Differences-in-Differences

(v. 3.3)

Oscar Torres-Reyna

otorres@princeton.edu

August 2015

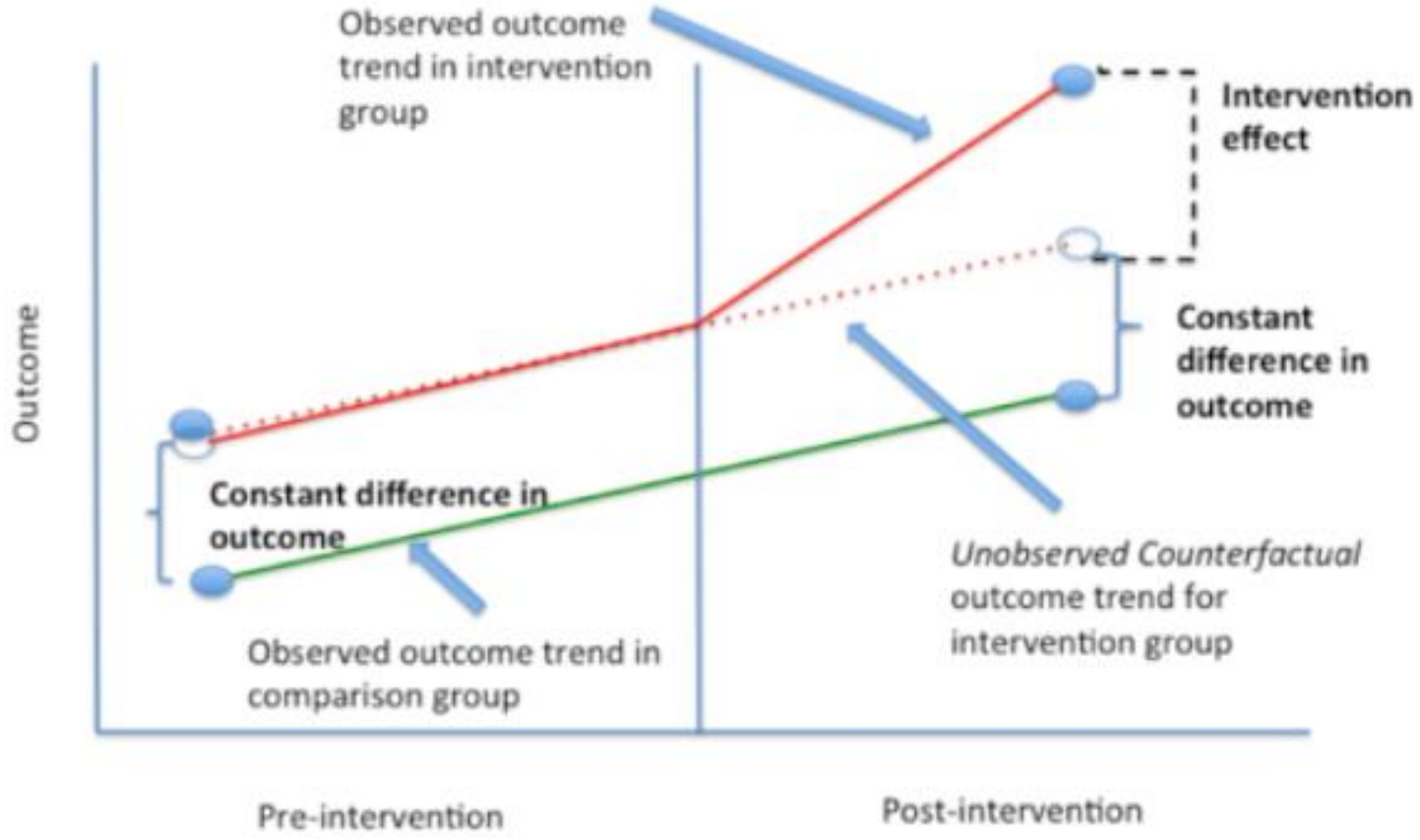
http://www.princeton.edu/~otorres/

Intro

Differences-in-Differences regression (DID) is used to asses the causal effect of an event by comparing the set of units where the event happened (treatment group) in relation to units where the event did not happen (control group).

The logic behind DID is that if the event never happens, the differences between treatment and control groups should stay the same overtime, see graph next slide.*

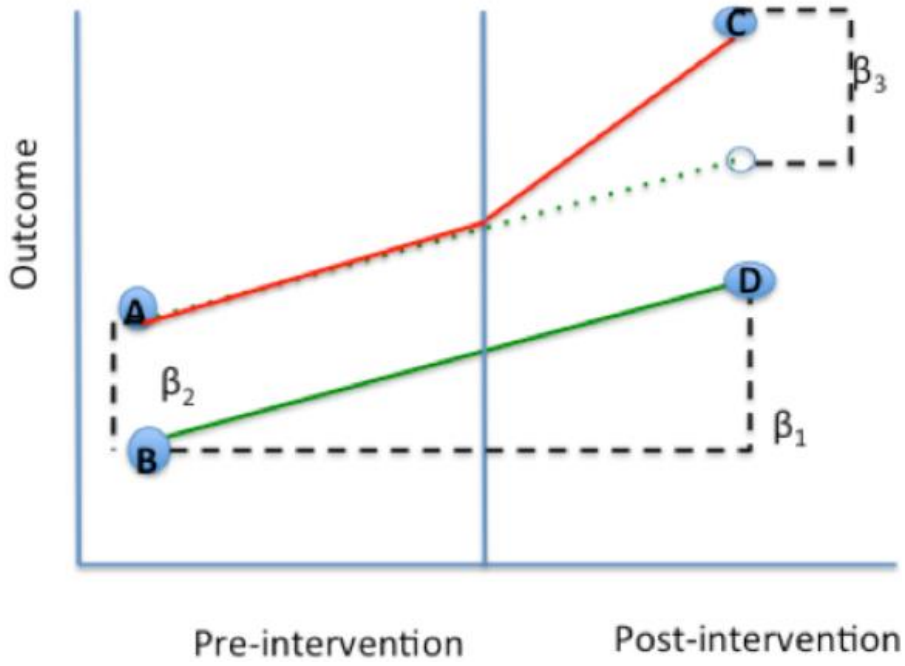
*See: https://www.publichealth.columbia.edu/research/population-health-methods/difference-difference-estimation



Source: https://www.publichealth.columbia.edu/research/population-health-methods/difference-difference-estimation

$$y = \beta_0 + \beta_1 time + \beta_2 treated + \beta_3 time * treated + \varepsilon$$

Coefficient	Calculation	Interpretation
β_0	B	Baseline average
β_1	D-B	Time trend in control group
β_2	A-B	Difference between two groups pre-intervention
β_3	(C-A)-(D-B)	Difference in changes over time



Source: https://www.publichealth.columbia.edu/research/population-health-methods/difference-difference-estimation

Intro

This document shows how to perform difference-in-differences regression in the following two situations:

- Event happened at the same time for all treated groups.

- Event is staggered across groups.

Event happens at the same time for all treated groups

Data preparation

The before/after variable

Create an indicator variable where:

- 0 = time before the event happens
- 1 = time when the event happens and after

Example:

```
use "http://www.princeton.edu/~otorres/WDI.dta", clear

* Fake event X happens in 2009 affecting all countries
* Creating the before/after dummy variable: 0 = before, 1 = after

gen after = (year >= 2009) if !missing(year)

*To check, type:

tab year after
```

Source of data: World Development Indicators, https://databank.worldbank.org/source/world-development-indicators

The treatment variable

Create an indicator variable to identify treated cases where:

- 0 = units in your data that were never treated, for example, states that never passed a policy of interest.
- 1 = units that where treated, for example, states that passed a policy of interest.

If, for example, states “abc”, “xyz”, and “cgi” are in the treatment group and in string format, you can create the treated variable as follows:

```
gen treated = (state == "abc" | ///
               state == "xyz" | ///
               state == "cgi")  if !missing(state)
```

The treatment variable

```
* For the example in this document, the treated
countries were saved in a separate fake Stata dataset
containing a variable "treated" = 1. Below we merge
that file to have the treatment variable.

merge m:1 country using
"http://www.princeton.edu/~otorres/Treated.dta",
gen(merge1)

*The untreated units will have a missing value (“.”)

replace treated = 0 if treated == .

*To check, type:

tab country treated
```

The diff-in-diff indicator

```
* The diff-in-diff indicator is an interaction between the
treatment and before/after variables.

* In this example we call the treatment variable "treated"
and the before/after variable "after" (replace with your own
variables as needed).


* Create the diff-in-diff indicator


gen did = after * treated


* Create a labeled numeric variable for the grouping or panel
variable. This is needed for Stata commands to identify the
panels in the data.

encode country, gen(country1)


* Set data as panel data (only for use with 'xt' commands).

xtset country1 year
```

Event happens at the same time for all treated groups

Using Stata's `xtdidregress` / `didregress`

Using Stata's `xtdidregress`

```
* Works only for Stata 17+ (see manual estimation few slides ahead).
* For details and examples on this command type: help xtdidregress

xtdidregress (gdppc) (did), group(country1) time(year)

Number of groups and treatment time

Time variable: year
Control:       did = 0
Treatment:     did = 1
-------------------------------------
             |   Control  Treatment
-------------+-----------------------
Group        |
    country1 |        58         68
-------------+-----------------------
Time         |
     Minimum |      2000       2009
     Maximum |      2000       2009
-------------------------------------

Difference-in-differences regression                 Number of obs = 2,772
Data type: Longitudinal

                            (Std. err. adjusted for 126 clusters in country1)
------------------------------------------------------------------------------
             |               Robust
       gdppc | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATET         |
         did |
   (1 vs 0)  |   1164.492   610.0838     1.91   0.059    -42.93971    2371.923
------------------------------------------------------------------------------
Note: ATET estimate adjusted for panel effects and time effects.
```

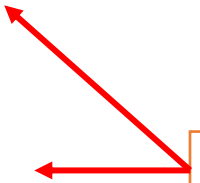
Use `xtdidregress` if panel data.
Use `didregress` if repeated cross-sectional data (i.e. surveys over time)

Not significant at 5%, event did not have a significant effect on GDPpc.

Using Stata's `xtidregress`: parallel trends

```
* For details and example on didregress postestimation commands type
help xtdidregress_postestimation
* Run xtdidregress first
xtdidregress (gdppc) (did), group(country1) time(year)
[OUTPUT OMITTED]
estat ptrends
Parallel-trends test (pretreatment time period)
H0: Linear trends are parallel

F(1, 125) =   3.94
 Prob > F = 0.0495
```



Linear trends are not parallel at 95% level.

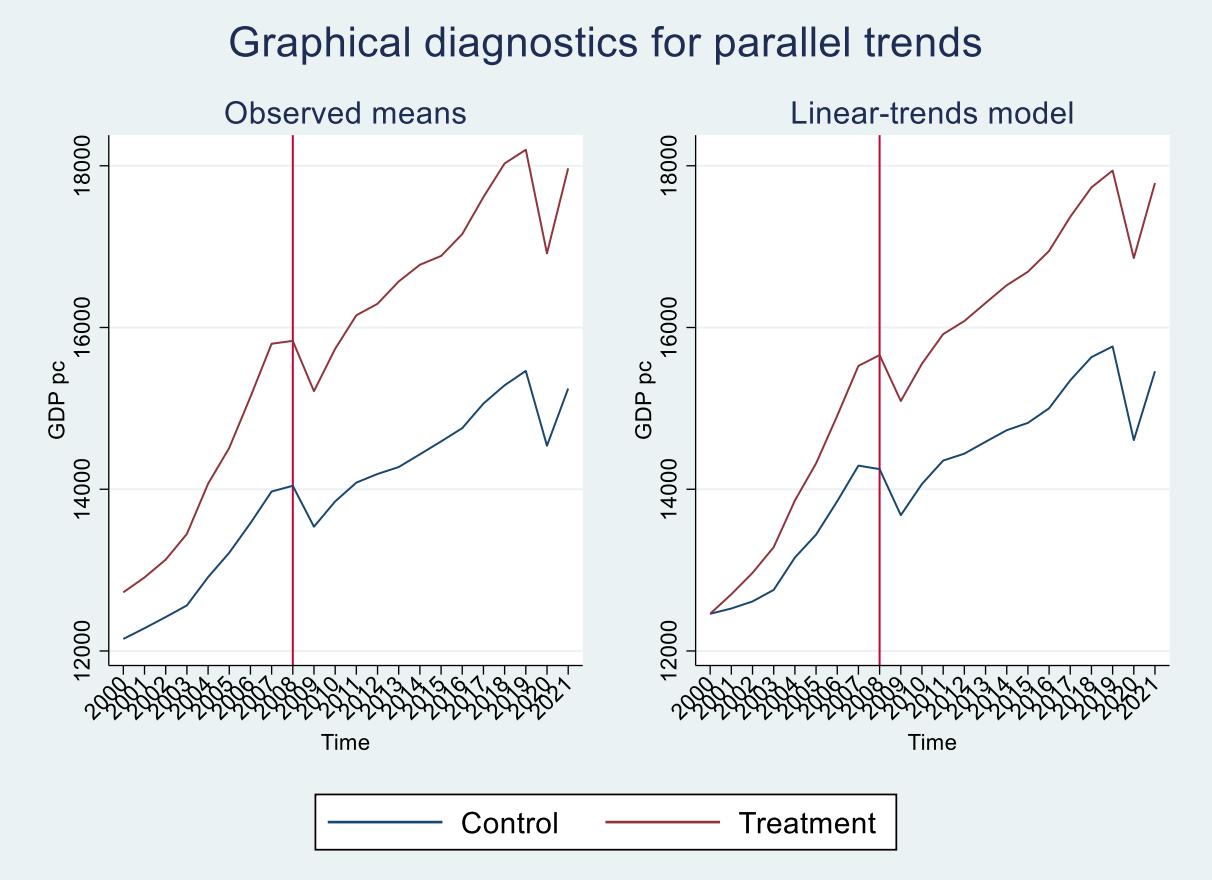
Using Stata's `xtdidregress`: visualization

```
* For details and example on didregress postestimation commands type
help xtdidregress_postestimation

* Run xtdidregress first

xtdidregress (gdppc) (did), group(country1) time(year)

estat trendplots, ytitle(GDP pc)
```



Event happens at the same time for all treated groups

Using OLS fixed effects regression (manual estimation)

Diff-in-diff basic regression: same event for all

```
* Create a labeled numeric variable for the grouping or panel variable.

encode country, gen(country1)

* DID regression (after and treated not needed due to the panel/time fixed effects).

xtreg gdppc did i.year, fe vce(cluster country1)

Fixed-effects (within) regression               Number of obs      =      2,772
Group variable: country1                        Number of groups   =        126

R-squared:                                      Obs per group:
     Within  = 0.2119                                         min =         22
     Between = 0.0023                                         avg =       22.0
     Overall = 0.0063                                         max =         22

                                                F(22,125)          =       7.55
corr(u_i, Xb) = 0.0072                          Prob > F           =     0.0000

                            (Std. err. adjusted for 126 clusters in country1)
------------------------------------------------------------------------------
             |               Robust
       gdppc | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         did |   1164.492   610.0838     1.91   0.059    -42.93971    2371.923
        year |
```

[YEAR FE OUTPUT OMITTED]

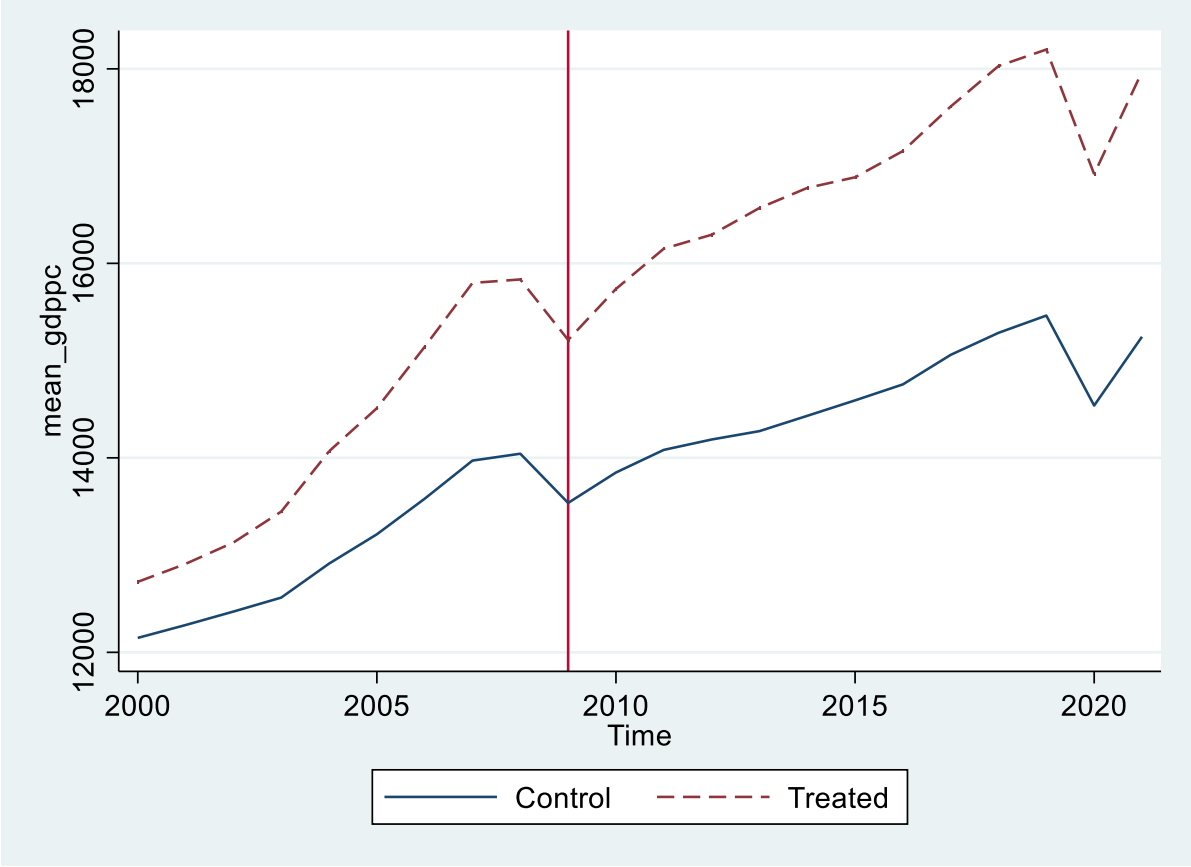
* The coefficient for ‘did’ is the differences-in-differences estimator. The effect is not significant at 95% (P>|t| > 0.05), therefore we conclude that the event did not have a significant effect on the response variable.

Thanks to Mateus Dias for useful feedback

Visualizing parallel trends

```
bysort year treated: egen mean_gdppc = mean(gdppc)

twoway line mean_gdppc year if treated == 0, sort || ///
       line mean_gdppc year if treated == 1, sort lpattern(dash) ///
           legend(label(1 "Control") label(2 "Treated")) ///
           xline(2009)
```



Testing for parallel trends (event happening at the same time)

```
reg gdppc treated##ibn.year if after == 0, vce(cluster country1) hascons
note: 1.treated#2008.year omitted because of collinearity.

Linear regression                               Number of obs     =      1,134
                                                F(18, 125)        =       7.91
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0037
                                                Root MSE          =      18443

                                (Std. err. adjusted for 126 clusters in country1)
------------------------------------------------------------------------------
             |               Robust
       gdppc |  Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
   1.treated |   1791.787   3498.167     0.51   0.609    -5131.519    8715.093
             |
        year |
        2000 |   12148.72   2113.707     5.75   0.000     79
        2001 |   12281.17   2137.624     5.75   0.000     80
        2002 |   12419.36   2158.998     5.75   0.000     81
        2003 |   12563.37   2168.074     5.79   0.000      8
        2004 |   12912.69   2219.854     5.82   0.000     85
        2005 |   13214.44   2257.931     5.85   0.000     87
        2006 |   13579.78   2305.982     5.89   0.000     90
        2007 |   13972.03   2352.561     5.94   0.000     93
        2008 |   14042.81   2344.425     5.99   0.000     94
             |
treated#year |
     1 2000  |  -1215.452   624.5541    -1.95   0.054    -2451.522    20.61814
     1 2001  |  -1163.395   594.2939    -1.96   0.053    -2339.576    12.78646
     1 2002  |  -1081.763   548.8365    -1.97   0.051    -2167.978    4.452972
     1 2003  |   -907.977   494.2322    -1.84   0.069    -1886.124    70.16994
     1 2004  |  -636.9737    350.322    -1.82   0.071    -1330.305    56.35703
     1 2005  |  -495.4092   283.8196    -1.75   0.083    -1057.123    66.30495
     1 2006  |  -229.9688   174.1906    -1.32   0.189    -574.7137    114.7761
     1 2007  |   35.19131   94.47967     0.37   0.710    -151.7957    222.1783
     1 2008  |           0  (omitted)
------------------------------------------------------------------------------
```

No significant difference (at 95%) between treatment and control groups per year, which may suggest parallel trends.

Creating a time-to-event variable

For illustration purposes, no needed when event happened at the same time

Creating time to event (single event)

The following procedure is not needed when testing for a single event. Showing here as FYI.

```
* Generating the time to event variable, assuming
event happened in 2009 for all treatment units.


gen time_to_event2009 = year - 2009 if treated == 1


replace time_to_event2009 = 0 if treated == 0


browse country year time_to_event2009
```

Time to event variable (single event)

```
. tab time_to_event2009

time_to_eve |
     nt2009 |      Freq.     Percent        Cum.
------------+-----------------------------------
         -9 |          68        2.45        2.45
         -8 |          68        2.45        4.91
         -7 |          68        2.45        7.36
         -6 |          68        2.45        9.81
         -5 |          68        2.45       12.27
         -4 |          68        2.45       14.72
         -3 |          68        2.45       17.17
         -2 |          68        2.45       19.62
         -1 |          68        2.45       22.08
          0 |       1,344       48.48       70.56
          1 |          68        2.45       73.02
          2 |          68        2.45       75.47
          3 |          68        2.45       77.92
          4 |          68        2.45       80.38
          5 |          68        2.45       82.83
          6 |          68        2.45       85.28
          7 |          68        2.45       87.73
          8 |          68        2.45       90.19
          9 |          68        2.45       92.64
         10 |          68        2.45       95.09
         11 |          68        2.45       97.55
         12 |          68        2.45      100.00
------------+-----------------------------------
      Total |       2,772      100.00
```

Time to event variable (single event)

```
* Creating dummies for each time_to_event2009

tab  time_to_event2009, gen(z)

* Removing the "time_to_event2009== " part of the label for each
dummy. Each dummy will have the prefix "z" [replace with your own]

sum time_to_event2009
local min = r(min)
local i = `min'
foreach var of varlist z1-z22 {
       label variable `var' "`i'"
       local i = `i'+1
}
```

z1	time_to_event2009==	-9.0000
z2	time_to_event2009==	-8.0000
z3	time_to_event2009==	-7.0000
z4	time_to_event2009==	-6.0000
z5	time_to_event2009==	-5.0000
z6	time_to_event2009==	-4.0000
z7	time_to_event2009==	-3.0000
z8	time_to_event2009==	-2.0000
z9	time_to_event2009==	-1.0000
z10	time_to_event2009==	0.0000
z11	time_to_event2009==	1.0000
z12	time_to_event2009==	2.0000
z13	time_to_event2009==	3.0000

z1	-9
z2	-8
z3	-7
z4	-6
z5	-5
z6	-4
z7	-3
z8	-2
z9	-1
z10	0
z11	1
z12	2
z13	3

Event is staggered across groups

Dynamic differences-in-differences

When the event happens…

Need a variable indicating the timing of the event. For example, if the event happened in country A in year 3, in country B in year 6, and in country C in year 9, then we create a variable called here 'eventX' (you can use any name you like):

```
gen eventX = .
replace eventX = 3 if country == "A"
replace eventX = 6 if country == "B"
replace eventX = 9 if country == "C"
```

*** For the example in this document, we saved the year the event happened for a random selection of countries in a separate data file (a fake dataset in this case).**

```
merge m:1 country using
"http://www.princeton.edu/~otorres/eventX.dta", gen(merge2)

order country year eventX
```

The time to event variable

*** Generating the time to event variable. In this example we have years, replace with your own time variable (i.e. months, quarters, etc.).**

```
gen time_to_event = year - eventX
```

[See next slide to check the variable]

Time to event variable

tab time_to_event

time_to_event	Freq.	Percent	Cum.
-15	6	0.40	0.40
-14	14	0.94	1.34
-13	22	1.47	2.81
-12	24	1.60	4.41
-11	35	2.34	6.75
-10	39	2.61	9.36
-9	41	2.74	12.10
-8	51	3.41	15.51
-7	55	3.68	19.18
-6	61	4.08	23.26
-5	68	4.55	27.81
-4	68	4.55	32.35
-3	68	4.55	36.90
-2	68	4.55	41.44
-1	68	4.55	45.99
0	68	4.55	50.53
1	68	4.55	55.08
2	68	4.55	59.63
3	68	4.55	64.17
4	68	4.55	68.72
5	68	4.55	73.26
6	68	4.55	77.81
7	62	4.14	81.95
8	54	3.61	85.56
9	46	3.07	88.64
10	44	2.94	91.58
11	33	2.21	93.78
12	29	1.94	95.72
13	27	1.80	97.53
14	17	1.14	98.66
15	13	0.87	99.53
16	7	0.47	100.00
Total	1,496	100.00	

***J* lags**

***K* leads**

NOTE: This only includes units where the event happened.

All treatment units experienced the event in this range.

Event is staggered across groups

Using `eventdd` command.

Source: https://docs.iza.org/dp13524.pdf

Using eventdd for staggered events

```
* See https://docs.iza.org/dp13524.pdf. Install the following:

ssc install eventdd
ssc install matsort
ssc install reghdfe
ssc install ftools

* Accumulating the periods beyond the specified leads/lags, J = -5, K = 6.

eventdd gdppc, hdfe absorb(country1) vce(cluster country1) timevar(time_to_event)
graph_op(xlabel(-5(1)6, labsize(3))) ci(rarea, color(gs14%33)) leads(5) lags(6) accum


HDFE Linear regression                            Number of obs    =      2,772
Absorbing 1 HDFE group                            F(  11,    125) =       5.09
Statistics robust to heteroskedasticity           Prob > F         =     0.0000
                                                  R-squared        =     0.9801
                                                  Adj R-squared    =     0.9791
                                                  Within R-sq.     =     0.1486
Number of clusters (country1) =        126        Root MSE         =  2830.3445

                            (Std. err. adjusted for 126 clusters in country1)
------------------------------------------------------------------------------
             |               Robust
       gdppc |  Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       lead5 |  -1788.375   408.8666    -4.37   0.000    -2597.573   -979.1776
       lead4 |  -567.0806   209.0783    -2.71   0.008    -980.8725   -153.2886
       lead3 |  -263.1147   165.3622    -1.59   0.114    -590.3869    64.15759
       lead2 |  -184.2048   90.43976    -2.04   0.044    -363.1963   -5.213244
        lag0 |   406.7118   203.5858     2.00   0.048     3.790269    809.6334
        lag1 |   960.3596   459.2116     2.09   0.039     51.52283    1869.196
        lag2 |   1357.789   559.8445     2.43   0.017     249.7869    2465.791
        lag3 |   1639.607   676.8084     2.42   0.017     300.1196    2979.095
        lag4 |   1809.949   657.0901     2.75   0.007      509.486    3110.412
        lag5 |   1868.826   509.0578     3.67   0.000     861.3375    2876.315
        lag6 |   2456.058   547.1238     4.49   0.000     1373.232    3538.884
       _cons |   14667.33   117.7002   124.62   0.000     14434.39    14900.27
------------------------------------------------------------------------------

Absorbed degrees of freedom:
-----------------------------------------------------+
 Absorbed FE | Categories  - Redundant  = Num. Coefs |
-------------+---------------------------------------|
    country1 |       126         126           0    *|
-----------------------------------------------------+
* = FE nested within cluster; treated as redundant for DoF computation
```

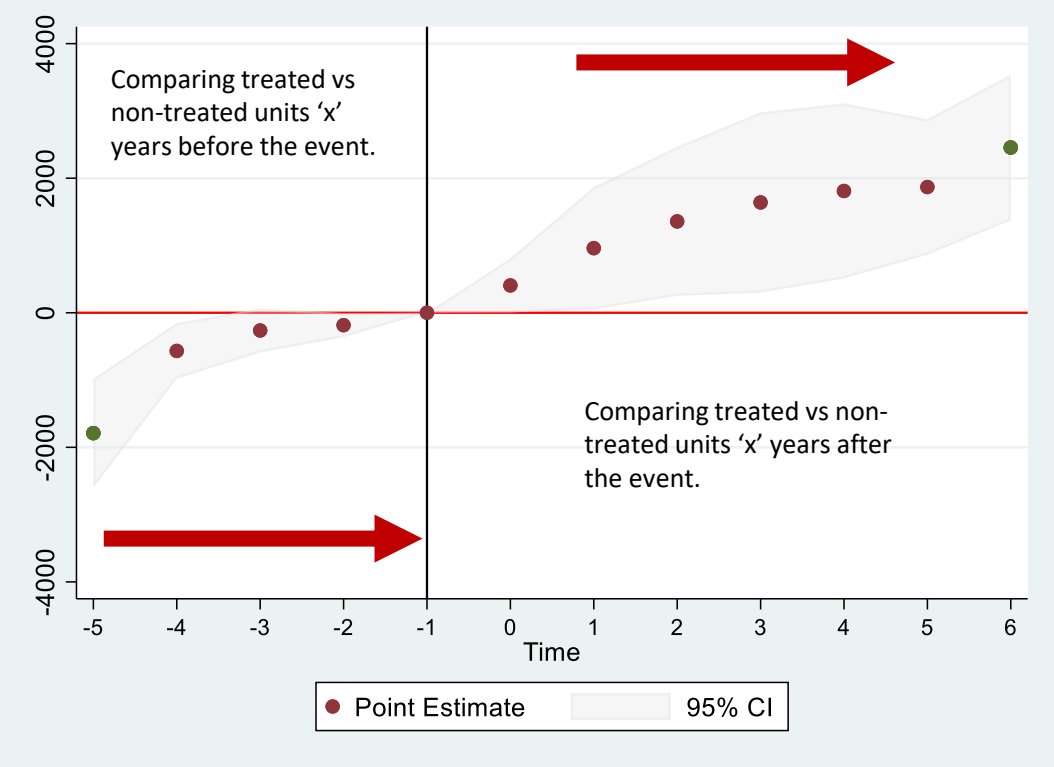
It will run the model only for the times where all units were treated, '`accum`' option

See graph next slide

Using eventdd for staggered events

```
* See https://docs.iza.org/dp13524.pdf
```

Each dot is the coefficient for the corresponding leads/lags. The shaded area show the 95% confidence intervals of the coefficients. The coefficients are significant as long as the shaded area does not cross the horizontal red line. Countries where the event never happened will served as controls (same for pre-time in the treatment). The country fixed effects will account for any unobserved heterogeneity across countries.



Event is staggered across groups

Manual estimation using OLS procedure

When the event happens…

Need a variable indicating the timing of the event. For example, if the event happened in country A in year 3, in country B in year 6, and in country C in year 9, then we create a variable called here 'eventX' (you can use any name you like):

```
gen eventX = .
replace eventX = 3 if country == "A"
replace eventX = 6 if country == "B"
replace eventX = 9 if country == "C"
```

*** For the example in this document, we saved the year the event happened for a random selection of countries in a separate data file (a fake dataset in this case).**

```
merge m:1 country using
"http://www.princeton.edu/~otorres/eventX.dta", gen(merge2)

order country year eventX
```

The time to event variable

```
* Generating the time to event variable. In this example we have years, replace with your own time variable (i.e. months, quarters, etc.).

gen time_to_event = year - eventX
```

[See next slide to check the variable]

Time to event variable

tab time_to_event

time_to_event	Freq.	Percent	Cum.
-15	6	0.40	0.40
-14	14	0.94	1.34
-13	22	1.47	2.81
-12	24	1.60	4.41
-11	35	2.34	6.75
-10	39	2.61	9.36
-9	41	2.74	12.10
-8	51	3.41	15.51
-7	55	3.68	19.18
-6	61	4.08	23.26
-5	68	4.55	27.81
-4	68	4.55	32.35
-3	68	4.55	36.90
-2	68	4.55	41.44
-1	68	4.55	45.99
0	68	4.55	50.53
1	68	4.55	55.08
2	68	4.55	59.63
3	68	4.55	64.17
4	68	4.55	68.72
5	68	4.55	73.26
6	68	4.55	77.81
7	62	4.14	81.95
8	54	3.61	85.56
9	46	3.07	88.64
10	44	2.94	91.58
11	33	2.21	93.78
12	29	1.94	95.72
13	27	1.80	97.53
14	17	1.14	98.66
15	13	0.87	99.53
16	7	0.47	100.00
Total	1,496	100.00	

***J* lags**

***K* leads**

NOTE: This only includes units where the event happened.

All treatment units experienced the event in this range.

Modified time to event variable

```
clonevar time_to_event_accum = time_to_event

replace time_to_event_accum = -5 if time_to_event_accum < -5 & ///
                                    time_to_event_accum !=.

replace time_to_event_accum = 6 if time_to_event_accum > 6 & ///
                                   time_to_event_accum !=.

tab time_to_event_accum
```

time_to_eve nt_accum	Freq.	Percent	Cum.
-5	416	27.81	27.81
-4	68	4.55	32.35
-3	68	4.55	36.90
-2	68	4.55	41.44
-1	68	4.55	45.99
0	68	4.55	50.53
1	68	4.55	55.08
2	68	4.55	59.63
3	68	4.55	64.17
4	68	4.55	68.72
5	68	4.55	73.26
6	400	26.74	100.00
Total	1,496	100.00	

All treatment units experienced the event at J = -5 and K = 6

Time to event indicators (modified variable)

*** Creating dummies for each time_to_event, each dummy will have the prefix "x" [replace with your own]**

```
tab  time_to_event_accum, gen(x)
```

*** Removing the "time_to_event_accum== " part of the label for each dummy.**

```
sum time_to_event_accum
local min = r(min)
local i = `min'
foreach var of varlist x1-x12 {
        label variable `var' "`i'"
        local i = `i'+1
}
```

Name	Label
x1	time_to_event_accum== -5.0000
x2	time_to_event_accum== -4.0000
x3	time_to_event_accum== -3.0000
x4	time_to_event_accum== -2.0000
x5	time_to_event_accum== -1.0000
x6	time_to_event_accum== 0.0000
x7	time_to_event_accum== 1.0000
x8	time_to_event_accum== 2.0000
x9	time_to_event_accum== 3.0000
x10	time_to_event_accum== 4.0000
x11	time_to_event_accum== 5.0000
x12	time_to_event_accum== 6.0000

Name	Label
x1	-5
x2	-4
x3	-3
x4	-2
x5	-1
x6	0
x7	1
x8	2
x9	3
x10	4
x11	5
x12	6

Event staggered across groups

```
* Using reghdfe, need to install:

ssc install reghdfe
ssc install ftools

* Event diff-in-diff regression where x5 is the reference(year before the event happened in each country)

reghdfe gdppc x1-x4 x6-x12, absorb(country1) vce(cluster country1)

(MWFE estimator converged in 1 iterations)

HDFE Linear regression                            Number of obs   =      1,496
Absorbing 1 HDFE group                            F(  11,     67) =       5.04
Statistics robust to heteroskedasticity           Prob > F        =     0.0000
                                                  R-squared       =     0.9739
                                                  Adj R-squared   =     0.9724
                                                  Within R-sq.    =     0.1738
Number of clusters (country1) =         68        Root MSE        =  3515.6856

                              (Std. err. adjusted for 68 clusters in country1)
------------------------------------------------------------------------------
             |               Robust
       gdppc |  Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
          x1 |  -1788.375   410.9684    -4.35   0.000    -2608.672   -968.0792
          x2 |  -567.0806   210.1531    -2.70   0.009    -986.5478   -147.6133
          x3 |  -263.1147   166.2122    -1.58   0.118    -594.8756    68.64625
          x4 |  -184.2048   90.90466    -2.03   0.047    -365.6512   -2.758306
          x6 |   406.7118   204.6323     1.99   0.051    -1.735955    815.1596
          x7 |   960.3596   461.5721     2.08   0.041     39.05779    1881.661
          x8 |   1357.789   562.7223     2.41   0.019     234.5902    2480.987
          x9 |   1639.607   680.2874     2.41   0.019      281.748    2997.467
         x10 |   1809.949   660.4678     2.74   0.008     491.6496    3128.248
         x11 |   1868.826   511.6746     3.65   0.001     847.5194    2890.133
         x12 |   2456.058   549.9362     4.47   0.000     1358.381    3553.735
       _cons |   15297.11   219.2126    69.78   0.000     14859.56    15734.66
------------------------------------------------------------------------------

Absorbed degrees of freedom:
-----------------------------------------------------+
 Absorbed FE | Categories  - Redundant  = Num. Coefs |
-------------+---------------------------------------|
    country1 |        68          68           0    *|
-----------------------------------------------------+
* = FE nested within cluster; treated as redundant for DoF computation
```

Visualizing the time to event coefficients

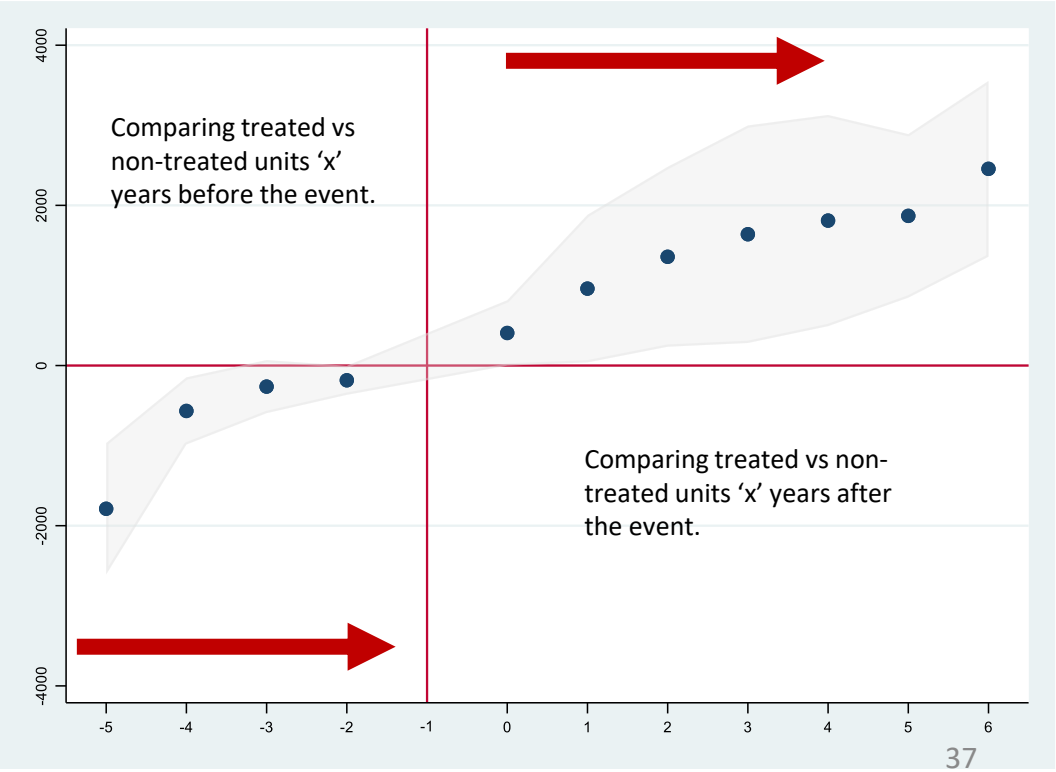
```
* Install user-written command –coefplot-
* See http://repec.sowi.unibe.ch/stata/coefplot/getting-started.html

ssc install coefplot

reghdfe gdppc x1-x4 x6-x12, absorb(country1) vce(cluster country1)

coefplot, keep(x*) order(x1 x2 x3 x4 . x6 x7 x8 x9 x10 x11 x12) vertical
drop(_cons) yline(0) xline(5) xlabel(, labsize(2)) ylabel(, labsize(2))
ciopts(recast(rarea) color(gs14%33)) ttext(-4500 5 "-1", size(2))
```

Each dot is the coefficient for the corresponding dummy.
The shaded area shows 95% confidence intervals of the coefficients.
The coefficients are significant as long as the shaded area does not cross the horizontal red line.
Countries where the event never happened will served as controls (same for pre-time in the treatment).
The country fixed effects will account for any unobserved heterogeneity across countries.



Additional references

- *Introduction to econometrics*, James H. Stock, Mark W. Watson. 2nd ed., Boston: Pearson Addison Wesley, 2007.

- "Difference-in-Differences Estimation", Imbens/Wooldridge, Lecture, Notes 10, summer 2007. http://www.nber.org/WNE/lect_10_diffindiffs.pdf

- "Lecture 3: Differences-in-Differences", Fabian Waldinger, https://www.fabianwaldinger.com/_files/ugd/0d0a02_6fef951d28064c8db2cf06d6dfa0cff6.pdf