Causal Inference using Difference-in-Differences

Lecture 6: Leveraging Advances in Machine Learning

Pedro H. C. Sant'Anna

Emory University

January 2025

Introduction

DiD procedures with Covariates

- We can include covariates into DiD to allow for covariate-specific trends.
- There are several “correct” ways of implementing conditional DiD:
  - Regression adjustments;
  - Inverse probability weighting;
  - Doubly Robust (augmented inverse probability weighting);
- DR DiD is my preferred method:
  - More robust against model misspecifications.
  - Can be semiparametrically efficient (confidence intervals are tighter).
- All these are implemented in **DRDID** and **did** R packages, and **drdid** and **csdid** Stata packages.

Implementations, so far, only allow for parametric first-step models.

What if I want to leverage Machine Learning procedures do to DiD?

We will focus on the 2x2 case with Panel Data.

Let’s review our assumptions

Assumptions in 2x2 setup

Assumption (Conditional Parallel Trends Assumption)

$$\mathbb{E}\left[Y_{t=2}(\infty)|G=2,X\right]-\mathbb{E}\left[Y_{t=1}(\infty)|G=2,X\right]=\mathbb{E}\left[Y_{t=2}(\infty)|G=\infty,X\right]-\mathbb{E}\left[Y_{t=1}(\infty)|G=\infty,X\right]\quad a.s.$$

Assumption (No-Anticipation)

For all units i, $Y_{i,t}(g) = Y_{i,t}(\infty)$ for all groups in their pre-treatment periods, i.e., for all $t < g$.

Assumption (Strong Overlap Assumption)

The conditional probability of belonging to the treatment group, given observed characteristics X, is uniformly bounded away from 1. That is, for some $\epsilon > 0$, $\mathbb{P}[G=2|X] < 1-\epsilon$ almost surely.

Different ATT formulations

Regression adjustment procedure

- Originally proposed by Heckman, Ichimura and Todd (1997) and Heckman, Ichimura, Smith and Todd (1998):

$$ATT = \mathbb{E}\left[Y_{t=2} - Y_{t=1} | G = 2\right] - \mathbb{E}\left[m_{\Delta}^{G=\infty}\left(X\right) | G = 2\right]$$

  where

$$m_{\Delta}^{G=\infty}\left(X\right) \equiv \mathbb{E}\left[Y_{t=2} - Y_{t=1} | G = \infty, X\right]$$

- Now, it is “only” a matter of modelling $m_{\Delta}^{G=\infty}\left(X\right)$ and applying the plug-in principle.

- What types of estimation methods can I use to estimate $m_{\Delta}^{G=\infty}\left(X\right)$? Parametric? Nonparametric? Semiparametric? Data-adaptive/Machine Learning?

Inverse probability weighting procedure

- Sant'Anna and Zhao (2020), building on Abadie (2005), considered the following IPW estimand when Panel data are available:

$$ATT_{std}^{ipw,p} = \mathbb{E}\left[\left(\frac{D}{\mathbb{E}[D]} - \frac{\frac{p(X)(1-D)}{1-p(X)}}{\mathbb{E}\left[\frac{p(X)(1-D)}{1-p(X)}\right]}\right)(Y_{t=2} - Y_{t=1})\right],$$

where

$$p(X) \equiv \mathbb{P}[G = 2|X]$$

- Now, it is "only" a matter of modelling $p(X)$ and applying the plug-in principle.
- What types of estimation methods can I use to estimate $p(X)$?
Parametric? Nonparametric? Semiparametric? Data-adaptive/Machine Learning?

Doubly Robust DiD procedure with Panel

- Sant'Anna and Zhao (2020) considered the following doubly robust estimand when panel data are available:

$$ATT^{dr,p} \quad = \quad \mathbb{E}\left[\left(\frac{D}{\mathbb{E}\left[D\right]} - \frac{\dfrac{p(X)\,(1-D)}{1-p(X)}}{\mathbb{E}\left[\dfrac{p(X)\,(1-D)}{1-p(X)}\right]}\right)\left(\left(Y_{t=2}-Y_{t=1}\right) - \left(m_{t=2}^{G=\infty}\left(X\right) - m_{t=1}^{G=\infty}\left(X\right)\right)\right.\right.$$

- Again, it is "only" a matter of modeling $p\left(X\right)$ and $m_{\Delta}^{G=\infty}\left(X\right)$ and applying the plug-in principle.

- What estimation methods can I use to estimate these nuisance models? Parametric? Nonparametric? Semiparametric? Data-adaptive/Machine Learning?

What if I want to use ML?

Being inspired by the recent developments in Causal ML

- In the last 15 years or so, we have seen many advances in Causal Machine Learning.
  - Belloni, Chernozhukov and Hansen (2014)
  - Farrell (2015)
  - Belloni, Chernozhukov, Fernández-Val and Hansen (2017),
  - Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins (2017)
  - Athey and Wager (2018)
  - Athey, Tibshirani and Wager (2019)
  - Chernozhukov, Demirer, Duflo and Fernández-Val (2022).
- All these papers propose estimators that are Doubly Robust/Neyman Orthogonal.
- These ideas have been explored in DiD setups only recently;
  see, e.g., Sant'Anna and Zhao (2020); Chang (2020); Callaway, Drukker, Liu and Sant'Anna (2023).
- Let's touch on some of the basics—only the basics!

Leveraging Machine Learning

What are the practical appeal and challenges?

- Nowadays, we are witnessing a boom in data availability.
- We should be happy about this since more data is more information.
- Maybe it makes conditional PT more plausible!
- OTOH, richer set of covariates can make the **estimation and inference** about the ATT much more challenging.
    - What if we have $n = 200$ but we have 300 different X's?
    - What if we do not know the functional form of the pscore and the outcome-regression?
    - More generally, what variables conditioning variables X should I include in my models?
    - Should we include X, or $1/X$, or $\exp(X)$ or $\log(X)$ or $X^{1/2}$, X^2, $\ldots$

Treatment effects in Data-Rich environments

- This is where machine learning techniques can help us!
- We want to estimate and make inferences about the ATT, allowing for the number of potential covariates, $k := \dim f(X)$, to be potentially larger than the number of cross-sectional units in the data, n.
- Of course, informative inference about **any** causal parameters cannot proceed allowing $k \gg n$ without further restrictions.
- Different machine learning procedures impose different restrictions.
- Here, we will follow the popular approach (at least in economics) of assuming that our nuisance functions, $p(X)$ and $m_{\Delta}^{G=\infty}(X)$, are approximately sparse.
  (This is not required in low dimensional settings; we can also make alternative assumptions).

Approximate Sparsity

- Approximate sparsity imposes that these nuisance functions can be represented up to a “small” approximation error as linear combinations of a number $s \ll n$ of variables $f(X)$, whose identity is a priori unknown to the researcher.
  - This is the case under which we don’t know how X should enter our models $(X^2, \log(X), \exp(X), \sin(X) \ldots)$ but we impose that only a “small” number of these transformations of X matter, though we do not know a priori which one.
- The approximate sparse approach imposes that we are unsure about what to do, so we must conduct some model selection.
- **Key challenge:** how to do valid inference following model selection is nontrivial.
- ML procedures were not originally built to be reliable for inference but to have good predictive properties.

Valid inference after model selection

- We should ignore the model selection step unless we are willing to assume additional structure to the model that imposes that **perfect** model selection is possible.
- Example allowing perfect model selection: “beta-min” condition
  - Requires that all but a small number of coefficients are **exactly** zero. The nonzero coefficients are large enough to be distinguished from zero with probability near 1 in finite samples.
- Such structure can be restrictive and seems unlikely to be satisfied in many applications.
- Rules out the possibility that some variables have moderate but nonzero effects.

Valid inference after model selection

- There are plenty of ML procedures one can use, including:
  1. LASSO
  2. Ridge-Regression
  3. Random Forest and Random Trees
  4. Boosting
  5. Support vector machine(SVM)
  6. Neural nets
- We will focus on LASSO because they are known to perform very well under (approximate) sparsity constraints; see, e.g., Chernozhukov et al. (2017) and Chang (2020) for additional discussions on other methods.
- With LASSO, the implementation is very easy and requires little modifications of available software (which is another reason why we are focusing on it)

Using LASSO regressions

LASSO

- A very popular data-adaptive procedure to estimate the nuisance parameters is the LASSO.
- LASSO stands for least absolute shrinkage and selection operator.
- Its a method that performs both **variable selection** and **regularization**.
  - Enhance prediction accuracy and interpretability of the resulting statistical model (Tibshirani, 1996).
- It has been successfully used in many causal inference procedures, see, e.g., Belloni et al. (2014), Farrell (2015), Chernozhukov et al. (2017), Belloni et al. (2017), among many others.
- More recently, Chang (2020) have built on it for DiD analysis, too!

But what do I need to do LASSO, in practice?

LASSO in practice

- First step, select a "dictionary" of transformations of your covariates X, $f(X)$.
- Now, generically speaking, LASSO becomes a penalized OLS regression (when you think OLS is appropriate):

$$\min_{b} \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\left(Y_i - f\left(X_i\right)' b\right)^2}{2} + \frac{\lambda}{n} \left\| \hat{\Psi} b \right\|_1 \right),$$

where, for a generic Z, $\|Z\|_p = \left(\sum_{l=1}^{n} |Z_l|^p\right)^{1/p}$ is the standard l_p-norm and $\hat{\Psi} = diag\left(\hat{l}_1, \ldots, \hat{l}_k\right)$ is a diagonal matrix of data-dependent penalty loading's.

Using LASSO to estimate $m_{\Delta}^{G=\infty}(X)$

- Construct a “dictionary” of transformations of your covariates X, $f(X)$.

- Next, we can fit penalized OLS regression using only untreated units:

$$\min_{b}\left(\frac{1}{n}\sum_{i:G_i=\infty}\frac{\left(\Delta Y_i - f\left(X_i\right)' b\right)^2}{2} + \frac{\lambda}{n}\left\|\hat{\Psi} b\right\|_1\right),$$

- Once we have our $\widehat{\beta}$'s, we can then estimate $m_{\Delta}^{G=\infty}(x)$ by $\widehat{\mu}_{\Delta}^{G=\infty}(x) = f(x)'\widehat{\beta}$.

Using LASSO to estimate $p(X)$

- OLS is not appropriate to estimate binary outcomes, as in the case with the propensity score.

- But we can easily modify the criterion function and fit a penalized maximum likelihood regression:

$$\min_b \left\{ \frac{1}{n} \sum_{i=1}^{n} - \left[1\left\{ D_i = 1 \right\} \log \Lambda \left(f\left(X_i \right)' b \right) + \right. \right.$$

$$\left. \left. +1\left\{ D_i = 0 \right\} \log \left(1 - \Lambda \left(f\left(X_i \right)' b \right) \right) \right] + \frac{\lambda}{n} \left\| \hat{\Psi} b \right\|_1 \right\},$$

  where, in our context, $D = 1\{G = 2\}$, and $\Lambda(\cdot)$ is a link function–in our case, a logistic function, $\Lambda(\cdot) = exp(\cdot)/(1 + exp(\cdot))$.

- Once we have our $\widehat{\beta}_{ps}$'s, we can then estimate $p(x)$ by $\widehat{\pi}(x) = \Lambda(f(x)'\widehat{\beta}_{ps})$.

Using LASSO regressions

How do we pick the penalty parameters?

Picking penalty parameters

- In the previous slides, you saw that using LASSO involves choosing tuning parameters λ and $\hat{\Psi} = diag\left(\hat{l}_1, \ldots, \hat{l}_k\right)$
- If λ is "too large" : we select "few" regressors
- If λ is "too small" : we select "too many" (perhaps noisy) regressors
- How should you choose the penalty λ and the loadings $\hat{l}_j, j = 1 \ldots, k$?
- They are selected to guarantee good theoretical properties of the method.
- But how?
  - Theory-driven way of picking these: Belloni et al. (2017)
  - More computationally expensive (but with good performances, too): cross-validation Chetverikov, Liao and Chernozhukov (2021)

“Problem” of LASSO

- Estimated LASSO parameters $\hat{\beta}_n^{LASSO}$ for β tend to be downward biased
- This is induced by the shrinkage (penalization)
- To avoid this problem, one can use Post-LASSO, which is a two-step procedure:
  1. Use LASSO as a model selection: that is, run LASSO and select all the variables such that $\hat{\beta}_{j,n}^{LASSO} \neq 0, j = 1, \ldots, k$.
  2. Run OLS (or Maximum likelihood) using only the selected variables.
- For references, see Belloni and Chernozhukov (2013) and Belloni, Chernozhukov and Wei (2016).
- You can include the union of selected covariates when using doubly robust procedures; see, e.g., Belloni et al. (2014).

Let’s see how these work in a DiD simulation exercise

Monte Carlo Simulations

Simulations

- Use LASSO to estimate all functions, using cross-validation to select penalty terms.
- Compare DR DiD estimators with standardized IPW, outcome regression, and unconditional DiD estimators.
- Samples sizes $n = 500$. 500 Monte Carlo repetitions.
- Available data are $\{Y_{t=2}, Y_{t=1}, D, X\}_{i=1}^{n}$, where $D_i = 1\{G_i = 2\}$.
- We estimate the pscore assuming a logit specification and the outcome regression models assuming a linear specification.
- We enter all X linearly (linear dictionary).
- Select “relevant” covariates using LASSO, them run equivalent “post-LASSO” procedure.

DGPs

- Let $X \sim N(0, \Sigma)$ be a $p = 300$ dimensional vector of covariates, with $\Sigma_{j,k} = 0.5^{|j-k|}$.
- Let $\gamma_0^{reg} = (\gamma_{0,1}^{reg}, \ldots, \gamma_{0,p}^{reg})'$, where $\gamma_{0,j}^{reg} = 0.1 \times 1\{j \leq 10\} + \frac{1}{j^2}$.
- Likewise $\gamma_0^{ps} = (\gamma_{0,1}^{ps}, \ldots, \gamma_{0,p}^{ps})'$, where $\gamma_{0,j}^{ps} = \frac{11-j}{10} \times 1\{j \leq 10\} - \frac{1}{j^2}$.
- In our DGPs, we do not have “exact” sparsity!

DGPs

$$f_{ps}\left(X\right) = X'\gamma_0^{ps}$$

$$f_{v}\left(X\right) = X'(1 + \gamma_0^{reg})$$

$$f_{trend}\left(X\right) = 5 \times exp(p(X)) + 5 \times X_4 + 10 * X_{10}$$

$$v\left(X, D\right) \overset{d}{\sim} N\left(D \cdot f_v\left(X\right), 1\right)$$

$$\varepsilon_{t=1}(\infty) \overset{d}{\sim} N\left(0, 1\right)$$

$$\varepsilon_{t=2}\left(2\right) \overset{d}{\sim} N\left(0, 1\right)$$

$$\varepsilon_{t=2}\left(\infty\right) \overset{d}{\sim} N\left(0, 1\right)$$

$$U \overset{d}{\sim} U\left(0, 1\right)$$

3 DGPs, varying the level of heterogeneity

DGP1 - Unconditional PT is valid

- *DGP*1:

$$\begin{aligned}
Y_{i,t=1}(\infty) &= f_v(X_i) + v_i(X_i, D_i) + \varepsilon_{i,t=1}(\infty) \\
Y_{i,t=2}(\infty) &= 1 + f_v(X_i) + v_i(X_i, D_i) + \varepsilon_{i,t=2}(\infty) \\
Y_{i,t=2}(2) &= 1 + f_v(X_i) + v_i(X_i, D_i) + \varepsilon_{i,t=2}(\infty) \\
p(X_i) &= \frac{\exp(0.5 \cdot f_{ps}(X_i))}{1 + \exp(0.5 \cdot f_{ps}(X_i))} \\
D_i &= 1\{p(X_i) \geq U\}
\end{aligned}$$

- ATT is constant across values of X, $ATT(X) = 0$ a.s.
- PT holds unconditionally on X–average trend equal to 1.
- Approx. sparsity is only there for the growth, not for the levels–the term $f_v(X)$ is not approximately sparse.

DGP2 - Conditional PT holds with $ATT(X) = 0$

- *DGP2:*

$$\begin{aligned}
Y_{i,t=1}(\infty) &= f_v(X_i) + v_i(X_i, D_i) + \varepsilon_{i,t=1}(\infty) \\
Y_{i,t=2}(\infty) &= f_v(X_i) + f_{trend}(X_i) + v_i(X_i, D_i) + \varepsilon_{i,t=2}(\infty) \\
Y_{i,t=2}(2) &= f_v(X_i) + f_{trend}(X_i) + v_i(X_i, D_i) + \varepsilon_{i,t=2}(\infty) \\
p(X_i) &= \frac{\exp(0.5 \cdot f_{ps}(X_i))}{1 + \exp(0.5 \cdot f_{ps}(X_i))} \\
D_i &= 1\{p(X_i) \geq U\}
\end{aligned}$$

- ATT is constant across values of X, $ATT(X) = 0$ a.s.
- PT holds conditionally on X but not unconditionally
- Approx. sparsity is only there for the growth, not for the levels–the term $f_v(X)$ is not approximately sparse.

DGP3 - Conditional PT holds with varying $ATT(X)$

- *DGP3*:

$$
\begin{aligned}
Y_{i,t=1}(\infty) &= f_v(X_i) + v_i(X_i, D_i) + \varepsilon_{i,t=1}(\infty) \\
Y_{i,t=2}(\infty) &= f_v(X_i) + f_{trend}(X_i) + v_i(X_i, D_i) + \varepsilon_{i,t=2}(\infty) \\
Y_{i,t=2}(2) &= 1.05 \times f_v(X_i) + f_{trend}(X_i) + v_i(X_i, D_i) + \varepsilon_{i,t=2}(\infty) \\
p(X_i) &= \frac{\exp(0.5 \cdot f_{ps}(X_i))}{1 + \exp(0.5 \cdot f_{ps}(X_i))} \\
D_i &= 1\{p(X_i) \geq U\}
\end{aligned}
$$

- ATT is varying across values of X, $ATT(X) = \mathbb{E}[f_v(X)|D = 1] = 0.13$.
- PT holds conditionally on X but not unconditionally
- ATT(X) is dense in X.

Table 1: Monte Carlo Simulations, DGP1: Unconditional PT

	Bias	RMSE	MC Std. Dev.	Coverage	CI length
$\widehat{\tau}^{unf}$	0.0026	0.0884	0.0884	NA	NA
$\widehat{\tau}^{unc}$	-0.0072	0.0884	0.1301	0.9460	0.4949
$\widehat{\tau}^{reg}$	-0.0070	0.1304	0.1302	0.9440	0.4950
$\widehat{\tau}^{ipw,p}_{std}$	-0.0106	0.1887	0.1884	0.9480	0.6790
$\widehat{\tau}^{dr}$	-0.0065	0.1896	0.1894	0.9400	0.6765

Figure 1: Monte Carlo for DID estimators, DGP1: Unconditional PT

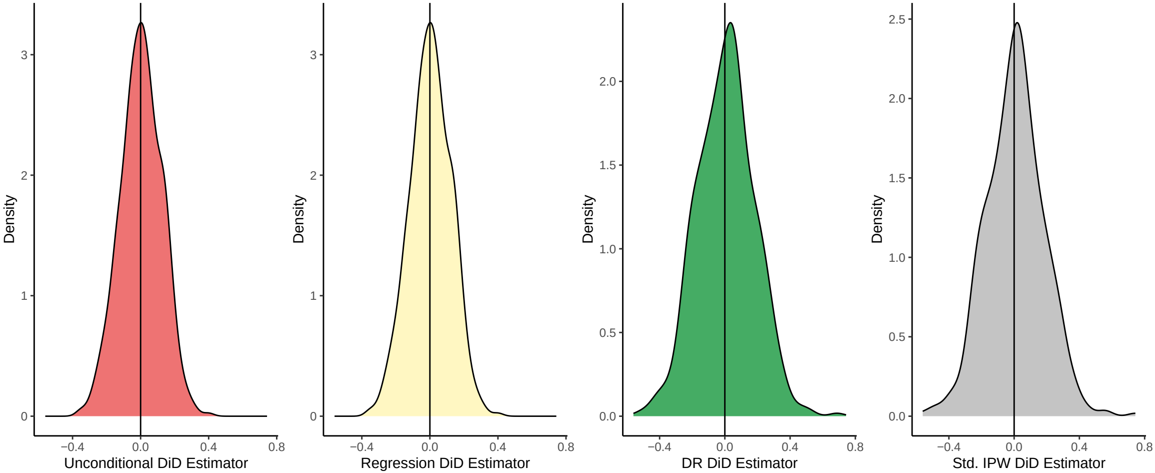


Table 2: Monte Carlo Simulations, DGP2: Conditional PT but homogeneous ATT across X

	Bias	RMSE	MC Std. Dev	Coverage	CI length
$\widehat{\tau}^{unf}$	-0.0039	0.0940	0.0940	NA	NA
$\widehat{\tau}^{unc}$	6.4718	6.4364	1.4666	0.0040	5.6945
$\widehat{\tau}^{reg}$	0.1875	0.2516	0.1677	0.768	0.6403
$\widehat{\tau}^{ipw,p}_{std}$	1.0821	2.2643	1.9890	0.8540	6.8466
$\widehat{\tau}^{dr}$	0.0253	0.1929	0.1913	0.9280	0.6790

Figure 2: Monte Carlo for DID estimators, DGP2: Conditional PT but homogeneous ATT across X

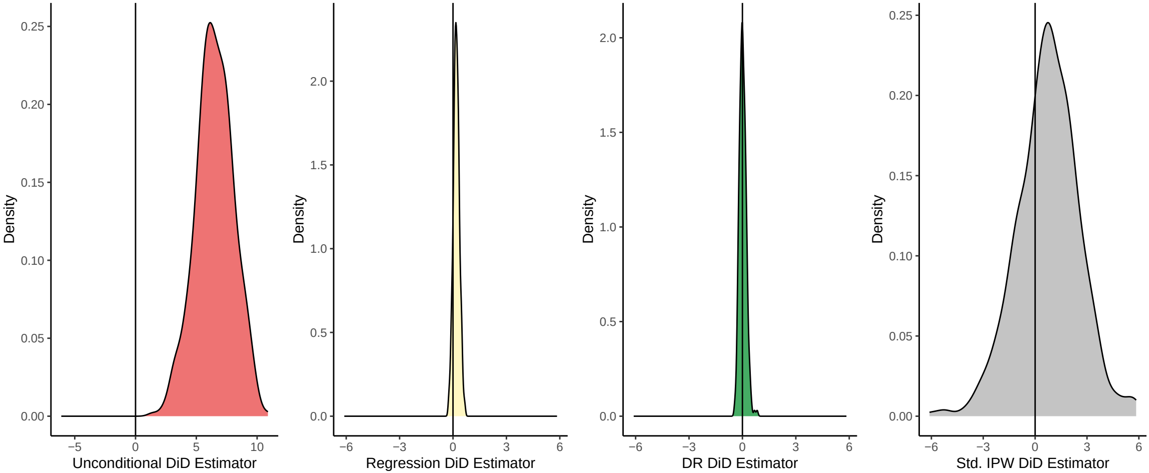


Figure 3: Monte Carlo for DID estimators, DGP2: Conditional PT but homogeneous ATT across X

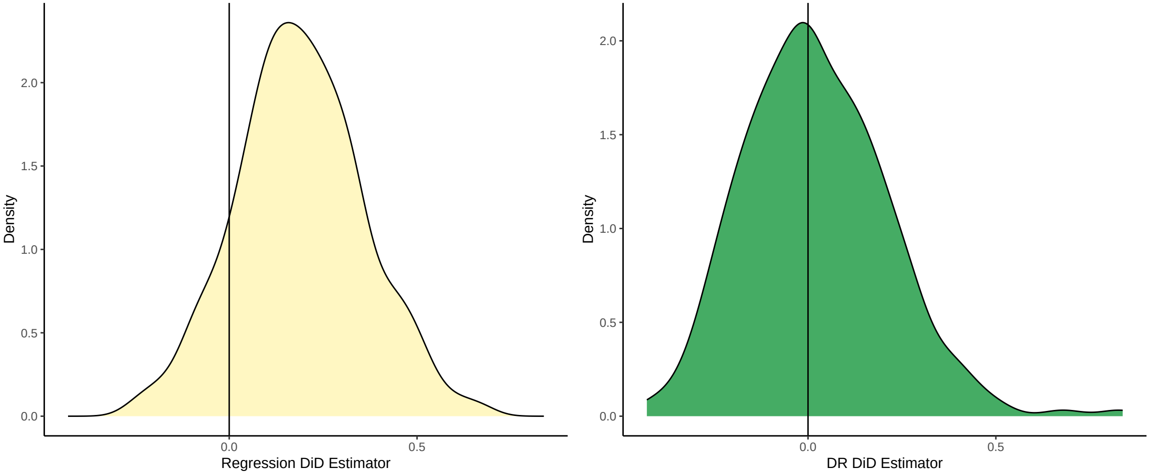


Table 3: Monte Carlo Simulations, DGP3: Conditional PT and heterogeneous ATT across X

	Bias	RMSE	MC Std. Dev,	Coverage	CI length
$\widehat{\tau}^{unf}$	-0.0062	0.1292	0.1290	NA	NA
$\widehat{\tau}^{unc}$	6.5815	6.7297	1.4045	0.0020	5.7457
$\widehat{\tau}^{reg}$	0.1959	0.2811	0.2015	0.8020	0.7481
$\widehat{\tau}^{ipw,p}_{std}$	1.3172	2.4383	2.0519	0.8180	6.8487
$\widehat{\tau}^{dr}$	0.0212	0.2192	0.2182	0.9260	0.7806

Figure 4: Monte Carlo for DID estimators, DGP3: Conditional PT and heterogeneous ATT across X

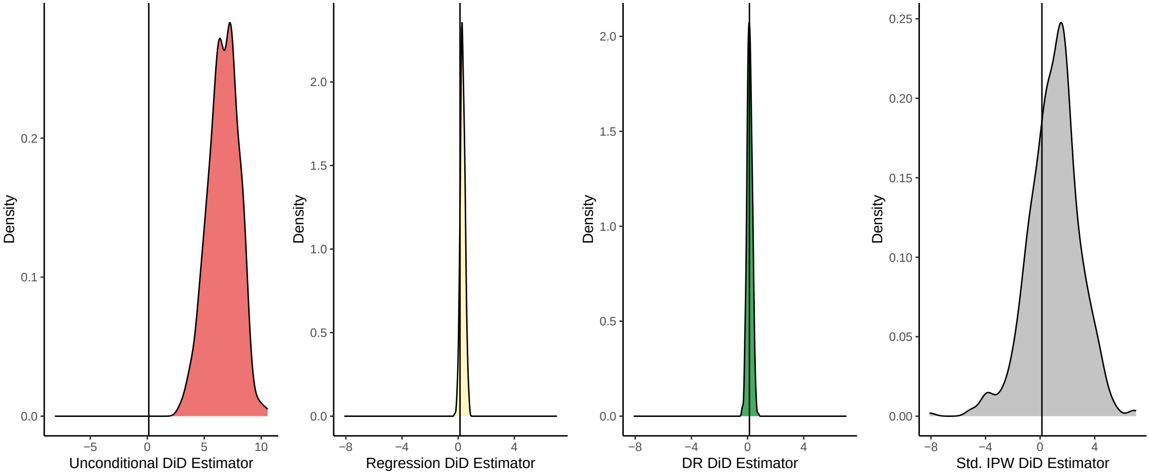
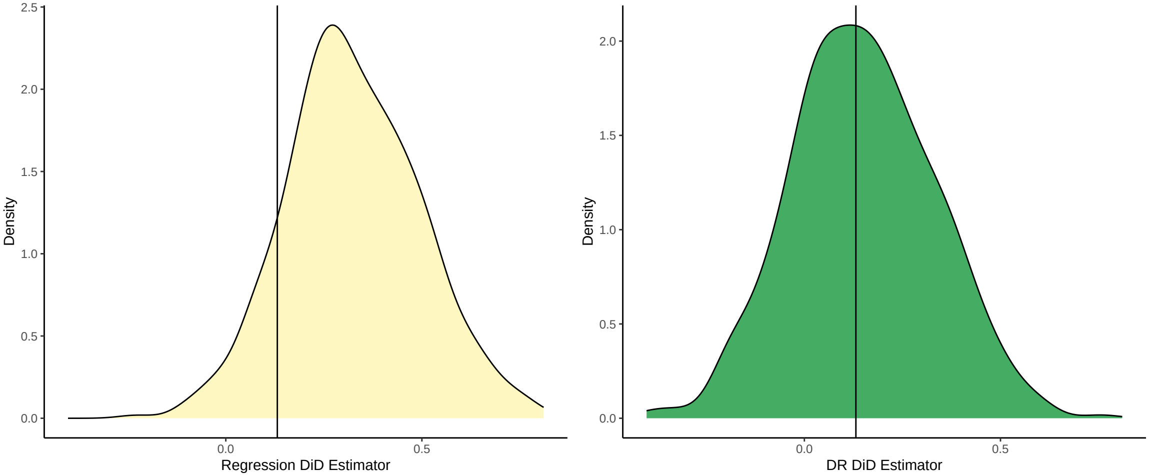


Figure 5: Monte Carlo for DID estimators, DGP3: Conditional PT and heterogeneous ATT across X



What are the requirements?

What are the requirements to use ML in the first step?

- We need to use "orthogonal" moment equations that are first-order (locally) insensitive to changes in the values of the nuisance parameters $m_\Delta^{G=\infty}(\cdot)$, and $p(\cdot)$ that are estimated using data-adaptive methods.
  - This is usually referred to as the "Neyman Orthogonality condition", which our Doubly-Robust formulation satisfies!
- We need to ensure that the model selection mistakes are "moderately" small for the underlying model.
  - It suffices that the product of errors are relatively small, that is,
    $$||(m_\Delta^{G=\infty}(\cdot) - \widehat{\mu}_\Delta^{G=\infty}(\cdot))(p(\cdot) - \widehat{\pi}(\cdot))||_2 = o\left(n^{-1/4}\right).$$
  - This usually comes from assumptions about the "complexity" of the model. Cross-fitting also helps to ensure this for some classes of models (relax some additional conditions when doing LASSO, too).

Take-way messages

Take-way message

- As long as you use the Doubly-Robust formula for DiD, you can use machine learning to estimate nuisance functions.
- Cross-fitting is unnecessary if you proceed with LASSO and have approximate sparsity.
- In some more sophisticated ML procedures, however, you do!
- See Chang (2020) for some results and discussions.
- Although we haven't covered it in detail here, it is easy to use Random Forests a la Athey and Wager (2018) and Athey et al. (2019) with DiD, too. Some tuning is needed, though.

References

Abadie, Alberto, “Semiparametric Difference-in-Differences Estimators,” *The Review of Economic Studies*, 2005, *72* (1), 1–19.

Athey, Susan and Stefan Wager, “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests,” *Journal of the American Statistical Association*, 2018, *113* (523), 1228 – 1242.

— , Julie Tibshirani, and Stefan Wager, “Generalized random forests,” *The Annals of Statistics*, 2019, *47* (2), 1148 – 1178.

Belloni, Alexandre and Victor Chernozhukov, “Least squares after model selection in high-dimensional sparse models,” *Bernoulli*, 2013, *19* (2), 521–547.

— , — , and Christian Hansen, “Inference on Treatment Effects after Selection among High-Dimensional Controls,” *The Review of Economic Studies*, apr 2014, *81* (2), 608–650.

— , — , and Ying Wei, “Post-Selection Inference for Generalized Linear Models With Many Controls,” *Journal of Business & Economic Statistics*, oct 2016, *34* (4), 606–619.

— , — , **Iván Fernández-Val, and Christian Hansen**, “Program Evaluation and Causal Inference With High-Dimensional Data,” *Econometrica*, 2017, *85* (1), 233–298.

Callaway, Brantly, David Drukker, Di Liu, and Pedro H. C. Sant’Anna, “Difference-in-Differences via Machine Learning,” *Working Paper*, 2023.

Chang, Neng-Chieh, “Double/debiased machine learning for difference-in-differences models,” *The Econometrics Journal*, 2020, *23* (2), 177––191.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins, “Double/debiased machine learning for treatment and structural parameters,” *The Econometrics Journal*, jun 2017, pp. 1–71.

— **, Mert Demirer, Esther Duflo, and Iván Fernández-Val**, “Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments ,” *arXiv:1712.04802*, 2022.

Chetverikov, Denis, Zhipeng Liao, and Victor Chernozhukov, “On cross-validated Lasso in high dimensions,” *The Annals of Statistics*, jun 2021, *49* (3), 1–25.

Farrell, Max H., "Robust inference on average treatment effects with possibly more covariates than observations," *Journal of Econometrics*, 2015, *189* (1), 1–23.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd, "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 1998, *66* (5), 1017–1098.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd, "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *The Review of Economic Studies*, October 1997, *64* (4), 605–654.

Sant'Anna, Pedro H. C. and Jun Zhao, "Doubly robust difference-in-differences estimators," *Journal of Econometrics*, November 2020, *219* (1), 101–122.

Tibshirani, Robert, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996, *58* (1).