Jorge Eduardo Pérez Pérez

Advances in Differences in Differences and Bartik instruments: Class Notes

Banco de México

October 8, 2024

Chapter 2
Staggered adoption, heterogeneity, and issues with the TWFE specification

In this chapter we introduce the staggered adoption setup and discuss issues with the TWFE specification for estimation of treatment effects under staggered adoption. We follow the setup in Roth et al. (2023).

2.1 Staggered adoption setup

The staggered adoption setup is motivated by units selectively being treated over time. For example, we can think of units as cities and treatment being the entry of a ride-sharing platform such as Uber to those cities. Uber does not enter all cities at the same time, and once Uber enters a city it does not leave (or rarely does).

Now we assume there are T periods indexed by $t = 1, \ldots, T$. The variable $D_{i,t}$ denoes treatment status for unit i at time t. For staggered adoption, we assume the following:

Assumption 2.1 *(Staggered adoption)*

1. *Treatment is binary:* $D_{i,t} \in \{0,1\}$.
2. *All units begin untreated:* $D_{i,1} = 0$ *for all i.*
3. *Treatment is absorbing:* $D_{i,t'} \geq D_{i,t}$ *for all* i *and* $t' \geq t$.

The absorbing treatment assumption is convenient because it lets us group the treated units by treatment cohorts. We define the treatment cohort G_i as $G_i = min\{t : D_{i,t} = 1\}$, the first period when an unit receives treatment. For control units (that never receive treatment), $G_i = \infty$.

With multiple time periods, we also need to extend the potential outcomes notation to allow for treatment histories. A treatment history is a vector of length T of zeros and ones. Moreover, in the staggered adoption setting, all the treatment history vectors will be of the form $(\mathbf{0}_s, \mathbf{1}_{T-s})$.

The potential outcome for unit i if they were treated for the first time at time g (that is, if their cohort $G_i = g$) is $Y_{i,t}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+})$. If it was never treated, its potential outcome is $Y_{i,t}(\mathbf{0}_T)$. We can simplify these in the staggered adoption setting as $Y_{i,t}(g) = Y_{i,t}(\mathbf{0}_{g-1}, \mathbf{1}_{T-g+})$ and $Y_{i,t}(\infty) = Y_{i,t}(\mathbf{0}_T)$.

We can also extend our notation for treatment effects. The **individual treatment effect** for unit i at time t if they belong to cohort g is

$$\tau_{i,t}(g) := Y_{i,t}(g) - Y_{i,t}(\infty). \tag{2.1}$$

Notice subtle differences with the treatment effect in the simple case. This is defined in terms of potential outcomes.

With the individual treatment effects as building blocks, you can build other quantities of interest. We will focus on the **average treatment effect on the treated for cohort g at time t ($ATT_{g,t}$)** is the average of these individual treatment effects for a particular treatment cohort:

$$\tau_{g,t} := \mathbb{E}[\tau_{i,t}(g)|G_i = g]. \tag{2.2}$$

You could also define an **average treatment effect on the treated at time t**:

$$\tau_t := \mathbb{E}[\mathbb{E}[\tau_{i,t}(g)|G_i = g]].$$

To identify $ATT_{g,t}$, we need to generalize the parallel trends assumption. In the basic case, we assumed that in absence of treatment, the outcomes of the treated and control groups would evolve in parallel. A natural extension is to require that the untreated potential outcomes of all cohorts evolve in parallel:

Assumption 2.2 *(Strong unconditional parallel trends) For all $t \neq t'$ and $g \neq g'$:*

$$\mathbb{E}[Y_{i,t}(\infty) - Y_{i,t'}(\infty)|G_i = g] = \mathbb{E}[Y_{i,t}(\infty) - Y_{i,t'}(\infty)|G_i = g']$$

(This assumption is stronger than needed: we only need parallel trends on average between control and treatment groups). We also extend the no anticipation assumption as

Assumption 2.3 *(Staggered no anticipation)*

$$Y_{i,t}(g) = Y_{i,t}(\infty) \text{ for all } i \text{ and } t < g$$

2.2 Two-way fixed effects estimation

To introduce two-way fixed effects estimation, assume that treatment effects are constant across time and across units:

$$\tau_{g,t} = \tau \text{ for all } t \geq g$$

In this scenario, a natural extension of (**??**) is a linear model with unit and time effects:

$$Y_{i,t} = \alpha_i + \theta_t + D_{i,t}\beta + \varepsilon_{i,t} \tag{2.3}$$

Under constant treatment effects, $\hat{\beta}$ from (2.3) is a consistent estimator for τ, and it can be estimated through fixed-effects estimation of (2.3) (For review of fixed effects estimation, see Wooldridge (2010).

2.3 Treatment effect heterogeneity and TWFE weights

If we are willing to assume that treatment effects are constant, there is nothing wrong with TWFE estimation. There may be settings where this is a plausible assumption: for example, all units are treated at the same time and economic theory suggests that there is not treatment heterogeneity. Or, units are treated at different times but treatments have a one-off, homogeneous impact.

However, in most applications, we may expect **heterogeneity in treatment effects**. Economic theory will often imply heterogeneous treatment effects across units. For example, the elasticity of labor supply will be different across individuals with different outside options.

We would expect that the estimate from (2.3) corresponds to a weighted average of $ATT_{g,t}$ with some reasonable weights, i.e., $\hat{\beta} = \sum_{t,g} \omega_{t,g}\tau_{t,g}$ with $\omega_{t,g} > 0$ and $\sum_g \omega_{t,g} = 1$. However, this turns out not to be the case. To show some mathematical intuition of this, note that by the FWL theorem, the OLS estimate of β from (2.3) equals the coefficient of a regression of Y_{it} on the residuals of a regression of $D_{i,t}$ on unit

and time effects. That is, run the regression $D_{i,t} = \tilde{\alpha}_i + \tilde{\delta}_t + u_{i,t}$ and obtain the residuals $D_{i,t} - \hat{D}_{i,t}$. Then the OLS estimate of β from (2.3) equals:

$$\hat{\beta} = \frac{\hat{\text{Cov}}\left(Y_{i,t}, D_{i,t} - \hat{D}_{i,t}\right)}{\hat{\text{Var}}(D_{i,t} - \hat{D}_{i,t})} = \frac{\sum_{i,t} Y_{i,t}(D_{i,t} - \hat{D}_{i,t})}{\sum_{i,t}(D_{i,t} - \hat{D}_{i,t})^2} \tag{2.4}$$

If we break down the numerator into observations where $D_{i,t} = 1$ and $D_{i,t} = 0$ we can write:

$$\hat{\beta} = \frac{\sum_{i,t,D_{i,t}=0} Y_{i,t}(-\hat{D}_{i,t}) + \sum_{i,t,D_{i,t}=1} Y_{i,t}(1 - \hat{D}_{i,t})}{\sum_{i,t}(D_{i,t} - \hat{D}_{i,t})^2}$$

Moreover, when $D_{i,t} = 1$, $\tau_{i,t}(g) = Y_{i,t}(g) - Y_{i,t}(\infty) = Y_{i,t} - Y_{i,t}(\infty)$. Replacing this value of $Y_{i,t}$ in the numerator for $D_{i,t} = 1$:

$$\hat{\beta} = \frac{\sum_{i,t,D_{i,t}=0} Y_{i,t}(-\hat{D}_{i,t}) + \sum_{i,t,D_{i,t}=1} (Y_{i,t}(\infty) + \tau_{i,t}(g))(1 - \hat{D}_{i,t})}{\sum_{i,t}(D_{i,t} - \hat{D}_{i,t})^2} \tag{2.5}$$

Here, we can see the that the OLS estimate of β equals a weighted average of treated and control observations. For treated observations, the weights are proportional to $1 - \hat{D}_{i,t}$. However, since $\hat{D}_{i,t}$ is a prediction from a linear model, nothing guarantees that $1 - \hat{D}_{i,t}$ will be positive! So some of the treatment effects $\tau_{i,t}(g)$ will get negative weights when building $\hat{\beta}$, which is counterintuitive. Moreover, the weights in (2.5) need not be proportional to the sample sizes of each cohort g.

The negative weights will tend to arise for early-treated units, that is, units with low value of g, late in the sample. The predicted value $\hat{D}_{i,t}$ equals $\bar{D}_i + \bar{D}_t - \bar{\bar{D}}$, where the bars denote sample means. Early treated units will have high values of $\bar{D}_i$ (because they are treated in almost every t) so $\bar{D}_i \approx 1$. If they are late in the sample then most units will have been treated, so $\bar{D}_t \approx 1$. Then $\hat{D}_{i,t} \approx 2 - \bar{\bar{D}}$. The average $\bar{\bar{D}}$ will be less than one if there are non-treated units. So $\hat{D}_{i,t}$ will be strictly higher than 1, and $1 - \hat{D}_{i,t}$ will be negative in those cases.

2.4 Goodman-Bacon Decomposition

Goodman-Bacon (2021) proposes an intuitive decomposition of the TWFE estimator that illustrates why it may give counterintuitive weights to some units and how it involves "forbidden" comparisons that may lead the TWFE estimate to be a biased estimate of the average treatment effect.

Suppose there are three treatment cohorts: an "early-treated" cohorts with $G_i = k$, a "late-treated" treatment cohort with $G_i = \ell > k$, and an untreated cohort with $G_i = \infty$. We will also denote this cohort by U for untreated. We call $PRE(k)$ as the time window before the k cohort is treated, $MID(k,\ell)$ as the time window when the k cohort has been treated but the ℓ cohort has not, and a $POST(\ell)$ window when both cohorts have been treated. We denote by $\bar{Y}_k^{PRE(k)}$ as the sample average of the outcome for units in the k cohort during the $PRE(k)$ time window, and define other sample averages $\bar{Y}_g^W, W \in \{PRE(k), MID(k,\ell), POST(\ell)\}, g \in \{k,\ell,\infty\}$ accordingly.

Let the fractions of *units* belonging to each cohort be n_k, n_ℓ, and n_U, respectively. Assume that we observe a balanced panel.

[GB Fig 1. o Cunningham II 3 65]

With these three groups we can define four 2x2 difference in difference estimators:

[Cunningham II - 66-69]

A. Early treated vs. untreated:

$$\hat{\beta}_{kU}^{2\times2} = \left(\bar{Y}_k^{POST(k)} - \bar{Y}_k^{PRE(k)}\right) - \left(\bar{Y}_U^{POST(k)} - \bar{Y}_U^{PRE(k)}\right) \tag{2.6}$$

This comparison uses a fraction $n_k + n_U$ of units, and since the panel is balanced, it also uses a fraction $n_k + n_U$ of the NT observations in the panel.

B. Late treated vs. untreated:

$$\hat{\beta}^{2\times2}_{\ell U} = \left(\bar{Y}^{POST(\ell)}_{\ell} - \bar{Y}^{PRE(\ell)}_{\ell}\right) - \left(\bar{Y}^{POST(\ell)}_{U} - \bar{Y}^{PRE(\ell)}_{U}\right) \tag{2.7}$$

This comparison uses a fraction $n_\ell + n_U$ of units, and since the panel is balanced, it also uses a fraction $n_\ell + n_U$ of the NT observations in the panel.

C. Early treated vs. late treated:

$$\hat{\beta}^{2\times2}_{k\ell} = \left(\bar{Y}^{MID(k,\ell)}_{k} - \bar{Y}^{PRE(k)}_{k}\right) - \left(\bar{Y}^{MID(k,\ell)}_{\ell} - \bar{Y}^{PRE(\ell)}_{\ell}\right) \tag{2.8}$$

This comparison uses a fraction $n_k + n_\ell$ of units. It does not use all the time periods, though: it only uses the periods in the $PRE(\ell)$ window. Letting $\bar{D}_k$ and $\bar{D}_\ell$ denote the fraction of periods in which each cohort is treated, this comparison uses a fraction $(n_k + n_\ell)(1 - \bar{D}_\ell)$ of the NT observations.

D. Late treated vs. early treated:

$$\hat{\beta}^{2\times2}_{\ell k} = \left(\bar{Y}^{POST(\ell)}_{\ell} - \bar{Y}^{MID(k,\ell)}_{\ell}\right) - \left(\bar{Y}^{POST(\ell)}_{k} - \bar{Y}^{MID(k,\ell)}_{k}\right) \tag{2.9}$$

This comparison uses a fraction $n_k + n_\ell$ of units. It does not use all the time periods, though: it only uses the periods in the $POST(k)$ window. Letting $\bar{D}_k$ and $\bar{D}_\ell$ denote the fraction of periods in which each cohort is treated, this comparison uses a fraction $(n_k + n_\ell)(\bar{D}_k)$ of the NT observations.

Goodman Bacon shows that the TWFE estimate is a weighted average of these four difference-in-difference estimates. To understand the weights, recall that the coefficient estimate of linear regression on a binary variable and covariates (e.g. on a treatment indicator) will put more weight on covariate cells where there is most variation in the treatment. (MHE 3.3) Here, the weights will be proportional to the variance of treatment in each one of the comparisons, after adjusting for unit and time effects.

The variance of treatment for each one of the comparisons is:

$$\begin{aligned}
\hat{V}^{D}_{jU} &= n_{jU}(1 - n_{jU})\bar{D}_j(1 - \bar{D}_j), j = k, \ell \\
\hat{V}^{D}_{k\ell} &= n_{k\ell}(1 - n_{k\ell})\frac{\bar{D}_k - \bar{D}_\ell}{1 - \bar{D}_\ell}\frac{1 - \bar{D}_k}{1 - \bar{D}_\ell} \\
\hat{V}^{D}_{\ell k} &= \frac{\bar{D}_\ell}{\bar{D}_k}\frac{\bar{D}_k - \bar{D}_\ell}{\bar{D}_k}
\end{aligned}$$

where $n_{ab} \equiv \frac{n_a}{n_a + n_b}$.

With these variances, we can write the decomposition of the TWFE estimate:

$$\hat{\beta} = \sum_{k \neq U} s_{kU}\hat{\beta}^{2\times2}_{kU} + \sum_{k \neq U}\sum_{\ell > k}\left[s_{kl}\hat{\beta}^{2\times2}_{k\ell} + s_{\ell k}\hat{\beta}^{2\times2}_{\ell k}\right] \tag{2.10}$$

with weights proportional to the variance of treatment and the sample size in each comparison:

$$s_{kU} = \frac{(n_k + n_U)^2 \hat{V}^D_{kU}}{\hat{V}^D},$$
$$s_{k\ell} = \frac{((n_k + n_\ell)(1-\bar{D}_\ell))^2 \hat{V}^D_{k\ell}}{\hat{V}^D},$$
$$s_{\ell k} = \frac{((n_k + n_\ell)(\bar{D}_k))^2 \hat{V}^D_{\ell k}}{\hat{V}^D}.$$

The weights tend to be higher for comparisons where there's more variance treatment, that is, groups where treatment occurs in the middle of the panel. This may be undesirable if we want to estimate ATT, because it will downweight some groups.

This decomposition tells us about what TWFE estimates, but it does not tell us whether it is unbiased for ATT or not. Let's assume dynamic treatment effects as earlier and $ATT_k(W)$ denote the average treatment effect in a treatment window for group k, e.g. $ATT_k(W) = \frac{1}{T_W}\sum_{t \in W} \mathbb{E}[Y_{it}(k) - Y_{it}(0)]$.

In this case each of the 2 by 2 estimates converges in probability to a different quantity:

$$\beta^{2\times 2}_{kU} = ATT_k(POST(k)) + \Delta Y^0_k(POST(k), PRE(k)) - \Delta Y^0_U(Post(k), Pre)$$
$$\beta^{2\times 2}_{k\ell} = ATT_k(MID) + \Delta Y^0_k(MID, PRE(k)) - \Delta Y^0_l(MID, PRE(k))$$
$$\beta^{2\times 2}_{\ell k} = ATT_\ell(POST(\ell)) + \Delta Y^0_\ell(POST(\ell), MID) - \Delta Y^0_k(POST(\ell), MID) - (ATT_k(POST(\ell)) - ATT_k(MID))$$

The two first comparisons are "unproblematic". Under parallel trends, these comparisons converge to ATTs for the given window. The third comparison, though, the later treated vs. early treated, is problematic under dynamic treatment effects. The difference-in-differences comparisons using the early treated units as controls is a "forbidden comparison", because under dynamic treatment effects, the early treated may still have treatment effects happening when they are used as a control group for the later-treated groups. This is the same point risen by Borusyak and Jaravel (2018). In extreme settings, the contamination in this comparison may even flip the sign of the TWFE estimator!

Bacon's decomposition has become a standard diagnostic tool to assess whether this contamination may be a potential issues. The decomposition consists of calculating the weights in (2.10) and see if there is a large weight on the late treated vs. early treated comparisons. Large weights on these comparisons are suggestive of potential issues in a scenario with dynamic treatment effects.

2.5 Callaway and Sant'Anna's / Sun and Abraham's estimator for difference in differences with multiple treatment periods

Callaway and Sant'Anna (2021) and Sun and Abraham (2021) take a different approach to estimating $ATT_{g,t}$ and ATT. Since TWFE estimators may suffer from bias to estimate the ATT when there is staggered treatment adoption, and this bias may appear because TWFE makes forbidden comparisons, why not estimate all the $ATT_{g,t}$'s separately and then aggregate them however we like? We can choose to use only the comparisons we like for aggregation. CS argue that by making this switch, we can focus on identification and easy to interpret estimates at the cost of harder implementation, instead of focusing on easy-to-implement but hard to interpret estimates such as those coming from TWFE.

CS's assumptions are the same as those in the beginning of the section, although they do allow for weaker versions of the no anticipation and parallel trends assumptions. They also consider "conditional" versions of their assumptions instead of unconditional versions, but we will work with unconditional versions for now.

Assumption 2.4 *(Limited anticipation) There is a known* $\delta \geq 0$ *such that*

$$\mathbb{E}[Y_{i,t}(g)|G_i = g] = \mathbb{E}[Y_{i,t}(\infty), G_i = g] \text{ for all } g \text{ such that } t < g - \delta$$

This allows for anticipation effects of the treatment up to δ periods before the treatment. For parallel trends, we can choose either of two assumptions:

Assumption 2.5 *(Parallel trends based on an untreated group) For all it, for each g such that* $t \geq g - \delta$

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

Assumption 2.6 *(Parallel trends based on a not-yet-treated group) For all it, for each g such that* $t \geq g - \delta$

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

Under the parallel trends based on an untreated group assumption, we can show that the difference in differences between each cohort and the untreated group equals $\tau_{g,t}$, using a similar argument as in the basic diff-in-diff design:

$$\tau_{g,t} = \mathbb{E}[Y_{i,t} - Y_{i,g-1}|G_i = g] - \mathbb{E}[Y_{i,t} - Y_{i,g-1}|G_i = \infty]$$

and we can estimate it replacing these expectations by their sample counterparts. We can then aggregate these building blocks to get other parameters of interest. For example, to estimate τ_t we only need to average across cohorts: $\tau_t = \sum_g (N_g/N)\tau_{t,g}$ where N_g is the number of *units* in cohort g.

2.6 Event Study

In staggered designs sometimes we are not interested in the effects $\tau_{g,t}$ for a particular cohort in a particular period. Instead, we may be interested in the effect of the policy a few years after it was adopted. We could obtain these estimates by aggregating $\tau_{g,t}$ appropriately. For example, assume there are two cohorts g and g', and we want to know the ATT two years after treatment. We could then average $\tau_{g,g+1}$ and $\tau_{g',g'+1}$ to obtain this estimate.

To obtain all the relative estimates directly, we could estimate this regression:

$$Y_{i,t} = \alpha_i + \phi_t + \sum_{r \neq 0} \mathbf{1}[R_{i,t} = r]\beta_r + \varepsilon_{i,t} \tag{2.11}$$

with $R_{i,t} = t - G_i + 1$ being the relative time to treatment. The β_r coefficients would then measure ATTs in relative time. Equation (2.12) is convenient because it allows for estimation of all effects at once. It also allows for estimation of ATTs in negative relative time. These provide for a simple falsification test of the parallel trends assumption: in negative relative time and under no anticipation, the potential outcomes are equal to the observed outcomes for control and treatment units, so the ATTs should be zero. Note, however, that while this is necessary for the parallel trends assumption to hold, it is not sufficient: the parallel trends assumption also requires potential untreated outcomes to evolve in parallel in positive relative times.

This specification, known as the TWFE event study specification, also suffers from possible negative weighting issues. Sun and Abraham (2021) propose to fix this problem by estimating it one at a time comparing treated cohorts to untreated or not yet treated cohorts, and then averaging the events per relative time and per cohort. This can be implemented with a regression as:

$$Y_{i,t} = \alpha_i + \phi_t + \sum_{r \neq 0} \mathbf{1}[R_{i,t} = r]\mathbf{1}[G_i = g]\beta_{r,g} + \varepsilon_{i,t} \tag{2.12}$$

References

Borusyak, Kirill and Xavier Jaravel (2018), *Revisiting event study designs*. SSRN.

Callaway, Brantly and Pedro HC Sant'Anna (2021), "Difference-in-differences with multiple time periods." *Journal of econometrics*, 225, 200–230.

Goodman-Bacon, Andrew (2021), "Difference-in-differences with variation in treatment timing." *Journal of econometrics*, 225, 254–277.

Roth, Jonathan, Pedro HC Sant'Anna, Alyssa Bilinski, and John Poe (2023), "What's trending in difference-in-differences? a synthesis of the recent econometrics literature." *Journal of Econometrics*.

Sun, Liyang and Sarah Abraham (2021), "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of econometrics*, 225, 175–199.

Wooldridge, Jeffrey M (2010), *Econometric analysis of cross section and panel data*. MIT press.