Two-way Fixed Effects Estimation and Staggered Difference-in-Differences

September 2021

Måns Söderbom
University of Gothenburg

Introduction

- In this lecture, I provide a non-technical overview of a new literature concerned with the estimation of treatment effects in a 'staggered' difference-in-differences setting. I emphasize intuituion and a practitioner's perspective.
- Throughout the lecture, I will make use of a real dataset that has been used to analyze the effects of reforms to the divorce law on female suicides in the US.
- The main reference for the lecture is the paper by Andrew Goodman-Bacon which is forthcoming in Journal of Econometrics. I provide a (short) list of other relevant references on my web-site:

  http://www.soderbom.net/Teaching_AAU.htm

Part I: Traditional Diff-in-diff estimation

- Recap: A Difference-in-differences (DD) estimate is the difference between the change in outcomes before and after a treatment (”difference one”) comparing a treatment group to a control group (”difference two”).
- The simplest structure for DiD estimation is a two-group / two period (2 x 2) setting:

$$(\bar{y}^{POST}_{TREAT} - \bar{y}^{PRE}_{TREAT}) - (\bar{y}^{POST}_{CONTROL} - \bar{y}^{PRE}_{CONTROL})$$

- This is equal to (you should be able to prove this) the estimated coefficient on the interaction of a treatment group dummy and a post-treatment period dummy in a regression of the following form:

$$y_{it} = \gamma + \gamma_{i.} TREAT_i + \gamma_{.t} POST_t + \beta^{2x2} TREAT_i \times POST_t + u_{it}.$$

- It is *also* equal to the estimated coefficient on the treatment group dummy in the following regression in first differences:

$$\Delta y_{it} = \gamma + \beta^{2\times 2} TREAT_i + \Delta u_{it}$$

- And it is *also* equal to the estimated coefficient on the treatment group dummy in the following **two-way fixed effects*** regression:

$$y_{it} = \gamma_i + \gamma_t + \beta^{2\times 2} D_{it} + u_{it}, \qquad t = 1, 2$$

where $D_{it} = TREAT_i \times POST_t$

is a time-varying dummy variable equal to 1 if unit *i* has received treatment at time *t* , and zero otherwise.

*Why is this called a ”two-way” fixed effects model?

Examples

Throughout this lecture I will use lots of examples. I will use a real dataset that has been used to study the effect of divorce reforms on female suicides in the US. This dataset can be obtained from within Stata by typing:

use http://pped.org/bacon_example.dta

These data were used by Stevenson and Wolers (2006; QJE), and subsequently by Goodman-Bacon (2021) to study the properties of the two-way fixed effects DD estimator. The dataset is a balanced panel dataset with N=49 states and T=33 time periods (annual data for the period 1964-1996).

I have created a few useful additional variables, so an extended version of the dataset can be obtained here:

http://soderbom.net/teaching/aau/bacon_example_extended.dta

The 'roll-out' of divorce law reform in the US:

No-fault divorce year (k)	Number of states
Non-reform states	5
Pre-1964 reform states	8
1969	2
1970	2
1971	7
1972	3
1973	10
1974	3
1975	2
1976	1
1977	3
1980	1
1984	1
1985	1

- Five states undertook no reform (untreated)
- 8 states had already reformed the law prior to the first sample period
- The remaining states reformed the law at some point between 1969 – 1985.

This structure of the data implies that many alternative "2 x 2" treatment vs. control group comparisons are possible. We will come back to this point later. For now, let's focus on comparing the states that reformed the law in 1973 to the non-reform states.

Comparison: 1973 reformers vs. non-reformers

No-fault divorce year (k)	Number of states
Non-reform states	5
1973	10

- Thus, there are 5 non-reform states, which will serve as a control group here, and 10 reform states, which will form the treatment group.
- There are 33 years of data (1964-1996), but before using the full time series, I will only use data for 1973 and 1972 i.e. the year of treatment, and the year before.
- The outcome variable is female suicide mortaility (number of suicides per 1 million women) and the variable name is asmrs. Means of asmrs are as follows, for the two groups and years:

Year: 1972

nonreform	mean	N
0	71.10677	10
1	47.10805	5
Total	63.1072	15

Year: 1973

nonreform	mean	N
0	77.26132	10
1	49.00167	5
Total	67.84144	15

- We observe an increase between 1972-1973 equal to 1.9 for the control group and 6.16 for the treatment group.
- Thus DiD = 4.26.

Estimation of DiD by regression:

```
Linear regression                               Number of obs     =         15
                                                F(1, 13)          =       0.27
                                                Prob > F          =     0.6143
                                                R-squared         =     0.0157
                                                Root MSE          =      17.06

------------------------------------------------------------------------------
             |               Robust
     D.asmrs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        post |   4.260928   8.253195     0.52   0.614    -13.56902    22.09087
       _cons |   1.893617   5.710826     0.33   0.745    -10.44387    14.23111
------------------------------------------------------------------------------

Fixed-effects (within) regression               Number of obs     =         30
Group variable: stfips                          Number of groups  =         15

R-sq:                                           Obs per group:
     within  = 0.0949                                         min =          2
     between = 0.3335                                         avg =        2.0
     overall = 0.0901                                         max =          2

                                                F(2,14)           =       0.59
corr(u_i, Xb)  = 0.1982                         Prob > F          =     0.5675

                               (Std. Err. adjusted for 15 clusters in stfips)
------------------------------------------------------------------------------
             |               Robust
       asmrs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        post |   4.260928   8.242271     0.52   0.613    -13.41698    21.93884
             |
        year |
       1973  |   1.893617   5.703267     0.33   0.745    -10.33867    14.12591
             |
       _cons |    63.1072   2.199485    28.69   0.000     58.38978    67.82463
------------------------------------------------------------------------------
```

- These results simply confirm the difference in the difference in the means shown on the previous slide (4.26).
- I leave it as an exercise to show that identical results will be obtained using pooled cross section with a ”treatment x post” interaction term.
- Estimation is based on a very small sample and the reform effect is statistically insignificant (and has the ”wrong” sign).

- In applied work, it is very common for there to be more than two periods. For example, we may have panel data where *N* cross-sectional units (e.g. households or firms) are observed over *T* time periods.
- In such settings, a very common approach to estimating a linear model is to include both unit and time fixed effects in OLS estimation. This estimator is often called the two-way fixed effects estimator.
- The two-way fixed effects estimator is sometimes used in a difference-in-differences setting, where some units form a treatment group and other units form a control group. We refer to this estimator as the two-way fixed effects difference-in-differences (TWFEDD) estimator :

$$y_{it} = \alpha_i + \alpha_t + \beta^{DD} D_{it} + e_{it}, \qquad t = 1, 2, ..., T$$

- Next, we will extend the analysis by making use of the **full time series** of data (33 years) for these two groups. The DiD estimator now compares means of asmrs before and after 1973, for the control group and the treatment group:
  - Mean(asmrs) for control group 1964-1972: 48.0
  - Mean(asmrs) for control group 1973-1996: 43.4
  - Mean(asmrs) for treatment group 1964-1972: 69.8
  - Mean(asmrs) for treatment group 1973-1996: 59.5
- The fixed effect difference-in-differences estimator confirms the DiD estimate implied by the above cell means:

```
Fixed-effects (within) regression               Number of obs     =        495
Group variable: stfips                          Number of groups  =         15

R-sq:                                           Obs per group:
     within  = 0.0813                                         min =         33
     between = 0.2031                                         avg =       33.0
     overall = 0.0001                                         max =         33

                                                F(2,14)           =       4.93
corr(u_i, Xb)  = -0.2344                        Prob > F          =     0.0239

                                 (Std. Err. adjusted for 15 clusters in stfips)
```

asmrs	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
post	-5.63777	4.742855	-1.19	0.254	-15.81018	4.534642
after	-4.595844	2.832167	-1.62	0.127	-10.67024	1.478549
_cons	62.53224	1.968198	31.77	0.000	58.31088	66.7536

- ➢ The DiD estimate of implies that the reform caused 5.6 fewer suicides per 1M people.
- ➢ The effect is still statistically insignificant.
- ➢ The dummy *after* is equal to 1 for years after the reform i.e. between 1973-96.

TWFEDD estimates (year dummies included):

```
. xtreg asmrs post i.year, fe cluster(stfips)

Fixed-effects (within) regression               Number of obs     =        495
Group variable: stfips                          Number of groups  =         15

R-sq:                                           Obs per group:
     within  = 0.3556                                         min =         33
     between = 0.2031                                         avg =       33.0
     overall = 0.0689                                         max =         33

                                                F(14,14)          =          .
corr(u_i, Xb)  = -0.1217                        Prob > F          =          .

                               (Std. Err. adjusted for 15 clusters in stfips)
```

asmrs	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
post	-5.63777	4.899727	-1.15	0.269	-16.14664	4.871101
year						
1965	4.761001	1.846686	2.58	0.022	.8002534	8.721749
1966	4.589173	4.287609	1.07	0.303	-4.606833	13.78518
1967	3.181033	3.974063	0.80	0.437	-5.342486	11.70455
(…….)						
1995	-11.63895	3.815102	-3.05	0.009	-19.82153	-3.456373
1996	-13.42542	4.909438	-2.73	0.016	-23.95511	-2.89572
_cons	58.64085	3.341113	17.55	0.000	51.47488	65.80682
sigma_u	20.087174					
sigma_e	11.569197					
rho	.75090986	(fraction of variance due to u_i)				

- Replacing the *after* dummy with a full set of year dummies doesn't affect the DiD estimate and only marginally affects the DiD standard error.
- Including a full set of year dummies is the standard design for two-way fixed effects difference-in-differences (TWFEDD) estimation.

Event study analysis

- With the data set up like this, we can easily do an 'event study' of the effect of the reform. This means that we track the treatment vs. control difference in the mean of the outcome variable, centered on the year of reform.
- To do this, I use the variable `reformyr` to create a bunch of dummies equal to 1 if, at a given point in time (year), the reform happened X years ago (lagX=1) or Z years from now (leadZ=1):

```
ge YRDR=year-reformyr
forvalues k=0(1)27{
          ge lag`k'=YRDR==`k'
}
forvalues k=0(1)21{
          ge lead`k'=YRDR==-`k'
}

xtreg asmrs lead9 lead8 lead7 lead6 lead5 lead4 lead3 lead2 /*lead1*/ lag0-lag23
i.year, fe cluster(stfips)
```

- Results on the next slide. Notice that, since I exclude the dummy `lead1`, the year just prior to the reform year becomes the base category.
- Notice also the inclusion of year dummies, in addition to the lag / lead dummies. Hence, this is just another way of using the TWFEDD estimator.

Results from event study, comparing 1973 reformers to nonreformers:

```
Fixed-effects (within) regression               Number of obs      =       495
Group variable: stfips                          Number of groups   =        15

R-sq:                                           Obs per group:
     within  = 0.4011                                         min =         33
     between = 0.2031                                         avg =       33.0
     overall = 0.0570                                         max =         33

                                                F(13,14)           =          .
corr(u_i, Xb)  = -0.1847                        Prob > F           =          .

                               (Std. Err. adjusted for 15 clusters in stfips)
```

asmrs	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
lead9	-5.570686	7.043121	-0.79	0.442	-20.67668	9.535306
lead8	-5.354075	7.65803	-0.70	0.496	-21.77892	11.07077
lead7	-15.16098	15.30093	-0.99	0.339	-47.97821	17.65625
lead6	-.1727524	7.694966	-0.02	0.982	-16.67681	16.33131
lead5	-2.96979	10.88192	-0.27	0.789	-26.30919	20.36961
lead4	1.983333	7.587417	0.26	0.798	-14.29006	18.25672
lead3	4.713074	8.110706	0.58	0.570	-12.68266	22.10881
lead2	2.276694	7.9207	0.29	0.778	-14.71152	19.26491
lag0	4.260928	8.524306	0.50	0.625	-14.02189	22.54374
lag1	2.6787	7.112427	0.38	0.712	-12.57594	17.93334
lag2	1.933709	13.0168	0.15	0.884	-25.98454	29.85196
lag3	.6193878	6.397099	0.10	0.924	-13.10103	14.3398
lag4	.1154079	5.872546	0.02	0.985	-12.47995	12.71077
lag5	-7.926484	8.116357	-0.98	0.345	-25.33434	9.481371
lag6	-5.112848	8.278863	-0.62	0.547	-22.86924	12.64355
lag7	-14.09038	7.526147	-1.87	0.082	-30.23236	2.051595
lag8	-3.336827	7.807234	-0.43	0.676	-20.08168	13.40802
lag9	-6.385329	11.96864	-0.53	0.602	-32.05552	19.28486
lag10	-4.066158	5.987714	-0.68	0.508	-16.90853	8.776212
lag11	-11.72706	6.522541	-1.80	0.094	-25.71652	2.262399
lag12	-4.856163	7.727067	-0.63	0.540	-21.42907	11.71675
lag13	-11.51541	5.003911	-2.30	0.037	-22.24773	-.7830837
lag14	-10.64217	8.581902	-1.24	0.235	-29.04852	7.764175
lag15	-11.75066	8.523151	-1.38	0.190	-30.03101	6.529678
lag16	-9.84265	8.209715	-1.20	0.250	-27.45074	7.765437
lag17	-11.73118	6.450461	-1.82	0.090	-25.56604	2.103686
lag18	-15.16588	7.15236	-2.12	0.052	-30.50617	.1744053
lag19	-20.73385	8.011516	-2.59	0.021	-37.91684	-3.550853
lag20	-11.34154	7.868179	-1.44	0.171	-28.21711	5.534023
lag21	-9.228827	6.256682	-1.48	0.162	-22.64808	4.19042
lag22	-13.70959	4.430507	-3.09	0.008	-23.21209	-4.207101
lag23	-15.76541	6.714804	-2.35	0.034	-30.16723	-1.363591
year						
1965	4.616594	1.557363	2.96	0.010	1.276382	7.956806
1966	10.9827	10.83966	1.01	0.328	-12.26605	34.23145
1967	-.4175896	3.422725	-0.12	0.905	-7.758605	6.923426
(...)						
1995	-9.971529	4.273507	-2.33	0.035	-19.13729	-.8057672
1996	-10.38745	5.910166	-1.76	0.101	-23.06349	2.2886
_cons	62.35464	4.455623	13.99	0.000	52.79828	71.911

Four years before the reform, the difference in mean(asmrs) between reforming and nonreforming states was 1.98 higher than one year before the reform.

At year of reform, the difference in mean(asmrs) between reforming and nonreforming states was 4.26 higher than one year before the reform. Recall that we obtained this result earlier, when we did a simple 2 x 2 DiD comparison of the two groups in 1972 and 1973.

10 years after reform, the difference in mean(asmrs) between reforming and nonreforming states was 4.07 lower than one year before the reform.

It is often helpful to show event study results such as these graphically. See Fig. 5 in Goodman-Bacon (2021) for an example of how this can be done.

The average of the lag coefficients (pre-reform) is -2.25 and the average of the lead coefficients (post-reform) is -7.89. Their difference is -5.64 i.e. the DiD estimate obtained above.

Testing the null of common trends

- Reference: Wooldridge (2021), Sections 7-8.
- As you know, the assumption that the treatment and control group have ”common trends” (CT) is important a DiD estimate can be interpreted as an estimate of an average treatment effect, provided that.
- That is, it is assumed that underlying trends in the outcome variable are the same for the treatment group and the control group.

Tests for common trends

- You might test the null hypothesis of a common trend (for the two comparison groups) by testing for the significance of a treatment group x year interaction term – either for the pre-reform period or for the entire sample period.
- Alternatively, you could create treatment group x year dummy interactions for the pre-reform period, and investigate whether they are significant.
- Some illustrations next.

i) Test H_0: Common trend, pre-reform

```
. xtreg asmrs post i.year T_yr if year<1973, fe cluster(stfips)
note: post omitted because of collinearity

Fixed-effects (within) regression               Number of obs      =       135
Group variable: stfips                          Number of groups   =        15

R-sq:                                           Obs per group:
     within  = 0.0997                                         min =          9
     between = 0.1957                                         avg =        9.0
     overall = 0.1612                                         max =          9

                                                F(9,14)            =      2.17
corr(u_i, Xb)  = -0.9999                        Prob > F           =    0.0939

                                (Std. Err. adjusted for 15 clusters in stfips)
------------------------------------------------------------------------------
             |               Robust
       asmrs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        post |          0  (omitted)
             |
        year |
        1965 |   3.793454   1.820547     2.08   0.056     -.111231    7.698139
        1966 |   2.654079    4.25215     0.62   0.543    -6.465875    11.77403
(...)
        1971 |   3.981615   5.390691     0.74   0.472    -7.580268     15.5435
        1972 |  -3.274027    5.43591    -0.60   0.557    -14.93289     8.38484
             |
        T_yr |   1.451321   .9038805     1.61   0.131    -.4873101    3.389952
       _cons |  -1841.622   1184.449    -1.55   0.142    -4382.012    698.7683
-------------+----------------------------------------------------------------
```

ii) Test H_0: Common trend, entire period

```
. xtreg asmrs post i.year T_yr , fe cluster(stfips)

Fixed-effects (within) regression               Number of obs      =       495
Group variable: stfips                          Number of groups   =        15

R-sq:                                           Obs per group:
     within  = 0.3742                                         min =         33
     between = 0.2031                                         avg =       33.0
     overall = 0.1274                                         max =         33

                                                F(14,14)           =         .
corr(u_i, Xb)  = -0.9995                        Prob > F           =         .

                                (Std. Err. adjusted for 15 clusters in stfips)
------------------------------------------------------------------------------
             |               Robust
       asmrs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        post |   5.156138   4.027699     1.28   0.221    -3.482417    13.79469
             |
        year |
        1965 |   5.197119   1.818339     2.86   0.013     1.297169    9.097068
        1966 |   5.461408   4.257365     1.28   0.220    -3.669731    14.59255
(...)
        1995 |  -5.315251    5.50139    -0.97   0.350    -17.11456    6.484058
        1996 |  -6.665596   5.890324    -1.13   0.277    -19.29908    5.967892
             |
        T_yr |  -.6541762   .3304069    -1.98   0.068    -1.362829    .0544761
       _cons |   915.1756   433.9592     2.11   0.053    -15.57436    1845.926
-------------+----------------------------------------------------------------
```

Note: The T_yr variable is an interaction term between treatment and year:
ge T_yr=(1-nonreform)*year

- Pre-reform: Some evidence of a positively deviating time trend for reform states. But not statistically significant, hence you could accept the null hypothesis of a common trend for treatment & control groups.
- Entire period: Some evidence of a negatively deviating time trend for reform states. Statistically significant at 10% but not at 5%.

iii) Test H_0: Common time effects, pre-reform

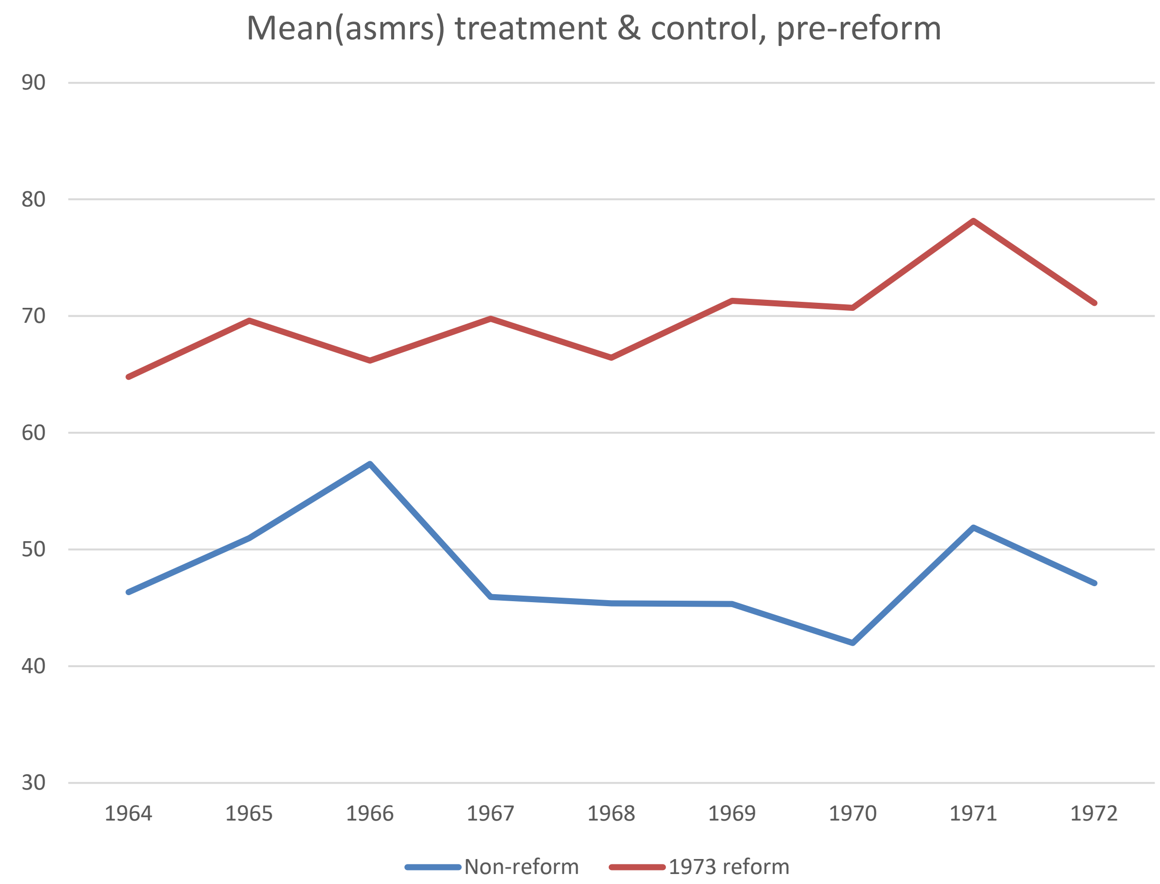
```
. xtreg asmrs post i.year Ty_1965-Ty_1972 , fe cluster(stfips)

Fixed-effects (within) regression               Number of obs      =       495
Group variable: stfips                          Number of groups   =        15

R-sq:                                           Obs per group:
     within  = 0.3658                                         min =         33
     between = 0.2031                                         avg =       33.0
     overall = 0.1183                                         max =         33

                                                F(14,14)           =         .
corr(u_i, Xb)  = -0.0207                        Prob > F           =         .

                                (Std. Err. adjusted for 15 clusters in stfips)
------------------------------------------------------------------------------
             |               Robust
       asmrs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        post |   -2.31766   6.895223    -0.34   0.742    -17.10644    12.47112
             |
        year |
        1965 |   4.616594   1.517313     3.04   0.009     1.362282    7.870906
        1966 |    10.9827   10.56089     1.04   0.316    -11.66816    33.63357
(...)|
        1995 |  -13.85236    3.97856    -3.48   0.004    -22.38552   -5.319198
        1996 |  -15.63882    5.05356    -3.09   0.008    -26.47763   -4.800016
             |
     Ty_1965 |   .2166107   3.087623     0.07   0.945    -6.405683    6.838904
     Ty_1966 |  -9.590297   11.05264    -0.87   0.400    -33.29585    14.11526
     Ty_1967 |   5.397933   6.585405     0.82   0.426    -8.726357    19.52222
     Ty_1968 |   2.600896   7.581892     0.34   0.737    -13.66065    18.86244
     Ty_1969 |   7.554019   6.504555     1.16   0.265    -6.396863     21.5049
     Ty_1970 |   10.28376   6.425647     1.60   0.132    -3.497883     24.0654
     Ty_1971 |   7.847379   6.846128     1.15   0.271    -6.836104    22.53086
     Ty_1972 |   5.570686   6.861993     0.81   0.430    -9.146825     20.2882
       _cons |   58.64085   3.342061    17.55   0.000     51.47284    65.80886
-------------+----------------------------------------------------------------
     sigma_u |  19.280436
     sigma_e |   11.58136
         rho |  .73485338   (fraction of variance due to u_i)
------------------------------------------------------------------------------

.
. test Ty_1965 Ty_1966 Ty_1967 Ty_1968 Ty_1969 Ty_1970 Ty_1971 Ty_1972

 ( 1)  Ty_1965 = 0
 ( 2)  Ty_1966 = 0
 ( 3)  Ty_1967 = 0
 ( 4)  Ty_1968 = 0
 ( 5)  Ty_1969 = 0
 ( 6)  Ty_1970 = 0
 ( 7)  Ty_1971 = 0
 ( 8)  Ty_1972 = 0

       F(  8,    14) =   11.36
            Prob > F =    0.0001
```

Note: The Ty_year variables are pre-reform time dummies interacted with a dummy for reform states:

```
forvalues k=1965(1)1972{
        ge Ty_`k'=(nonreform==0 & year==`k')
}
```

- To test the null hypothesis of common time effects for treatment & control groups in the pre-reform, we carry out a Wald test. For this test, the null hypothesis is that all the coefficients on the Ty_year interaction terms are equal to zero.
- The result from the Wald test strongly indicates that we should reject the null hypothesis of common time effects in the pre-reform period.

Exercise: The graph below shows mean values of asmrs by year for non-reform states and 1973 reform states during the pre-treatment period. Explain how these mean values relate to the the T_year interaction effects in the regression shown on the previous page.



Part II:
Staggered Diff-in-diff estimation

- Consider the two-way fixed effects DiD model (TWFEDD) introduced above:

$$y_{it} = \alpha_i + \alpha_t + \beta^{DD} D_{it} + e_{it}, \qquad t = 1, 2, ..., T$$

- We have just seen how this estimator can be used to obtain DiD estimates comparing a treatment group to a control group.
- In the previous examples, the treatment group included states that reformed the divorce law in 1973 and the control group consisted of states that did not reform divorce law over the sampling period.
- For the treatment group, the dummy variable D_{it} switched from 0 to 1 in 1973 (and remained = 1 after 1973), while for the control group D_{it} = 0 throughout the period of analysis.

- In many datasets (including the dataset that I have introduced above), treatments occur at different times. Using such a dataset to estimate treatment effects is sometimes referred to as staggered difference-in-differences estimation. This is currently a very active area of research.
- In such cases, in order to estimate the causal effect of treatment on outcomes, researchers usually estimate a two-way fixed effects regression

$$y_{it} = \alpha_i + \alpha_t + \beta \cdot D_{it} + x_{it}\lambda + e_{it}, \qquad t = 1, 2, ..., T$$

  (the vector x_{it} includes control variables but I will abstract from control variables throughout this lecture).
- Two-way fixed effects results using the full panel dataset on divorce law reform and female suicide rates are shown on the next page.

```
. xtreg asmrs post i.year, fe

Fixed-effects (within) regression               Number of obs     =      1,617
Group variable: stfips                          Number of groups  =         49

R-sq:                                           Obs per group:
     within  = 0.3461                                         min =         33
     between = 0.0293                                         avg =       33.0
     overall = 0.1462                                         max =         33

                                                F(33,1535)        =      24.62
corr(u_i, Xb)  = -0.0240                        Prob > F          =     0.0000

------------------------------------------------------------------------------
       asmrs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        post |  -3.079926   1.111656    -2.77   0.006    -5.260452   -.8993999
             |
        year |
       1965  |   5.461577    2.22522     2.45   0.014     1.096784     9.82637
       1966  |   2.624452    2.22522     1.18   0.238    -1.740341    6.989244
(...)
       1996  |  -11.38105   2.370369    -4.80   0.000    -16.03055   -6.731542
-------------+----------------------------------------------------------------
     sigma_u |   14.76249
     sigma_e |  11.014277
         rho |  .64239957   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(48, 1535) = 57.63                   Prob > F = 0.0000
```

- The two-way fixed effects estimate of beta is -3.08 (std err 1.11). This is the result reported by Goodman-Bacon on pp.12-13 of his paper.
- Can this be interpreted as a diff-in-diff estimate? If so, what are the treatment and control groups here? Until recently, these issues have not been entirely clear…

- A number of recent studies show:
  - How the two-way fixed effects estimator compares mean outcomes across groups in a difference-in-differences fashion
  - What treatment effect parameter is identified through this approach, and potential sources of bias of the estimator
  - How and why alternative specifications change estimates
- The paper by Andrew Goodman-Bacon (forthcoming, Jnl of Econometrics) provides a very clear analysis of these issues, and I will draw on it extensively during the remainder of this lecture.

- Recall that divorce law reform was 'rolled out' at different times during 1969-1985:

No-fault divorce year (k)	Number of states
Non-reform states	5
Pre-1964 reform states	8
1969	2
1970	2
1971	7
1972	3
1973	10
1974	3
1975	2
1976	1
1977	3
1980	1
1984	1
1985	1

- In the first part of this lecture, we looked specifically at the 1973 reform states and compared them to the non-reform states. But the structure of the data implies that many alternative "2 x 2" treatment vs. control group comparisons are possible........

- You could obtain DD estimates by comparing the states that reformed in 1969 (or 1970, or in any other year) to the **non-reformers**.
  - The non-reformers constitute the control group here. Notice that for the control group, the treatment dummy D_{it} is constant at **zero** throughout the period of analysis.
- This is precisely what we did in the first part of this lecture: we compared 1973 reform states to the non-reformers.

- You could also compare 'early' reformers to 'late' reformers over a period when 'late' reformers had not yet reformed. In this case, the 'late' reformers constitute the control group.
  - For example, compare the 1973 reformers to the 1985 reformers, over the period up until, but not including, 1985:

Year	**1964**	**1965**	**1966**	**(...)**	**1972**	**1973**	**1974**	**(...)**	**1983**	**1984**	1985
Treatment group:											
1973 reformers	**0**	**0**	**0**	**(...)**	**0**	**1**	**1**	**(...)**	**1**	**1**	1
Control group:											
1985 reformers	**0**	**0**	**0**	**(...)**	**0**	**0**	**0**	**(...)**	**0**	**0**	1

Note: The table shows D_{it} for the two groups. Notice that D_{it} switches from 0 to 1 in 1973 for 1973 reformers ('early' reformers) while for the 1985 reformers ('late' reformers) , D_{it} remains constant at zero until 1985. The 1985 observation is thus excluded from the present comparison.

- It would also be possible to compare states that reformed at a given point in time between 1969-1985 to the **pre-1964 reform states**.
  - The pre-1964 reform states form the control group in this case. For this control group, D_{it} is constant at **one** throughout the period of analysis.
  - The latter point is potentially confusing: we usually think about cases where $D_{it} = 1$ as treatment observations.
  - However, in general we identify DiD estimates by comparing states for which D_{it} *changes* from zero to one to states for which D_{it} *doesn't change*.
  - We thus define control group to mean states for which D_{it} doesn't change. This is an important point that we need to keep in mind in order to understand the material below: *Early treatment observations can potentially be used to form a control group.*

- In a similar spirit, you could also compare ’late’ reformers to ’early’ reformers over a period when ’early’ reformers already had reformed. In this case, the ’early’ reformers constitute the control group.
  - For example, compare the 1985 reformers to the 1973 reformers, over the period from 1973 to 1989:

Year	1964	(...)	1972	1973	1974	(...)	1984	1985	1986	(...)	1996
Control group:											
1973 reformers	0	(...)	0	1	1	(...)	1	1	1	(...)	1
Treatment group:											
1985 reformers	0	(...)	0	0	0	(...)	0	1	1	(...)	1

Note: The table shows D_{it} for the two groups. Notice that D_{it} switches from 0 to 1 in 1985 for 1985 reformers (’late’ reformers) while for the 1973 reformers (’early’ reformers) , D_{it} remains constant at one from 1973 and onwards. The observations for 1964-1972 are thus excluded from the present comparison.

- Results based on a comparison of the 1973 and 1985 reformers, where they alternate as treatment and control group as explained above, are shown on the next page (’early’ treatment vs ’late’ control; ’late’ treatment vs. ’early’ control).

i) 'early' (1973 reform) treatment vs. 'late' (1985 reform) control:

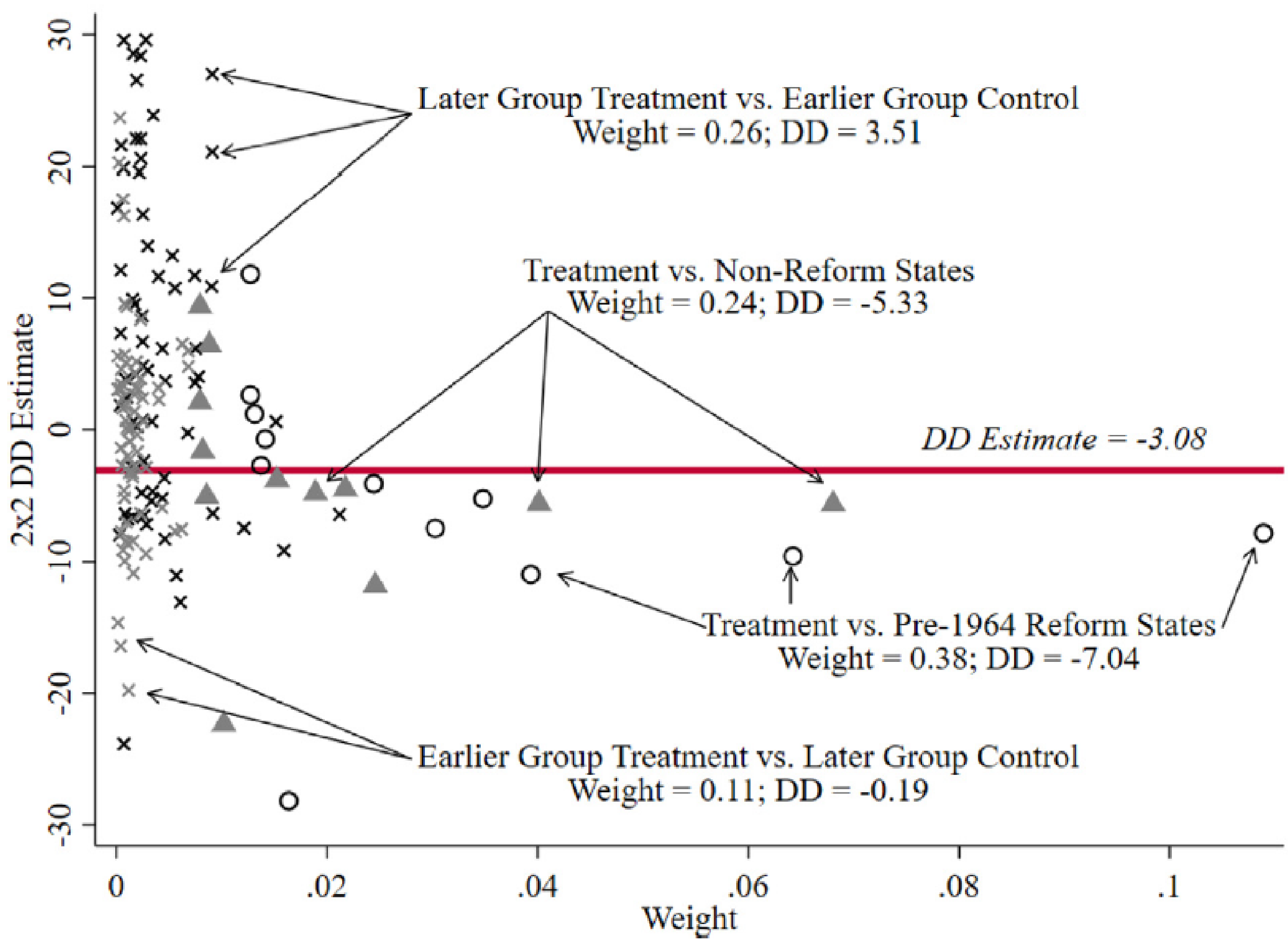
```
xtreg asmrs post i.year if (rperiod==5 | rperiod==12) & year<=1984, fe

Fixed-effects (within) regression               Number of obs      =       231
Group variable: stfips                          Number of groups   =        11

------------------------------------------------------------------------------
       asmrs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        post |   6.036513   5.600634     1.08   0.282    -5.007695    17.08072
             |
        year |
       1965  |   4.445432   5.163709     0.86   0.390    -5.737178    14.62804
       1966  |    4.76939   5.163709     0.92   0.357     -5.41322      14.952
(...)
       1984  |    -12.148   7.251698    -1.68   0.095    -26.44804     2.15203
------------------------------------------------------------------------------
```

ii) 'late' (1985 reform) treatment vs. 'early' (1973 reform) control:

```
xtreg asmrs post i.year if (rperiod==5 | rperiod==12) & year>=1973, fe

Fixed-effects (within) regression               Number of obs      =       264
Group variable: stfips                          Number of groups   =        11

------------------------------------------------------------------------------
       asmrs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        post |    21.1303   4.535741     4.66   0.000     12.19317    30.06742
             |
        year |
       1974  |  -.9724079   4.516959    -0.22   0.830    -9.872522    7.927706
       1975  |   3.125864   4.516959     0.69   0.490     -5.77425    12.02598
(...)
       1996  |  -30.43772   4.535741    -6.71   0.000    -39.37484    -21.5006
------------------------------------------------------------------------------
```

- Thus, there are 4 types of DD comparisons available here:
  - Timing vs. Non-reformers
  - Timing vs. Pre-reformers
  - Early reformers (treatment) vs. Late reformers (control)
  - Late reformers (treatment) vs. Early reformers (control)
- The key result shown by Goodman-Bacon is that the TWFE estimate based on the full panel dataset is a weighted average of all DD comparisons available in the dataset.
- In this dataset, there are 156 distinct DD components: 12 comparisons between timing groups and pre-reform states, 12 comparisons between timing and non-reform states, and $(12^2-12)/2 = 66$ comparisons between an early switcher (treatment) and late switcher (control), and 66 comparisons between a late switcher (treatment) and an early switcher (control).

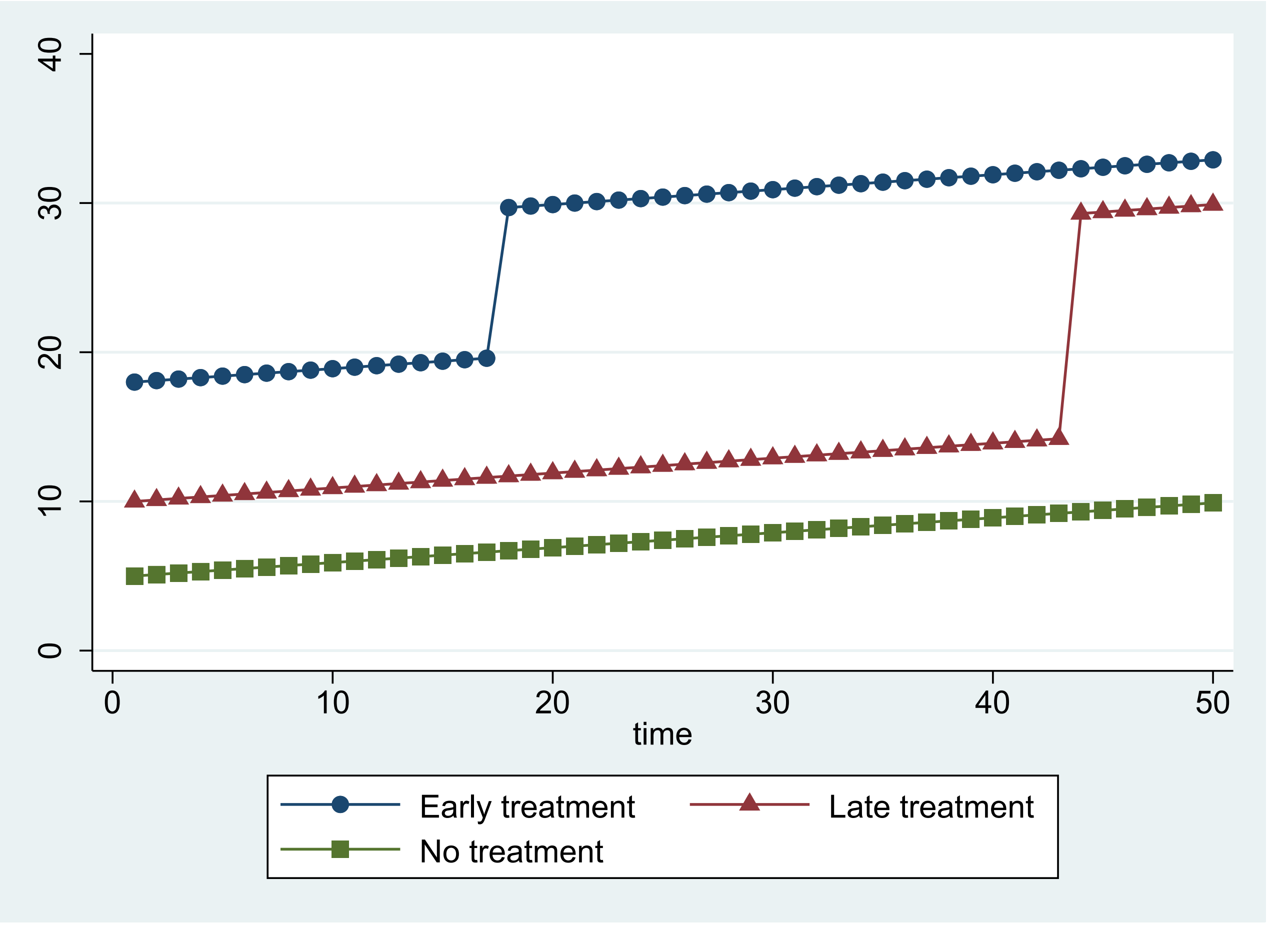
The following graph is taken from G-B's paper (page 13). It shows all 156 2x2 DD estimates and the associated weights (we will come back to how these weights are computed). The TWFEDD estimate of -3.08 is simply the weighted average of all these 2x2 DD estimates.



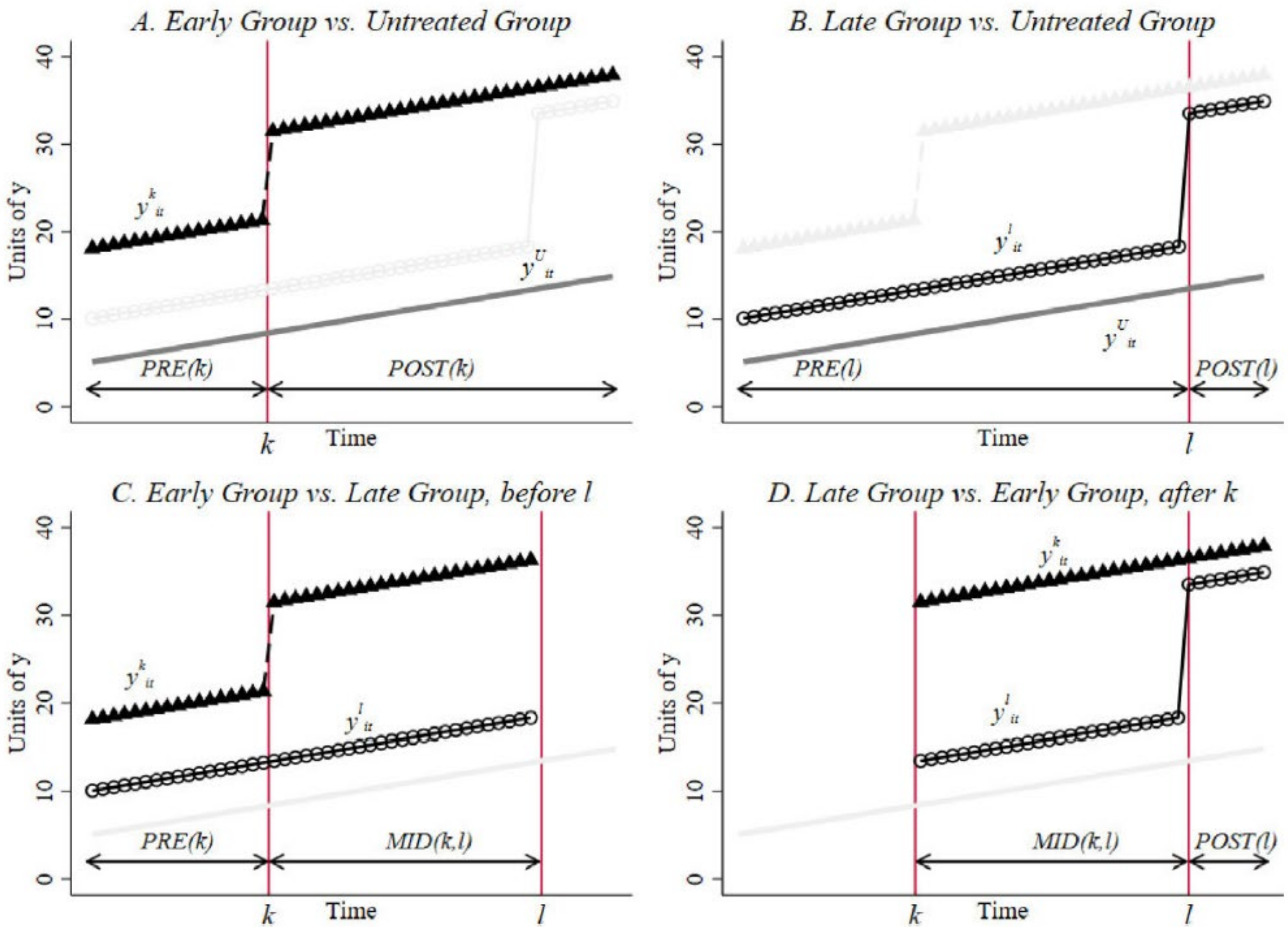
Numerical illustration

- In Section 2 of his paper, G-B provides a nice numerical illustration of the main results of his study. I will undertake a similar exercise here.
- Imagine we are analyzing an outcome variable *y* on a balanced panel dataset with T=50 time periods and N units (e.g. households). There are three distinct treatment groups:
  - One group of units receiving treatment 'early' at period 17. Observed y will thus be y(0) (outcome under no treatment) during periods 1-16 and y(1) (outcome under treatment) during periods 17-50. For this group, y(1) = y(0) + 10, i.e. the true treatment effect is 10.
  - One group of units receiving treatment 'late' at period 43. For this group, the true treatment effect is 15. Observed y will thus be y(0) during periods 1-43 and y(1) during periods 44-50. For this group, y(1) = y(0) + 15, i.e. the true treatment effect is 15.
  - One group of units receiving no treatment.
- For all groups, there is a weak positive trend in y(0).

Numerical illustration continued



There are four 2x2 DD:s…



- In two of these, the treatment effect is 10; in the other two it is 15.
- What would be the TWFEDD estimate? You might think that the answer is 12.5, but that is not the case. The answer is in fact 11.6. Thus, the lower ('early') treatment effect of 10 gets a higher weight than the higher ('late') treatment effect of 15. Why might this be?

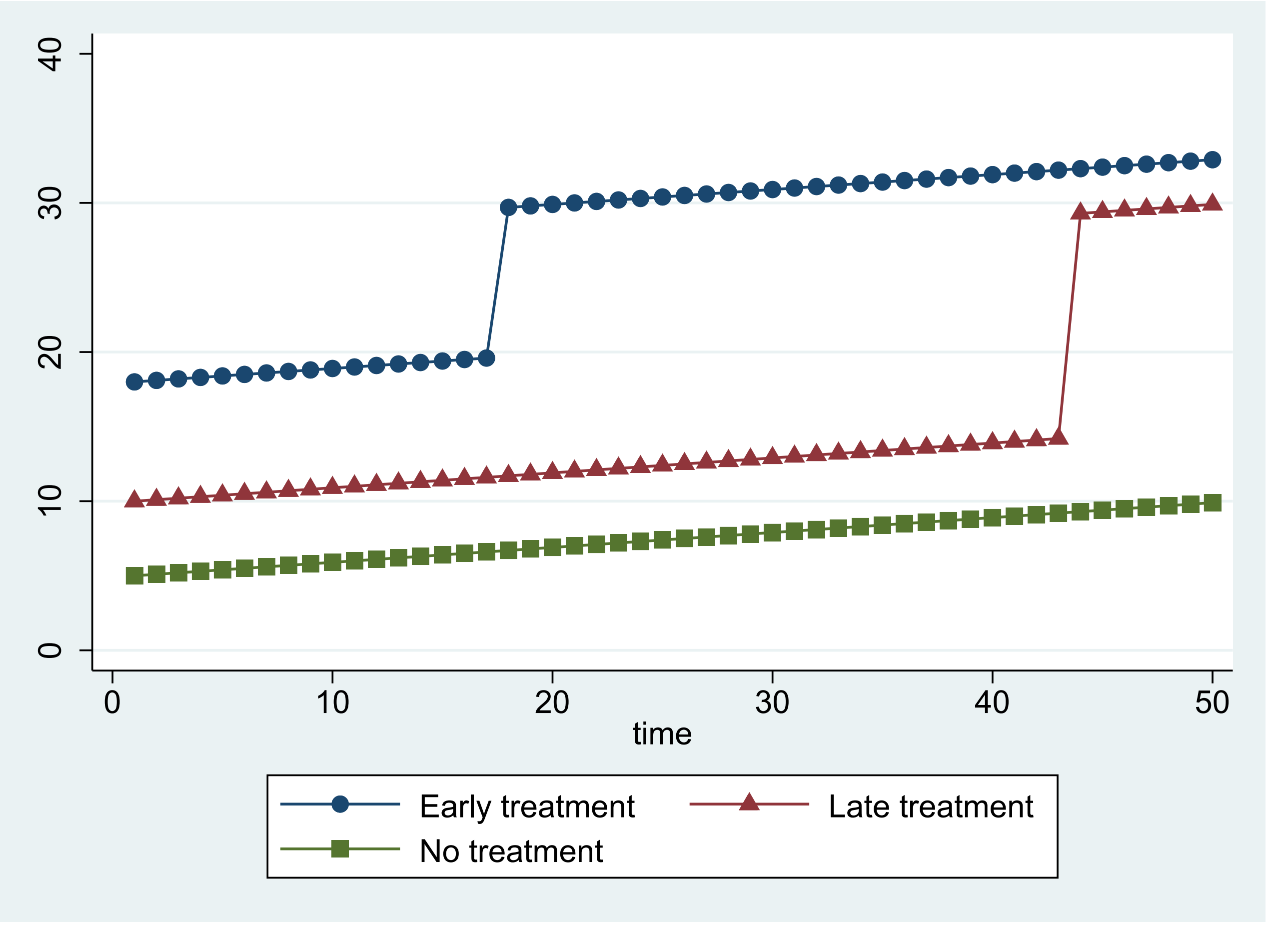
Goodman-Bacon shows that the overall TWFEDD estimator can be decomposed into the underlying 2x2 DD estimators. In the present case, we have four such estimators, in which case:

$$\hat{\beta}^{DD} = s_{kU}\hat{\beta}_{kU}^{2x2} + s_{\ell U}\hat{\beta}_{\ell U}^{2x2} + s_{k\ell}^{k}\hat{\beta}_{k\ell}^{2x2,k} + s_{k\ell}^{\ell}\hat{\beta}_{k\ell}^{2x2,\ell}$$

where the *s*__ terms are weights that depend on the variance of in D_{it} within groups and the relative size of the treatment groups. See equations (10e)-(10g) in Goodman-Bacon for the exact expressions.

Intuitively, the weight associated with a particular 2x2 DD estimate will be higher if…

- ….the within group average of D_{it} is relatively 'close' to 0.5. That is, if there is roughly the same number of ones and zeros in the series of D_{it}; which in this setting means that the switch from 0 to 1 happens at or near the middle of the sample period.
- …..a relatively large share of the sample is used for the estimation of a particular 2x2 DD estimator.



The variance of the treatment dummy is higher for the ’early treatment’ group (treatment effect 10) than for the ’late treatment’ group, simply because the switch from 0 to 1 happens closer to the middle of the period. Hence, the lower treatment effect 10 gets a higher weight in this case.

What parameter does TWFEDD identify?

- The TWFEDD estimator is a variance-weighted average of all available 2x2 DD estimators.
- The probability limit ('plim') of the TWFEDD estimator is a variance-weighted average of the average treatment effects (ATTs) for the units and periods that get used in estimation of the 2x2 DD estimators - *provided* these 2x2 DD estimators are themselves consistent estimators of their respective ATTs…
- It follows that if some of the underlying 2x2 DD estimators of ATT are in fact biased & inconsistent, then the TWFEDD won't identify the variance-weighted ATT.

What parameter does TWFEDD identify?

- Goodman-Bacon writes the probability limit of the TWFEDD estimator as follows:

$$\underset{N\to\infty}{plim}\,\hat{\beta}^{DD} = \beta^{DD} = VWATT + VWCT - \Delta ATT.$$

where:

- *VWATT* = variance-weighted ATT. This is a causal parameter of interest.
- *VWCT* = variance-weighted common trends. This term captures differences in counterfactual trends between comparison groups. This term captures the possibility that different groups might have different underlying trends in the outcome variable, which, as you know, will bias DD estimates.
- ΔATT = a weighted sum of the *change* in treatment effects within each timing group's post-period, with respect to antoher unit's treatment timing.

Some diagostics

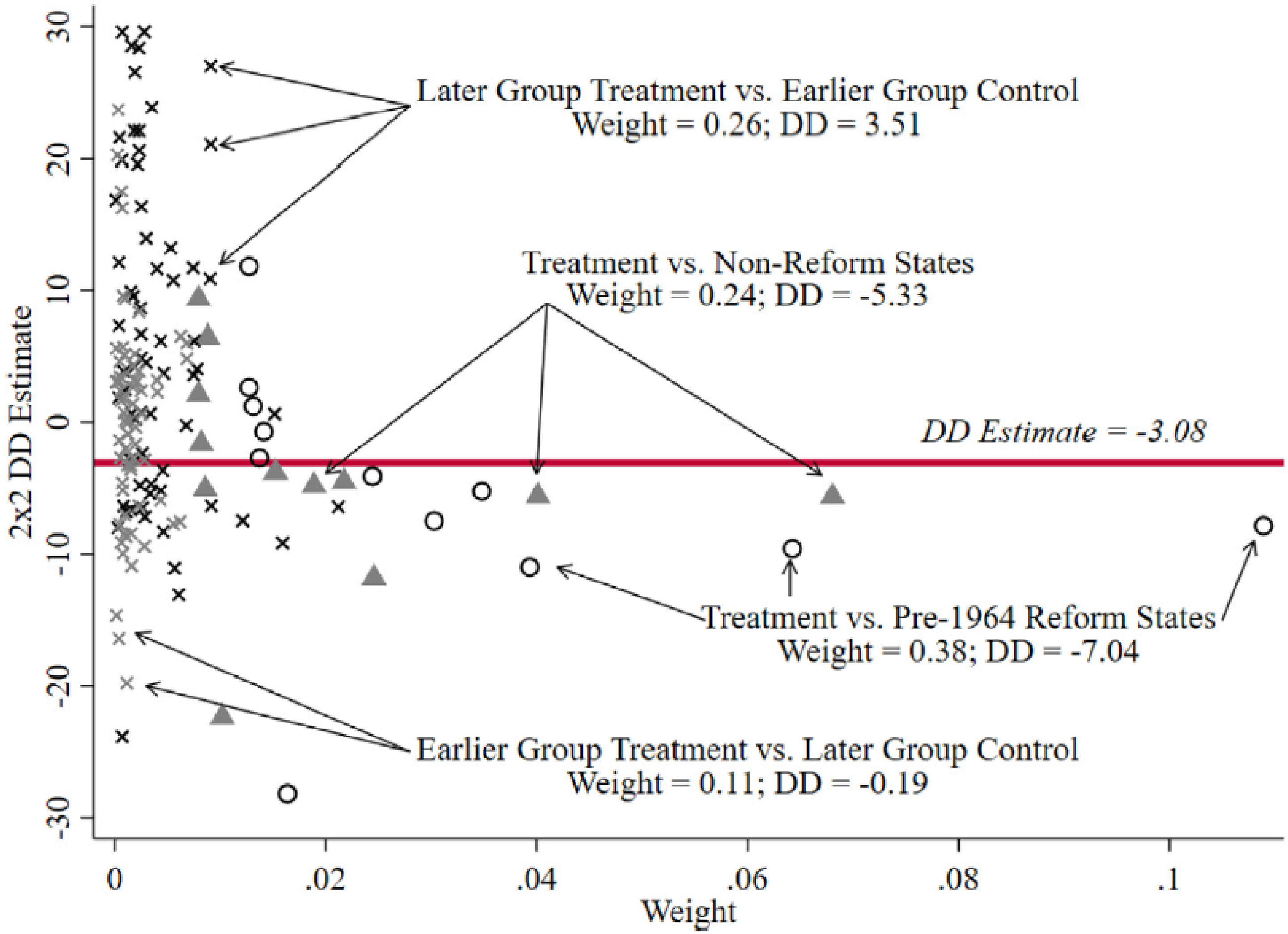
- Recall:

$$\underset{N \to \infty}{plim}\, \hat{\beta}^{DD} = \beta^{DD} = VWATT + VWCT - \Delta ATT.$$

  Obviously, none of the terms on the right-hand side of this equation is directly observable, so we can never know for certain whether our estimator would be close to VWATT in a large N sample or not…

  However, we can use the data to try to shed some light on whether our TWFEDD estimator can be interpreted as a credible estimator of VWATT.

- Using the dataset on divorce law reform and suicides, Goodman-Bacon plots each 2x2 DD component against their weight (see Figure 6 in his paper). We saw this graph earlier. This is a good way of understanding the underlying variation in the 2x2 DD components and their relative influence on the overall TWFEDD estimate. Let’s have another look at this graph.



Some observations:

(a) While the overall DD estimate is negative, there are many *positive* 2x2 DD estimates. Wrong sign?

(b) The 'late' treatment vs. 'early' control comparisons seem particularly problematic; the average of all such DD estimates is 3.51 (wrong sign?).

- I expect diagnostics for the TWFEDD estimator to be growing area of research in the near future.
- Since the TWFEDD estimate is an average of lots of 2x2 DD estimates, we can report standard diagnostic tests for the 2x2 DD estimates, e.g. common trend tests.
- As an example, I carried out 156 common trend tests – one for each 2x2 DD estimator – and found that the null hypothesis of common trends can be rejected at the 5% level in 28% of the cases (for these cases the weights sum to 36%).

Some advice for applied researchers

- Be transparent with respect to treatment timing. Show the distribution of treatment over time, and understand that groups for whom treatment happens in the middle of the sample period will get more weight in the TWFEDD estimate than groups with very early or very late treatment.
- Scrutinize the underlying 2x2 DD estimates, e.g. by using the Goodman-Bacon graph, or by carrying out diagnostic tests.
- As we have seen, the TWFEDD uses an earlier treated group as a control for a later treated group. Intuitively, this seems quite unattractive (why?), and hardly not something we would do in a 2x2 DD setting. But, like it or not, that’s what the TWFEDD estimator does. Pay special attention to these comparisons (e.g. use the G-B graph).

- You may want to take a look at the papers by Callaway & Sant'Anna (2020) and Sun & Abraham (2020), see reading list. These papers focus on generating unbiased DD estimates by removing dubious control units.
- This is an active area of research, and best-practice will likely change quickly. Keep up with the literature!