EC 709: Dealing with Covariates in the Observational Study

Liang Zhong

Boston University

samzl@bu.edu

Novemeber 2023

Overview

Table of Contents

Covariates in the identification assumptions

1. Unconfoundness: $(Y_i(1), Y_i(0)) \perp D_i | X_i \quad \forall i$ for some X_i

   E.g,. Cross-Sectional data
   - Usually omit notation i as $(Y(1), Y(0)) \perp D|X$

2. Conditional Random Assignment of IV: $((Y(1), Y(0), D_1, D_0)) \perp Z|X$
   - 81% of papers using IV included at least one covariate X (Blandhol et al, 2022)

3. Conditional PT:
   $E[Y_{i,t=2}(0) - Y_{i,t=1}(0)|D_i = 1, X] = E[Y_{i,t=2}(0) - Y_{i,t=1}(0)|D_i = 0, X]$

   E.g,. Panel data
   - Allow for covariate-specific trends
   - Not necessarily weaker than canonical PT

? How to handle these covariates in implementation?

- The Rest of the talk assumed common support: For some $\epsilon > 0$, $\epsilon < P(D = 1|X) < 1 - \epsilon$ a.s.

Covariates and Unconfoundness

- Parameter of interest: ATE

$= E(Y(1) - Y(0)) = E_X(E(Y(1) - Y(0)|X)) \equiv E_X(CATE(x))$

$= E_X(E(Y(1)|D = 1, X) - E(Y(0)|D = 0, X))$ (Unconfoundness)

$= E_X(E(Y|D = 1, X) - E(Y|D = 0, X))$ (Observable in practice)

- What if we ran a simple regression: $Y = \beta D + \gamma X + u$?
- Denote $L[D_i|X_i]$ as the best linear predictor of D, $\tilde{D}_i = D_i - L(D_i|X_i)$

  $\Rightarrow E(\tilde{D}_i) = 0, E(\tilde{D}_i X_i) = 0$

- By Frisch-Waugh-Lovell theorem, or called Regression anatomy formula in MHE:

$\Rightarrow$

$$\beta = \frac{E(Y_i \tilde{D}_i)}{E(\tilde{D}_i^2)}$$

Decompose the OLS coefficient

$$\beta = \frac{E(Y_i \tilde{D}_i)}{E(\tilde{D}_i^2)} = \frac{E_{D,X}(E(Y_i|D_i, X_i)\tilde{D}_i)}{E(\tilde{D}_i^2)} \text{(Law of Iterated Expectation)}$$

$$= \frac{E_{D,X}(E(D_i Y_i(1) + (1 - D_i) Y_i(0)|D_i, X_i)\tilde{D}_i)}{E(\tilde{D}_i^2)} \text{(Definition of } Y_i\text{)}$$

$$= \frac{E_X(D_i CATE(X)\tilde{D}_i))}{E(\tilde{D}_i^2)} + \frac{E_X(E(Y_i(0)|X_i)\tilde{D}_i)}{E(\tilde{D}_i^2)} \text{(Unconfoundness)}$$

$$\equiv E_X[\omega(X_i) CATE(X_i)] + \delta$$

- $\omega(X_i) \equiv \frac{D_i(D_i - L(D_i|X_i))}{E((D_i - L(D_i|X_i))^2)} \Rightarrow$ By $E(\tilde{D}_i X_i) = 0$, $E[\omega(X_i)] = 1$
- $\delta \equiv \frac{E(E(Y_i(0)|X_i)\tilde{D}_i)}{E(\tilde{D}_i^2)}$

How to interpret the OLS coefficient?

- $\beta = E[\omega(X_i)CATE(X_i)] + \delta$

1. If $E(D_i|X_i) = L[D_i|X_i] \Rightarrow E(\tilde{D}_i|X_i) = 0 \Rightarrow \beta = E[\omega(X_i)CATE(X_i)]$
   - Can write $\omega(X_i) = \frac{Var(D_i|X_i=x)}{E(Var(D_i|X_i))} \geq 0$
   - $\beta = ATE$ when $CATE(X_i)$ and $\omega(X_i)$ are uncorrelated (e.g,. either of them are constant)
   - More weights on X_i that has a lot of variation on D_i

   $\Rightarrow$ Not the desired weight for ATE

2. Typically $E(D_i|X_i) \neq L[D_i|X_i]$

$\Rightarrow$ β no longer a convex weighted average of $CATE(X_i)$

- Moreover, weight can be negative when $L(D_i|X_i) > 1$

$\star$ In general, either cases would lead to $\beta \neq ATE$

Covariates and IV

- Parameter of interest: LATE; Denote compliers as C:

$= E(Y(1) - Y(0)|C) = E_{X|C}(E(Y(1) - Y(0)|C, X)) \equiv E_{X|C}(LATE(X))$

$= E_{X|C}(\frac{E(Y|X,Z=1)-E(Y|X,Z=0)}{E(D|X,Z=1)-E(D|X,Z=0)})$ (Similar procedure as the no covariate case)

  - Using the fact that $P(C|X) = E(D|X, Z = 1) - E(D|X, Z = 0)$ (Frolich, 2007)

$= \frac{E_X(E(Y|X,Z=1)-E(Y|X,Z=0))}{E_X(E(D|X,Z=1)-E(D|X,Z=0))}$ (Observable in practice)

  - Same as Wald estimator if no covariates

- Blandhol et al (2022): 2SLS with covariates doesn't give us LATE
  - Discussed last time
  - Similar intuition to the OLS case above

Covariates and Conditional PT

- Parameter of interest: ATT

$= E(Y_{t=2}(1) - Y_{t=2}(0)|D=1) = E_{X|D=1}(E(Y_{t=2}(1) - Y_{t=2}(0)|D=1,X)) \equiv E_{X|D=1}(ATT(X))$

$= E_{X|D=1}(E(Y_{t=2}(1) - Y_{t=1}(0)|D=1,X) - E(Y_{t=2}(0) - Y_{t=1}(0)|D=0,X))$ (By conditional PT)

$= E[Y_{t=2} - Y_{t=1}|D=1] - E_{X|D=1}(E[Y_{t=2} - Y_{t=1}|D=0,X])$ (Observable in practice)

- In practice, people often use TWFE with covariates
  - Causing huge bias if potential outcome and treatment assignments are related to covariates
  - Simulation exercise can be found in the previous slides
  - The weighted sum formula like the OLS case can be found in Lin and Zhang (Economics Letters, 2022)

Choice of covariates in Conditional PT

- For panel data, a more tricky issue is the choice of X: X_i or $\{X_{it}, t = 1, ..., T\}$?
- X_i: Conditional on pre-treatment covariates only
  - Safe choice, but might not be enough for PT to hold
- $\{X_{it}, t = 1, ..., T\}$: Conditional on time-varying covariates in all time periods
  - Need to make sure the temporal change of $X_{i,t}$ is not caused by the policy
  - Post-treatment bias: covariates measured after treatment may obscure the causal effect (Caetano et al, 2022)
- In general, adding more controls need not bring you closer to identification!
  - See MHE discussion of "bad controls," or more sophisticated discussions of "collider bias" from the DAG literature (Cinelli, Forney, and Pearl, 2022)

Summary of the problem

1. Linear regressions implicitly restrict the effects to be homogeneous

$\Rightarrow$ Constant effect is a strong assumption; we'd like to avoid it when possible

- Seems likely that effects vary across both observables & unobservables
- For binary Y_i (or other limited support), the constant effect is impossible

2. All the parameters of interest are related to the conditional average potential outcome $E(Y|D=1,X) - E(Y|D=0,X)$

- In LATE, "D" is the IV Z, "Y" can be either Y or D
- In ATT of DID, "Y" is $Y_{t=2} - Y_{t=1}$, and we also need to take care of $E_{X|D=1}$
- How to deal with them in practice?

Table of Contents

How about matching using covariates?

- How to estimate $E(Y|D=1,X) - E(Y|D=0,X)$?
  1. Match treated and control observations with the same value of X_i
  2. Estimate $E[Y_i|D_i=1,X_i=x] - E[Y_i|D_i=0,X_i=x]$ for each x
  3. Average these estimates together by the marginal distribution of X_i
- Simple? Matching can be tricky when X_i takes on many values / has many rows (The Curse of Dimensionality)
- Denote the dimension of X as k
- Larger $k \Rightarrow$ more plausible identification

  $\Rightarrow$ Larger matching discrepancies due to limited sample size

  E.g,. Subclassification: divide each covariates into 2 coarse categories (e.g., age would be "young" or "old", and income would be "low" or "high")
  - Number of subclassification cells is 2^k. With $k=10$, we obtain $2^{10}=1024$, not enough observations in each cell

  $\Rightarrow$ Estimators for $E[Y|D=0,X]$ might not be able to converge
- Smaller $k \Rightarrow$ each "cells" are "too coarse", make identification problematic
- What to do next?

Solution 1: Regression-adjusted estimators

- How about estimate $E(Y|D=1,X)$ and $E(Y|D=0,X)$ separately?
  1. Estimate $E[Y|X,D=1]$ and $E[Y|X,D=0]$ using your favorite method. Denote these by $\hat{\mu}_{1,n}(X_i)$ and $\hat{\mu}_{0,n}(X_i)$, respectively
  2. Use estimated regressions to produce analog estimators: $\hat{ATE}_n = \frac{1}{n}\Sigma_{i=1}^{n}(\hat{\mu}_{1,n}(X_i) - \hat{\mu}_{0,n}(X_i))$; $\hat{ATT}_n = \frac{1}{n_1}\Sigma_{i=1}^{n} D_i(Y_i - \hat{\mu}_{0,n}(X_i))$

- Rely on researchers' ability to model the potential outcome
  - In the ATT case, basically impute $Y(0)$ for the treated groups using the model estimated from the control groups
  - People often Use linear regression model, but other semi-parametric or non-parametric models can be used as well
  - See a nice discussion in Wooldridge (2010) for the choice of models (Quasi-MLE has been mentioned a lot)
- Similar idea for DID and LATE, Take DID as an example:

$$ATT = E[Y_{t=2} - Y_{t=1}|D=1] - E_{X|D=1}(\hat{\mu}_{t=2}^{D=0}(X_i) - \hat{\mu}_{t=1}^{D=0}(X_i))$$

Solution 2: Using Propensity Scores

- Rather than matching on X_i, it's enough to match on the scalar propensity score $p(X_i) = Pr(D_i = 1|X_i)$ (Rosenbaum & Rubin, 1983)

Prop: $(Y_i(0), Y_i(1)) \perp D_i|X_i$ implies $(Y_i(0), Y_i(1)) \perp D_i|p(X_i)$

Proof:
$$\begin{aligned} & Pr(D_i = 1|p(X_i), Y_i(0), Y_i(1)) = E[D_i|p(X_i), Y_i(0), Y_i(1)] \\ = \; & E[E[D_i|X_i, p(X_i), Y_i(0), Y_i(1)]|p(X_i), Y_i(0), Y_i(1)] \\ = \; & E[E[D_i|X_i]|p(X_i), Y_i(0), Y_i(1)] \text{ (By Unconfoundness)} \\ = \; & E[p(X_i)|p(X_i), Y_i(0), Y_i(1)] \\ = \; & p(X_i) = Pr(D_i = 1|p(X_i)) \end{aligned}$$

- This suggests a two-step procedure to estimate causal effects under the unconfoundedness setup:
  1. Estimate the propensity score $p(X)$, using e.g. logit regression
  2. Conduct matching or subclassification on the estimated propensity score

⋆ Substantial dimension reduction (as long as we know $p(X)$)

- Rely on researchers' ability to model the propensity score

Inverse Probability Weighting Estimators

- Can also weight inversely by $p(X_i)$

Prop: For any function ϕ, $E[\phi(Y(1)) - \phi(Y(0))] = E[\frac{D}{p(X)}\phi(Y)] - E[\frac{1-D}{1-p(X)}\phi(Y)]$

Proof: Let $\tau^{\phi}(X) \equiv E[\frac{D}{p(X)}\phi(Y)|X] - E[\frac{1-D}{1-p(X)}\phi(Y)|X]$

$= E[\frac{1}{p(X)}\phi(Y)|X, D=1]p(X) - E[\frac{1}{1-p(X)}\phi(Y)|X, D=0](1-p(X))$
(By definition of Expectation)

$= E[\phi(Y)|X, D=1] - E[\phi(Y)|X, D=0]$

$= E[\phi(Y(1))|X, D=1] - E[\phi(Y(0))|X, D=0]$

$= E[\phi(Y(1)) - \phi(Y(0))|X]$ (By unconfoundness)

- Comparison between propensity score matching v.s. weighting:
  - Matching method tends to have lower bias but higher variance
  - Weighting method tends to have higher bias but lower variance
  - No one dominates the other (See Busso, DiNardo, and McCrary (2009, 2014))

Inverse Probability Weighting for ATT

Prop: $E[\phi(Y(1)) - \phi(Y(0))|D=1] = \frac{1}{P(D=1)}(E[D\phi(Y)] - E[p(X)\frac{1-D}{1-p(X)}\phi(Y)])$

Proof: $E(\tau^{\phi}(X)|D=1) = \int \tau^{\phi}(x)F(dx|D=1)$

$= \frac{\int \tau^{\phi}(x)P(D=1|X=x)F(dx)}{\int P(D=1|X=x)F(dx)}$ (By Bayes' Theorem)

$= \frac{\int \tau^{\phi}(x)p(x)F(dx)}{P(D=1)}$

$= \frac{E(\tau^{\phi}(X)p(X))}{P(D=1)}$

- Might prefer normalized weights, and replace $P(D=1) = E[p(X)\frac{1-D}{1-p(X)}]$
  - It is often called Hájek (1971)-type estimators can be more stable
- Similar for DID and LATE, still take DID as an example:

$$ATT = E[(\frac{D}{E(D)} - \frac{p(X)\frac{1-D}{1-p(X)}}{E(p(X)\frac{1-D}{1-p(X)})})(Y_{t=2} - Y_{t=1})]$$

Estimate Propensity score in Practice

1. Try to approximate the treatment assignment process as closely as possible

   E.g., Logit/Probit, mostly works fine
   - See Abadie and Imbens (2016) for matching and Hirano, Imbens, and Ridder (2003) for weighting

- Also a large/growing ML literature to flexibly model the propensity score:

i. Bayesian Additive Regression Trees (BART, Hill et al., 2011)
   - BART is a sum-of-trees-approach that uses a Bayesian prior to prevent overfitting while allowing the model to be very flexible

ii. SuperLearner: A stacking method that allows you to supply many different machine learning methods (Pirrachio et al., 2015)
   - Either picks the best one or takes an optimally weighted combination of them

- JJ will talk about it later
- Not all ML methods can work with propensity score weighting: irregular functional form would make inference very hard

Balance weighting method

- The whole idea of propensity score relies on the finding that balancing on a well-formed propensity score balances **all pre-treatment covariates fully**
  - In practice, even if we tried the most flexible model, there is almost no hope of correctly modeling the treatment process to obtain propensity scores

2. Try to obtain propensity scores that yield covariate balance

- Imai and Ratkovic (2014, JRSS): Estimate γ use a logit model by restricting:

$$E(X') = E(\frac{D}{p(X\gamma)}X') = E(\frac{1-D}{1-p(X\gamma)}X')$$

  - Ensures the weighted means of all covariates are the same in control and treated subsamples
  - With misspecification, tends to work better than MLE-based weights
  - User-written Stata command available: **psweight**

$\star$ Regardless of which approach you choose, you should assess balance on your covariates

Table of Contents

Challenge 1: Misspecification

- Regression based treatment effects estimators requires correctly specified regression model for the outcome of interest
- Inverse probability weighting based treatment effects estimators requires correctly specified propensity score model for $p(X)$
- Although in practice, both models are likely to be misspecified, anything we can do to make us feel more comfortable?

$\Rightarrow$ Doubly robust (DR) estimator: combines the regression and the IPW approach

  - Also called augmented inverse probability weighting
  - the estimator will be consistent if either putative regression or propensity score model is correctly specified

- Even though both model misspecified, DR is more efficient as long as the overlap condition holds
  - See Busso, DiNardo, and McCrary (2009, 2014) for simulation results
  - Sant'Anna and Zhao (2020) also shown that DR for did is "semi-parametrically efficient"(confidence interval are tighter)

Doubly Robust estimation

- ATE:

$$\frac{1}{n}\sum_{i=1}^{n}[\frac{D_i}{\hat{p}_n(X_i)}Y_i + (1-\frac{D_i}{\hat{p}_n(X_i)})\hat{\mu}_{1,n}(X_i)]$$

$$-\frac{1}{n}\sum_{j=1}^{n}[\frac{1-D_j}{1-\hat{p}_n(X_j)}Y_j + (1-\frac{1-D_j}{1-\hat{p}_n(X_j)})\hat{\mu}_{0,n}(X_j)]$$

  - Taking the IPW estimator and "augmenting" it by a second term
  - When $Y_i = \hat{\mu}_n(X_i)$, back to the regression-based method
  - When $\hat{p}_n(X_i) = p(X_i)$, $E(\frac{D_i}{\hat{p}_n(X_i)}) = 1$, back to the IPW estimator
- Functional form for LATE and DID are following the same idea
- DID package: DRDID
- LATE: STATA command *drlate* is available

Challenge 2: Limited Overlap

- The ATE is only identified when $p(X_i)$ is bounded away from zero and one
  - $\Leftarrow$ Intuitively, can't identify effects at X_i where $D_i = 0$ or $D_i = 1$ always
- ATE estimators are likely to be very noisy of $p(X_i)$ is ever near zero or one
  - $\Leftarrow$ Intuitively, need a lot of data to estimate effects at such X_i

- The finite-sample performance of ATE estimators under limited overlap can be improved by "trimming" propensity scores near 0 and 1
  - $\Leftarrow$ Trimming in large samples changes the estimand, from ATE to a weighted-average $CATE(X_i)$ among X_i with non-trimmed $p(X_i)$ (Crump et al. 2009)
    - Without trimming, all matching methods have bad performances

Beyond Matching

- Even without the practical challenge above, in practice P-score matching/weighting can be a little involved
  - How to conduct inference?
  - Some packages exist (e.g. teffects in Stata), but results highly depends on the method you choose, even if overlap is decent (recall your experience in the problem set)
- What else can we try?
- Recall the key issue in linear regression is $L(D|X) \neq E(D|X)$
- What if we try regression $Y = \beta D + g(X) + u$, where $g(X)$ is a nonparametric function of X?
- $L(D|g(X)) \approx E(D|X)$? Then we might at least have a convex combination of CATE

$\Rightarrow$ One of the motivation for the usage of Semiparametric methods and Double Machine Learning

Thank You!