

## Eco311 Optional Reading: Event Study

(Jing Li, Miami University)

1. This note assumes you have a solid understanding of dummy variable and interaction term. Knowledge of difference-in-difference (DID) is helpful.
2. In the simplest DID research design, there are only one treated group and one untreated group. Plus, there are only two periods—one before treatment, one after treatment. The simplest DID regression looks like

$$y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_1 D_2 + control + u \quad (1)$$

where  $y$  is the outcome variable;  $D_1$  is the dummy variable that equals one for treated group;  $D_2$  is the dummy variable that equals one for post-treatment periods;  $D_1 D_2$  is their product (interaction term). The DID estimate is  $\beta_3$ , which measures the after-treatment gap between the two groups minus the before-treatment gap:

$$\beta_3 = (\bar{y}^{treated,after} - \bar{y}^{untreated,after}) - (\bar{y}^{treated,before} - \bar{y}^{untreated,before}) \quad (2)$$

3. In general, there can be multiple pre-treatment periods, and multiple post-treatment periods. Or, there can be multiple untreated groups and multiple treated groups. The dates of treatment can be the same or different for those treated groups (patients may start taking a new medicine at different dates), the latter is called staggered treatment. The magnitude of treatment effect may vary across treated groups (a new medicine may be more effective for some patients than others). In those general cases, event study or dynamic DID regression can be applied.

### Two Groups, Multiple Periods, Static DID

4. We use Monte Carlo to illustrate the idea, which allows us to compare results to truth. We assume there are one treated group and one untreated group. There are ten periods  $t = 2001, 2002, \dots, 2010$ . In each period, there are 100 observations for each group<sup>1</sup>.

---

<sup>1</sup>The data are panel data if those 100 entities are identical over time. In that case, fixed effect may be accounted for.

5. The outcome variable  $y$  follows  $N(2, 1)$  distribution for the untreated group. There is no treatment or intervention.
6. For the treated group, the distribution is  $N(\mu_t, 1)$ , where  $\mu_t = 3$  for  $t = 2001, 2002, \dots, 2005$ ,  $\mu_t = 4$  for  $t = 2006$ ,  $\mu_t = 5$  for  $t = 2007$ ,  $\mu_t = 6$  for  $t = 2008$ ,  $\mu_t = 7$  for  $t = 2009$ , and  $\mu_t = 8$  for  $t = 2010$ . In words, the intervention occurs in 2006, and the intervention causes the mean value to increase by one in each year after 2005. Note that we allow the treated group differs from the untreated group in terms of mean value before treatment.
7. The R codes that generate data are

```

> set.seed(1234)
> ns = 100
> nt = 10
> n = ns*nt
> tinde = 2000+sort(rep(seq(1,nt),ns))
> miuc = 2
> yc = miuc+rnorm(n)
> miut = 3
> yt = miut+rnorm(n)

> cut = 0.5
> treatment = sort(rep(seq(1,5),ns))
> yt[tinde>(2000+cut*nt)]= yt[tinde>(2000+cut*nt)]+treatment

> y = c(yc,yt)
> t = c(tinde,tinde)
> j = sort(rep(seq(1,2),ns*nt))
> data = data.frame(y,t,j)
> head(data)
      y     t j
1 0.7929343 2001 1
2 2.2774292 2001 1
3 3.0844412 2001 1
4 -0.3456977 2001 1

```

```

5 2.4291247 2001 1
6 2.5060559 2001 1

```

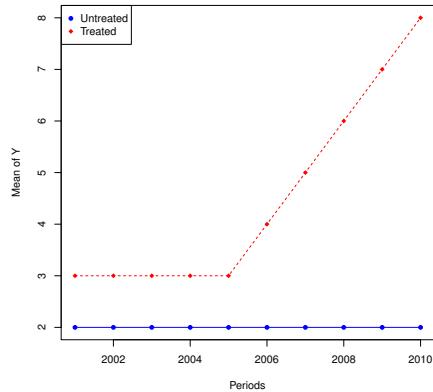
where  $j = 1$  for untreated group, and  $j = 2$  for treated group<sup>2</sup>.

8. The graph below plots the  $E(y)$  for the two groups over time. It is clear that the Parallel Trend Assumption is satisfied before 2006 since the blue and red lines are parallel before 2006. Note that we do not require the pre-treatment trends overlap (common trends)<sup>3</sup>. We only require that the gaps between two groups be constant before the treatment. The treatment effect is indicated by the widening gap between the two lines after 2006.

```

mc = rep(2,10)
mt = c(rep(3,5),4,5,6,7,8)
matplot(seq(1,10)+2000,cbind(mc,mt),type = "o", pch = c(16, 18), col = c("blue", "red"))
legend("topleft", legend = c("Untreated", "Treated"), col = c("blue", "red"), pch =

```



9. Alternatively, we can use boxplot to show distribution

```

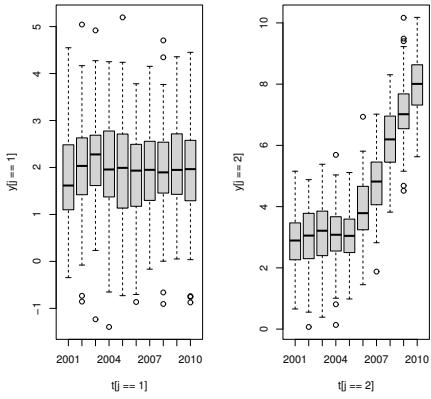
par(mfrow = c(1, 2))
boxplot(y[j==1]~t[j==1])
boxplot(y[j==2]~t[j==2])

```

---

<sup>2</sup>Data here are pooled cross sections. For panel data, there should be an ID variable.

<sup>3</sup>In a synthetic control research design, we are able to achieve common trends.



10. To see how traditional or static DID works,

- (a) Let's first compare  $E(y)$  in 2001 across the two groups

```
> mean(y[j==1&t==2001])
[1] 1.843238
> mean(y[j==2&t==2001])
[1] 2.915798
```

As expected, the sample means 1.843238, 2.915798 are close to the population means 2, 3. The difference is  $2.915798 - 1.843238 = 1.07256$ .

- (b) The same result can be obtained by regressing  $y$  onto a treatment dummy (called dummy-variable-regression I or DVR I)

```
> summary(lm(y~as.factor(j),data=data,subset=(t==2001)))$coef
            Estimate Std. Error    t value   Pr(>|t|)
(Intercept)  1.843238  0.09601828 19.196744 4.438171e-47
as.factor(j)2 1.072560  0.13579035  7.898645 1.905275e-13
```

Note that the slope coefficient  $\hat{\beta}_1 = 1.072560$  is the difference in means, or  $\hat{\beta}_1$  measures the gap. Basically, the result indicates that there is a pre-treatment gap of 1.072560 across the two groups in 2001.

- (c) The pre-treatment gap in 2002 is similar:

```
> summary(lm(y~as.factor(j),data=data,subset=(t==2002)))$coef
            Estimate Std. Error    t value   Pr(>|t|)
(Intercept)  2.0412432  0.1005297 20.304871 2.748589e-50
as.factor(j)2 0.9512149  0.1421705  6.690662 2.225836e-10
```

Focus on the slope coefficients. The fact that 0.9512149 is close to 1.072560 implies constant gaps, an evidence supporting the Parallel Trend Assumption.

- (d) Next we estimate the post-treatment gap in 2006

```
> summary(lm(y~as.factor(j),data=data,subset=(t==2006)))$coef
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.863123 0.09652508 19.30196 2.188116e-47
as.factor(j)2 2.065830 0.13650708 15.13350 6.945076e-35
```

Notice that the 2006 gap 2.065830 differs from the 2001 gap 1.07256, which suggests treatment effect.

- (e) Static DID essentially takes pre-treatment gap into account, and compares 2006 gap to 2001 gap:

$$DID^{2006vs2001} = 2.065830 - 1.072560 = 0.99327$$

- (f) The same DID result can be obtained from the simplest DID regression (1)

```
> d1 = (j==2)
> d2 = (t==2006)
> d1d2 = d1*d2
> summary(lm(y~d1+d2+d1d2,data=data,subset=(t==2006|t==2001)))$coef
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.84323826 0.09627201 19.1461483 2.584681e-58
d1TRUE      1.07255977 0.13614919  7.8778271 3.244560e-14
d2TRUE      0.01988474 0.13614919  0.1460511 8.839554e-01
d1d2        0.99327049 0.19254403  5.1586668 3.940032e-07
```

Equivalently, we can use asterisk shortcut in lm function

```
> summary(lm(y~d1*d2,data=data,subset=(t==2006|t==2001)))$coef
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.84323826 0.09627201 19.1461483 2.584681e-58
d1TRUE      1.07255977 0.13614919  7.8778271 3.244560e-14
d2TRUE      0.01988474 0.13614919  0.1460511 8.839554e-01
d1TRUE:d2TRUE 0.99327049 0.19254403  5.1586668 3.940032e-07
```

The DID estimates is given by  $\hat{\beta}_3 = 0.99327049$ . Its t value 5.1586668 exceeds 1.96. So the treatment effect is significant in 2006. The interpretations of other coefficients are

```
> mean(y[j==1&t==2001])
[1] 1.843238
> mean(y[j==2&t==2001])-mean(y[j==1&t==2001])
[1] 1.07256
> mean(y[j==1&t==2006])-mean(y[j==1&t==2001])
[1] 0.01988474
```

Can you explain why?

## Two Groups, Multiple Periods, Dynamic DID

- Now we can get a series of DID called dynamic DID—one for each year in the set of (2002, 2003, ...2010) relative to the base year 2001

```
> # Dynamic DID using 2001 as base year
> for (year in unique(t)[-1]) {
+ cat("****year is ***",year,"\n")
+ d2 = (t==year)
+ print(summary(lm(y~d1*d2,data=data,subset=(t==year|t==2001)))$coef[4,])
+ }
****year is *** 2002
  Estimate Std. Error      t value    Pr(>|t|)
-0.1213449  0.1965998 -0.6172179  0.5374457
****year is *** 2003
  Estimate Std. Error      t value    Pr(>|t|)
-0.09353037 0.19142638 -0.48859705  0.62539747
****year is *** 2004
  Estimate Std. Error      t value    Pr(>|t|)
0.02128641 0.19626575 0.10845706 0.91368808
****year is *** 2005
  Estimate Std. Error      t value    Pr(>|t|)
-0.0200076 0.1974564 -0.1013266  0.9193425
```

```

***year is *** 2006
      Estimate Std. Error      t value      Pr(>|t|)
9.932705e-01 1.925440e-01 5.158667e+00 3.940032e-07

***year is *** 2007
      Estimate Std. Error      t value      Pr(>|t|)
1.790889e+00 1.925543e-01 9.300695e+00 9.590704e-19

***year is *** 2008
      Estimate Std. Error      t value      Pr(>|t|)
3.093386e+00 1.975013e-01 1.566261e+01 2.253937e-43

***year is *** 2009
      Estimate Std. Error      t value      Pr(>|t|)
4.038495e+00 1.963129e-01 2.057173e+01 1.730128e-64

***year is *** 2010
      Estimate Std. Error      t value      Pr(>|t|)
4.983012e+00 1.960998e-01 2.541059e+01 3.441708e-85

```

The pattern is

- (a) For the pre-treatment periods  $t = 2002, 2003, 2004, 2005$ , the DID estimates are all insignificant. For instance, the t value is  $-0.6172179$  for  $DID^{2002vs2001}$ . This finding implies constant pre-treatment gaps or parallel trends.
  - (b) By contrast, for the post-treatment periods  $t = 2006, 2007, 2008, 2009, 2010$ , the DID estimates are all significant. For instance, the t value is  $5.158667$  for  $DID^{2006vs2001}$ . This finding indicates treatment effect.
  - (c) Moreover, the DID estimate is close to the true value. For instance, consider 2008 vs 2001. The true DID is  $(6 - 2) - (3 - 2) = 3$ , and the estimate is 3.093386.
12. Amazingly, we can get those dynamic DID by running just one regression, called Dynamic DID regression (DDR)

```

> d1 = (j==2)
> summary(lm(y~d1*factor(t),data=data))$coef
              Estimate Std. Error      t value      Pr(>|t|)
(Intercept) 1.84323826 0.09885437 18.6459974 1.291033e-71
d1TRUE       1.07255977 0.13980119  7.6720363 2.637733e-14

```

factor(t)2002	0.19800492	0.13980119	1.4163322	1.568356e-01
factor(t)2003	0.31136541	0.13980119	2.2272015	2.604584e-02
factor(t)2004	0.14865661	0.13980119	1.0633430	2.877561e-01
factor(t)2005	0.13497587	0.13980119	0.9654845	3.344201e-01
factor(t)2006	0.01988474	0.13980119	0.1422358	8.869082e-01
factor(t)2007	0.06889995	0.13980119	0.4928424	6.221786e-01
factor(t)2008	0.15592455	0.13980119	1.1153307	2.648441e-01
factor(t)2009	0.17488618	0.13980119	1.2509635	2.110956e-01
factor(t)2010	0.08904719	0.13980119	0.6369559	5.242272e-01
d1TRUE:factor(t)2002	-0.12134490	0.19770873	-0.6137559	5.394471e-01
d1TRUE:factor(t)2003	-0.09353037	0.19770873	-0.4730715	6.362143e-01
d1TRUE:factor(t)2004	0.02128641	0.19770873	0.1076655	9.142720e-01
d1TRUE:factor(t)2005	-0.02000760	0.19770873	-0.1011973	9.194041e-01
d1TRUE:factor(t)2006	0.99327049	0.19770873	5.0239080	5.517715e-07
d1TRUE:factor(t)2007	1.79088912	0.19770873	9.0582196	3.094442e-19
d1TRUE:factor(t)2008	3.09338643	0.19770873	15.6461801	4.082606e-52
d1TRUE:factor(t)2009	4.03849541	0.19770873	20.4264898	2.606467e-84
d1TRUE:factor(t)2010	4.98301178	0.19770873	25.2038022	8.387616e-122

where nesting the `factor` function inside the `lm` function effectively create a series of dummy variables, one for each value of 2001, 2002, ... 2010. To avoid dummy variable trap, the 2001 dummy is excluded (or 2001 is the base or reference group). Note that the coefficients of those interaction terms are dynamic DID:  $DID^{2002vs2001} = -0.12134490$ ,  $DID^{2003vs2001} = -0.09353037$ , and so on. It is clear that before 2006 those interaction terms are insignificant, while after 2006 they become significant. The interpretations of other coefficients are

- (a)  $1.84323826 = E(y|j = 1, t = 2001)$
- (b)  $1.07255977 = E(y|j = 2, t = 2001) - E(y|j = 1, t = 2001)$ . Its significant t-value 7.6720363 implies a 2001 gap
- (c)  $0.19800492 = E(y|j = 1, t = 2002) - E(y|j = 2, t = 2001)$  Its insignificant t-value 1.4163322 implies no change for the untreated group from 2001 to 2002
- (d)  $0.31136541 = E(y|j = 1, t = 2003) - E(y|j = 2, t = 2001)$  Its significant t-value 2.2272015 is an example of type-I-error.

```

> mean(y[j==1&t==2001])
[1] 1.843238
> mean(y[j==2&t==2001])-mean(y[j==1&t==2001])
[1] 1.07256
> mean(y[j==1&t==2002])-mean(y[j==1&t==2001])
[1] 0.1980049
> mean(y[j==1&t==2003])-mean(y[j==1&t==2001])
[1] 0.3113654

```

13. So far 2001 is used as the base year, and all static DID comparison is relative to 2001. In practice, we prefer using 2005, one period before the treatment, as the base year. The new dynamic DID are

```

> for (year in unique(t)[-5]) {
+   cat("*****year is",year,"\n")
+   d2 = (t==year)
+   print(summary(lm(y~d1+d2+d1:d2,data=data,subset=(t==year|t==2005)))$coef[4,])
+ }
*****year is 2001
  Estimate Std. Error    t value  Pr(>|t|)
0.0200076  0.1974564  0.1013266  0.9193425
*****year is 2002
  Estimate Std. Error    t value  Pr(>|t|)
-0.1013373  0.2018972 -0.5019253  0.6159990
*****year is 2003
  Estimate Std. Error    t value  Pr(>|t|)
-0.07352277 0.19686307 -0.37347162  0.70899732
*****year is 2004
  Estimate Std. Error    t value  Pr(>|t|)
0.0412940  0.2015720  0.2048599  0.8377869
*****year is 2006
  Estimate Std. Error    t value  Pr(>|t|)
1.013278e+00 1.979500e-01 5.118858e+00 4.805360e-07
*****year is 2007
  Estimate Std. Error    t value  Pr(>|t|)

```

```

1.810897e+00 1.979600e-01 9.147789e+00 3.103285e-18
*****year is 2008
      Estimate   Std. Error      t value      Pr(>|t|)
3.113394e+00 2.027752e-01 1.535392e+01 4.484569e-42
*****year is 2009
      Estimate   Std. Error      t value      Pr(>|t|)
4.058503e+00 2.016178e-01 2.012968e+01 1.420013e-62
*****year is 2010
      Estimate   Std. Error      t value      Pr(>|t|)
5.003019e+00 2.014104e-01 2.483993e+01 8.727578e-83

```

Or equivalently, we can run a new DDR after we relevel the period

```

> d1 = (j==2)
> rt = relevel(as.factor(t), ref="2005")
> summary(lm(y~d1*factor(rt),data=data))$coef
      Estimate Std. Error      t value      Pr(>|t|)
(Intercept) 1.97821413 0.09885437 20.0113987 2.782891e-81
d1TRUE       1.05255217 0.13980119  7.5289216 7.725425e-14
factor(rt)2001 -0.13497587 0.13980119 -0.9654845 3.344201e-01
factor(rt)2002  0.06302905 0.13980119  0.4508478 6.521486e-01
factor(rt)2003  0.17638954 0.13980119  1.2617171 2.071992e-01
factor(rt)2004  0.01368073 0.13980119  0.0978585 9.220546e-01
factor(rt)2006 -0.11509114 0.13980119 -0.8232486 4.104658e-01
factor(rt)2007 -0.06607593 0.13980119 -0.4726421 6.365206e-01
factor(rt)2008  0.02094868 0.13980119  0.1498462 8.809012e-01
factor(rt)2009  0.03991031 0.13980119  0.2854790 7.753070e-01
factor(rt)2010 -0.04592868 0.13980119 -0.3285286 7.425468e-01
d1TRUE:factor(rt)2001 0.02000760 0.19770873  0.1011973 9.194041e-01
d1TRUE:factor(rt)2002 -0.10133731 0.19770873 -0.5125586 6.083173e-01
d1TRUE:factor(rt)2003 -0.07352277 0.19770873 -0.3718742 7.100263e-01
d1TRUE:factor(rt)2004  0.04129400 0.19770873  0.2088628 8.345768e-01
d1TRUE:factor(rt)2006  1.01327809 0.19770873  5.1251054 3.263027e-07
d1TRUE:factor(rt)2007  1.81089672 0.19770873  9.1594170 1.264255e-19
d1TRUE:factor(rt)2008  3.11339402 0.19770873 15.7473774 9.884420e-53

```

```
d1TRUE:factor(rt)2009 4.05850301 0.19770873 20.5276872 4.694769e-85
d1TRUE:factor(rt)2010 5.00301938 0.19770873 25.3049995 1.210114e-122
```

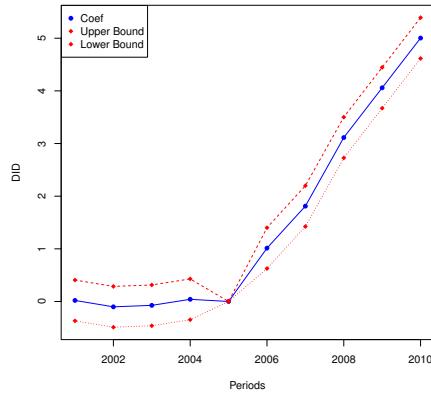
- (a) Keep in mind that all comparisons are relative to 2005. For instance, below codes show how to interpret the first three coefficients

```
> mean(y[j==1&t==2005])
[1] 1.978214
> mean(y[j==2&t==2005])-mean(y[j==1&t==2005])
[1] 1.052552
> mean(y[j==1&t==2001])-mean(y[j==1&t==2005])
[1] -0.1349759
```

- (b) Look at the interaction term coefficient of  $d1TRUE : factor(rt)2006$ , which is  $DID^{2006vs2005} = 1.01327809$ . It is equivalent to

```
> (mean(y[j==2&t==2006])-mean(y[j==1&t==2006]))-(mean(y[j==2&t==2005])-mean(y[j==1&t==2005]))
[1] 1.013278
```

14. In the literature it becomes increasingly popular to use coefficient plot other than a table to summarize the DDR. Basically we need to plot the dynamic DID estimates (coefficients of interaction terms, blue line) and their 95 percent confidence intervals (two red lines). A trick is that we need to insert the value 0 for the base year 2005, which is excluded from the regression to avoid dummy variable trap<sup>4</sup>.



The R codes to draw that coefficient plot are

<sup>4</sup>If we include all dummies, one for each year including 2005, then the sum of those dummies equal the constant term, a situation called perfect multicollinearity. Similarly, if we include all interaction terms of year dummies and the treatment dummy, their sum would duplicate the treatment dummy.

```

> didcoef = summary(lm(y~d1*factor(rt),data=data))$coef[12:20,1]
> didcoef = c(didcoef[1:4],0,didcoef[5:9])
> didse = summary(lm(y~d1*factor(rt),data=data))$coef[12:20,2]
> didse = c(didse[1:4],0,didse[5:9])
> didub = didcoef+1.96*didse
> didlb = didcoef-1.96*didse
>
> yind = seq(2001,2010)
> matplot(yind,cbind(didcoef,didub,didlb),type = "o", pch = c(16, 18, 18), col = c(
> legend("topleft", legend = c("Coef", "Upper Bound", "Lower Bound"), col = c("blue"

```

In short, we need to look for these patterns in the coefficient plot

- (a) Before the treatment date, the parallel trend assumption holds if value 0 is inside the 95 percent confidence intervals (two red lines)
- (b) After the treatment date, evidence for treatment effect is present if value 0 is outside the 95 percent confidence intervals

### Three Groups, Staggered Treatment, Heterogenous Treatment Effect

15. Now suppose there is another treated group. So in total there are two treated groups and one untreated group. For the second treated group, the treatment occurs in 2008, later than the first treated group. This is an example of staggered treatment. Furthermore, we let the treatment effects of the two treated groups be different or heterogenous:

```

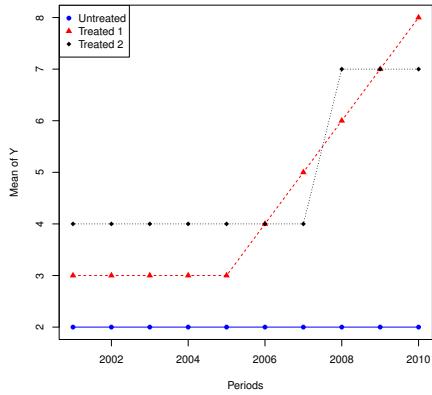
> miut2 = 4
> yt2 = miut2+rnorm(n)
>
> cut2 = 0.7
> treatment2 = sort(rep(c(3,3,3),ns))
> yt2[tinde>(2000+cut2*nt)]= yt2[tinde>(2000+cut2*nt)]+treatment2
>
> y = c(yc,yt,yt2)
> t = c(tinde,tinde,tinde)

```

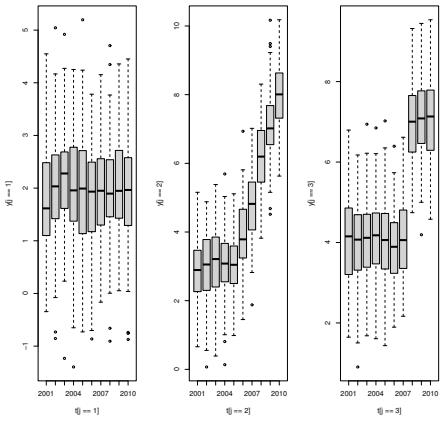
```

> j = sort(rep(seq(1,3),ns*nt))
> data = data.frame(y,t,j)
>
> mc = rep(2,10)
> mt = c(rep(3,5),4,5,6,7,8)
> mt2 = c(rep(4,7),7,7,7)
> matplot(seq(1,10)+2000,cbind(mc,mt,mt2),type = "o", pch = c(16, 17, 18), col = c(
> legend("topleft", legend = c("Untreated", "Treated 1", "Treated 2"), col = c("blue",
> par(mfrow = c(1, 3))
> boxplot(y[j==1]^t[j==1])
> boxplot(y[j==2]^t[j==2])
> boxplot(y[j==3]^t[j==3])

```



where the black line represents the second treated group. Can you see staggered treatment and heterogeneous treatment in the diagram above? Their boxplots are below



16. In this case, the most straightforward approach is to run two DDRs that compare separately each treated group to the untreated group. For instance, the DDR below uses 2007 as the base year, and compares the second treated group to the untreated group. To save space, only the coefficients of interaction terms are shown here

```
> # group-wise DDR
> d1 = (j==3)
> rt = relevel(as.factor(t), ref="2007")
> summary(lm(y~d1*factor(rt),data=data, subset=(j==1|j==3)))$coef[12:20,]
              Estimate Std. Error      t value    Pr(>|t|)
d1TRUE:factor(rt)2001  0.05650582  0.2013342  0.2806568 7.790030e-01
d1TRUE:factor(rt)2002 -0.20002712  0.2013342 -0.9935077 3.205840e-01
d1TRUE:factor(rt)2003 -0.21926917  0.2013342 -1.0890804 2.762510e-01
d1TRUE:factor(rt)2004 -0.06961283  0.2013342 -0.3457576 7.295616e-01
d1TRUE:factor(rt)2005 -0.10289971  0.2013342 -0.5110890 6.093457e-01
d1TRUE:factor(rt)2006 -0.14639876  0.2013342 -0.7271429 4.672244e-01
d1TRUE:factor(rt)2008  2.85251240  0.2013342 14.1680447 1.783939e-43
d1TRUE:factor(rt)2009  2.86809196  0.2013342 14.2454262 6.546137e-44
d1TRUE:factor(rt)2010  2.95565828  0.2013342 14.6803563 2.155293e-46
```

We see that coefficients of interaction terms are all insignificant before 2008, and all significant after 2008. The estimated treatment effects 2.8525124, 2.86809196, 2.95565828 are all close to true values 3, 3, 3. For easy comparison, the DDR that compares the first treated group and untreated group are duplicated below

```

> d1 = (j==2)
> rt = relevel(as.factor(t), ref="2005")
> summary(lm(y~d1*factor(rt), data=data, subset=(j==1 | j==2)))$coef[12:20,]
              Estimate Std. Error    t value   Pr(>|t|)
d1TRUE:factor(rt)2001  0.02000760  0.1977087  0.1011973 9.194041e-01
d1TRUE:factor(rt)2002 -0.10133731  0.1977087 -0.5125586 6.083173e-01
d1TRUE:factor(rt)2003 -0.07352277  0.1977087 -0.3718742 7.100263e-01
d1TRUE:factor(rt)2004  0.04129400  0.1977087  0.2088628 8.345768e-01
d1TRUE:factor(rt)2006  1.01327809  0.1977087  5.1251054 3.263027e-07
d1TRUE:factor(rt)2007  1.81089672  0.1977087  9.1594170 1.264255e-19
d1TRUE:factor(rt)2008  3.11339402  0.1977087 15.7473774 9.884420e-53
d1TRUE:factor(rt)2009  4.05850301  0.1977087 20.5276872 4.694769e-85
d1TRUE:factor(rt)2010  5.00301938  0.1977087 25.3049995 1.210114e-122

```

The coefficients of interaction terms are all insignificant before 2006, and all significant after 2006. The estimated treatment effects 1.01327809, 1.81089672, 3.11339402, 4.05850301, 5.00301938 are all close to true values 1, 2, 3, 4, 5.

17. We are ready to compute the average treatment effect ATE. For example, the immediate or lag 0 ATE is the average of treatment effect in 2008 for the 2nd treated group and treatment effect in 2006 for the 1st treated group

```

> (2.85251240+1.01327809)/2
[1] 1.932895

```

The first lagged ATE and second lagged ATE are

```

> (2.86809196+1.81089672)/2
[1] 2.339494
> (2.95565828+3.11339402)/2
[1] 3.034526

```

18. It turns out those ATE can be approximately obtained based on an even-study-regression ESR given by

```
> # ESR
```

```

> ttr = rep(2011,3*n)
> ttr[j==2] = 2006
> ttr[j==3] = 2008
> tgap = t-ttr
> unique(tgap)
[1] -10  -9  -8  -7  -6  -5  -4  -3  -2  -1   0   1   2   3   4
> d1 = (j==2|j==3)
> rgap = relevel(as.factor(tgap), ref=c("-1"))
> summary(lm(y~d1*factor(rgap),data=data))$coef
Coefficients: (8 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.932285 0.115574 16.719 <2e-16 ***
d1TRUE       1.617213 0.141549 11.425 <2e-16 ***
factor(rgap)-10 -0.089047 0.163447 -0.545 0.5859
factor(rgap)-9  0.108958 0.163447  0.667 0.5051
factor(rgap)-8  0.222318 0.163447  1.360 0.1739
factor(rgap)-7  0.059609 0.163447  0.365 0.7154
factor(rgap)-6  0.045929 0.163447  0.281 0.7787
factor(rgap)-5 -0.069162 0.163447 -0.423 0.6722
factor(rgap)-4 -0.020147 0.163447 -0.123 0.9019
factor(rgap)-3  0.066877 0.163447  0.409 0.6824
factor(rgap)-2  0.085839 0.163447  0.525 0.5995
factor(rgap)0   1.918862 0.115574 16.603 <2e-16 ***
factor(rgap)1   2.359449 0.115574 20.415 <2e-16 ***
factor(rgap)2   3.055074 0.115574 26.434 <2e-16 ***
factor(rgap)3   3.579681 0.141549 25.289 <2e-16 ***
factor(rgap)4   4.438359 0.141549 31.356 <2e-16 ***
d1TRUE:factor(rgap)-10      NA      NA      NA      NA
d1TRUE:factor(rgap)-9       NA      NA      NA      NA
d1TRUE:factor(rgap)-8       NA      NA      NA      NA
d1TRUE:factor(rgap)-7       0.446729 0.216220  2.066  0.0389 *
d1TRUE:factor(rgap)-6       0.401881 0.216220  1.859  0.0632 .
d1TRUE:factor(rgap)-5       0.023276 0.200181  0.116  0.9074
d1TRUE:factor(rgap)-4       0.006065 0.200181  0.030  0.9758

```

d1TRUE:factor(rgap)-3	-0.033856	0.200181	-0.169	0.8657
d1TRUE:factor(rgap)-2	-0.156059	0.200181	-0.780	0.4357
d1TRUE:factor(rgap)0	NA	NA	NA	NA
d1TRUE:factor(rgap)1	NA	NA	NA	NA
d1TRUE:factor(rgap)2	NA	NA	NA	NA
d1TRUE:factor(rgap)3	NA	NA	NA	NA
d1TRUE:factor(rgap)4	NA	NA	NA	NA

- (a) We generate a variable ttr, which indicates the treatment dates—the year when the treatment occurs. A trick is that for the untreated group, we let ttr = 2011. That is, we assume that group’s treatment does not happen until one period beyond the sample<sup>5</sup>.
- (b) We then generate tgap, the gap between t and ttr. The tgap can take values of  $-10, -9, \dots, 4$ . Not every value is taken by every group

```
> unique(tgap[j==1])
[1] -10 -9 -8 -7 -6 -5 -4 -3 -2 -1
> unique(tgap[j==2])
[1] -5 -4 -3 -2 -1  0  1  2  3  4
> unique(tgap[j==3])
[1] -7 -6 -5 -4 -3 -2 -1  0  1  2
```

For instance, only untreated group can have  $-10, -9, -8$ ; only the first treated group can have  $3, 4$ .

- (c) The treatment dummy d1 equals one for  $j = 2$  or  $j = 3$ .
- (d) Now I can explain why there are so many NAs in the estimation results. For instance, the variable  $d1TRUE : factor(rgap)-10$  is zero for all observations, so its coefficient can not be estimated. The coefficient of  $d1TRUE : factor(rgap)0$  can not be estimated for a different reason—that variable is identical to  $factor(rgap)0$  since d1 equals one in all the observations where  $factor(rgap)0$  equals one. The codes below explicitly generate the dummy variables for each value of tgap, and their interaction terms with the treatment dummy. We can verify that  $d1TRUE : factor(rgap) - 10$  only takes zero value, and  $factor(rgap)0$  and  $d1TRUE : factor(rgap)0$  are identical.

---

<sup>5</sup>Rigorously we should let ttr be NA for the untreated group. But R handles missing values in an awkward way, so we use this trick.

```

> k = 1
> for (i in unique(tgap)) {
+     assign(paste0("D", k), as.numeric(tgap==i))
+     assign(paste0("DD", k), as.numeric(tgap==i)*as.numeric(j!=1))
+     k = k + 1
+ }
> identical(DD1, rep(0, length(DD1)))
[1] TRUE
> identical(D11,DD11)
[1] TRUE

```

R calls those cases singularities

19. The ESR regresses  $y$  onto  $d1$ , the dummies for each value of  $tgap$ , and their interaction terms, after choosing the reference level of -1 of  $tgap$

- (a) The intercept captures the base group for which  $d1 = 0$  and  $tgap = -1$

```

> mean(y[tgap===-1&d1==0])
[1] 1.932285
> mean(y[j==1&t==2010])
[1] 1.932285

```

Note that only the untreated group in 2010 satisfies  $d1 = 0$  and  $tgap = -1$ .

- (b)  $d1$  is dummy, so its coefficient is a difference

```

> mean(y[tgap===-1&d1==1])-mean(y[tgap===-1&d1==0])
[1] 1.617213
> 0.5*mean(y[j==2&t==2005])+0.5*mean(y[j==3&t==2007])-mean(y[j==1&t==2010])
[1] 1.617213

```

Note that both the 1st treated group in 2005 and the 2nd treated group in 2007 satisfy  $d1 = 1$  and  $tgap = -1$ .

- (c) The interpretations of coefficients of  $factor(rgap)-10$ ,  $factor(rgap)-9$ , ...  $factor(rgap)-2$  are

```

> mean(y[tgap===-10&d1==0])-mean(y[tgap===-1&d1==0])
[1] -0.08904719

```

```

> mean(y[tgap===-9&d1==0])-mean(y[tgap===-1&d1==0])
[1] 0.1089577
> mean(y[tgap===-2&d1==0])-mean(y[tgap===-1&d1==0])
[1] 0.08583899

```

- (d) Most importantly, interpretations of coefficients of  $\text{factor}(rgap)0, \text{factor}(rgap)1, \dots \text{factor}(rgap)4$  are based on the fact that (explained before) those variables are identical to  $d1TRUE : \text{factor}(rgap)0, d1TRUE : \text{factor}(rgap)1, \dots d1TRUE : \text{factor}(rgap)4$ . That is, we hold constant  $d1 = 1$ , and compare different value of tgap. Thus, the interpretations are

```

> mean(y[tgap==0&d1==1])-mean(y[tgap===-1&d1==1])
[1] 1.918862
> mean(y[tgap==1&d1==1])-mean(y[tgap===-1&d1==1])
[1] 2.359449
> mean(y[tgap==2&d1==1])-mean(y[tgap===-1&d1==1])
[1] 3.055074
> mean(y[tgap==3&d1==1])-mean(y[tgap===-1&d1==1])
[1] 3.579681
> mean(y[tgap==4&d1==1])-mean(y[tgap===-1&d1==1])
[1] 4.438359

```

Focus on the immediate ATE 1.918862. We can show it equals to

```

> 0.5*mean(y[j==2&t==2006])+0.5*mean(y[j==3&t==2008])
[1] 5.46836
> 0.5*mean(y[j==2&t==2005])+0.5*mean(y[j==3&t==2007])
[1] 3.549498
> 5.46836-3.549498
[1] 1.918862

```

This immediate ATE differs slightly from the immediate ATE 1.932895 that we obtain from two DDRs because here we implicitly use the same base group:

$$1.918862 = \left( \frac{\bar{y}^{j=2,t=2006} + \bar{y}^{j=3,t=2008}}{2} - base \right) - \left( \frac{\bar{y}^{j=2,t=2005} + \bar{y}^{j=3,t=2007}}{2} - base \right) \quad (3)$$

- (e) Interpretations of coefficients of those interaction terms are

```

> mean(y[tgap===-7&d1==1])-mean(y[tgap===-1&d1==1])-(mean(y[tgap===-7&d1==0])-mean(y[tgap===-1&d1==0]))
[1] 0.4467285

> mean(y[tgap===-6&d1==1])-mean(y[tgap===-1&d1==1])-(mean(y[tgap===-6&d1==0])-mean(y[tgap===-1&d1==0]))
[1] 0.4018813

...

```

20. To obtain the pre-treatment gap directly, we better modify ESR as

```

> d1 = (j==1)
> rgap = relevel(as.factor(tgap), ref=c("-1"))
> summary(lm(y~d1*factor(rgap),data=data))$coef
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.549498   0.081723  43.433 < 2e-16 ***
d1TRUE          -1.617213   0.141549 -11.425 < 2e-16 ***
factor(rgap)-10 -0.089047   0.163447  -0.545 0.585927
factor(rgap)-9   0.108958   0.163447   0.667 0.505064
factor(rgap)-8   0.222318   0.163447   1.360 0.173874
factor(rgap)-7   0.506338   0.141549   3.577 0.000353 ***
factor(rgap)-6   0.447810   0.141549   3.164 0.001574 **
factor(rgap)-5  -0.045886   0.115574  -0.397 0.691377
factor(rgap)-4  -0.014082   0.115574  -0.122 0.903030
factor(rgap)-3   0.033021   0.115574   0.286 0.775115
factor(rgap)-2  -0.070220   0.115574  -0.608 0.543519
factor(rgap)0    1.918862   0.115574  16.603 < 2e-16 ***
factor(rgap)1    2.359449   0.115574  20.415 < 2e-16 ***
factor(rgap)2    3.055074   0.115574  26.434 < 2e-16 ***
factor(rgap)3    3.579681   0.141549  25.289 < 2e-16 ***
factor(rgap)4    4.438359   0.141549  31.356 < 2e-16 ***
d1TRUE:factor(rgap)-10     NA       NA       NA       NA
d1TRUE:factor(rgap)-9       NA       NA       NA       NA
d1TRUE:factor(rgap)-8       NA       NA       NA       NA
d1TRUE:factor(rgap)-7   -0.446729   0.216220 -2.066 0.038907 *
d1TRUE:factor(rgap)-6   -0.401881   0.216220 -1.859 0.063173 .
d1TRUE:factor(rgap)-5   -0.023276   0.200181 -0.116 0.907441
d1TRUE:factor(rgap)-4   -0.006065   0.200181 -0.030 0.975832

```

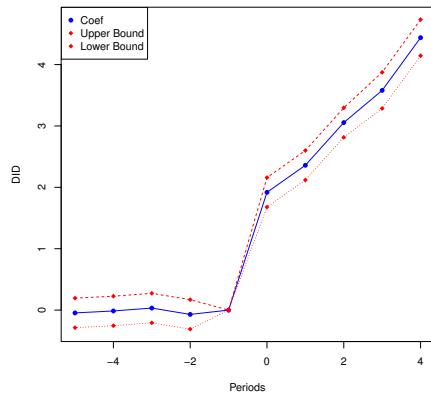
d1TRUE:factor(rgap)-3	0.033856	0.200181	0.169	0.865709
d1TRUE:factor(rgap)-2	0.156059	0.200181	0.780	0.435695
d1TRUE:factor(rgap)0	NA	NA	NA	NA
d1TRUE:factor(rgap)1	NA	NA	NA	NA
d1TRUE:factor(rgap)2	NA	NA	NA	NA
d1TRUE:factor(rgap)3	NA	NA	NA	NA
d1TRUE:factor(rgap)4	NA	NA	NA	NA

21. Now we can draw the coefficient plot

```

> didcoef= summary(lm(y~d1*factor(rgap),data=data))$coef[8:16,1]
> didcoef = c(didcoef[1:4],0,didcoef[5:9])
> didse = summary(lm(y~d1*factor(rgap),data=data))$coef[8:16,2]
> didse = c(didse[1:4],0,didse[5:9])
> didub = didcoef+1.96*didse
> didlb = didcoef-1.96*didse
> yind = seq(-5,4)
> par(mfrow = c(1, 1))
> matplot(yind,cbind(didcoef,didub,didlb),type = "o", pch = c(16, 18, 18), col = c(
> legend("topleft", legend = c("Coef", "Upper Bound", "Lower Bound"), col = c("blue"

```



Exercise 1: how to interpret the coefficient of  $factor(rgap) - 7$ , which is 0.506338. Why is its t value greater than 1.96? Why is it a good idea to exclude this coefficient in the coefficient plot above?

Exercise 2: how to use (some of) those D and DD variables on page 18 to run an ESR so that no NA (no singularities) is reported in the results?