

# Better Understanding Triple Differences Estimators

---

**Marcelo Ortiz-Villavicencio**

Emory University

Oct 29, 2024

Field Paper Presentation

# Motivation

---

# Triple Differences

- Triple Differences (DDD) extend to cases where the parallel trends (PT) assumption in DiD *may not hold*.

# Triple Differences

- Triple Differences (DDD) extend to cases where the parallel trends (PT) assumption in DiD *may not hold*.
  - ▶ **Ex:** PT can be violated due to the presence of a *time-varying confounder* that *changes differently across states*.

# Triple Differences

- Triple Differences (DDD) extend to cases where the parallel trends (PT) assumption in DiD *may not hold*.
  - ▶ **Ex:** PT can be violated due to the presence of a *time-varying confounder* that *changes differently across states*.
- When the PT assumption is questionable, researchers often augment the design by adding another *placebo* comparison group to recover the treatment effect of interest.

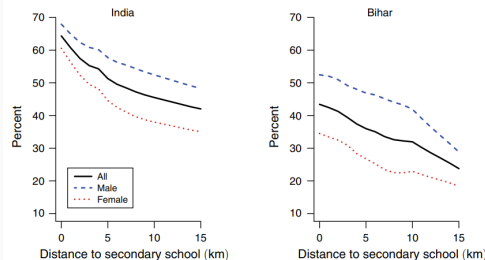
# Triple Differences

- Triple Differences (DDD) extend to cases where the parallel trends (PT) assumption in DiD *may not hold*.
  - ▶ **Ex:** PT can be violated due to the presence of a *time-varying confounder* that *changes differently across states*.
- When the PT assumption is questionable, researchers often augment the design by adding another *placebo* comparison group to recover the treatment effect of interest.
- DDD designs address this issue by finding a *within-state comparison group* that **is not exposed** to the treatment but is **affected** by the time-varying confounder.

# Running Example: Muralidharan and Prakash (2017)

- The researchers examine how giving bicycles to *girls* influences their *enrollment in secondary schools*.
- **Distance effect:** Female enrollment drops significantly with increased distance to school (Panel B).
- **Pre-program gap:** The attrition for female in secondary school enrollment is higher than male enrollment.
- The Cycle program in *Bihar, India* aimed to reduce the gender gap in secondary school enrollment.

Panel B: 16- and 17-year-olds enrolled in or completed grade 9 by distance and gender



## Running Example: Muralidharan and Prakash (2017)

- PT assumption is questionable because the bicycle program coincided with *rapid growth* and *increased education spending* in Bihar, unlike the control state, Jharkhand.
- Changes in girls' secondary school enrollment cannot be attributed to the *Cycle Program* directly.



## Running Example: Muralidharan and Prakash (2017)

- PT assumption is questionable because the bicycle program coincided with *rapid growth* and *increased education spending* in Bihar, unlike the control state, Jharkhand.
- Changes in girls' secondary school enrollment cannot be attributed to the *Cycle Program* directly.
- DDD to the rescue!

## Running Example: Muralidharan and Prakash (2017)

- PT assumption is questionable because the bicycle program coincided with *rapid growth* and *increased education spending* in Bihar, unlike the control state, Jharkhand.
- Changes in girls' secondary school enrollment cannot be attributed to the *Cycle Program* directly.
- DDD to the rescue!
- Boys in Bihar **are not** exposed to the policy but are **affected** by the expansion in education spending.

# What's the appeal of DDD compared to DiD?

■ DDD allow us to take into account:

1. **Location**-specific trends: Those related to the difference in education spending in Bihar vs. Jharkhand.

This would be ruled out if we were to do DiD by dropping observations of **boys** across states.

# What's the appeal of DDD compared to DiD?

- DDD allow us to take into account:

1. **Location**-specific trends: Those related to the difference in education spending in Bihar vs. Jharkhand.

This would be ruled out if we were to do DiD by dropping observations of **boys** across states.

2. **Partition**-specific trends: Those related to the inherent disparities in girls' access to education in India.

This would be ruled out if we were to do DiD by dropping observations of Jharkhand.

## What's the appeal of DDD compared to DiD?

Putting everything together, in DDD we allow to use all the information to control for **location-specific trends** and **partition-specific trends**, which otherwise would arise questionable results using DiD.

	No Eligible	Eligible
Treated	(Bihar, boys)	(Bihar, girls)
Control	(Jharkhand, boys)	(Jharkhand, girls)

- Although it is also widely used in empirical work, DDD hasn't received as much attention as DiD.
- The key question in this paper is: How can we leverage our DiD knowledge to approach DDD?
  - ▶ We study identification, estimation, and inference procedures for DDD designs.
  - ▶ We derive the semiparametric efficiency bound for DDD designs and demonstrate that DDD estimators using a doubly robust representation reach this bound.
  - ▶ We extend our framework to staggered DDD designs.

# Framework

---

## Some notation

We have access to a sample of  $n$  units available,  $i = 1, 2, \dots, n$

- T time periods:  $t = 1, 2, \dots, T$ .
- Different groups adopt a policy in different time periods  $g$ . Let  $G \in \mathcal{G} \subset \{2, \dots, T\} \cup \{\infty\}$  denote the time when group  $g$  is first-adopt the policy, with the notion that if a group is "never-treated",  $G = \infty$ .
- Within each set of groups, we have two partitions (defined by some well-known criterion),  $\ell \in P \equiv \{0, 1\}$ . This determines *eligibility status*.
- Let  $D_i \in \mathcal{D} \subset \{2, \dots, T\} \cup \{\infty\}$  denote the time unit  $i$  is first-treated, with the notion that if a unit is "never-treated",  $D_i = \infty$ .
- Note that  $\mathcal{D} = \mathcal{G}$  such that:

$$D = \begin{cases} d & \text{if } G = g \equiv d \wedge P = 1, \\ \infty & \text{if } (G = g \wedge P = 0) \text{ or } (G = g' \wedge P \in \{0, 1\}, \text{ with } g' > g) \end{cases}$$

**Ex:** Units in  $G = 2$  with  $P = 1$  are treated at time  $D = 2$ , otherwise the unit remains untreated ( $D = \infty$ ).



## Building block of the analysis

- Let  $Y_{i,t}(d)$  be the potential outcome for unit  $i$ , at time  $t$ , if this unit is first treated at time period  $d$ .
- A parameter that is interesting and has clear economic interpretation is the  $ATT(g, t)$  (Callaway and Sant'Anna, 2021).

### Definition (Parameter of interest: $ATT(g,t)$ )

Average Treatment Effect at time  $t$  of starting treatment at time  $g$ , among the units that indeed started treatment at time  $g$ .

$$ATT(g, t) := \mathbb{E} [Y_t(d) - Y_t(\infty) | G = g, P = 1], \text{ for } t \geq g.$$

## Building block of the analysis

- Let  $Y_{i,t}(d)$  be the potential outcome for unit  $i$ , at time  $t$ , if this unit is first treated at time period  $d$ .
- A parameter that is interesting and has clear economic interpretation is the  $ATT(g, t)$  (Callaway and Sant'Anna, 2021).

### Definition (Parameter of interest: $ATT(g,t)$ )

Average Treatment Effect at time  $t$  of starting treatment at time  $g$ , among the units that indeed started treatment at time  $g$ .

$$ATT(g, t) := \mathbb{E} [Y_t(d) - Y_t(\infty) | G = g, P = 1], \text{ for } t \geq g.$$

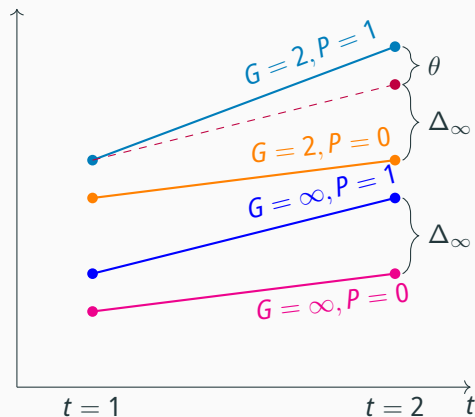
- Then, our identification problem comes from the fact that we never observe  $\mathbb{E} [Y_t(\infty) | G = g, P = 1]$  in  $t \geq g$ .

# Canonical DDD Design

---

## 2x2x2 DDD without covariates

For simplicity, let's start with the canonical 2x2x2 DDD. Thus,  $g \in G = \{2, \infty\}$  and  $\ell \in P = \{0, 1\}$ .



$$\begin{aligned} \theta = & \left[ \left( \mathbb{E}[Y_{t=2}|G=2, P=1] - \mathbb{E}[Y_{t=1}|G=2, P=1] \right) \right. \\ & \left. - \left( \mathbb{E}[Y_{t=2}|G=2, P=0] - \mathbb{E}[Y_{t=1}|G=2, P=0] \right) \right] \\ & - \left[ \left( \mathbb{E}[Y_{t=2}|G=\infty, P=1] - \mathbb{E}[Y_{t=1}|G=\infty, P=1] \right) \right. \\ & \left. - \left( \mathbb{E}[Y_{t=2}|G=\infty, P=0] - \mathbb{E}[Y_{t=1}|G=\infty, P=0] \right) \right] \end{aligned}$$

- Note that this is the difference of two DiD's: one among  $G=2$  across eligible groups, and one among  $G=\infty$  across eligible groups.

# Recovering the ATT using 3WFE Regression

- When there are only 2 time periods and *no covariates*, the following three-way fixed-effects (3WFE) regression specification can be used to recover the ATT:

$$\begin{aligned} Y_{i,t} = & \alpha_0 + \gamma_{0,1} \mathbf{1}_{\{G_i=2\}} + \gamma_{0,2} \mathbf{1}_{\{P_i=1\}} + \gamma_{0,3} \mathbf{1}_{\{T_i=2\}} \\ & + \gamma_{0,4} \mathbf{1}_{\{G_i=2\}} \mathbf{1}_{\{P_i=1\}} + \gamma_{0,5} \mathbf{1}_{\{G_i=2\}} \mathbf{1}_{\{T_i=2\}} + \gamma_{0,6} \mathbf{1}_{\{P_i=1\}} \mathbf{1}_{\{T_i=2\}} \\ & + \beta_0^{3wfe} \mathbf{1}_{\{G_i=2\}} \mathbf{1}_{\{P_i=1\}} \mathbf{1}_{\{T_i=2\}} + \varepsilon_{i,t}, \end{aligned}$$

- We can show that  $\beta_0^{3wfe} = \theta$  (Olden and Møen, 2022).

# Recovering the ATT using 3WFE Regression

- When there are only 2 time periods and *no covariates*, the following three-way fixed-effects (3WFE) regression specification can be used to recover the ATT:

$$\begin{aligned} Y_{i,t} = & \alpha_0 + \gamma_{0,1} \mathbf{1}_{\{G_i=2\}} + \gamma_{0,2} \mathbf{1}_{\{P_i=1\}} + \gamma_{0,3} \mathbf{1}_{\{T_i=2\}} \\ & + \gamma_{0,4} \mathbf{1}_{\{G_i=2\}} \mathbf{1}_{\{P_i=1\}} + \gamma_{0,5} \mathbf{1}_{\{G_i=2\}} \mathbf{1}_{\{T_i=2\}} + \gamma_{0,6} \mathbf{1}_{\{P_i=1\}} \mathbf{1}_{\{T_i=2\}} \\ & + \beta_0^{3wfe} \mathbf{1}_{\{G_i=2\}} \mathbf{1}_{\{P_i=1\}} \mathbf{1}_{\{T_i=2\}} + \varepsilon_{i,t}, \end{aligned}$$

- We can show that  $\beta_0^{3wfe} = \theta$  (Olden and Møen, 2022).

**These results suggest that we can estimate ATT in the canonical DDD either (i) by a difference between two DiDs or (2) by a saturated 3WFE regression.**

**What happens when covariates  
play an important role?**

---

# Assumptions

- Adding covariates in the above 3WFE specification would imply additional restrictions to the DGP:
  - ▶ Homogeneous treatment effects in covariates.
  - ▶ Rule out covariate-specific trends in both the treated and comparison groups.



# Assumptions

- Adding covariates in the above 3WFE specification would imply additional restrictions to the DGP:
  - ▶ Homogeneous treatment effects in covariates.
  - ▶ Rule out covariate-specific trends in both the treated and comparison groups.
- Our goal is to introduce an estimator for DDD under the condition that the PT assumption is valid after controlling for covariates, i.e.,  $X \in \mathcal{X} \subseteq \mathbb{R}^d$ .

# Assumptions

- Adding covariates in the above 3WFE specification would imply additional restrictions to the DGP:
  - ▶ Homogeneous treatment effects in covariates.
  - ▶ Rule out covariate-specific trends in both the treated and comparison groups.
- Our goal is to introduce an estimator for DDD under the condition that the PT assumption is valid after controlling for covariates, i.e.,  $X \in \mathcal{X} \subseteq \mathbb{R}^d$ .

## Assumption (Conditional Parallel Trends Assumption for DDD)

$$\begin{aligned} \mathbb{E}[Y_{t=2}(\infty) - Y_{t=1}(\infty) | G = 2, P = 1, X] &- \mathbb{E}[Y_{t=2}(\infty) - Y_{t=1}(\infty) | G = 2, P = 0, X] \\ &= \\ \mathbb{E}[Y_{t=2}(\infty) - Y_{t=1}(\infty) | G = \infty, P = 1, X] &- \mathbb{E}[Y_{t=2}(\infty) - Y_{t=1}(\infty) | G = \infty, P = 0, X] \text{ a.s.} \end{aligned}$$

# Assumptions

- Adding covariates in the above 3WFE specification would imply additional restrictions to the DGP:
  - ▶ Homogeneous treatment effects in covariates.
  - ▶ Rule out covariate-specific trends in both the treated and comparison groups.
- Our goal is to introduce an estimator for DDD under the condition that the PT assumption is valid after controlling for covariates, i.e.,  $X \in \mathcal{X} \subseteq \mathbb{R}^d$ .

## Assumption (Conditional Parallel Trends Assumption for DDD)

$$\begin{aligned} \mathbb{E}[Y_{t=2}(\infty) - Y_{t=1}(\infty) | G = 2, P = 1, X] &= \mathbb{E}[Y_{t=2}(\infty) - Y_{t=1}(\infty) | G = 2, P = 0, X] \\ &= \mathbb{E}[Y_{t=2}(\infty) - Y_{t=1}(\infty) | G = \infty, P = 1, X] = \mathbb{E}[Y_{t=2}(\infty) - Y_{t=1}(\infty) | G = \infty, P = 0, X] \text{ a.s.} \end{aligned}$$

**Additional Assumptions:** Strong Overlap & No Anticipation.

# Doubly Robust and Semiparametric Efficiency

- Under the previous assumptions, the ATT can be identified via Regression Adjustments or IPW or any convex combination between them. RA & IPW
- However, this depends on the researcher's ability to accurately model outcome regression or propensity scores.
  - ▶ How would you choose this combination if the goal was to achieve **Doubly Robustness** (DR)?
  - ▶ How would you choose this combination if the goal was **efficiency**?
- We tackle these questions by deriving the *semiparametric efficiency bound* for the ATT in DDD setups.
- That usually leads to DR estimands, too.

# Semiparametric Efficiency Bound

## Proposition (Semiparametric Efficiency Bound for DDD)

Suppose that conditional PT, no-anticipation, and strong overlap assumptions are satisfied, and balanced panel data is available. Let  $\theta(\Delta Y, X) := \Delta Y - m_{\Delta}^{G=2, P=0}(X) - m_{\Delta}^{G=\infty, P=1}(X) + m_{\Delta}^{G=\infty, P=0}(X)$ ,  $S := (\Delta Y, G, P, X)$ . Then, the efficient influence function for the ATT is given by

$$\begin{aligned}\eta_{\text{eff}}(S) = & \omega_1^{G=2, P=1} \cdot \left( \theta(\Delta Y, X) - \theta \right) \\ & - \omega_0^{G=2, P=0}(X) \cdot \left( \Delta Y - m_{\Delta}^{G=2, P=0}(X) \right) \\ & - \omega_0^{G=\infty, P=1}(X) \cdot \left( \Delta Y - m_{\Delta}^{G=\infty, P=1}(X) \right) \\ & + \omega_0^{G=\infty, P=0}(X) \cdot \left( \Delta Y - m_{\Delta}^{G=\infty, P=0}(X) \right).\end{aligned}$$

Furthermore, the semiparametric efficiency bound for the set of all regular, and asymptotic linear estimators of the ATT is  $\mathbb{E}[\eta_{\text{eff}}(S)^2]$ .

# DR DDD as a function of 3 DiDs

- We can take the expected value of  $\eta_{\text{eff}}(W)$  and isolate  $\theta$  given that any influence function has mean zero.
- Let's conveniently rewrite the propensity scores  $\forall (g', \ell) \in \mathcal{S}_c$  as

$$p_{g', \ell}(X) = \mathbb{P}[G = 2, P = 1 | X, (G = 2, P = 1) \cup (G = g', P = \ell)],$$

- Finally, we get the following **DR-DDD** estimand for the ATT,

$$\begin{aligned} \theta^{DR} = & \mathbb{E} \left[ \left( \omega_1^{G=2, P=1} - \omega_0^{G=2, P=0} (p_{2,0}(X)) \right) \left( \Delta Y - m_{\Delta}^{G=2, P=0}(X) \right) \right] \Leftarrow DRDiD_{(2,1)}^{(2,0)} \\ & + \mathbb{E} \left[ \left( \omega_1^{G=2, P=1} - \omega_0^{G=\infty, P=1} (p_{\infty,1}(X)) \right) \left( \Delta Y - m_{\Delta}^{G=\infty, P=1}(X) \right) \right] \Leftarrow DRDiD_{(2,1)}^{(\infty,1)} \\ & - \mathbb{E} \left[ \left( \omega_1^{G=2, P=1} - \omega_0^{G=\infty, P=0} (p_{\infty,0}(X)) \right) \left( \Delta Y - m_{\Delta}^{G=\infty, P=0}(X) \right) \right] \Leftarrow DRDiD_{(2,1)}^{(\infty,0)} \end{aligned}$$

Expanded Version

# Monte Carlo Simulations

---

## Simulations for $T = 2$ with covariates

- For simplicity, we consider the scenario for panel data with  $T = 2$  and we have access to generic data  $W = (W_1, W_2, W_3, W_4)'$ .
- WLOG, consider that the *eligibility of treatment* is given by binary well-know criterion  $P = \{0, 1\}$  and let  $(g, \ell) \in \mathcal{S}_c := \{(\infty, 0), (\infty, 1), (2, 0)\}$  and  $\mathcal{S} := \mathcal{S}_c \cup \{(2, 1)\}$ .
- Since we have 4 partitions in the data, we can model the selection into treatment as multinomial logistic link function.
- Outcome process can be modeled as linear regression onto  $W$ .
- We consider 4 DGPs:
  - ▶ **both** models are *correctly* specified;
  - ▶ Only **propensity score** is *correctly* specified;
  - ▶ Only **outcome model** is *correctly* specified;
  - ▶ **both** models are *wrong*.
- We compare our DR DDD estimator with:
  - ▶ 3WFE specification.
  - ▶ Difference between 2 Doubly Robust DiD (Sant'Anna & Zhao, 2020).



# Results

	DGP 1: $\mathbb{E} [\eta_{\text{eff}}(W)^2] = 32.82$			DGP 2: $\mathbb{E} [\eta_{\text{eff}}(W)^2] = 32.52$			DGP 3: $\mathbb{E} [\eta_{\text{eff}}(W)^2] = 32.82$			DGP 4: $\mathbb{E} [\eta_{\text{eff}}(W)^2] = 32.52$		
	$\hat{\theta}_{ddd}$	$\hat{\theta}_{3wfe}$	$\hat{\theta}_{dr}$	$\hat{\theta}_{ddd}$	$\hat{\theta}_{3wfe}$	$\hat{\theta}_{dr}$	$\hat{\theta}_{ddd}$	$\hat{\theta}_{3wfe}$	$\hat{\theta}_{dr}$	$\hat{\theta}_{ddd}$	$\hat{\theta}_{3wfe}$	$\hat{\theta}_{dr}$
$n = 1000$												
Bias	<b>0.0007</b>	-7.3298	-4.0199	<b>-0.0029</b>	-6.3178	-3.4484	<b>0.0255</b>	-3.8366	-1.9826	<b>0.0421</b>	-5.6247	-3.4447
RMSE	<b>0.1823</b>	8.4185	5.0331	<b>0.1799</b>	7.4545	4.5122	<b>1.4384</b>	5.1171	3.3235	<b>1.4498</b>	6.5893	4.3858
$\mathbb{E}[\text{Var}]$	43.5075	47341.2395	.	43.1213	47906.9263	.	2121.5766	45057.4548	.	2163.0339	45102.7697	.
Cov. 95	0.9650	0.9240	.	0.9730	0.9550	.	0.9600	0.9970	.	0.9530	0.9860	.
avg. length	<b>0.8110</b>	26.9471	.	<b>0.8071</b>	27.1033	.	<b>5.7012</b>	26.2999	.	<b>5.7553</b>	26.3134	.
$n = 50000$												
Bias	<b>0.0008</b>	-7.3101	-4.0285	0.0007	-6.2891	-3.4389	-0.0039	-3.9647	-2.1148	0.1039	-5.4834	-3.3654
RMSE	<b>0.0257</b>	7.3348	4.0522	0.0257	6.3154	3.4636	0.2011	3.9929	2.1469	0.2345	5.5038	3.3857
$\mathbb{E}[\text{Var}]$	34.3777	47502.9417	.	33.9857	48240.7801	.	2059.7784	45339.0209	.	2113.8460	45453.0303	.
Cov. 95	0.9550	0.0000	.	0.9470	0.0000	.	0.9540	0.0000	.	0.9130	0.0000	.
avg. length	<b>0.1028</b>	3.8208	.	<b>0.1022</b>	3.8503	.	<b>0.7956</b>	3.7328	.	<b>0.8060</b>	3.7375	.

# Conclusion

---

# Highlights

- DDD is widely used in empirical research, but its properties have receive little attention.
- In its basic format, it is equivalent to running two separate DiD estimators and subtracting one from another.
  - ▶ This equivalence breaks down when covariates play an important role in the analysis.
- We derived semiparametric efficiency bound for DDD and proposed DR DDD estimands.
- We can leverage these results to tackle staggered treatment setups, too.

## Next steps:

- Looking for an empirical application to illustrate the performance of the proposed estimator.
- A very fast package implementation in R is under construction.

# Thanks!

✉ marcelo.ortiz@emory.edu

🔗 marcelortiz.com

🐦 @marcelortizv

## Some additional notation

- For  $(g, \ell) \in \{2, \infty\} \times \{0, 1\}$ , let  $\Delta Y = Y_{t=2} - Y_{t=1}$ , and

$$m_{\Delta}^{G=g, P=\ell}(X) := \mathbb{E}[\Delta Y | G = g, P = \ell, X], \quad (\text{outcome regression}).$$

$$p^{G=g, P=\ell}(X) := \mathbb{P}[G = g, P = \ell | X] \quad (\text{multi-valued propensity score}).$$

- For  $(g, \ell) \in \mathcal{S}_c \equiv \{(\infty, 0), (\infty, 1), (2, 0)\}$ , let

$$\omega_1^{G=2, P=1} := \frac{1_{\{G=2, P=1\}}}{\mathbb{E}[1_{\{G=2, P=1\}}]},$$

$$\omega_0^{G=g, P=\ell}(X) := \frac{\frac{1_{\{G=g, P=\ell\}} \cdot p^{G=2, P=1}(X)}{p^{G=g, P=\ell}(X)}}{\mathbb{E}\left[\frac{1_{\{G=g, P=\ell\}} \cdot p^{G=2, P=1}(X)}{p^{G=g, P=\ell}(X)}\right]}$$

- Let  $\theta(\Delta Y, X) := \Delta Y - m_{\Delta}^{G=2, P=0}(X) - m_{\Delta}^{G=\infty, P=1}(X) + m_{\Delta}^{G=\infty, P=0}(X)$ ,  $S := (\Delta Y, G, P, X)$ .

# Regression Adjustment and IPW identification

- We can show that if conditional PT, no-anticipation, and strong overlap assumptions are satisfied and balanced panel data is available, the ATT is identified via regression adjustments or IPW:

$$\theta = ATT^{RA} = ATT^{IPW},$$

where

$$ATT^{RA} := \mathbb{E} [\Delta Y | G = 2, P = 1] - \mathbb{E} \left[ m_{\Delta}^{G=2, P=0}(X) + \left( m_{\Delta}^{G=\infty, P=1}(X) - m_{\Delta}^{G=\infty, P=0}(X) \right) \middle| G = 2, P = 1 \right],$$

$$ATT^{IPW} := \mathbb{E} \left[ \left( \left( w^{G=2, P=1}(G, P) - w^{G=2, P=0}(G, P, X) \right) - \left( w^{G=\infty, P=1}(G, P, X) - w^{G=\infty, P=0}(G, P, X) \right) \right) \Delta Y \right].$$

Go Back

# DR DDD as a function of 3 DiDs

- To get a *DR-DDD* estimand for the ATT, isolate  $\theta$  given that any influence function has mean zero. [Go Back](#)
- We can conveniently rewrite the propensity scores  $\forall (g', \ell) \in \mathcal{S}_c$  as

$$p_{g', \ell}(X) = \mathbb{P}[G = 2, P = 1 | X, (G = 2, P = 1) \cup (G = g', P = \ell)],$$

$$\begin{aligned} \Rightarrow \theta^{DR} &= \mathbb{E} \left[ \left( \frac{1_{\{G=2, P=1\}}}{\mathbb{E}[1_{\{G=2, P=1\}}]} - \frac{\frac{p_{2,0}(X) \cdot 1_{\{G=2, P=0\}}}{1 - p_{2,0}(X)}}{\mathbb{E}\left[\frac{p_{2,0}(X) \cdot 1_{\{G=2, P=0\}}}{1 - p_{2,0}(X)}\right]} \right) (\Delta Y - m_{\Delta}^{G=2, P=0}(X)) \right] \\ &+ \mathbb{E} \left[ \left( \frac{1_{\{G=2, P=1\}}}{\mathbb{E}[1_{\{G=2, P=1\}}]} - \frac{\frac{p_{\infty,1}(X) \cdot 1_{\{G=\infty, P=1\}}}{1 - p_{\infty,1}(X)}}{\mathbb{E}\left[\frac{p_{\infty,1}(X) \cdot 1_{\{G=\infty, P=1\}}}{1 - p_{\infty,1}(X)}\right]} \right) (\Delta Y - m_{\Delta}^{G=\infty, P=1}(X)) \right] \\ &- \mathbb{E} \left[ \left( \frac{1_{\{G=2, P=1\}}}{\mathbb{E}[1_{\{G=2, P=1\}}]} - \frac{\frac{p_{\infty,0}(X) \cdot 1_{\{G=\infty, P=0\}}}{1 - p_{\infty,0}(X)}}{\mathbb{E}\left[\frac{p_{\infty,0}(X) \cdot 1_{\{G=\infty, P=0\}}}{1 - p_{\infty,0}(X)}\right]} \right) (\Delta Y - m_{\Delta}^{G=\infty, P=0}(X)) \right] \end{aligned}$$

- Since we have 4 partitions in the data, we consider the following PS using a multinomial logistic link function:

$$p^{G=g,P=\ell}(W) = \begin{cases} \frac{\exp(f_{ps}^{g,\ell}(W))}{1 + \sum_{(g,\ell) \in \mathcal{S}_c} \exp(f_{ps}^{g,\ell}(W))}, & \text{if } (g,\ell) \in \mathcal{S}_c \\ \frac{1}{1 + \sum_{(g,\ell) \in \mathcal{S}_c} \exp(f_{ps}^{g,\ell}(W))}, & \text{if } (g,\ell) = (2,1). \end{cases}$$

where,  $f_{ps}^{g,\ell}(W) = \alpha_1^{g,\ell} W_1 + \alpha_2^{g,\ell} W_2 + \alpha_3^{g,\ell} W_3 + \alpha_4^{g,\ell} W_4$



- Let  $U \sim \text{Uniform}[0, 1]$ . The partition groups are assigned as follows

$$(g, \ell) := \begin{cases} (\infty, 0), & \text{if } U \leq p^{G=\infty, P=0}(W), \\ (\infty, 1), & \text{if } p^{G=\infty, P=0}(W) < U \leq p^{G=\infty, P=0}(W) + p^{G=\infty, P=1}(W), \\ (2, 0), & \text{if } p^{G=\infty, P=0}(W) + p^{G=\infty, P=1}(W) < U \leq 1 - p^{G=2, P=1}(W), \\ (2, 1), & \text{if } 1 - p^{G=2, P=1}(W) < U. \end{cases}$$

- For the Outcome Regression process, define

$$f_{reg, G=2}^{g, \ell}(W) = \beta_{11}^{g, \ell} W_1 + \beta_{21}^{g, \ell} W_2 + \beta_{31}^{g, \ell} W_3 + \beta_{41}^{g, \ell} W_4, \forall (g, \ell) \in \{(2, \ell)\}$$

$$f_{reg, G=\infty}^{g, \ell}(W) = \beta_{10}^{g, \ell} W_1 + \beta_{20}^{g, \ell} W_2 + \beta_{30}^{g, \ell} W_3 + \beta_{40}^{g, \ell} W_4, \forall (g, \ell) \in \{(\infty, \ell)\}$$

- Let *time-invariant unobserved heterogeneity* be defined as

$$\nu(W, G, P) \sim N\left(f_{het}^{g, \ell}(W), 1\right), \forall (g, \ell) \in \mathcal{S} \text{ where,}$$

$$f_{het}^{g, \ell}(W) = 1_{\{G=2\}} \cdot 1_{\{P=1\}} \cdot f_{reg, G=2}^{g, \ell}(W) + (1 - 1_{\{G=2\}}) \cdot 1_{\{P=1\}} \cdot f_{reg, G=\infty}^{g, \ell}(W)$$