Lecture Notes: Causal Inference Fall 2020

Claire Duquennois

7/28/2020

Contents

1 Difference-in-Differences

Sometimes randomizing treatment by a program or policy is impossible because the intervention has already happened. In other situations it may be that randomization my be unfeasible. There are also many important questions where randomizing treatment would be unethical: the idea of randomizing exposure to pollutants for instance, is clearly problematic.

When we cannot implement a randomized control trial, and haven't been able to assign the treatment (or program, or policy, etc.) randomly ourselves, then we have to assume that the treatment was not randomly assigned: that it depends on either observable or unobservable characteristics of the people, firms, cities... under consideration. In this case their are important differences between our treated and untreated units that we cannot control for in a regression. Leaving these variables out in u_i will cause OVB, which totally foils our attempts at estimating the causal effect. Differences-in-differences is a way of getting around non-random assignment of a treatment.

1.1 Example and intuition

[1]

In 2002, Craigslist opened a new section on its front page called "erotic services" in San Francisco California. The section would end up being used by sex workers exclusively to advertise to and solicit clients. Sex workers claimed it made them safer, because instead of working on street corners and for pimps, they could solicit indoors from their computers, which as a bonus, also gave them the chance to learn more about the men contacting them. But activists and law enforcement worried that it was facilitating sex trafficking and increasing violence against women. Which was it? Was erotic services (ERS) making women safer, or was it placing them in harm's way?

This is ultimately an empirical question. We want to know the effect of ERS on female safety, but the fundamental problem of causal inference says that we can't know what effect it had because we are missing the data necessary to make the calculation. That is,

[1] This section is borrowed from scunning.com

$$E[\tau] = E[\underbrace{Y_{SF,2003}(D_{SF,2003}=1)}_{\text{observed}}] - E[\underbrace{Y_{SF,2003}(D_{SF,2003}=0)}_{\text{unobserved}}]$$

where $Y_{SF,2003}(D_{SF,2003}=1)$ is women murdered in a world where San Francisco has ERS in 2003 and $Y_{SF,2003}(D_{SF,2003}=0)$ is women murdered in a world where San Francisco does not have ERS in 2002, *at the exact same moment in time*. In 2002 only the first occurs, and the second is a counterfactual. So how do we proceed?

The standard way to evaluate interventions such as this is the standard differences-in-differences strategy (DiD). DiD is basically a version of panel fixed effects, but can also be used with repeated cross-sections. Let's look at this example using some tables, which hopefully will help give you an idea of the intuition behind DiD, as well as some of it's identifying assumptions. Let's say the intervention, D, is the availability of the Craigslist erotic services webpage, and we want to know the causal effect, τ of D on Y, female murders. Couldn't we just compare San Francisco murders in, say, 2003 with some other city, like Pittsburgh?

City	Outcome
San Francisco	$Y_{SF,2003} = \alpha_{SF} + \tau$
Pittsburgh	$Y_{P,2003} = \alpha_P$

where α_{SF} is an unobserved San Francisco fixed effect and $\alpha_P$is a Pittsburgh fixed effect. When we make a simple comparison between Pittsburgh and San Francisco, $Y_{SF,2003} - Y_{P,2003} = \alpha_{SF} + \tau - \alpha_P$ we get a causal effect that equals $\tilde{\tau} = \tau + \alpha_{SF} - \alpha_P$. Thus the simple difference is biased because of $\alpha_{SF} - \alpha_P$, the difference in the underlying murder rates between the two cities in a world where neither gets treated. So if our goal is to get an unbiased estimate of τ, then the simple difference won't work.

But what if we compare San Francisco to itself? Say compare it in 2003 to two years earlier in 2001? Let's look at that simple before and after difference.

City	Time	Outcome
San Francisco	Before	$Y_{SF,2001} = \alpha_{SF}$
	After	$Y_{SF,2003} = \alpha_{SF} + \lambda_{03} + \tau$

Again, this doesn't lead to an unbiased estimate of τ, even if it does eliminate the city fixed effect. That's because such differences can't control for, or net out, natural changes in the murder rate over time. I can't compare San Francisco before and after because to the time effects, λ_{03} since $Y_{SF,2003} - Y_{SF,2001} = \alpha_{SF} + \lambda_{03} + \tau - \alpha_{SF} = \lambda_{03} + \tau$. However if I could identify and control for these time effects then I would be fine.

This is the intuition of the DiD strategy. I combine both these two simpler approaches, so that I can eliminate both the city effects and the time effects. Let's look at the following table.

City	Time	Outcome	1st Difference	2nd Difference
San Francisco	Before	$Y_{SF,2001} = \alpha_{SF}$		
	After	$Y_{SF,2003} = \alpha_{SF} + \lambda_{03} + \tau$	$\lambda_{03} + \tau$	
				τ
Pittsburgh	Before	$Y_{P,2001} = \alpha_P$		
	After	$Y_{P,2003} = \alpha_P + \lambda_{03}$	λ_{03}	

The first difference does the simple before and after difference. This ultimately eliminates the unit specific fixed effects. Then, once those differences are made, we difference the differences (hence the name) to get the

unbiased estimate of τ.

1.2 The mechanics

The basic idea is, that sometimes treatment and control group outcomes move in parallel in the absence of treatment. When they do, the divergence of a post-treatment path from the trend established by a comparison group may signal a treatment effect.

Difference-in-differences can be implemented as follows:

1) Compute the difference in the mean outcome variable Y in the post treatment period ($t = 1$) and the before treatment period ($t = 0$) for the control group C:

$$\bar{Y}_{C,1} - \bar{Y}_{C,0} = \Delta\bar{Y}_C$$

which allows us to cancel out the control group fixed effect and identify the time fixed effect since

$$\bar{Y}_{C,1} - \bar{Y}_{C,0} = \alpha_C + \lambda_1 - \alpha_c = \lambda_1 = \Delta\bar{Y}_C$$

2) Compute the difference in the mean outcome variable Y in the post treatment period ($t = 1$) and the before treatment period ($t = 0$) for the treated group T:

$$\bar{Y}_{T,1} - \bar{Y}_{T,0} = \Delta\bar{Y}_T$$

which allows us to cancel out the treated group fixed effect

$$\bar{Y}_{T,1} - \bar{Y}_{T,0} = \alpha_T + \lambda_1 + \tau - \alpha_T = \lambda_1 + \tau = \Delta\bar{Y}_T$$

3) Treatment impact is then measured by the difference-in-differences:

$$(\bar{Y}_{T,1} - \bar{Y}_{T,0}) - (\bar{Y}_{C,1} - \bar{Y}_{C,0}) = (\Delta\bar{Y}_T - \Delta\bar{Y}_C)$$

since by comparing the differences we can cancel out the time fixed effect and isolate the treatment effect of interest:

$$\Delta\bar{Y}_T - \Delta\bar{Y}_C = \lambda_1 + \tau - \lambda_1 = \tau$$

In a regression framework, we estimate the following,

$$Y_{it} = \beta_0 + \beta_1 Post_t + \beta_2 GetsTreat_i + \beta_3 Post_t \times GetsTreat_i + u_{it}$$

where $Post_t$ is an indicator for the post treatment period, and $GetsTreat_i$ is an indicator for observations in the treatment group that eventually gets treated.

The regression framework has the benefit of providing us with standard errors so we can test for the significance of our estimators. We call β_3 the difference-in-differences estimator as it illustrates the differential impact of being in the post treatment period if you are in the treated group since, following the difference in mean procedure above, we have

$$\begin{aligned} E[Y_{C,1}] - E[Y_{C,0}] &= (\beta_0 + \beta_1) - (\beta_0) = \beta_1 \\ E[Y_{T,1}] - E[Y_{T,0}] &= (\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_2) = \beta_1 + \beta_3 \\ (E[Y_{T,1}] - E[Y_{T,0}]) - (E[Y_{C,1}] - E[Y_{C,0}]) &= \beta_3 \end{aligned}$$

1.2.1 A simulation:

Suppose you are a principal of a school with ten 4th grade classrooms of 30 students each. At the beginning of the 2001 school year, teachers were able to enroll their class in the scholastic book club, a new program that was not offered in prior years. 4 of your fourth grade teachers opted to enroll. You are interested in estimating the effect of participation in the book club on 4th grade 2001 reading scores.

```
set.seed(6000)
scores<-as.data.frame(rep(c(1,2,3,4,5,6,7,8,9,10),times=30))
names(scores)<-c("class")
scores <- fastDummies::dummy_cols(scores, select_columns = "class")

scores$error<-rnorm(300, mean=0, sd=10)

#suppose teachers in the better performing classes (classes, 7,8,9,10) select to participate in the boo
scores$treat<-0
scores$treat[scores$class%in%c(7,8,9,10)]<-1

tau<-10

#the data generating process
scores$read4<-(85+tau*scores$treat
               +(-10)*scores$class_1+(-15)*scores$class_2+(-5)*scores$class_3
              +(-8)*scores$class_4+(-7)*scores$class_5+(-13)*scores$class_6
              +(11)*scores$class_7+(8)*scores$class_8+(10)*scores$class_9
              +(12)*scores$class_10
              +scores$error)

scores$year<-"2001"

scores01<-scores

rm(scores)

scores<-as.data.frame(rep(c(1,2,3,4,5,6,7,8,9,10),times=30))
names(scores)<-c("class")
scores <- fastDummies::dummy_cols(scores, select_columns = "class")

scores$error<-rnorm(300, mean=0, sd=10)

scores$treat<-0
scores$treat[scores$class%in%c(7,8,9,10)]<-1

#the data generating process
scores$read4<-(78
              +(-10)*scores$class_1+(-15)*scores$class_2+(-5)*scores$class_3
              +(-8)*scores$class_4+(-7)*scores$class_5+(-13)*scores$class_6
              +(11)*scores$class_7+(8)*scores$class_8+(10)*scores$class_9
              +(12)*scores$class_10
              +scores$error)

scores$year<-"2000"

scores00<-scores
rm(scores)
```

```
scores<-rbind(scores01, scores00)

regnodid<-felm(read4~treat,scores[scores$year=="2001",])

scores$post<-0
scores$post[scores$year=="2001"]<-1
regdid<-felm(read4~post+treat+post*treat,scores)


regdidfe<-felm(read4~post+treat+post*treat|class,scores)
```

```
## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or indefinite
```

```
stargazer(regnodid, regdid,regdidfe, type="latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Oct 25, 2020 - 12:05:48 PM

Table 4:

	Dependent variable:		
	read4		
	(1)	(2)	(3)
post		7.612***	7.612***
		(1.073)	(1.023)
treat	27.482***	18.387***	
	(1.141)	(1.200)	
post:treat		9.095***	9.095***
		(1.697)	(1.617)
Constant	76.775***	69.163***	
	(0.721)	(0.759)	
Observations	300	600	600
R^2	0.661	0.613	0.653
Adjusted R^2	0.660	0.611	0.646
Residual Std. Error	9.679 (df = 298)	10.182 (df = 596)	9.704 (df = 588)
Note:			*p<0.1; **p<0.05; ***p<0.01

1.3 Identifying Assumption:

The key assumption for the validity of this method is the *the difference between before and after in the comparison group is a good counterfactual for the treatment group.* In other words, the trend in outcomes of the comparison group is what we would have observed in the treatment group absent the treatment (which could be exposure to an event, policy, intervention. . .).

The idea behind this assumption is easily illustrated graphically.

In green we see how the outcome variable changes in the control group between $t = 0$ and $t = 1$, a change that is estimated by β_1.

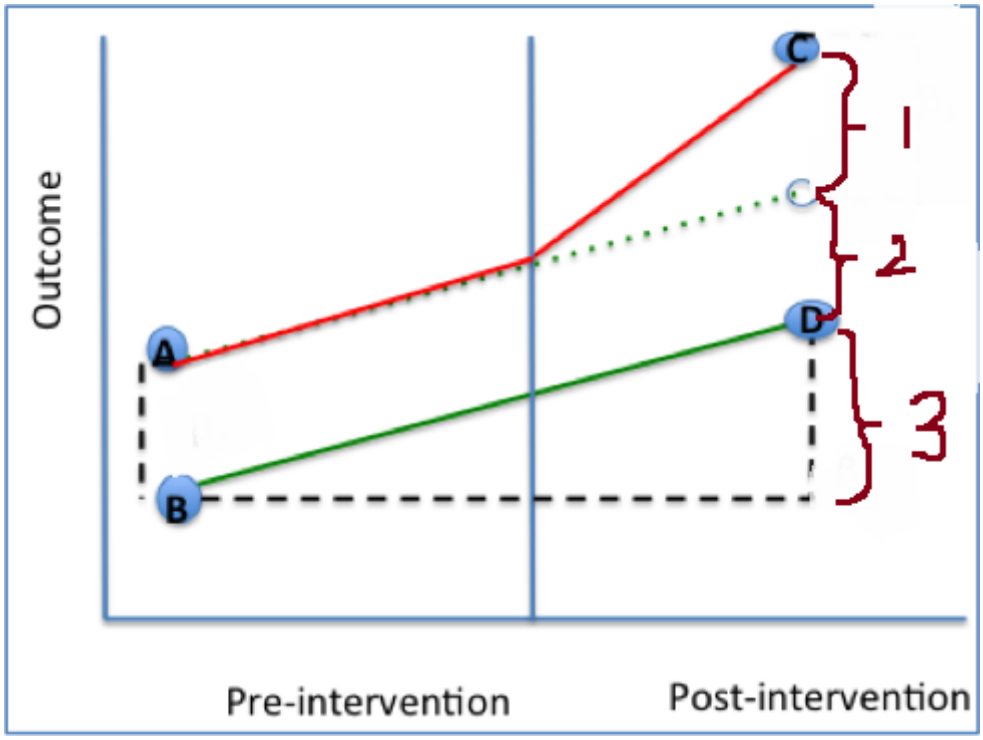


Figure 1: ...

Notice that at $t = 0$, there is a difference in the outcome variable between the control group and the observations that will eventually get treated. This difference is estimated by β_2.

The key assumption in a DID model is that absent treatment, the outcome of the treated group would have followed a trend that was parallel to that of the control group, ie the dashed green line in the figure. Thus we are treating the dashed green line as the counterfactual for the treated group so that any deviation from this counterfactual is attributed to the treatment effect, β_3. Though straightforward, this assumption is fundamentally untestable, because we will never actually observe this counterfactual of what would have happened to the treated group had they not been treated. Maybe it would have followed the same trend as the control. But then maybe it wouldn't have. Ultimately, we have no way of knowing for sure.

1.4 Robustness

1.4.1 Parallel Trends

The assumption that the treated group would have followed a trend that was parallel to that of the control group is actually a fairly big assumption in many circumstances. Indeed, policymakers for instance will often select into treatment or control based on differences in the anticipated effects of treatment, or pre-existing differences in outcomes, which basically implies that the parallel trends assumption does not hold. A classic example of this is the "Ashenfelter dip", named after Orley Ashenfelter, a labor economist at Princeton. Individuals who are "Treated" by job training programs are often individuals who just experienced a "dip" in earnings due to a job loss. When they get rehired their earnings increase substantially. If I compare their change in earnings to the change experienced by people who did not sign up for job training, I will see a large differential increase in earnings associated with program participation. It would be a mistake however to attribute this to the causal effect of the training as it is primarily due to who selects into training. Similarly, when policy makers select individuals or locations to apply a new policy, they will likely select places or people who have the most to gain from the policy. This is a form of selection bias that violates parallel trends.

How can we check to see if the parallel trends assumption is likely to hold? Empiricists faced with this

untestable assumption typically choose to use deduction as a way to check this assumptions validity. We reason that if the pre-treatment trends were parallel between the two groups, then wouldn't it stand to reason that the post-treatment trends would have been too? Note that it is important to understand that just because the pre-treatment trend are parallel does not *prove* that the post-treatment trends would have evolved in the same way. But given that we see that the pre-treatment trends evolved similarly, it does give some confidence that the post treatment trends would have too (absent some unobserved group specific time shock).

Including leads into the DiD model is an easy way to check the pre-treatment trends. Lags can also be included to analyze whether the treatment effect changes over time after the treatment assignment too. This is done by estimating some form of the following specification

$$Y_{it} = \beta_0 + \beta_2 GetsTreat_i + \sum_{t=-n}^{m} \beta_{3t} Period_t \times GetsTreat_i + \lambda_t + u_{it}$$

where treatment occurs right after period $t = 0$. Thus if parallel trends holds, $E[\hat{\beta}_{3,t<1}] = 0$ since there is no treatment in these time periods. Whereas the estimated coefficients $\hat{\beta}_{3,t>0}$ could be different from 0 depending on the treatment effect and it's lagged impacts. These results are typically best presented graphically in a graph that

a)either plots the mean outcome for both the treatment and control for several periods before and after treatment,

b)and/or a graph that plots the $\hat{\beta}_{3t}$ estimates from the specification above.

```
set.seed(1999)
scoresbase<-as.data.frame(rep(c(1,2,3,4,5,6,7,8,9,10),times=30))
names(scoresbase)<-c("class")
scoresbase <- fastDummies::dummy_cols(scoresbase, select_columns = "class")

#suppose teachers in the better performing classes (classes, 7,8,9,10) select to participate in the boo
scoresbase$treat<-0
scoresbase$treat[scoresbase$class%in%c(7,8,9,10)]<-1

yr<-c(1995,1996,1997,1998,1999,2000,2001,2002,2003,2004,2005)
tauyr<-c(0,0,0,0,0,0,10,10,10,10,10)
yrfe<-c(72,77,75,79,81,79,83,77,82,84,81)

for(i in 1:11){
  name<-paste("scores", yr[i], sep="_")
  scores<-scoresbase
  scores$error<-rnorm(300, mean=0, sd=10)
  tau<-tauyr[i]
  yearfe<-yrfe[i]
  #the data generating process
  scores$read4<-(yearfe+tau*scores$treat+scores$error
                 +(-10)*scores$class_1+(-15)*scores$class_2+(-5)*scores$class_3
                +(-8)*scores$class_4+(-7)*scores$class_5+(-13)*scores$class_6
                +(11)*scores$class_7+(8)*scores$class_8+(10)*scores$class_9
                +(12)*scores$class_10)

  scores$year<-yr[i]
  assign(name, scores)
  rm(scores)
}
```

```
allscores<-rbind(scores_1995,scores_1996,scores_1997,scores_1998,scores_1999,
                 scores_2000,scores_2001,scores_2002,scores_2003,scores_2004,scores_2005)

allscores$post<-0
allscores$post[allscores$year%in%c(2001,2002,2003,2004,2005)]<-1

allscores <- fastDummies::dummy_cols(allscores, select_columns = "year")

regdidall2<-felm(read4~treat
                 +year_1995*treat+year_1996*treat+year_1997*treat+year_1998*treat
                 +year_1999*treat+year_2001*treat+year_2002*treat+year_2003*treat
                 +year_2004*treat+year_2005*treat,
                 allscores)

stargazer( regdidall2, type="latex",no.space=TRUE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Oct 25, 2020 - 12:05:48 PM

```
#start with plot of group means

#calculateing the mean score for each year by treatment status
grp_mean<-allscores%>%
    group_by(year,treat)%>%
    dplyr::summarize(groupmean = mean(read4, na.rm=TRUE))
```

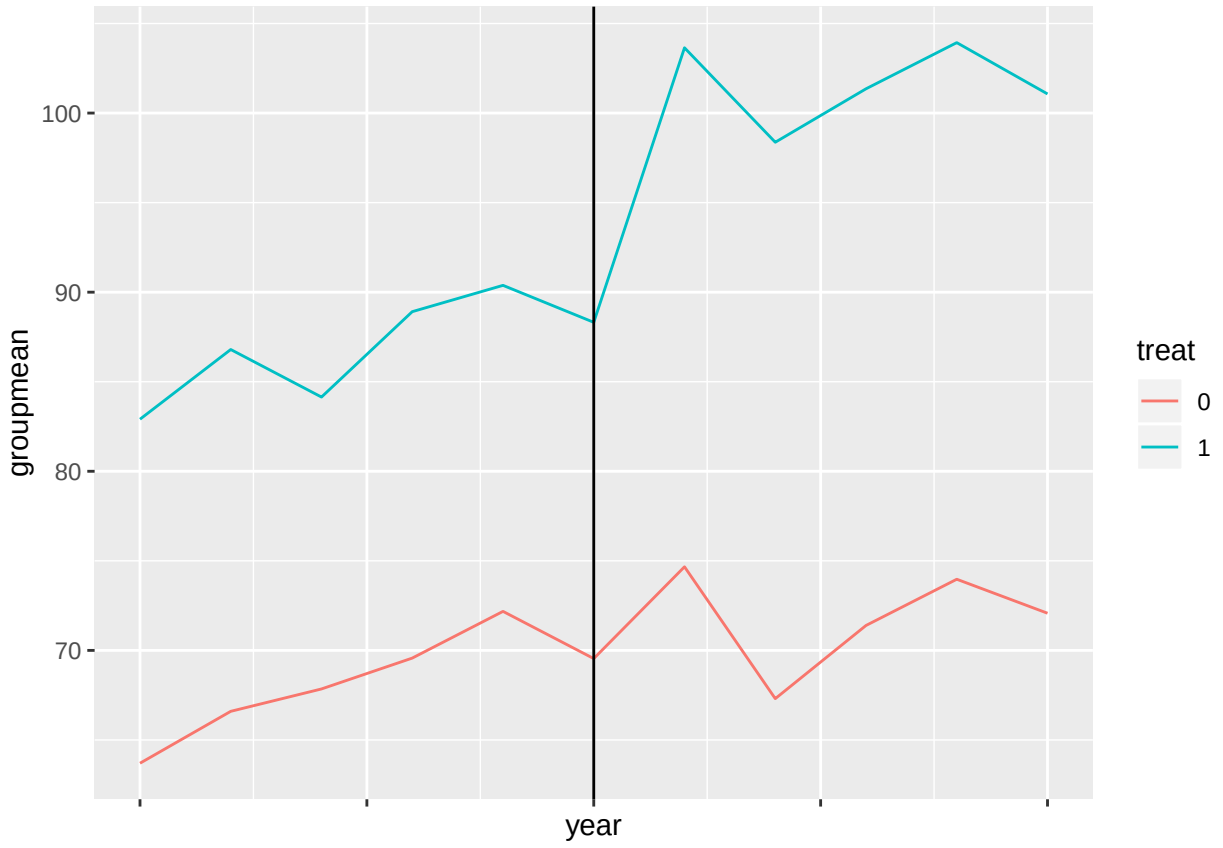
```
## `summarise()` regrouping output by 'year' (override with `.groups` argument)
```

```
grp_mean$treat<-as.factor(grp_mean$treat)

#difference in means plot
didmeans<-ggplot(grp_mean, aes(year, groupmean, group=treat, color = treat)) +
    stat_summary(geom = 'line') +
    geom_vline(xintercept = 2000) +
    theme( axis.text.x = element_blank())

didmeans
```

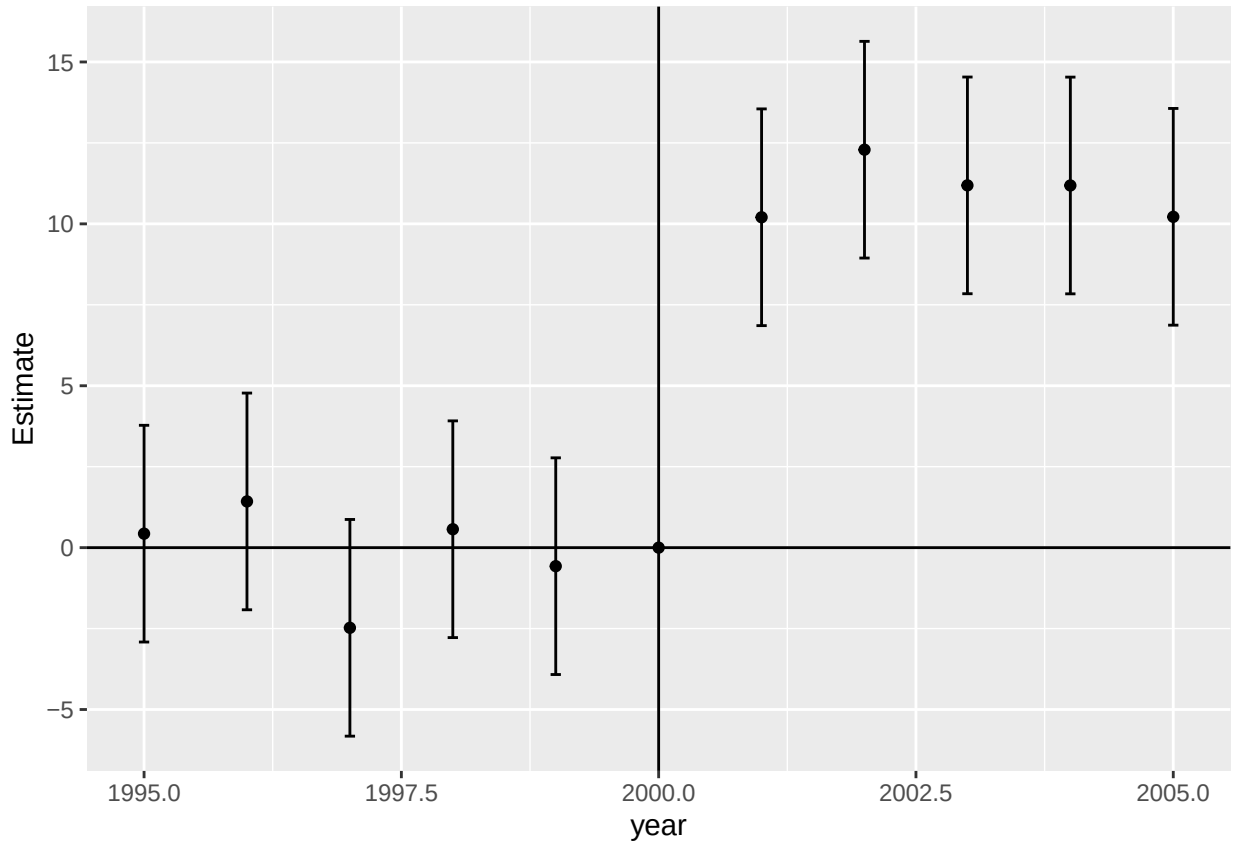
```
## No summary function supplied, defaulting to `mean_se()
```

Table 5:

	Dependent variable:
	read4
treat	18.774***
	(1.207)
year_1995	−5.842***
	(1.080)
year_1996	−2.945***
	(1.080)
year_1997	−1.696
	(1.080)
year_1998	0.027
	(1.080)
year_1999	2.636**
	(1.080)
year_2001	5.125***
	(1.080)
year_2002	−2.235**
	(1.080)
year_2003	1.849*
	(1.080)
year_2004	4.430***
	(1.080)
year_2005	2.533**
	(1.080)
treat:year_1995	0.431
	(1.707)
treat:year_1996	1.427
	(1.707)
treat:year_1997	−2.476
	(1.707)
treat:year_1998	0.568
	(1.707)
treat:year_1999	−0.573
	(1.707)
treat:year_2001	10.204***
	(1.707)
treat:year_2002	12.290***
	(1.707)
treat:year_2003	11.188***
	(1.707)
treat:year_2004	11.186***
	(1.707)
treat:year_2005	10.217***
	(1.707)
Constant	69.542***
	(0.764)
Observations	3,300
R^2	0.613
Adjusted R^2	0.610
Residual Std. Error	10.244 (df = 3278)
Note:	*p<0.1; **p<0.05; ***p<0.01




```
#plot of differences coefficients

res<-coef(summary(regdidall2))
res<-as.data.frame(res)

res<-res[13:22,]

a<-c(0,0,0,0)

res<-rbind(res,a)

year<-c(1995,1996,1997,1998,1999,2001,2002,2003,2004,2005,2000)
res<-cbind(res,year)
res$ci<-1.96*res$`Std. Error`

names(res)<-c("Estimate","se", "t",  "p", "year", "ci")

# Use 95% confidence interval instead of SEM
didplot2<-ggplot(res, aes(x=year, y=Estimate)) +
    geom_errorbar(aes(ymin=Estimate-ci, ymax=Estimate+ci),width=.1) +
    geom_vline(xintercept = 2000)+
    geom_hline(yintercept = 0)+
    geom_point()

didplot2
```



1.4.2 A note on Inference

In many DID setups, the variable of interest only varies at a "group" level (ex: at the state or class level) meaning that the outcome variables are often serially correlated. This means that conventional standard errors will often severely understate the standard deviation of the estimators and so standard errors are biased downward (i.e. incorrectly small) which can lead to inference errors. The most commonly applied solution to this is to cluster standard errors at the group level (more on clustering at the end of the semester). Clustering standard errors can be done in R using the felm function as illustrated below.

```
regdidall3<-felm(read4~treat
                 +year_1995*treat+year_1996*treat+year_1997*treat+year_1998*treat
                 +year_1999*treat+year_2001*treat+year_2002*treat+year_2003*treat
                 +year_2004*treat+year_2005*treat
                 |0
                 |0
                 |class,
                 allscores)

stargazer( regdidall2,regdidall3, type="latex",no.space=TRUE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Oct 25, 2020 - 12:05:49 PM

1.4.3 Falsification Tests: Alternative outcomes

Suppose you are still concerned that the parallel trends assumption is violated. In addition to the leads and lags check described above, you can also check and see if there is a pattern of parallel trends for an alternative

Table 6:

	Dependent variable:	
	read4	
	(1)	(2)
treat	18.774***	18.774***
	(1.207)	(2.268)
year_1995	−5.842***	−5.842***
	(1.080)	(0.879)
year_1996	−2.945***	−2.945***
	(1.080)	(1.070)
year_1997	−1.696	−1.696
	(1.080)	(1.032)
year_1998	0.027	0.027
	(1.080)	(0.257)
year_1999	2.636**	2.636**
	(1.080)	(1.237)
year_2001	5.125***	5.125***
	(1.080)	(0.941)
year_2002	−2.235**	−2.235
	(1.080)	(1.582)
year_2003	1.849*	1.849***
	(1.080)	(0.654)
year_2004	4.430***	4.430***
	(1.080)	(1.206)
year_2005	2.533**	2.533***
	(1.080)	(0.746)
treat:year_1995	0.431	0.431
	(1.707)	(1.768)
treat:year_1996	1.427	1.427
	(1.707)	(2.108)
treat:year_1997	−2.476	−2.476*
	(1.707)	(1.287)
treat:year_1998	0.568	0.568
	(1.707)	(1.570)
treat:year_1999	−0.573	−0.573
	(1.707)	(1.477)
treat:year_2001	10.204***	10.204***
	(1.707)	(1.602)
treat:year_2002	12.290***	12.290***
	(1.707)	(2.310)
treat:year_2003	11.188***	11.188***
	(1.707)	(1.824)
treat:year_2004	11.186***	11.186***
	(1.707)	(2.216)
treat:year_2005	10.217***	10.217***
	(1.707)	(1.510)
Constant	69.542***	69.542***
	(0.764)	(1.965)
Observations	3,300	3,300
R^2	0.613	0.613
Adjusted R^2	0.610	0.610
Residual Std. Error (df = 3278)	10.244	10.244
Note:	*p<0.1; **p<0.05; ***p<0.01	

outcome measure that should not be affected by treatment. For the ERS example, you could estimate the same model with manslaughter or male murders as the outcome variable. Though these variables would be affected by the underlying violent crime in the city, neither of these variables are predicted to be affected by ERS. We would anticipate a pattern of parallel trends with the control cities and no discontinuous change in San Francisco when ERS comes online.

1.4.4 Falsification Tests: Alternative controls and triple differences

A DiD research design can sometimes be made more compelling by adding another layer of differencing to the estimator, resulting in a triple-diffs estimator. For example, consider our simulated example estimating the effect of joining the scholastic book club on reading scores. Now suppose that a subset of students already had access to the book club through an after school program. In this case, we can use the students in the after school program as an additional "control" group, since their performance should not change when the book club is introduced to the classrooms.

In practice, we would implement this with a triple difference estimator. Let $\bar{Y}_{a,g,t}$ represent the mean reading score of students in group g (control of treatment classes), in year t, that participate ($a = AS$) or are not ($a = NAS$) in the after school program. The triple difference estimator is then

$$[(\bar{Y}_{T,1,AS} - \bar{Y}_{T,1,NAS}) - (\bar{Y}_{T,0,NAS} - \bar{Y}_{T,0,NAS})] - [(\bar{Y}_{C,1,AS} - \bar{Y}_{C,1,NAS}) - (\bar{Y}_{C,0,AS} - \bar{Y}_{C,0,NAS})]$$

In other words, we compare the evolution of the gap between the after school kids and the others in the treated classrooms to the evolution of the gap between after school kids and the others in the control classrooms.

The advantage of this triple diff structure is that it allows us to relax some assumptions. We no longer need to assume that outcomes for treated and control would evolve similarly in expectation- we now only need to assume that, to the extent that outcomes evolve differently in treated and control classes, the difference affects participants and non-participants in the after school program similarly.

We can easily implement this triple diffs estimator within the regression framework. the key is to put in an indicator for every main effect or interaction up to, but not including, the level at which the treatment varies. Thus we include the main effects for time, class and after-school program participation as well as all possible two way interactions between each of these indicators. The regression looks like this:

$$\begin{aligned} Y_{i,a,g,t} =& \alpha_0 + \alpha_1 GetT_g + \alpha_2 AS_a + \alpha_3 Post_t + \alpha_4 (GetT \times Post)_{gt} + \alpha_5 (GetT \times AS)_{ga} + \alpha_6 (Post \times AS)_{ta} \\ &+ \alpha_7 (GetT \times Post \times AS)_{agt} + u_{igta} \end{aligned}$$

In this case, since it is only the students who are not in the after school program ($AS = 0$) that get treated when the the treated classes ($GetT = 1$) adopt the book club when $Post = 1$, we would expect the treatment effect to manifest as $\alpha_4 > 0$ and $\alpha_4 + \alpha_7 = 0$ since the after school kids experience no change in the post period.

I simulate a new set of reading scores that based on a DGP that includes the after school program effects. I estimate a simple DID specification as well as a triple differences specification.

```
set.seed(8000)
scores<-as.data.frame(rep(c(1,2,3,4,5,6,7,8,9,10),times=30))
names(scores)<-c("class")
scores <- fastDummies::dummy_cols(scores, select_columns = "class")

scores$error<-rnorm(300, mean=0, sd=10)
scores$aftsch<-rbinom(300,1,0.4)
#suppose teachers in the better performing classes (classes, 7,8,9,10) select to participate in the boo
scores$treat<-0
scores$treat[scores$class%in%c(7,8,9,10)]<-1
```

```
scores$treatnotaftsch<-0
scores$treatnotaftsch[scores$treat==1 & scores$aftsch==0]<-1

tau<-10

#the data generating process
scores$read4<-85+13*scores$aftsch+tau*scores$treatnotaftsch+(-10)*scores$class_1+(-15)*scores$class_2+(

scores$year<-"2001"

scores01<-scores

rm(scores)

scores<-as.data.frame(rep(c(1,2,3,4,5,6,7,8,9,10),times=30))
names(scores)<-c("class")
scores <- fastDummies::dummy_cols(scores, select_columns = "class")

scores$error<-rnorm(300, mean=0, sd=10)
scores$aftsch<-rbinom(300,1,0.5)

scores$treat<-0
scores$treat[scores$class%in%c(7,8,9,10)]<-1

scores$treatnotaftsch<-0
scores$treatnotaftsch[scores$treat==1 & scores$aftsch==0]<-1

#the data generating process
scores$read4<-78+18*scores$aftsch+(-10)*scores$class_1+(-15)*scores$class_2+(-5)*scores$class_3+(-8)*sc

scores$year<-"2000"

scores00<-scores
rm(scores)

scores<-rbind(scores01, scores00)


scores$post<-0
scores$post[scores$year=="2001"]<-1
regdid3d<-felm(read4~post+treat+post*treat,scores)



regdid3dinter<-felm(read4~post+treat+aftsch+post*treat+treat*aftsch+post*aftsch+post*aftsch*treat,scores



stargazer( regdid3d,regdid3dinter,  type="latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Oct 25, 2020 - 12:05:49 PM

The first thing to notice is that the simple difference in difference specification now gives me an underestimate

Table 7:

	Dependent variable:	
	read4	
	(1)	(2)
post	0.060	5.448***
	(1.427)	(1.566)
treat	13.829***	14.809***
	(1.596)	(1.852)
aftsch		19.780***
		(1.677)
post:treat	7.922***	10.612***
	(2.257)	(2.490)
treat:aftsch		−1.574
		(2.652)
post:aftsch		−6.819***
		(2.439)
post:treat:aftsch		−9.504**
		(3.825)
Constant	82.195***	72.415***
	(1.009)	(1.179)
Observations	600	600
R^2	0.311	0.528
Adjusted R^2	0.308	0.523
Residual Std. Error	13.542 (df = 596)	11.247 (df = 592)
Note:		*p<0.1; **p<0.05; ***p<0.01

of the true treatment effect. This is because the coefficient will give me the average treatment effect on the treated classes in the post period. However now that we generated data where 50% of students were already treated through the after school program and experience no change between t=0 and t=1, the average treatment effect is halved. This is not unlike how what we saw in the previous chapter where we looked at heterogeneity in treatment effects and saw that the average treatment effect can hide vast differences in treatment effects across groups. In order to reveal these, we need to add additional interaction terms. When I estimate the triple difference specification, the full picture emerges. As predicted, $E[\hat{\alpha}_4] = \tau = 10$, as modeled, and $E[\hat{\alpha}_7] = -\tau = -10$, since the treatment that starts in period 1 does not affect the after school participants.

This result can act as a good robustness test. We have just shown that the jump in test scores in 2001 is driven by students who just gained access to the book club. Students **in the same classes**, whose access to the book club did not change, did not experience this jump in scores. Their performance continues to parallel that of the control group. This is convincing because

1) it provides support for the parallel trends assumption and

2)it helps rule out that something else changed in the classes that select into treatment that could explain the shift in scores.

1.5 Generalized DID: Staggered events

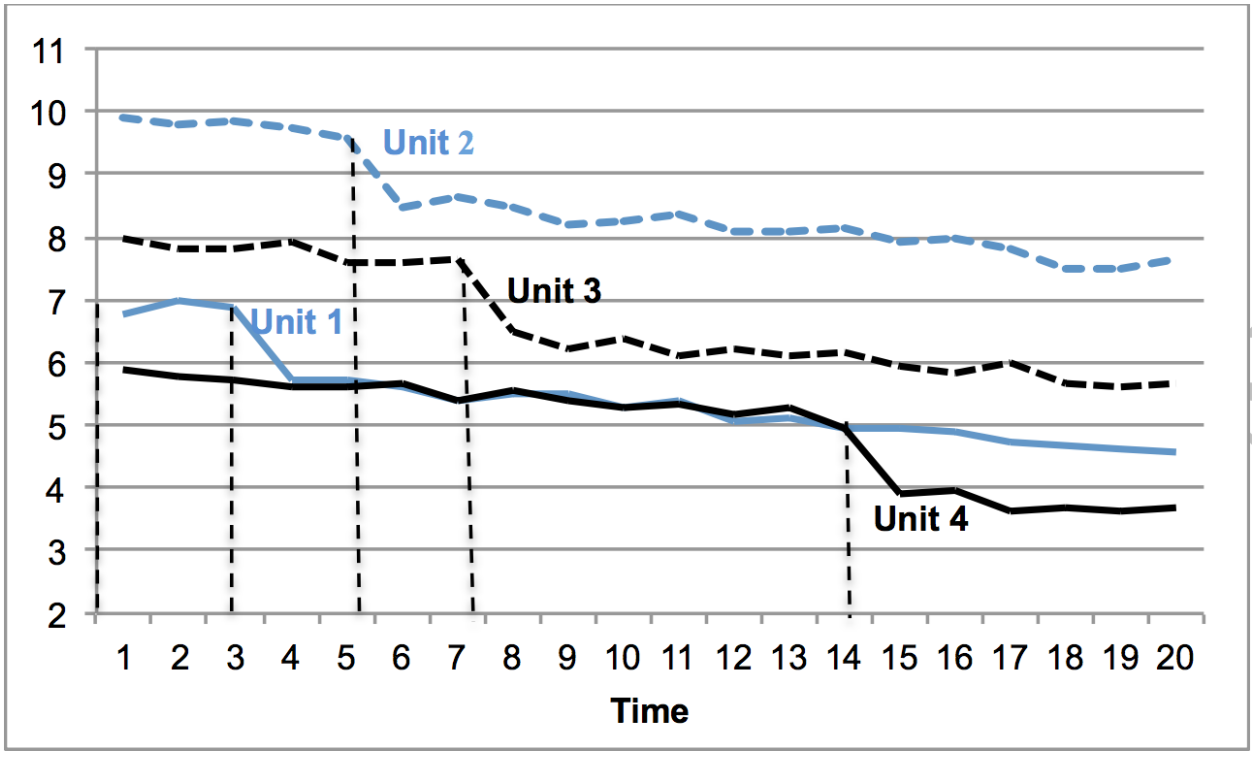


Figure 2: ...

Sometimes there are natural experiments in which there is a "pure control" that never gets treated and can serve as a counterfactual to treatment. Other times we do not have a true control, but can use our DID model for another type of comparison if a treatment or reform is rolled out over a number of months/years across various treatment units. In these cases we can use the staggered nature of the roll out to identify the causal effect of the program.

In the figure above, unit 1 is treated at time 3, unit 2 at time 5, unit 3 at time 7, etc. The units that have not yet been treated at time t serve as the control group for the units that are being treated at time t. The

impact we are trying to capture is the change in the outcome (here the decline) measured on the vertical axis that seems to happen whenever a unit enters treatment.

In a regression framework, we would estimate the following,

$$y_{it} = \beta_0 + \beta D_{it} + \gamma_i + \phi_t + u_{it}$$

Here too, the key assumption is that changes in the control group are a good counterfactual for the annual change in the treatment group and the way you would test these assumptions are very similar to the methods we have already discussed. The main difference is that transition into treatment does not happen at a fixed point in time. In order to do the robustness test described earlier in this context, it usually makes sense to generate a new set of "relative to event time" indicators that are centered around the treatment year of a particular unit. Thus $t = 0$ would be year 3 for unit 1, year 5 for unit 2, 7 for unit 3 and 14 for unit 4. $t - 1$ would be year 2 for unit 1, year 4 for unit 2 etc. These event time indicators can then be used to "line up" the treatment years to visualize and test the parallel trends assumptions.