# Introduction to DiD with Multiple Time Periods

Brantly Callaway and Pedro H.C. Sant'Anna

2022-07-19

## Introduction

Difference-in-differences is one of the most common approaches for identifying and estimating the causal effect of participating in a treatment on some outcome.

The "canonical" version of DiD involves two periods and two groups. The *untreated group* never participates in the treatment, and the *treated group* becomes treated in the second period.

However, much applied work deals with cases where there are more than two time periods and different units can become treated at different points in time. Regardless of the number of time periods, by far the leading approach in applied work is to try to estimate the effect of the treatment using a two-way fixed effects (TWFE) linear regression. This works great in the case with two periods, but there are a number of recent methodological papers that suggest that there may be substantial drawbacks to using TWFE with multiple time periods.

This vignette briefly discusses the emerging literature on DiD with multiple time periods – both issues with standard approaches as well as remedies for these potential problems. The `did` package implements a number of these remedies. A vignette for how to use the `did` package is available here. The background article for these vignettes is Callaway and Sant'Anna (2021), "Difference-in-Differences with Multiple Time Periods".

## Background

To start with, we'll consider some background material in this section. First, we'll discuss DiD with two time periods and two groups – this is the "canonical" case of DiD. Second, we briefly consider issues with TWFE linear regressions when there are multiple time periods.

### DiD with 2 Periods and 2 Groups

The baseline case for DiD is the one with two periods (let's call these periods $t$ and $t-1$) and two groups (a treated group and an untreated group).

**Notation / Setup**

- For $s \in \{t, t-1\}$, $Y_{is}(0)$ is unit $i$'s *untreated potential outcomes* – this is the outcome that unit $i$ would experience in period $s$ if they *did not* participate in the treatment
- For $s \in \{t, t-1\}$, $Y_{is}(1)$ is unit $i$'s *treated potential outcome* – this is the outcome that unit $i$ would experience in period $s$ if they *did* participate in the treatment.
- Set $D = 1$ for units in the treated group and $D = 0$ for units in the untreated group

- In the first period, no one participates in the treatment. In the second period, units in the treated group become treated. This means that observed outcomes are given by

  $$Y_{it-1} = Y_{it-1}(0) \quad \text{and} \quad Y_{it} = D_i Y_{it}(1) + (1 - D_i) Y_{it}(0)$$

  In other words, in the first period, we observe untreated potential outcomes for everyone (there is a no-anticipation assumption built in here). In the second period, we observe treated potential outcomes for units that actually participate in the treatment and untreated potential outcomes for units that do not participate in the treatment.

- The main parameter of interest in most DiD designs is the Average Treatment Effect on the Treated (ATT). It is given by

  $$ATT = E[Y_t(1) - Y_t(0) | D = 1]$$

  This is the difference between treated and untreated potential outcomes, on average, for units in the treated group.

The main assumption in DiD designs is called the parallel trends assumption:

**Parallel Trends Assumption**

$$E[Y_t(0) - Y_{t-1}(0) | D = 1] = E[Y_t(0) - Y_{t-1} | D = 0]$$

In words, this assumption says that the change (or “path”) in outcomes over time that units in the treated group *would have experienced if they had not participated in the treatment* is the same as the path of outcomes that units in the untreated group actually experienced. The parallel trends assumption allows for the level of untreated potential outcomes to differ across groups and is consistent with, for example, fixed effects models for untreated potential outcomes where the mean of the unobserved fixed effect can be different across groups.

This assumption is potentially useful because the path of untreated potential outcomes for units in the treated group (the term on the left in the above equation) is not known, but the researcher does observe the path of untreated potential outcomes for units in the untreated group (term on the right in the above equation). In fact, it is straightforward to show that, under the parallel trends assumption, the $ATT$ is identified and given by

$$ATT = E[Y_t - Y_{t-1} | D = 1] - E[Y_t - Y_{t-1} | D = 0]$$

That is, the $ATT$ is the difference between the mean change in outcomes over time experienced by units in the treated group adjusted by the mean change in outcomes over time experienced by units in the untreated group; the latter term, under the parallel trends assumption, is what the path of outcomes for units in the treated group would have been if they had not participated in the treatment.

## Two way fixed effects regressions

Now let’s move to a more general case where there are $\mathcal{T}$ total time periods. Denote particular time periods by $t$ where $t = 1, \dots, \mathcal{T}$.

By far the most common approach to *trying* to estimate the effect of a binary treatment in this setup is the TWFE linear regression. This is a regression like

$$Y_{it} = \theta_t + \eta_i + \alpha D_{it} + v_{it}$$

where $\theta_t$ is a time fixed effect, $\eta_i$ is a unit fixed effect, $D_{it}$ is a treatment dummy variable, $v_{it}$ are time varying unobservables that are mean independent of everything else, and $\alpha$ is presumably the parameter of interest. $\alpha$ is often interpreted as the *average effect* of participating in the treatment.

Although this is essentially a standard approach in applied work, there are a number of recent papers that point out potentially severe drawbacks of using the TWFE estimation procedure. These include: Borusyak and Jaravel (2018), Goodman-Bacon (2021), de Chaisemartin and D'Haultfoeuille (2020), and Sun and Abraham (2021).

**When will TWFE work?**

1. Effects really aren't heterogeneous. If the effect of participating in the treatment really is $\alpha$ for all units, TWFE will work great. That being said, in many applications, treatment effects are very likely to be heterogeneous – they may vary across different units or exhibit dynamics or change across different time periods. In particular applications, this is worth thinking about, but, at least in our view, we think that heterogeneous effects of participating in some treatment is the leading case.
2. There are only two time periods. This is the canonical case (2 periods, one group becomes treated in the second period, the other is never treated). In this case, under parallel trends an no-anticipation, $\alpha$ is going to be numerically equal to the $ATT$. In other words, in this case, even though it looks like you have restricted the effect of participating in the treatment to be the same across all units, TWFE exhibits *robustness* to treatment effect heterogeneity. Unfortunately, this robustness to treatment effect heterogeneity does not continue to hold when there are more periods and groups become treated at different points in time.

**Why is TWFE not robust to treatment effect heterogeneity?**

There are entire papers written about this, see, e.g., Borusyak and Jaravel (2018), Goodman-Bacon (2021), de Chaisemartin and D'Haultfoeuille (2020), and Sun and Abraham (2021). But here is the short version: in a TWFE regression, units whose treatment status doesn't change over time serve as the comparison group for units whose treatment status does change over time. With multiple time periods and variation of treatment timing, some of these comparisons are:

- newly treated units relative to ``never treated" units (good!)
- newly treated units relative to ``not-yet treated" units (good!)
- newly treated units relative to already treated units (bad!!!)

The first of these two comparisons are good (or at least in the spirit of DiD) in that they take the path of outcomes experienced by units that become treated and adjust it by the path of outcomes experienced by units that are not participating in the treatment. The third comparison is different though: it adjusts the path of outcomes for newly treated units by the path of outcomes for already treated units. But this is not the path of untreated potential outcomes, it includes *treatment effect dynamics*. Thus, these dynamics appear in $\alpha$, *making it very hard to give a clear causal interpretation*.

And this issue can have potentially severe consequences. For example, it is possible to come up with examples where the effect of participating in the treatment is positive for all units in all time periods, but the TWFE estimation procedure leads to estimating a negative effect of participating in the treatment. Even in the case where ``negative weights" can be ruled out, $\alpha$ recover a weighted average of $ATT's$, though these weights are hard to interpret.

## Treatment Effects in Difference in Differences Designs with Multiple Periods

In light of the potential problems with TWFE regressions in DiD designs with multiple periods, are there alternative approaches that can be used in this case?

Yes, and it turns out that it is not all that complicated! It is just a matter of using the ``good/desirable" comparisons between groups instead of all possible comparisons.

To fix ideas, let's provide some extended notation and be clear about the identifying assumptions that we are going to make.

**Notation**

- $Y_{it}(0)$ is unit $i$'s untreated potential outcome. This is the outcome that unit $i$ would experience in period $t$ if they do not participate in the treatment.
- $Y_{it}(g)$ is unit $i$'s potential outcome in time period $t$ if they become treated in period $g$.
- $G_i$ is the time period when unit $i$ becomes treated (often *groups* are defined by the time period when a unit becomes treated; hence, the $G$ notation).
- $C_i$ is an indicator variable for whether unit $i$ is in a **never-treated** group.
- $D_{it}$ is an indicator variable for whether unit $i$ has been treated by time $t$.
- $Y_{it}$ is unit $i$'s observed outcome in time period $t$. For units in the never-treated group, $Y_{it} = Y_{it}(0)$ in all time periods. For units in other groups, we observe $Y_{it} = \mathbf{1}\{G_i > t\}Y_{it}(0) + \mathbf{1}\{G_i \leq t\}Y_{it}(G_i)$. The notation here is a bit complicated, but in words, we observe untreated potential outcomes for units that have not yet participated in the treatment, and we observe treated potential outcomes for units once they start to participate in the treatment (and these can depend on *when* they became treated). Implicit in this notation there is a **no treatment anticipation** assumption, which can be relaxed as discussed in Callaway and Sant'Anna (2021), "Difference-in-Differences with Multiple Time Periods".
- $X_i$ vector of pre-treatment covariates.

## Main Assumptions

**Staggered Treatment Adoption Assumption** Recall that $D_{it} = 1$ if a unit $i$ has been treated by time $t$ and $D_{it} = 0$ otherwise. Then, for $t = 1, \dots, \mathcal{T} - 1$, $D_{it} = 1 \implies D_{it+1} = 1$.

Staggered treatment adoption implies that once a unit participates in the treatment, they remain treated. In other words, units do not "forget" about their treatment experience. This is a leading case in many applications in economics. For example, it would be the case for policies that roll out to different locations over some period of time. It would also be the case for many unit-level treatments that have a "scarring" effect. For example, in the context of job training, many applications consider participating in the treatment *ever* as defining treatment.

Within the DiD context, we believe it is hard to analyze non-staggered treatment setups **without** further restricting treatment effect heterogeneity across time, groups, treatment sequences, etc. That is the main reason we focus on this leading case.

**Parallel Trends Assumption based on never-treated units** For all $g = 2, \dots, \mathcal{T}$, $t = 2, \dots, \mathcal{T}$ with $t \geq g$,

$$E[Y_t(0) - Y_{t-1}(0)|G = g] = E[Y_t(0) - Y_{t-1}(0)|C = 1]$$

This is a natural extension of the parallel trends assumption in the two periods and two groups case. It says that, in the absence of treatment, average untreated potential outcomes for the group first treated in time $g$ and for the "never treated" group would have followed parallel paths in all post-treatment periods $t \geq g$.

Note that the aforementioned parallel trend assumption rely on using the ``never treated" units as comparison group for all "eventually treated" groups. This presumes that (i) a (large enough) "never-treated" group is available in the data, and (ii) these units are "similar enough" to the eventually treated units such that they can indeed be used as a valid comparison group. In situations where these conditions are not satisfied, one can use an alternative parallel trends assumption that uses the **not-yet treated** units as valid comparison groups.

**Parallel Trends Assumption based on not-yet treated units** For all $g = 2, \dots, \mathcal{T}$, $s, t = 2, \dots, \mathcal{T}$ with $t \geq g$ and $s \geq t$

$$E[Y_t(0) - Y_{t-1}(0) | G = g] = E[Y_t(0) - Y_{t-1}(0) | D_s = 0, G \neq g]$$

In plain English, this assumption states that one can use the not-yet-treated by time $s$ ($s \geq t$) units as valid comparison groups when computing the average treatment effect for the group first treated in time $g$. In general, this assumption uses more data when constructing comparison groups. However, as noted in Marcus and Sant'Anna (2021), this assumption does restrict some pre-treatment trends across different groups. In other words, there is no free-lunch.

## Group-Time Average Treatment Effects

The above assumptions are natural extensions of the identifying assumptions in the two periods and two groups case to the multiple periods case.

Likewise, a natural way to generalize the parameter of interest (the ATT) from the two periods and two groups case to the multiple periods case is to define **group-time average treatment effects**:

$$ATT(g, t) = E[Y_t(g) - Y_t(0) | G = g]$$

This is the average effect of participating in the treatment for units in group $g$ at time period $t$. Notice that when there are two time periods and two groups (the canonical case), the average treatment effect on the treated is given by $ATT = ATT(g = 2, t = 2)$.

To give a couple more examples, suppose that a researcher has access to three time periods. Then, $ATT(g = 2, t = 3)$ is the average effect of participating in the treatment for the group of units that become treated in time period 2, in time period 3. Similarly, $ATT(g = 3, t = 3)$ is the average effect of participating in the treatment for the group of units that become treated in time period 3, in time period 3.

### Identification of Group-Time Average Treatment Effects

Under either version of the parallel trends assumptions mentioned above, it is straightforward to show that group-time average treatment effects are identified. For instance, when one impose the parallel trends assumption based on "never-treated units", we have that, for all $t \geq g$

$$ATT(g, t) = E[Y_t - Y_{g-1} | G = g] - E[Y_t - Y_{g-1} | C = 1].$$

Alternatively, when one impose the parallel trends assumption based on "not-yet-treated units", we have that, for all $t \geq g$

$$ATT(g, t) = E[Y_t - Y_{g-1} | G = g] - E[Y_t - Y_{g-1} | D_t = 0, G \neq g].$$

These group-time average treatment effects are the building blocks of understanding the effect of participating in a treatment in DiD designs with multiple time periods.

## Parallel Trends Conditional on Covariates

In many cases, the parallel trends assumption is substantially more plausible if it holds after conditioning on observed pre-treatment covariates. In other words, if the parallel trends assumptions are modified to be

**Conditional Parallel Trends Assumption based on never-treated units** For all $g = 2, \dots, \mathcal{T}$, $t = 2, \dots, \mathcal{T}$ with $t \geq g$,

$$E[Y_t(0) - Y_{t-1}(0)|X, G = g] = E[Y_t(0) - Y_{t-1}(0)|X, C = 1]$$

**Parallel Trends Assumption based on not-yet treated units** For all $g = 2, \dots, \mathcal{T}$, $s, t = 2, \dots, \mathcal{T}$ with $t \geq g$ and $s \geq t$

$$E[Y_t(0) - Y_{t-1}(0)|X, G = g] = E[Y_t(0) - Y_{t-1}(0)|X, D_s = 0, G \neq g]$$

These parallel trends assumptions are the conditional analogues of previous ones. Importantly, they allow for covariate-specific trends in outcomes across groups, which can be particularly important in setups where the distribution of covariates varies across groups.

An example of a case where this assumption is attractive is one where a researcher is interested in estimating the effect of participating in job training on earnings. In that case, if the path of earnings (in the absence of participating in job training) depends on things like education, previous occupation, or years of experience (which it almost certainly does), then it would be important to condition on these types of variables in order to make parallel trends more credible.

In this case, the parameter of interest is still often the $ATT(g,t)'s$ (or their aggregation). It is still straightforward to identify and estimate the $ATT$ in this case. Basically, one needs to estimate the change in outcomes for units in the untreated group conditional on $X$, but average out $X$ over the distribution of covariates for individuals in group $g$ to obtain $ATT(g,t)$ (see Callaway and Sant'Anna (2021) and references therein for many more details). In practice, you can use different approaches to recover these parameters. More precisely, you can estimate the $ATT(g,t)'s$ using outcome-regressions, inverse probability weighting, or doubly-robust methods. But the **did** package automates all of this for the user.

## Aggregating Group-Time Average Treatment Effects

Group-time average treatment effects are natural parameters to identify in the context of DiD with multiple periods and multiple groups. But in many applications, there may be a lot of them. There are some benefits and costs here. The main benefit is that it is relatively straightforward to think about heterogeneous effects across groups and time using group-time average treatment effects. On the other hand, it can be hard to summarize them (e.g., they are not just a single number).

In our paper, Callaway and Sant'Anna (2021), "Difference-in-Differences with Multiple Time Periods", we propose a number of ways to aggregate group-time average treatment effects. Here, we will just consider a few important ones that we think applied researchers are most often interested in. First, consider the average effect of participating in the treatment, separately for each group. This is given by

$$\theta_S(g) = \frac{1}{\mathcal{T} - g + 1} \sum_{t=2}^{\mathcal{T}} \mathbf{1}\{g \leq t\} ATT(g,t).$$

This parameter may be of interest in its own right, since it allows one to highlight treatment effect heterogeneity with respect to treatment adoption period. Furthermore, it is fairly straightforward to further aggregate $\theta_S(g)$ to get an easy-to-interpret overall effect parameter,

$$\theta_S^O := \sum_{g=2}^{\mathcal{T}} \theta_S(g) P(G=g).$$

$\theta_S^O$ is the overall effect of participating in the treatment across all groups that have ever participated in the treatment. In our view, this is close to being a multi-period analogue of the $ATT$ in the two period case. Thus, if a researcher is constrained to report a single treatment effect summary parameter, we recommend reporting $\theta_S^O$.

In DiD setups with multiple periods, it is natural to ask "How does treatment effects vary with elapsed treatment time?" Here, note that researchers are interested in understanding treatment effect dynamics. This is at the heart of event-study-type of analysis that is widespread in applied work.

In this case, a natural way to aggregate the group-time average treatment effect to highlight treatment effect dynamics is given by

$$\theta_D(e) := \sum_{g=2}^{\mathcal{T}} \mathbf{1}\{g+e \leq \mathcal{T}\} ATT(g, g+e) P(G=g | G+e \leq \mathcal{T}).$$

This is the average effect of participating in the treatment for the group of units that have been exposed to the treatment for exactly $e$ time periods.

All of these aggregations are available in the **did** package and examples with real data are available in our Getting Started with the **did** Package vignette. In Callaway and Sant'Anna (2021), we also discuss additional aggregation schemes. We encourage you to take a look!

## Conclusion

This vignette has covered basic background issues on DiD with multiple periods. Callaway and Sant'Anna (2021) discusses many extensions and these are all provided in the **did** package as well. See our User Guides for more details.

## References

- Borusyak, Kirill, and Xavier Jaravel. "Revisiting Event Study Designs". Available at SSRN 2826228 (2018)
- Callaway, Brantly, and Pedro H. C. Sant'Anna. "Difference-in-differences with multiple time periods." Journal of Econometrics, Vol. 225, No. 2, pp. 200-230, 2021.
- de Chaisemartin, Clement, and Xavier d'Haultfoeuille. "Two-way fixed effects estimators with heterogeneous treatment effects." American Economic Review 110.9 (2020): 2964-96.
- Goodman-Bacon, Andrew. Difference-in-differences with variation in treatment timing." Journal of Econometrics, Vol. 225, No. 2, pp. 254-277, 2021
- Marcus, Michelle, and Pedro H. C. Sant'Anna. "The Role of Parallel Trends in Event Study Settings: An Application to Environmental Economics". Journal of the Association of Environmental and Resource Economists, Vol. 8, No. 2, pp. 235-275, 2021

- Sun, Liyang, and Sarah Abraham. “Estimating dynamic treatment effects in event studies with heterogeneous treatment effects.” Journal of Econometrics, Vol. 225, No. 2, pp. 175-199, 2021