8 Difference-in-Differences

History and background

One of the first recorded (and also the most famous) examples of using the difference-in-differences (DD) design to answer a causal question is John Snow's work on the origins of cholera. He challenged the received knowledge at that time that cholera was spread through air and showed instead that cholera was spread by fecally-contaminated water in his 1855 work titled "On the Mode of Communication of Cholera".

Cholera is a bacterial disease of the small intestine with acute symptoms such as vomiting and diarrhea. In the 19th century, it was usually fatal. There were three main epidemics that hit London at that time, tens of thousands of people suffered and died from the disease, and doctors could not help the victims.

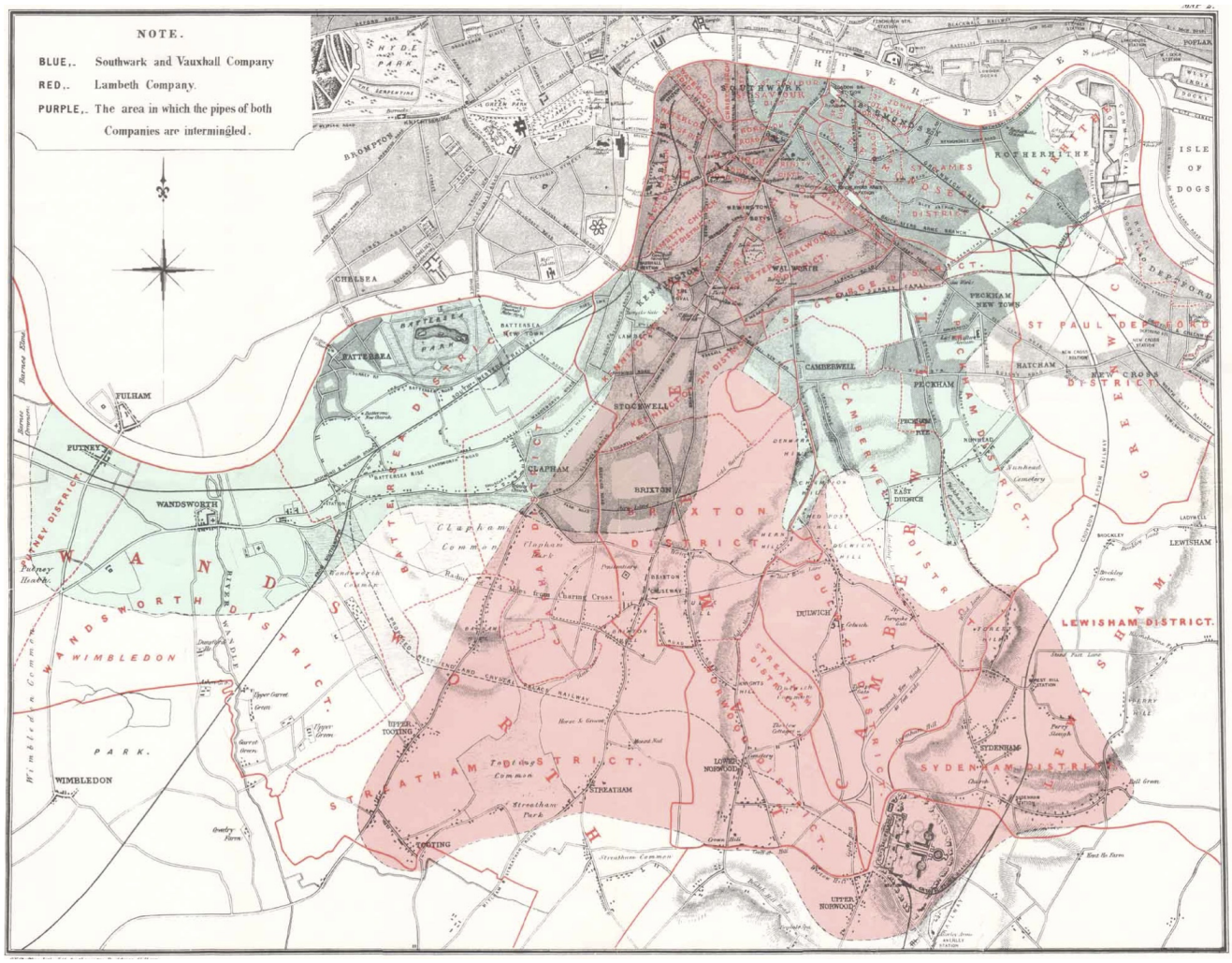
The dominant theory about cholera transmission at that time was the *miasma theory*, which said diseases were spread by microscopic poisonous particles that infected people by floating through the air. These particles were thought to be inanimate, and because microscopes at that time had incredibly poor resolution, it would be years before microorganisms would be seen. Treatments, therefore, tended to be designed to stop poisonous dirt from spreading through the air. But tried and true methods like quarantining the sick were strangely ineffective at slowing down cholera.

John Snow worked in London during these epidemics. Originally he accepted the miasma theory and tried many solutions based on the theory to block the airborne poisons from reaching other people. He covered the sick with burlap bags, for instance, but the disease still spread. Faced with the theory's failure to explain and fight against cholera, he changed his mind and began looking for a new explanation.

John Snow came to believe that cholera spread by dirty drinking water. He had a few ways of providing evidence, one of which is very similar to a modern-day difference-in-differences research design, and can be easily discussed in those terms (see Coleman (2019) for a review). He used a natural experiment in London during which one of the water companies changed its water intake.

London's water needs were served by a number of competing companies, who got their water intake from different parts of the Thames river. Water taken in from the parts of the Thames that were downstream of London contained everything that Londoners dumped in the river, including plenty of fecal matter from people infected with cholera. But in 1849, the Lambeth water company had moved its intake pipes upstream higher up the Thames, above the main sewage discharge point, thus giving its customers uncontaminated water. They did this to obtain cleaner water, but it had the added benefit of being too high up the Thames to be infected with cholera.

Snow realized that it had given him a natural experiment to test his hypothesis. If his theory was right, then the Lambeth houses should have lower cholera death rates than some other set of households whose water was infected. He found his explicit counterfactual in the Southwark and Vauxhall Waterworks Company that have not moved their intake point upstream and who served similar households. The map below is taken from Snow's work and shows the neighborhoods served by each company.



Snow collected the data on death rates (per 10,000 people in 1851), which are summarized in the table below.

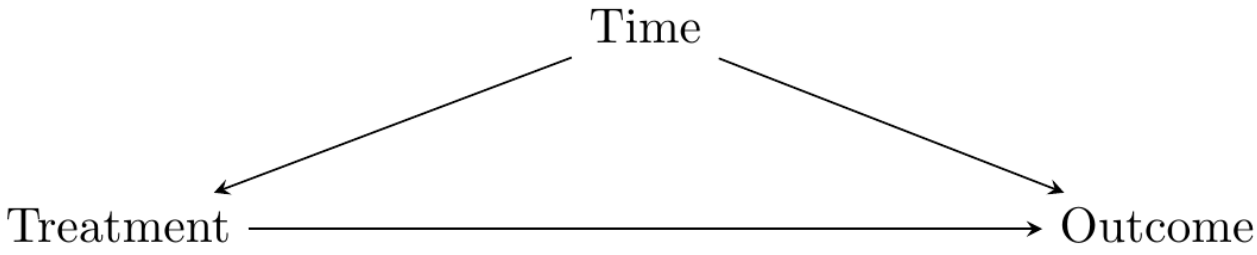
Company	Death rate in 1849	Death rate in 1854
Southwark and Vauxhall	135	147
Lambeth	85	19

If we compare cholera death rate across the two years for Southwark and Vauxhall, we see that the rate increased by 12. During the same time period, the death rate for Lambeth dropped by 66. If Lambeth followed the same time trend as Southwark and Vauxhall, then its death rate should have increased by 12 to 97 (85+12). Then the predicted (counterfactual) death rate in 1854 for Lambeth is 97, while the actual death rate is 19, which gives us an average effect of -78 (19 - 97).

We can follow an alternative logical path to arrive at the same conclusion. Notice that there are pre-treatment differences between Southwark + Vauxhall and Lambeth. The death rate in Lambeth in 1849 is lower than the death rate in Southwark and Vauxhall by 50 (135-85). If we assume that the difference between the companies remains the same over time, then in 1854 we should have expected the death rate for Lambeth at 97 (147-50). This is, again, the counterfactual death rate. Instead, the observed death rate for Lambeth was 19, which again gives us an effect of -78.

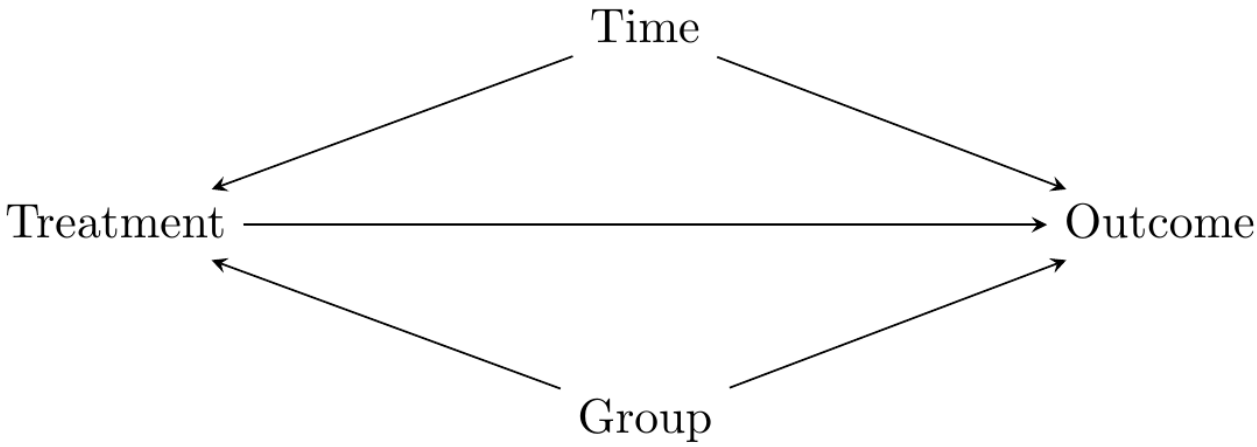
Causal diagram of DD

Many treatments occur at a particular time. We can see the outcomes before the treatment is applied, and after. We want to know how much of the change in the outcome is due to that treatment. The obvious back door path we have to deal with is through *time*.



Identifying the effect of *Treatment* on *Outcome* requires us to close the back door that goes through *Time*. But we can't do this entirely, because all the variation in *Treatment* is explained by *Time*. You're either in a pre-treatment time and untreated, or in a post-treatment time and treated.

DD design brings in another group that is never treated. So now in the data we have both the group that receives treatment at a certain point, and another group that never receives treatment. At first this seems counterintuitive—that untreated group may be different from the treated group! We have introduced a second back door path.



Seems like we've made things worse by introducing the control group. The key is this, though: now that we have that untreated group, even though we've added a new back door, we can now close both back doors. First, we isolate the within variation (across time) for both the treated group and untreated group. Because we have isolated within variation,

we are controlling for group differences and closing the back door through *Group*. Second, we compare the within variation in the treated group to the within variation in the untreated group. Because the within variation in the untreated group is affected by time, doing this comparison controls for time differences and closes the back door through *Time*.

We are looking for how much more the treated group changed than the untreated group when going from before to after. The change in the untreated group represents how much change we would have expected in the treated group if no treatment had occurred. So any additional change beyond that amount must be the effect of the treatment.

Two-by-two DD

Since we have variation across both groups and time, we will index the outcomes accordingly. Let $\mathbb{E}[Y_{it}]$ be the expected outcome of group i at time t. In a simple case with two groups and two dates, the group index i will take values of T (treated, the group that experienced treatment between the two dates) or U (untreated, the group that did not experience treatment), and the time index t will take values of 0 (pre-treatment date) and 1 (post-treatment date). This gives us four expected outcomes in total that we can arrange in a 2x2 table.

	$t=0$	$t=1$	**Difference**
$i=U$	$\mathbb{E}[Y_{U0}]$	$\mathbb{E}[Y_{U1}]$	$\mathbb{E}[Y_{U1}]-\mathbb{E}[Y_{U0}]$
$i=T$	$\mathbb{E}[Y_{T0}]$	$\mathbb{E}[Y_{T1}]$	$\mathbb{E}[Y_{T1}]-\mathbb{E}[Y_{T0}]$
Difference	$\mathbb{E}[Y_{T0}]-\mathbb{E}[Y_{U0}]$	$\mathbb{E}[Y_{T1}]-\mathbb{E}[Y_{U1}]$	$\mathbb{E}[Y_{T1}]+\mathbb{E}[Y_{U0}]-\mathbb{E}[Y_{T0}]-\mathbb{E}[Y_{U1}]$

Within each row of the table, we can take the time difference (post-treatment minus pre-treatment). Within each column of the table, we can take the group difference (treated minus untreated). To get at the treatment effect, we difference those differences. We can difference either the Difference row (post-treatment minus pre-treatment) or the Difference column (treated minus untreated). The end result in the bottom-right cell is the same for each of these differences.

To simplify the formulas a little, suppose that the difference in expected outcomes at time $t=0$ between the treated and untreated groups is

$$\lambda \equiv \mathbb{E}[Y_{T0}]-\mathbb{E}[Y_{U0}].$$

Likewise, assume that the difference for the untreated group between the post- and pre-treatment dates is

$$\tau \equiv \mathbb{E}[Y_{U1}]-\mathbb{E}[Y_{U0}].$$

Then we can find the post-treatment outcome for the treated group in one of the two ways. First, we can assume that the difference in expected outcomes for the treated group between the dates is caused by the treatment (δ) and time (τ) effects

$$\mathbb{E}[Y_{T1}]-\mathbb{E}[Y_{T0}]=\delta+\tau.$$

Alternatively, we can assume that the difference in expected outcomes at the post-treatment date between the treated and untreated groups is caused by the treatment (δ) and group (λ) effects

$$\mathbb{E}[Y_{T1}] - \mathbb{E}[Y_{U1}] = \delta + \lambda.$$

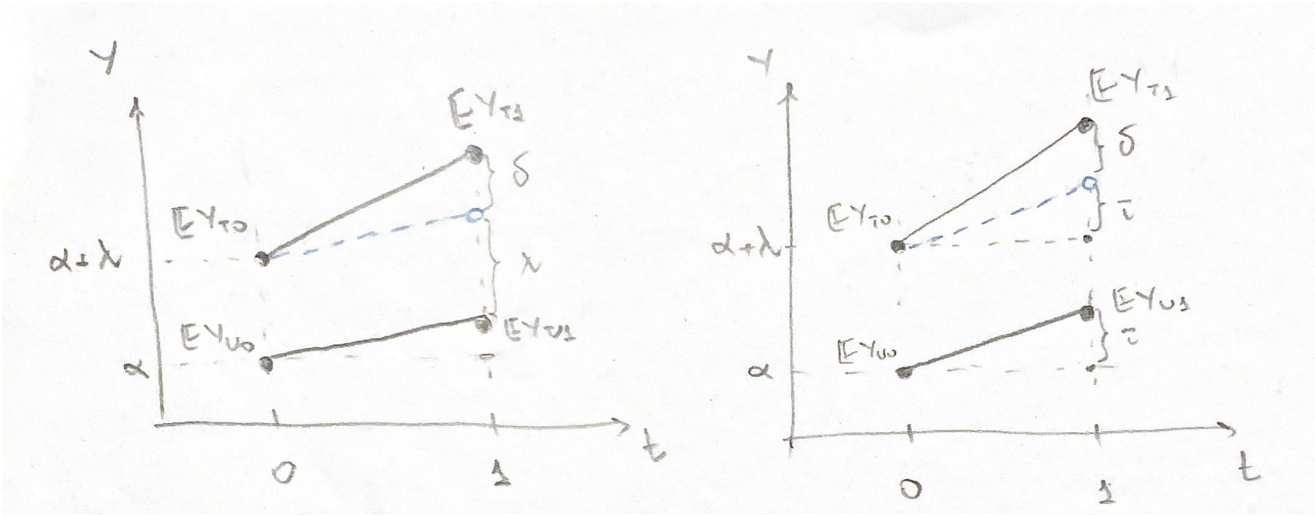
These two formulations for $\mathbb{E}[Y_{T1}]$ are equivalent.

Consider what happens to the table when we make the proposed substitutions.

	$t=0$	$t=1$	**Difference**
$i=U$	$\mathbb{E}[Y_{U0}]$	$\mathbb{E}[Y_{U1}]$	τ
$i=T$	$\mathbb{E}[Y_{T0}]$	$\mathbb{E}[Y_{T1}]$	$\delta+\tau$
Difference	λ	$\delta+\lambda$	δ

Regardless of how you compute the difference-in-differences (within-rows first, within-column second or within-columns first, within-row second), the end result is δ, the treatment effect.

Drawing time



We can represent the DD design visually. In the picture above (left panel), we compare the difference in expected outcomes within each date. At the post-treatment date, the difference in outcomes is due to the effect of the treatment (δ) and the group differences. So we use the pre-treatment group differences (λ) to construct the counterfactual (hollow dot) for what the treated group's post-treatment outcome would have been in the absence of treatment. The treatment effect that we get is the difference between the observed post-treatment outcome for the treated group, $\mathbb{E}[Y^1_{T1}]$, and that counterfactual, $\mathbb{E}[Y^0_{T1}]$.

Alternatively, the right panel shows that we can construct a counterfactual using the time differences. The difference between the post- and pre-treatment outcomes for the treated group is due to the effect of the treatment (δ) and the time differences. So we use the time difference for the untreated group (τ) to construct the counterfactual (hollow dot). The treatment effect is computed as in the previous case. Notice, however, that the two ways to

construct the counterfactual are identical, which can be seen from the geometry of the graph.

DD in action

A classic example of DD in labor economics is Card and Krueger (1994) who studied the effect of a minimum wage increase on employment. In theory, in a competitive labor market, increases in the minimum wage move us up a downward-sloping demand curve. Higher minimum wage therefore reduces employment, perhaps hurting the very workers these policies were designed to help.

Card and Krueger (1994) use a dramatic change in the New Jersey state minimum wage to see if this is true. On April 1, 1992, New Jersey raised the state minimum from $4.25 to $5.05. Card and Krueger collected data on employment at fast food restaurants in New Jersey in February 1992 and again in November 1992. These restaurants (Burger King, Wendy's, and so on) are big minimum-wage employers. Card and Krueger collected data from the same type of restaurants in eastern Pennsylvania, just across the Delaware river. The minimum wage in Pennsylvania stayed at $4.25 throughout this period. They used their data set to compute the DD effect of the New Jersey minimum wage increase.

The table below shows the average full-time equivalent (FTE) employment at restaurants in Pennsylvania (PA, untreated) and New Jersey (NJ, treated) before (February, $t = 0$) and after (November, $t = 1$) a minimum wage increase in New Jersey.

	February	November	Difference
$i = PA$	23.33	21.17	-2.16
$i = NJ$	20.44	21.03	0.59
Difference	-2.89	-0.14	2.75

We see that employment in PA restaurants is somewhat higher than in NJ in February but falls by November. Employment in New Jersey, in contrast, increases slightly. These two changes produce a positive DD effect, the opposite of what we might expect if a higher minimum wage pushes businesses up the labor demand curve.

Two-way fixed effects

In practice, the DD is estimated using a regression. For a simple two-by-two case, the DD estimated is equivalent to an interaction effect in the following model.

$$Y = \alpha + \lambda \mathbb{I}(i = T) + \tau \mathbb{I}(t = 1) + \delta \mathbb{I}(i = T)\mathbb{I}(t = 1) + \epsilon,$$

where $\mathbb{I}(\cdot)$ is the indicator function. You can easily verify that plugging all the four possible combinations of i and t will yield the four outcomes from the table, where α denotes the outcome $\mathbb{E}[Y_{U0}]$. The interaction effect tells us by how much bigger the group effect is when $t = 1$ relative to $t = 0$. Equivalently, it tells us by how much bigger the time effect is when

$i = T$ relative to $i = U$. Both of these interpretations are identical to the DD effect we introduced earlier using differences in expected outcomes.

The regression framework makes it easy to extend the model to include multiple periods or groups. The model will take the following form, which is a generalization of the model above

$$Y = \alpha_g + \alpha_t + \delta D + \epsilon.$$

This model is known as two-way fixed effects model. The "two-way" part comes from the fact that we are including the so-called fixed effects for both groups (α_g) and time (α_t). These fixed effects allow different groups and different dates to have their own conditional means. The D variable is an indicator equal to 1 if the observation is in the treated group after the treatment occurred.

> **Note**
>
> When estimating the standard errors in this model, it is common to cluster at the group level. Clustering accounts for the fact that we expect errors to be related within group over time, which can make standard errors a little overconfident (i.e., too small) if not accounted for.

The model can also include additional controls. However, notice that any control variables that vary over group but don't change over time are unnecessary and would drop out—we already have group fixed effects. The inclusion of time-varying controls, however, imposes some statistical problems related to whether those controls impact treated and untreated similarly, and, importantly, the assumption that treatment doesn't affect later values of covariates. If you need to include covariates, it's often a good idea to show your results both with and without them.

If you have a single treatment period, the regression can be an easy way to estimate difference-in-differences. The downside of this regression approach for estimating DD is that it doesn't work very well for the so-called *rollout designs* or *staggered treatment timing designs*, where the treatment is applied at different times to different groups. Researchers used the two-way fixed effects for these cases for a long time, but it turns out to not work as expected.

Treatment effect and assumptions

The graphical analysis we conducted already suggests an answer to a question of what kind of treatment effect DD produces. But let's look at it more formally. The DD effect is given by

$$\begin{aligned}\mathbb{E}[Y_{T1}] + \mathbb{E}[Y_{U0}] - \mathbb{E}[Y_{T0}] - \mathbb{E}[Y_{U1}] &= \mathbb{E}[Y^1_{T1}] + \mathbb{E}[Y^0_{U0}] - \mathbb{E}[Y^0_{T0}] - \mathbb{E}[Y^0_{U1}] \\ &\quad + \mathbb{E}[Y^0_{T1}] - \mathbb{E}[Y^0_{T1}] \\ &= \underbrace{\mathbb{E}[Y^1_{T1}] - \mathbb{E}[Y^0_{T1}]}_{ATT} \\ &\quad + \underbrace{\mathbb{E}[Y^0_{T1}] - \mathbb{E}[Y^0_{T0}]}_{\tau_T} - \underbrace{(\mathbb{E}[Y^0_{U1}] - \mathbb{E}[Y^0_{U0}])}_{\tau_U}.\end{aligned}$$

Therefore, the DD estimates an ATT (specifically, an ATT in the post-treatment period, unless the treatment effect is constant over time) plus a term that is a difference in the time effects for the treated (τ_T) and untreated (τ_U) groups. If the two time effects are not equal, the DD estimate is biased relative to the true ATT. Hence, the main identifying assumption of DD is that $\tau_T = \tau_U$. This assumption is called *parallel trends*.

> **Note**
>
> Another way to understand that DD identifies an ATT and not, say, ATU, is to recall that in the data only the treated group has control-state and treated-state outcomes. We only observe the control-state outcomes for the untreated group. Hence, we cannot get the ATU.

Notice that due to the symmetry of the design, we can rewrite the expression above as

$$\begin{aligned}\mathbb{E}[Y_{T1}] + \mathbb{E}[Y_{U0}] - \mathbb{E}[Y_{T0}] - \mathbb{E}[Y_{U1}] &= \underbrace{\mathbb{E}[Y^1_{T1}] - \mathbb{E}[Y^0_{T1}]}_{ATT} \\ &\quad + \underbrace{\mathbb{E}[Y^0_{T1}] - \mathbb{E}[Y^0_{U1}]}_{\lambda_1} - \underbrace{(\mathbb{E}[Y^0_{T0}] - \mathbb{E}[Y^0_{U0}])}_{\lambda_0}.\end{aligned}$$

In other words, the DD effect is ATT plus the difference between the groups effects post-treatment (λ_1) and pre-treatment (λ_0). Another way to state the parallel trends assumption is then to assume that the group effects do not change over time, $\lambda_1 = \lambda_0$.

The parallel trends assumption is what allows us to difference out the time and group effects in the table to get δ. It also allows us to construct the counterfactual in the graph. Without this assumption, DD does not identify any meaningful effect. Notice that the parallel trends assumption involves the counterfactual $\mathbb{E}[Y^0_{T1}]$. This is the control-state outcome for the treated group in the post-treatment period. We never observe this value, hence the identifying assumption of DD is inherently untestable. We can, however, use several diagnostic tests: the *test of prior trends* and the *placebo test*.

The test of prior trends looks at whether the treated and untreated groups already had differing trends in the lead-up to the treatment date. There are two ways to perform this test. The first is to graph the average outcomes over time in the pre-treatment period (this assumes that we have observations over multiple dates before treatment) and see if they look different.

The second is to perform a statistical test to see if the trends are different, and if so, how much different. The simplest form of this uses the regression model

$$Y = \alpha_g + \beta_1 t + \beta_2 t \times Group + \epsilon$$

estimated using only data from before the treatment period, where the interaction term allows the time trend to be different for each group. A test of $\beta_2 = 0$ provides information on whether the trends are different. This is the simplest specification, and you could look for more complex time trends by adding polynomial terms or other nonlinearities to the model.

When failing a prior trends test, some researchers will see this as a reason to add "controls for trends" to salvage their research design by including the *Time* variable in the model directly, rather than the time fixed effects. However, this can have the unfortunate effect of controlling away some of the actual treatment effect, especially for treatments with effects that get stronger or weaker over time.

For the placebo test, we can follow these steps:

1. Use only the data that came before the treatment went into effect
2. Pick a fake treatment period
3. Estimate the same DD model you were planning to use, but create the *Treated* variable as equal to 1 you're in the treated group and after the fake treatment date you picked
4. If you find an "effect" for that treatment date where there really shouldn't be one, that's evidence that there's something wrong with your design, which may imply a violation of parallel trends.

Another way to do this if you have multiple untreated groups is to use all the data, but drop the data from the treated groups. Then, assign different untreated groups to be fake treated groups, and estimate the DD effect for them. This approach is less common since it doesn't address parallel trends directly (and it's not really a problem if parallel trends fails among your untreated groups). However, this approach is a very common placebo test for the *synthetic control method*.