Good Controls Gone Bad:
Difference-in-Differences with Covariates[*]

Sunny Karim[†]   Matthew D. Webb [‡]

November 26, 2025

This paper introduces the two-way common causality of covariates (CCC) assumption in Difference-in-Differences (DiD), an implicit assumption requiring that covariates have the same effect on the untreated potential outcome across regions and time. We show that while conditional parallel trends (CPT) may still hold when the two-way CCC assumption is violated, this has not been recognized in the literature, leading researchers to incorrectly impose common covariate coefficients. Through theoretical analysis and Monte Carlo simulations, we demonstrate that many existing DiD estimators become inconsistent when covariates are modeled with a common coefficient while the two-way CCC assumption is violated. We further show that DiD estimators incorporating propensity scores cannot accommodate two-way CCC violations. To address these issues, we propose the Intersection Difference-in-Differences (DID-INT) estimator, which remains consistent under two-way CCC violations, staggered treatment adoption, and heterogeneous treatment effects. We also recommend researchers to assess the plausibility of the Intersection Parallel Trends (IPT), which remains valid under two-way CCC violations.

1 Introduction

Difference-in-differences (DiD) is a widely used method for assessing the effectiveness of a policy which is implemented non-randomly at a provincial level. In the simplest two group and two period setting, DiD compares the difference in outcomes before and after treatment between the group which received treatment and the group which did not (Bertrand et al., 2004). This simple setup serves as the building block for estimating the average treatment effect on the treated (ATT) within the more complex staggered treatment rollout framework in methods

proposed by Callaway and Sant'Anna (2021); De Chaisemartin and d'Haultfoeuille (2023) and Sun and Abraham (2021).

Both conventional and modern DiD approaches rely on well-documented assumptions to support consistent estimation of the ATT. Among the key identifying assumptions which include no anticipation and homogeneous treatment effects, the strong parallel trends assumption is the most crucial (Roth et al., 2022; Abadie, 2005; De Chaisemartin and d'Haultfoeuille, 2020a; Callaway and Sant'Anna, 2021). It asserts that, in the absence of treatment, the average outcomes between the treated groups and control groups would have moved parallel to each other in the absence of treatment (Abadie, 2005). Since we do not observe the untreated potential outcomes for the treated group, researchers examine pre-intervention trends between the treated and the control groups to assess the plausibility of parallel trends after intervention. To improve the plausibility of parallel trends, researchers impose the parallel trends assumption to hold only conditional on covariates (Roth et al., 2022). Conventional DiD estimation strategies with $s = 1, 2, ..., S$ groups and $t = 1, 2, ..., T$ involve running the following two-way fixed effects (TWFE) regression with covariates (Bertrand et al., 2004):

$$Y_{i,s,t} = \alpha_s + \delta_t + \beta^{DD} D_{i,s,t} + \sum_k \gamma^k X_{i,s,t}^k + \epsilon_{i,s,t}. \tag{1}$$

where, $Y_{i,s,t}$ is the outcome variable of interest for individual i in region s in period t. α_s represents region fixed-effects that accounts for unobserved heterogeneity, δ_t denotes time fixed effects, $D_{i,s,t}$ is the treatment indicator , and $X_{i,s,t}^k$ are covariates which can either be time invariant or time varying. In this model, there are a total of K covariates. Current methods of including covariates implicitly assume that the effect of the covariates on the untreated potential outcome remain constant between regions and periods, but do not address this explicitly. These methods are discussed in Section 2. In this paper, we explicitly introduce this assumption, which we call the **common causality of covariates (CCC)**. Specifically, we introduce three types of CCC assumptions: region-invariant CCC, time-invariant CCC and the two-way CCC. In this paper, we show that these assumptions are necessary in both conventional and newer DiD methods to obtain a consistent estimate of the ATT. However, using the Labour Force Survey (LFS), we demonstrate a case where the CCC assumption appears to be violated. We also show, through both theoretical proofs and a Monte Carlo Simulation Study, that the TWFE, the CS-DID, the imputatation and the FLEX estimators can be inconsistent when the CCC assumption is violated.

We show that the observed inconsistency in the TWFE, CS-DID, Imputation and FLEX estimators results from implicitly assuming that the two-way CCC holds whenever CPT seems plausible. The consistency of these estimators rely on this implicit assumption that covariates share a common coefficient similar to the two-way fixed effects specification shown in Equation (1). We show that conditional parallel trends can still hold when the two-way CCC assumption is violated, which is not explicitly known in the existing literature. In addition, we posit

2

that one of the necessary conditions for conditional parallel trends is hard to interpret without additional assumptions on covariates, and is very unlikely to hold except by coincidence. Unfortunately, most of the existing estimators are not easily modified to accommodate CCC violations.

Building on this observation, we introduce an alternative to the conditional parallel trends assumption which we call **Intersection Parallel Trends (IPT)**. The intersection parallel trends assumption remains plausible when the two-way CCC assumption is violated, and it requires fewer conditions to hold compared to conditional parallel trends. To formally demonstrate this, we derive the necessary conditions required for both the standard and the conditional parallel trends to hold. We show that the two-way common causality of covariates assumption, while not strictly necessary for conditional parallel trends, strengthens its plausibility. In addition, we demonstrate why incorporating covariates that vary with time can complicate DID analysis. We also derive the necessary conditions for intersection parallel trends and show that it requires fewer necessary conditions to hold compared to the conditional parallel trends assumption. Moreover, it does not impose any additional restrictions on either the distribution of covariates or their effects on the untreated potential outcome.

We propose a new estimator called the **Intersection Difference-in-differences (DID-INT)** which can provide a consistent estimate of the ATT under violations of the CCC assumption. The DID-INT estimator is also applicable in settings with staggered treatment adoption. Our estimator only requires the intersection parallel trends assumption to hold, but remains consistent under conditional parallel trends when the two-way CCC condition is violated, without imposing additional restrictions on covariates. This paper brings both negative and positive results to the literature on difference-in-differences. The negative result is that if the two-way CCC assumption is violated, then existing estimators can be inconsistent. The more positive result is that correcting for these violations can result in consistent estimates. Preliminary results from our Monte Carlo experiments suggest that very severe violations of the two-way CCC assumption "appear" in parallel trends figures. Currently, many researchers will just abandon a project when the parallel trends figures do not "look" parallel. Or, they will examine parallel trends conditional on covariates (but under the two-way CCC assumption), again abandoning the project if those trends do not look parallel. Plotting the residuals of the outcome variable regressed on flexible versions of the covariates can yield parallel trends, which are not present when the less flexible, and incorrect, version of the model for covariates is used. Figure 1 shows an example from our Monte Carlo in Section 11. These data come from a DGP where the two-way CCC is violated. The figure on the left plots conditional trends that are clearly not parallel if we assume CCC holds. The right plots trends in residuals after controlling for the covariates in the correct manner, these trends appear to be more plausibly parallel. This approach broadens the set of applications in which parallel trends can be found. This paper does not look at strategies to partially identify the ATT when parallel trends are violated, which is explored in more details in Rambachan and Roth (2023).
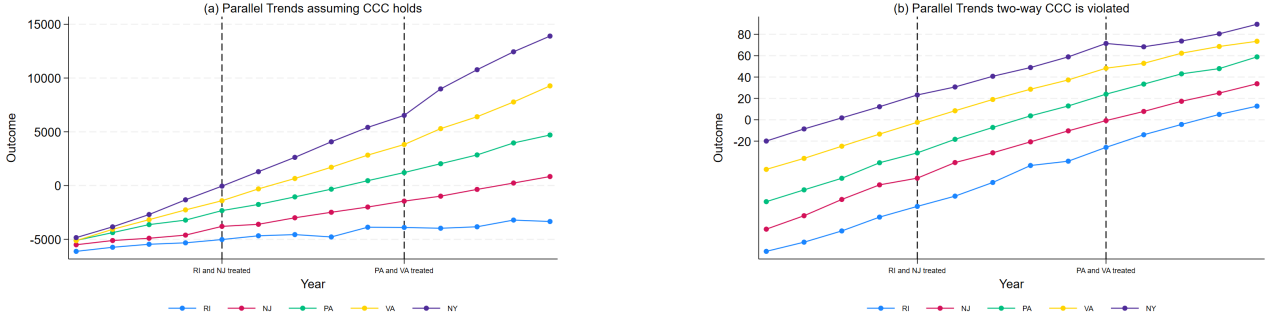
Figure 1: Unconditional and Corrected Parallel Trends in DGP with CCC Violations

The rest of the paper is as follows Section 2 explores existing approaches to including covariates. Section 3 introduces the key terminology used in the paper. Section 4 introduces the three types of CCC assumptions, and discusses several examples where violations of the CCC is plausible. Section 5 categorizes covariates into five types, using DAGs, based on the specific CCC assumption applied to them. In Section 6, we derive the necessary and sufficient conditions required for parallel trends, conditional parallel trends and intersection parallel trends, and discuss the key differences between them. Section 7 generalizes the analysis to multiple regions and periods, and introduces the staggered adoption design. Section 8 introduces the Intersection Difference-in-differences estimator, and describes four functional form of covariates that can be used in this estimator. It also introduces the model selection algorithm which can be used to select the appropriate functional form of covariates using parallel trends figures. Section 9 theoretically proves that TWFE is inconsistent under CCC violations. Section 10 discusses other estimators, namely the Callaway and Sant'Anna estimator in sub-section 10.1, the imputation estimator in sub-section 10.2 and the FLEX estimator in sub-section 10.3. Section 11 describes the design and results of Monte Carlo experiments and Section 12 concludes.

2 Existing Approaches for Including Covariates

While empirical researchers usually include covariates to ensure the plausibility of conditional parallel trends in DiD, they might also use them to control for factors that can influence outcomes. The literature highlights the need to choose covariates carefully in DiD analyses. Notably, covariates that are affected by participation in treatment, known as **bad controls**, should be excluded (Caetano and Callaway, 2024). The DiD literature also recommends using either time-invariant or pre-treatment covariates when the covariates change with time (Caetano and Callaway, 2024). However, researchers might still want to include time varying covariates that change with time, even though they are not needed to maintain parallel trends. For instance, consider a study where we are interested in the effect of a hypothetical treatment in reducing cardiac arrests, and the treatment is implemented at a provincial level. In such a study, researchers may want to control for time varying covariates like age and smoking status. Age, in particular, is unlikely to be affected by the treatment, and being older increases the

frequency of cardiac arrests. Including pre-treatment values of age in this analysis may lead to counter-intuitive results, as we are unable to capture the effect of a higher age on the frequency of having cardiac arrests. Additionally, many datasets are repeated cross-sections, rather than true panels, and pre-treatment values are typically not available in these datasets.

Caetano and Callaway (2024) uncovers the shortcomings of TWFE regressions using both time-invariant and time-varying covariates, even in setups with two periods without staggered treatment rollout. In particular, Caetano and Callaway (2024) pointed out three additional assumptions we need to make in order to completely rule out the inconsistency. The first assumption requires the path of untreated potential outcomes to be independent of time-invariant covariates. The second assumption states that this path should depend solely on the change in time-varying covariates. Lastly, the third assumption specifies that the relationship between the path of untreated potential outcomes and the change in time-varying covariates is linear. In staggered adoption designs with heterogeneous treatment effects, the inconsistency in TWFE is further exacerbated due to negative weighting issues and forbidden comparisons (Goodman-Bacon, 2021).

To address these shortcomings, the DID literature has proposed several alternatives to obtaining a consistent estimate of the ATT. In the absence of staggered adoption, Caetano et al. (2022) recommends using the Doubly-Robust DiD method proposed by Sant'Anna and Zhao (2020). However, it still relies on two distributional assumptions on covariates, namely covariate exogeneity and covariate unconfoundedness. Covariate exogeneity stipulates that the distribution of covariates for the treated group are not changed due to participation in treatment. Conversely, covariate unconfoundedness requires that the distribution of the untreated potential covariates are the same for the treated and the untreated groups. Another alternative in the absence of staggered adoption is the Augmented inverse probability weighting (AIPW) estimator proposed by Caetano and Callaway (2024). However, this method requires dimension reduction of covariates, which poses an additional challenge in its implementation.

For designs involving staggered adoption, the Callaway and Sant'Anna DiD (CSDID) offers a feasible alternative, which estimates the ATT without the forbidden comparisons (Callaway and Sant'Anna, 2021). The estimation of the ATT using this method involves two steps. First, the dataset is partitioned into several "2x2 comparison" blocks, each consisting of a treated group and an untreated (or not yet treated) group. The ATT for each "2x2 comparison" block, denoted as $ATT(g,t)$, is estimated using the doubly-robust DiD estimator developed by Sant'Anna and Zhao (2020). Second, these estimates are aggregated together by taking a weighted average of the $ATT(g,t)$'s from the first step to obtain an overall estimate of the ATT. Since the CS-DID uses the DR-DID by default to estimate the $ATT(g,t)$'s for each 2X2 "blocks", it still requires covariate exogeneity to hold. Note g here comprises all individuals who are first treated in the same period. This is usually, but not always, a collection of s with the same treatment timing.

A second alternate, called the static imputation estimator (Borusyak and Jaravel, 2017), is done in four steps. First, a TWFE model is estimated using data for the control group. Next, this fitted model is used to predict the untreated counterfactuals for the treated observations in the post-intervention period. In step 3, the unit specific treatment effects are estimated by taking the difference between the observed outcome and the predicted counterfactual from the previous step. Finally, these unit specific treatment effects are aggregated together to get an overall estimate of the ATT. While Borusyak et al. (2024) primarily focuses on event study designs, the paper also includes a section on the static version of the imputation estimator, which we examine in our paper.

A third alternative, called FLEX, uses a linear regression with a flexible functional form. This estimator was developed in part to allow for time-varying covariates. The estimator includes group-by-time treatment effects, two-way fixed effects, and interaction terms with leads and lags (Deb et al., 2024).

	Time-invariant covariates	Staggered adoption	Time-varying covariates	CCC violations
TWFE	✓	✗	✗	✗
CSDID	✓	✓	✗	✗
Imputation	✓	✓	✓	✗
FLEX	✓	✓	✓	✗
DID-INT	✓	✓	✓	✓

Note: ✓ = supports feature; ✗ = does not support feature

Table 1: Comparison of Estimators

Table 1 compares the features of the estimators we address in this paper. The TWFE estimator can use time-invariant covariates, but fails to identify the ATT in settings with staggered treatment rollout (Goodman-Bacon, 2021), time varying covariates (Caetano et al., 2022), and violations of the CCC. The CS-DID and the imputation estimators improve on TWFE by enabling researchers to estimate the ATT in settings with staggered treatment adoption. The CS-DID can use time-varying covariates, provided covariate exogeneity and covariate unconfoundedness assumptions hold. However, in the absence of these additional assumptions, CS-DID does not account for time-varying covariates. In contrast, the imputation estimator can incorporate time-varying covariates without additional assumptions, as long as they are unaffected by treatment. FLEX allows time-varying covariates to be used, but like TWFE and CS-DID, does not address violations of the CCC assumption. Through a simulation study, we show that the CS-DID, imputation estimator, and FLEX fails to identify the ATT under CCC violations. By comparison, DID-INT supports all four features, including CCC violations, but at a loss of efficiency.

3 Setup

In this section, we introduce the notation used throughout this paper. We consider a dataset with S regions and T periods, where some regions implement treatment at different times. Within each period $t \in \{1, 2, \cdots, T\}$, each region $s \in \{1, 2, \cdots, S\}$ can be classified into two broad categories: the treated group and the control group. The control group includes either **never-treated** regions, or both **never-treated** and **not-yet-treated** regions. Never-treated regions are regions which have not been treated throughout the available dataset, while not not-yet-treated regions have not yet implemented treatment in a period t. In this paper, we give researchers the flexibility to choose whether they want to use never-treated regions, or both never-treated and not-yet-treated regions as controls. For each period t, let D_t be a dummy variable equal to 1 for observations in treated regions, 0 otherwise.

$$D_t = \begin{cases} 1 & \text{if observation is treated in period } t. \\ 0 & \text{if observation is never treated or not-yet treated in period } t. \end{cases}$$

In a staggered design, it is possible for different regions to be treated in a given period t. Therefore, we define $\mathbf{s}^t = \{s \in S \mid D_t = 1\}$ as a **vector** of regions which are treated in period t. Similarly, $\mathbf{s}^c = \{s \in S \mid D_t = 0\}$ is a **vector** of regions which are either **never treated** or both **never treated and not-yet-treated** in period t, depending on what types of controls researchers want to use. Similar to Callaway and Sant'Anna (2021), we use the period right before a region s is treated as the pre-intervention period, which we index by a **scalar** $t^s - 1$.

To clearly present the notation used, consider the simple three region and three period case shown in Figure 2. In this example, the set of regions $S = \{A, B, C\}$ and the set of periods $T = \{1, 2, 3\}$. Region A receives treatment in period 2, and region B receives treatment in period 3. In period 2, Region B is categorized as a **not-yet-treated** region since the region receives treatment in a later period in the available data. Region C, which remains untreated throughout the dataset, is categorized as a **never-treated** region. The treatment indicator for period 2, D_2, is equal to 1 if observation is in region A, and 0 otherwise and the vector of treated regions $\mathbf{s}^2 = (A)$, since only A is treated in period 2. Similarly, in period 3, D_3 equal 1 for observations in region A and B, and 0 otherwise and $\mathbf{s}^3 = (A, B)$, since both A and B are treated in period 3.

4 Common Causality of Covariates

In this section, we formalize the common causality of covariates assumption, and distinguish between three types of CCC assumptions: the two-way CC, the region-invariant CCC, the time-invariant CCC, and the two-way CCC. Each assumption imposes different restrictions on the effect of the covariates across regions and periods. To clarify the CCC assumptions, we first
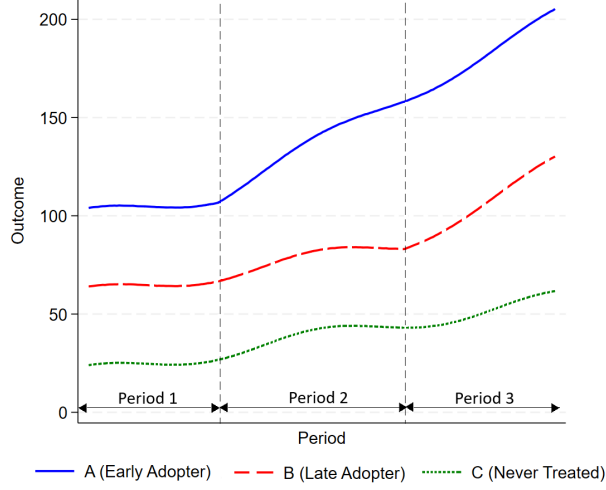
Figure 2: A three region, three period staggered design

present a model of untreated potential outcomes and the introduce the underlying assumptions made for this model. We then define the three CCC assumptions in details and illustrate how these assumptions may be violated in real-world datasets using an example.

Let the untreated potential outcome of an individual i in region s at period t be given by:

$$Y_{i,s,t}(0) = \sum_{s,t} \theta_{s,t}^{\circ} Z_{i,s,t} + \sum_{k=1}^{K} \gamma_{s,t}^{\circ,k} X_{i,s,t}^{k}(0) + \epsilon_{i,s,t} \tag{2}$$

Here, there are a total of K covariates, and $X_{i,s,t}^{k}(0)$ represents the kth untreated potential covariates. $Z_{i,s,t}$ are the group-time indicator variables. The parameters $\theta_{s,t}^{\circ}$ are the coefficients associated with these group-time indicators, and $\gamma_{s,t}^{\circ,k}$ represents the true effect of the kth covariate on the untreated potential outcome. The model for the untreated potential outcome in Equation (2) imposes two key assumptions. $\epsilon_{i,s,t}$ are the error terms which are **independently and identically distributed**.

Assumption 1 (Random Sampling). The data is a random sample from the population, conditional on covariates and are independently and identically distributed.

$$\{Y_{i,s,t}, D_i, X_{i,s,t}\}_{i=1}^{N}. \tag{3}$$

Assumption 2 (Linearity). The untreated potential outcome $Y_{i,s,t}(0)$ can be expressed as a linear combination of covariates $X_{i,s,t}(0)$.

Assumption 3 (Covariate Exogeneity). Participating in treatment does not change the distribution of covariates in the treated group (Caetano et al., 2022).

$$(X_t(0)|D=1) \sim (X_t(1)|D=1) \tag{4}$$

8

Assumption 3 implies that $\mathbb{E}\left[X_t(0) \mid D = 1\right] = \mathbb{E}\left[X_t(1) \mid D = 1\right]$, and rules out **bad controls**, or controls which are affected by treatment. For simplicity, let us assume that there is a single covariate $X_{i,s,t}(0)$, and omit the (0) since the distribution of covariate is unaffected by treatment due to assumption 3. Therefore, we re-write equation (2) as:

$$Y_{i,s,t}(0) = \sum_{s,t} \theta_{s,t}^{\circ} Z_{i,s,t} + \gamma_{s,t}^{\circ} X_{i,s,t} + \epsilon_{i,s,t} \tag{5}$$

Similarly, the model for $Y(1)_{i,s,t}$ is:

$$Y_{i,s,t}(1) = \tau_{i,s,t} + \sum_{s,t} \theta_{s,t}^{\circ} Z_{i,s,t} + \gamma_{s,t}^{\circ} X_{i,s,t} + \epsilon_{i,s,t} \tag{6}$$

Here, $\tau_{i,s,t}$ is the additive treatment effect, and is the parameter of interest. The **Average Treatment Effect of the Treated (ATT)** in terms of potential outcomes is given by:

$$ATT := \mathbb{E}\left[Y_i(1) - Y_i(0)|D = 1\right] = \mathbb{E}\left[\tau_{i,s,t}|D = 1\right] \tag{7}$$

Here, $Y_i(1)$ represents the treated potential outcome, and $Y_i(0)$ represents the untreated potential outcome of individual i. To further streamline the analysis, we adopt the homogeneous treatment effects assumption, formally introduced below. While not required for identification, this assumption is imposed solely to simplify notations. Under Assumption (4), the causal parameter of interest simplifies to τ.

Assumption 4 (Homogeneous treatment effect)**.** All treated units have the same treatment effect across both time and individuals.

$$\begin{aligned}
&\left[E[Y_{i,s,t}(1)|D = 1] - E[Y_{i,s,t}(0)|D = 1]\right] \\
&= \left[E[Y_{i,s',t}(1)|D = 1] - E[Y_{i,s',t}(0)|D = 1]\right] a.s. \forall s; s \neq s'
\end{aligned} \tag{8}$$

In DiD analyses, researchers include covariates to ensure that parallel trends are more plausible, discussed in more details in Section 6. Despite the recommendations of the DID literature, empirical researchers still continue to use covariates, particularly time varying covariates, without considering the assumptions required to justify their inclusion. In this paper, we explicitly introduce yet another assumption researchers need to consider when including covariates in DiD, which we call the **common causality of covariates (CCC)**. While all methods mentioned in Section 2 implicitly assume that the effect of the covariates are stable across regions and time, this assumption is not explicitly stated or analyzed directly.

Assumption 5 (Two-way Common Causality of Covariate)**.** The effect of the covariate is equal

between regions and across all periods.

$$\gamma^{\circ}_{s,t} = \gamma^{\circ}_{s',t'}.$$

Under Assumption 5, the true coefficient of $X_{i,s,t}$ on the untreated potential outcome is assumed to be the same between regions and across periods. Therefore, we can re-write Equation (5) with a common γ°, as shown in Equation (9).

$$Y_{i,s,t}(0) = \sum_{s,t} \theta^{\circ}_{s,t} Z_{i,s,t} + \gamma^{\circ} X_{i,s,t} + \epsilon_{i,s,t} \tag{9}$$

Assumption 6 (Region-invariant Common Causality of Covariate)**.** The effect of the covariate is equal between regions, but can vary across periods.

$$\gamma^{\circ}_{s,t} = \gamma^{\circ}_{s',t}$$
$$\gamma^{\circ}_{s,t} \neq \gamma^{\circ}_{s,t'}$$

Assumption 7 (Time-invariant Common Causality of Covariate)**.** The effect of the covariate is equal across all periods, but can vary between regions.

$$\gamma^{\circ}_{s,t} = \gamma^{\circ}_{s,t'}$$
$$\gamma^{\circ}_{s,t} \neq \gamma^{\circ}_{s',t}$$

In a simple setting with two periods (1 and 2) and two regions (A and B), the **two-way CCC** implies $\gamma^{\circ}_{A,1} = \gamma^{\circ}_{A,2} = \gamma^{\circ}_{B,1} = \gamma^{\circ}_{B,2} = \gamma$. The **region-invariant CCC** implies $\gamma^{\circ}_{A,1} = \gamma^{\circ}_{B,1}$ and $\gamma^{\circ}_{A,2} = \gamma^{\circ}_{B,2}$, and the **time-invariant CCC** implies $\gamma^{\circ}_{A,1} = \gamma^{\circ}_{A,2}$ and $\gamma^{\circ}_{B,1} = \gamma^{\circ}_{B,2}$. The two-way CCC assumption is more restrictive compared to Assumptions (6) and (7), requiring that the effect of the covariates are the same across both regions and time. When the two-way CCC assumption holds, both the region-invariant and time-invariant CCC assumptions holds as well. However, if the two-way CCC is violated, either the region-invariant CCC, or the time-invariant CCC, or both may be violated.

To better understand the CCC, consider an example where we want to estimate the ATT of a region-level policy on wage outcomes. Given the well-known link between education and wages (Mincer, 1958), researchers may want to include a dummy variable for education (BA or higher) as a control. Since the 1990s, the number of college graduates has increased (Denning et al., 2022). This rise has lead to a decline in the relative value of a college degree, especially in jobs that require less cognitive effort (Horowitz, 2018). As a result, the coefficient of the dummy variable may decrease over time, because of a shift in the educational attainment levels. This scenario suggests that the time-invariant CCC is unlikely to hold. Additionally, the coefficient of the dummy variable may be higher in regions with historically better higher education policies

and higher enrollment rates (Fortin, 2006). Regions with more capital-intensive industries may also show a higher coefficient compared to regions that rely more on hospitality, education, and health industries (Card et al., 2024). This indicates that the region-invariant CCC is unlikely to hold. If we analyze the data for multiple regions and periods together, the two-way CCC may not hold. The two-way CCC is difficult to verify empirically, since the two-way CCC is an assumption on the unobserved untreated potential outcome.

4.1 Motivating Examples

To motivate the practical importance of the CCC assumption we consider a number of empirically relevant situations. Specifically, we consider examples where the coefficient is not the same across regions, across time, and across both regions and time.

4.1.1 Region Varying Coefficients

This example comes from Bray (2025). This paper examines the impact of water use restrictions on water use following a water main break in the City of Calgary. The restrictions were in place in the summer, and primarily applied to outdoor water use. The author wished to use urban neighbourhoods as the control group, and suburban neighbourhoods as the treated group. This was motivated by the fact that the majority of outdoor water use is for watering lawns, and those are concentrated in the surburban neighbourhoods.

Figure 3a shows the (unconditional) trends for the treated and control groups. In this case, the pre-trends are clearly not parallel, as the seasonality is far more pronounced in the treatment group than in the control group. Given this, one might be tempted to model the seasonality, assuming a common covariate, by including month-of-year dummies. Figure 3b shows the conditional trends of water use, by treatment and control, after residualizing out the monthly dummies. In this figure the pre-trends are clearly not parallel either. In fact, the residualizing based on assuming common covariates has made things worse. The estimates of the common monthly coefficients are an over-estimate of the seasonality in the control group and an under-estimate in the treated groups. Accordingly, the residualized series show opposite seasonality across the treated and control groups, contrary to the raw data. At this point, many researchers would be tempted to conclude that CPT is likely implausible, given that the conditional pre-trends are not parallel.

Instead, we could relax the CCC assumption and estimate separate month-of-year dummies for both the treated and control groups. Figure 3c plots these residuals. The pre-trends in this figure, while not perfect, suggest that IPT assumption is plausible. This example highlights how the new IPT example can expand the scope of settings in which valid ATTs can be estimated.
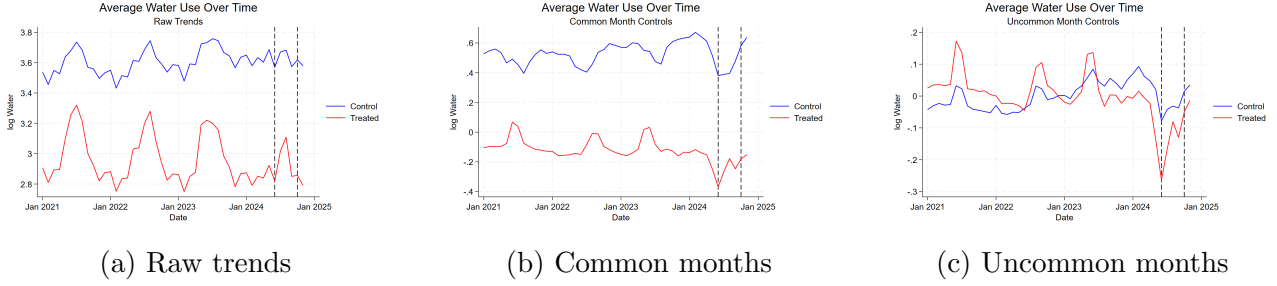
11

(a) Raw trends   (b) Common months   (c) Uncommon months

Figure 3: Comparison of raw and adjusted trends.

4.1.2 Time Varying Covariates or Coefficients

Consider a hypothetical intervention to reduce deaths from heart attacks. Further suppose we observe individuals over the period 2010–2020, where we have a relatively small panel of individuals in both a treated and a control group. Figure 4 shows the the relationship between heart attack deaths and age using Centre for Disease Control Data from 2018-2023. Given the small sample, and the importance of age on the prevalence of heart attacks, a researcher would likely want to add age as a control variable.

However, there are a number of problems with controlling for age in this example. The first is that age is a time-varying covariate. Conventionally, there are two ways around this, either to recode age as birth year, or to use (fixed) pre-treatment values. The birth-year approach changes a time-varying variable into a time-invariant one. However, this directly assumes that the the covariate is constant over time. For example, consider a person with a birth year of 1940. In 2010, that person is 70 and in 2020 that person is 80. Given the rapid increase in the hazard rate shown in Figure 4, the common coefficient assumption is clearly violated here.

The other approach, following Caetano et al. (2022) is to use pre-treatment values of covariates. Unfortunately, this approach suffers from the same problem. The 1940 birth year person now gets a value for "age" of 70 in both 2010 and 2020. The estimate of the coefficient would then be an average of the relationship between age and heart attacks over the entire period, rather than the age-by-age relationship found in the data. Again, the figure suggests that assuming a common coefficient for age (in 2010) on heart attacks across the period of 2010 to 2020 is likely to be violated.

4.1.3 Region and Time Varying Covariates

Consider an example where we want to estimate the ATT of a region-level policy on wage outcomes. Given the well-known link between education and wages (Mincer, 1958), researchers may want to include a dummy variable for education (BA or higher) as a control. Since the 1990s, the number of college graduates has increased (Denning et al., 2022). This rise has lead to a decline in the relative value of a college degree, especially in jobs that require less cognitive effort (Horowitz, 2018). As a result, the coefficient of the dummy variable may decrease over time, because of a shift in the educational attainment levels. This scenario suggests that the
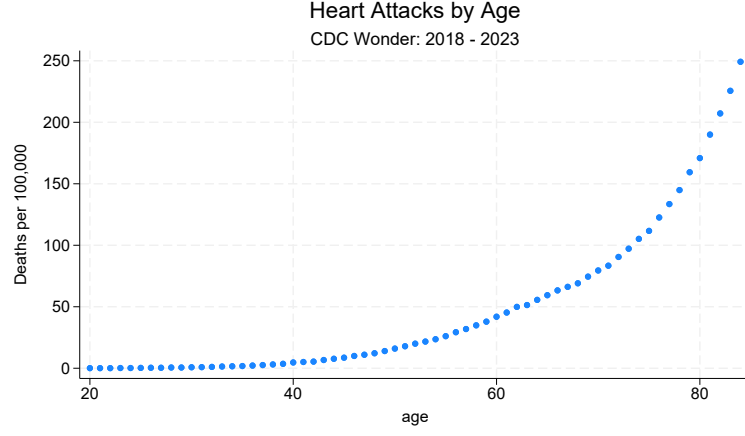
Figure 4: Heart attack deaths over time.

time-invariant CCC is unlikely to hold. Additionally, the coefficient of the dummy variable may be higher in regions with historically better higher education policies and higher enrollment rates (Fortin, 2006). Regions with more capital-intensive industries may also show a higher coefficient compared to regions that rely more on hospitality, education, and health industries (Card et al., 2024). This indicates that the region-invariant CCC is unlikely to hold. If we analyze the data for multiple regions and periods together, the two-way CCC may not hold.
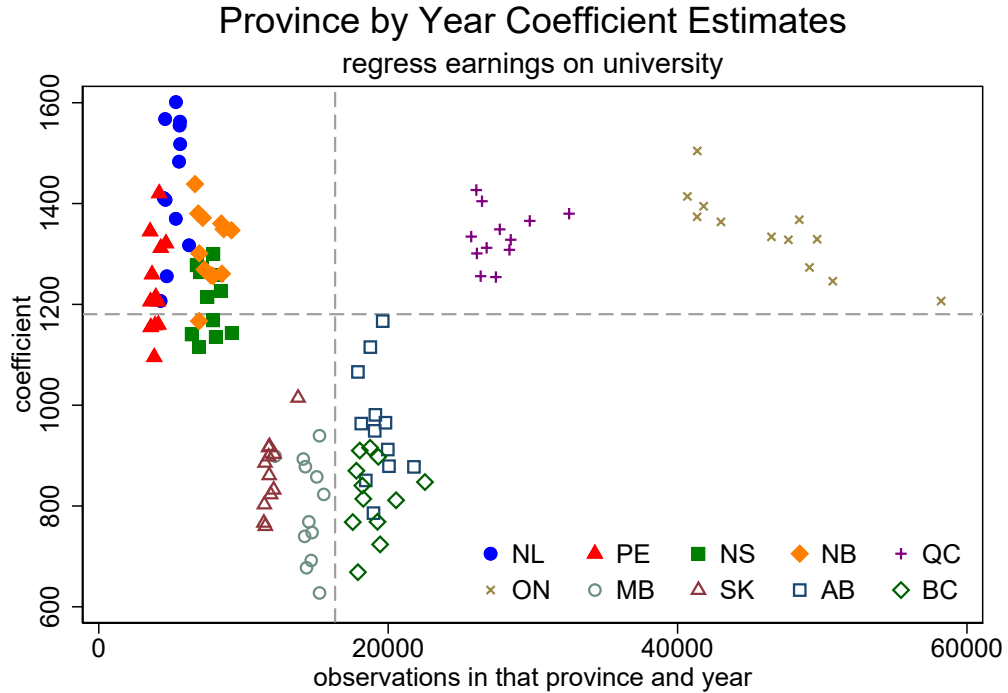
Figure 5: Region×Year coefficients of earnings on a dummy for university (BA or better)

To highlight that this assumption may be violated in actual datasets we consider a simple analysis using the Labour Force Survey (LFS) dataset for Canada. This survey is used to calculate the official unemployment rate in Canada and surveys 100,000 people per month, following individuals for 6 months in total. While we do not analyze an actual intervention

13

here, we could imagine someone wanting to estimate the ATT of a set of policies which were thought to effect earnings. Moreover, the anticipated ATT is small, so controlling for variation in earnings due to education is important. The sample is restricted to men between the ages of 35 to 55 who are employed from all 10 provinces between years 2008 to 2019. The final sample has 1,573,585 observations. In line with the discussion above, we estimate the following regression:

$$\text{earnings}_{i,p,t} = \alpha + \beta \text{college}_{i,p,t} + \epsilon_{i,p,t}. \tag{10}$$

Here, $\text{earnings}_{i,p,t}$ is the dollar earnings for person i in province p in year t, and $\text{college}_{i,p,t}$ is a dummy variable which takes on a value of one if the person has a college degree (BA or higher). We then re-estimate the model for each province and year separately. For each province and year pair, we record both the $\hat{\beta}_{p,t}$ coefficient estimate, but also the number of observations in that province×year pair that have a college degree.

The results are shown in Figure 6. This plot shows that there is considerable variation in the coefficient estimates between provinces. There are also considerable variations in the coefficients across years for some regions, like Ontario and Quebec. The average number of observations per cell is around 16,332, and there is considerable variation in the number of observations ranging from 3,549 to 58,215. However, even the smallest counts represent a fairly large sample. This suggests that the variation in the coefficients is not coming from small sample sizes either. Obviously, these are just estimates of the coefficients, and not the underlying causal parameters, but taken together this figure suggests that the assumption that the relationship between earnings and college being constant in all provinces and years is implausible.

Taken together, these three examples show why the assumptions of both region and/or invariant coefficients may be implausible.

5   Nature of Covariates

In this section, we distinguish between 5 types of covariates in DiD analysis, each based on the specific CCC assumption applied to them. We classify covariates for which the two-way CCC holds as **good controls**, the DAG for which is shown in Panel (a) of Figure (7). In other words, we assume $\gamma^\circ_{s,t} = \gamma^\circ_{s',t'}$, implying that the effect of the covariate is the same across all groups and time periods. If the covariate is truly "good" in the DGP, we can get consistent estimates of the ATT using TWFE and CS-DID, provided the assumptions for the respective estimators hold.

The second type of covariates, which we refer to as the **temporally shifting good controls gone bad**, include covariates that violate both the region-invariant and the time-invariant CCC (or the two-way CCC) assumptions. The DAG for this type of covariates is shown in Panel (b) of Figure (7). In a simple case, with only two groups (A and B) and two periods (1 and 2), the effect of X on Y varies both across groups and over time. Therefore, $\gamma^\circ_{A,1} \neq \gamma^\circ_{A,2} \neq \gamma^\circ_{B,1} \neq$

Province by Year Coefficient Estimates
regress earnings on university
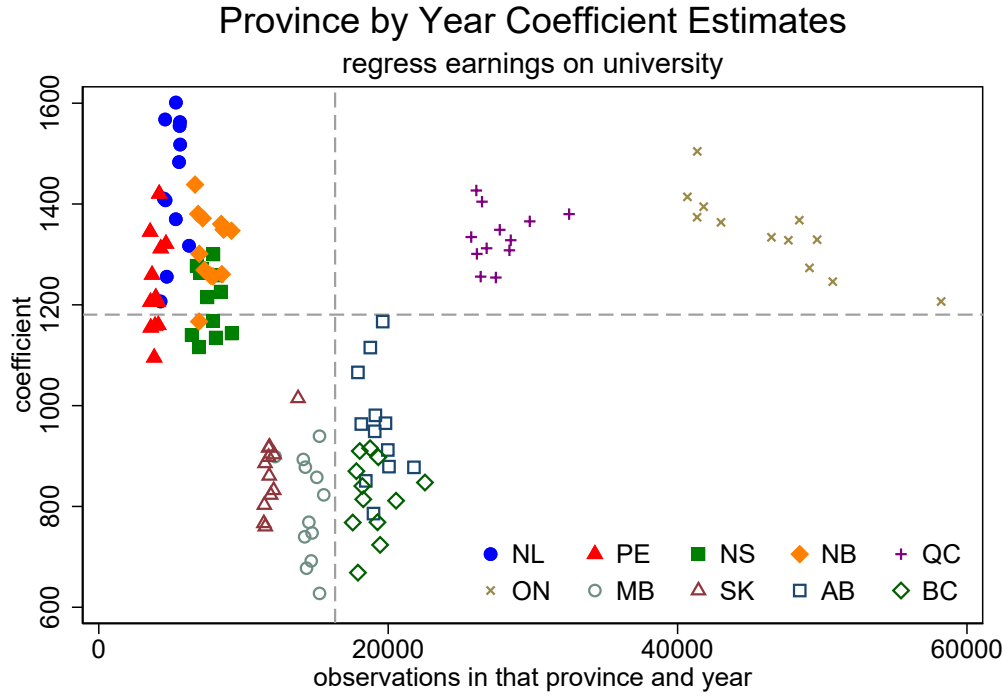
Figure 6: Region×Year coefficients of earnings on a dummy for university (BA or better)

(a) Good controls

(b) Temporally shifting good controls gone bad

(c) Good controls gone bad

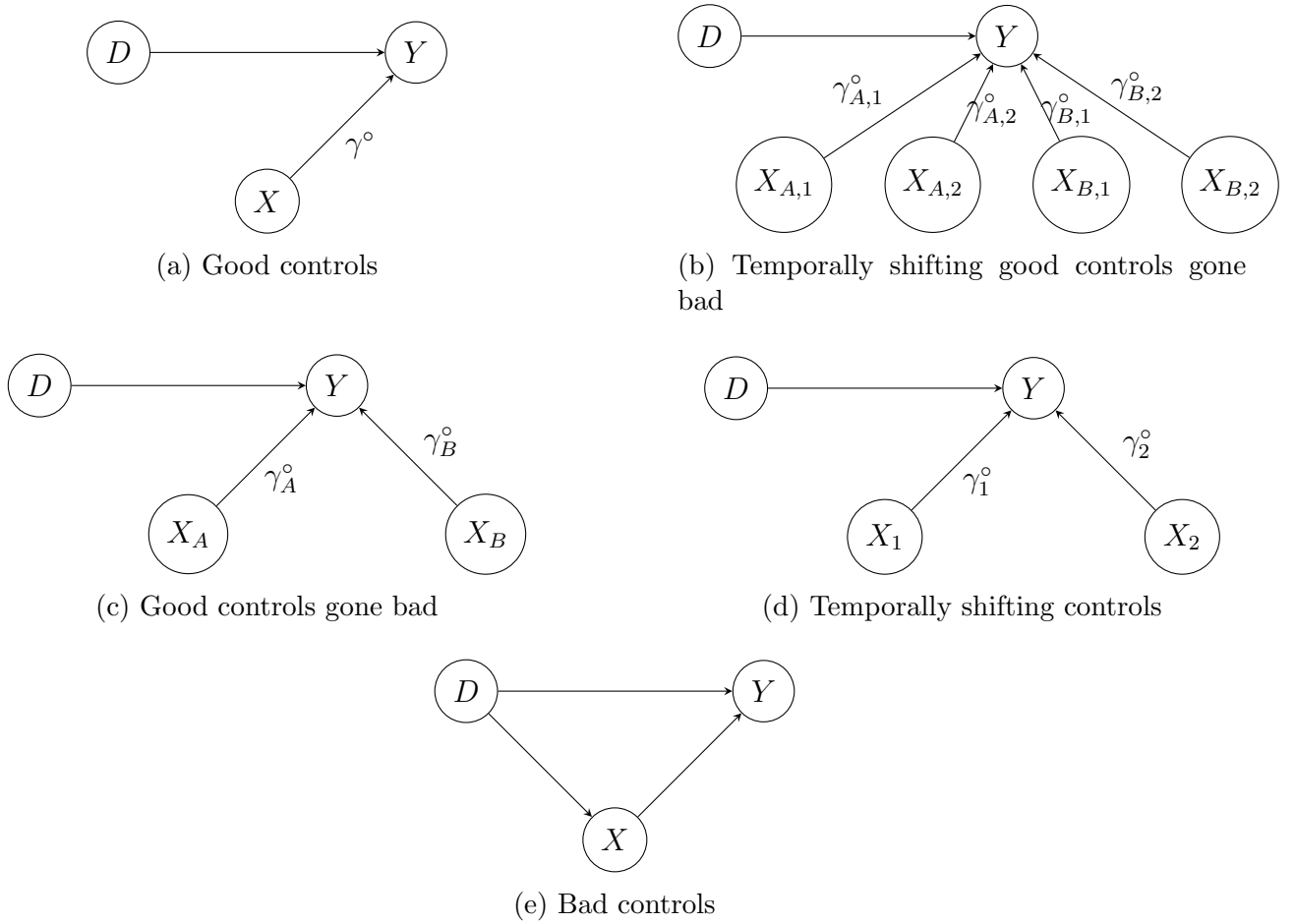(d) Temporally shifting controls

(e) Bad controls

Figure 7: Illustration of different types of covariates

$\gamma_{B,2}^{\circ}$.

The third type of covariates, which we refer to as **good controls gone bad**, are covariates for which the region-invariant CCC assumption is violated. The DAG for good controls gone bad is shown in Panel (c) of Figure 7. In a simple case where there are only two groups, A and B, the effect of X on Y is different for A compared to B. In other words, this violation occurs when $\gamma_A^{\circ} \neq \gamma_B^{\circ}$. However, the effect of the covariate remains the same across time.

The fourth classification, **temporally shifting controls**, refers to covariates that violate the time-invariant CCC assumption. The DAG for good controls gone temporal is shown in Panel (d) of Figure (7). In this case, the effect of the control variable X on Y is the same across groups but changes over time. Consider two distinct periods 1 and 2. If the relationship between X and Y differs between these periods while remaining the same for each group, we observe a violation of time-invariant CCC. In this case, $\gamma_1^{\circ} \neq \gamma_2^{\circ}$. The DAG for **bad controls** are shown in Panel (e) of Figure (7). Bad controls are controls which violate Assumption 3.

6 Parallel Trends

In this section, we provide a detailed discussion of the parallel trends assumption, a key identifying assumption in DID. We begin by deriving the necessary conditions required for both the standard and the conditional parallel trends to hold. We show that the *two-way common causality of covariates* assumption, while not strictly necessary for conditional parallel trends, strengthens its plausibility. In addition, we demonstrate why incorporating covariates that vary with time can complicate DID analysis, and derive the necessary conditions required for conditional parallel trends to hold with such covariates. Finally, we introduce an alternative to the parallel trends assumption which we call **Intersection Parallel Trends (IPT)**. The intersection parallel trends assumption remains plausible when the two-way CCC assumption is violated, and it requires fewer conditions to hold compared to conditional parallel trends.
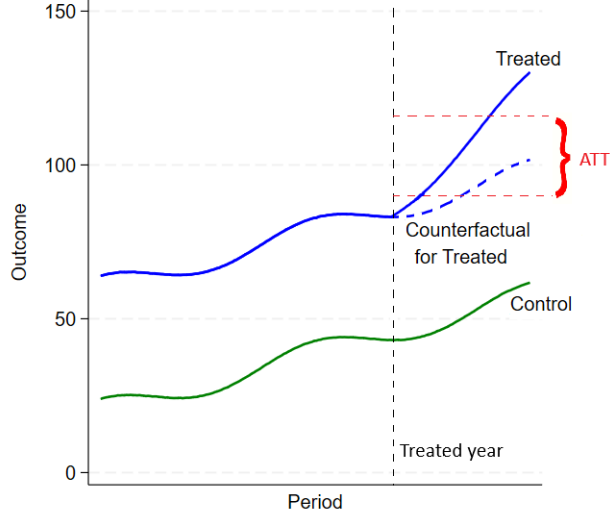
Figure 8: (Strong) Parallel Trends

To formally illustrate the importance of the parallel trends assumption, consider a simple 2×2 setting with two periods (1 and 2) and two regions (A and B), where B is treated in period 2. Let D be a dummy variable equal to 1 for observations in the treated group B, and 0 otherwise.

$$D = \begin{cases} 1 & \text{if observation is in region B (treated).} \\ 0 & \text{if observation is in region A (control).} \end{cases}$$

In Figure 8, the treated potential outcome is given by the solid blue line, which we observe. The untreated potential outcome is denoted by the dashed blue line, which we do not observe. Since the untreated potential outcome is unobserved, we cannot directly estimate the ATT using the sample analogues of Equation (7). As a result, we require a control group and pre-treatment data for both the treated and control groups. We also rely on the *parallel trends assumption*, which states that the treated and control groups would have followed the same trends in their outcomes in the absence of treatment. We use the trend of the control group to "impute" the counterfactual outcomes of the treated group, which allows us to estimate the ATT.

DID is commonly used to evaluate policies or interventions implemented in a region, where these policies are introduced in some regions (treated regions) and not in others (control regions). Regions could be provinces in Canada, states in the US, cities, municipalities, health authorities, etc. Since these regions may vary in demographic, economic, social and political characteristics, parallel trends without covariates is unlikely to hold. As a result, researchers use covariates to improve the plausibility of parallel trends, leading to the conditional parallel trends assumption (CPT). Under CPT, treated and control groups are assumed to follow similar trends in the absence of treatment, conditional on covariates.

6.1   Parallel Trends without covariates

In this subsection, we will derive the necessary conditions required for Parallel Trends to hold. For simplicity, we focus on the two region, two period setting, which we will continue to use in this and the following subsections. In the absence of covariates, when $\gamma_{s,t}^\circ = 0$, the parallel trends assumption in a 2×2 setup can be expressed as:

$$\mathbb{E}[Y_2(0) - Y_1(0) \mid s = A] = \mathbb{E}[Y_2(0) - Y_1(0) \mid s = B]. \tag{11}$$

Or equivalently,

$$\mathbb{E}[\Delta Y(0) \mid s = A] = \mathbb{E}[\Delta Y(0) \mid s = B]. \tag{12}$$

Theorem 1 (Necessary conditions for parallel trends without covariates).

$$\textit{Condition 1: } \theta_{A,2}^\circ - \theta_{A,1}^\circ = \theta_{B,2}^\circ - \theta_{B,1}^\circ. \tag{13}$$

The proof of Theorem 1 can be found in Appendix A. In this paper, we do not aim to simplify **Condition 1** further by imposing a particular functional form on $\theta_{s,t}^\circ$. However, to intuitively explain what Condition 1 means, let us assume a two-way fixed effects type functional form for $\theta_{s,t}^\circ$ shown in Equation (14).

$$\theta_{s,t}^\circ \equiv \alpha_s^\circ + \delta_t^\circ \tag{14}$$

Here, α_s° are the region fixed effect for region s, δ_t° are the period fixed effects for period t for region s. In the standard two-way fixed effects specification, these time effects are assumed to be common across regions since parallel trends hold. To demonstrate this formally, we allow δ_t° to vary across regions, denoting them as $\delta_{s,t}^\circ$. This allows us to recognize that different regions may experience different time-specific trends. Therefore, we can re-write Equation (14) as:

$$\theta_{s,t}^\circ \equiv \alpha_s^\circ + \delta_{s,t}^\circ \tag{15}$$

Plugging Equation (15) into Equation (13) for each period and time, we get:

$$\alpha_A^\circ + \delta_{A,2}^\circ - \alpha_A^\circ - \delta_{A,1}^\circ = \alpha_B^\circ + \delta_{B,2}^\circ - \alpha_B^\circ - \delta_{B,1}^\circ$$

$$\delta_{A,2}^\circ - \delta_{A,1}^\circ = \delta_{B,2}^\circ - \delta_{B,1}^\circ \tag{16}$$

In this simple case, condition 1 captures the **parallel trends in time specific shocks** as shown in Equation (16). Therefore, in a two-way fixed effects type structure for $\theta_{s,t}^\circ$, it is safe

to assume that $\delta_t^\circ = \delta_{s,t}^\circ$.

6.2 Conditional Parallel Trends

Due to differences in characteristics across regions, parallel trends without covariates are un-likely to hold. As a result, researchers include covariates to improve the credibility of parallel trends, resulting in the conditional parallel trends assumption. When covariates matter in parallel trends, $\gamma_{s,t}^\circ$ in Equation (5) is no longer assumed to be 0. Conditional Parallel Trends in a 2×2 setup can be expressed as:

$$
\begin{aligned}
&\mathbb{E}[Y_2(0) \mid s = A, X_{A,2}] - \mathbb{E}[Y_1(0) \mid s = A, X_{A,1}] \\
=&\mathbb{E}[Y_2(0) \mid s = B, X_{B,2}] - \mathbb{E}[Y_1(0) \mid s = B, X_{B,1}].
\end{aligned}
\tag{17}
$$

Theorem 2 (Necessary conditions for conditional parallel trends)**.**

$$
\textbf{\textit{Condition 1:}} \ \theta_{A,2}^\circ - \theta_{A,1}^\circ = \theta_{B,2}^\circ - \theta_{B,1}^\circ.
\tag{18}
$$

$$
\begin{aligned}
\textbf{\textit{Condition 2:}} \ &\gamma_{A,2}^\circ \mathbb{E}[X_{A,2}(0) \mid s = A] - \gamma_{A,1}^\circ \mathbb{E}[X_{A,1}(0) \mid s = A] \\
&= \gamma_{B,2}^\circ \mathbb{E}[X_{B,2}(0) \mid s = B] - \gamma_{B,1}^\circ \mathbb{E}[X_{B,1}(0) \mid s = B].
\end{aligned}
\tag{19}
$$

The proof of Theorem 2 can be found in Appendix B. Theorem 2 shows that although Conditions 1 and 2 are each necessary for conditional parallel trends, the assumption holds only when both conditions are satisfied jointly. Let us analyze condition 2 in more details. **Condition 2 can hold when two-way CCC is violated**, since condition 2 allows for the coefficients of the covariates to vary across regions and over time. However, this condition is hard to interpret and has **not been explicitly recognized** in the existing literature in its current form. In fact, the literature implicitly assume the two-way CCC, which simplifies condition 2 and makes it more interpretable. Under the two-way CCC, we can re-write condition 2 as:

$$
\begin{aligned}
&\gamma^\circ \mathbb{E}[X_{A,2}(0) \mid s = A] - \gamma^\circ \mathbb{E}[X_{A,1}(0) \mid s = A] \\
&= \gamma^\circ \mathbb{E}[X_{B,2}(0) \mid s = B] - \gamma^\circ \mathbb{E}[X_{B,1}(0) \mid s = B].
\end{aligned}
\tag{20}
$$

The above relies on the following **sufficient condition**:

$$
\textbf{Sufficient Condition 1:} \ \gamma_{A,1}^\circ = \gamma_{A,2}^\circ = \gamma_{B,1}^\circ = \gamma_{B,2}^\circ = \gamma.
\tag{21}
$$

Cancelling out the γ°, we get:

Assumption	Meaning	Sufficient Condtion 2?
Time-invariant covariates	$\Delta X_{s,t} = \Delta X_{s,t'}$	Yes
Invariant distributions of covariates	$\Delta X_{s,t} \sim \Delta X_{s',t}$	Yes
Varying distributions of covariates	$\Delta X_{s,t} \nsim \Delta X_{s',t}$	No

Table 2: Assumptions required for Sufficient Condition 2

$$\textbf{Sufficient Condition 2: } \mathbb{E}[X_{A,2}(0) \mid s = A] - \mathbb{E}[X_{A,1}(0) \mid s = A]$$
$$= \mathbb{E}[X_{B,2}(0) \mid s = B] - \mathbb{E}[X_{B,1}(0) \mid s = B]. \tag{22}$$

Equation (22) is well-known in the DID literature, and is called **parallel trends of untreated potential covariates**. It states that, in order for conditional parallel trends to hold, we require the untreated potential covariates to evolve in parallel to each other as well. Table 2 summarizes the different assumptions required for this to hold. First, it holds when **the covariates are time-invariant**, meaning that covariate values do not change across periods. This case is well recognized in the literature and has been recommended in many early DID studies (Caetano et al., 2022). The second assumption, called the **invariant distribution of covariates**, allows covariates to vary over time as long as their distributions remain stable across regions and periods. Finally, **the varying distribution of covariates** assumption allows the distribution of covariates to change over time and across regions. While this assumption is the most flexible and realistic, sufficient condition 2 fails to hold in this setting.

To summarize, in this sub-section we derive two necessary conditions required for conditional parallel trends to hold. The two conditions are shown in Equation (21) and (22). Since the second necessary condition is difficult to interpret, we further simplify it by identifying two sufficient conditions for conditional parallel trends to hold: the two-way CCC and the parallel trends of untreated potential covariates. The latter condition is interpretable and is well-known in literature. We demonstrate that both the two-way CCC and the parallel trends of untreated potential covariates **must hold jointly** for CPT to be satisfied. However, it is important to note that these two-way CCC and the conditional trends of covariates are sufficient, but not necessary, conditions. Condition 2 may still hold even if the two sufficient conditions fail, purely by coincidence.

6.3 Intersection Parallel Trends (IPT)

In this subsection, we introduce *Intersection Parallel Trends (IPT)*, which allows violations of the two-way CCC, and requires only the first necessary condition to hold. Since the second necessary condition required for conditional parallel trends is no longer imposed, researchers are not required to justify additional assumptions regarding either the covariates or their coefficients. We do not aim to establish whether IPT is stronger or weaker than CPT; rather, our objective is to show that IPT holds under fewer necessary conditions compared to the CPT. In a 2×2 setup, IPT is defined as:

$$
\begin{aligned}
&\mathbb{E}[\tilde{Y}_2(0) \mid s = A, X_{A,2}] - \mathbb{E}[\tilde{Y}_1(0) \mid s = A, X_{A,1}] \\
=&\mathbb{E}[\tilde{Y}_2(0) \mid s = B, X_{B,2}] - \mathbb{E}[\tilde{Y}_1(0) \mid s = B, X_{B,1}].
\end{aligned}
\tag{23}
$$

where $\tilde{Y}_{i,s,t}(0)$ denotes the residuals from regressing $Y_{i,s,t}(0)$ on the covariates:

$$
\tilde{Y}_{i,s,t}(0) = Y_{i,s,t}(0) - \gamma_{s,t}^{\circ} X_{i,s,t}(0).
\tag{24}
$$

IPT requires that, after residualizing untreated potential outcomes on covariates, the resulting trends are the same between treated and control groups.

Theorem 3 (Necessary conditions for intersection parallel trends)**.**

$$
\textit{\textbf{Condition 1:}} \ \theta_{A,2}^{\circ} - \theta_{A,1}^{\circ} = \theta_{B,2}^{\circ} - \theta_{B,1}^{\circ}.
\tag{25}
$$

The proof of Theorem 3 can be found in Appendix C. Both conditional parallel trends and intersection parallel trends are assumptions about the untreated potential outcomes, and therefore cannot be directly tested. In practice, researchers rely on pre-trend tests as a plausibility check.

From the perspective of a theorist, the difference between conditional parallel trends when the two-way is violated and intersection parallel trends may not be apparent. **If the conditional parallel trends holds under necessary condition 2, the intersection parallel trends will hold as well.** The difference becomes more prominent when we think of it in terms of how applied researchers **check for the plausibility of parallel trends**. Since the conditional parallel trends is an assumption on the untreated potential outcome which we do not observe, researchers examine pre-intervention trends between the treated and the control groups to assess the plausibility of parallel trends after intervention. With covariates, researchers typically regress the observed outcomes on the covariates in the pre-intervention period assuming a common coefficient, and plot the residuals of these outcomes for each region on a common figure. The residuals are shown in Equation (26).

$$\tilde{Y}_{i,s,t} = Y_{i,s,t} - \widehat{\gamma} X_{i,s,t}. \tag{26}$$

In the pre-intervention period, $\tilde{Y}_{i,s,t}(0) = \tilde{Y}_{i,s,t}$, since we observe the untreated potential outcomes for the treated group. Comparing Equations (24) and (26), $\widehat{\gamma}$ is not a consistent estimate of $\gamma_{s,t}^{\circ}$, since in the former, the true effect of the covariates on outcome varies based on region and period. As a result, we may obtain pre-trends similar to Panel (a) in Figure (1), where conditional parallel trends does not seem plausible, although in truth it holds. This may cause researchers to abandon the project altogether since they fail to find evidence in favour of conditional parallel trends. In Appendix E, we further show that imposing a common coefficient when the true effects of covariates vary across regions or periods can lead to inconsistent estimates of the ATT using two-way fixed effects.

This limitation highlights the practical advantage of intersection parallel trends. The intersection parallel trends examines the trend of the residualized untreated potential outcomes where the effect of the covariates are partialled out. Therefore, it no longer imposes a common coefficient on covariates. To check for the plausibility of **intersection parallel trends** in a two-region and two-period setting, researchers can run **separate** regressions for each region and period combinations, and plot the residuals. In other words, researchers will regress the outcomes on covariates for region A in period 1, region A in period 2, region B in period 1 and region B in period 2 separately, and store the residuals for each of these regressions. The residuals of each regression are shown in Equation (27), where $\widehat{\gamma}_{s,t}$ can consistently estimate $\gamma_{s,t}^{\circ}$. They can them plot the residuals for each of these outcomes for each region on a common figure, similar to Panel (b) of Figure (26). This allows researchers to account for the differing effects of the covariates on outcomes for each region and period.

$$\tilde{Y}_{i,s,t} = Y_{i,s,t} - \widehat{\gamma}_{s,t} X_{i,s,t}. \tag{27}$$

A simpler approach that requires fewer regressions is to estimate a single model where the covariates are interacted with region and time indicators, and then plot the residuals from this regression on a common figure. The regression is shown in Equation (28), and the residuals will be the same as the residuals shown in Equation (27), since $\widehat{\gamma}_{s,t}$ can consistently estimate $\gamma_{s,t}^{\circ}$ for each region and time.

$$Y_{i,s,t} = \gamma_0 + \gamma_{s,t} I(s) I(t) X_{i,s,t} + \eta_{s,t}. \tag{28}$$

Checking for the plausibility of parallel trends in this way allows researchers to model covariates flexibly, allowing their coefficients to vary across regions and time. In Panel (b) of Figure 1, plotting the residuals of the outcome variable regressed on flexible versions of the covariates can yield a plot where parallel trends seem plausible. This allows researchers to **recover previously**

hidden parallel trends, which were previously hidden because researchers plotted trends after residualizing outcomes using a model that imposed a common coefficient on the covariates. By instead allowing the covariate effects to vary across regions and periods, the intersection parallel trends approach corrects this issue and can uncover the true underlying parallel trends.

The intersection parallel trends also removes the need for condition 2 in theory 2 to hold, which depends on assumptions about the covariates themselves. Although flexible modelling of covariates is easy to incorporate in the two-way fixed effects model, it becomes challenging to do when we introduce propensity score re-weighting in DID, such as the semi-parametric DID estimator by Abadie (2005) and the doubly robust DID estimator by Sant'Anna and Zhao (2020). In the subsequent sections, we show that many DID estimators which rely on propensity scores do not permit this flexibility. Therefore, these estimators assumes conditional parallel trends holds such that both **sufficient conditions 1 and 2 hold**. In other words, these models implicitly assume that **the two-way CCC holds**, although CPT may hold when this assumption is violated. This can lead to inconsistent estimates when the two-way CCC is violated in truth. We also show that a "modified" two-way fixed effects model can allow flexible modeling of covariates. However, this modified model is not robust to forbidden comparisons and negative weighting in staggered adoption settings. Consequently, this leaves researchers with an important research gap: how can we estimate the ATT when the two-way CCC is violated. Addressing this question motivates the development of Intersection Difference-in-differences introduced in the next section.

For the remainder of the paper, we focus on staggered treatment designs with S regions and T periods. Therefore, we restate the parallel trends, conditional parallel trends , and intersection parallel trends assumptions in this more general setting, as shown below.

7   Multiple Groups and Staggered Adoption

In the previous section, we have focused on a simple 2×2 case to derive the necessary and sufficient conditions for parallel trends, conditional parallel trends and intersection parallel trends. For the remainder of the paper, we focus on more general settings with S regions and T periods. Therefore, we restate the parallel trends, conditional parallel trends and intersection parallel trends assumptions in this more general setting, as shown below. For simplicity of notation, we continue to assume that there is a single time-varying covariate $X_{i,s,t}$.

Assumption 8 (Parallel Trends)**.** The treated and control groups would have followed the same trends in their outcomes in the absence of treatment.

$$\mathbb{E}[Y_t(0) - Y_{t^s-1}(0) \mid s \in \mathbf{s}^t] = \mathbb{E}[Y_t(0) - Y_{t^s-1}(0) \mid s \in \mathbf{s}^c]. \tag{29}$$

Assumption 9 (Conditional Parallel Trends)**.** The treated and control groups would have fol-

lowed the same trends in their outcomes in the absence of treatment, conditional on covariates.

$$
\begin{aligned}
&\mathbb{E}[Y_t(0) \mid s \in \mathbf{s}^t, X_{s,t}] - \mathbb{E}[Y_{t^s-1}(0) \mid s \in \mathbf{s}^t, X_{s,t^s-1}] \\
&= \mathbb{E}[Y_t(0) \mid s' \in \mathbf{s}^t, X_{s',t}] - \mathbb{E}[Y_{t^s-1}(0) \mid s' \in \mathbf{s}^c, X_{s',t^s-1}].
\end{aligned}
\tag{30}
$$

Assumption 10 (Intersection Parallel Trends)**.** The treated and control groups would have followed the same trends in their residualized outcomes in the absence of treatment, conditional on covariates.

$$
\begin{aligned}
&\mathbb{E}[\tilde{Y}_t(0) \mid s \in \mathbf{s}^t, X_{s,t}] - \mathbb{E}[\tilde{Y}_{t^s-1}(0) \mid s \in \mathbf{s}^t, X_{s,t^s-1}] \\
&= \mathbb{E}[\tilde{Y}_t(0) \mid s' \in \mathbf{s}^t, X_{s',t}] - \mathbb{E}[\tilde{Y}_{t^s-1}(0) \mid s' \in \mathbf{s}^t, X_{s',t^s-1}].
\end{aligned}
\tag{31}
$$

In order to write the ATT conditional on $X_{i,s,t}$ shown in terms of **observed outcomes**, we also require the **no anticipation** assumption, which implies that treated units do not change behavior before treatment occurs (Abadie, 2005; De Chaisemartin and d'Haultfoeuille, 2020a). Violation of no anticipation can lead to deviations in parallel trends in periods right before treatment. We also require Assumption **overlap**, which states that for each treated unit, there exists untreated units with the same covariate values.

Assumption 11 (No anticipation)**.** The treated potential outcome is equal to the untreated potential outcome for all units in the treated group in the pre-intervention period.

$$
Y_{i,s,t}(1) = Y_{i,s,t}(0) \ \ \forall i \quad a.s. \ \text{for all} \ \ t < t^s.
\tag{32}
$$

Assumption 12 (Overlap)**.** There exists some ϵ such that $P(D = 1) > \epsilon$ and $P(D = 1 | X_{i,s,t}) < 1 - \epsilon$.

With covariates, and using assumptions random sampling, linearity, covariate exogeneity, conditional parallel trends or intersection parallel trends, no anticipation and overlap, the conditional ATT can be written in the form shown in Equation (33) (Abadie, 2005; Heckman et al., 1997; Bertrand et al., 2004; Callaway and Sant'Anna, 2021; Caetano et al., 2022; Caetano and Callaway, 2024; Goodman-Bacon, 2021). In common treatment adoption settings, $t^s - 1$ includes the entire span of the pre-intervention period.

$$
\begin{aligned}
ATT_X &= \mathbb{E}[Y_t(0) \mid s \in \mathbf{s}^t, X_{s,t}] - \mathbb{E}[Y_{t^s-1}(0) \mid s \in \mathbf{s}^t, X_{s,t^s-1}] \\
&= \mathbb{E}[Y_t(0) \mid s' \in \mathbf{s}^t, X_{s',t}] - \mathbb{E}[Y_{t^s-1}(0) \mid s' \in \mathbf{s}^c, X_{s',t^s-1}].
\end{aligned}
\tag{33}
$$

In staggered treatment adoption designs, where different regions can implement treatment at different periods, the two-way fixed effects estimator is known be be inconsistent under heterogeneous treatment effects due to forbidden comparisons and negative weighting issues Goodman-Bacon (2021). To address the problem of "forbidden comparisons". Callaway and Sant'Anna (2021) proposes a framework that decomposes staggered adoption designs into a

series of simpler "2×2" comparisons, which we call *stacks*. Each stack isolates one treated cohort and a appropriate control cohort, which can either be never treated or not yet treated. In the absence of covariates, conditional ATT for each stack, called $ATT_X(s,t)$ is written as:

$$
\begin{aligned}
ATT_X(s,t) &= \mathbb{E}[Y_t(0) \mid s \in \mathbf{s}^t, X_{s,t}] - \mathbb{E}[Y_{t^s-1}(0) \mid s \in \mathbf{s}^t, X_{s,t^s-1}] \\
&= \mathbb{E}[Y_t(0) \mid s' \in \mathbf{s}^t, X_{s',t}] - \mathbb{E}[Y_{t^s-1}(0) \mid s' \in \mathbf{s}^c, X_{s',t^s-1}].
\end{aligned}
\tag{34}
$$

Here, $t^s - 1$ includes only the period right before cohort s is treated in staggered treatment adoption settings (Callaway and Sant'Anna, 2021). In this paper, we use regions to define a cohorts rather than pooling regions by treatment timing, similar to (Callaway and Sant'Anna, 2021). Benefits of this approach is explored in more detail in Karim et al. (2025a).

8   Intersection Difference-in-differences (DID-INT)

In this section, we introduce a new estimator called the Intersection Difference-in-Differences (DID-INT), which can provide a consistent estimate of the ATT, and is robust to the four types of CCC violations we discuss in this paper (described in details in Section 4). The ATT is estimated in five steps. In the first step, we employ a **model selection algorithm** to plot a sequence of parallel trends figure, which is used to determine the functional form of covariates $f(X_{i,s,t}^k)$. Details of this algorithm is shown in Section 8.2. In the second step, we propose running the following regression without a constant:

$$
Y_{i,s,t} = \sum_s \sum_t \lambda_{s,t} I(s,t) + f(X_{i,s,t}) + \epsilon_{i,s,t},
\tag{35}
$$

where, $I(s,t)$ is a dummy variable that takes a value of 1 if the observation is in region s in period t, or the region×time intersection, hence the name. $f(X_{i,s,t})$ represents a function of covariates, which varies according to the specific CCC violations researchers intend to account for in their analysis. Depending on the function of $f(X_{i,s,t})$, we also generate two types of dummy variables: $I(s)$ which takes on a value of 1 if the observation is in region s; and $I(t)$ which takes on a value of 1 if the observation is from period t.

In the third step, we store the differences in $\lambda_{s,t}$ in a matrix for each period after region s is first treated, using the period right before treatment ($t^s - 1$) as the pre-intervention period.

$$
\widehat{\Delta}_s = (\widehat{\lambda}_{s,t} - \widehat{\lambda}_{s,t^{-s}}).
\tag{36}
$$

In the fourth step, we estimate the ATT for group s in period t, denoted by $\widehat{\theta}_{s,t}$ as follows:

$$
\widehat{\theta}_{s,t} = \Delta_s - \Delta_{s'}.
\tag{37}
$$

Here, s' is a relevant control group for region s, and t^s the period when region s is first treated.

These are drawn from the matrix of differences in the third step. In the fifth step, we estimate the overall ATT by taking a weighted average of the $\widehat{\theta}_{s,t}$'s estimated in the second step. The expression of the overall ATT is:

$$\widehat{\theta} = \sum_{s=2}^{S} \sum_{t=2}^{\mathcal{T}} 1\{t^s \leq t\} w_{s,t} \widehat{\theta}_{s,t}. \tag{38}$$

In the above expression, the forbidden comparisons highlighted by Goodman-Bacon (2021) are excluded from the calculation. Cluster robust inference on the ATT can be done on the ATT using a cluster jackknife and randomization inference. See Karim et al. (2024) for details, which uses the cluster jackknife and randomization inference for a similar multi-step DiD estimator designed for unpoolable data.

Theorem 4a (DID-INT under Assumptions 4, 9, 11 and 12). *The causal parameter of interest τ is identified under conditional parallel trends, homogeneous treatment effects, no anticipation and overlap assumptions for each stack.*

$$\theta_{s,t} = \tau. \tag{39}$$

Theorem 4b (DID-INT under Assumptions 4, 10, 11 and 12). *The causal parameter of interest τ is identified under intersection parallel trends, homogeneous treatment effects, no anticipation and overlap assumptions for each stack.*

$$\theta_{s,t} = \tau. \tag{40}$$

A proof of the above Theorem is shown in Appendix D.

8.1 Additional Implementation Details and Software

When there is more than one control region, the fourth step of the DID-INT estimator is easiest to do in the form of a regression. This allows for $\widehat{\Delta}_{(s)}$ terms from multiple regions to contribute to the estimate of $\widehat{\theta}_{s,t}$. This regression based approach is identical to the one proposed in Karim et al. (2024), except for how the $\widehat{\Delta}_{(s,t)}$ are calculated, see Section 5.2 of that paper for details. To ease in the implementation of the DID-INT estimator we have a package available in `Julia` which is available at https://github.com/ebjamieson97/DiDInt.jl. We also have a wrapper for `Stata` which calls the `Julia` program to perform the calculations, using the approach in Roodman (2025). The `Stata` program is available at https://github.com/ebjamieson97/didintjl. A wrapper in `R` is forthcoming. The software package allows for cluster robust inference using both a cluster-jackknife and randomization inference. The details of these routines, and their finite sample performance is discussed in the companion paper Karim et al. (2025b).

8.2 Variants of the DID-INT Estimator and a Model Selection Algorithm

In this section, we will explore the four distinct ways to model covariates in DID-INT, depending on the type of CCC violations researchers want to account for. When the two-way CCC seems plausible, we recommend modeling the covariates as $f(X_{i,s,t}) = \gamma X_{i,s,t}$. This version of DID-INT will be referred to as the **homogeneous DID-INT**. If the time-invariant CCC assumption is plausible but the region-invariant CCC is not, we recommend researchers to interact the covariates with the $I(s)$ dummies and include the interacted terms as covariates in the model. Therefore, $f(X_{i,s,t}) = \sum_{s=1}^{S} \gamma_s I(s) X_{i,s,t}$, which adjusts for potential violations of the region-invariant CCC. This approach is referred to as the **region-varying DID-INT**. The third approach, referred to as the **time-varying DID-INT**, accounts for plausible time-invariant CCC violations when the region-invariant CCC assumption is plausible. Potential violations in region-invariant CCC is accounted for by interacting the covariates with the $I(t)$ dummy variables. This implies: $f(X_{i,s,t}) = \sum_{t=1}^{T} \gamma_t I(t) X_{i,s,t}$. Lastly, the **two-way DID-INT** allows for two-way CCC violations, where $f(X_{i,s,t}) = \sum_{t=1}^{T} \sum_{s=1}^{S} \gamma_{s,t} I(s) I(t) X_{i,s,t}$. Here, the covariates are interacted with both the $I(s)$ and the $I(t)$ dummy variables and included as covariates in the model. Figure 9 provides a summary. Here A and B are two groups, 1 and 2 are two time periods. The true γ terms, $\gamma°$, are allowed to potentially vary either across groups, across periods, or across both groups and periods. In the next section, we will introduce the **model selection algorithm** to determine the appropriate functional form of covariates to use in DID-INT.

I - Homogeneous:

	A	B
1	γ^0	γ^0
2	γ^0	γ^0

$$f(X_{i,s,t}) = \gamma X_{i,s,t}$$

II - Region Variation:

	A	B
1	$\gamma_A°$	$\gamma_B°$
2	$\gamma_A°$	$\gamma_B°$

$$f(X_{i,s,t}) = \sum_{s=1}^{S^T} \gamma_s I(s) X_{i,s,t}$$

III - Year Variation:

	A	B
1	$\gamma_1°$	$\gamma_1°$
2	$\gamma_2°$	$\gamma_2°$

$$f(X_{i,s,t}) = \sum_{t=1}^{T} \gamma_t I(t) X_{i,s,t}$$

IV - Region & Year (Two-way):

	A	B
1	$\gamma_{A,1}°$	$\gamma_{B,1}°$
2	$\gamma_{A,2}°$	$\gamma_{B,2}°$

$$f(X_{i,s,t}) = \sum_{t=1}^{T} \sum_{s=1}^{S^T} \gamma_{s,t} I(s) I(t) X_{i,s,t}$$

Figure 9: Modeling Covariates in DID-INT

We have introduced four versions of DID-INT that researchers can employ based on the func-

tional form of the covariates, $f(X_{i,s,t})$. In practice, the true functional form is unobserved. In this section, we present a model selection algorithm designed to help researchers identify the correct functional form of the covariates. This, in turn, will guide the selection of the most appropriate version of DID-INT, as outlined earlier. The algorithm relies on plotting a sequence of parallel trends figures, conditional on the appropriate functional form of covariates, and stopping where pre-trends seem plausible through visual inspection. The algorithm is presented in Figure 10.

The model selection process begins by plotting the unconditional trends of outcomes without covariates for each group separately. If the unconditional pre-trends appear visually plausible, researchers may stop there and proceed to use DID-INT without covariates. Researchers may use other methods as well, since the assumptions needed to identify the ATT with covariates are no longer necessary. However, the TWFE should not be used for designs with staggered adoption, since its not robust to forbidden comparisons and negative weights with heterogeneous treatment effects (Goodman-Bacon, 2021).

If pre-trends do not seem plausible, the next step is to include covariates. Conventionally, this is done by regressing the outcome on covariates, storing residuals and plotting the trends of these residuals separately for each group. If these trends seem plausible, researchers should use the homogeneous DID-INT, or alternative methods since the CCC assumptions holds. However, the TWFE should still be avoided in this case. When pre-trends with covariates appear implausible, researchers typically refrain from proceeding with the ATT estimation, since this assumption is key to identifying the ATT. In this paper, we recommend researchers do go a few steps further to check for the plausibility of parallel trends by allowing for CCC violations.
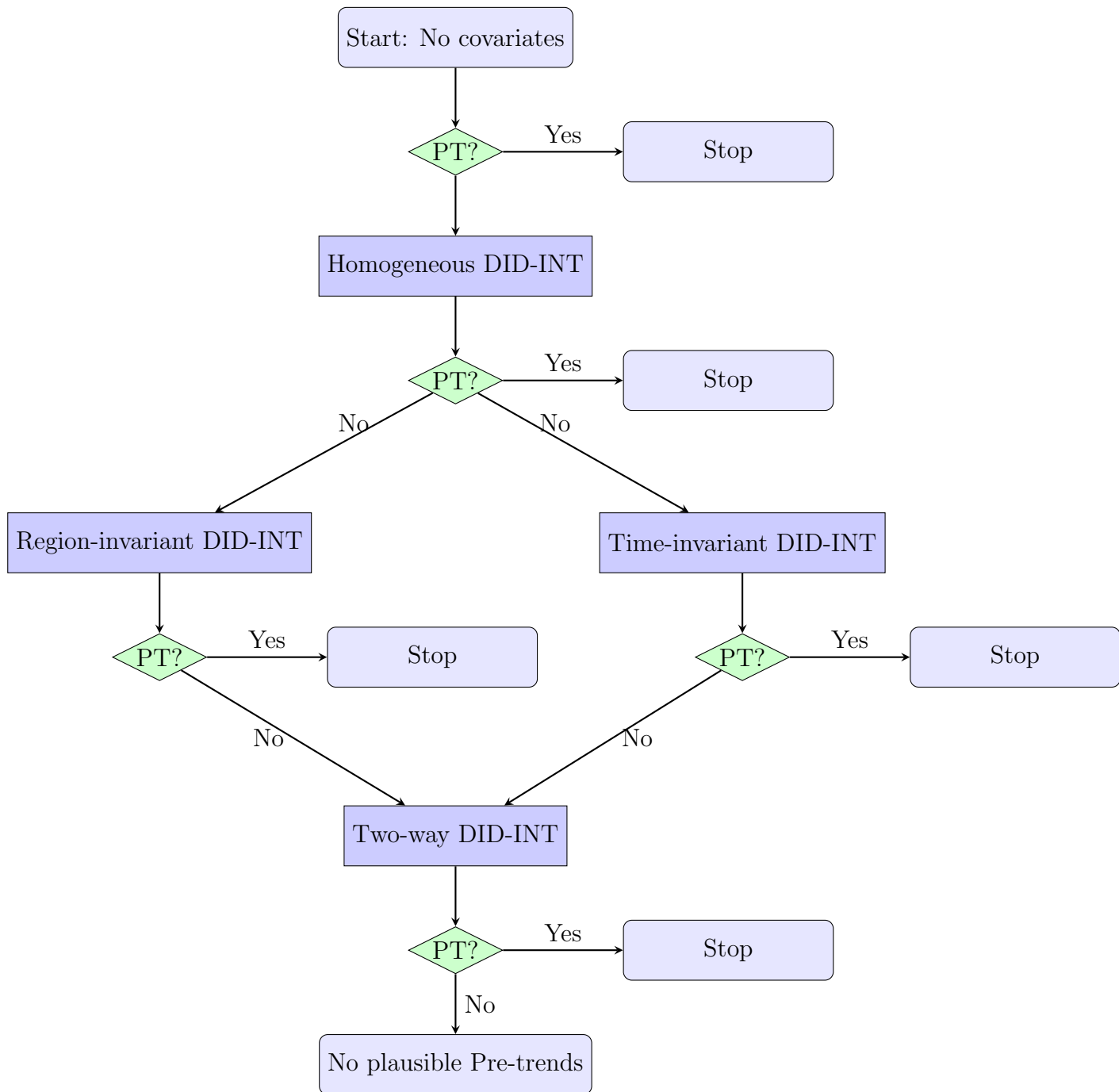
Figure 10: Model Selection Algorithm

In the next step, we recommend interacting the covariates with either the $I(s)$ or $I(t)$ dummies, and plotting trends of the resulting residualized outcomes. These residualized unconditional outcomes are obtained by regressing the outcome on the interacted covariates. If the trends look plausible with these interacted covariates, researchers should use the region-varying or time-varying DID-INT, depending on which interaction yields parallel trends. When the trends remain implausible, the final step is to interact the covariates with both $I(s)$ and $I(t)$ dummies, regress the outcome on these interacted covariates, and check the trends of the resulting residuals. If the trends now appear plausible, researchers should use the two-way DID-INT. However, if none of these steps yield plausible pre-trends, then the conditional parallel trends and intersection parallel trends assumptions are both likely violated, and using any DID method is discouraged. Now we illustrate the model selection algorithm in practice using two examples.
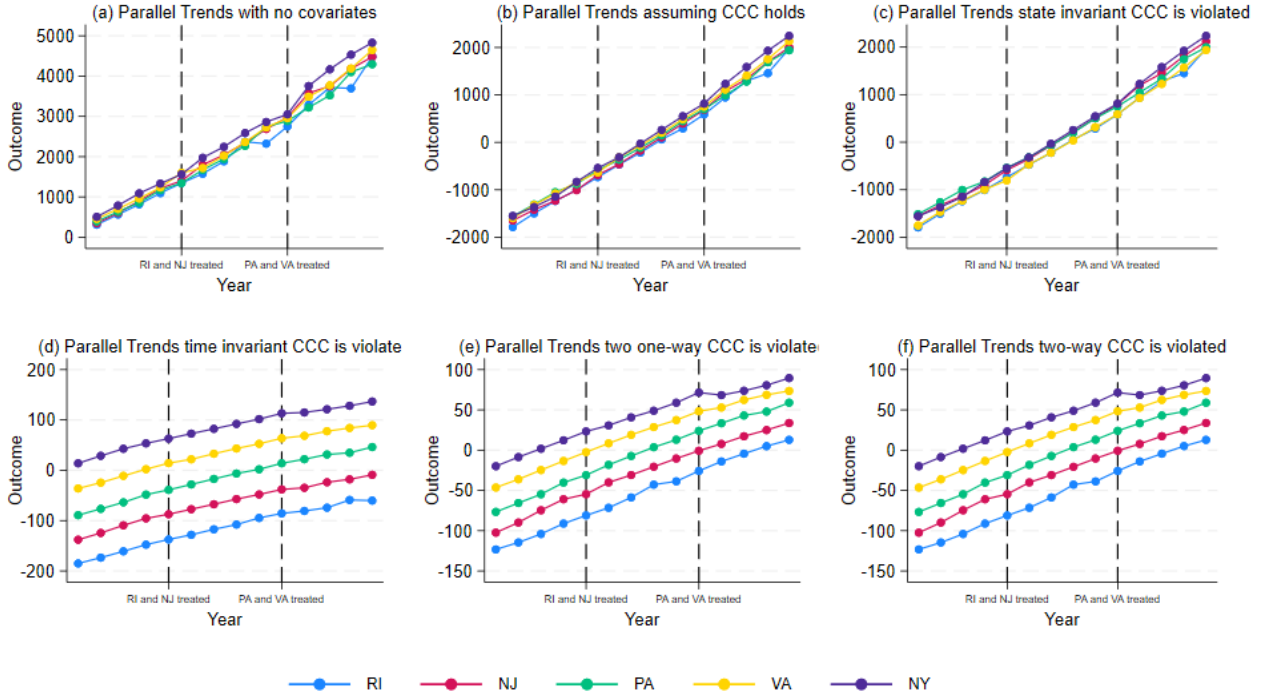

Figure 11: Example 1: DGP with time-invariant CCC violations

Figure 11 uses a DGP where the time-invariant CCC is violated. We begin by assessing pre-trends without covariates and then incorporate different functional forms of covariates, according to the sequence specified by the model selection algorithm. In panels (a), (b) and (c), the parallel trends do not seem plausible. However, in panel (d), after interacting covariates with $I(t)$ dummies, the trend of residualized outcomes become visually parallel. This indicates that the time-varying DID-INT is appropriate, as it is the first specification in the sequence that satisfies the conditional parallel trends assumption. Figure 12 uses a DGP where the two-way CCC is violated. In this example, we observe that parallel trends seem implausible in all panels except panel (f). Therefore, the two-way DID-INT is appropriate based on the model selection algorithm. The software program which implements the model selection algorithm includes an

additional covariate specification option, referred to as the two one-way DID-INT. Details of this specification is provided in Appendix H.
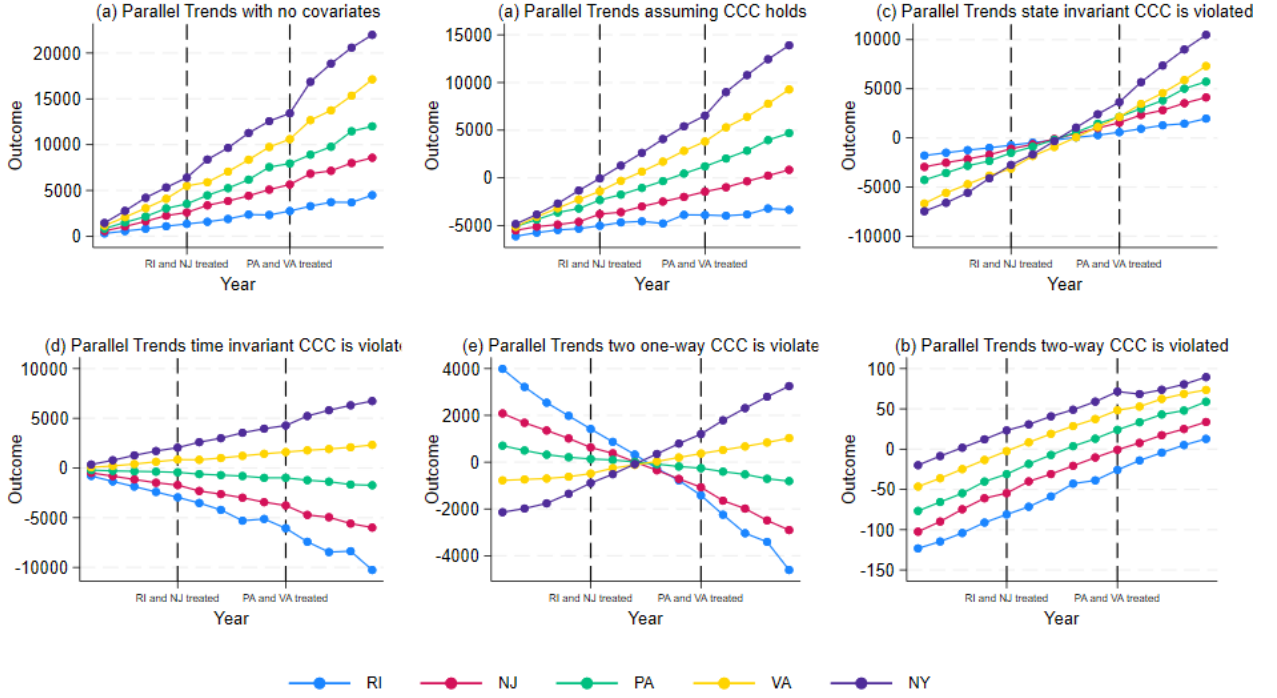

Figure 12: Example 2: DGP with two-way CCC violations

9   The Two-way Fixed Effects (TWFE) Estimator

In this section, we demonstrate that the two-way fixed effects (TWFE) estimator becomes inconsistent when the two-way common causal covariates (CCC) assumption is violated. Because a violation of the two-way CCC necessarily implies that both the region-invariant and time-invariant CCC assumptions are also violated, we focus our analysis on the case of a two-way CCC violation. The results, however, extend directly to the other cases as well. In this proof, we assume that the **conditional parallel trends** hold, but the two-way CCC assumption is violated. In other words, the **necessary conditions** shown in Theorem 2 holds, but the **sufficient conditions** do not. We maintain the homogeneous treatment effects assumption to isolate the inconsistency arising solely from the violation of the CCC assumption in the TWFE regression. The presence of heterogeneous treatment effects would further contribute to inconsistency through forbidden comparisons and negative weighting issues, as discussed by Goodman-Bacon (2021) and De Chaisemartin and d'Haultfoeuille (2020a). The two-way fixed effects model can be written as:

$$Y_{i,s,t} = \alpha_s + \delta_t + \beta^{DD} D_{i,s,t} + \gamma X_{i,s,t} + \epsilon_{i,s,t} \tag{41}$$

The key parameter of interest is β^{DD}, which we derive in this section. For simplicity, we assume

that there is a single time-varying covariate $X_{i,s,t}$ required for **conditional parallel trends** to hold. **It is already well-known in the literature that β^{DD} is inconsistent when conditional parallel trends is violated.** Therefore, we will not analyze the case where conditional parallel trends does not hold. Importantly, as discussed in Section 6.2, conditional parallel trends can still hold when the two-way CCC assumption is violated. However, researchers are unaware of this possibility and rely on the implicit assumption that covariates share a common coefficient across regions and time. As a result, they incorrectly model covariates in the TWFE regression, which introduces inconsistency in β^{DD} even though the underlying CPT assumption may be valid. A proof of Theorem 5 is shown in Appendix E.

Theorem 5 (Two-way fixed effects is inconsistent even when conditional parallel trends hold)**.**

$$\beta^{DD} = E[\widehat{\beta}^{DD}] \neq \tau \tag{42}$$

In order to accomodate two-way CCC violations, the TWFE model shown in Equation (41) can be modified to include interacted covariates with both the $I(s)$ and the $I(t)$ dummies. Researchers typically do not use interacted covariates in practice, since the model selection algorithm we proposed in Section 8.2 was not previously available to guide model specifications. The TWFE with the correct functional form of covariates can be written as follows:

$$Y_{i,s,t} = \alpha_s + \delta_t + \beta^{DD}_{modified}D_{i,s,t} + \gamma_{s,t}I(s)I(t)X_{i,s,t} + \epsilon_{i,s,t} \tag{43}$$

We call this the **modified two-way fixed effects** estimator. This estimator can provide consistent estimates of the ATT when both the conditional parallel trends and the intersection parallel trends hold, and the two-way CCC is violated. A proof of Theorem 6 is shown in Appendix F.

Theorem 6 (Two-way fixed effects is inconsistent even when conditional parallel trends hold)**.**

$$\beta^{DD} = E[\widehat{\beta}^{DD}_{modified}] = \tau \tag{44}$$

It is important to note that, both the TWFE and the modified TWFE estimators will include forbidden comparisons and negative weighting issues under heterogeneous treatment effects in staggered adoption designed. Note that, although the TWFE can be modified to allow for CCC violations, other alternative estimators cannot. We discuss these alternative estimators in the next section.

10 Other Difference-in-Difference Estimators

In this section we discuss three alternative difference-in-difference estimators. Specifically, we discuss: the widely used Callaway and Sant'Anna (2021) estimator for staggered adoption in Section 10.1; the imputation estimator from Borusyak et al. (2024) in Section 10.2; and the new FLEX estimator from Deb et al. (2024) which can handle time varying covariates in Section 10.3.

10.1 Callaway and Sant'Anna (2021) DiD estimator

In this section, we will explore the potential inconsistency that arises in the Callaway and Sant'Anna (2021) DiD estimator (CS-DID) when the two-way CCC assumption is violated. The CS-DID is a semi-parametric method that estimates the ATT without forbidden comparisons, as demonstrated by Goodman-Bacon (2021) and De Chaisemartin and d'Haultfoeuille (2020a). The estimation of the ATT involves two steps. In the first step, the dataset is decomposed into several "2x2 comparison" blocks, each containing a treated group and an untreated (or not yet treated) group. The pre-intervention period is the period right before the treated group is treated. Without covariates, the ATT of each of the "2x2 comparison" blocks, known as $ATT(g,t)$, is estimated non-parametrically as follows:

$$\widehat{ATT(g,t)} = \left(\overline{Y_{i,g,t}} - \overline{Y_{i,g,g-1}}\right) - \left(\overline{Y_{i,g',t}} - \overline{Y_{i,g',g-1}}\right) \tag{45}$$

The groups or cohorts are determined by the period they were first treated (g), which differs from DID-INT. The second step involves taking a weighted average of all the $ATT(g,t)$'s to get an overall estimate of the ATT:

$$\widehat{ATT} = \sum_{g=2}^{G} \sum_{t=2}^{\mathcal{T}} 1\{g \leq t\} w_{g,t} \widehat{ATT(g,t)} \tag{46}$$

The above avoids all the forbidden comparisons demonstrated by Goodman-Bacon (2021) and De Chaisemartin and d'Haultfoeuille (2020a). With covariates, the first step is estimated using the Doubly Robust DiD (DR-DID) approach first proposed by Sant'Anna and Zhao (2020) by default. The DR-DID approach combines the inverse probability weighting (IPW) approach proposed by Abadie (2005) and the outcome regression (OR) approach proposed by Heckman et al. (1997) to derive a doubly robust estimator. This estimator is robust to misidentification, provided either the propensity score model or the outcome regression model is correctly specified. The CS-DID can also estimate the $ATT(g,t)$'s using other approaches like the inverse probability weighting or regression adjustment (Rios-Avila et al., 2021). However, using the DR-DID can be advantageous if the propensity score and outcome regressions depends on time varying covariates in both periods, due to the property of double robustness (Caetano et al., 2022).

Matching methods such as IPW, OR and DR-DID are used when the conditional parallel trends assumption is likely to be implausible. To ensure a cleaner comparison group, units in the control group are re-weighted so that observations with covariates more similar to the treatment group receive a higher weight than those that do not. There are three disadvantages to using these estimators. The first disadvantage is that semi-parametric approaches can only eliminate inconsistencies if conditional parallel trends seems implausible. However, these methods can provide inconsistent estimates of the ATT when CPT holds, and lead to inefficiencies by dropping (or giving less weight on) observations in the control group that differ from the treated group in terms of covariates (O'Neill et al., 2016).

The second disadvantage is that, semi-parametric approaches like the CS-DID, DR-DID and IPW cannot incorporate interacted covariates as controls, unlike the modified TWFE, due to violations of the overlap assumption. Therefore, we do not have a modified model for CS-DID using the default settings. In addition, the IPW and OR require strictly time invariant covariates to estimate the ATT (Abadie, 2005; Heckman et al., 1997). The DR-DID can estimate the ATT with time varying covariates, provided covariate exogeneity hold.

A third disadvantage is that the propensity scores in IPW and DR-DID tend to place more weight on observations with substantial overlap in covariate values between treated and control groups. In this paper, we refer to this issue as the problem of **implicit sub-sampling** with propensity scores. When the ATT conditional on covariates, $ATT(x)$, varies with the value of these covariates, such sub-sampling can introduce inconsistencies. To illustrate this, consider an example where we want to estimate the ATT of a job training program on wages. Suppose the program is designed to benefit individuals with lower cognitive ability, as measured by test scores. In this case, the $ATT(x)$ would be higher for individuals with low cognitive abilities compared to mid or high level cognitive abilities. However, the aggregate ATT is positive when averaged over the full distribution of x. Now imagine the distribution of cognitive abilities is different between the treatment and control groups, such that the majority of the overlap occurs for individuals with mid level cognitive abilities. In that case, the propensity scores are non-zero for the middle of the x distribution, and quite small elsewhere. Hence the ATT being estimated is much closer to the $ATT(x) \in$ mid, and may be inconsistent compared to the desired ATT. In contrast, DID-INT does not rely on propensity scores, and is therefore robust to the implicit sub-sampling problem. The difference essentially comes from propensity scores attempting to estimate $ATT(x)$ and then integrating over x, while DID-INT attempts to residualize the outcome by the covariate before estimating the ATT.

Since the DR-DID approach is used to estimate the ATT of each of the "2x2" comparison blocks in the CS-DID by default, let us analyze the DR-DID estimator in a two group, two period framework. For the treated group g, the treatment dummy D_i is assigned a value of 1. For the control group g', D_i is assigned a value of 0. In this canonical framework, the DR-DID estimand of the ATT is shown in Equation (47) (Caetano et al., 2022).

Corollary 1 (DR-DID Estimand). The DR-DID estimand under conditional parallel trends, no anticipation and no overlap:

$$\theta^{DRDID} = E\left[\left(\frac{D}{E[D]} - \frac{P(X_{i,g,t})(1-D)}{E[D](1-P(X_{i,g,t}))}\right)\left(Y_{i,g,t} - Y_{i,g,g-1} - E[Y_{i,g',t} - Y_{i,g',g-1}|X_{i,g',t}, X_{i,g',g-1}, G = g']\right)\right]$$
(47)

Note, the period right before treatment in for group g in Callaway and Sant'Anna (2021) is $g-1$. Theorems 7, 8 and 9 demonstrates how identification of the key causal parameter of interest τ for DR-DID depends on the type of covariates used, and the plausibility of the two-way CCC assumption.

Theorem 7 (Identification: DR-DID with time-invariant covariates). *The DR-DID is consistent under overlap, conditional parallel trends, no anticipation, homogeneous treatment effects, two-way CCC assumptions and when the covariates are time-invariant.*

$$\theta^{DRDID} = E[\widehat{\theta}^{DRDID}] = \tau$$
(48)

Theorem 8 (Identification: DR-DID with time-varying covariates). *The DR-DID is inconsistent under overlap, conditional parallel trends, no anticipation, homogeneous treatment effects, two-way CCC assumptions, and when the covariates are time-varying.*

$$\theta^{DRDID} = E[\widehat{\theta}^{DRDID}] \neq \tau$$
(49)

Theorem 9 (Identification: DR-DID with time-varying covariates and violations of the two-way CCC). *The DR-DID is inconsistent under overlap, conditional parallel trends, no anticipation, homogeneous treatment effects assumptions, but the two-way CCC assumption is violated and covariates are time-varying.*

$$\theta^{DRDID} = E[\widehat{\theta}^{DRDID}] \neq \tau$$
(50)

The proofs are Theorem 7, 8 and 9 are shown in Appendix G states that the DR-DID is consistent only when overlap, conditional parallel trends, no anticipation, homogeneous treatment effects and the two-way CCC assumptions hold, and time invariant covariates are used. However, DR-DID is inconsistent with time-varying covariates regardless of whether the two-way CCC assumption holds or is violated. This result means that if the two-way CCC assumption is violated, then the CS-DID estimator does not produce a consistent estimate of the ATT, since it uses the DR-DID by default. Unfortunately, unlike TWFE, it is not possible to modify the CS-DID estimator to allow for these violations.

10.2 Imputation Estimator

In this section we compare the (static) imputation estimator proposed by Borusyak et al. (2024), also designed to estimate the ATT under staggered treatment adoption. The imputation

estimator is done in four steps. In the first step, the unit (α_i) and period (β_t) fixed effects, and coefficients of covariates (γ) are fitted by using a TWFE model on the untreated units only. The regression is shown in Equation (51).

$$Y_{i,s,t} = \alpha_i + \beta_t + \tau^{static} D_{i,s,t} + \gamma X_{i,s,t} + \epsilon_{i,s,t} \tag{51}$$

Note, for all observations in the control group, $D_{i,s,t} = 0$. In the second step, the estimated fixed effects from this model are used to impute the untreated potential outcomes of the treated group, shown in Equation (52).

$$\widehat{Y}_{i,s,t}(0) = \widehat{\alpha}_i + \widehat{\beta}_t + \widehat{\gamma} X_{i,s,t} + \epsilon_{i,s,t} \tag{52}$$

In the third step, the individual treatment effects are estimated by taking a difference of the observed outcome of each treated unit and the imputed counterfactuals. This is shown in Equation (53).

$$\widehat{\tau}_{i,s,t} = Y_{i,s,t} - \widehat{\alpha}_i + \widehat{\beta}_t - \widehat{\gamma} X_{i,s,t} \ \forall i \in S^T \tag{53}$$

In the final step, the overall ATT is estimated by taking a weighted average of $\tau_{i,s,t}$ across all treated units and periods. Although the imputation estimator provides a viable method for estimating the ATT in the presence of time-varying covariates and staggered treatment designs, it is not robust to CCC violations. Specifically, step 2 relies on the assumption that the coefficients of the covariates are identical for the treated and control groups in order to impute the untreated potential outcomes of the treated observations. As a result, the overall estimate of the ATT obtained in step 4 may be inconsistent under CCC violations.

This means that the imputation estimator does not provide a consistent estimate of the ATT when the CCC assumption is violated. Unfortunately, it is also not possible to easily modify the estimator as the current version fits all parameters using regressions with only untreated units.

10.3 The FLEX model

In this section, we compare the DID-INT to the flexible linear model or FLEX proposed by Deb et al. (2024). The FLEX model also interacts the covariates with a group dummy and a time time dummy. However, FLEX model generates three types of variables: one where the covariates are interacted with the group dummies ($I(g)X_{i,g,t}^k$); one where the covariates are interacted with the time dummies ($I(t)X_{i,g,t}^k$); and the third where the covariates are interacted with both time and group dummies ($\sum_{g \neq \infty} \sum_{t \geq r} \sum_k \beta_{gtk} I(g) I(t) X_{gtk}$). Importantly, these "intersection" dummies are only for the treated units in either the post period, or all periods, depending on whether or not the 'leads' option is specified. These covariates are then included in the FLEX model in an additive way: $\sum_{g \neq \infty} \sum_{t \geq t^*} \sum_k \beta_{gtk} I(g) I(t) X_{gtk} + \sum_g \sum_k \beta_{gk} I(g) X_{gk} + \sum_t \sum_k \beta_{tk} I(t) X_{tk}$. The regression for the FLEX model is shown below:

$$y_{gt} = \sum_{g \neq \infty} \sum_{t \geq t^*} \tau_{gt} I(g) I(t) + \sum_{g \neq \infty} \sum_{t \geq t^*} \sum_k \beta_{gtk} I(g) I(t) X_{gtk}$$
$$+ \sum_g \sum_k \beta_{gk} I(g) X_{gk} + \sum_t \sum_k \beta_{tk} I(t) X_{tk} \tag{54}$$
$$+ \sum_k \beta_k X_k + \sum_t \phi_t I(t) + \sum_g \psi_g I(g) + \epsilon_{gt}.$$

The second step involves taking a weighted average of the estimated treatment effect similar to the DID-INT. We highlight a few key differences between the FLEX and the two-way DID-INT. First, the FLEX model includes the three types of interacted covariates in the regression specification shown above, in addition to non-interacted covariates. In contrast, the DID-INT only includes the covariates interacted with the time or group dummies, or both. Second, the FLEX model includes two-way interactions of covariates for only a subset of the 'intersections'. As mentioned, these are only estimated for the treated groups. They are either estimated only for the post-intervention period when 'leads' is not specified, and both pre-intervention and post-intervention periods when it is not. This implies that the DID-INT can capture the variations across time and group in both treatment and control groups. Whether DID-INT or FLEX is estimating more parameters depends on the number of groups, the number of time periods, and the number of covariates. Finally, FLEX is based on the TWFE model and tries to model the untreated outcomes with group and time fixed effects. Our simulations show that FLEX is inconsistent when CCC is violated, see Section 11.4.

11 Monte Carlo Simulation Study

In this section, we introduce the design of our Monte Carlo Simulation Study. This is first used to analyze the properties of the standard TWFE and the modified TWFE described in section 9. To keep the constructed dataset as realistic as possible, we use data from the Current Population Survey (CPS) covering the years 2000 to 2014. The CPS is a repeated cross-sectional dataset that includes information on employment status, earnings, education, and demographic trends of individuals. Similar to Bertrand et al. (2004), we restrict our sample to women between the ages of 24 and 55 in their fourth interview month.

To generate our constructed outcome, we start by estimating coefficient for selected covariates based on individual's weekly earnings. We limit our analysis to Rhode Island, New Jersey, Pennsylvania, Virginia, and New York, where parallel trends seem plausible. The parallel trends figures are shown in Figure (13). The chosen covariates include age, race, education and marital status, which are known to influence weekly wages. Race, education, and marital status are transformed into binary variables, while age remains continuous. When the two-way CCC holds, the coefficients of covariates which are to be used in the DGP are estimated using
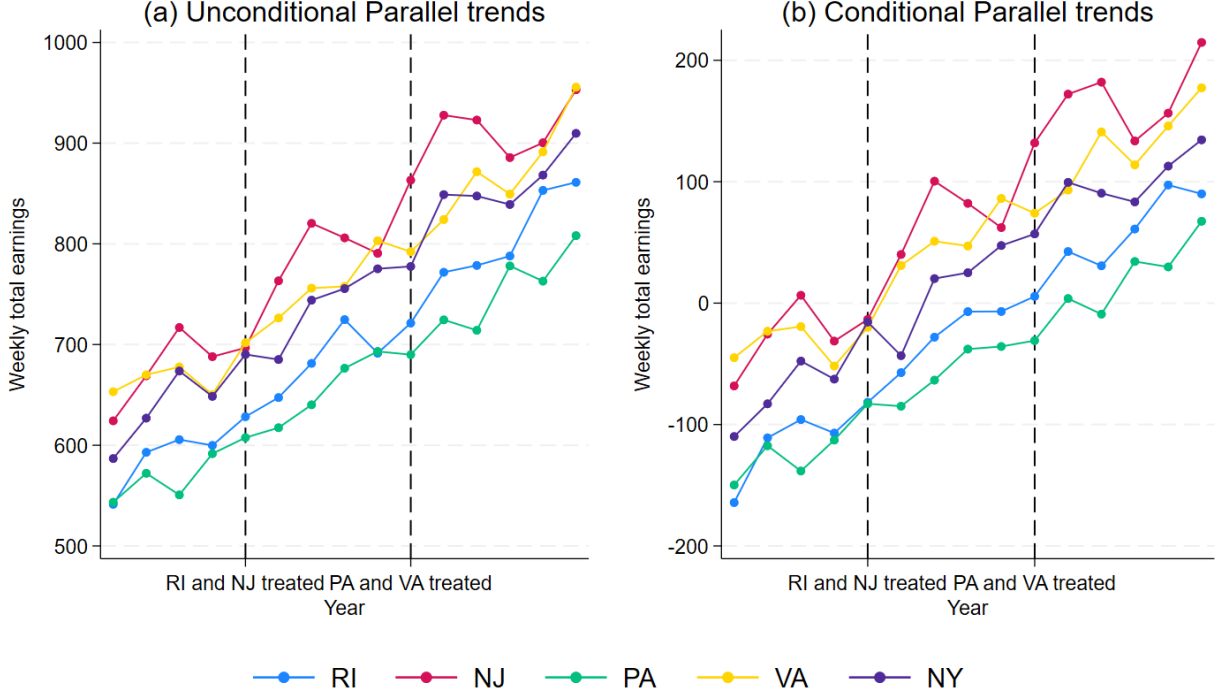
Figure 13: Parallel trends for weekly earnings

the following regression:

$$\text{earnings}_{i,s,t} = \phi_0 + \sum_k \gamma^k X_{i,s,t}^k + \epsilon_{i,s,t}. \tag{55}$$

When the two-way CCC assumption is violated, we estimate the coefficients by running a separate regression for each group and period. The regression is shown in Equation (56).

$$\text{earnings}_{i,s,t} = \phi_0 + \sum_k \gamma_{s,t}^k X_{i,s,t}^k + \epsilon_{i,s,t} \quad \text{if group} = s \ \text{ and } \ \text{year} = t. \tag{56}$$

We generate two types of outcomes, one where the two-way CCC assumption holds ($Y_{i,s,t}^1$) and one where the two-way CCC assumption is violated ($Y_{i,s,t}^2$). We begin by generating a baseline earning variable called y_0, which is generated using the following formula:

$$y_0 = y_{init} + \widehat{\beta_s^0}\text{year} \quad \text{if group} = s, \tag{57}$$

where, y_{init} follows a normal distribution, with the mean being the average weekly earnings for all individuals in group s in the year 2000. The time trend β_t^0 is estimated from the following regression:

$$\text{earnings}_{i,t} = \alpha_0 + \beta_t^0\text{year} + \epsilon_{i,t}. \tag{58}$$

When the two-way CCC holds, the known-DGP outcome is generated as follows:

$$Y_{i,s,t}^1 = y_0 + \sum_k \widehat{\gamma^k} X_{i,s,t}^k. \tag{59}$$

where, $\widehat{\gamma^k}$'s are the estimated coefficients from the regression in Equation (55). Conversely, when the two-way CCC is violated, the known-DGP outcome is generated as follows:

$$Y_{i,s,t}^2 = y_0 + \sum_k \widehat{\gamma_{s,t}^k} X_{i,s,t}^k \quad \text{if group} = s \;\; \text{and} \;\; \text{year} = t. \tag{60}$$

where, $\widehat{\gamma_{s,t}^k}$'s are the estimated coefficients from the regression in Equation (56).

To incorporate a staggered adoption design, Rhode Island and Pennsylvania are treated in 2004, while New Jersey and Virginia are treated in 2009. The true ATT (ATT^0) is set to be zero, which implies that the **homogeneous treatment effects assumption** holds. In this study, we maintain this assumption to rule out inconsistency which arises from negative weighting issues and forbidden comparisons in a staggered treatment rollout framework as highlighted by Goodman-Bacon (2021). This will help us isolate the inconsistency which arises from violations of the two-way CCC assumption. Once the dataset has been constructed, we estimate the ATT using the standard TWFE, the modified TWFE and the two-way DID-INT repeat the process a 1000 times. We then explore the kernel densities of the ATT estimates from each estimator to explored the consistency and efficiency of the two estimators. In addition to **homogeneous treatment effects**, **conditional parallel trends** and **no anticipation** assumptions hold as well. The two-way CCC assumption holds for $Y_{i,s,t}^1$, but is violated for $Y_{i,s,t}^2$.

The results are shown in Figure 14. Panel (a) shows the the kernel densities for the standard TWFE, the modified TWFE and the two-way DID-INT when the two-way CCC assumption holds, while panel (b) shows the densities when the two-way CCC assumption is violated. In panel (a), all three estimators are consistent, with their densities centered around the true ATT value of 0. In panel (b), we observe that the modified TWFE and the two-way DID-INT remains consistent, while the standard TWFE is inconsistent. It is worth noting that, when the homogeneous treatment effects assumption is violated, both the TWFE and the modified TWFE will be inconsistent due to negative weighting issues and forbidden comparisons (Goodman-Bacon, 2021). However, the two-way DID-INT is robust to these issues, since the forbidden comparisons are excluded in the fifth step where all the "valid" $ATT(s,t)$'s are aggregated together to get an overall estimate of the ATT. Since inconsistency implies that the estimator does not converge to the true parameter value as the sample size grows, an inconsistent estimator will generally be biased, even in large samples. Therefore, Figure 14 demonstrates the **bias-variance tradeoff** across estimators: reducing bias often requires using estimators with greater variance. When the two-way CCC is violated, we observe that the TWFE has a higher bias, but a low variance. In contrast, both the modified TWFE and two-way DID-INT have a low bias, at the cost of higher variance. This is illustrated by the wider kernel densities for the
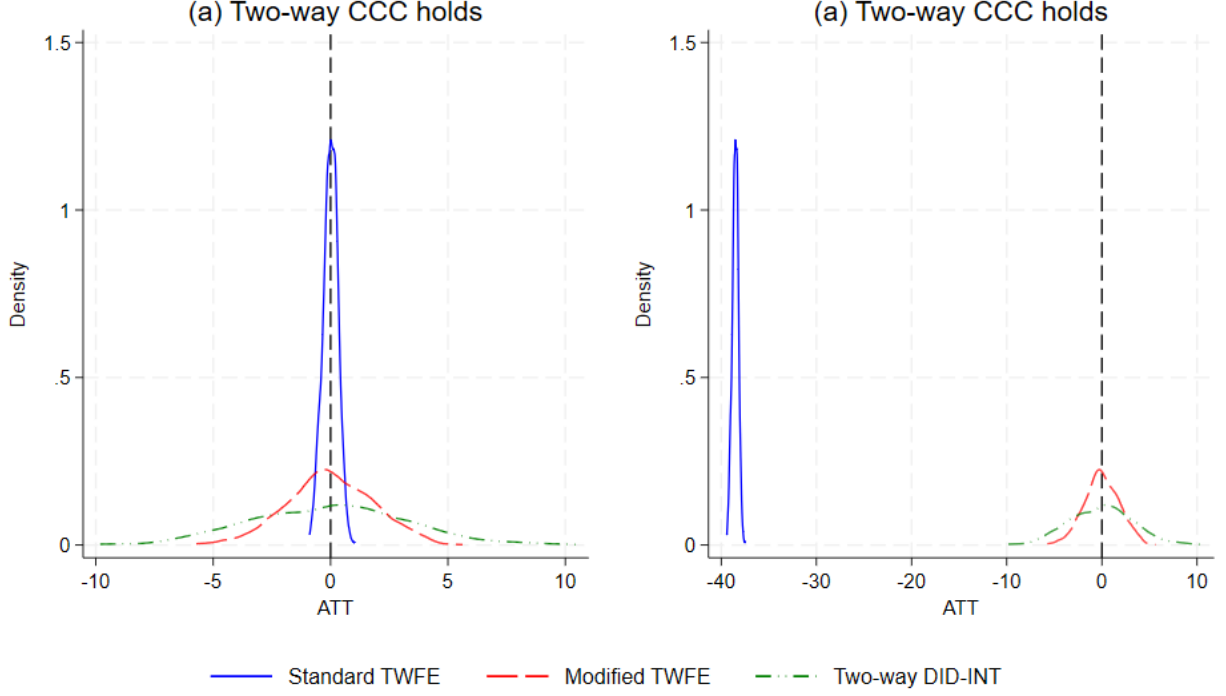
Figure 14: Kernel Densities of the standard TWFE, modified TWFE and two-way DID-INT

modified TWFE and DID-INT, indicating that these estimators are less efficient compared to the TWFE. The two-way DID-INT is the most inefficient among the three, as reflected by the widest kernel density among the three.

11.1 DID-INT vs DID-INT

In this section, we explore the performance of the four types of DID-INT highlighted in section 8.2 across all possible DGPs that may arise in empirical settings. To do so, we incorporate two additional constructed outcomes. In the first, denoted as $Y_{i,s,t}^3$, only the region-invariant CCC is violated but the time-invariant CCC holds. In this DGP, the coefficients of covariates are estimated from the CPS data using the following regression:

$$\text{earnings}_{i,s,t} = \phi_0 + \sum_k \gamma_s^k X_{i,s,t}^k + \epsilon_{i,s,t} \quad \text{if group} = s. \tag{61}$$

We then generate $Y_{i,s,t}^3$ using the following:

$$Y_{i,s,t}^3 = y_0 + \sum_k \widehat{\gamma_s^k} X_{i,s,t}^k \quad \text{if group} = s. \tag{62}$$

where, y_0 baseline income variable. In the the second additional constructed outcome, labeled $Y_{i,s,t}^4$, only the time-invariant CCC is violated. Similar to the previous DGP, the coefficients

are estimated from CPS data, using the following regression:

$$\text{earnings}_{i,s,t} = \phi_0 + \sum_k \gamma_t^k X_{i,s,t}^k + \epsilon_{i,s,t} \quad \text{if year} = t. \tag{63}$$

We then generate $Y_{i,s,t}^4$ using the following:

$$Y_{i,s,t}^4 = y_0 + \sum_k \widehat{\gamma_t^k} X_{i,s,t}^k \quad \text{if year} = t. \tag{64}$$

For the four possible DGPs, we run the region-varying DID-INT, the time-varying DID-INT and the two-way DID-INT and compare the kernel densities across methods. The results are shown in Figure 15. Panel (a) uses a DGP $(Y_{i,s,t}^1)$ where assumptions conditional parallel trends, no anticipation and the two-way CCC hold. Panel (b) depicts a DGP $(Y_{i,s,t}^3)$ where assumptions conditional parallel trends and no anticipation, but the region-invariant CCC is violated. In panel (c), a DGP $(Y_{i,s,t}^4)$ is used where the conditional parallel trends and no anticipation assumptions holds, but the time-invariant CCC assumption is violated. Lastly, panel (d) illustrates the DGP $(Y_{i,s,t}^4)$ where the conditional parallel trends and no anticipation assumptions holds, but the two-way CCC assumption is violated. Two-way violations imply that neither region-invariant or time-invariant CCC holds.
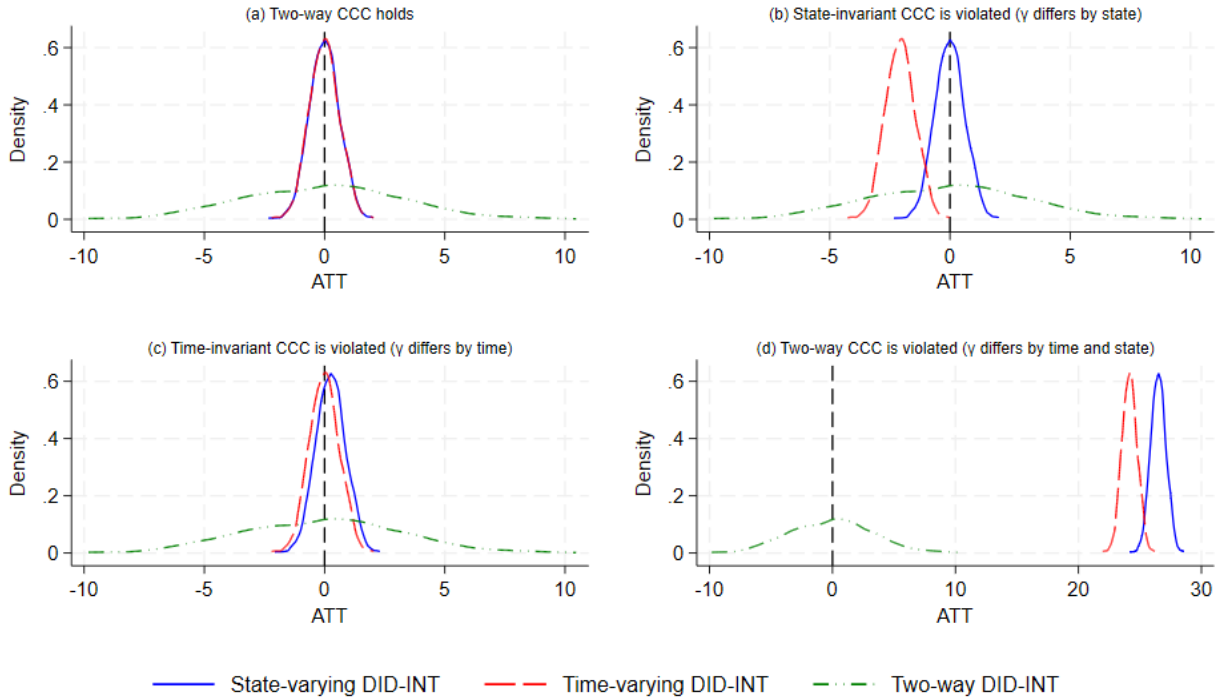

Figure 15: Kernel Densities of region-varying, time varying and two-way DID-INT

In Panel (a) we observe that all three estimators are consistent. However, the two-way DID-INT is less efficient compared to the region-varying and the time-varying versions of DID-INT. Since DID-INT estimates each group and time interactions separately for each covariate, we

expect the variance of the estimate to be higher compared to the versions of DID-INT with just group or time interactions. Furthermore, the higher number of estimated parameters in this specification lowers the degrees of freedom.

In Panel (b), the region-varying DID-INT is consistent, while the time-varying DID-INT is inconsistent. The inconsistency in the time-varying DID-INT arises from model mis-specification, as it fails to capture the variation of the covariates accross regions. Conversely, in Panel (c), the time-varying DID-INT is consistent and the state-varying DID-INT is inconsistent due to mis-specification. In this case, the region-varying CCC is inconsistent as it does not capture the variations of the covariates over time. In Panel (d), both the region-varying and time-varying DID-INT are inconsistent.

The Two-way DID-INT model is consistent across all types of DGPs. However, this consistency comes at the cost of efficiency. In Panel (b), the two-way DID-INT estimator is less efficient compared to the region-varying DID-INT. Similarly, in Panel (c), the two-way DID-INT is less efficient compared to the time-varying DID-INT. This is an example of the bias-variance trade off, which highlights the efficiency loss from ensuring accurate parameter estimates. In most empirical settings, the true underlying DGP is unknown. Therefore, we recommend that researchers either: A) use the two-way DID-INT as default, since it is consistent across all possible DGPs, or B) use the model selection algorith presented in Section 8.2 to investigate parallel trends under different CCC assumptions and select the most parsimonious model which satisfies parallel trends.

11.2   Callaway and Sant'Anna Monte Carlo

Similar to the preceding sections, we will analyze the kernel densities of the CS-DID estimator from the Monte Carlo simulation study to evaluate its performance relative to the two-way DID-INT estimator. We will examine these kernel densities under the DGPs where two-way CCC holds and where it is violated. The results are shown in Figure 16. Since the DGP contains time-varying covariates, we observe that the CS-DID is inconsistent when the two way CCC holds. In panel (b), the inconsistency is amplified due to violations of the two-way CCC assumption. In both panels, the two-way DID-INT is consistent. These findings reflect the theoretical findings in Section 10.1. We also observe a bias-variance tradeoff between the two-way DID-INT and the CS-DID estimators. The two-way DID-INT has a wider distribution compared to the CS-DID, reflecting its lower efficiency.
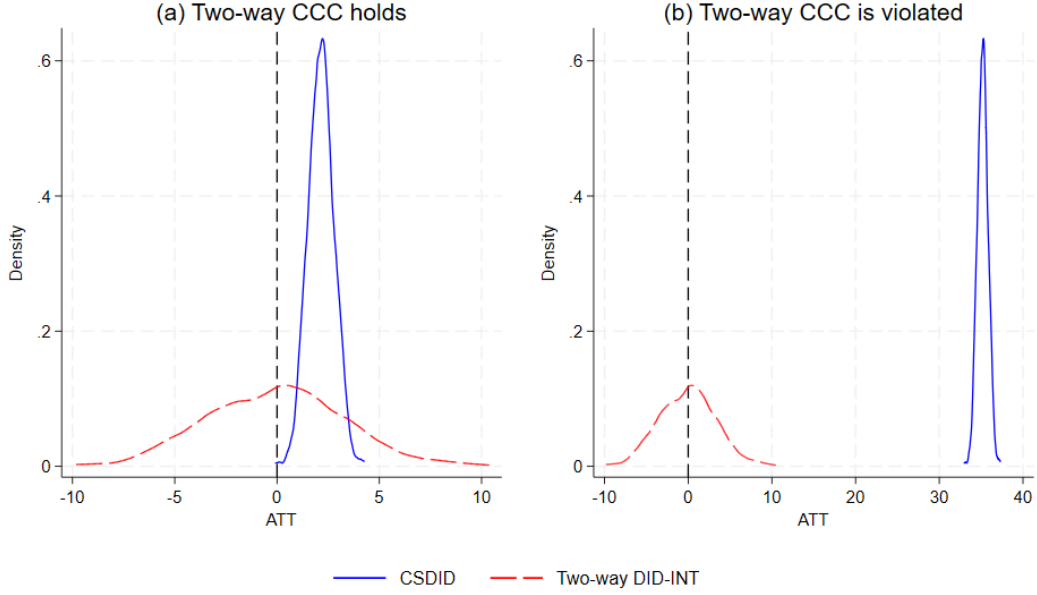
Figure 16: Kernel Densities of the CS-DID and two-way DID-INT

11.3   Imputation Monte Carlo

In this subsection, we run the static Imputation estimator on the same Monte Carlo design described in Section 11, and compare its kernel density to the kernel density of the two-way DID-INT. The results are shown in Figure 17. In Panel (a), we observe that both the Imputation estimator and the two-way DID-INT are inconsistent when there are no two-way CCC violations. In Panel (b), the two-way DID-INT is consistent, while the Imputation estimator is inconsistent when there are two-way CCC violations. This inconsistency in the imputation estimator arises from step 2, which uses the same value of the coefficient of $X_{i,s,t}$ ($\widehat{\gamma}$) from the control group to impute the unobserved counterfactual for the treated group in the post intervention period.

11.4   FLEX Monte Carlo

To compare the performance of the DID-INT to the FLEX, we compare the kernel densities of the two estimators using the same Monte Carlo simulation design described in Section 11. The results are shown in Figure 18. In Panel (a), we observe that both the two-way DID-INT and the FLEX model are consistent. In Panel (b), the two-way DID-INT is consistent, but the FLEX model is inconsistent. This inconsistency results from the inability of the FLEX model to capture within group variations of the coefficients for the control groups. The two-way DID-INT is less efficient compared to the FLEX, reflected by its wider kernel density.
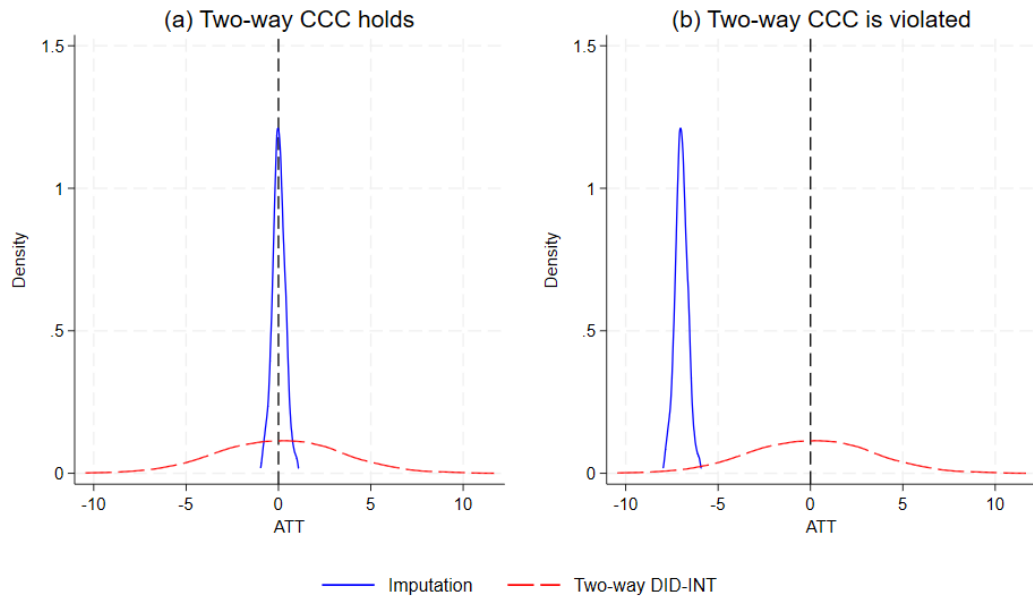
43

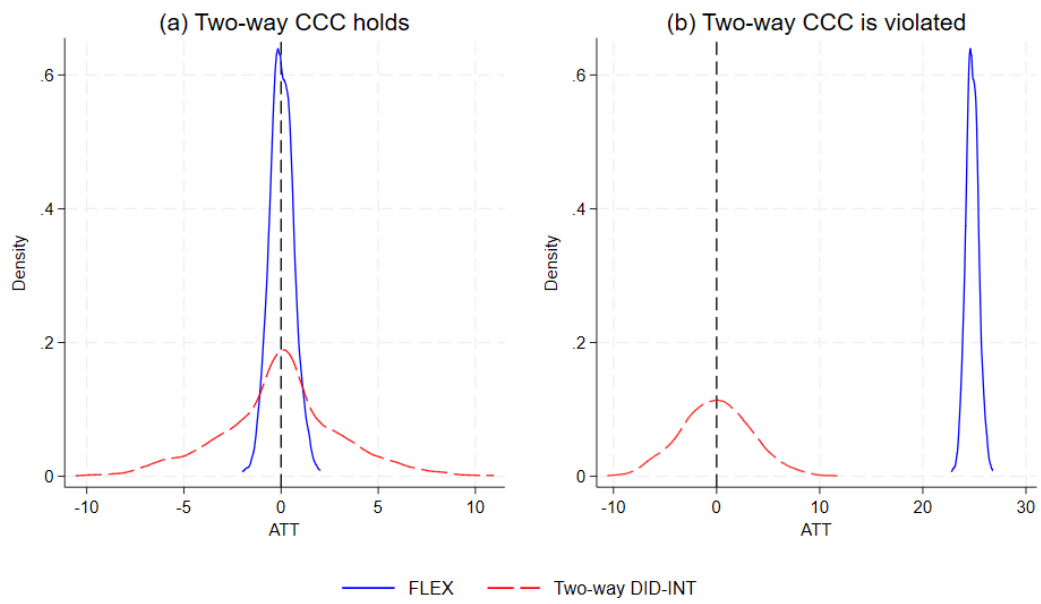Figure 17: Kernel Densities of the Imputation estimator and two-way DID-INT

Figure 18: Kernel Densities of FLEX and the two-way DID-INT

12   Conclusion

Difference-in-differences (DiD) is widely used in estimating treatment effects for policies which have been implemented at a jurisdictional level. However, existing DiD methods require careful selection of covariates to recover an consistent estimate of the ATT. The literature recommends using either time-invariant covariates or pre-treatment covariates when the covariates change with time. Nonetheless, researchers may still want to include time varying covariates in DiD analysis, even though they are not necessary for parallel trends to be plausible. The study contributes to existing literature by providing researchers with a tool to obtain a consistent estimate of the ATT when time varying covariates are used, called the Intersection Difference-in-differences (DID-INT).

We began the analysis by introducing a new assumption called the common causality covariates (CCC) assumption, which has been implied in the existing literature. In particular, we introduce three types of CCC assumptions called the region-invariant CCC, time-invariant CCC and the two-way CCC. The region-invariant CCC assumes that the effects of the covariates are the same between regions, while the time-invariant CCC assumes that these effects remain stable across time. The two-way CCC combines both, implying that the effect of the covariates remain constant across both regions and time. When the two-way CCC holds, both region-invariant CCC and time-invariant CCC holds as well.

We propose three versions of DID-INT depending on the assumptions we make on the covariates. The region-varying CCC accounts for region-invariant CCC violations by interacting time-varying covariates with region dummies. Conversely, the time-varying DID-INT accounts for time-invariant violations by interacting covariates with time dummies. Finally, the two-way DID-INT adjusts for two-way CCC violations, by interacting the covariates with both region and time dummies. This new estimator relies on parallel trends of the residualized outcome variable, with a flexible functional form for the covariates called intersection parallel trends. This can recover parallel trends that can be missed by less flexible functional form.

We show, through theoretical proofs and a Monte Carlo simulation study, that the TWFE is inconsistent when the two-way CCC assumption is violated since researchers model covariates by assuming a common effect. This is demonstrated in a staggered rollout setting with additional homogeneity assumption of treatment. We also show that the a modified TWFE with interacted covariates can provide a consistent estimate of the ATT when the two-way CCC is violated, at the cost of a loss of efficiency. Moreover, we show that the two-way DID-INT can provide a consistent estimate of the ATT, but with efficiency losses over both the standard TWFE and the modified TWFE. The DID-INT is robust to the forbidden comparisons and negative weighting issues prevalent in both the conventional and modified TWFE estimators when the homogeneity assumption of treatment is relaxed.

Additionally, we compare the region-varying, time-varying and two-way DID-INT across four DGPs to assess the consistency and efficiency of the estimators. Our findings demonstrate

that the two-way DID-INT is inconsistent across all DGPs, but it is less efficient compared to the other estimators. When only the region-invariant CCC is violated, the region-varying DID-INT is consistent, while the time-varying DID-INT is inconsistent. Conversely, the time-varying DID-INT is consistent, while the region-varying DID-INT is inconsistent when only the time-invariant CCC is violated. Since researchers are unable to observe the DGP in empirical settings, we recommend the two-way DID-INT as default, since it is consistent across all DGPs.

Finally, we compare the performance of the two-way DID-INT to CS-DID, Imputation and FLEX, all of which are robust to forbidden comparisons and negative weighting issues in staggered treatment rollout settings with heterogeneous treatment effects. We show that the CS-DID is inconsistent both when the two-way CCC assumption is violated and when it holds, on account of time varying covariates in the latter case. The imputation and FLEX estimators are consistent when the two-way CCC holds. However, both are inconsistent when the two-way CCC is violated, but is more efficient than DID-INT.

Evidently, the role of covariates is more complicated than simply knowing *which* covariates are in the DGP. *How* the covariates enter the DGP also matters. We can try to establish this using the model specification algorithm from Section 8.2, or remain agnostic about this, at the expense of efficiency, and use the two-way DID-INT as a default.

References

Abadie, A. (2005) 'Semiparametric difference-in-differences estimators,' *The review of economic studies* 72(1), 1–19

Bertrand, M., E. Duflo, and S. Mullainathan (2004) 'How much should we trust differences-in-differences estimates?,' *The Quarterly journal of economics* 119(1), 249–275

Borusyak, K., and X. Jaravel (2017) 'Revisiting event study designs,' *Available at SSRN 2826228*

Borusyak, K., X. Jaravel, and J. Spiess (2024) 'Revisiting event-study designs: robust and efficient estimation,' *Review of Economic Studies* 91(6), 3253–3285

Bray, B. (2025) 'Estimating the effectiveness of the city of calgary's response to the bearspaw south feeder main rupture,' Mimeo

Caetano, C., and B. Callaway (2024) 'Difference-in-differences when parallel trends holds conditional on covariates,' *arXiv preprint arXiv:2406.15288*

Caetano, C., B. Callaway, S. Payne, and H. S. Rodrigues (2022) 'Difference in differences with time-varying covariates,' *arXiv preprint arXiv:2202.02903*

Callaway, B., and P. H. Sant'Anna (2021) 'Difference-in-differences with multiple time periods,' *Journal of Econometrics* 225(2), 200–230

Card, D., J. Rothstein, and M. Yi (2024) 'Industry wage differentials: A firm-based approach,' *Journal of Labor Economics* 42(S1), S11–S59

De Chaisemartin, C., and X. d'Haultfoeuille (2020a) 'Two-way fixed effects estimators with heterogeneous treatment effects,' *American Economic Review* 110(9), 2964–2996

——— (2023) 'Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey,' *The Econometrics Journal* 26(3), C1–C30

Deb, P., E. C. Norton, J. M. Wooldridge, and J. E. Zabel (2024) 'A flexible, heterogeneous treatment effects difference-in-differences estimator for repeated cross-sections,' Technical report, National Bureau of Economic Research

Denning, J. T., E. R. Eide, K. J. Mumford, R. W. Patterson, and M. Warnick (2022) 'Why have college completion rates increased?,' *American Economic Journal: Applied Economics* 14(3), 1–29

Fortin, N. M. (2006) 'Higher-education policies and the college wage premium: Cross-state evidence from the 1990s,' *American Economic Review* 96(4), 959–987

Goodman-Bacon, A. (2021) 'Difference-in-differences with variation in treatment timing,' *Journal of Econometrics* 225(2), 254–277

Heckman, J. J., H. Ichimura, and P. E. Todd (1997) 'Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme,' *The review of economic studies* 64(4), 605–654

Horowitz, J. (2018) 'Relative education and the advantage of a college degree,' *American Sociological Review* 83(4), 771–801

Karim, S., E. Strumpf, N. Austin, and M. Webb (2025a) 'Which policy works and where? estimation and inference of state level treatment effects using difference-in-differences,' in *Slides presented at the Canadian Economics Association Annual Conference, accessed July*, volume 11, p. 2025

Karim, S., E. Strumpf, N. Austin, and M. D. Webb (2025b) 'Which policy works and where? estimation and inference of state level treatment effects using difference-in-differences,' Slides presented at the Canadian Economics Association Annual Conference, accessed July 11, 2025

Karim, S., M. D. Webb, N. Austin, and E. Strumpf (2024) 'Difference-in-differences with unpoolable data,' *arXiv preprint arXiv:2403.15910*

Mincer, J. (1958) 'Investment in human capital and personal income distribution,' *Journal of political economy* 66(4), 281–302

O'Neill, S., N. Kreif, R. Grieve, M. Sutton, and J. S. Sekhon (2016) 'Estimating causal effects: considering three alternatives to difference-in-differences estimation,' *Health Services and Outcomes Research Methodology* 16, 1–21

Rambachan, A., and J. Roth (2023) 'A more credible approach to parallel trends,' *Review of Economic Studies* 90(5), 2555–2591

Rios-Avila, F., P. H. Sant'Anna, and B. Callaway (2021) 'CSDID: Stata module for the estimation of Difference-in-Difference models with multiple time periods,' Statistical Software Components, Boston College Department of Economics

Roodman, D. (2025) 'Julia as a universal platform for statistical software development,' *The Stata Journal* 25(2), 255–284

Roth, J., P. H. Sant'Anna, A. Bilinski, and J. Poe (2022) 'What's trending in difference-in-differences? a synthesis of the recent econometrics literature,' *arXiv preprint arXiv:2201.01194*

Sant'Anna, P. H., and J. Zhao (2020) 'Doubly robust difference-in-differences estimators,' *Journal of Econometrics* 219(1), 101–122

Sun, L., and S. Abraham (2021) 'Estimating dynamic treatment effects in event studies with heterogeneous treatment effects,' *Journal of Econometrics* 225(2), 175–199

A Proof of necessary conditions for parallel trends without covariates

The parallel trends assumption can be written as:

$$\mathbb{E}[Y_2(0) - Y_1(0) \mid s = A] = \mathbb{E}[Y_2(0) - Y_1(0) \mid s = B]. \tag{1}$$

When $\gamma_{s,t}^{\circ} = 0$, Equation (5) can be written as:

$$Y_{i,s,t}(0) = \sum_{s,t} \theta_{s,t}^{\circ} Z_{i,s,t} + \epsilon_{i,s,t} \tag{2}$$

Taking the expectation on either side conditional on S, we can write:

$$\mathbb{E}[Y_t(0) \mid S] = \theta_{s,t}^{\circ} \mathbb{E}[Z_t \mid S]. \tag{3}$$

In the above equation, $\mathbb{E}[\epsilon_{i,s,t} \mid S] = 0$, and both the i and the s subscripts are absorbed by the expectation. Since there is just one region and one period in these conditional expectation, we can drop the summation as well, and since Z_t are indicators, $\mathbb{E}[Z_t \mid S] = 1$. Therefore, we can simply Equation (3) as:

$$\mathbb{E}[Y_t(0) \mid S] = \theta_{s,t}^{\circ}. \tag{4}$$

We can plug in versions of Equation (3) for each potential outcome in Equation (1) and get:

$$\theta_{A,2}^{\circ} - \theta_{A,1}^{\circ} = \theta_{B,2}^{\circ} - \theta_{B,1}^{\circ}. \tag{5}$$

B Proof of necessary conditions for parallel trends with covariates

The conditional parallel trends assumption can be written as:

$$
\begin{aligned}
&\mathbb{E}[Y_2(0) \mid s = A, X_{A,2}] - \mathbb{E}[Y_1(0) \mid s = A, X_{A,1}] \\
=&\mathbb{E}[Y_2(0) \mid s = B, X_{B,2}] - \mathbb{E}[Y_1(0) \mid s = B, X_{B,1}].
\end{aligned} \tag{6}
$$

When $\gamma_{s,t}^{\circ} \neq 0$, Equation (5) can be written as:

$$Y_{i,s,t}(0) = \sum_{s,t} \theta_{s,t}^{\circ} Z_{i,s,t} + \gamma_{s,t}^{\circ} X_{i,s,t} + \epsilon_{i,s,t} \tag{7}$$

Taking the expectation on either side conditional on S and $X_{s,t}$, we can write:

$$\mathbb{E}[Y_t(0) \mid S, X_{s,t}] = \sum_{s,t} \theta_{s,t}^{\circ} \mathbb{E}[Z_t \mid S, X_{s,t}] + \gamma_{s,t}^{\circ} \mathbb{E}[X_{s,t} \mid S, X_{s,t}]. \tag{8}$$

In the above equation, $\mathbb{E}[\epsilon_{i,s,t} \mid S] = 0$, and both the i and the s subscripts are absorbed by the expectation. Since there is just one region and one period in these conditional expectations, we can drop the summation as well, and since Z_t are indicators, $\mathbb{E}[Z_t \mid S] = 1$. Lastly, $\mathbb{E}[X_{s,t} \mid S, X_{s,t}] = \mathbb{E}[X_t \mid S]$ since we are conditioning $X_{s,t}$ on itself. Therefore, we can simply Equation (8) as:

$$\mathbb{E}[Y_t(0) \mid S] = \theta_{s,t}^{\circ} + \gamma_{s,t}^{\circ} \mathbb{E}[X_t \mid S]. \tag{9}$$

We can plug in versions of Equation (9) for each potential outcome in Equation (7) as follows:

$$\theta_{A,2}^{\circ} + \gamma_{A,2}^{\circ} \mathbb{E}[X_2 \mid S = A] - \theta_{A,1}^{\circ} - \gamma_{A,1}^{\circ} \mathbb{E}[X_1 \mid S = A]$$
$$= \theta_{B,2}^{\circ} + \gamma_{B,2}^{\circ} \mathbb{E}[X_2 \mid S = B] - \theta_{B,1}^{\circ} - \gamma_{B,1}^{\circ} \mathbb{E}[X_1 \mid S = B] \tag{10}$$

The equality in Equation (10) is the following necessary conditions are satisfied:

$$\textbf{Condition 1: } \theta_{A,2}^{\circ} - \theta_{A,1}^{\circ} = \theta_{B,2}^{\circ} - \theta_{B,1}^{\circ}. \tag{11}$$

$$\textbf{Condition 2: } \gamma_{A,2}^{\circ} \mathbb{E}[X_{A,2}(0) \mid s = A] - \gamma_{A,1}^{\circ} \mathbb{E}[X_{A,1}(0) \mid s = A]$$
$$= \gamma_{B,2}^{\circ} \mathbb{E}[X_{B,2}(0) \mid s = B] - \gamma_{B,1}^{\circ} \mathbb{E}[X_{B,1}(0) \mid s = B]. \tag{12}$$

C Proof of necessary conditions for intersection trends with covariates

The Intersection Parallel Trends can be written as:

$$\mathbb{E}[\tilde{Y}_2(0) \mid s = A, X_{A,2}] - \mathbb{E}[\tilde{Y}_1(0) \mid s = A, X_{A,1}]$$
$$= \mathbb{E}[\tilde{Y}_2(0) \mid s = B, X_{B,2}] - \mathbb{E}[\tilde{Y}_1(0) \mid s = B, X_{B,1}]. \tag{13}$$

where $\tilde{Y}_{i,s,t}(0)$ denotes the residuals from regressing $Y_{i,s,t}(0)$ on the covariates:

$$\tilde{Y}_{i,s,t}(0) = Y_{i,s,t}(0) - \gamma_{s,t} X_{i,s,t}(0). \tag{14}$$

Taking the expectation on either side conditional on S and $X_{s,t}$, we can write:

$$\mathbb{E}[\tilde{Y}_{s,t}(0) \mid S, X_{s,t}] = \mathbb{E}[Y_t(0) \mid S, X_{s,t}] - \gamma_{s,t} \mathbb{E}[X_t \mid S]. \tag{15}$$

In the above equation, $\mathbb{E}[\epsilon_{i,s,t} \mid S] = 0$, and both the i and the s subscripts are absorbed by the expectation. $\mathbb{E}[X_{s,t} \mid S, X_{s,t}] = \mathbb{E}[X_t \mid S]$ since we are conditioning $X_{s,t}$ on itself. Replacing $\mathbb{E}[Y_t(0) \mid S, X_{s,t}]$ from Equation (9) into Equation (15), we get:

$$\mathbb{E}[\tilde{Y}_t(0) \mid S, X_{s,t}] = \theta_{s,t}^{\circ}. \tag{16}$$

We can plug in versions of Equation (16) for each potential outcome in Equation (13) and get:

$$\theta_{A,2}^{\circ} - \theta_{A,1}^{\circ} = \theta_{B,2}^{\circ} - \theta_{B,1}^{\circ}. \tag{17}$$

D Intersection Difference-in-differences proof of consistency

In the first step of the two-way Intersection Difference-in-differences, we run the following regression:

$$Y_{i,s,t} = \sum_s \sum_t \lambda_{s,t} I(s,t) + \gamma_{s,t} I(s) I(t) X_{i,s,t} + \epsilon_{i,s,t}, \tag{18}$$

For simplicity, we continue to assume that there is a single time-varying covariate $X_{i,s,t}$ required for intersection parallel trends. Taking the expectation conditional on $s \in \mathbf{s}^t$ and $X_{s,t}$ of the regression shown in Equation (18) and simplifying, we get:

$$E[Y_t \mid s \in \mathbf{s}^t, X_{s,t}] = \lambda_{s,t} + \gamma_{s,t} E[X_t \mid s \in \mathbf{s}^t] \tag{19}$$

In the above equation, $\mathbb{E}[\epsilon_{i,s,t} \mid S, X_{s,t}] = 0$, and both the i and the s subscripts are absorbed by the expectation. The indicators $I(s,t)$, $I(s)$ and $I(t)$ are excluded because they take the value of one for all observations in the sample under consideration. Moreover, $E[X_{s,t} \mid s \in \mathbf{s}^t, X_{s,t}] = E[X_t \mid s \in \mathbf{s}^t]$ since $X_{s,t}$ is conditioned on itself. After re-arranging, $\lambda_{s,t}$ can be expressed as:

$$\lambda_{s,t} = E[Y_t \mid s \in \mathbf{s}^t, X_{s,t}] - \gamma_{s,t} E[X_t \mid s \in \mathbf{s}^t, X_{s,t}] \tag{20}$$

Similarly, we can derive λ_{s,t^s-1}, $\lambda_{s',t}$, λ_{s',t^s-1}:

$$\lambda_{s',t} = E[Y_t \mid s' \in \mathbf{s}^c, X_{s',t}] - \gamma_{s',t} E[X_t \mid s' \in \mathbf{s}^c, X_{s',t}] \tag{21}$$

$$\lambda_{s,t^s-1} = E[Y_{t^s-1} \mid s \in \mathbf{s}^t, X_{s,t^s-1}] - \gamma_{s,t^s-1} E[X_{t^s-1} \mid s \in \mathbf{s}^t, X_{s,t^s-1}] \tag{22}$$

$$\lambda_{s',t^s-1} = E[Y_{t^s-1} \mid s' \in \mathbf{s}^c, X_{s',t^s-1}] - \gamma_{s',t^s-1}E[X_{t^s-1} \mid s' \in \mathbf{s}^c, X_{s',t^s-1}] \tag{23}$$

Here, $s' \in \mathbf{s}^c$ is a relevant control group for period group s in period t, which can either be a region which is never treated, or a region which is never treated and not-yet treated. Imposing the **SUTVA assumption** formally stated below, we can write observed outcomes in terms of their observed potential outcome. Since only region s is treated and s' is untreated in period t, we can re-write Equations (20) and (21) as:

$$\lambda_{s,t} = E[Y_t(1) \mid s \in \mathbf{s}^t, X_{s,t}] - \gamma_{s,t}E[X_t \mid s \in \mathbf{s}^t, X_{s,t}] \tag{24}$$

$$\lambda_{s',t} = E[Y_t(0) \mid s' \in \mathbf{s}^c, X_{s',t}] - \gamma_{s',t}E[X_t \mid s' \in \mathbf{s}^c, X_{s',t}] \tag{25}$$

Assumption 1 (Stable Unit Treatment Value Assumption (SUTVA)). Observed outcomes at time t are realized as:

$$Y_{i,s,t} = 1\{D_i = 1\}Y_{i,s,t}(1) + 1\{D_i = 0\}Y_{i,s,t}(0) \tag{26}$$

Since both regions s and s' are untreated in period $t^s - 1$, we can re-write Equations (22) and (23) as:

$$\lambda_{s,t^s-1} = E[Y_{t^s-1}(0) \mid s \in \mathbf{s}^t, X_{s,t^s-1}] - \gamma_{s,t^s-1}E[X_{t^s-1} \mid s \in \mathbf{s}^t, X_{s,t^s-1}] \tag{27}$$

$$\lambda_{s',t^s-1} = E[Y_{t^s-1}(0) \mid s' \in \mathbf{s}^c, X_{s',t^s-1}] - \gamma_{s',t^s-1}E[X_{t^s-1} \mid s' \in \mathbf{s}^c, X_{s',t^s-1}] \tag{28}$$

In the second step of the two-way DID-INT, we hypothesize that the estimand of the $ATT(s,t)$ using DID-INT is:

$$\theta_{s,t} = \Delta_s - \Delta_{s'} \tag{29}$$

Here,

$$\Delta_s = \left(\lambda_{s,t} - \lambda_{s,t^s-1}\right) \tag{30}$$

and

$$\Delta_{s'} = \left(\lambda_{s',t} - \lambda_{s',t^s-1}\right) \tag{31}$$

Plugging in the corresponding values from Equations (24), (25) in Equation (30), we get:

$$\Delta_s = \left(E[Y_t(1) \mid s \in \mathbf{s}^t, X_{s,t}] - \gamma_{s,t} E[X_t \mid s \in \mathbf{s}^t, X_{s,t}] \right)$$
$$- \left(E[Y_{t^s-1}(0) \mid s \in \mathbf{s}^t, X_{s,t^s-1}] - \gamma_{s,t^s-1} E[X_{t^s-1} \mid s \in \mathbf{s}^t, X_{s,t^s-1}] \right) \tag{32}$$

Adding and subtracting $E[Y_t(0) \mid s \in \mathbf{s}^t, X_{s,t}]$ in Equation (32), we get:

$$\Delta_s = \left(E[Y_t(1) \mid s \in \mathbf{s}^t, X_{s,t}] - E[Y_t(0) \mid s \in \mathbf{s}^t, X_{s,t}] + E[Y_t(0) \mid s \in \mathbf{s}^t, X_{s,t}] - \gamma_{s,t} E[X_t \mid s \in \mathbf{s}^t, X_{s,t}] \right)$$
$$- \left(E[Y_{t^s-1}(0) \mid s \in \mathbf{s}^t, X_{s,t^s-1}] - \gamma_{s,t^s-1} E[X_{t^s-1} \mid s \in \mathbf{s}^t, X_{s,t^s-1}] \right)$$

We know, $E[Y_t(1) \mid s \in \mathbf{s}^t, X_{s,t}] - E[Y_t(0) \mid s \in \mathbf{s}^t, X_{s,t}] = \tau$ under **homogeneous treatment effects**. From Equation (15), we also know that $\mathbb{E}[\tilde{Y}_t(0) \mid S, X_{s,t}] = \mathbb{E}[Y_t(0) \mid S, X_{s,t}] - \gamma_{s,t} \mathbb{E}[X_t \mid S]$. Therefore, we can further simplify the above equation as:

$$\Delta_s = \tau + \left(\mathbb{E}[\tilde{Y}_t(0) \mid s \in \mathbf{s}^t, X_{s,t}] \right) - \left(\mathbb{E}[\tilde{Y}_{t^s-1}(0) \mid s \in \mathbf{s}^t, X_{s,t^s-1}] \right) \tag{33}$$

Plugging in the corresponding values from Equations (27), (28) in Equation (31), and simplifying we get:

$$\Delta_{s'} = \left(\mathbb{E}[\tilde{Y}_t(0) \mid s' \in \mathbf{s}^c, X_{s',t}] \right) - \left(\mathbb{E}[\tilde{Y}_{t^s-1}(0) \mid s' \in \mathbf{s}^c, X_{s',t^s-1}] \right) \tag{34}$$

Imposing the **intersection parallel trends assumption**, we know that:

$$\left(\mathbb{E}[\tilde{Y}_t(0) \mid s \in \mathbf{s}^t, X_{s,t}] \right) - \left(\mathbb{E}[\tilde{Y}_{t^s-1}(0) \mid s \in \mathbf{s}^t, X_{s,t^s-1}] \right)$$
$$= \left(\mathbb{E}[\tilde{Y}_t(0) \mid s' \in \mathbf{s}^c, X_{s',t}] \right) - \left(\mathbb{E}[\tilde{Y}_{t^s-1}(0) \mid s' \in \mathbf{s}^c, X_{s',t^s-1}] \right) \tag{35}$$

Therefore, a difference between Equations (34) and (33) gives:

$$\theta_{s,t} = \Delta_s - \Delta_{s'} = \tau \tag{36}$$

E Two-way fixed effects proof of consistency

In this proof, we assume that the **conditional parallel trends** hold, but the two-way CCC assumption is violated. In other words, the **necessary conditions** shown in Theorem 2 holds, but the **sufficient conditions** do not. The two-way fixed effects model can be written as:

$$Y_{i,s,t} = \alpha_s + \delta_t + \beta^{DD} D_{i,s,t} + \gamma X_{i,s,t} + \epsilon_{i,s,t} \tag{37}$$

The key parameter of interest is β^{DD}, which we derive in this section. For simplicity, we assume that there is a single time-varying covariate $X_{i,s,t}$ required for **conditional parallel trends** to hold. **It is already well-known in the literature that β^{DD} is inconsistent when conditional parallel trends is violated.** Therefore, we will not analyze the case where conditional parallel trends does not hold. Taking the expectation conditional on S and $X_{s,t}$ of the regression shown in Equation (37) and simplifying, we get:

$$\mathbb{E}[Y_t \mid S, X_{s,t}] = \alpha_s + \delta_t + \beta^{DD}\mathbb{E}[D_t \mid S, X_{s,t}] + \gamma\mathbb{E}[X_t \mid S] \tag{38}$$

In the above equation, $\mathbb{E}[\epsilon_{i,s,t} \mid S, X_{s,t}] = 0$, and both the i and the s subscripts are absorbed by the expectation. $E[X_{s,t} \mid S, X_{s,t}] = E[X_t \mid S]$ since $X_{s,t}$ is conditioned on itself.

For the treated region in the post-intervention period, Equation (38) can be expressed as:

$$\mathbb{E}[Y_t(1) \mid s \in \mathbf{s}^t, X_{s,t}] = \theta_{s,t} + \beta^{DD} + \gamma\mathbb{E}[X_t \mid s \in \mathbf{s}^t] \tag{39}$$

In the above expression, we can write $\mathbb{E}[Y_t \mid s \in \mathbf{s}^t, X_{s,t}] = \mathbb{E}[Y_t(1) \mid s \in \mathbf{s}^t, X_{s,t}]$ imposing **SUTVA**, since we observe the treated potential outcome for the treated region in the post-intervention period. Similarly, $\mathbb{E}[D_t \mid s \in \mathbf{s}^t, X_{s,t}] = 1$ for all units in the treated region in the post-intervention period. We can also write $\theta_{s,t} \equiv \alpha_s + \delta_t$, based on the potential outcome model shown in Equation (5). For the treated region in the pre-intervention period, Equation (38) can be expressed as:

$$\mathbb{E}[Y_{t'}(0) \mid s \in \mathbf{s}^t, X_{s,t'}] = \theta_{s,t'} + \gamma\mathbb{E}[X_{t'} \mid s \in \mathbf{s}^t] \tag{40}$$

In the above expression, the pre-intervention period is no longer the period right before treatment as used in Appendix D. Instead, all the years before treatment is bunched together as the pre-intervention period, which we index by t'. To intuitively understand why only the period right before treatment is used as the pre-period for DID methods in staggered adoption designs, please refer to Callaway and Sant'Anna (2021). The two-way fixed effect model uses data for all the periods before treatment as the pre-treatment period, but is inconsistent in staggered designs with heterogeneous treatment effects due to negative weighting issues and forbidden comparisons (Goodman-Bacon, 2021). Imposing SUTVA, we can write $\mathbb{E}[Y_t \mid s, X_{s,t}] = \mathbb{E}[Y_t(0) \mid s, X_{s,t}]$, since we observe the untreated potential outcome for the treated region in the pre-intervention period. Similarly, $\mathbb{E}[D_t \mid s, X_{s,t}] = 0$ for all units in the treated region in the pre-intervention period, which is why β^{DD} no longer appears in the equation. We can similarly derive expressions for the control group in the post and pre-periods,

shown in Equations (41) and (42) respectively.

$$\mathbb{E}[Y_t(0) \mid s' \in \mathbf{s}^c, X_{s,t}] = \theta_{s',t} + \gamma\mathbb{E}[X_t \mid s' \in \mathbf{s}^c] \tag{41}$$

$$\mathbb{E}[Y_{t'}(0) \mid s' \in \mathbf{s}^c, X_{s',t'}] = \theta_{s',t'} + \gamma\mathbb{E}[X_{t'} \mid s \in \mathbf{s}'^c] \tag{42}$$

In the above expressions, we can write $\mathbb{E}[Y_t \mid s', X_{s',t}] = \mathbb{E}[Y_t(0) \mid s', X_{s',t}]$, since we observe the untreated potential outcome for the control regions in both periods. Taking a difference of Equations (39) and (40), we get:

$$\mathbb{E}[Y_t(1) \mid s \in \mathbf{s}^t, X_{s,t}] - \mathbb{E}[Y_{t'}(0) \mid s \in \mathbf{s}^t, X_{s,t'}] = \theta_{s,t} - \theta_{s,t'} + \beta^{DD} + \gamma\left(\mathbb{E}[X_t \mid s \in \mathbf{s}^t] - \mathbb{E}[X_{t'} \mid s \in \mathbf{s}^t]\right)$$

Adding and subtracting $\mathbb{E}[Y_t(0) \mid s \in \mathbf{s}^t, X_{s,t}]$ in the above expression, and setting $\mathbb{E}[Y_t(1) \mid s \in \mathbf{s}^t, X_{s,t}] - \mathbb{E}[Y_t(0) \mid s \in \mathbf{s}^t, X_{s,t}] = \tau$, we get:

$$\begin{aligned}
\tau + \mathbb{E}[Y_t(0) \mid s \in \mathbf{s}^t, X_{s,t}] &- \mathbb{E}[Y_{t'}(0) \mid s \in \mathbf{s}^t, X_{s,t'}] \\
&= \theta_{s,t} - \theta_{s,t'} + \beta^{DD} + \gamma\left(\mathbb{E}[X_t \mid s \in \mathbf{s}^t] - \mathbb{E}[X_t \mid s \in \mathbf{s}^t]\right)
\end{aligned} \tag{43}$$

Similarly, taking a difference of Equations (41) and (42), we get:

$$\begin{aligned}
\mathbb{E}[Y_t(0) \mid s' \in \mathbf{s}^c, X_{s,t}] &- \mathbb{E}[Y_{t'}(0) \mid s' \in \mathbf{s}^c, X_{s',t'}] \\
&= \theta_{s',t} - \theta_{s',t'} + \gamma\left(\mathbb{E}[X_t \mid s' \in \mathbf{s}^c] - \mathbb{E}[X_{t'} \mid s \in \mathbf{s}'^c]\right)
\end{aligned} \tag{44}$$

Taking a difference of Equations (43) and (44), and simplifying, we get:

$$\tau = \beta^{DD} + \gamma\left(\mathbb{E}[X_t \mid s \in \mathbf{s}^t] - \mathbb{E}[X_t \mid s \in \mathbf{s}^t]\right) - \gamma\left(\mathbb{E}[X_t \mid s' \in \mathbf{s}^c] - \mathbb{E}[X_{t'} \mid s \in \mathbf{s}'^c]\right) \tag{45}$$

Since the **conditional parallel trends assumption holds**, the above expression sets:

$$\left(\mathbb{E}[Y_t(0) \mid s \in \mathbf{s}^t, X_{s,t}] - \mathbb{E}[Y_{t'}(0) \mid s \in \mathbf{s}^t, X_{s,t'}]\right) - \left(\mathbb{E}[Y_t(0) \mid s' \in \mathbf{s}^c, X_{s,t}] - \mathbb{E}[Y_{t'}(0) \mid s' \in \mathbf{s}^c, X_{s',t'}]\right) = 0$$

and

$$\left(\theta_{s,t} - \theta_{s,t'}\right) - \left(\theta_{s',t} - \theta_{s',t'}\right) = 0$$

The second expression is one of the **necessary conditions** for conditional parallel trends to hold. The expression in Equation (61) implies that $\beta^{DD} \neq \tau$, and is therefore inconsistent. This inconsistency arises from the term shown in Equation (46), which emerges from modeling $X_{i,s,t}$'s with a common coefficient γ, whereas in reality, the coefficients vary across regions and

periods.

$$\gamma\left(\mathbb{E}[X_t \mid s \in \mathbf{s}^t] - \mathbb{E}[X_t \mid s \in \mathbf{s}^t]\right) - \gamma\left(\mathbb{E}[X_t \mid s' \in \mathbf{s}^c] - \mathbb{E}[X_{t'} \mid s \in \mathbf{s}'^c]\right) \quad (46)$$

It is important to note that when **two-way CCC** and the parallel trends in covariates assumption holds (sufficient conditions 1 and 2 in Theorem (2)), the above expression will be 0. In that case, β^{DD} will be consistent. **Conditional parallel trends may still hold if the two sufficient conditions are not met**.

F   Modified two-way fixed effects proof of consistency

In this proof, we consider two cases. In the first case, we assume that intersection parallel trends hold. In the second case, we assume that conditional parallel trends holds, such that two-way CCC is violated but **condition 2** holds. The *modified* two-way fixed effects model can be written as:

$$Y_{i,s,t} = \alpha_s + \delta_t + \beta^{DD}_{modified}D_{i,s,t} + \gamma_{s,t}I(s)I(t)X_{i,s,t} + \epsilon_{i,s,t} \quad (47)$$

The key parameter of interest is $\beta^{DD}_{modified}$, which we derive in this section. For simplicity, we assume that there is a single time-varying covariate $X_{i,s,t}$ required for either **intersection parallel trends** for the first case or **conditional parallel trends** for the second case to hold. Taking the expectation conditional on S and $X_{s,t}$ of the regression shown in Equation (47) and simplifying, we get:

$$\mathbb{E}[Y_t \mid S, X_{s,t}] = \alpha_s + \delta_t + \beta^{DD}\mathbb{E}[D_t \mid S, X_{s,t}] + \gamma_{s,t}\mathbb{E}[X_{i,s,t} \mid S] \quad (48)$$

In the above equation, $\mathbb{E}[\epsilon_{i,s,t} \mid S, X_{s,t}] = 0$, and both the i and the s subscripts are absorbed by the expectation. $E[X_{s,t} \mid S, X_{s,t}] = E[X_t \mid S]$ since $X_{s,t}$ is conditioned on itself. The indicators $I(s,t)$, $I(s)$ and $I(t)$ are excluded because they take the value of one for all observations in the sample under consideration. For the treated region in the post-intervention period, Equation (48) can be expressed as:

$$\mathbb{E}[Y_t(1) \mid s \in \mathbf{s}^t, X_{s,t}] = \theta_{s,t} + \beta^{DD} + \gamma_{s,t}\mathbb{E}[X_t \mid s \in \mathbf{s}^t] \quad (49)$$

In the above expression, we can write $\mathbb{E}[Y_t \mid s \in \mathbf{s}^t, X_{s,t}] = \mathbb{E}[Y_t(1) \mid s \in \mathbf{s}^t, X_{s,t}]$ imposing **SUTVA**, since we observe the treated potential outcome for the treated region in the post-intervention period. Similarly, $\mathbb{E}[D_t \mid s \in \mathbf{s}^t, X_{s,t}] = 1$ for all units in the treated region in the post-intervention period. We can also write $\theta_{s,t} = \alpha_s + \delta_t$, based on the potential outcome model shown in Equation (5).

Let us analyze the **first case** where **intersection parallel trends** hold. Adding and subtracting $\mathbb{E}[Y_t(1) \mid s \in \mathbf{s}^t, X_{s,t}]$ in the left hand side of Equation (49), we get:

$$\mathbb{E}[Y_t(1) \mid s \in \mathbf{s}^t, X_{s,t}] - \mathbb{E}[Y_t(0) \mid s \in \mathbf{s}^t, X_{s,t}] + \mathbb{E}[Y_t(0) \mid s \in \mathbf{s}^t, X_{s,t}] = \theta_{s,t} + \beta^{DD} + \gamma_{s,t}\mathbb{E}[X_t \mid s \in \mathbf{s}^t]$$

This can be further simplified to:

$$\tau + \mathbb{E}[\tilde{Y}_t(0) \mid s \in \mathbf{s}^t, X_{s,t}] = \theta_{s,t} + \beta^{DD}_{modified} \tag{50}$$

This follows from $\mathbb{E}[Y_t(1) \mid s \in \mathbf{s}^t, X_{s,t}] - \mathbb{E}[Y_t(0) \mid s \in \mathbf{s}^t, X_{s,t}] = \tau$ under **homogeneous treatment effects**, and the definition of $\tilde{Y}_t(0)$ from Equation (24). For the treated region in the pre-intervention period, Equation (48) can be expressed as:

$$\mathbb{E}[\tilde{Y}_{t'}(0) \mid s \in \mathbf{s}^t, X_{s,t'}] = \theta_{s,t'} \tag{51}$$

Imposing SUTVA, we can write $\mathbb{E}[Y_t \mid s, X_{s,t}] = \mathbb{E}[Y_t(0) \mid s, X_{s,t}]$, since we observe the untreated potential outcome for the treated region in the pre-intervention period. Similarly, $\mathbb{E}[D_t \mid s, X_{s,t}] = 0$ for all units in the treated region in the pre-intervention period, which is why $\beta^{DD}_{modified}$ no longer appears in the equation. We can similarly derive expressions for the control group in the post and pre-periods, shown in Equations (52) and (53) respectively.

$$\mathbb{E}[\tilde{Y}_t(0) \mid s' \in \mathbf{s}^c, X_{s,t}] = \theta_{s',t} \tag{52}$$

$$\mathbb{E}[\tilde{Y}_{t'}(0) \mid s' \in \mathbf{s}^c, X_{s',t'}] = \theta_{s',t'} \tag{53}$$

Taking a difference-in-differences of Equations (50), (51), (52) and (53), we get:

$$\tau + \left(\mathbb{E}[\tilde{Y}_t(0) \mid s \in \mathbf{s}^t, X_{s,t}] - \mathbb{E}[\tilde{Y}_{t'}(0) \mid s \in \mathbf{s}^t, X_{s,t'}]\right) - \left(\mathbb{E}[\tilde{Y}_t(0) \mid s' \in \mathbf{s}^c, X_{s,t}] - \mathbb{E}[\tilde{Y}_{t'}(0) \mid s' \in \mathbf{s}^c, X_{s',t'}]\right)$$
$$= \left(\theta_{s,t} - \theta_{s,t'}\right) - \left(\theta_{s',t} - \theta_{s',t'}\right) + \beta^{DD}_{modified} \tag{54}$$

Since the **Intersection parallel trends assumption holds**, the above expression sets:

$$\left(\mathbb{E}[\tilde{Y}_t(0) \mid s \in \mathbf{s}^t, X_{s,t}] - \mathbb{E}[\tilde{Y}_{t'}(0) \mid s \in \mathbf{s}^t, X_{s,t'}]\right) - \left(\mathbb{E}[\tilde{Y}_t(0) \mid s' \in \mathbf{s}^c, X_{s,t}] - \mathbb{E}[\tilde{Y}_{t'}(0) \mid s' \in \mathbf{s}^c, X_{s',t'}]\right) = 0$$

and

$$\left(\theta_{s,t} - \theta_{s,t'}\right) - \left(\theta_{s',t} - \theta_{s',t'}\right) = 0$$

The second expression is the **necessary conditions** for intersection parallel trends to hold.

Therefore, Equation (54) can be written as:

$$\tau = \beta^{DD}_{modified} \tag{55}$$

Now, let us analyze the **second case** where **conditional parallel trends** hold, but two-way CCC is violated. We leave the expression for the treated region in the post intervention period shown in Equation (49) as it is. For the treated region in the pre-intervention period, Equation (49) can be expressed as:

$$\mathbb{E}[Y_{t'}(0) \mid s \in \mathbf{s}^t, X_{s,t'}] = \theta_{s,t'} + \gamma_{s,t'} \mathbb{E}[X_{t'} \mid s \in \mathbf{s}^t] \tag{56}$$

We can similarly derive expressions for the control group in the post and pre-periods, shown in Equations (57) and (58) respectively.

$$\mathbb{E}[Y_t(0) \mid s' \in \mathbf{s}^c, X_{s,t}] = \theta_{s',t} + \gamma_{s',t} \mathbb{E}[X_t \mid s' \in \mathbf{s}^c] \tag{57}$$

$$\mathbb{E}[Y_{t'}(0) \mid s' \in \mathbf{s}^c, X_{s',t'}] = \theta_{s',t'} + \gamma_{s',t'} \mathbb{E}[X_{t'} \mid s' \in \mathbf{s}'^c] \tag{58}$$

Taking a difference of Equations (49) and (56), we get:

$$\mathbb{E}[Y_t(1) \mid s \in \mathbf{s}^t, X_{s,t}] - \mathbb{E}[Y_{t'}(0) \mid s \in \mathbf{s}^t, X_{s,t'}] = \theta_{s,t} - \theta_{s,t'} + \beta^{DD}_{modified} + \gamma_{s,t} \mathbb{E}[X_t \mid s \in \mathbf{s}^t] - \gamma_{s,t'} \mathbb{E}[X_{t'} \mid s \in \mathbf{s}^t]$$

Adding and subtracting $\mathbb{E}[Y_t(0) \mid s \in \mathbf{s}^t, X_{s,t}]$ in the above expression, and setting $\mathbb{E}[Y_t(1) \mid s \in \mathbf{s}^t, X_{s,t}] - \mathbb{E}[Y_t(0) \mid s \in \mathbf{s}^t, X_{s,t}] = \tau$, we get:

$$\begin{aligned}
\tau + \mathbb{E}[Y_t(0) \mid s \in \mathbf{s}^t, X_{s,t}] &- \mathbb{E}[Y_{t'}(0) \mid s \in \mathbf{s}^t, X_{s,t'}] \\
&= \theta_{s,t} - \theta_{s,t'} + \beta^{DD}_{modified} + \gamma_{s,t} \mathbb{E}[X_t \mid s \in \mathbf{s}^t] - \gamma_{s,t'} \mathbb{E}[X_t \mid s \in \mathbf{s}^t]
\end{aligned} \tag{59}$$

Similarly, taking a difference of Equations (52) and (53), we get:

$$\begin{aligned}
\mathbb{E}[Y_t(0) \mid s' \in \mathbf{s}^c, X_{s,t}] &- \mathbb{E}[Y_{t'}(0) \mid s' \in \mathbf{s}^c, X_{s',t'}] \\
&= \theta_{s',t} - \theta_{s',t'} + \gamma_{s',t} \mathbb{E}[X_t \mid s' \in \mathbf{s}^c] - \gamma_{s',t'} \mathbb{E}[X_{t'} \mid s \in \mathbf{s}'^c]
\end{aligned} \tag{60}$$

Taking a difference of Equations (59) and (60), and simplifying, we get:

$$\tau = \beta^{DD}_{modified} \tag{61}$$

Since the **conditional parallel trends assumption holds**, the above expression sets:

$$\left(\mathbb{E}[Y_t(0) \mid s \in \mathbf{s}^t, X_{s,t}] - \mathbb{E}[Y_{t'}(0) \mid s \in \mathbf{s}^t, X_{s,t'}] \right) - \left(\mathbb{E}[Y_t(0) \mid s' \in \mathbf{s}^c, X_{s,t}] - \mathbb{E}[Y_{t'}(0) \mid s' \in \mathbf{s}^c, X_{s',t'}] \right) = 0$$

and

$$\left(\theta_{s,t} - \theta_{s,t'}\right) - \left(\theta_{s',t} - \theta_{s',t'}\right) = 0$$

$$\left(\gamma_{s,t}\mathbb{E}[X_t \mid s \in \mathbf{s}^t] - \gamma_{s,t'}\mathbb{E}[X_t \mid s \in \mathbf{s}^t]\right) - \left(\gamma_{s',t}\mathbb{E}[X_t \mid s' \in \mathbf{s}^c] - \gamma_{s',t'}\mathbb{E}[X_{t'} \mid s \in \mathbf{s}'^c]\right)$$

The two expressions above are the **necessary conditions** for conditional parallel trends to hold.

G Doubly Robust Difference-in-differences proof of consistency

In this section, we will analyze whether the DR-DID can identify the key causal parameter of interest τ. To begin, let us first derive the outcome regression component of the above estimand: $E[Y_{i,g',t} - Y_{i,g',g-1}|X_{i,g',t}, X_{i,g',g-1}, G = g']$. An estimate of $E[Y_{i,g',t}|X_{i,g',t}, G = g']$ can be obtained from the fitted values of the following regression:

$$Y_{i,g',t} = \sum_k \gamma_{i,g',t}^k X_{i,g,t}^k + \nu_{i,g',t} \tag{62}$$

Note: The above regression is run using observations in the control group in period t, which is the post intervention period. Similarly, using data for the control group in period $g-1$, which is the pre-intervention period, we can estimate $E[Y_{i,g',g-1}|X_{i,g',g-1}, G = g']$ from the fitted values of the following regression:

$$Y_{i,g',g-1} = \sum_k \gamma_{i,g',g-1}^k X_{i,g,g-1}^k + \nu_{i,g',g-1} \tag{63}$$

The difference between the fitted values from Equations (62) and (63) will be an estimate of the outcome regression component, shown below.

$$E[Y_{i,g',t} - Y_{i,g',g-1}|X_{i,g',t}, X_{i,g',t-1}, G = g'] = \sum_k \gamma_{i,g',t}^k X_{i,g',t}^k - \sum_k \gamma_{i,g',g-1}^k X_{i,g',g-1}^k \tag{64}$$

Since the observed outcomes of the control groups in both periods is the same as the potential outcome of the control group in the absence of treatment, a difference between equation (2) between periods t and $g-1$ is the same as Equation (64). Now, let us derive $Y_{i,g,t} - Y_{i,g,g-1}$ from Equation (47). In period t, the observed outcome of the treated group is the same as the potential outcome of the treated group when treated, as shown in Equation (6). Similarly, the observed outcome of the treated group in period $g-1$ (pre-intervention period) is the same as the potential outcome of the treated group in the absence of treatment, as shown in Equation (2). Therefore, taking a difference of Equation (6) and (2) yields the following:

$$Y_{i,g,t} - Y_{i,g,g-1} = \tau + \sum_k \gamma_{i,g,t}^k X_{i,g,t}^k - \sum_k \gamma_{i,g,g-1}^k X_{i,g,g-1}^k \tag{65}$$

Plugging in Equations (64) and (65) into Equation (47) and re-arranging:

$$E\left[\frac{D}{E[D]}\tau\right] + E\left[\frac{D}{E[D]}\left(\sum_k \gamma_{i,g,t}^k X_{i,g,t}^k - \sum_k \gamma_{i,g,g-1}^k X_{i,g,g-1}^k\right) - \frac{P(X_{i,g,t}^k)(1-D)}{E[D](1-P(X_{i,g,t}^k))}\left(\sum_k \gamma_{i,g',t}^k X_{i,g',t}^k - \sum_k \gamma_{i,g',g-1}^k X_{i,g',g-1}^k\right)\right] \tag{66}$$

Under homogeneous treatment effect assumption, the above equation can be further simplified to:

$$\tau + E\left[\frac{D}{E[D]}\left(\sum_k \gamma_{i,g,t}^k X_{i,g,t}^k - \sum_k \gamma_{i,g,g-1}^k X_{i,g,g-1}^k\right) - \frac{P(X_{i,g,t}^k)(1-D)}{E[D](1-P(X_{i,g,t}^k))}\left(\sum_k \gamma_{i,g',t}^k X_{i,g',t}^k - \sum_k \gamma_{i,g',g-1}^k X_{i,g',g-1}^k\right)\right] \tag{67}$$

Equation (67) shows that, under no additional assumptions on covariates, the estimand of the ATT includes τ, the key parameter of interest and an added bias term. When the CCC assumption holds, and the covariates are time invariant, we can simplify the above expression, as shown in Equation (68).

$$\tau + E\left[\frac{D}{E[D]}\underbrace{\left(\sum_k \gamma^k X_{i,g}^k - \sum_k \gamma^k X_{i,g}^k\right)}_{0} - \frac{P(X_{i,g,t}^k)(1-D)}{E[D](1-P(X_{i,g,t}^k))}\underbrace{\left(\sum_k \gamma^k X_{i,g'}^k - \sum_k \gamma^k X_{i,g'}^k\right)}_{0}\right] = \tau \tag{68}$$

Cancelling out the like terms, and simplifying, we get:

$$\theta^{DRDID} = E[\widehat{\theta}^{DRDID}] = \tau \tag{69}$$

However, the binconsistency persists when time-varying covariates are used, and the CCC assumption holds. This is shown in Equation (70).

$$\tau + E\left[\frac{D}{E[D]}\underbrace{\left(\sum_k \gamma^k X_{i,g,t} - \sum_k \gamma^k X_{i,g,g-1}^k\right)}_{\neq 0} - \frac{P(X_{i,g,t}^k)(1-D)}{E[D](1-P(X_{i,g,t}^k))}\underbrace{\left(\sum_k \gamma^k X_{i,g',t} - \sum_k \gamma^k X_{i,g',g-1}^k\right)}_{\neq 0}\right] \tag{70}$$

$$\therefore \theta^{DRDID} = E[\widehat{\theta}^{DRDID}] \neq \tau \tag{71}$$

The inconsistency is amplified when there are violations of (two-way) CCC in addition to using time varying covariates. This is shown in Equation (72).

$$\tau + E\left[\frac{D}{E[D]}\underbrace{\left(\sum_k \gamma_{i,g,t}^k X_{i,g,t} - \sum_k \gamma_{i,g,g-1}^k X_{i,g,g-1}^k\right)}_{\neq 0} - \frac{P(X_{i,g,t}^k)(1-D)}{E[D](1-P(X_{i,g,t}^k))}\underbrace{\left(\sum_k \gamma_{i,g',t}^k X_{i,g',t} - \sum_k \gamma_{i,g',g-1}^k X_{i,g',g-1}^k\right)}_{\neq 0}\right] \tag{72}$$

$$\therefore \theta^{DRDID} = E[\widehat{\theta}^{DRDID}] = \tau \tag{73}$$

	A	B
1	$\gamma_A^0 + \gamma_1^0$	$\gamma_B^0 + \gamma_1^0$
2	$\gamma_A^0 + \gamma_2^0$	$\gamma_B^0 + \gamma_2^0$

$$f(X_{i,s,t}) = \sum_{s=1}^{S^T} \sum_{k=1}^{K} \gamma_s^k I(s) X_{i,s,t}^k + \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_t^k I(t) X_{i,s,t}^k$$

Figure 19: Example of Two One-Way CCC Violations

H   Two one-way DID-INT

In this section, we introduce an additional way to model covariates in DID-INT, which we call the **two one-way DID-INT**. This version accounts for another possible type of CCC violation, which we call the two one-way CCC. If the two one-way CCC assumption is violated, we recommend researchers to interact the covariates with both the with the $I(s)$ and the $I(t)$ dummies separately, and including both interactions as covariates in the model. Therefore, $f(X_{i,s,t}) = \sum_{s=1}^{S^T} \sum_{k=1}^{K} \gamma_s^k I(s) X_{i,s,t}^k + \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_t^k I(t) X_{i,s,t}^k$, which takes into account two-one way CCC violations.

Assumption 2 (Two One-way Common Causal Covariates)**.** The effect of the covariates is additive and separable, with group-invariant and time-invariant coefficients.

$$\gamma^{s,t} = \gamma^{s',t'} \text{ where, } \{s, s' = 1, 2, ..., S\}; \{t, t' = 1, 2, ..., T\} \text{ \& } s \neq s'; t \neq t'.$$
$$f(\tilde{X}_{i,s,t}) = \sum_k \widehat{\gamma_s^k} I(s) X_{i,s,t}^k + \sum_k \widehat{\gamma_t^k} I(t) X_{i,s,t}^k$$

To explore the performance of the two-way DID-INT and the two-one way DID-INT, we incorporate an additional constructed outcome, denoted by $Y_{i,s,t}^5$, where there are two-one way violations. The region-varying coefficient is obtained from Equation (61) and the time-varying coefficient is obtained from Equations (63). We then generate $Y_{i,s,t}^5$ using the following:

$$Y_{i,s,t}^5 = y_0 + \sum_k \widehat{\gamma_s^k} X_{i,s,t}^k + \sum_k \widehat{\gamma_t^k} X_{i,s,t}^k \quad \text{if group} = s \text{ \& time} = t. \tag{74}$$

For the DGP where there are two-way CCC violations $(Y_{i,s,t}^2)$ and two one-way violations $(Y_{i,s,t}^5)$, we estimate both the two-way DID-INT and the two one-way DID-INT a 1000 times, and compare their kernel densities. The results are shown in Figure (20). In Panel (a), where the two-way CCC holds, both estimators are unbiased. In contrast, Panel (b) shows that when the two-way CCC is violated, the two-way DID-INT remains consistent, while the two one-way DID-INT is inconsistent. Panels (c) and (d) considers cases where the two one-way CCC

holds and is violated respectively. In both cases, both two-way and two-one way estimators are consistent. However, the two-way DID-INT is more inefficient compared to the two one-way DID-INT.
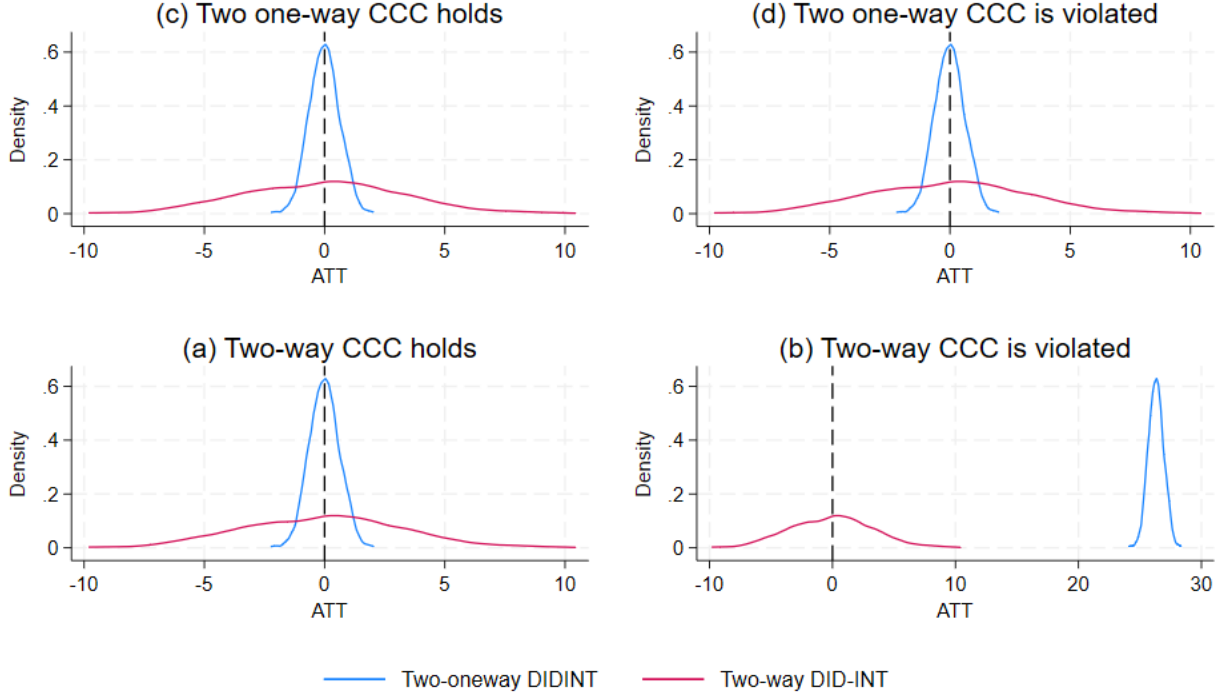

Figure 20: Kernel Densities of two one-way and two-way DID-INT

I Degree of CCC violation

In this section, we investigate the effects of increasing the degree of CCC violation on the bias of TWFE and the CS-DID estimators. A higher "degree" of CCC violation corresponds to a larger difference between $\gamma_{i,s}$ between groups and/or periods. To do so, we use the variable education from the CPS to construct a series of fake outcome variables, $Y_{i,s,t}$, based on the type and "degree" of CCC violation. The sample remains restricted to women aged 24 to 55 in their fourth interview month from New York, Pennsylvania, New Jersey, Virginia and Rhode Island from 2000 to 2014.

We generate data using four types of DGPs. In DGP 1, the effect of the covariate education (γ_i) differs by region but remains the same across time. In other words, the region-invariant CCC is violated. In DGP 2, the time-invariant CCC is violated, which implies γ_s differs by years, but remains the same across regions. The DGP3, $\gamma_{i,s}$ differs by both region and years, implying a two-way CCC violation. Within each DGP, we introduce five "degrees" of CCC violation in order to assess its effect on bias. The degrees are classified as very low, low, medium, high and very high; each corresponding to a difference of 10, 50, 100, 250 and 500 between γs. Lastly, in DGP4, the CCC assumption holds.

After generating the data, we proceed to run the TWFE, CS-DID, region-varying DID-INT, time-varying DID-INT and the two-way DID-INT and repeat this a 1000 times. We then calculate the bias for each estimator across all DGPs and "degrees" of CCC violation using the following formula:

$$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta^0 \tag{75}$$

where, $\hat{\theta}$ is the estimated ATT from each method, and θ^0 is the true ATT (set to 0). The results are shown in Table (3). We observe that the bias of the TWFE estimator increases as the "degree" of CCC violation increases across the first three DGPs. A similar pattern for the CS-DID is observed in DGPs 1 and 3. However, in DGP 2, CSDID maintains a constant bias of 0.2 across all "degrees" of CCC violations. The two-way DID-INT remains relatively unbiased across all DGPs and "degrees" of CCC violations, while the region-varying DID-INT remains relatively unbiased for DGP1 and the time-varying DID-INT remains relatively unbiased for DGP2. In DGP 3, the bias for both region-varying and time-varying DID-INT increases as the "degree" of CCC violations increases. When there are no CCC violations, all estimators are unbised, except for the CS-DID.

DGP1 (Region-varying coefficients)					
CCC Violation	**TWFE**	**CSDID**	**Region-varying DID-INT**	**Time-varying DID-INT**	**Two-way DID-INT**
Very Low	0.822	0.528	0.005	1.471	0.005
Low	14.045	0.204	0.008	7.388	0.008
Medium	8.221	3.295	0.003	14.763	0.004
High	20.552	7.92	0.004	36.903	0.006
Very High	41.105	15.637	0.004	73.795	0.005
DGP2 (Time-varying coefficients)					
CCC Violation	**TWFE**	**CSDID**	**Region-varying DID-INT**	**Time-varying DID-INT**	**Two-way DID-INT**
Very Low	2.809	0.219	1.673	0.005	0.006
Low	14.045	0.205	8.399	0.008	0.008
Medium	28.091	0.211	16.785	0.003	0.004
High	70.227	0.21	41.959	0.004	0.006
Very High	140.455	0.217	83.906	0.004	0.005
DGP3 (Two-way varying coefficients)					
CCC Violation	**TWFE**	**CSDID**	**Region-varying DID-INT**	**Time-varying DID-INT**	**Two-way DID-INT**
Very Low	67.577	67.799	77.937	82.522	0.006
Low	337.884	339.884	389.719	412.647	0.008
Medium	675.768	679.967	779.425	825.28	0.0035
High	1689.419	1700.236	1948.56	2063.197	0.006
Very High	3378.84	3400.675	3897.109	4126.383	0.005
No CCC Violation					
CCC Violation	**TWFE**	**CSDID**	**Region-varying DID-INT**	**Time-varying DID-INT**	**Two-way DID-INT**
	0.001	0.217	0.004	0.004	0.005

Table 3: Absolute Bias