1 Two-way fixed effect regression and difference in difference estimator

1.1 Overview

The **Two-Way Fixed Effects Regression** (TWFEr) regression is often used in Economics and other social sciences to analyze causality from panel data. The main reasons to use this method are its ability to adjust for time, and unit effects and its resemblance to a generalization of the **Difference-in-Difference** (DiD) estimator in the two groups and two periods design. This section aims to show that outside of that design, the two estimators (TWFEr and DiD estimator) are very different, each with its own strengths and weaknesses. A better understanding of how these two estimators work and why they could give different results is a useful tool for applied research. To draw a proper comparison between the two, we will follow what has been done by Imai and Kim ((2021)) and proceed by using their equivalent matching estimators[1] as they excel in displaying counterfactuals for observation. Pictures from Imai and Kim ((2021)) will be used to show visually how the estimators use different observations for the counterfactual.

An observation's counterfactual is the value it would have reached if it belonged to the opposite treatment status. For example, if an observation is treated, its counterfactual is the value it would have reached if it was not treated. The opposite applies to control observations. By underlining the differences in how DiD and TWFEr compute counterfactuals, a clearer view of the behaviour of the two estimators will be given.

All of the analysis will be conducted in a different framework than before. The treatment variable is homogeneous, non-staggered (units can enter and leave the treatment at any time), all units can receive treatment, and our analysis will not be limited to two time periods.

The model that will be used in all of the paper will be:

$$Y_{i,t} = u_i + v_t + \theta X_{i,t} + e_{i,t} \tag{1}$$

[1]Part 2 and 3 are taken from the Imai and Kim ((2021)), while the idea of the leaving effect from Imai et al. ((2021)) its formalization, for what it counts, is my own while the later simulations and the combination they express are my own.

Our model is composed of a time-invariant unit effect u_i, a unit invariant time effect v_t, an error term e_{it} and a dummy variable X_{it} (one if treated, zero if untreated, treatment is group independent and assigned casually to observations, all observations with $X_{it} = 1$ will be called "treated". In contrast, all the ones with $X_{it} = 0$ "untreated") this is the only covariate present in the model and the only one the matching estimators will care about when matching observations.

1.2 Two-way fixed effect regression

$$\begin{aligned}\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}[\{(Y_{i,t} - \bar{Y}) - (\bar{Y_i} - \bar{Y}) - (\bar{Y_t} - \bar{Y})\} \\ - \theta\{(X_{i,t} - \bar{X}) - (\bar{X_i} - \bar{X}) - (\bar{X_t} - \bar{X})\}]^2\end{aligned} \tag{2}$$

This estimator has been analyzed in Section 3.3. We already stressed how it can adjust for time-invariant unit effects and unit-invariant time effects and their limits. Instead, the focus will be on how the method estimates the treatment effect θ. To do so, we reach the equivalent matching method through a series of equations (the equivalence has been shown in Imai and Kim ((2021)) and is outside this paper's scope).

$$\hat{\theta} = \frac{1}{K}\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{i,t}(Y_{i,t} - \widehat{Y_{i,t}(0)}) + (1 - X_{i,t})(\widehat{Y_{i,t}(1)} - Y_{i,t})\right)\right] \tag{3}$$

As shown above, matching methods try to compare each observation, given its treatment status, with its **potential outcome of opposite treatment status (counterfactual)**[2] "what the outcome variable would have been if everything was the same but the treatment". Treated observations will use the first difference inside the formula ($X_{i,t} = 1$) and untreated will use the second ($1 - X_{i,t} = 1$), the difference between the observation and its counterfactual represent the treatment effect. To estimate the counterfactual of an observation, we use the other observations available in our sample.

It is important to notice that the model object of our analysis is composed of unit effect, time effect, treatment and error. No other covariates are present. The only covariate available, as a consequence, the only one that the matching estimator cares about, is the

[2] $\widehat{Y_{i,t}(x)}, x\epsilon\{0,1\}$ is the potential outcome of unit i at time t with treatment status x, it is called counterfactual when for an observation the potential outcome has opposite treatment status compared to $X_{i,t}$.

treatment status. To build a proper counterfactual, we would need the potential outcome for that observation (hence, the same unit and time) but with the opposite treatment status. This is why it is possible to find an **equivalent matching method** without relying on propensity score matching or other techniques to make our counterfactual closer in covariates values to the original observation.

Here below, we present how the matching-TWFEr equivalent method computes the counterfactual of a specific observation, with x being the treatment status ($x \in \{0, 1\}$).

$$\widehat{Y_{it}(x)} = \frac{1}{T-1}\sum_{t' \neq t} Y_{i,t'} + \frac{1}{N-1}\sum_{i' \neq i} Y_{i',t} - \frac{1}{(T-1)(N-1)}\sum_{i' \neq i}\sum_{t' \neq t} Y_{i',t'} \tag{4}$$

Equation (4) computes the potential value of[3] of Y_{it} by averaging over all the observations with the same unit (excluding same time) plus all of the observations with the same time (excluding same unit) and by subtracting the mean of all the observations that share neither unit nor time.

The model (1) is made of unit FE, time FE, treatment status (multiplied by θ), and the counterfactual of a treated observation would ideally contain all of the terms as above but with the opposite treatment status. A graphical illustration Figure 1 will make clearer which observations are used for the counterfactual estimate of $Y_{4,3}$.

As shown in formula 4, TWFEr uses all but the same observation to compute its counterfactual, this implies that even units with the same treatment status (so-called mismatches[4]) do matter for our counterfactual. The treatment status of the counterfactual will not be the opposite, but the sum of the three averages[5]. This leads to a biased counterfactual.

Including observations with the same treatment status as the selected observation implies that the difference between a unit and its counterfactual will be smaller than expected. For this reason, Imai and Kim ((2021)) named it **"attenuation bias"**.

Even though the TWFEr counterfactual could be biased, this will not be the case for

[3]By using as an example a 2 by 2 design (easy to expand to a larger design) with $X_{1,1}, X_{1,2}, X_{2,1} = 0$ and $X_{2,2} = 1$ the counterfactual $\widehat{Y_{2,2}(0)} = t_2 + u_1 + u_2 + t_1 - u_1 - t_1$, if a treated unit were to be present, the treatment effect θ would have been inside the past formula.

[4]A mismatch is an observation that is used to estimate the counterfactual of another observation with which shares the same treatment status.

[5]In the example above the treatment status of $\widehat{Y_{4,3}(0)}$ is $\frac{1}{3} + \frac{2}{4} - \frac{7}{12}$

Figure 1: Two-way Fixed Effect Estimator

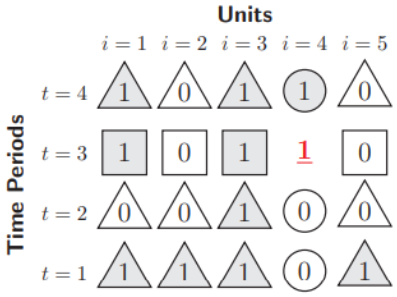
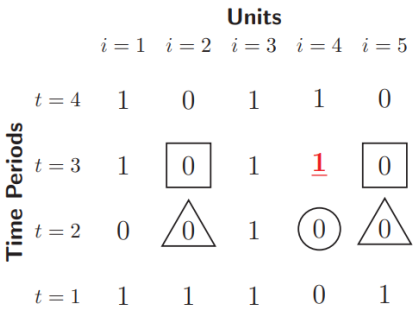


Figure 2: Differences in Differences Estimator



In the first figure one represent treated units while zero control units. The observation for which we want to compute the counterfactual is the red one. In this case, observations that share the same units are circles (first member in equation (4)), squares share the same time (second term in (4)), while triangles do not share either time or unit (last member in (4)). For example in this figure, $Y_{4,3}$ has many mismatches, one shares the same unit ($Y_{4,4}$), two with the same time $Y_{3,3}, Y_{1,3}$ and more in the adjustment set. On the other the second pictures does not have any mismatch, but relies on a smaller number of observations.

the estimate of the treatment effect ($\hat{\theta}$). In the $\hat{\theta}$ of the equivalent matching estimator, TWFEr adjusts for this bias through K[6], counting where and how many correct matches are present when computing $\hat{\theta}$ and rebalancing[7]. The following points are relevant to the analysis and are the object of a comparison with the DiD estimator. All the observations in the dataset contribute to the estimate of a single counterfactual, even the ones with the same treatment status.

Even if observations do not contribute with the same weight for a single counterfactual (sharing the same time or same unit implies a bigger weight), they will have the same weight since all the counterfactuals of all observations matter. The estimator's variance is mainly given by the error in the realized $Y_{i,t}$. The errors in the counterfactual are less relevant to our ends.

1.3 Differences in Differences

Since our dataset is made of more than two periods, the DiD estimator will be different from the one we have seen before in the paper (2×2 design), even though the idea will be the same.

With DiD, we also need to assume parallel trends. This assumption has already been

[6] formula is left to the appendix

[7] The following statement comes from OLS regression estimating a dummy variable without any bias

investigated in Section 5 before. In the DiD estimator comparing each observation with its counterfactual is intuitive[8]. It is possible to estimate the treatment effect only on units that change their treatment status in the sample. Those units can be considered only if at least one unit had the same treatment status in t-1 and kept it in t. Without the presence of this unit, it would be impossible to build a counterfactual and to have a comparison to our unit.

We use control observation with the same time, with the same unit (but one time before) and other control observations that share unit and time with the one before. Since the DiD estimator computes a counterfactual only for units that received treatment, our method will compute the average treatment effect on the treated (ATT).

$$\tau = E[Y_{i,t}(1) - Y_{i,t}(0)|X_{i,t} = 1, X_{i,t-1} = 0]$$

To compute the counterfactual, we introduce some notation:
$M_{it}^{DiD} = \{(i', t') : i' = i, t' = t - 1, X_{i't'} = 0\}$ the set containing the observation with the same unit, but that in a time period before and only if it was an untreated observation.
$N_{it}^{DiD} = \{(i', t') : i' \neq i, t' = t, X_{i't'} = 0, X_{i',t'-1} = 0\}$ the set containing observation with the same time, but different units, an observation belongs to this set if and only if it was untreated in time period t and t-1.
$A_{it}^{DiD} = \{(i', t') : i' \neq i, t' = t - 1, X_{i't'} = 0, X_{i',t'+1} = 0\}$ the set containing all the observation that are untreated in t-1 and t and have time t-1 and unit different from i.

$$\widehat{Y_{it}(0)} = Y_{it-1} + \frac{1}{|N_{it}^{DiD}|} \sum_{(i',t)\epsilon N_{it}^{DiD}} Y_{i't} - \frac{1}{|A_{it}^{DiD}|} \sum_{(i',t')\epsilon A_{it}^{DiD}} Y_{i't'} \tag{5}$$

The counterfactual of an observation (in (5) is for treated observations, as they are the only relevant for the DiD estimator to compute a potential outcome) in (5) is made of the outcome of the dependent variable of the same unit in a period before (the member of M_{it}^{DiD}) plus the average of the members of N_{it}^{DiD} minus the average of the members of A_{it}^{DiD}. It is coherent with the 2×2 design.

[8]Differences in Differences could be already considered as a matching estimator; by comparing a unit that goes from no treatment to treatment with one that stays untreated, we are building a counterfactual for that very same unit. Assume $X_{1,1}, X_{1,2}, X_{2,1} = 0 and X_{2,2} = 1$ then $Y_{2,2} - \widehat{Y_{2,2}(0)} = Y_{2,2} - (Y_{1,2} + Y_{2,1} - Y_{1,1}) = u_2 + v_2 + \theta - (u_1 + v_2 + u_2 - v_1 - u_1 + v_1) = \theta$

The variable $D_{i,t}$ tells us if an observation is suited to compute the ATT ($\hat{\theta}$) by taking into account if the observation is treated and if there exist other observations such that the two sets N_{it}^{DiD} and M_{it}^{DiD} are not empty.

Let's define function f for convenience: $f(x) = 1$, if $x > 0$, $f(x) = 0$ else

$$t = 1, D_{it} = 0,$$

$$t \neq 1, D_{it} = X_{it} * f(|N_{it}^{DiD}||M_{it}^{DiD}|)$$

As a consequence, the ATT formula will be:

$$\hat{\tau} = \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left(Y_{it} - \widehat{Y_{it}(0)}\right) \tag{6}$$

The ATT formula sums over all observations (through unit and time sums) for the product between the variable D_{it} and the difference between the outcome variable and its counterfactual. All is divided by the number of observations suitable for the ATT. From (5), it is possible to notice that not all the observations carry the same weight in computing the treatment effect. Some will not show up there ($D_{i,t} = 0$), while others will but with a smaller weight (used only in counterfactuals).

The following characteristics are of the DiD estimator. Not all observations matter for computing the counterfactual of an observation, only the one with the opposite treatment status, this makes the counterfactual unbiased. The number of observations whose counterfactual is available is always inferior to the full sample, this leads to a bigger variance in the estimator. It could happen that an observation cannot be found in computing any counterfactual; in that case, it will not carry any weight to compute the treatment effect.

1.4 Treatment effect

In the last section, the treatment effect on the treated has been estimated when units received the treatment moving from a no treatment status to a treatment status, and this is always the case if the treatment is staggered. But, by being in a context where units can enter and leave the treatment, it is possible to compute the treatment effect

when units are leaving the treatment. In this way, it is possible to separate the **entering effect** and the **leaving effect**. What has been explained in the DiD section was the entering effect. Now it will be shown how to compute the leaving effect; the procedure will be symmetrical to what was explained before.

$$ATT_{leave} = E[Y_{i,t}(1) - Y_{i,t}(0)|X_{i,t} = 0, X_{i,t-1} = 1] \tag{7}$$

What we want to compute is the difference between a unit that goes from treatment to no treatment. The counterfactual for this is a treated unit that remained in the treatment. The same notation as before will be used. The observations object of the ATT_{leave} will be the ones that are now untreated but were treated in the period before.
$M_{it}^{DiD_leave} = \{(i', t') : i' = i, t' = t - 1, X_{i't'} = 1\}$the set containing the observation with the same unit, but that in a time period before and only if it was a treated observation.
$N_{it}^{DiD\_leave} = \{(i', t') : i' \neq i, t' = t, X_{i',t'} = 1, X_{i',t'-1} = 1\}$ the set containing observations with the same time, but different units, an observation belongs to this set if and only if it was treated in time period t and t-1
$A_{it}^{DiD\_leave} = \{(i', t') : i' \neq i, t' = t - 1, X_{i't'} = 1, X_{i't'+1} = 1\}$ the set containing all the observation that are untreated in t-1 and t and have time t-1 and unit different from i.

$$\widehat{Y_{it}(1)} = Y_{it-1} + \frac{1}{|N_{it}^{DiD\_leave}|} \sum_{(i',t)\epsilon N_{it}^{DiD\_leave}} Y_{i't} - \frac{1}{|A_{it}^{DiD\_leave}|} \sum_{(i',t')\epsilon A_{it}^{DiD}} Y_{i't'} \tag{8}$$

The procedure for computing the counterfactual in (8) is the same as it was in (5) but with the new sets. The variable $D_{i,t}$ tells us if an observation is suited to compute the ATT ($\widehat{\theta_{leave}}$), by taking into account if the observation is untreated and if there exist other observations such that the two sets $N_{it}^{DiD\_leave}$ and $M_{it}^{DiD\_leave}$ are not empty.

$$t = 1, D_{it} = 0,$$

$$t \neq 1, D_{it} = (1 - X_{it}) * f(|N_{it}^{DiD}||M_{it}^{DiD}|)$$

using the same f as before.

As a consequence, the ATT_{leave} formula will be:

$$\widehat{ATT_{leave}} = \frac{1}{\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}} \sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}\left(\widehat{Y_{it}(1)} - Y_{it}\right)$$

Notice that the ATT_{leave} computes the difference between being treated and untreated. A positive number implies that going from treated to untreated is a loss of value.

By using a TWFEr it is impossible to separate the entering and leaving effects. They are both present in the treatment effect estimate; this is due to the ordinary least square assumption of linearity and leads to assuming symmetric effects for entering and leaving the treatment. It could be the case that the magnitude and significance of the treatment are due to just one of the two. An interesting view on this topic has been given in a replication study in Imai et al. ((2021)) where they used this very same method compared to TWFEr replicating the work of Acemoglu et al. ((2019)) and more general example will be given later.

A note on the side should be made: the leaving effect can be computed if and only if there are observations leaving the treatment. This cannot happen in a non-staggered design.

1.5 Simulations

The main difference between the two estimators is the observations used to estimate the treatment effect and how this exposes the two methods to specific weaknesses.

All the simulations will be conducted in the following way. A dataset is generated following a specific model shown below. First, DiD and TWFEr models are estimated on the standard model present in the Overview (1), and all of it is repeated 30 times using $\theta = 5$. Then we look at all the DiD and TWFEr estimates and compare them. Time and unit coefficient are set to be equal to t and i, for example: $Y_{2,3} = 2 + 3 + 5 * X_{2,3} + e_{i,t}$ All the parameters to generate the data will be shown. The number of units and time periods are always set to 20. The probability of treatment is set at 40%. The error is normally distributed (mean is zero, and variance amounts to one).

The first simulation study regards the existence of a difference in entering the treatment (θ_{enter}) and leaving the treatment (θ_{leave}). Our model (1) does not specify a difference between the two and is missing these parameters. We expect TWFEr to generate an estimate between the two θs, while DiD to generate a precise estimate of θ_{enter}. We will

also use DiD to compute an unbiased ATT_{leave} to show the opportunities of this powerful estimator.

The model (1) is assumed when computing the estimates, but the true one is presented below.

$$te_{i,t} = |\{X_{i',t'}|i' = i, t'\epsilon[2,t]X_{i',t'} = 1, X_{i',t'-1} = 0\}|\text{times the treatment was entered}$$
$$tl_{i,t} = |\{X_{i',t'}|i' = i, t'\epsilon[2,t]X_{i',t'} = 0, X_{i',t'-1} = 1\}|\text{times the treatment was left}$$

$$t \neq 1, Y_{i,t} = u_i + v_t + \theta_{enter}te_{i,t} - \theta_{leave}tl_{i,t} + e_{i,t}$$

$$t = 1, Y_{i,t} = u_i + v_t + \theta_{enter}X_{i,t} + e_{i,t}$$

Unit, time fixed effects and treatment are the same as in the first model introduce (1). However, every time a unit i enters the treatment, a factor θ_{enter} is added, and every time it leaves the treatment, a factor θ_{leave} is subtracted.

How is **Two-Way Fixed Effects** (TWFE) behaving with this model? Let us simplify and ignore time and unit FE for a moment[9]. For example, assume T=4 and focus on a single unit and $X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0$. Then, the treatment effect would be:

$$\widehat{\theta} = \bar{Y(1)} - \bar{Y(0)} = \frac{Y_1 + Y_3}{2} - \frac{Y_2 + Y_4}{2} =$$
$$= \frac{(\theta_{enter}) + (2\theta_{enter} - \theta_{leave})}{2} - \frac{(\theta_{enter} - \theta_{leave}) + 2(\theta_{enter} - \theta_{leave})}{2} = \theta_{leave}$$

If units never left treatment, then there would be no θ_{leave} in the equation and $\theta = \theta_{enter}$. The estimate is dependent on how many untreated units are before being treated for the first time and how many are after[10]. If they are all after, θ_{leave} will dominate; in the opposite case θ_{enter} will. In the first case, the difference between treated and untreated amounts up to θ_{enter}, while in the second one is only θ_{leave}. The estimate of the linear regression will always be between the two. In the simulations, it will be the average (random designs are repeated)[11]. In the DiD estimator:

[9] In this model misspecification, we are not stressing the ability to adjust for time and unit effects, but the linear regression assumptions of linearity and its consequence of assuming symmetry when receiving and leaving treatment if treatment is specified as dummy variable

[10] Imagine 2 observations of the same unit, $X_1 = 0$ and $X_2 = 1$, the treatment would be $Y_1 - Y_0 = \theta_{enter}$, instead if it was to be the opposite case, first a treated observation and then untreated $Y_0 - Y_1 = \theta_{enter} - (\theta_{enter} - \theta_{leave}) = \theta_l eave$

[11] In a TWFE environment, our treatment effect estimate would also be influenced by how many

θ_{leave} will be perfectly computed by ATT_{leave}

θ_{enter} will be perfectly computed by ATT_{enter}

Assuming a bigger θ_{leave} than θ_{enter} will imply that all observations after they left the treatment will be smaller than if they never received it. This will make our treatment effect estimates bigger. On the other hand, a smaller θ_{leave} will make all of our untreated observations after leaving treatment bigger than if they never received it.

Figure 3: Leaving Treatment

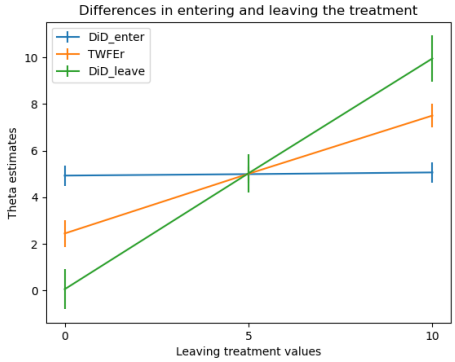
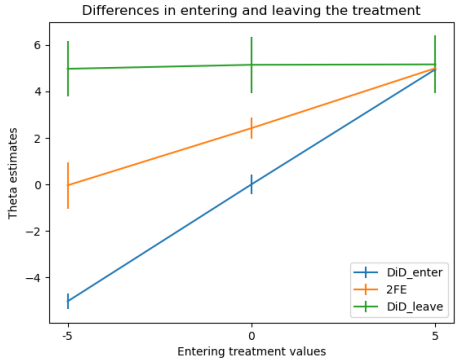


Figure 4: Entering Treatment



The simulations are run with the values used on the x axis, in this case three for each figure.

These two simulations showed how varying both of them will always change values in the TWFEr estimate.

It is interesting in the first picture that if our θ_{leave} is smaller (bigger) than θ_{enter} effect, the TWFEr will compute a smaller (bigger) treatment estimate even if adopting the policy is better (worse) compared to a case with a bigger (smaller) θ_{leave} that will make our TWFEr estimate bigger (smaller). It is better (worse) because if tomorrow the policy cannot continue anymore, leaving it will result in smaller (bigger) damage compared to the case with symmetric treatment effects (OLS assumption), implying that overall the dependent untreated variable will be higher (lower) by $\theta_{enter} - \theta_{leave}$. Therefore, it is really important to detach the view of the treatment effect (in non-staggered contexts) as only entering effect or as treatment will make our variable bigger by θ.

In the second figure, it should be noticed how TWFEr has significant estimates even if adopting the treatment is not significant at all ($\theta_{enter} = 0$). If the average of the two treatment effects is zero, TWFEr is estimating a non-significant treatment effect while

observations have never been in treated in units other than the one the counterfactual is estimated

both of them are significant ($\theta_{enter} - \theta_{leave} = -10$), meaning that deciding to adopt the policy will make us worse by θ_{enter} and leaving it would do it by $-\theta_{leave}$, it would be better not to adopt the policy, but TWFEr suggests that is not going to be significantly different from zero and same applies for positive values. If the act of changing policies leads to a positive or negative effect, this cannot be always identified by TWFEr. A second simulation study will be left in the Appendix.

Bibliography

D. Acemoglu, S. Naidu, P. Restrepo, and J. A. Robinson. Democracy does cause growth. *Journal of political economy*, 127(1):47–100, 2019.

K. Imai and I. S. Kim. On the use of two-way fixed effects regression models for causal inference with panel data. *Political Analysis*, 29(3):405–415, 2021.

K. Imai, I. S. Kim, and E. H. Wang. Matching methods for causal inference with time-series cross-sectional data. *American Journal of Political Science*, 2021.

2 Appendix

2.1 K-factor TWFEr

$$K = \frac{1}{NT}\sum_{i'\neq i}^{N}\sum_{t'\neq t}^{T} X_{i,t}\left(\frac{\sum_{t'\neq t}(1-X_{it'})}{T-1} + \frac{\sum_{i'\neq i}(1-X_{i',t})}{N-1} - \frac{\sum_{i'\neq i}\sum_{t'\neq t}(1-X_{i',t'})}{(T-1)(N-1)}\right) \\ +(1-X_{i,t})\left(\frac{\sum_{t'\neq t}(X_{it'})}{T-1} + \frac{\sum_{i'\neq i}(X_{i',t})}{N-1} - \frac{\sum_{i'\neq i}\sum_{t'\neq t}(X_{i',t'})}{(T-1)(N-1)}\right) \tag{9}$$

2.2 Carry-over effect

The carry-over effect is an effect that is manifested if a treated unit in time t stays in treatment in t+1, however this effect will be lost if the unit leaves the treatment.

$$Y_{i,t} = u_i + v_t + \theta X_{i,t} + carry\_over Z_{i,t} + e_{i,t}$$

The model is the same as (1) with the addition of the carry over effect explained below.

$$t = 1 \rightarrow Z_{i,t} = 0$$

$$t \neq 1 \rightarrow Z_{i,t} = |\{X_{i,t}|X_{i,t} = 1, X_{i,t-1} = 1\}|$$

The variable $Z_{i,t}$ is equal to one if a treated observations was treated in the period before, in any other case the variable will be zero. When this variable is equal to one a "carry over" effect is manifested, a premium for staying in treatment, this effect disappear once the treatment is left. While carry over effect itself has not been object of your analysis this is a recurrent case of model misspecification and it is useful to show how the different estimators reacts, an example is given.

$$X_{i,t} = 1, Z_{i,t} = 1, X_{i,t+1} = 0, Z_{i,t+1} = 0 \tag{10}$$

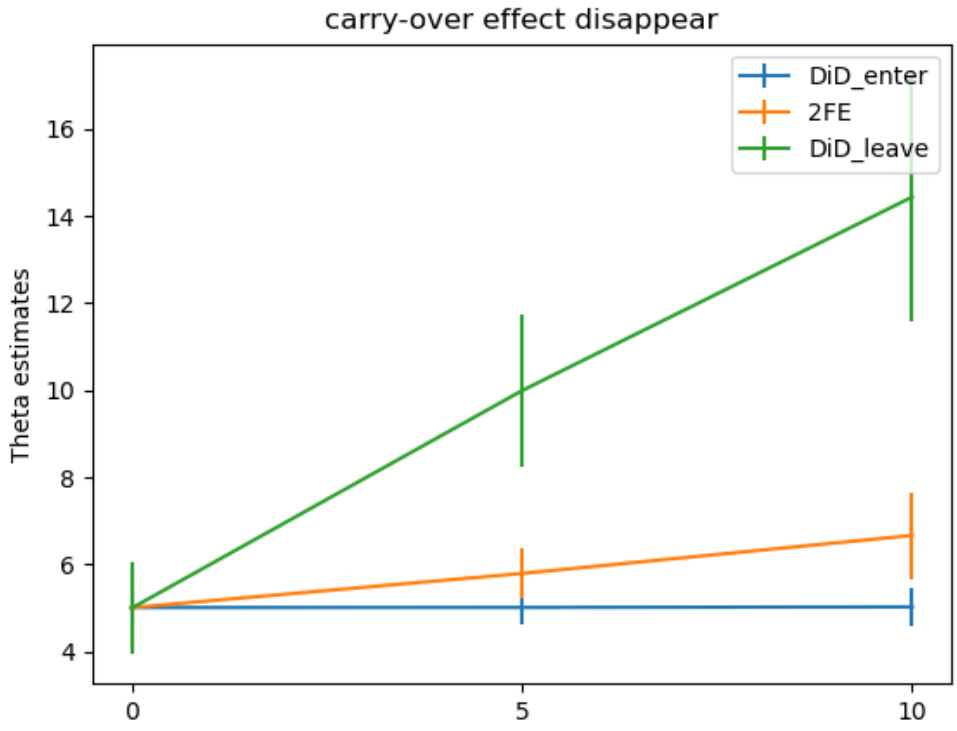
$$Y_{i,t} - Y_{i,t+1} = v_t - v_{t+1} + carry\_over + \theta \tag{11}$$

The estimates of the TWFE will be affected since the treated units are on average larger (if the carry over effect is positive, if not it would be smaller) than what it is in the estimator model (1), however the magnitude of the effect depends on how many units are recipients of the carry over effect. It is clear that this effect will not affect the ATT_{enter}

since the counterfactual is not changing compared to the standard setting. However, it will affect the counterfactual of units leaving the treatment (they are compared with units that stayed there) ATT_{leave} will be biased. With the increase of the carry-over effect also the variance of ATT_{leave} increases.By looking at what an observation and its counterfactual are it will be clearer:

$$\widehat{\theta_{leave}} = (Y_{c,t+1} - Y_{c,t}) - (Y_{i,t+1} - Y_{i,t})$$

The first difference belongs to $0, carry\_over$ while the second to $-\theta, -(\theta + carry\_over)$, the ATT_{leave} can range between $[\min\{carry\_over, \theta\}, 2carry\_over + \theta]$, this ranges only become wider with an increase in $carry\_over$,hence more variant.



It is possible to notice that with a $carry\_over$ effect of 5, the TWFE is already significantly different than the treatment effect value of 5.