Difference in Differences

Isaac Shon

September 18, 2024

1 Introduction

Oftentimes, we are interested in exploring the effect of an intervention *over a period of time*. For example, we may be interested in the effect of a one-time, non-random policy change that affects an entire jurisdiction or a wide set of units. If we are able to simultaneously observe a set of untreated units over time, we are able to see how units exposed and unexposed to the treatment experience a change in their outcomes over time. This gives us an opportunity to find whether an intervention rolled out had a credible, causal effect on some outcome Y.

Although we can simply compare differences in before and after measures of our outcome of interest, such a comparison is naive. We do not know if such a difference can be attributed to our treatment, or if such a change would have already occurred without the treatment. In a setting where we have access to panel data (i.e., the same units are observed over a period of time) where some units are treated and others are untreated, difference in differences (DID) is a popular strategy to estimate the causal effects of an intervention. In order to obtain treatment effects using DID, we must make a key identifying assumption known as the *parallel trends assumption*. This assumption asserts that had the treated units not experienced the treatment, their outcome evolution would have occurred in a parallel trend as the untreated units.

Let us first briefly discuss the canonical, 2x2 difference in differences model, where we simply consider two groups during pre-treatment and post-treatment periods. Let Y_{it} be the outcome variable of interest for unit i, observed over t periods where $t = 1$ or $t = 2$. Additionally, let T_i represent the treatment status for a given unit i, Post_t represent a dummy variable indicating the post-treatment period, and u_{it} represent the time-and-unit-varying unobservable error term. In its most basic form, the difference-in-differences can be represented with the following regression equation:

$$Y_{it} = \beta_0 + \beta_1(\text{Post}_t \times T_i) + \beta_2\text{Post}_t + \beta_3 T_i + u_{it} \tag{1}$$

In this equation, β_1 represents the causal effect of the treatment, and it represents the average treatment effect on the treated (ATT). OLS estimates of Equation 1 provide consistent and asymptotically valid confidence intervals of the ATT, provided that (1) the parallel trends assumption holds, (2) the no-anticipation assumption is satisfied, and (3) the stable-unit treatment value assumption (SUTVA) holds. In particular, SUTVA is an important assumption in the identification of the ATT, as it requires that the outcomes of a given unit i do not depend on the outcome of unit $j \neq i$, which effectively rules out potential spillover and general equilibrium effects. Additionally, the no-anticipation assumption (i.e., units do not anticipate the treatment of interest and that the treatment has no causal effect prior to its actual implementation), is another assumption necessary for identification, as changes in outcomes between periods $t = 1$ and $t = 2$ for both groups might include both the causal effect *and* anticipatory effects in period $t = 1$.[1]

[1] This also relates to the "Ashenfelter dips" often seen in program evaluation (Ashenfelter, 1978).

In the canonical 2x2 case, where we have two groups observed pre- and post-treatment, the DID estimator is given as the equation below. Note that this is a *"difference in differences"*:

$$\beta_1 = (\mathbb{E}[Y_{it}|P_t = 1, T_i = 1] - \mathbb{E}[Y_{it}|P_t = 0, T_i = 1]) - (\mathbb{E}[Y_{it}|P_t = 1, T_i = 0] - \mathbb{E}[Y_{it}|P_t = 0, T_i = 0])$$

1.1 An Early Empirical Example: Card & Krueger (1994)

One particularly famous example of a natural experiment using DID comes from Card and Krueger (1994)'s minimum wage study of fast-food restaurant chains in New Jersey and Pennsylvania. In this study, the authors were interested in seeing whether minimum wage changes had an impact on fast food industry employment. The policy change of interest took place in New Jersey, where the minimum wage rose from \$4.25/hr to \$5.05/hr beginning in April 1, 1992. Here, the authors collected data on a sample of fast-food restaurants in New Jersey and Eastern Pennsylvania (where the policy did not take effect) in two waves: the first in February/March 1992 ("pre-treatment") and again in November/December 1992 ("post-treatment").[2]

It is clear that under this setup, we have also two clearly defined groups: the "treated" restaurants in NJ that experienced the policy change and the "untreated" restaurants in eastern PA. In Figure 1, I reproduce Figure 1 of the original AER article, using Bruce Hansen's slightly abridged .dta file. As seen from the plots, starting wages for all of the surveyed restaurants in New Jersey were at least \$5.05/hr following the minimum wage hike, whereas starting wages in surveyed restaurants in PA remained relatively unchanged.

We can take the conditional means of full time employment counts to compute our 2x2 DID estimate. As stated in Card and Krueger (1994), full-time equivalent (FTE) employment counts includes the number of full time workers, managers, and $\frac{1}{2}$ of part-time workers. The conditional FTE means are provided in Table 1. Whereas the average employment per store in the sampled restaurants in PA decreased between the two survey waves, the average employment per store in the sampled NJ stores increased between the two waves. This finding famously challenged the widely supported, textbook prediction that minimum wage hikes decrease employment. Based on the means given in Table 1, the manually-computed DID estimate comes out to around 2.75. Alternatively, we can also use OLS to compute the DID estimate (that take the form of Equation 1). However, as stated previously, OLS is consistent and asymptotically valid only under satisfied assumptions.

1.2 Potential Outcomes Framework

Let us now more formally discuss the difference-in-differences approach through potential outcomes (Rubin, 1976). To do this, we will now introduce new notation and define potential outcomes as seen in Roth et al. (2023). We now consider a panel data structure of units $i \in \{1, 2, ..., N\}$ observed over the time periods $t \in \{1, 2, ..., T\}$. Let $Y_{i,t}(g)$ be the potential outcome for unit i at time period t, if this unit was first treated at time g. $g_i \in \mathbb{G} \subset \{1, ..., T\} \cup \{\infty\}$ denotes the time period unit i is first-treated. If a unit is "never-treated", we let $G_i = \infty$. The observed outcome for unit i in time period t is given by $Y_{i,t} = \sum_{g \in G} \mathbb{1}\{G_i = g\} Y_{i,t}(g)$. In the 2x2 example, a subset of all units are treated at time $g = 2$, and a subset of units remain untreated at time $t = 2$. As such, $G = \{2, \infty\}$.

For units that are treated at $t = 2$, we observe $Y_{i,t=1}(2)$ and $Y_{i,t=2}(2)$. Likewise, for our untreated units we observe $Y_{i,t=1}(\infty)$ and $Y_{i,t=2}(\infty)$. At period t, actual realized outcomes are defined as:

$$Y_{i,t} = \mathbb{1}\{G_i = 2\} Y_{i,t}(2) + \mathbb{1}\{G_i = \infty\} Y_{i,t}(\infty) \tag{2}$$

The causal effect of the treatment for unit i treated at period t is simply the difference in potential outcomes, $Y_{i,t}(2) - Y_{i,t}(\infty)$. Of course, the fundamental problem in causal inference is that

[2] The original replication can be found on Prof. David Card's website here. Data in .dta format was retrieved through Prof. Bruce Hansen's Econometrics Data folder, which can be found here.

for both treated and untreated groups, we cannot simultaneously observe both the potential outcomes $Y_{i,t}(2)$ and $Y_{i,t}(\infty)$. The causal parameter of interest, the ATT, is defined as:

$$\text{ATT} \equiv \mathbb{E}[Y_{i,t}(2)|G=2] - \mathbb{E}[Y_{i,t}(\infty)|G=2], \tag{3}$$

where the term $\mathbb{E}[Y_{i,t}(\infty)|G=2]$ is our unobserved counterfactual that we would like to approximate. Recall that identification of the ATT requires 3 crucial assumptions to hold: (1) parallel trends, (2) no anticipation, and (3) SUTVA. Let us now formalize these assumptions:

Definition 1 (SUTVA)**.** *Observed outcomes at time period t are realized as $Y_{i,t} = \sum_{g\in G} \mathbb{1}\{G_i = g\}Y_{i,t}(g)$.*

Definition 2 (No anticipation)**.** *For all units i, $Y_{i,t}(g) = Y_{i,t}(\infty)$ for all groups in their pre-treatment periods (i.e., for all $t < g$).*

Definition 3 (Parallel trends)**.**

$$\mathbb{E}[Y_{i,t=2}(\infty)|G_i=2] - \mathbb{E}[Y_{i,t=1}(\infty)|G_i=2] = \mathbb{E}[Y_{i,t=2}(\infty)|G_i=\infty] - \mathbb{E}[Y_{i,t=1}(\infty)|G_i=\infty]$$

From our formal definitions of SUTVA and the ATT, we have that:

$$\begin{aligned} \text{ATT} &\equiv \mathbb{E}[Y_{i,t=2}(2)|G=2] - \mathbb{E}[Y_{i,t=2}(\infty)|G=2] \\ &= \mathbb{E}[Y_{i,t=2}|G=2] - \mathbb{E}[Y_{i,t=2}(\infty)|G=2] \end{aligned}$$

Under SUTVA, the green object above is estimable from our data. Under the parallel trends assumption and the no-anticipation assumption, we are also able to rewrite the (unobserved) red object. Under SUTVA, no-anticipation, and parallel trends, the ATT can be re-written as a *difference-in-differences* in potential outcome notation:

$$\begin{aligned} \text{ATT} &\equiv \mathbb{E}[Y_{i,t}(2)|G=2] - \mathbb{E}[Y_{i,t=2}(\infty)|G=2] \\ &= \mathbb{E}[Y_{i,t=2}|G=2] - \mathbb{E}[Y_{i,t=2}(\infty)|G=2] \\ &= \mathbb{E}[Y_{i,t=2}|G=2] - (\mathbb{E}[Y_{i,t=1}|G_i=2] + (\mathbb{E}[Y_{i,t=2}|G_i=\infty] - \mathbb{E}[Y_{i,t=1}|G_i=\infty])) \\ &= (\mathbb{E}[Y_{i,t=2}|G=2] - \mathbb{E}[Y_{i,t=1}|G_i=2]) - (\mathbb{E}[Y_{i,t=2}|G_i=\infty] - \mathbb{E}[Y_{i,t=1}|G_i=\infty]) \\ &= \mathbb{E}[Y_{i,t=2} - Y_{i,t=1}|G=2] - \mathbb{E}[Y_{i,t=2} - Y_{i,t=1}|G=\infty] \end{aligned}$$

2 Two-Way Fixed Effects

Oftentimes in empirical work, we rarely see papers that manually compute DID estimates. In modern practice, we are often used to seeing more generalized versions of difference-in-differences being incorporated into regression models. Let us now turn to discussion on the two-way fixed effects (TWFE) regression estimator, a type of generalization of difference-in-differences. Suppose that we observe N units over t periods, where t can take on values greater than 2. For simplicity, we'll assume that we have a balanced panel data set (i.e., all of our units are observed over each time period). We will consider the following simple two-way linear fixed effects model without covariates:

$$Y_{it} = \alpha_i + \gamma_t + \beta D_{it} + \varepsilon_{it}, \tag{4}$$

for $i = 1, 2, ..., N$ and $t = 1, 2, ..., T$ with unit fixed-effects (α_i) and time fixed-effects (γ_t). In the binary treatment setting, $D_{it} = \mathbb{1}\{\text{Treated in period } t\}$. The inclusion of α_i accounts for unit-specific but time-invariant unobserved confounders, whereas γ_t accounts for time-specific but unit-invariant unobserved confounders in a flexible manner. More sprecifically, $\alpha_i = h(U_i)$ and

$\gamma_t = f(V_t)$, where U_i and V_t represent unit-specific and time-specific unobservable confounders. Note that the exact functional forms of $h(.)$ and $f(.)$ are unknown.

When $T = 2$, TWFE is the same as the traditional 2-by-2 difference-in-differences model. The following TWFE estimator formally defines how β is estimated in the 2x2 setting:

$$\beta = \arg\min_{\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} [\{(Y_{it} - \bar{Y}) - (\bar{Y}_i - \bar{Y}) - (\bar{Y}_t - \bar{Y})\} - \{(D_{it} - \bar{D}) - (\bar{D}_i - \bar{D}) - (\bar{D}_t - \bar{D})\}]^2,$$

where $\bar{Y}_i = \sum_{t=1}^{T} Y_{it}/T$ and $\bar{D}_i = \sum_{t=1}^{T} D_{it}/T$ are unit-specific means, and $\bar{Y}_t = \sum_{i=1}^{N} Y_{it}/N$ and $\bar{D}_t = \sum_{i=1}^{N} D_{it}/N$ are time-specific means. Additionally, $\bar{Y} = \sum_{i=1}^{N} \sum_{t=1}^{T} Y_{it}/NT$ and $\bar{D} = \sum_{i=1}^{N} \sum_{t=1}^{T} D_{it}/NT$ both represent overall means. In the above equation, we can clearly see that least squares estimation is applied after the variation within-units and within-time periods are subtracted from the overall variation for both the outcome Y and treatment D.

2.1 Covariate Selection

In difference-in-difference studies, we typically only require information at the group level (i.e., the treatment and control groups) to identify treatment effects. However, there are several reasons why one may consider including a set of covariates in a difference-in-differences regression. First, including covariates may help in reducing the error variance. By controlling for within-group variation, we may be able to increase the precision of our estimates and reduce our standard errors.

Typically, one common diagnostic tool to assess whether the traditional parallel trends assumption might hold is to plot pre-trends between treatment and control groups and offer a justification why such an assumption may hold.[3] However, there are often situations where, absent treatment, units in the treatment and control groups with different observed characteristics may evolve differently over time. For example, is it reasonable to assume that absent a hypothetical workforce training program, the wages of younger workers would evolve similarly to the wages of older workers? Or, would it be reasonable to assume, that absent minimum wage hikes, states in New England would experience similar employment evolution to states in the Southeast? In such cases, there may not be a clear justification for the traditional parallel trends assumption to hold. As such, researchers may choose to include covariates in regression models to instead justify a *conditional* parallel trends assumption.

Generally, the traditional parallel trends assumption may be implausible when pre-treatment characteristics that are thought to be associated with the dynamics of the outcome variable are "unbalanced" between the treated and the untreated group (Abadie, 2005). When this is the case, given a vector of pre-treatment characteristics, we may be able to relax the traditional parallel trends assumption with the conditional parallel trends assumption. This assumption states that in the absence of treatment, conditional on X, the evolution of the outcome among the treated units is, on average, the same as the evolution of the outcome among the untreated units. Of course, the researcher must exercise caution in selecting which exact covariate(s) to use, but an added benefit of conditioning on pre-treatment covariates is that it allows for analyzing covariate-specific trends.

2.2 Clustered Standard Errors

Recall from introductory econometrics that the choice between using different types of standard errors depends on what assumptions we would like to impose on the variance of our error term, u_{it}. However, when using difference-in-differences, Bertrand et al. (2004) point out that

[3] It is important to remember that the parallel trends assumption is in itself an untestable assumption. Examining pre-trends cannot tell us information about a counterfactual that we cannot observe.

oftentimes even conventional standard errors are biased *downward* and understate the true standard deviation of DID estimators. They note that standard error biases in DID estimation may possibly be due to severe serial correlation problems. Heteroskedasticity-robust standard errors traditionally used in OLS estimation, for example, assume that the $N \times N$ matrix $\mathbb{E}[\varepsilon\varepsilon'|X]$ is diagonal (i.e., no correlation between the errors across observations). However, in panel data settings this assumption may not necessarily hold. Given the fact that many DID studies make use of panel data over multiple time periods, often use dependent variables that are "highly positively serially correlated", and use treatment variables that change little within a state over time, together these issues often result in mis-measurement of the standard errors when using OLS estimation.

In difference-in-difference designs, it is usually the case that we cluster standard errors at the level of how our treatment was assigned. Clustered standard errors (also known as Liang-Zeger standard errors) are another type of standard error the researcher may choose to use, and unlike traditional standard errors, differ in that they allow for unrestricted forms of serial correlation and heteroskedasticity within certain sub-groups in the data. These standard errors are typically consistent in the presence of cluster-based sampling or treatment assignment. Abadie et al. (2022) point out that what matters when deciding to cluster standard errors lies in how the sample was selected, and whether there are clusters in the population of interest that are not represented in the sample. For example, it may not be necessary to cluster standard errors when a sample of individuals is randomly drawn from the population of interest. However, when randomly assigning individuals within a randomly-selected subset of villages or townships within a country, it may be necessary to cluster at the village/township level as there are other potential units beyond the sample in the population of interest.

3 Staggered Difference-in-Differences

Motivated by the fact that in the 2x2 setup previously discussed, DID is equal to the treatment coefficient in a TWFE regression, researchers have also begun to estimate TWFE regressions in more complicated designs. These include settings in which there are many groups and time periods, variation in treatment timing, non-binary treatments, and treatments that "switch" on and off. In this section, we will focus and discuss in particular some of the potential issues that arise with the traditional TWFE estimator when considering *staggered treatment regimes*. In staggered treatment regimes, researchers consider settings when the treatment of interest is irreversible and rolled out to units over an extended period of time. That is, instead of a treatment being assigned at a single period of time, some or all units may receive the treatment at different dates. Here, we are typically working with panel data extending over multiple time periods.

Given that the variable D_{it} is simply an indicator representing whether a unit i was treated at time period t, it is indeed tempting to use some sort of variation of the TWFE specification outlined in Equation 4 in staggered settings. After all, in the traditional 2x2 setting, the TWFE estimator is consistent under satisfied assumptions and offers a relatively straightforward interpretation of the causal parameter of interest, β. Overall, much of the literature has suggested that standard TWFE (with $T = 2$) only works when there are homogeneous treatment effects across units and time periods, the parallel trends assumption holds, and the linear additive effects assumption holds. However, in more complicated staggered settings, several potential issues could introduce bias to the TWFE estimator.

In the basic 2x2 TWFE model, we have previously established that the coefficient β represents the average treatment effect on the treated units. In principle, we could run a regression similar to the 2x2 case to obtain our estimates in staggered treatment regimes, but it is not exactly clear what the parameter β represents or what comparisons are being made in the first place. Goodman-Bacon (2021) provides a helpful and intuitive way to unpack the TWFE estimator in staggered settings. Here, we will consider a setting where we have balanced panel data with T periods and N cross-sectional units i that fall into 3 general categories: an untreated group U, an early-treated group k that receives treatment at period t_k^*, and a late-treated group l which receives the binary treatment at $t_l^* > t_k^*$. The primary finding of the paper is the Bacon-Decomposition Theorem, outlined in the following box:

Theorem 1 (Goodman-Bacon (2021) Decomposition)**.** *Assume that there are $k = 1, ..., K$ groups of units that receive a binary treatment ordered by treatment time $t_k^* \in (1, T]$. There may be one "never-treated" group, U, that is never exposed to the treatment. Also, let the share of units in a particular group k be n_k, and their respective share of periods spent under treatment be $\bar{D}_k$. The regression estimate of a static TWFE, $\hat{\beta}$, is a weighted average of all possible 2x2 difference-in-difference estimators:*

$$\hat{\beta} = \sum_{k \neq U}(s_{kU}\hat{\beta}_{kU}) + \sum_{k \neq U}\sum_{l>k}(s_{kl}\hat{\beta}_{kl} + s_{lk}\hat{\beta}_{lk}),$$

where the 2x2 difference-in-difference estimators are:

$$\begin{aligned}\hat{\beta}_{kU} &\equiv (\bar{y}_k^{POST(k)} - \bar{y}_k^{PRE(k)}) - (\bar{y}_U^{POST(k)} - \bar{y}_U^{PRE(k)})\\ \hat{\beta}_{kl} &\equiv (\bar{y}_k^{MID(kl)} - \bar{y}_k^{PRE(k)}) - (\bar{y}_l^{MID(kl)} - \bar{y}_l^{PRE(k)})\\ \hat{\beta}_{lk} &\equiv (\bar{y}_l^{POST(l)} - \bar{y}_l^{MID(kl)}) - (\bar{y}_k^{POST(l)} - \bar{y}_k^{MID(kl)}),\end{aligned}$$

and the weights are given by:

$$s_{kU} = \frac{(n_k + n_U)^2\hat{V}_{kU}}{\hat{V}(\widetilde{D_{i,t}})}, s_{kl} = \frac{((n_k + n_l)(1 - \bar{D}_l)\hat{V}_{kU})^2}{\hat{V}(\widetilde{D_{i,t}})}, s_{lk} = \frac{((n_k + n_l)(\bar{D}_k)\hat{V}_{lk})^2}{\hat{V}(\widetilde{D_{i,t}})},$$

such that $\sum_{k \neq U} s_{kU} + \sum_{k \neq U}\sum_{l>k}(s_{kl} + s_{lk}) = 1$

In short, Goodman-Bacon (2021) shows that the TWFE estimator is essentially a weighted average of all possible 2x2 DID estimators. Three specific types of 2x2 comparisons are made, all of which depend on what "cohort" a given unit falls under. These include treated v.s. never-treated comparisons, early-treated v.s. later-treated comparisons, as well as later-treated v.s. early-treated comparisons. While units that never receive treatment are considered part of the control group, we can see that in staggered settings some treated units serve as control comparisons as well.

Although all of the weights attached to each 2x2 DID estimator according to the Bacon Decomposition theorem are positive *by definition*, it may be the case that some comparisons are weighed *negatively* in the presence of treatment effect heterogeneity across units or time. Negative weighting may be problematic and could bias TWFE estimates. This is particularly concerning, considering that in many applications, heterogeneity in treatment effects is likely to hold. Additionally, the presence of treatment effect dynamics (i.e., treatment effects evolving over time) implies different trends from what would have happened absent the treatment. Because the "control group" in some of the 2x2 comparisons may be already-treated at both periods, changes in their treatment effects over time will be subtracted from the 2x2 DID, leading to negative weights. In certain instances, this might lead to a complete sign reversal of the overall $\hat{\beta}$ estimate even if all 2x2 comparisons are positive, and make causal interpretation difficult. Another implication of Theorem 1 is that, even when weights are not negative, the weights on underlying parameters may be influenced by other factors such as cohort size, treatment timing, and the total number of time periods in the data.

In light of these issues, researchers have developed a wide variety of approaches to improve

TWFE estimation. Much of recent DID work have proposed new heterogeneity-robut estimators and attempt to understand what exactly the causal parameter β in Equation 4 recovers in staggered treatment regimes (e.g., see De Chaisemartin and D'Haultfœuille, 2020; Athey and Imbens, 2022; Borusyak et al., 2024). These proposed strategies are inherently similar in the sense they seek to provide remedies to the negative weighting issues that arise in traditional TWFE models with treatment effect heterogeneity/treatment effect dynamics. However, many of these estimation strategies impose different types of assumptions and essentially make some sort of trade-off in robustness and efficiency. When deciding on which strategy to apply, the researcher should ask themselves ex-ante what particular research question they are interested in answering, what setting they would like to examine, and what restrictions are they willing to impose.

While there are a variety of new estimators in the literature, we will now discuss one specific strategy developed by Callaway and Sant'Anna (2021). Similar to other recently proposed solutions, this proposed strategy essentially seeks to dis-aggregate the overall ATT to make more "desirable" comparisons between groups instead of all possible comparisons. When aggregating certain types of comparisons into new parameters, researchers are then able to more concisely summarize heterogeneity with respect to some particular dimension, whether it be based on how long units receive exposure to treatment or a single overall treatment effect. Let us now provide a brief overview of Callaway and Sant'Anna (2021). **It is highly encouraged for the reader to thoroughly read the original JoE manuscript for more precise definitions, explanations, and proofs.**[4]

Identification Results

Like most staggered DID settings, they consider the case where there are $i = 1, ..., N$ units observed over $t = 1, ..., T$ periods and a treatment variable $D_{it} = \mathbb{1}\{\text{treated in period } t\}$. The core "building block" the authors consider is the *group-time* average treatment effect on the treated, $ATT(g, t)$. This parameter in itself has a clear economic interpretation: the group-time ATT gives the average treatment effect at time t for the cohort first treated in time g. For example, $ATT(2020, 2024)$ gives the treatment effect in the year 2024 for units that were treated in 2020.

Given that we never observe the counterfactual potential outcome $Y(\infty)$ in post-treatment periods among units that have been treated, the authors first lay out several identifying assumptions for our $ATT(g, t)$'s. These assumptions are: (1) irreversible treatment, (2) random sampling, (3) limited treatment anticipation, (4) conditional parallel trends for never-treated, (5) conditional parallel trends for not-yet treated, and (6) strong overlap. Under the first assumption, we have that every unit, once treated, does not change their treatment status at some later period in the data. Under the second assumption, we also assume that we have access to a balanced panel data set that is $i.i.d.$. Additionally, Assumption 3 restricts anticipation of the treatment for all "eventually treated" cohorts, though they do allow for some limited anticipation by some $\delta \geq 0$ time periods.

Assumptions 4 and 5 make two types of parallel trends assumptions, depending on the particular comparison group one chooses to use. Assumption 4 states that the average outcomes for the group that receives treatment for the first time in period g and for the group that never receives treatment would have followed parallel trends in the absence of treatment, conditional on pre-treatment characteristics X. Likewise, Assumption 5 states that the average outcomes for the group that receives treatment for the first time at period g and for the group that are "not-yet treated" at period $t + \delta$, conditional on pre-treatment characteristics X. The final

[4] In the following text, I lay out the assumptions and definitions from Callaway and Sant'Anna (2021) as-is and refrain from defining some of the new notation introduced. I do this so as to avoid unintentional plagiarism and excessive copying from the original JoE article.

identifying assumption imposed is an overlap condition, which states that a positive fraction of the population is first-treated in period g, and that for all g and t the generalized propensity score (the probability of being first treated at time g, conditional on pre-treatment covariates X and on being a member of either cohort g or the "not-yet-treated" cohort by time s) is uniformly bounded away from one.

Callaway and Sant'Anna (2021) show that a wide family of group-time ATT's are "non-parametrically point-identified" when these assumptions hold, and that one is able to use outcome regression, inverse probability weighting, and doubly robust estimands to recover the ATT(g,t)'s.[5] For example, the doubly-robust estimands when $\delta = 0$ (i.e., when no anticipation holds) are given as:

Never-Treated as Comparison Group

$$ATT_{dr}^{nev}(g,t) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}[G_g]} - \frac{\frac{p_g(X)G_\infty}{1-p_g(X)}}{\mathbb{E}\left[\frac{p_g(X)G_\infty}{1-p_g(X)}\right]}\right)(Y_t - Y_{g-1} - m_{g,t}^{nev}(X))\right],$$

where $G_g = \mathbb{1}\{G = g\}$, $m_{g,t}^{nev}(X) = \mathbb{E}[Y_t - Y_{g-1}|X, G = \infty]$, and $p_g(X) = \mathbb{E}[\mathbb{1}\{G = g\}|X, G = \{\infty, g\}]$.

Not-Yet-Treated as Comparison Group

$$ATT_{dr}^{ny}(g,t) = \mathbb{E}\left[\left(\frac{G_g}{\mathbb{E}[G_g]} - \frac{\frac{p_{g,t}(X)(1-D_t)}{1-p_{g,t}(X)}}{\mathbb{E}\left[\frac{p_{g,t}(X)(1-D_t)}{1-p_{g,t}(X)}\right]}\right)(Y_t - Y_{g-1} - m_{g,t}^{ny}(X))\right],$$

where $G_g = \mathbb{1}\{G = g\}$ and $D_t = \mathbb{1}\{G \leq t\}$, $m_{g,t}^{ny}(X) = \mathbb{E}[Y_t - Y_{g-1}|X, D_t = 0, G_g = 0]$, and $p_{g,t}(X) = \mathbb{P}[G = g|X, G_g + (1 - D_t)(1 - G_g) = 1]$.

Overall, the main result of their paper is the following theorem:

Theorem 2 (Callaway and Sant'Anna (2021) Identification of $ATT(g,t)$)**.** *Let A1-A3 and A6 hold. If A4 holds, then for all g and t such that $g \in \mathbb{G}_\delta$, $t \in \{2, ..., T - \delta\}$ and $t \geq g - \delta$,*

$$ATT(g,t) = ATT_{ipw}^{nev}(g,t;\delta) = ATT_{dr}^{nev}(g,t;\delta) = ATT_{or}^{nev}(g,t;\delta)$$

If A5 holds, then for all g and t such that $g \in \mathbb{G}_\delta$, $t \in \{2, ..., T - \delta\}$ and $g - \delta \leq t < \bar{g} - \delta$,

$$ATT(g,t) = ATT_{ipw}^{ny}(g,t;\delta) = ATT_{dr}^{ny}(g,t;\delta) = ATT_{or}^{ny}(g,t;\delta)$$

ATT(g,t) Aggregation Schemes

Given the identification of $ATT(g,t)$, we can now aggregate these parameters to summarize treatment effects across groups and time. Of course, we can obtain a single-summary treatment effect estimate by taking some weighted average of the $ATT(g,t)$ estimates. However, several different types of aggregation schemes exist, including cohort-based summaries (the average treatment effect that units treated in cohort g experienced):

$$\theta_s(g) = \frac{1}{T - g + 1} \sum_{t=2}^{T} \mathbb{1}\{g \leq t\} ATT(g,t)$$

[5] The authors also establish consistency and asymptotic normality of proposed sample-analouge estimators and prove the validity of a multiplier bootstrap procedure that constructs simultaneous confidence intervals for $ATT(g,t)$.

...time-based summaries (the average treatment effect in time period t for cohorts that were treated by time t):

$$\theta_c(t) = \sum_{t=2}^{T} \mathbb{1}\{g \leq t\} ATT(g,t) \mathbb{P}(G = g | G \leq t, C \neq 1)$$

... or in dynamic TWFE specifications, exposure length-based summaries (the average treatment effect for units that were exposed to the treatment for $e = t - g$ periods):

$$\theta_D(e) = \sum_{t=2}^{T} \mathbb{1}\{g + e \leq t\} ATT(g, g+e) \mathbb{P}(G = g | G + e \leq t, C \neq 1)$$

Simulation Exercise: Implementing Callaway & Sant'Anna (2021)

In the following simulation exercise, I generate a balanced panel data set in which units experience a binary treatment that is rolled out over a period of time. Similar to Equation 4, the outcome variable, y_{it}, includes unit fixed effects $\alpha_i \sim N(0.025, 0.5)$, time fixed effects $\gamma_t \sim N(0.05, 0.32)$, as well as a time-and-unit varying error term $\varepsilon_{it} \sim N(0, 5)$. However, instead of explicitly setting a fixed β in the underlying data-generating process, I allow for treatment effect heterogeneity across groups and over time. More specifically, for a given unit i first-treated at time period g, the effect of being treated in time period t is some random integer between 3 and 7, multiplied by the number of periods it was exposed to the treatment. The original STATA code for the data simulation as well as R code for estimation and analysis can be found in the accompanying project folder.

In Table 2, I provide summary statistics based on treatment status. In the simulated data, there are $n = 1,000$ units observed over the years 2000-2009, resulting in a total of 10,000 observations. In terms of treatment assignment, I randomly selected half of the units to fall under as either "never-treated" units or "eventually-treated". Among the "eventually-treated" units, I randomly assign a particular treatment date from the years 2003-2006. The remaining units are considered to have been never-treated, so $D_{i,t} = 0$ for all the periods in the data set. We can think of this as some hypothetical policy that was rolled out to a certain fraction of units during the years 2003-2006, and that we have access the other pre-treatment and post-treatment periods outside of this window. Thus, there are a total of 5 groups/cohorts: those first-treated in 2003, 2004, 2005, 2006, or those that never experienced treatment.

In Figure 2a, I present a plot of how many units in the simulation sample eventually receive treatment. The vertical axis represents the portion of units in the sample, and the horizontal axis is divided into each time period. As shown in the figure, an increasing proportion of units receive the treatment over time, beginning in 2003 and ending in 2006. In Figure 2b, I also plot the mean outcomes over time for each cohort. Unlike classical DID settings, there are multiple groups that receive treatment exposure at different periods. As can be seen from the line plots, for each cohort that is eventually treated, there is a noticeable jump in the outcome $y_{i,t}$ on the date of their respective treatment date (marked by the vertical dotted lines). Had there been evidence of treatment anticipation, we would see these sudden "jumps" prior to the date in which units were first given treatment. Additionally, there is also a group of units that are never treated (in red), and do not appear to have any noticeable "jumps" in their outcome over time.

We now turn to estimating our treatment effect under this staggered treatment regime. Under a traditional TWFE specification, we obtain an estimated treatment effect of 9.90722. However, as previously discussed, the estimated treatment effect from TWFE can be viewed as a weighted average of all possible 2x2 DID estimators. These weights and estimates can be seen in Figure 4, where I visualize all the 2x2 DID comparisons and weights for the Bacon decomposition using the *bacondecomp* package in R. The horizontal dashed line represents the treatment effect estimate under the traditional TWFE specification. In addition to treated vs untreated comparisons, traditional TWFE also makes earlier treated vs later treated comparisons, as well as later-treated v.s. earlier-treated comparisons. We can see from this plot that many of the later-treated v.s. earlier-treated comparison estimates are negative, as we are essentially making this control group consist of already-treated units.

We will now estimate group-time average treatment effects. Recall that these estimates represent the average effect of participating in the treatment for units that were first-treated in period g, at time period t. In Figure 4, I plot the point estimates of each group-time average treatment effects for each cohort in each time period, bounded by a simultaneous confidence interval. In the pre-treatment time period, we should naturally expect that the estimated group-time treatment effects fall close to 0. We should not expect cohorts to experience treatment effects before they ever receive treatment. Although we see that there appears to be no treatment effect in the first treated period for each cohort, we can also see that the group-time average treatment effect for each cohort increases over time. That is, for each cohort in our data, the treatment effect appears to intensify over time, which is consistent with how we previously set it in our data set. In the first column of Table 3, I present the aggregated average treatment effects that are specific to each cohort. What we are essentially doing is reporting the average effect of participating in the treatment among units in group g, aggregated across all post-treatment periods. Likewise, in the second column of Table 3, I present the aggregated average treatment effects specific to each post-treatment period. This aggregation scheme reports the effect of participating in the treatment at a particular time period, for all groups that were considered “treated" at that period.

4 Concluding Remarks

Difference-in-differences is an incredibly powerful tool that lets researchers investigate the effect of policies and programs over time. At its very essence, we are studying the differential effect of a specific treatment/intervention on a ’treatment group’ versus another ’control group’ that did not experience the treatment. However, when working with longitudinal data, the researcher must carefully consider what type of analysis is appropriate given their setting and data, and ask themselves what restrictions are they willing to impose. In this writeup, we first started our discussion with the canonical 2x2 DID setup, some of its identifying assumptions, and consider an early empirical example of its use. We then discussed the classical two-way fixed effects estimator and how it is, in a sense, a type of generalization of the canonical 2x2 DID estimator.

While it is tempting to use the TWFE setup in more complicated settings, several issues arise that may potentially bias our regression estimates. We discuss in particular that in staggered treatment regimes, the TWFE estimator is essentially a weighted average of different 2x2 DID comparisons. However, some of these comparisons use already-treated units as the control group, which may lead to negative weighting issues when treatment effects vary over time. While there are a variety of heterogeneity-robust estimators that seek to resolve these issues, we focus attention to the strategy proposed by Callaway and Sant’Anna (2021). Their strategy considers the group-time ATT as the central building block, which allow us to better understand treatment effect heterogeneity. From this, we can aggregate these parameters to other dimensions of interest, such as by group or over time.

References

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

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2022). When Should You Adjust Standard Errors for Clustering? *The Quarterly Journal of Economics*, 138(1):1–35.

Ashenfelter, O. (1978). Estimating the Effect of Training Programs on Earnings. *The Review of Economics and Statistics*, 60(1):47.

Athey, S. and Imbens, G. W. (2022). Design-based analysis in Difference-In-Differences settings with staggered adoption. *Journal of Econometrics*, 226(1):62–79.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How Much Should We Trust Differences-In-Differences Estimates? *The Quarterly Journal of Economics*, 119(1):249–275.

Borusyak, K., Jaravel, X., and Spiess, J. (2024). Revisiting Event-Study Designs: Robust and Efficient Estimation. *Review of Economic Studies*, page rdae007.

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

Card, D. and Krueger, A. B. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *The American Economic Review*, 84(4):772–793.

De Chaisemartin, C. and D'Haultfœuille, X. (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review*, 110(9):2964–2996.

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

Roth, J., Sant'Anna, P. H., Bilinski, A., and Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2):2218–2244.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Table 1: Conditional Means of FTE

	Post-NJ	Pre-NJ	Post-PA	Pre-PA
fte	21.0274	20.4394	21.1656	23.3312
N	319	321	77	77

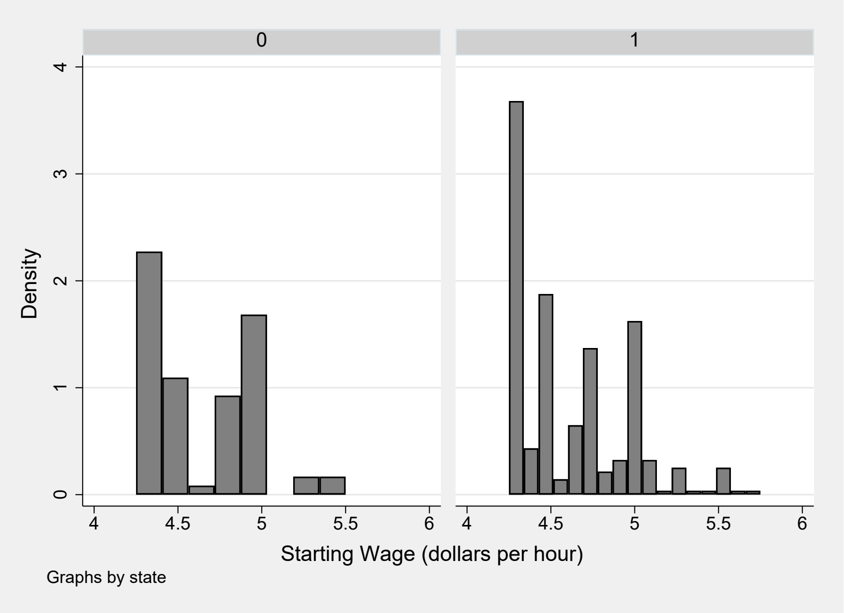
Table 2: Aggregate Summary Statistics by Treatment Status

Variable	Mean	Median	Min.	Max.	No. Obs.
		D_it: 0			
y_it	0.29	0.28	-19	19	7249
alpha_i	0.031	0.036	-2.2	1.6	7249
e_it	0.039	0.025	-20	18	7249
		D_it: 1			
y_it	13	13	-13	46	2751
alpha_i	0.042	0.039	-1.3	1.4	2751
e_it	0.1	0.2	-19	16	2751

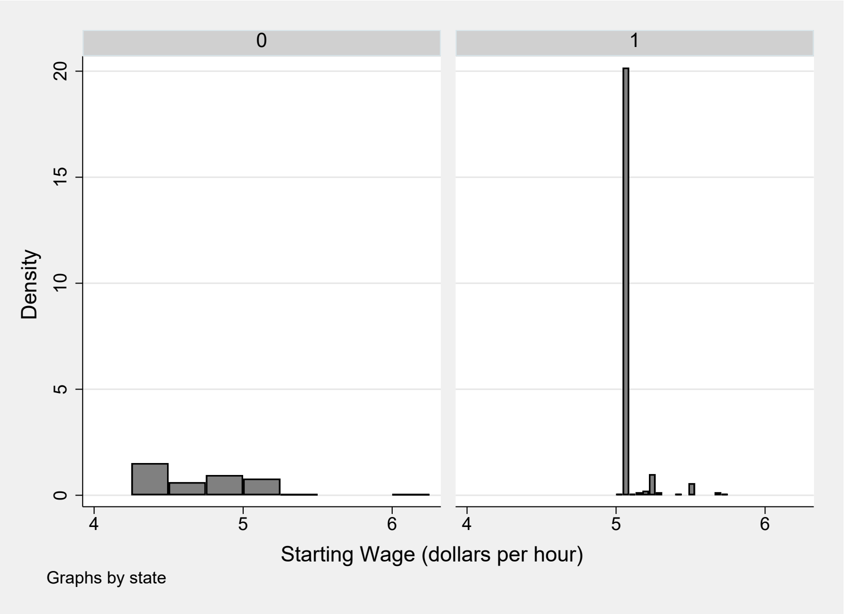
Table 3: Aggregated ATT(g,t) Estimates

	(1)	(2)
	group	calendar
ATT(Average)	12.844	
	[12.320, 13.368]	
ATT(2003)	15.379	0.597
	[14.033, 16.726]	[-1.304, 2.498]
ATT(2004)	14.661	2.903
	[13.128, 16.193]	[1.503, 4.304]
ATT(2005)	9.734	4.625
	[8.327, 11.141]	[3.360, 5.890]
ATT(2006)	11.752	8.372
	[10.405, 13.099]	[7.133, 9.612]
ATT(2007)		13.977
		[12.772, 15.182]
ATT(2008)		19.711
		[18.464, 20.959]
ATT(2009)		25.487
		[24.316, 26.658]
Num.Obs.	1000	1000
Std.Errors	by: i	by: i
type	group	calendar
ngroup	4	4
ntime	10	10
control.group	nevertreated	nevertreated
est.method	dr	dr

Figure 1: Distribution of Starting Wages Across Survey Waves (0 = PA, 1 = NJ)

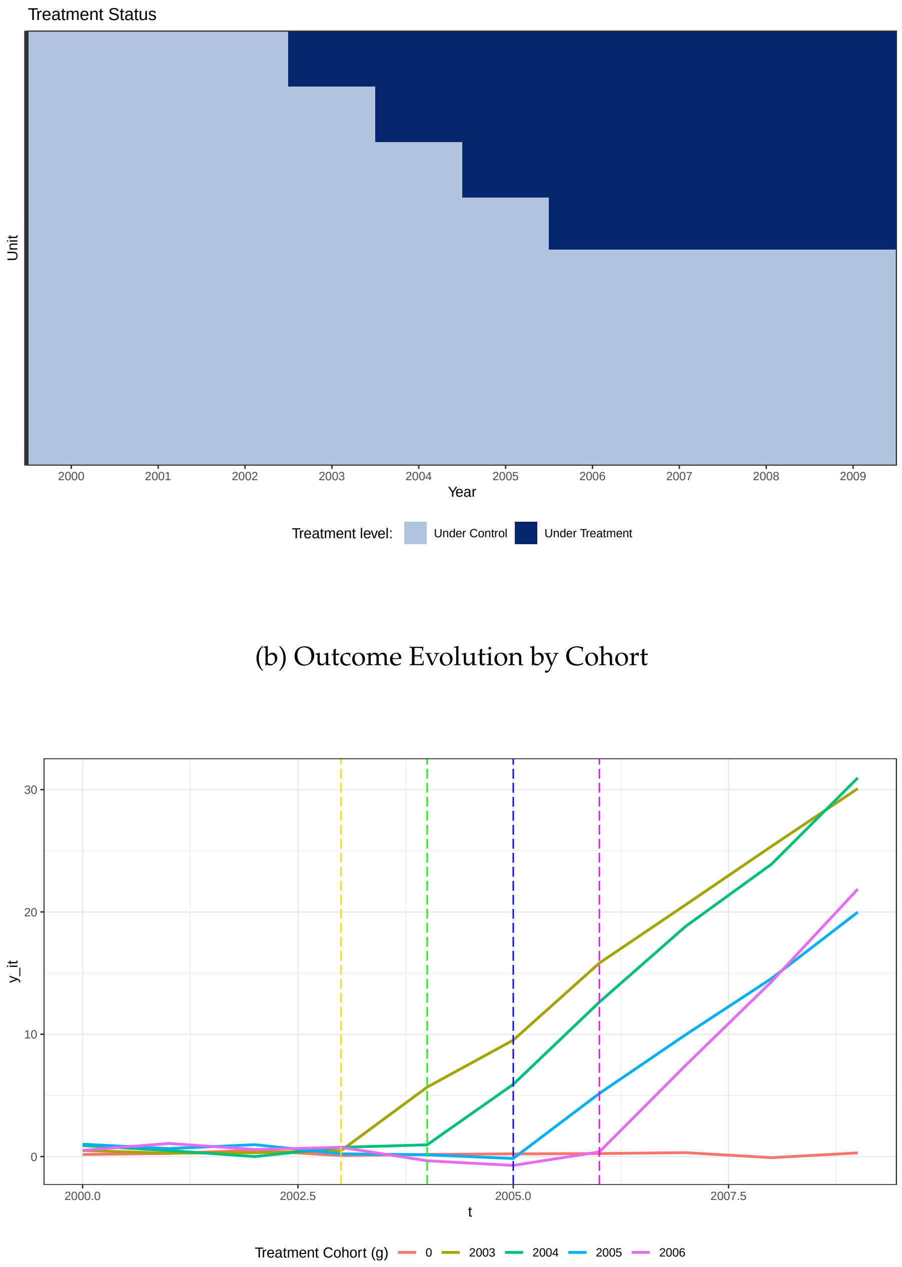


(a) February 1992



(b) November 1992

Figure 2: Staggered Treatment Rollout



(a) Treatment Distribution Over Time

(b) Outcome Evolution by Cohort

Figure 3: DID Weights

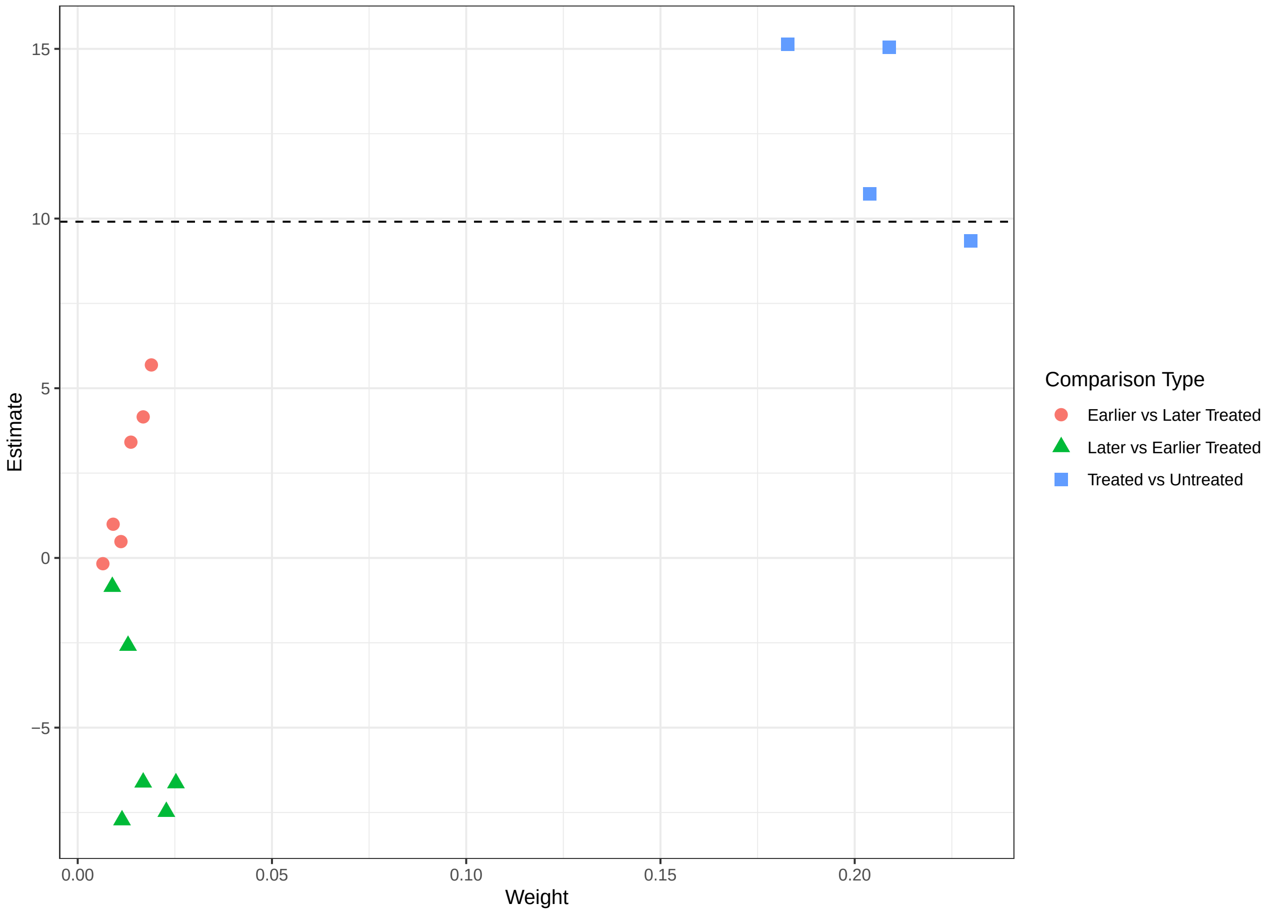
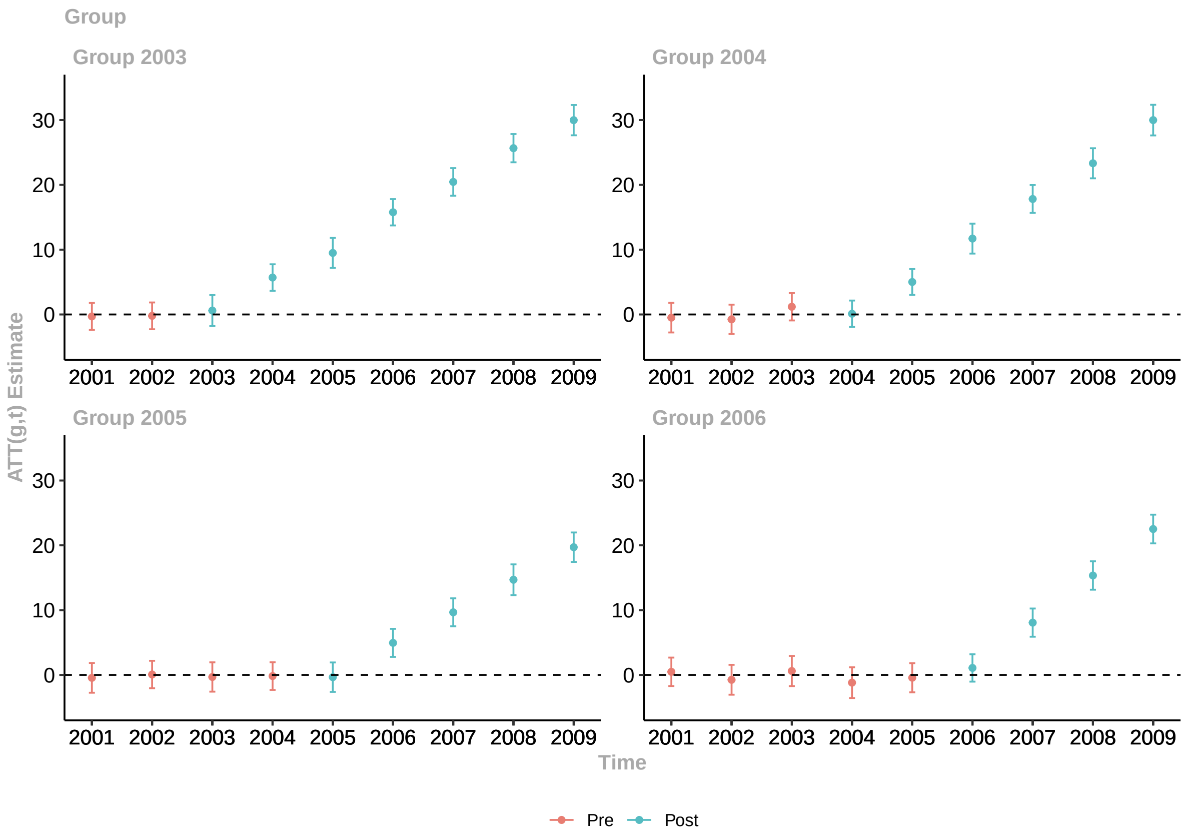


Figure 4: Group-Time Average Treatment on Treated Effects


Group
Group 2003
Group 2004
Group 2005
Group 2006
ATT(g,t) Estimate
Time
Pre
Post