071	0.284	-.161	0.106	-1.327	0.932
Double-selection	-.171	0.117	-.061	0.057	-.189	0.177

Source: Belloni, Chernozhukov, and Hansen (2014), $N = 600$

Summary Double Selection Procedure

Advantages:

- Asymptotic results available
- Standard inference
- Computationally fast

Disadvantages:

- Effect homogeneity
- Restrictive assumptions required
- Potentially too many covariates selected

2. Some General Thoughts

- Partial Linear Model:

$$Y_i = D_i\delta + g(X_i) + U_i \text{ and } D_i = m(X_i) + V_i$$

- Split sample in partitions S and S^c with sample sizes $n = N/2$
- Use ML to estimate $\hat{g}(X_i)$ in sample S^c
- Estimate $\hat{\delta}$ in sample S

$$\hat{\delta} = \left(\frac{1}{n} \sum_{i \in S} D_i^2 \right)^{-1} \frac{1}{n} \sum_{i \in S} D_i (Y_i - \hat{g}(X_i))$$

- Regularisation bias

$$\sqrt{n}(\hat{\delta} - \delta) = \left(E[D_i^2] \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i \in S} D_i \left(U_i + (g(X_i) - \hat{g}(X_i)) \right)$$

- $\hat{g}(X_i)$ converges to $g(X_i)$ at rate $n^{-\varphi_d}$, with $\varphi_d < 1/2$ for ML methods
- $\hat{\delta}$ has a convergence rate below $\sqrt{n}$: $|\sqrt{n}(\hat{\delta} - \delta)| \xrightarrow{p} \infty$

Some General Thoughts (cont.)

- Orthogonalised regressor: $\hat{V}_i = D_i - \hat{m}(X_i)$
- Estimate $\hat{\delta}$ in sample S

$$\hat{\delta} = \left(\frac{1}{n} \sum_{i \in S} \hat{V}_i D_i \right)^{-1} \frac{1}{n} \sum_{i \in S} \hat{V}_i (Y_i - \hat{g}(X_i))$$

- Estimation error

$$\sqrt{n}(\hat{\delta} - \delta) = \left(E[V_i^2] \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i \in S} (V_i U_i + (m(X_i) - \hat{m}(X_i))(g(X_i) - \hat{g}(X_i))) + c^*$$

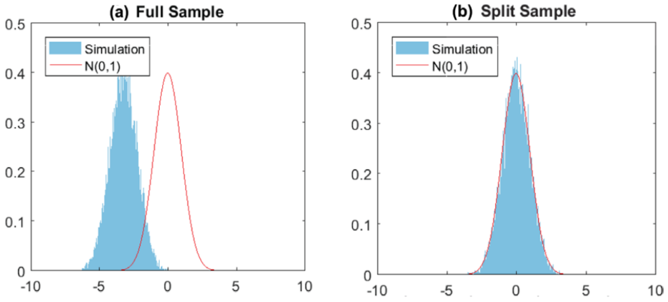
- $\hat{m}(X_i)$ converges to $m(X_i)$ at rate $n^{-\varphi_m}$
- The regularisation bias will vanish at $\sqrt{n}$-rate when $\varphi_g + \varphi_m \geq 1/2$
- Double-robustness property

Role of Sample Splitting

- Remainder term:

$$c^* = \left(E[V_i^2]\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i \in S} \left(U_i(m(X_i) - \hat{m}(X_i)) + V_i(g(X_i) - \hat{g}(X_i))\right)$$

- Vanishes because of sample splitting



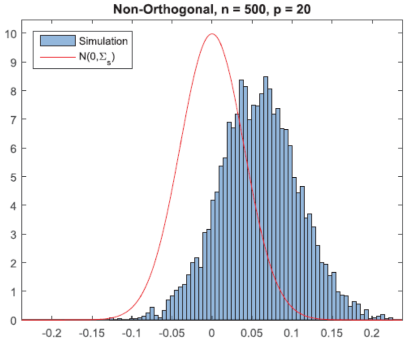
- Loss of efficiency because of sample splitting $\rightarrow$ cross-fitting

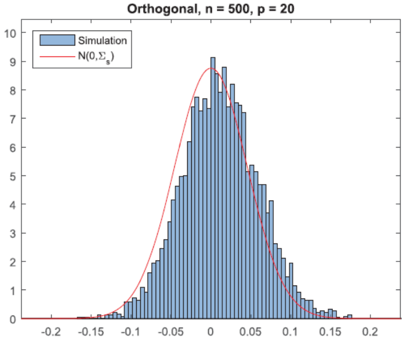
Source: Chernozhukov et al. (2018)

Neyman-Orthogonality

- **General Condition:**
  - Moment Condition:
$$\frac{1}{n}\sum_{i\in S}\psi(W;\hat{\delta}_0,\hat{\eta}_0)=0$$
  - Gateaux derivative:
$$\partial_\eta E[\psi(W;\delta_0,\eta_0][\eta-\eta_0]=0$$
- **Example OLS without orthogonalisation:**
  - Score: $\psi_i = D_i(Y_i - D_i\hat{\delta} - X_i\hat{\beta}_g)$
  - Jacobian: $-E[D_iX_i]\neq 0$
- **Example OLS with orthogonalisation:**
  - Score: $\psi_i = \hat{V}_i(Y_i - D_i\hat{\delta} - X_i\hat{\beta}_g)$
  - Jacobian: $-E[\hat{V}_iX_i]=0$

Simulation Exercise





Lessons learned from general thoughts:

→ Sample splitting is important

→ Orthogonalisation is important

Source: Chernozhukov et al. (2018)

3. Modified Outcome Method

Notation:

- D_i binary treatment dummy (e.g., assignment to training program)
- $Y_i(1)$ potential outcome under treatment (e.g., earnings under participation in training)
- $Y_i(0)$ potential outcome under non-treatment (e.g., earnings under non-participation in training)

Infeasible parameter:

- Individual causal effect: $\delta_i = Y_i(1) - Y_i(0)$

Feasible parameters:

- Average Treatment Effect (ATE): $\delta = E[Y_i(1) - Y_i(0)] = E[\delta_i]$
- Average Treatment Effect on the Treated (ATET): $\rho = E[\delta_i | D_i = 1]$
- Local Average Treatment Effect (LATE): $\gamma = E[\delta_i | Compliers]$

Identifying Assumptions for ATE

- **Stable Unit Treatment Value Assumption (SUTVA):**

$$Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$$

- **Exogeneity of Covariates:**

$$X_i(1) = X_i(0)$$

- **No Support Problems:**

$$\varepsilon < Pr(D_i = 1|X_i = x) = p(x) < 1 - \varepsilon$$

  for some small $\varepsilon \geq 0$ and all x in the support of X_i

- **Conditional Independence Assumption (CIA):**

$$Y_i(1), Y_i(0) \perp\!\!\!\perp D_i|X_i = x$$

  for all x in the support of X_i

Inverse Probability Weighting (IPW)

$$
\begin{aligned}
\delta = E[Y_i(1)] - E[Y_i(0)] &\overset{LIE}{=} \int E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x] f_X(x) dx \\
&\overset{CIA}{=} \int E[Y_i(1)|D_i = 1, X_i = x] - E[Y_i(0)|D_i = 0, X_i = x] f_X(x) dx \\
&= \int E[Y_i|D_i = 1, X_i = x] - E[Y_i|D_i = 0, X_i = x] f_X(x) dx \\
&= \int E[D_i Y_i|D_i = 1, X_i = x] - E[(1 - D_i) Y_i|D_i = 0, X_i = x] f_X(x) dx \\
&\overset{LIE}{=} \int E\left[\frac{D_i Y_i}{p(x)}\middle| X_i = x\right] - E\left[\frac{(1 - D_i) Y_i}{1 - p(x)}\middle| X_i = x\right] f_X(x) dx \\
&= \int E\left[\frac{D_i Y_i}{p(x)} - \frac{(1 - D_i) Y_i}{1 - p(x)}\middle| X_i = x\right] f_X(x) dx \\
&= \int E\left[\frac{D_i - p(x)}{p(x)(1 - p(x))} Y_i\middle| X_i = x\right] f_X(x) dx \overset{LIE}{=} E\left[\frac{D_i - p(x)}{p(x)(1 - p(x))} Y_i\right]
\end{aligned}
$$

with $p(x) = Pr(D_i = 1|X_i = x)$

Reference: Horvitz and Thompson (1952)

Modified Outcome Method

- $Y^*_{i,IPW} = W_i Y_i$ with $W_i = (D_i - p(x))/(p(x)(1-p(x)))$
- ATE: $\delta = E[Y^*_{i,IPW}]$
- We can use standard ML methods to estimate $\hat{p}(x)$ (possibly combined with cross-fitting)
- Goller, Lechner, Moczall, Wolff (2019): “Does the Estimation of the Propensity Score by Machine Learning Improve Matching Estimation? The Case of Germany’s Programmes for Long Term Unemployed”
- **Advantages:**
  - Generic approach
- **Disadvantages:**
  - Potentially omitting “weak outcome confounders” (sparsity assumption on selection equation)
  - Shows weak performance in simulations and applications
  - Moments are not Neyman-orthogonal

Orthogonal Score

$$
\begin{aligned}
\delta =& E\left[\mu_1(x)-\mu_0(x)+\frac{D_i(Y_i-\mu_1(x))}{p(x)}-\frac{(1-D_i)(Y_i-\mu_0(x))}{1-p(x)}\right]\\
=& E\left[\frac{D_i-p(x)}{p(x)(1-p(x))}Y_i+\frac{(p(x)-D_i)\mu_1(x)}{p(x)}-\frac{(D_i-p(x))\mu_0(x)}{1-p(x)}\right]\\
=& \int E\left[\frac{D_i-p(x)}{p(x)(1-p(x))}Y_i+\frac{(p(x)-D_i)\mu_1(x)}{p(x)}-\frac{(D_i-p(x))\mu_0(x)}{1-p(x)}\middle| X_i=x\right]f_X(x)dx\\
=& \int\left(E\left[\frac{D_i-p(x)}{p(x)(1-p(x))}Y_i\middle| X_i=x\right]+\frac{E[p(x)-D_i|X_i=x]}{p(x)}\mu_1(x)\right.\\
&\left.-\frac{E[D_i-p(x)|X_i=x]}{1-p(x)}\mu_0(x)\right)f_X(x)dx\\
=& \int E\left[\frac{D_i-p(x)}{p(x)(1-p(x))}Y_i\middle| X_i=x\right]f_X(x)dx=E\left[Y_i(1)-Y_i(0)\right]
\end{aligned}
$$

with $\mu_1=E[Y_i(1)|X_i=x]$ and $\mu_0=E[Y_i(0)|X_i=x]$

Reference: Robins and Rotnitzki (1995)

Double/Debiased Machine Learning (DML)

- $Y^*_{i,DML} = \mu_1(X_i) - \mu_0(X_i) + \frac{D_i(Y_i - \mu_1(X_i))}{p(X_i)} - \frac{(1-D_i)(Y_i - \mu_0(X_i))}{1-p(X_i)}$
- We can use standard ML methods to estimate $\hat{\mu}_1(x)$, $\hat{\mu}_0(x)$, and $\hat{p}(x)$ (possibly in different samples using cross-fitting)
- **Advantages:**
  - Treatment and outcome equations are modelled explicitly
  - Neyman orthogonality
  - Double robustness properties
  - $\sqrt{N}$-consistent and asymptotically normal
  - More robust than IPW when $p(x)$ is close to zero or one

DML Cross-Fitting Algorithm

1. Split data in samples S^A and S^B
2. Estimate the nuisance parameters $\mu_1^A(x)$,$\mu_0^A(x)$, and $p^A(x)$ in S^A; and $\mu_1^B(x)$,$\mu_0^B(x)$, and $p^B(x)$ in S^B with ML
3. Construct the efficient scores

$$Y_{i,DML}^{A*} = \mu_1^B(X_i) - \mu_0^B(X_i) + \frac{D_i(Y_i - \mu_1^B(X_i))}{p^B(X_i)} - \frac{(1-D_i)(Y_i - \mu_0^B(X_i))}{1-p^B(X_i)}$$

$$Y_{i,DML}^{B*} = \mu_1^A(X_i) - \mu_0^A(X_i) + \frac{D_i(Y_i - \mu_1^A(X_i))}{p^A(X_i)} - \frac{(1-D_i)(Y_i - \mu_0^A(X_i))}{1-p^A(X_i)}$$

4. Calculate ATE,

$$\hat{\delta} = \frac{1}{2}(\underbrace{\hat{E}[Y_{i,DML}^{A*}|S^A]}_{=\hat{\delta}_A} + \underbrace{\hat{E}[Y_{i,DML}^{B*}|S^B]}_{=\hat{\delta}_B}),$$

Asymptotic Results for ATE

- Regularity Condition: Convergence of tuning parameters $\varphi_g + \varphi_m \geq 1/2$
- ATE (and other group averages) can be estimated $\sqrt{N}$-consistently

$$\sqrt{N}(\hat{\delta} - \delta) \xrightarrow{d} N(0, \sigma)$$

with $\sigma^2 = Var(Y^*_{i,DML})$ and $Var(\hat{\delta}) = \sigma^2/N$

- Split sample estimator of σ^2

$$\hat{\sigma}^2 = \frac{1}{2}\left(\hat{\sigma}_A^2 + (\hat{\delta}_A - \hat{\delta})^2\right) + \frac{1}{2}\left(\hat{\sigma}_B^2 + (\hat{\delta}_B - \hat{\delta})^2\right)$$

for $\hat{\delta} = 1/2(\hat{\delta}_A + \hat{\delta}_B)$

Identifying Assumptions for ATET

- **No Support Problems:**

$$Pr(D_i = 1|X_i = x) = p(x) < 1 - \varepsilon$$

for some small $\varepsilon \geq 0$ and all x in the support of X_i

- **Conditional Independence Assumption (CIA):**

$$Y_i(0) \perp\!\!\!\perp D_i|X_i = x$$

for all x in the support of X_i

Orthogonal Score for ATET

$$
\begin{aligned}
\rho =& E\left[\frac{D_i(Y_i-\mu_0(x))}{p}-\frac{p(x)(1-D_i)(Y_i-\mu_0(x))}{p(1-p(x))}\right] \\
=& E\left[\frac{D_iY_i}{p}-\frac{p(x)(1-D_i)Y_i}{p(1-p(x))}-\frac{(D_i-p(x))\mu_0(x)}{p(1-p(x))}\right] \\
=& \int E\left[\frac{D_iY_i}{p}-\frac{p(x)(1-D_i)Y_i}{p(1-p(x))}-\frac{(D_i-p(x))\mu_0(x)}{p(1-p(x))}\middle| X_i=x\right]f_X(x)dx \\
=& \int\left(\frac{E[D_iY_i|X_i=x]}{p}-\frac{p(x)E[(1-D_i)Y_i|X_i=x]}{p(1-p(x))}\right. \\
& \left.-\frac{E[D_i-p(x)|X_i=x]}{p(1-p(x))}\mu_0(x)\right)f_X(x)dx \\
=& \int\left(\frac{E[D_iY_i|X_i=x]}{p}-\frac{p(x)E[(1-D_i)Y_i|X_i=x]}{p(1-p(x))}\right)f_X(x)dx
\end{aligned}
$$

with $p=Pr(D_i=1)$

Orthogonal Score for ATET (cont.)

$$
\begin{aligned}
\rho &= \int \frac{p(x)}{p} \left(E[D_i Y_i | D_i = 1, X_i = x] - E[(1-D_i) Y_i | D_i = 0, X_i = x]\right) f_X(x) dx \\
&= \int \left(E[Y_i(1) | D_i = 1, X_i = x] - E[Y_i(0) | D_i = 0, X_i = x]\right) f_{X|D=1}(x) dx \\
&= \int \left(E[Y_i(1) | D_i = 1, X_i = x] - E[Y_i(0) | D_i = 1, X_i = x]\right) f_{X|D=1}(x) dx \\
&= E[Y_i(1) - Y_i(0) | D_i = 1]
\end{aligned}
$$

- Asymptotic results similar to ATE
- Variance estimator:

$$
\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{D_i(Y_i - \hat{\mu}_0(x))}{\hat{p}} - \frac{\hat{p}(x)(1-D_i)(Y_i - \hat{\mu}_0(x))}{\hat{p}(1-\hat{p}(x))} - \frac{D\hat{\rho}}{\hat{p}} \right)^2
$$

- Calculation of cross-fitted variance corresponding to previous slide (slide 25)

Reference: Chernozhukov et al., 2018

LATE Notation

- Z_i is a binary instrument
- $D(1)$ and $D(0)$ denote the potential treatment states corresponding to the assignment status of the instrument
- Sample can be stratified in four groups denoted by τ_i:
  - a: always-takers $D(1) = D(0) = 1$
  - c: compliers $D(1) > D(0)$
  - n: never-takers $D(1) = D(0) = 0$
  - d: defiers $D(1) < D(0)$
- $Y(1)$ and $Y(0)$ denote the potential outcomes corresponding to the assignment status of the instrument

Reference: Frölich, 2007

Identifying Assumptions for LATE

- **Monotonicity:** $Pr(\tau_i = d) = 0$
- **Existence of Compliers:** $Pr(\tau_i = c) > 0$
- **No Support Problems:**

$$\varepsilon < Pr(Z_i = 1|X_i = x) = e(x) < 1 - \varepsilon$$

for some small $\varepsilon \geq 0$ and all x in the support of X_i

- **Conditional Independence Assumption (CIA):**

$$(Y_i(1), Y_i(0), D_i(1), D_i(0)) \perp\!\!\!\perp Z_i | X_i = x$$

for all x in the support of X_i

LATE

- LATE for binary instrument $Z_i \in \{0,1\}$ (Chernozhukov et al., 2018):
  - First Stage:

$$\gamma_F = E\left[\nu_1(x) - \nu_0(x) + \frac{Z_i(D_i - \nu_1(x))}{e(x)} - \frac{(1-Z_i)(D_i - \nu_0(x))}{1-e(x)}\right]$$

  - Second Stage:

$$\gamma_S = E\left[\omega_1(x) - \omega_0(x) + \frac{Z_i(Y_i - \omega_1(x))}{e(x)} - \frac{(1-Z_i)(Y_i - \omega_0(x))}{1-e(x)}\right]$$

with $e(x) = Pr(Z_i = 1|X_i = x)$, $\nu_1(x) = E[D_i|Z_i = 1, X_i = x]$, $\nu_0(x) = E[D_i|Z_i = 0, X_i = x]$, $\omega_1(x) = E[Y_i|Z_i = 1, X_i = x]$, and $\omega_0(x) = E[Y_i|Z_i = 0, X_i = x]$

LATE

- First Stage:

$$\begin{aligned}\gamma_F =& E[D(1)-D(0)]\\ =& E[D(1)-D(0)|\tau=a]Pr(\tau=a)+E[D(1)-D(0)|\tau=c]Pr(\tau=c)\\ &+E[D(1)-D(0)|\tau=n]Pr(\tau=n)\\ &=Pr(\tau=c)\end{aligned}$$

- Second Stage:

$$\begin{aligned}\gamma_S =& E[Y(1)-Y(0)]\\ =& E[Y(1)-Y(0)|\tau=a]Pr(\tau=a)+E[Y(1)-Y(0)|\tau=c]Pr(\tau=c)\\ &+E[Y(1)-Y(0)|\tau=n]Pr(\tau=n)\\ =& E[Y(1)-Y(0)|\tau=c]Pr(\tau=c)\end{aligned}$$

$\rightarrow$ Apply Wald-estimator: $\gamma=\gamma_S/\gamma_F=E[Y(1)-Y(0)|\tau=c]$

$\rightarrow$ Note: For compliers $Z=D$. Accordingly, γ identifies the effect of D.

Variance of LATE

- Variance estimator:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\Bigg(\Bigg(\hat{\omega}_1(x) - \hat{\omega}_0(x) + \frac{Z_i(Y_i - \hat{\omega}_1(x))}{\hat{e}(x)} - \frac{(1-Z_i)(Y_i - \hat{\omega}_0(x))}{1-\hat{e}(x)}\Bigg)$$
$$-\hat{\gamma}\Bigg(\hat{v}_1(x) - \hat{v}_0(x) + \frac{Z_i(D_i - \hat{v}_1(x))}{\hat{e}(x)} - \frac{(1-Z_i)(D_i - \hat{v}_0(x))}{1-\hat{e}(x)}\Bigg)\Bigg)^2$$

- Calculation of cross-fitted variance corresponding to previous slide (slide 25)

Reference: Chernozhukov et al., 2018

Identifying Assumptions for Difference-in-Differences

- $Y_t(d)$ potential outcome under treatment status d and time period t
- **No Anticipation:**

$$E[Y_0(1) - Y_0(0)|D = 1] = 0$$

- **Conditional Common Trend:**

$$E[Y_1(0) - Y_0(0)|D = 1, X = x] = E[Y_1(0) - Y_0(0)|D = 0, X = x]$$

  for all x in the support of X_i

- **No Support Problems:**

$$\varepsilon < Pr(D_i = 1|X_i = x) = p(x) < 1 - \varepsilon$$

  for some small $\varepsilon \geq 0$ and all x in the support of X_i

Difference-in-Differences

- ATET:

$$\rho = \underbrace{E[Y_1(1)|D=1]}_{\text{Observable}} - \underbrace{E[Y_1(0)|D=1]}_{\text{Counterfactual}}$$

- Identification:

$$\begin{aligned} E[Y_1(0)|D=1,X=x] =& E[Y_0(0)|D=1,X=x] - E[Y_1(0)|D=0,X=x] \\ & + E[Y_0(0)|D=0,X=x] \\ =& E[Y_0(1)|D=1,X=x] - E[Y_1(0)|D=0,X=x] \\ & + E[Y_0(0)|D=0,X=x] \\ =& E[Y|D=1,T=0,X=x] - E[Y|D=0,T=1,X=x] \\ & + E[Y|D=0,T=0,X=x] \end{aligned}$$

- Standard Estimation Model:

$$Y = \beta_0 + \beta_1 D + \beta_2 T + \rho DT + X\beta_3$$

Orthogonal Score for Difference-in-Differences

- Zimmert, 2018:

$$\rho = E\left[\frac{T - p_t}{p_t(1-p_t)}\frac{D_i - p(x)}{p(1-p(x))}(Y_i - \theta_0(x,t))\right]$$

with $\theta_0(x,t) = E[Y_i | D = 0, T = t, X = x]$ and $p_t = Pr(T = 1)$

Other Orthogonal Scores

- Multiple treatments $d \in \{1,2,3,...,\}$ (e.g., Farrell, 2015) :

$$E[Y(d)] = E\left[\frac{1\{D_i = d\}(Y_i - \hat{\mu}_d(x))}{Pr(D_i = d|X_i = x)} + \hat{\mu}_d(x)\right]$$

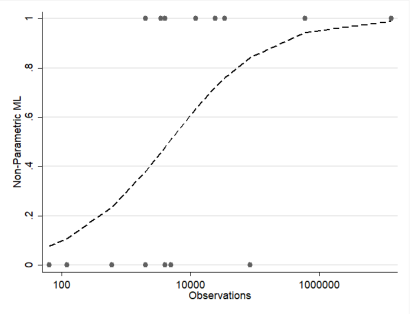
- Continuous treatments see, e.g., Graham and Pinto (2018)
- Mediation analysis see Tchetgen Tchetgen and Shpitser (2012)

4. Practical Considerations

- Lasso or Forest (or other ML method)?
- Sample size?
- Sample partitions (crossvalidation, cross-fitting, honest inference)?
- 1 standard error rule
- Bagging?
- Categorical variables?

Sample size?

Used Observations Sizes	Lasso Application
64	1
120	1
600	1
2,000	0
2,000	1
3,500	0
4,000	0
4,000	1
5,000	1
12,000	0
24,000	0
34,000	0
84,000	1
600,000	0
13,000,000	0



Some Remaining Challenges

- How to deal with support problems?
- Outcomes with limited support?