Differences in Differences

Compass-Lexecon

Raquel Carrasco

Universidad Carlos III de Madrid

May-June 2024

Outline Session 3

- Semiparametric Approaches in Causal Analysis: adding covariates
  - TWFE Assumptions for DD with covariates
    - Restrictions implied by the TWFE Framework
  - Alternative estimations:
    - Abadie (2005)
    - Sant'Anna and Zaho (2020)

Adding Covariates in the TWFE model

- A strand of the literature has focused on the possibility that the canonical parallel trends assumption may not hold exactly.
- One way to increase the credibility of the parallel trends assumption is to require that it holds only conditional on covariates.
- The idea is to compare the paths of outcomes among treated and untreated units conditional on having the same characteristics.
- By including covariates in a DD design we are trying to address a parallel trends violation
- Generally, the motivation is different in regression

Adding Covariates in the TWFE model

- Regression: we include covariates to get "unconfoundedness"

$$(Y_0, Y_1) \perp D \mid X$$

  - It means that conditional on X (e.g., age, gender, race...), D is assigned to units independent of their potential outcomes.
  - With the inclusion of the covariates, you have "isolated" a randomized experiment
- DD: The inclusion of covariates is not about trying to find random variation in the treatment within values of the dimension of X
  - It is based on the claim that the inclusion of covariates is necessary to re-establish parallel trends
- TWFE can only incorporate time varying covariates, and places restrictions on the model.

Adding Covariates

- The advantage of the TWFE model (regression formulation) is the easiness of obtaining the final estimates and their standard errors.
  - We can easily extend the model to cover more periods and more treatments, and add additional covariates.
- Disadvantages:
  - (i) The way how control variables are included: Including the control variables in a linear in parameters fashion implies the assumption of common trends conditional on the linear index, $\mathbf{X}'\pi$, which is more restrictive than assuming common trends conditional on $\mathbf{X}$.
  - (ii) Any deviation from the linear index is again absorbed in the regression error term and may introduce bias in the estimates.
- Ideally, covariates should be treated non-parametrically, so that any potential inconsistency created by functional form misspecification is avoided.

Adding Covariates

- But, not only that¡
- TWFE specification without covariates:

$$Y_{it} = \gamma_1 + \gamma_2 T_t + \gamma_3 D_i + \delta(T_t \times D_i) + e_{it}$$

- Under PT: $ATT = \delta$
- Is just adding covariates right?

$$Y_{it} = \gamma_1 + \gamma_2 T_t + \gamma_3 D_i + \delta(T_t \times D_i) + \theta X_{it} + e_{it}$$

  - Yes, if your are willing to impose extra assumption, in addition to conditional PT
  - (General recommendation is to include only "pre-treatment" controls, unless you have a strong argument that post treatment characteristics are strictly exogenous)

Basic Assumptions for DD with covariates

- Assumption 1: Assume panel data or repeated cross sections
- Assumption 2: Conditional Parallel Trends:

$$E\left[Y_0(2) - Y_0(1) \middle| D = 1, X\right] = E\left[Y_0(2) - Y_0(1) \middle| D = 0, X\right]$$

- Assumption 3: Common support or overlap (this is just about the distribution of treated and controls across values of X): $0 < Pr(D = 1 \mid X) < 1$
  - For any value X a unit i can be potentially observed with treatment $(D = 1)$ and without treatment $(D = 0)$.

TWFE with covariates

- TWFE places restrictions on the DGP
- Let's allow for the possibility that the effect of X is different for different periods and for treated and controls:
  - θ_{12} : the effect of X for treated in the second period. Similarly: $\theta_{11}, \theta_{02}, \theta_{01}$
- TWFE under Assumptions 1-3 implies:

$$\underbrace{E\left[Y_1(2) | D=1, X\right]}_{\text{observed}} = \gamma_1 + \gamma_2 + \gamma_3 + \delta + \theta_{12} X_{12}$$

- Conditional PT implies:

$$\begin{aligned} & E\left[Y_0(2) | D=1, X\right] \\ = \; & E\left[Y(1) | D=1, X\right] + E\left[Y(2) | D=0, X\right] - E\left[Y(1) | D=0, X\right] \end{aligned}$$

TWFE with covariates

- This gives:

$$\begin{aligned} & E\left[Y_0(2)|\, D=1, X\right] = \\ & (\gamma_1+\gamma_3+\theta_{11}X_{11})+(\gamma_1+\gamma_2+\theta_{02}X_{02})-(\gamma_1+\theta_{01}X_{01}) \\ = \quad & \gamma_1+\gamma_2+\gamma_3+\theta_{11}X_{11}+\theta_{02}X_{02}-\theta_{01}X_{01} \end{aligned}$$

- By replacing and collecting terms:

$$\begin{aligned} ATT \quad = \quad & E\left[Y_1(2)|\, D=1, X\right]-E\left[Y_0(2)|\, D=1, X\right] = \\ & \delta+\underbrace{\left[(\theta_{12}X_{12}-\theta_{11}X_{11})-(\theta_{02}X_{02}-\theta_{01}X_{01})\right]}_{\text{cancels out?}} \end{aligned}$$

TWFE with covariates

- Thus

$$\begin{aligned} ATT &= E\left[Y_1(2)|D=1,X\right] - E\left[Y_0(2)|D=1,X\right] \\ &= \delta + \left[(\theta_{12}X_{12-}\theta_{11}X_{11}) - (\theta_{02}X_{02-}\theta_{01}X_{01})\right] \end{aligned}$$

- Bias from TWFE, which is zero under additional assumptions:
- Assume that the effect of X for the treated (controls) is the same in the two periods:

$$\theta_{12} = \theta_{11} = \theta_1$$

$$\theta_{02} = \theta_{01} = \theta_0$$

Now take DD:

$$\begin{aligned} ATT &= \delta + \left[(\theta_1(X_{12-}X_{11}) - \theta_0(X_{02} - X_{01})\right] \\ &= 0 \text{ if trends in } X \text{ for } D=1 \text{ equals trends in } X \text{ for } D=0 \\ \text{and } \theta_1 &= \theta_0 \end{aligned}$$

TWFE with covariates

- Assumption 4: The effect of X on the outcome is homogenous:

$$\theta_1 = \theta_0 = \theta$$

- But other assumptions required: what about X specific trends?
- Now take DD:

$$\begin{aligned} ATE &= \delta + \theta \underbrace{[(X_{12} - X_{11}) - (X_{02} - X_{01})]} \\ &= 0 \text{ if trends in } X \text{ for } D = 1 \text{ equals trends in } X \text{ for } D = 0 \end{aligned}$$

- We need parallel trends in X

TWFE with covariates

- Assumptions 5 and 6: For $D = 0, 1$ we need "no $X-$specific trends for the treatment group (Assumption 5) and comparison group (Assumption 6):

$$E\left[Y(2) - Y(1) \middle| D = d, X\right] = E\left[Y(2) - Y(1) \middle| D = d\right]$$

- Intuition: No $X-$specific trends means you cannot allow women to be on a different trend than men, for instance.
- Without these additional assumptions, in general TWFE will not identify ATT.
  - It will be biased of unknown magnitude and sign.

DD with covariates: proposals

- If we need X for conditional PT, we have 4 options:

(1) TWFE: Assumptions 1-6

(2) Outcome Regression (OR), (Heckman et al. 1997): Assumptions 1-3

(3) Propensity Score-Inverse Probability Weighting (IPW), (Abadie, 2005): Assumptions 1-3

  - (2) and (3) need certain models to be correctly specified

(4) Double Robust DD (Sant'Anna and Zhao, 2020): combines (2) and (3) into one specification.

  - It works as far one of the two, propensity score or outcome regression, to be correctly specified.

Outcome Regression (OR): Heckman et al. (1997)

- This is the Heckman et al. (1997) approach: uses baseline X and control group only to impute the missing counterfactual $Y_0(2)$ for treatment group in a DD equation (the outcome evolution is modelled with a regression):

$$DD^{OR} = \overline{Y}_{1,2} - \overline{Y}_{1,1} - \left[\frac{1}{n_{treat}} \sum_{i|D_i=1} \left(\widehat{\mu}_{0,2}(X_i) - \widehat{\mu}_{0,1}(X_i) \right) \right],$$

  where $\overline{Y}_{1,t}$ is the sample average of Y for the treated in period t, $\widehat{\mu}_{0,t}$ is an estimator of $E[Y(t)|D=0,X]$.

- $\widehat{\mu}_{0,t}$ is the predicted outcome from a regression for the control group given its value X :
  - Estimate a regression for controls only
  - Get fitted values of Y for each individual based on her X values
    - Use the betas for the control to make prediction for the treated
- OR needs $\widehat{\mu}(X_i)$ to be correctly specified.

IPW: Abadie (2005)

- Abadie (2005) proposed a DD-type estimator, but it is not using TWFE.
  - His solution was based on re-weighting the comparison group using a propensity score based on X (Semiparametric DD, SDD).
- The SDD is a re-weighting technique that addresses the imbalance of characteristics between treated and untreated groups.
  - Hence, it makes the parallel trend assumption more credible.
- Three step method:
  - Compute each unit "after minus before", which is the DD part.
  - Estimate a propensity score which you will use to weight each unit.
  - Compare weighted changes in "after minus before" for treatment versus comparison groups.
  - Since ATT is only missing $Y_0(2)$ for treated, we only have to apply weights to the controls

IPW: Abadie (2005)

- Abadie (2005) assumes that, conditional on the covariates, the average outcomes for treated and controls would have followed parallel paths in absence of the treatment:

$$E\left[Y_0(2) - Y_0(1) | \mathbf{X}, D = 1\right] = E\left[Y_0(2) - Y_0(1) | \mathbf{X}, D = 0\right]$$

- Under previous assumption, the effect of the treatment on the treated conditional on $\mathbf{X}$ can be expressed as:

$$\begin{aligned} &E\left[Y_1(2) - Y_0(2) | \mathbf{X}, D = 1\right] \\ &= \left\{E\left[Y(2) | \mathbf{X}, D = 1\right] - E\left[Y(2) | \mathbf{X}, D = 0\right]\right\} \\ &- \left\{E\left[Y(1) | \mathbf{X}, D = 1\right] - E\left[Y(1) | \mathbf{X}, D = 0\right]\right\} \end{aligned}$$

- In principle, the identification result in previous equation can be used to estimate $E\left[Y_1(2) - Y_0(2) | \mathbf{X}, D = 1\right]$ by producing non-parametric estimates of each one of the four expectations on the RHS of the equation.

IPW: Abadie (2005)

- But in practice, the number of observations required to attain an acceptable precision for this type of non-parametric estimator increases very rapidly with the dimension of $\mathbf{X}$.
- This problem, often called the curse of dimensionality, may prevent us from using non-parametric estimators in many practical instances.
- Abadie (2005) estimator is semi-parametric in the sense that it imposes parametric restrictions on the propensity score (estimated with series logit or linear probability models), but the covariates are treated non-parametrically for identification.

IPW: Abadie (2005)

- Abadie (2005) proposes simple weighting schemes to produce estimators of the average effect on the treated $E\left[Y_1(2) - Y_0(2) \middle| D = 1\right]$.
  - The weighting scheme is directly based on the propensity score, $\Pr\left(D = 1 \middle| \mathbf{X}\right)$, which reduces the dimensionality of X into a single scalar.
- Think in the propensity score as dimension reduction method
  - Propensity score theorem (Rosenbaum and Rubin 1983) showed that if you need X to satisfy some assumption, the propensity score will satisfy too
  - Propensity score essentially transform your large dimensional problem into a single scalar called the propensity score, which is the conditional probability of treatment (conditional on X)
  - But we need to estimate the propensity score because we don't usually know it.

IPW: Abadie (2005)

- Since the object of study is the effect of the treatment on the treated, we need that some fraction of the population is exposed to the treatment.

$$\Pr(D = 1) > 0$$

- Common-Support assumption: Since identification is attained after controlling for the effect of some covariates, it will be required that for each given value of the covariates there is some fraction of the population that remains untreated and can be used as controls:

$$0 < \Pr(D = 1 \mid X) < 1.$$

  - This means that within all combinations of the covariates (e.g., across age, gender, race...) there are units in treatment and control groups

IPW: Abadie (2005)

- Abadie (2005) shows that the average effect of the treatment for the treated is given by:

$$E\left[Y_1(2) - Y_0(2) \mid D=1\right] = E\left[\frac{Y(2)-Y(1)}{\Pr(D=1)} \times \frac{D - \Pr(D=1 \mid X)}{1-\Pr(D=1 \mid X)}\right]$$

- For $D=1$: $E\left[\frac{Y(2)-Y(1)}{\Pr(D=1)}\right]$
- For$D=0$: $E\left[-\frac{Y(2)-Y(1)}{\Pr(D=1)} \times \frac{\Pr(D=1|X)}{1-\Pr(D=1|X)}\right]$

IPW: Abadie (2005)

- In words, under previous assumptions, a simple weighted average of temporal differences in the outcome variable recovers the average effect of the treatment for the treated.
  - The weights depend on the propensity score.
- The estimator is a weighted average of the difference of trend across groups (for the treated the weight is equal to 1).
  - It weights the trend for the untreated participants based on their propensity score $\Pr(D = 1 \mid X)$.
  - As $\frac{\Pr(D=1|X)}{1-\Pr(D=1|X)}$ is an increasing function of $\Pr(D = 1 \mid X)$, untreated with higher propensity score are given higher weight.
  - Example: if $\Pr(D = 1 \mid X) = 0.8$, weight $\frac{0.8}{1-0.8} = 4$; if $\Pr(D = 1 \mid X) = 0.2$, weight $\frac{0.2}{1-0.2} = 0.25$

IPW: Abadie (2005)

- Two-step method: (i) estimate the propensity score, $\Pr(D = 1 \mid X)$, and compute the fitted values; (ii) plug the fitted values into the equation to obtain an estimate of $E\left[Y_1 (2) - Y_0 (2) \middle| D = 1 \right]$.
- In particular,

$$DD^{IPW} = \widehat{p}^{-1} N^{-1} \sum_{i=1}^{N} \left[\frac{D_i - \widehat{p}(X_i)}{1 - \widehat{p}(X_i)} \left(Y_i (2) - Y_i (1) \right) \right]$$

  where $\widehat{p}$ is the fraction of units treated and the propensity score $\widehat{p}(X_i)$ is the propensity score
- The command `asdid` in Stata implements this estimator.

IPW: Abadie (2005)

- The PS can be approximated semiparametrically using a polynomial series of the predictors (LPM) or with a logit estimator:

$$\begin{aligned} \Pr(D &= 1 \mid X) = F(\alpha_0 + \alpha_1 X + \alpha_2 X^2 ...) \\ \Pr(D &= 1 \mid X) = \lambda_0 + \lambda_1 X + \lambda_2 X^2 ... \end{aligned}$$

- How to choose the PS?
  - Theoretical Considerations
  - Goodness-of-Fit Measures: Use measures like Akaike Information Criterion to compare models with different polynomial levels.
  - Cross-Validation: Split your data into training and testing sets, fit the model on the training set, and evaluate its performance on the testing set. This helps in assessing the model's out-of-sample predictive power.
  - `pstest`: This Stata package implements the class of specification test for the PS proposed in Sant'Anna and Song (2019).
- `absdid` uses a LPM to estimate the propensity score.

IPW: Abadie (2005)

- Previous discussion assumes constant ATT.
- But this estimator allows also the effect of the treatment to differ among individuals
  - In many practical instances, the desired level of aggregation is lower than the entire treated population and we want to study how the treatment affects the treated for different groups of the population:
  - Consider the situation in which we need to condition on some vector of random variables X to attain identification, but we are interested in $E\left[Y_1(2) - Y_0(2) \middle| X_k, D = 1\right]$, where X_k is some deterministic function of X.

OR or IPW: Caveat

- Outcome Regression needs $\widehat{\mu}$ to be correctly specified.
- IPW needs $\widehat{p}(X_i)$ to be correctly specified.
- Each is inconsistent when their own models are misspecified
- It's hard to "rank" these two in practice with regards to model misspecification because each is inconsistent when their own models are misspecified.

OR or IPW: Caveat

- Double Robust combines them.
  - Doubly Robust estimators are consistent if either the outcome model or the propensity score is correctly specified.
  - Double robust means the method does both of these at the same time so that you don't have to choose between them
  - DR basically accounts for X twice: with a linear regression, with a propensity score, just in case...Gives you two chances to be wrong
- The command `drdid` in Stata implements this estimator.

DR: Sant'Anna and Zhao (2020)

- The double robustness is achieved by combining the outcomes of two models:
  - A propensity score model, which predicts the probability of treatment assignment, and
  - An outcome regression model, which estimates the relationship between treatment and outcome variables.
- By combining information from both models, the proposed estimator improves efficiency compared to traditional DD estimators.
  - This means more precise estimates of treatment effects can be obtained with the same amount of data, or equally precise estimates can be obtained with less data.

DR: Sant'Anna and Zhao (2020)

$$DD^{DR} = E\left[\left(\frac{D}{E[D]} - \frac{\frac{p(X)(1-D)}{(1-p(X))}}{E\left(\frac{p(X)(1-D)}{(1-p(X))}\right)}\right)\left(\Delta Y - \mu_{0,\Delta}(X)\right)\right]$$

$$\begin{aligned} p(X) &: \text{ propensity score model} \\ \Delta Y &= Y(2) - Y(1) = Y_{post} - Y_{pre} \\ \mu_{0,\Delta} &= \mu_{0,1}(X) - \mu_{0,0}(X), \text{ where } \mu(X) \text{ is a model for} \\ & \quad m_{0,t} = E[Y_t | D = 0, X = x] \end{aligned}$$

So that means $\mu_{0,\Delta}$ is just the control group's change in average Y for each $X = x$

DR: Sant'Anna and Zhao (2020)

$$DD^{DR} = E\left[\left(\frac{D}{E[D]} - \frac{\frac{p(X)(1-D)}{(1-p(X))}}{E\left(\frac{p(X)(1-D)}{(1-p(X))}\right)}\right)\left(\Delta Y - \mu_{0,\Delta}(X)\right)\right]$$

- Notice how the model controls for X: you're weighting the adjusted outcomes using the propensity score
- The reason you control for X twice is because you don't know which model is right.
- The idea of the doubly robust – you only need one of these models to be correctly specified, not both.
- For $D = 1$: $\Delta Y - \mu_{0,\Delta}(X)$
- For$D = 0 : E\left[\left(-\frac{\frac{p(X)}{(1-p(X))}}{E\left(\frac{p(X)}{(1-p(X))}\right)}\right)\left(\Delta Y - \mu_{0,\Delta}(X)\right)\right]$

DR: Sant'Anna and Zhao (2020)

- Some intuition:

$$DD^{DR} = E\left(\Delta Y - \mu_{0,\Delta}(X) \mid D=1\right) - E\left[\left(\frac{\frac{p(X)}{(1-p(X))}}{E\left(\frac{p(X)}{(1-p(X))}\right)}\right)\left(\Delta Y - \mu_{0,\Delta}(X) \mid D=0\right)\right]$$

- If the outcome model is well specified, then $\Delta Y - \mu_{0,\Delta}(X) = 0$ and the treatment effect is driven by the first component.
- If the IPW is well specified, $(\mu_{0,\Delta}(X) \mid D=1 = \mu_{0,\Delta}(X) \mid D=0)$, and the ATT is driven by the IPW component.

DR: Sant'Anna and Zhao (2020)

- Monte Carlo study:
  - Compare DR with TWFE, OR and IPW
  - Sample size is 1,000
  - 10,000 Monte Carlo experiments
  - Propensity score estimated with logit; OR estimated using linear specification

DR: Sant'Anna and Zhao (2020)

- Monte Carlo Simulations, DGP1: Both OR and Propensity score correct

	Bias		RMESE
TWFE	-20.9518		21.1227
OR	-0.0012		0.1005
IPW	0.0257		2.7743
DR	-0.0014		0.1059

- First, the TWFE estimator is severely biased.
  - Not surprising, because,it rules out covariate-specific trends, and these are relevant in this DGP.
  - All semiparametric estimators show little to no Monte Carlo bias, but IPW estimator seem to be less efficient.

DR: Sant'Anna and Zhao (2020)

- Monte Carlo Simulations, DGP4, Neither OR and Propensity score correct

	Bias		RMESE
TWFE	-16.3846		16.5383
OR	-5.2045		5.3641
IPW	-1.0846		2.6557
DR	-3.1878		3.4544

- When both OR and PS are misspecified, all estimators have important biases.
- The IPW estimator has smaller biases and RMSE than the OR and the DR estimators.

DR: Sant'Anna and Zhao (2020)

- Concluding remarks
  - Including covariates in a DD design is done for reasons that are different than in regressions more generally – we are trying to address a parallel trends violation.
  - TWFE can only incorporate time varying covariates, and that places restrictions on the model, whereas other methods will not.
  - Doubly robust and IPW incorporate covariates through propensity scores and outcome regressions (or both)