A primer on staggered DD

John Gardner

Abstract

This is a quick introduction to DD and event studies with staggered adoption. I provide intuition for and simplified derivations of the key results in the literature, and discuss a simple approach to estimation when adoption is staggered and average treatment effects are heterogeneous across treatment cohorts and durations.

1 A review of basic DD

Suppose that there are two times periods (pre and post) and two groups (treatment and control). In this 2×2 case, differences in differences (DD) identifies the average effect of the treatment on the treated (ATT) under the parallel trends assumption, which holds that average outcomes would have changed by the same amount between the pre and post periods for both groups if it weren't for the treatment.

Mathematically, let i denote individuals, t denote time, and Y_{it}, Y_{0it} and Y_{1it} denote observed and potential untreated and treated outcomes. Then parallel trends holds that

$$E(Y_{it}|Post_{it}, Treat_{it}) = \gamma Post_{it} + \lambda Treat_{it} + \beta Post_{it} \times Treat_{it},$$

where $\beta = E(Y_{1it} - Y_{0it}|Post_{it} = 1, Treat_{it} = 1)$ is the average effect of the treatment on the treated. The ATT can be estimated by running the regression

$$Y_{it} = \gamma Post_{it} + \lambda Treat_{it} + \beta Post_{it} \times Treat_{it} + \varepsilon_{it}.$$

Note that even though this is a parametric regression model, the coefficient on $Post_{it} \times Treat_{it}$ identifies the average of potentially heterogenous causal effects. The reason why this works is that even though treatment effects are heterogeneous, our regression is a correctly specified model for the *conditional expectation* of Y_{it}, and the difference β between the treatment and control groups in post-pre average outcome differences represents the ATT.

There are a few variations on this basic approach. The above regression is equivalent to one that replaces the post and treat dummies with individual and time fixed effects:

$$Y_{it} = \lambda_i + \gamma_t + \beta D_{it} + \varepsilon_{it}, \tag{1}$$

where $D_{it} = Post_{it} \times Treat_{it}$ is an indicator for whether unit i is treated *at time* t.[1] It is also common to include time-varying covariates in the regression, to allow for the possibility that trends are parallel conditional on covariates.[2]

Another variation replaces the overall treatment indicator D_{it} with treatment duration indicators D_{pit}, equal to one if unit i has been treated for p periods as of time t:

$$Y_{it} = \lambda_i + \gamma_t + \sum_{p=1}^{P} \beta_r D_{pit} + \varepsilon_{it}. \tag{2}$$

These *event-study* regressions are useful for examining how the effect of the treatment depends on its duration, since β_p represents the average effect on the treated of being treated for exactly p periods. Event-study regressions can also be used to test the plausibility of parallel trends by including leads of treatment adoption (that is, replacing $p = 1$ with $p = -L$ in the sum, which is the same as including indicators for whether the treatment is adopted in L, $L - 1$, $\ldots$ periods). Nonzero coefficients on the adoption leads imply that

[1] To see that these regressions are equivalent, note that by the Frisch-Waugh-Lovell (FWL) theorem, the coefficient on D_{it} is the same as the coefficient on the residual from a regression of D_{it} on the other variables. Since D_{it} is the same for all members of the treatment group, it doesn't matter if we include individual and year effects or indicators for post and treat.

[2] This assumes that the treatment effect does not vary by covariates and that the covariates are not affected by the treatment. For more general approaches, see Wooldridge (2021) and Abadie (2005).

parallel trends does not hold (i.e., that untreated outcomes evolve differently for the treatment and control groups).[3]

2 DD with staggered adoption

A natural extension of 2×2 DD is to settings where there are multiple treatment cohorts, indexed by $c = 1, \ldots, C$, which adopt the treatment at different times. This is known as *staggered adoption.* In this case, under parallel trends, we can use a cohort-specific event-study regression to express outcomes as

$$Y_{it} = \lambda_i + \gamma_g + \sum_{c=1}^{C} \sum_{p=1}^{P_c} \beta_{cp} D_{cpit} + \varepsilon_{it}, \tag{3}$$

where P_c is the longest observed duration for members of cohort c, D_{cpit} is an indicator for whether an observation belongs to cohort c and has been treated for p periods, β_{cp} is the average effect of being treated for p periods for members of cohort c, and ε_{it} is mean zero conditional on unit effects, time effects, and all of the cohort-duration indicators D_{cpit}.

If the average effect of the the treatment doesn't depend on the duration of the treatment or the cohort, we can replace $\sum_{c=1}^{C} \sum_{p=1}^{P_c} \beta_{cp} D_{cpit}$ in the above with βD_{it}, and we are back to the usual DD regression specification (1). This has been the traditional approach to differences in differences with staggered adoption. However, one might be concerned that average treatment effects are heterogeneous across treatment cohorts and durations, in addition to individuals. Several papers have shown that the standard regression DD specification is invalid in this case (Borusyak and Jaravel, 2017; Goodman-Bacon, 2018; de Chaisemartin and D'Haltfouille, 2020).

A simple way to see the problem is to think of the term $\sum_{c=1}^{C} \sum_{p=1}^{P_c} \beta_{cp} D_{cpit}$ from the correct specification (3) as an omitted variable in the usual DD regression, (1). In this case, the population *regression* coefficients from the

[3] It isn't possible to include all duration-specific treatment effects and adoption leads in the model, since the sum of these will equal one for all members of the treatment group and zero for all members of the control group, and hence will be collinear with the individual fixed effects. Common practice is to exclude D_{0it} (an indicator for adopting the treatment in the next period).

misspecified model (1) identify the sum of the *true* coefficients from (3), plus the population linear projection of the omitted variable onto the included regressors (treatment status and unit and time indicators). Since the omitted variable is a sum, this is also the sum of the projections of the $\beta_{cp}D_{cpit}$ onto the included regressors. Thus, since the true coefficient on D_{it} (which doesn't belong in the model) is zero, the population regression coefficient on D_{it} identifies

$$\sum_{c=1}^{C}\sum_{p=1}^{P_c}\omega_{cp}\beta_{cp},$$

where ω_{cp} is the coefficient from a population regression of D_{cpit} on treatment status D_{it} and unit and time effects.

Hence, the usual regression DD specification identifies a regression-weighted average of the cohort×period-specific effects. At first blush, this might not sound so bad. We can even show that the weights must sum to one.[4] However, it turns out that the weights can also be negative, which makes the regression DD coefficient difficult to interpret.

To see how this might happen, recall that ω_{cp} is the coefficient from a population regression of D_{cpit} on D_{it} and unit and time fixed effects. By the FWL theorem, this is the same as a regression of D_{cpit} on the residual from a regression of D_{it} on unit and time effects. Using the two-way-within (aka double-demeaned) transformation, this regression can be expressed as[5]

$$D_{cpit} = \omega_{cp}\{[D_{it} - P(D_{it}=1|i)] - [P(D_{it}=1|t) - P(D_{it}=1)]\} + e_{it}.$$

For earlier cohorts, $P(D_{it}|i)$ will tend to be large, while in later periods (when more units are treated), $P(D_{it}=1|t) - P(D_{it}=1)$ will also be large. Consequently, regression DD will put less (and possibly even negative) weight

[4] Here is one proof: Since $\sum_{c,p} D_{cpit} = D_{it}$, the coefficient on D_{it} from a population regression of $\sum_{c,p} D_{cpit}$ on D_{it} and unit and time effects must be one. But by the FWL theorem, this coefficient also equals $Cov(\sum_{c,p} D_{cpit}, \tilde{D}_{it})/Var(\tilde{D}_{it}) = \sum_{c,p} Cov(D_{cpit}, \tilde{D}_{it})/Var(\tilde{D}_{it}) = \sum_{c,p}\omega_{cp}$, where $\tilde{D}_{it}$ is the residual from a regression of $\tilde{D}_{it}$ on unit and time fixed effects.

[5] Suppose that $D_{it} = \lambda_{di} + \gamma_{dt} + \nu_{dit}$, with $E(\nu_{dit}) = 0$. The within-unit variation in D_{it} is $D_{it} - E(D_{it}|i) = \gamma_{dt} + \nu_{dit} - [E(\gamma_{dt}) + E(\nu_{dit}|i)]$, and the within-time variation in this variation is $D_{it} - E(D_{it}|i) - [E(D_{it}|t) - E(D_{it})] = \nu_{dit} - E(\nu_{dit}|i) - [E(\nu_{dit}|t) - E(\nu_{dit})] = \nu_{dit}$. Hence, this transformation removes both unit and time fixed effects.

on earlier cohorts and later periods. Goodman-Bacon (2018) and de Chaisemartin and D'Haltfouille (2020) provide even more detailed decompositions of DD regressions, although the regression-weighting interpretation conveys the basic idea and remains valid when covariates are included.[6]

There are (at least) two ways to provide intuition for this result. One is that the usual DD regression specification compares newly treated units to units who have already been treated, which produces a misleading picture of the effect of the treatment when that effect is changing across cohorts and over the duration of the treatment. Another is that since DD regression assumes a constant ATT, it attributes some of the heterogeneous effects of the treatment to unit and time effects (with more of the treatment effect being absorbed by units that have been treated for more periods and at times when more units are treated).

3 Event studies with staggered adoption

Sun and Abraham (2020) show that this same phenomenon also applies to event-study regressions. This can be seen using the same linear projection trick as above. If the correct model is (3), but specification (2) is used instead, the coefficient on D_{pit} identifies the sum of the projections of the omitted $\beta_{cp}D_{cpit}$ onto the included D_{pit} and time and unit effects, so the coefficient on D_{pit} identifies

$$\sum_{c=1}^{C}\sum_{q=1}^{P_c}\omega_{cq|p}\beta_{cp},$$

where $\omega_{cq|p}$ is the coefficient on D_{pit} from a regression of D_{cqit} on all of the D_{qit}, $q \in \{1, \ldots, P\}$, as well as unit and time fixed effects.

Thus, when duration-specific treatment effects vary across cohorts, the coefficients on the adoption leads and duration indicators identify weighted averages of all of the cohort×period-specific average treatment effects. Here,

[6] Goodman-Bacon's decomposition shows, amazingly, that the regression DD identifies a weighted average of all possible 2×2 DD regressions.

the weights have the properties that[7]

$$\sum_{c=1}^{C} \omega_{cp|p} = 1 \quad \text{and} \quad \sum_{c=1}^{C} \omega_{cq|p\neq q} = 0.$$

As Sun and Abraham note, an important consequence of this is that the coefficients on the treatment-adoption leads may be nonzero even if trends are, in fact, parallel.[8]

4 What to do instead

A simple way to do DD with staggered adoption and heterogeneous ATTs is to estimate specification (3), which allows for separate treatment effects for each cohort and treatment duration. This produces estimates of several cohort×duration average treatment effects. An obvious way to summarize these effects is to report

$$\sum_{c=1}^{C} \sum_{p=1}^{P_c} \beta_{cp} P(D_{cpit} = 1 | D_{it} = 1).$$

This is the *overall* average effect of the treatment on the treated (i.e., averaged across all observed cohorts and durations). Other weighted averages of

[7] To derive these properties, note that $\sum_{c=1}^{C} D_{cpit} = 1 \cdot D_{pit} + 0 \cdot \sum_{q \neq p} D_{qit}$. Thus, the coefficient on D_{pit} from a population regression of $\sum_{c=1}^{C} D_{cpit}$ on D_{pit}, all of the $D_{q\neq p,it}$, and unit and time fixed effects must be one. By the FWL theorem, this coefficient is the same as the slope coefficient from a regression of $\sum_{c=1}^{C} D_{cpit}$ on the residual $\tilde{D}_{pit}$ from a regression of D_{pit} on the $D_{q\neq p,it}$ and unit and time effects. But this slope coefficient also equals $Cov(\sum_{c=1}^{C} D_{cpit}, \tilde{D}_{pit})/Var(\tilde{D}_{pit}) = \sum_{c=1}^{C} Cov(D_{cpit}, \tilde{D}_{pit})/Var(\tilde{D}_{pit}) = \sum_{c=1}^{C} \omega_{cp|p} = 1$. The other property is proved similarly.

[8] In the staggered adoption context, event studies are sometimes conducted without a control group that never receives the treatment, in which case not-yet-treated units serve as controls for newly treated units. As Borusyak and Jaravel (2017) note, the relationship between cohort, duration and time introduces a second source of collinearity in this case. If D_{cit} is an indicator for belonging to cohort c (i.e., the sum of all unit indicators for members of that cohort) and $D_{\tau it}$ is an indicator for time τ, then $\sum_{p=-R}^{P} pD_{pit} = \sum_{\tau=1}^{T} \tau D_{\tau it} - \sum_{c=1}^{T} (c-1) D_{cit}$. Since one of the coefficients in the sum on the left is zero, the time and unit indicators are a linear combination of $P + R$ of the $P + R + 1$ possible duration indicators, making it necessary to omit another adoption lead.

the cohort×duration-specific effects can be reported as well.[9]

Modern statistical software makes this easy. In Stata, if `d` is a dummy for (time-varying) treatment status, `dur` is a categorical variable for the (time-varying) treatment duration, and `cohort` is a categorical variable for the treatment cohort, we can estimate (3) and the overall ATT (including delta-method-based standard errors) using the syntax

```
xtreg y i.year ibn.dur#ibn.cohort#c.d, fe vce(cluster id)
margins, dydx(d) subpop(if d==1)
```

Similar commands can be used to produce event-study regressions that average over cohorts (see the attached Stata files, which is modeled on similar files by Jeffrey Wooldridge, for an example).

There are several alternative approaches. The above is similar to the approach suggested by Sun and Abraham (2020). Callaway and Sant'anna (2020) develop a related approach that can handle covariates more flexibly, which Wooldridge (2001) shows can be approximated using regression. There is also the stacked DD approach (see, e..g, Cengiz et al., 2019), which attempts to transform staggered DD to 2×2 DD by aligning observations for different cohorts by relative time instead of calendar time.

References

Abadie. 2005. Semiparametric difference-in-differences estimators. RESTUD.

Borusyak and Jaravel. 2017. "Revisiting event study designs, with an application to the estimation of the marginal propensity to consume." Working paper.

Callaway and Sant'anna. 2020. "Difference-in-differences with multiple time periods and an application on the minimum wage and employment." JOE.

[9] For example, since completed treatment durations will vary across cohorts, we might want to report the overall average effect of being treated for no more than a certain number of periods).

Cengiz, Dube, Lindner and Zipperer. 2019. “The effect of minimum wages on low-wage jobs.” QJE.

de Chaisemartin and D’Haltfouille. 2020. “Two-way fixed effects estimators with heterogeneous treatment effects.” AER.

Goodman-Bacon. 2018. “Difference-in-differences with variation in treatment timing.” Working paper.

Sun and Abraham. 2020. “Estimating dynamic treatment effects in event studies with heterogeneous treatment effects.” JOE.

Wooldridge. 2021. “Two-way fixed effects, the two-way Mundlak regression, and event study estimators.” Working paper.