Causal Inference Crash Course
Part 6:
Panel Models

Julian Hsu

Causal Inference Series

1) Foundations
2) Defining Some ATE/ATET Causal Models
3) ATE/ATET Inference, Asymptotic Theory, and Bootstrapping
4) Best Practices: Outliers, Class Imbalance, Feature Selection, and Bad Control
5) Heterogeneous Treatment Effect Models and Inference
6) **Difference-in-Difference Models for Panel Data**
7) Regression Discontinuity Models
8) Arguable Validation

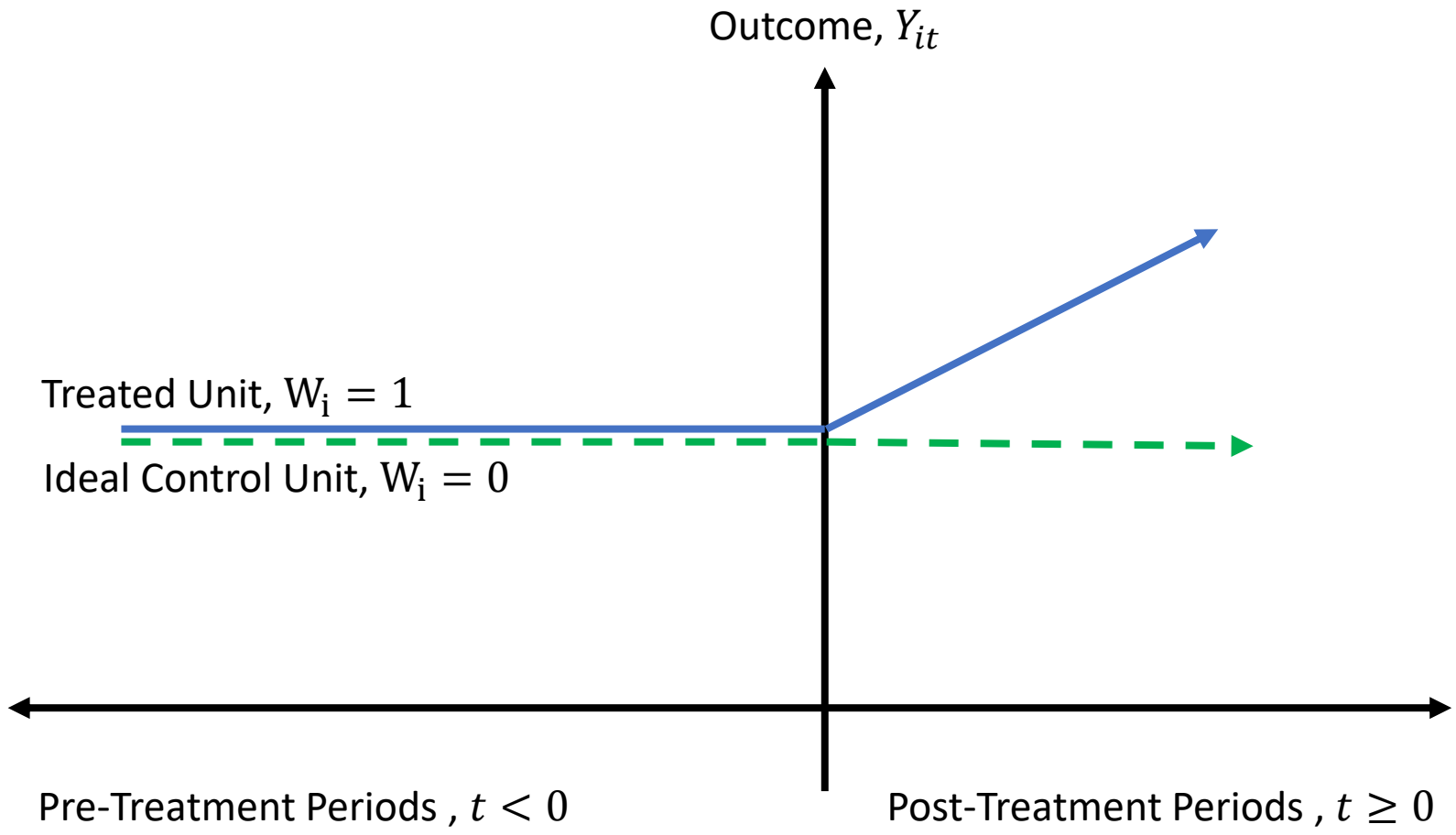
Overview

- This presentation will primarily cover how panel models can be used for causal inference, particularly difference-in-difference (DiD) and synthetic control-style models (SC).
- We will also discuss the role of prediction in panel models
- DiD is the most popular quasi-experimental design in economics for causal inference.
  - One quarter of NBER Working Paper series used diff-in-diff; and 16% of articles in top five economic journals (Currie et al, 2020)
- It exploits panel data to estimate causal impacts

Panel Data

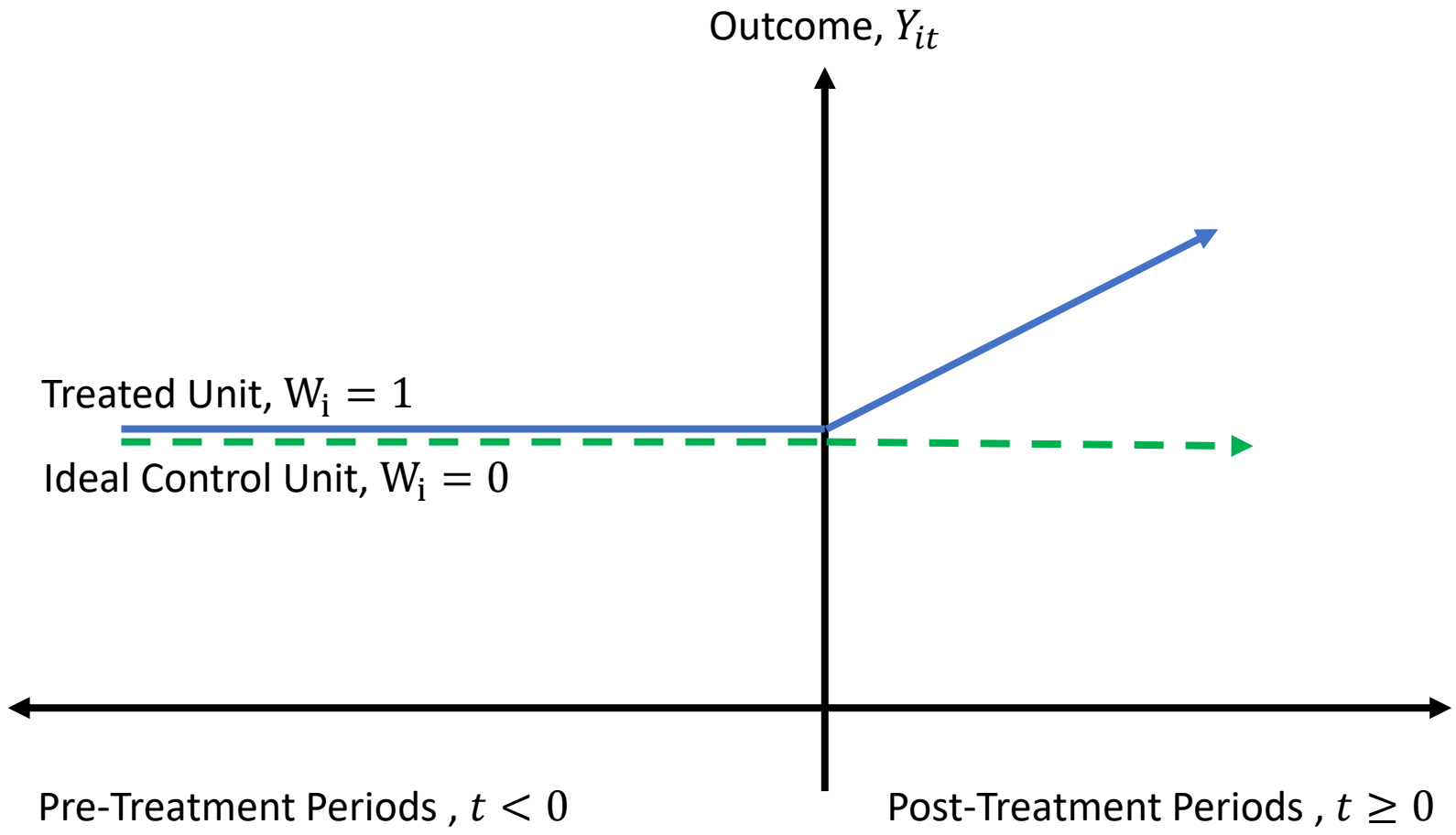
The big picture

- We track treated and control units over time, and see their outcomes before and after they are treated.
- Before treatment, their outcomes have the exact same trend.
- We assume that the difference in trends after treatment is due to treatment.



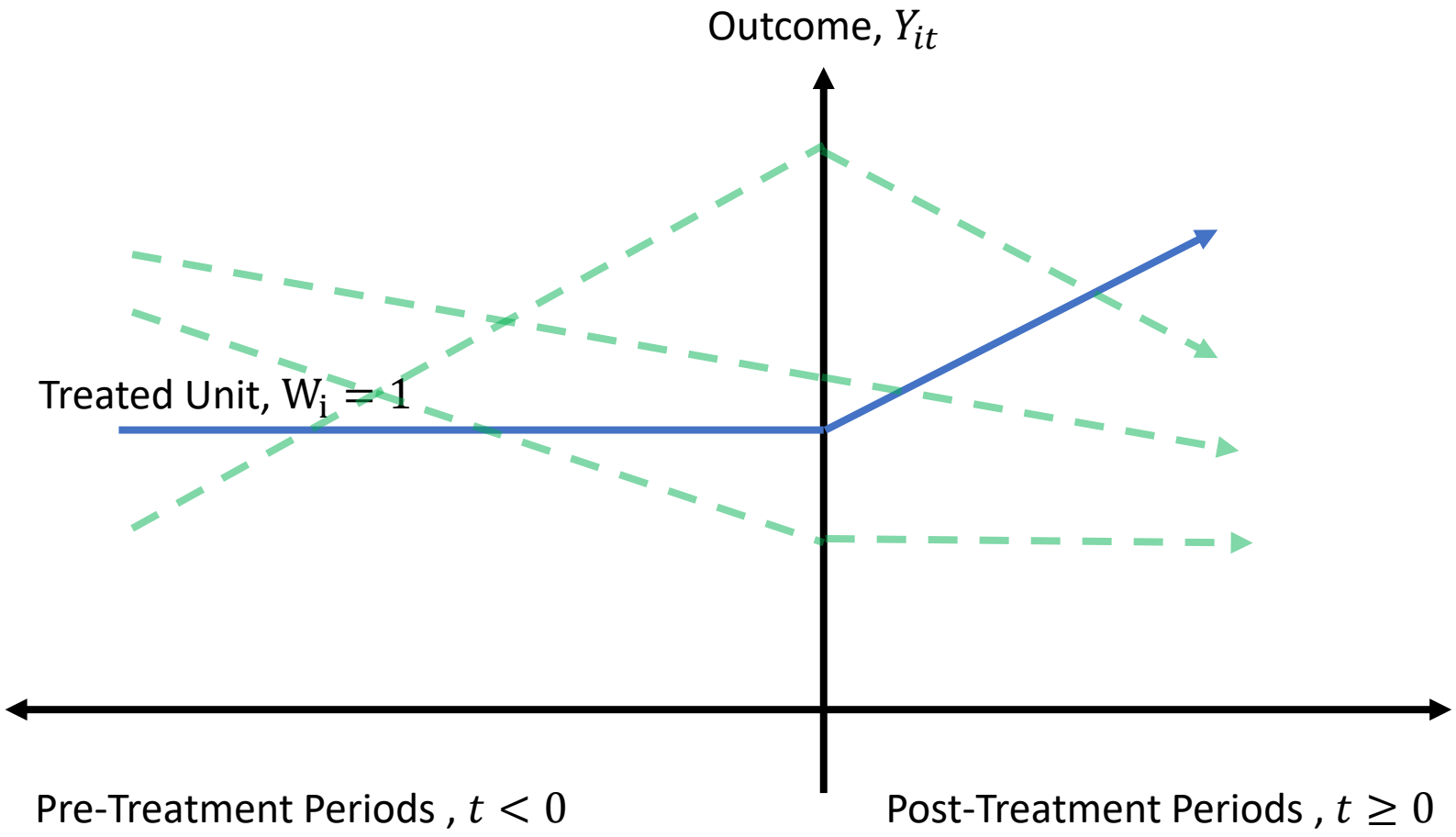
Comparing trends lets us arguably validate

- We want to know how the treated unit would behave if we did not treat it.
- The more similar pre-treatment trends are, the more we think that the control unit's post-treatment outcomes represent this.



Where DiD and SC models come in

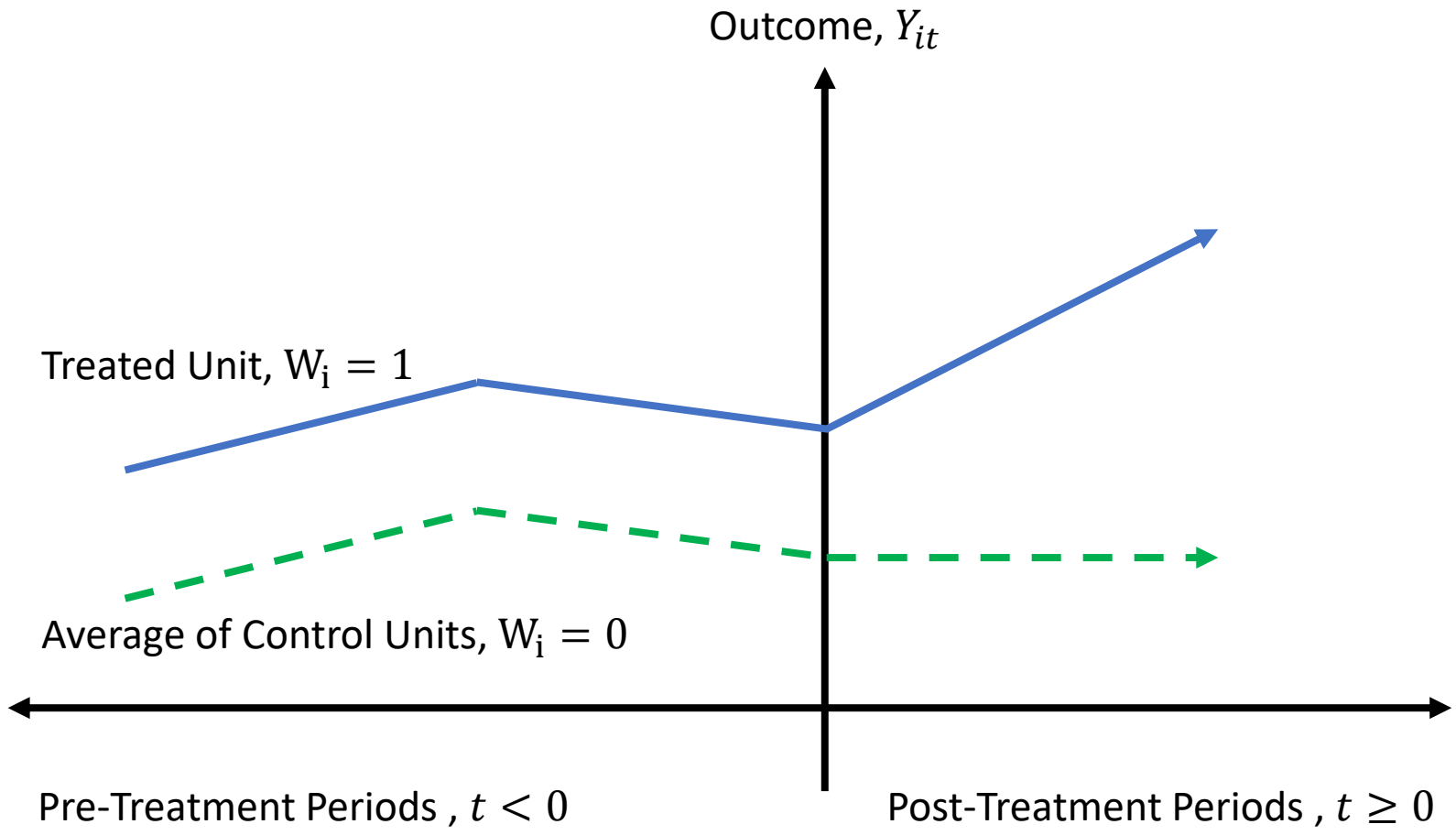
- How do we find the ideal control unit on the right? What if we can't?
- What if treatment is staggered (ie, roll out of a new algorithm over states or stores)?



Difference-in-Difference (DiD) Models

DiD Model

- We predict the treated unit's outcome if it were not treated ($Y_{W=1}(0)$) with the average of the control units, **and** assume this outcome is biased by a time-invariant constant



DiD Setup

- Let's setup our notation:
  - Before treatment: $t < 0$;
  - After treatment: $t \geq 0$;
  - Unit was treated: $W_i = 1$; and
  - Unit after it was treated: $D_{it} = 1$ iff $t \geq 0$ and $W_i = 1$

$$Y_{it} = \tau D_{it} + \epsilon_{it}$$

- We are interested in τ. DiD assumes ϵ_{it} has a time-invariant and a unit-invariant component:

$$Y_{it} = \tau D_{it} + \gamma_t + \eta_i + \zeta_{it}$$

DiD comes from combining two differences

$$Y_{it} = \tau D_{it} + \gamma_t + \eta_i + \zeta_{it}$$

- **Post-Pre Among Treated and Control:** Among treated and control subjects, compare Y_{it} before and after treatment.

$$\Delta_T = \mathrm{E}[Y_{it} | t \geq 0, W_i = 1] - E[Y_{it} | t < 0, W_i = 1]$$
$$\Delta_C = \mathrm{E}[Y_{it} | t \geq 0, W_i = 0] - E[Y_{it} | t < 0, W_i = 0]$$

- Notice that Δ_T cancels out the "η_i" part for treated units, similarly for Δ_C
- This means that Δ_T and Δ_C only have time-variant errors:

$$\Delta_T = \tau + \gamma_{t|t\geq 0} - \gamma_{t|t<0} + \zeta_{it|t\geq 0, W_i=1} - \zeta_{it|t<0, W_i=1}$$
$$\Delta_C = \gamma_{t|t\geq 0} - \gamma_{t|t<0} + \zeta_{it|t\geq 0, W_i=0} - \zeta_{it|t<0, W_i=0}$$

Deriving DiD

$$\Delta_T = \tau + \gamma_{t|t\geq 0} - \gamma_{t|t<0} + \zeta_{it|t\geq 0, W_i=1} - \zeta_{it|t<0, W_i=1}$$
$$\Delta_C = \gamma_{t|t\geq 0} - \gamma_{t|t<0} + \zeta_{it|t\geq 0, W_i=0} - \zeta_{it|t<0, W_i=0}$$

- First, we can see that the time terms, γ_t , are the same in Δ_T and Δ_C.
- Second, do we think that the ζ_{it} are the same, on average, between treatment and control too? In other words, do we think that treatment and control differ in time-varying ways?
- If so, then we can take the second difference to estimate τ :

$$\Delta_T - \Delta_C = \tau$$

Estimating DiD

- A simple model is where we aggregated fixed effects to be between treatment and control groups, and the time fixed effects of being before or after treatment periods:

$$Y_{it} = \tau D_{it} + \gamma 1\{t \geq 0\} + \eta W_i + \epsilon_{it}$$

- The current standard is a two-way fixed effects model where we control for all individual fixed effects, η_i and time fixed effects, γ_t:

$$Y_{it} = \tau D_{it} + \gamma_t + \eta_i + \epsilon_{it}$$

  - This gives the same estimate as above, but with additional precision.

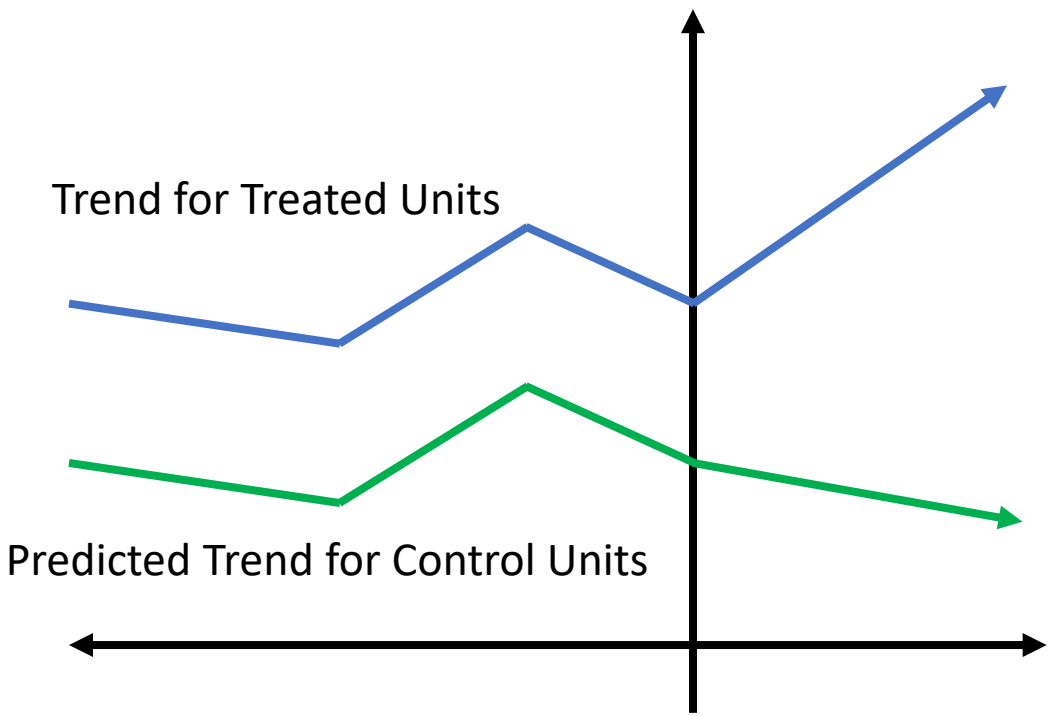
Validating DiD

- If treated and control units vary in time-varying ways that influence their outcomes, then we cannot trust the DiD estimate.
  - We do not think that the average control trend accurately represents how treated units would perform absent treatment.

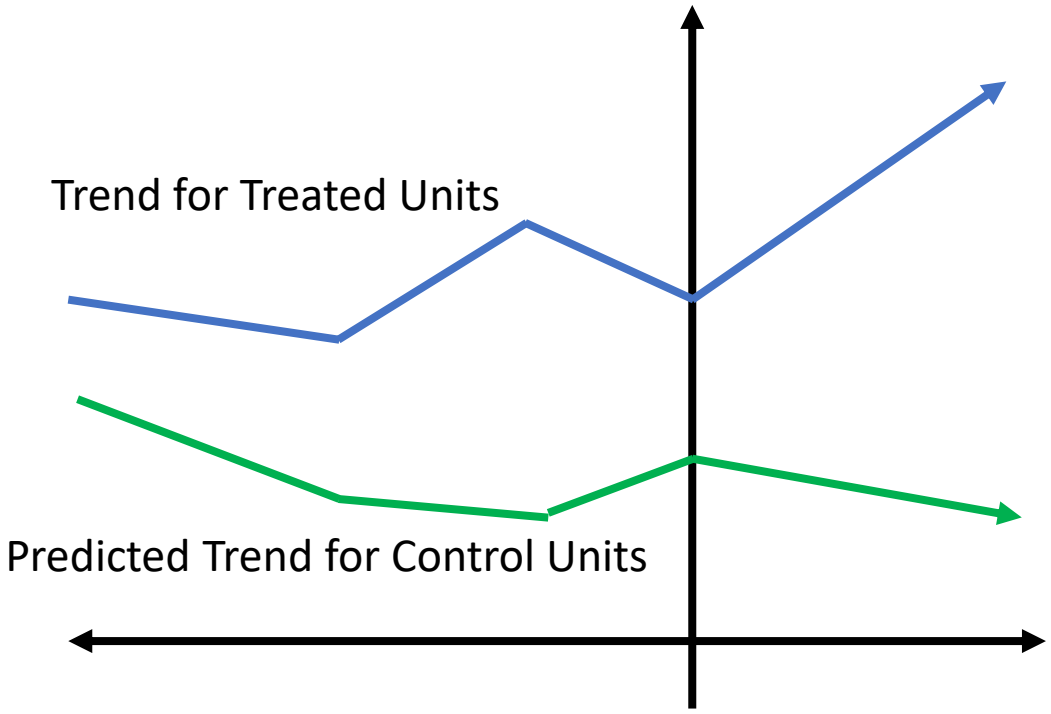
- We can arguably assess this by looking at whether the predicted control trend is parallel to the treated units.
- This is the Parallel Trends test.

Example of Parallel and Non-Parallel Trends

Parallel Trends Test: Yes



Parallel Trends Test: No



Two ways to Assess this

1. **Eye-ball check**: plot the treatment and control trends and see if they look similar. The drawback is that you cannot see if trends are statistically different;
2. **Event-study test**: estimate an altered version of the two-way fixed effects model where you allow for the impact of treatment to vary before and after treatment.

$$Y_{it} = \sum_{t' < 0} \tau_{t'} W_i + \sum_{t'' \geq 0} \tau_{t''} W_i + \gamma_t + \eta_i + \epsilon_{it}$$

If $\tau_{t'}$ is statistically indistinguishable from zero, the trends are parallel.

This is a-kin to a placebo test. Before the treatment took place, the treatment effect should be zero.

Including Time-Varying Covariates

- We can potentially improve on the two-ways fixed effects model by controlling for time-varying covariates:

$$Y_{it} = \tau D_{it} + \pi X_{it} + \gamma_t + \eta_i + \epsilon_{it}$$

- Improvement comes from (1) additional precision from reducing noise; and (2) greater likelihood to meet parallel trends assumption by conditioning on covariates.
- You run the risk of model misspecification bias if you X_{it} enters linearly. Sant'Anna and Zhao (2018) discusses how to incorporate ML models.

More DiD Extended Topics

- Roth, Sant'Anna, Bilinski, Poe (2022) for a comprehensive overview of the current literature. Some highlights below:
- Staggered Treatment Effects
  - Chaisemartin and D'Haultfœuille (2020)
  - Callaway and Sant'Anna (2018)
- Continuous Treatment
  - Callaway, Goodman-Bacon, Sant'Anna
- Doubly Robust DiD
  - Sant'Anna and Zhao (2018)

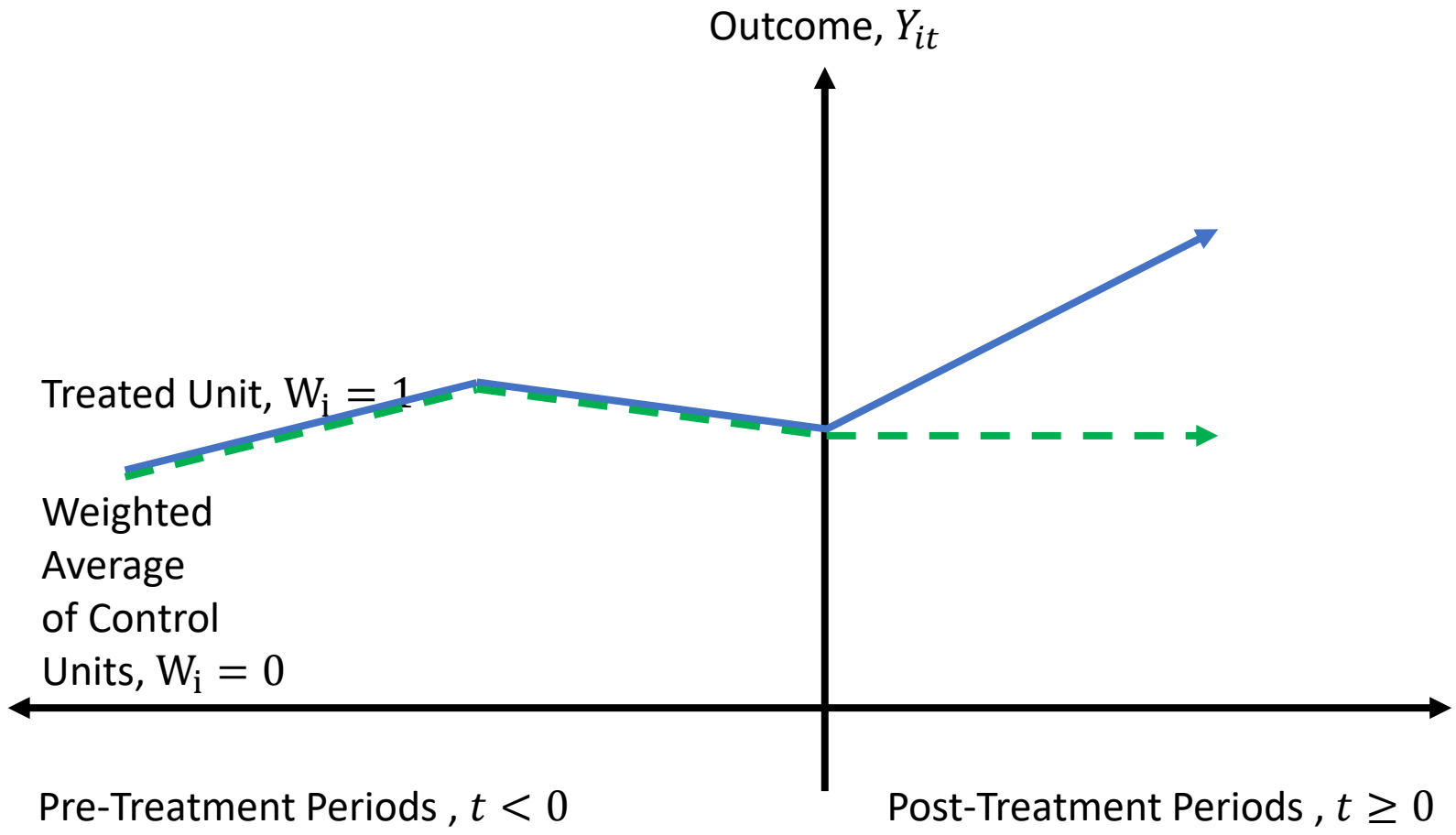
Synthetic Control

Abadie, Diamond, Hainmueller (2010)

Doudchenko and Imbens (2016)

SC Model

- We predict the treated unit's outcome if it were not treated with a weighted average of the control units
- Weights allow for a data-driven selection of control units



SC Setup

- Let $i = 0$ indicate the treated unit, so Y_{0t} is the trend of treated units
- We want to predict what the treated units' outcomes would be, if they had not been treated: $Y_{0t}(0)$

$$\tau = \sum_{t''\geq 0} Y_{0t''}(1) - \sum_{t''\geq 0} Y_{0t''}(0)$$

- SC proposes estimating $Y_{0t}(0)$ with a weighted average of all other control units

$$\widehat{Y_{0t}}(0) = \hat{\mu} + \sum_{i>0} \widehat{\omega_i} Y_{it}$$

- This flexible notation may alarm some of you; we will cover different constraints on (μ, ω_i).

Validating SC Models

- We will cover how different SC models go about estimating (μ, ω_i) in more detail soon, but at a high-level, an optimizer estimates (μ, ω_i) to best predict the pre-treatment of the treated unit.
- We can validate the prediction by having a hold-out sample of recent pre-treatment outcomes.
  1. Estimate (μ, ω_i) using data on potential control units $i > 0$ and for time periods $t < 0 \ - T_p$ where T_p is the number of hold-out sample periods.
  2. Evaluate whether (μ, ω_i) does a good job predicting outcomes for T_p periods.

SC, the big idea

- We want to have a data-driven way of identifying the ideal control group.
- Since DiD uses all the control units in our data, we may find ourselves on a time-consuming and likely non-rigorous data-mining exercise to find units that pass the parallel trends test.
- SC models have different approaches to identify the relevant control units and how important they are.
- We will now go over two models that place more and less restrictions on estimating (μ, ω_i)

Abadie, Diamond, Hainmueller 2010 (ADH)

- This is a more restrictive approach to estimating the weights (μ, ω_i). These restrictions allow us to find a unique solution for (μ, ω_i) to predict pre-trends.
- Our restrictions are (1) $\mu = 0$; (2) $\omega_i > 0$; and (3) $\sum_i \omega_i = 1$.
- These restrictions mean that only a few units will have strictly positive weights, giving us a more interpretable result.
  - For example, we can find that out of 100 stores, only three stores are needed to predict the outcome of the treated store.

Doudchenko and Imbens 2016 – DI

- A less restrictive SC model uses cross-validation to allow a more flexible way of estimating (μ, ω_i).
- Compared to ADH:

1. DI allows for μ to take on any value, allowing us to predict the trend of a treatment unit that is outside the range of other control units (ie, stores with the lowest or highest sales); and
2. ω_i can be positive or negative, and do not necessarily need to add up to one. This allows additional precision to estimating the pre-trend. A regularization term via elastic net allows a unique solution, compared to ADH's constraints.

How does inference work?

- Once we have a set of estimates (μ, ω_i), we can estimate the treatment effect on the treated:

$$\tau = \sum_{t''\geq 0} Y_{0t''}(1) - \sum_{t''\geq 0} \widehat{Y_{0t''}}(0)$$

- The main way to do inference is to do permute over units or time periods. (Permutation / Fischer Exact Test approach)
- Pretend other control units $i = 1, \dots, N$ are treated and estimate a corresponding placebo treatment $\tau_1, \dots, \tau_N$. Compare τ to these placebo treatments. P-value is how many placebo treatments are less than τ.
  - A modified version is to weight each placebo treatment with its propensity score.

Including Time-Varying Covariates

- We may want to incorporate time-varying covariates to increase our prediction of the pre-treatment and post-treatment outcome.
- We essentially do this by forcing matching the trend of time-varying covariates and outcome.
  - In ADH, you can estimate ω_i that are a function of covariate specific weights λ_k for covariate k.
  - In DI, you can residualize Y_{it} as a function of covariates.
- We should first consider whether we need to impose these additional restrictions.

Extended SC Topics

- Check out Abadie 2021 for a review of the literature.
- Other ways of doing inference
  - Prediction Intervals via Cattaneo, Feng, Titiunik (2019)
  - Conformal Inference via Chernozhukov, Wuthrich, and Zhu (2017)
- K-Fold approach to estimating SC Models
  - Chernozhukov, Wuthrich, Zhu 2018
- Adding time-specific weights via synthetic DiD
  - Arkhangelsky, Athey, Hirschberg, Imbens, Wager 2021

DiD vs SC

	DiD	SC
Control Group Definition	Unweighted average of all potential controls, allowing for a time-invariant difference	Weighted average for a subset of all potential controls to exactly match treatment
Computation Speed	OLS is fast.	Optimizers can take a while.
Inference Procedure	Done with OLS	Permutation; or others (see Chernozhukov, Wuthrich, Zhu (2018), Cattangeo, Feng, Titiunik (2019).)
Validation	Parallel Trends Test with event study model	Prediction validation of hold-out pre-treatment outcomes
Time-Varying Covariates	Linearly, or with ML models via doubly robust methods.	Depends on the SC model

Appendix and Old Slides

Defining the difference-in-difference estimator

Foundation using panel data

- Panel data is when each subject is tracked across multiple time periods. For example, knowing the purchase history of a given customer or account for each calendar day.
- With greater data storage capacity, panel data is becoming more available to scientists for analysis.

Simple panel data set

$$Y_{it} = \tau W_{it} + \epsilon_{it}$$

- Each subject i is tracked for time periods t. The outcome of interest is $\mathrm{Y_{it}}$.
- Suppose that all subjects are untreated before $t = 0$ ($\mathrm{W_{it}} = 0 \; \forall \; t < 0$), and some are treated starting at $t = 0$. Starting from $t = 0$, treatment and control groups are mutually exclusive and are permanently assigned.

$$W_{it} = \begin{cases} 0, t < 0 \\ 1, t \geq 0 \, , i \in (1, \dots, N_T) \\ 0, t \geq 0 \, , i \in (N_{T+1}, \dots, N) \end{cases}$$

- Treatment groups never switch to control.
- Treatment group: subjects that are eventually assigned treatment
- Control group: subjects that are never assigned treatment

Two naïve comparisons

- $Y_{it} = \tau W_{it} + \epsilon_{it}$

1. **Post-Pre:** Among treated subjects, compare Y_{it} before and after treatment. This compares across time.
   - $\hat{\tau} = \mathrm{E}[Y_{it} | t \geq 0, i \in Treat] - E[Y_{it} | t < 0, i \in Treat]$
   - Problematic because you do not know how much of $\hat{\tau}$ is due to the true difference τ and ϵ_{it}
   - For example, suppose that there was a change in a feature X_{it} around the same time.
2. **Treatment-Control:** Among observations after treatment assignment, compare treatment and control. This compares across units.
   - $\hat{\tau} = \mathrm{E}[Y_{it} | t \geq 0, i \in Treat] - E[Y_{it} | t \geq 0, i \in Control]$
   - Problematic for the same reason. Treatment and control can differ in terms of ϵ_{it}.

Instead, let's compare across time and units

$$Y_{it} = \tau W_{it} + \epsilon_{it}$$
$$Y_{it} = \tau W_{it} + \gamma_t + \eta_i + \epsilon_{it}$$

- Build additional structure to the model
  - Include fixed effects, or parameters, for each subject (η_i) and time period (γ_t)
- This additional structure shows us why the two naïve approaches are problematic and provides a solution for an unbiased estimate of τ, under conditions similar to the cross-sectional models.

Revising the naïve comparison: **Post-Pre**

- $Y_{it} = \tau W_{it} + \gamma_t + \eta_i + \epsilon_{it}$
- $\text{Post}-\text{Pre} = \text{E}[Y_{it}|t \geq 0, i \in T] \; - E[Y_{it}|t < 0, i \in T]$
- $\left(\tau + \gamma_{t|t\geq 0} + \eta_{i|i\in T} + \epsilon_{it \,|\, t\geq 0\,, i\in T}\right) - \left(\gamma_{t|t<0} + \eta_{i|i\in T} + \epsilon_{it \,|\, t<0\,, i\in T}\right)$
- These cancel out, leaving us:
- $\text{Post}-\text{Pre} = \tau + \left(\gamma_{t|t\geq 0} - \gamma_{t|t<0}\right) + \left(\epsilon_{it \,|\, t\geq 0\,, i\in T} \; - \epsilon_{it \,|\, t<0\,, i\in T}\right)$
- Problem is that we cannot distinguish the true treatment effect from time trend changes

Revising the naïve comparison: **Treatment-Control**

- $Y_{it} = \tau W_{it} + \gamma_t + \eta_i + \epsilon_{it}$
- $\text{Treatment} - \text{Control} = \text{E}[Y_{it} | t \geq 0, i \in T] - E[Y_{it} | t \geq 0, i \in C]$
- $\left(\tau + \gamma_{t|t\geq 0} + \eta_{i|i\in T} + \epsilon_{it \mid t\geq 0, i\in T}\right) - \left(\gamma_{t|t\geq 0} + \eta_{i|i\in C} + \epsilon_{it \mid t\geq 0, i\in C}\right)$
- These cancel out, leaving us:
- $\text{Treatment} - \text{Control} = \tau + \left(\eta_{i|i\in T} - \eta_{i|i\in C}\right) + \left(\epsilon_{it \mid t\geq 0, i\in T} - \epsilon_{it \mid t\geq 0, i\in C}\right)$
- The problem is that we cannot distinguish the true treatment effect from other time-invariant differences between treatment and control

Construct the diff-in-diff estimator

- The naïve comparisons suffer from being unable to distinguish between time-variant and time-invariant differences between treatment and control.
- Propose combining these two approaches to compensate for each others shortcomings:
- Intuitively, take two differences:
  1. Treatment subjects before and after treatment time; and
  2. Control subjects before and after treatment time.
- and then take the difference between them.

Diff-in-diff estimator

$$\Delta_T = \left[\mathrm{E}[Y_{it} | t \geq 0, i \in T] - \mathrm{E}[Y_{it} | t < 0, i \in T] \right]$$
$$\Delta_C = \left[E[Y_{it} | t \geq 0, i \in C] - E[Y_{it} | t < 0, i \in C] \right]$$

- We can show that:
- $\Delta_T = \tau + \gamma_{t|t\geq 0} - \gamma_{t|t<0} + \epsilon_{it|t\geq 0, i\in T} - \epsilon_{it|t<0, i\in T}$
- $\Delta_C = \gamma_{t|t\geq 0} - \gamma_{t|t<0} + \epsilon_{it|t\geq 0, i\in C} - \epsilon_{it|t<0, i\in C}$
- Therefore:
- $\Delta_T - \Delta_C = \tau + \left(\epsilon_{it|t\geq 0, i\in T} - \epsilon_{it|t<0, i\in T}\right) - \left(\epsilon_{it|t\geq 0, i\in C} - \epsilon_{it|t<0, i\in C}\right)$
- The ϵ_{it} part looks monstrous, but we have seen this before when we studied the fundamentals of causal inference for cross-sectional, or propensity-based models

Unconfoundedness, again

$$\Delta_T - \Delta_C = \tau + \left(\epsilon_{it|t\geq 0, i\in T} - \epsilon_{it|t<0, i\in T}\right) - \left(\epsilon_{it|t\geq 0, i\in C} - \epsilon_{it|t<0, i\in C}\right)$$

- We want to assume the differences in errors to be zero in expectation. This means that after controlling for time specific and subject specific effects, we assume treatment is exogeneous. Similar to the unconfoundedness assumption.
- This means there are inherently eight counterfactuals. For treated subjects, we only observe outcomes before and after treatment. For control subjects, we only observe outcome before and after control.

For subjects actually treated	For subjects actually in control
$\mathrm{E}[Y_{it}\|t \geq 0, i \in T]$	$\mathrm{E}[Y_{it}\|t \geq 0, i \in T]$
$\mathrm{E}[Y_{it}\|t < 0, i \in T]$	$\mathrm{E}[Y_{it}\|t < 0, i \in T]$
$E[Y_{it}\|t \geq 0, i \in C]$	$E[Y_{it}\|t \geq 0, i \in C]$
$E[Y_{it}\|t < 0, i \in C]$	$E[Y_{it}\|t < 0, i \in C]$

Observed
Not observed

Panel data's advantage over cross-sectional data

$$Y_{it} = \tau W_{it} + \gamma_t + \eta_i + \epsilon_{it}$$

- With cross-sectional, we cannot control for individual fixed effects η_i. We would have to use some flexible function of X_i.
- We could instead of η_i, control for X_{it}.

$$Y_{it} = \tau W_{it} + \gamma_t + g(X_{it}) + \epsilon_{it}$$

- This has implications for whether we believe the unconfoundedness assumption is true with X_{it}, and why we are not using η_i instead.

Difference-in-difference model

- Two-way fixed effects model where we control for all individual fixed effects, η_i and time fixed effects, γ_t:

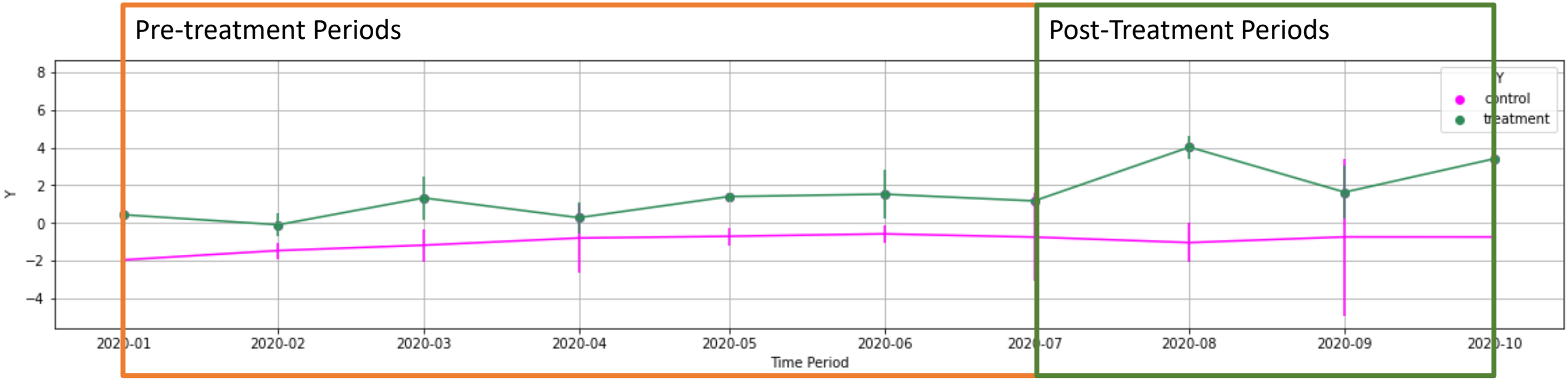
$$Y_{it} = \tau W_{it} + \gamma_t + \eta_i + \epsilon_{it}$$

- A simpler model is where we aggregated fixed effects to be between treatment and control groups, and the time fixed effects of being before or after treatment periods:

$$Y_{it} = \tau W_{it} + \gamma 1\{t \geq 0\} + \eta 1\{i \in T\} + \epsilon_{it}$$

Let’s see this come out in simulation evidence

- We generate a panel of outcome data, with seven pre-treatment periods, and three post-treatment periods.



Compare performance of cross-sectional to panel methods

- What if we ignored the panel structure, and estimated impacts only using pre-treatment period:

$$Y_{iT_1} = \tau W_i + X_{iT_0} + \epsilon_{iT_0}$$

- Where T_1 is a post-treatment period and T_0 is a pre-treatment period, so X_{iT_0} are pre-treatment features?
- If we control for enough features to meet the unconfoundedness assumption, then we should have the same performance as a difference-in-difference model will.

Simulation evidence of cross-sectional and panel data

		Estimate of τ	True Value
Cross-Sectional Estimates	First Pre-Period	2.679[1.502]	2.500
	Middle Pre-Period	3.510[0.848]	2.500
	Last Pre-Period	0.476[1.397]	2.500
Panel Estimate	Two-Way Fixed Effects	1.936 [0.577]	2.500
	Aggregated Fixed Effects	1.936 [0.537]	2.500

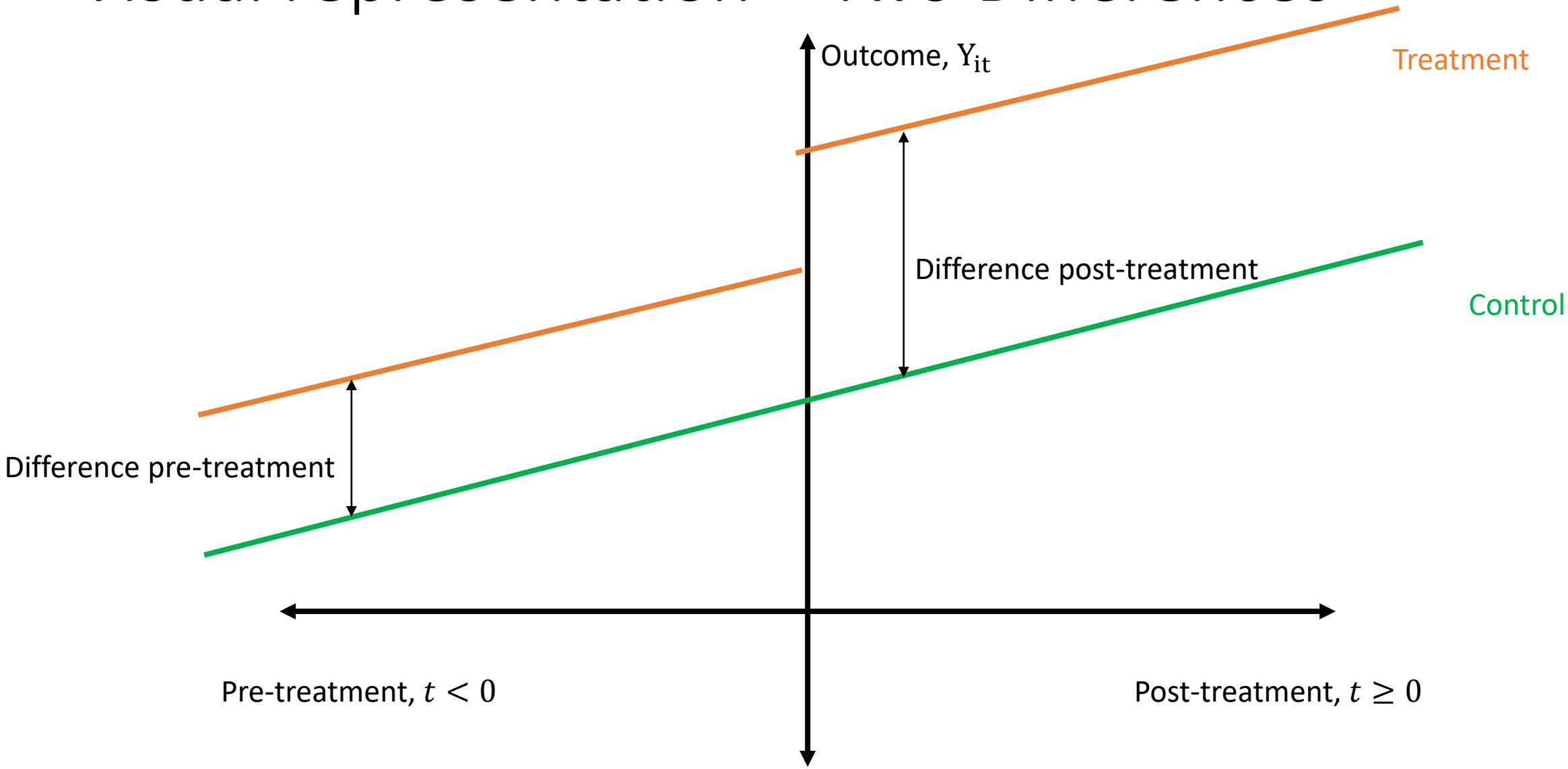
- Performance of cross-sectional models can vary over pre-treatment periods, but we have more consistent performance

Arguable validation using panel data

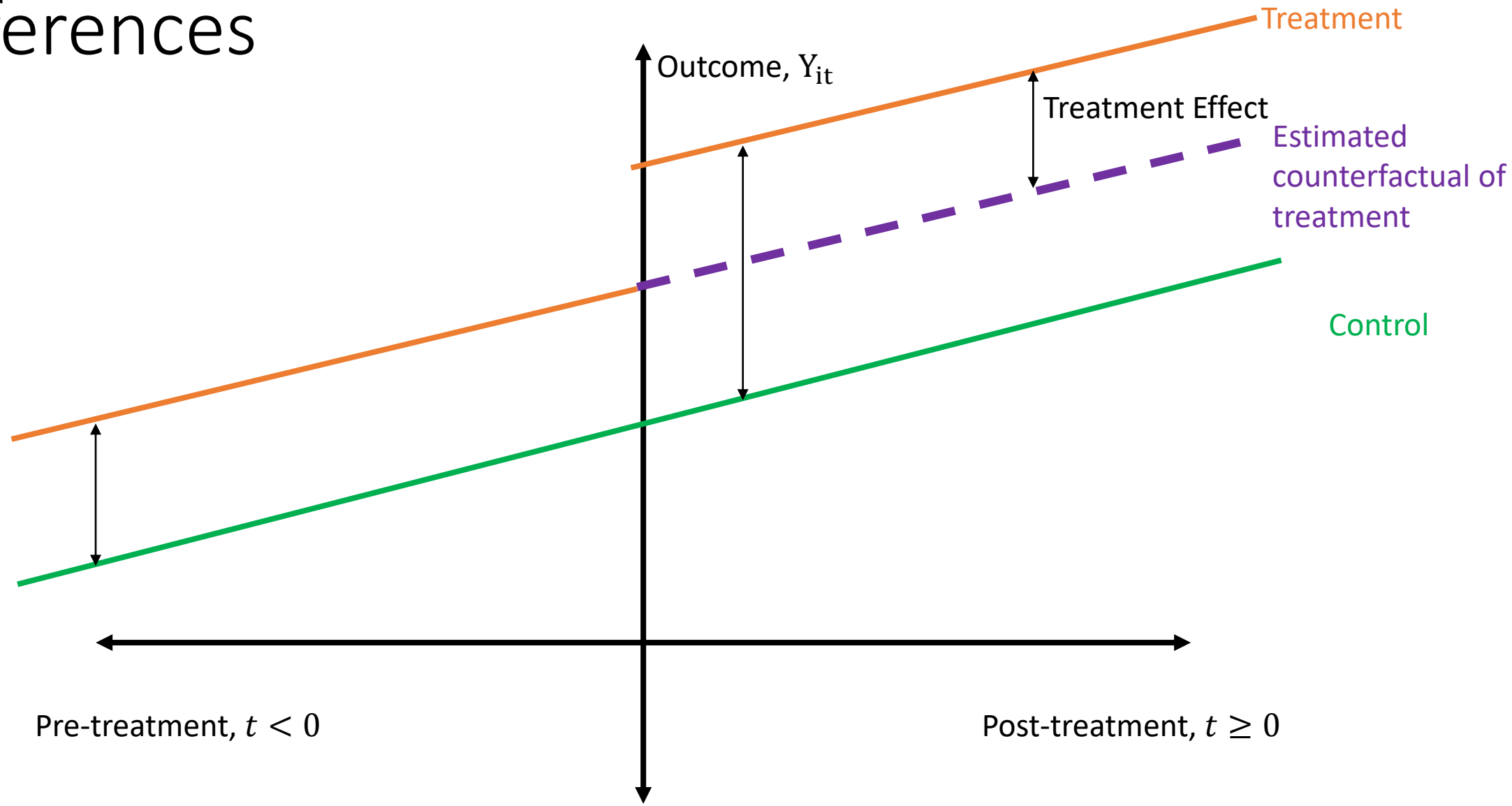
Difference-in-difference

- The diff-in-diff estimator comes from comparing the difference over time of the treatment group, to the difference over time of the control group.
- Therefore, the diff-in-diff estimator is identified – or yields an unbiased estimate – if the treatment and control have similar trends before treatment time.
- Differences in the outcome after treatment time ($t \geq 0$) are interpreted as due to treatment and what would have happened anyway.

Visual representation – Two Differences



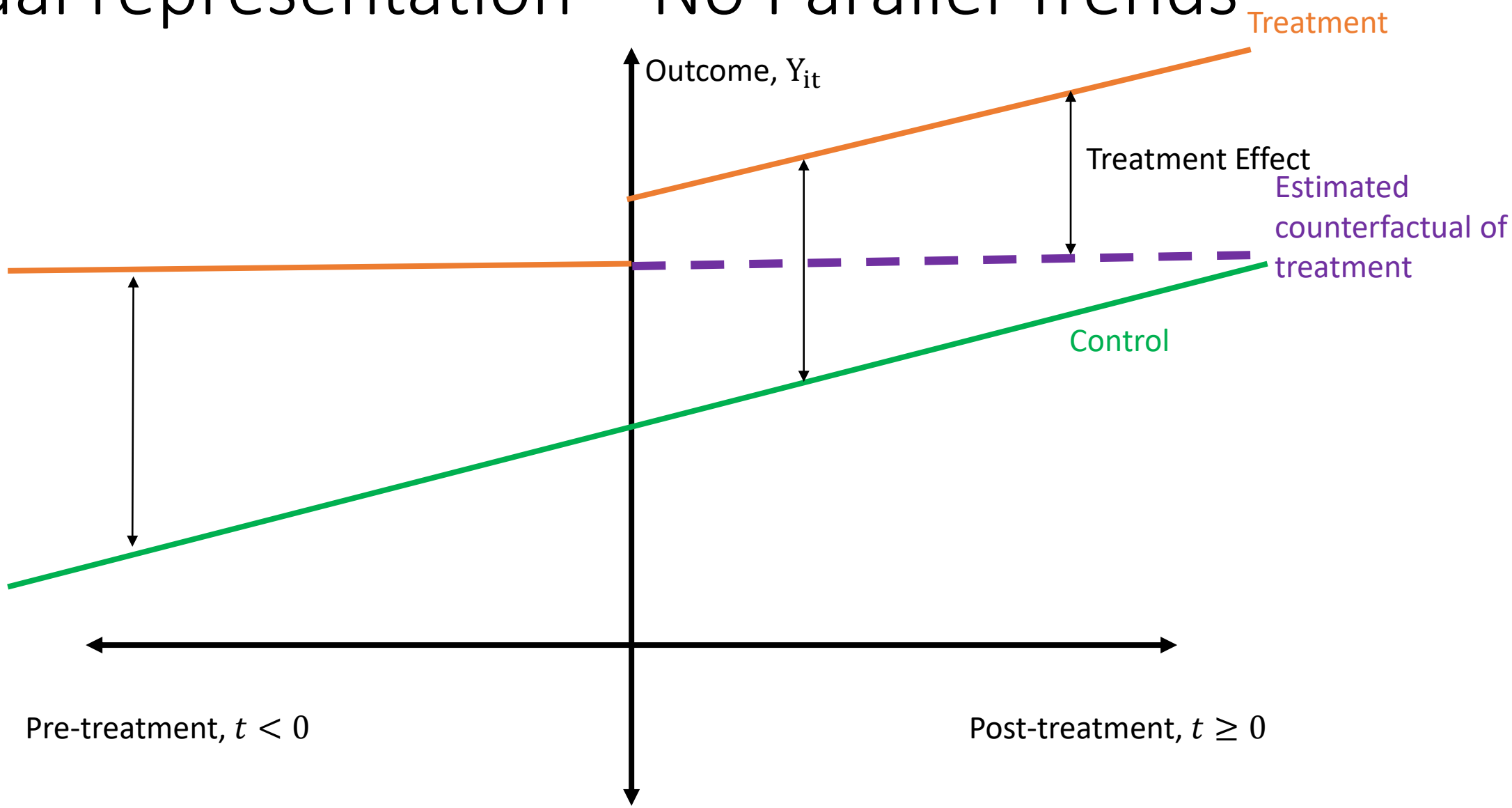
Visual representation – Difference in Differences



Pre-trend tests

- If treatment and control groups do not have parallel trends in outcomes pre-treatment, then we cannot distinguish how much of the post-treatment difference is due to the true treatment effect and what would have happened anyway.
- If the treatment was on a different trajectory of the control, then how can we be sure how the treatment would have behaved if it were not treated?

Visual representation – No Parallel Trends



Pre-trend tests

- Estimate the impact of the treatment on pre-treatment outcomes

$$Y_{it} = \sum_{t' \in (t<0)} \beta_{t'} W_{it'} + \sum_{s' \in (t \geq 0)} \alpha_{s'} W_{is'} + \gamma_t + \eta_i + \epsilon_{it}$$

- Where $W_{it'}$ is whether a subject is ever treated interacted with an indicator of time t'
- Then $\beta_{t'}$ estimates the impact of the treatment on pre-treatment outcomes, and $\alpha_{s'}$ estimates the impact on post-treatment outcomes.

Staggered treatment time

Identification deep dive

Identification deep dive

- Controlling for lagged outcome or not, and its implication for inference
- Controlling for features instead of fixed effects

Continuous treatments