Panel Data Analysis

Lecture 1: From Randomized Controlled Trials to Diff-in-Diff

28 April, 2015

Prof. Andrew Eggers

What are we talking about?

Generally, we're talking about

- **causal inference** (cf descriptive, predictive analysis)

  => we focus on a single treatment that varies across units

- for **grouped data**, e.g.
  - multiple classrooms, each with many students
  - multiple judges, each deciding many cases
  - multiple countries, each with several years of data (or, multiple years, each with multiple countries)

  => counterfactuals can be drawn from comparison with same group ("within" or "fixed-effect" estimator), comparison across groups ("between" estimator), both ("random effects")

  => challenges with inference: basically, clustered sampling

Goals

Focus on **intuition & connections** among research designs.

- What analysis to run in your own research
- What results really mean
- What questions to ask about other people's research
- How to answer questions about research design through simulation

Not:

- A set of commands to run
- A set of rules to follow
- A set of formulas to memorize

Applying what we learn

What dataset and research question have you brought?

- What is the structure of dataset? What are the groupings?
- What is the main independent variable of interest (i.e. treatment)? What values does it take?
- What is your question? Why is it important and interesting?

The first diff-in-diff?

Source: John Snow (1855), *On the communication of cholera*

In 1852, the Lambeth Company changed the source of its water from Hungerford Bridge to Thames Ditton.

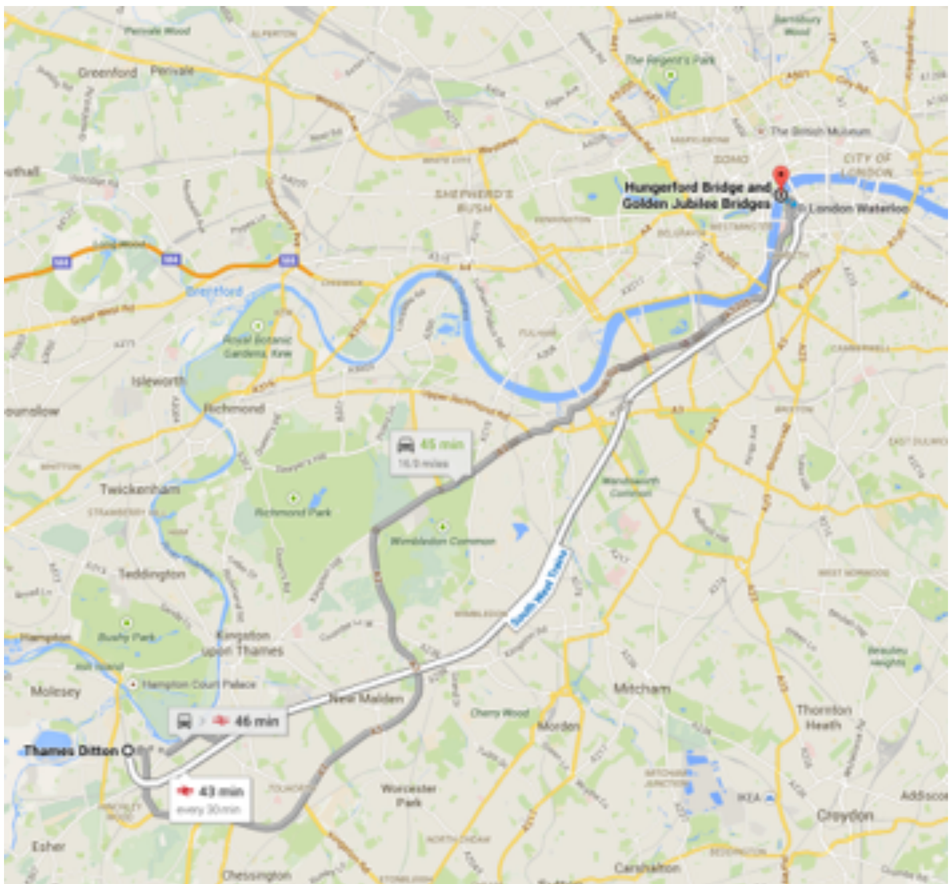


TABLE XII.

Sub-Districts.	Deaths from Cholera in 1849.	Deaths from Cholera in 1854.	Water Supply.
St. Saviour, Southwark	283	371	Southwark & Vauxhall Company only.
St. Olave	157	161	
St. John, Horsleydown	192	148	
St. James, Bermondsey	249	362	
St. Mary Magdalen	259	244	
Leather Market	226	237	
Rotherhithe*	352	282	
Wandsworth	97	59	
Battersea	111	171	
Putney	8	9	
Camberwell	235	240	
Peckham	92	174	
Christchurch, Southwark	256	113	Lambeth Company, and Southwark and Vauxhall Compy.
Kent Road	267	174	
Borough Road	312	270	
London Road	257	93	
Trinity, Newington	318	210	
St. Peter, Walworth	446	388	
St. Mary, Newington	143	92	
Waterloo Road (1st)	193	58	
Waterloo Road (2nd)	243	117	
Lambeth Church (1st)	215	49	
Lambeth Church (2nd)	544	193	
Kennington (1st)	187	303	
Kennington (2nd)	153	142	
Brixton	81	48	
Clapham	114	165	
St. George, Camberwell	176	132	
Norwood	2	10	Lambeth Company only.
Streatham	154	15	
Dulwich	1	—	
Sydenham	5	12	
First 12 sub-districts	2261	2458	Southwk. & Vauxhall.
Next 16 sub-districts	3905	2547	Both Companies.
Last 4 sub-districts	162	37	Lambeth Company.

Starting point: randomized experiment

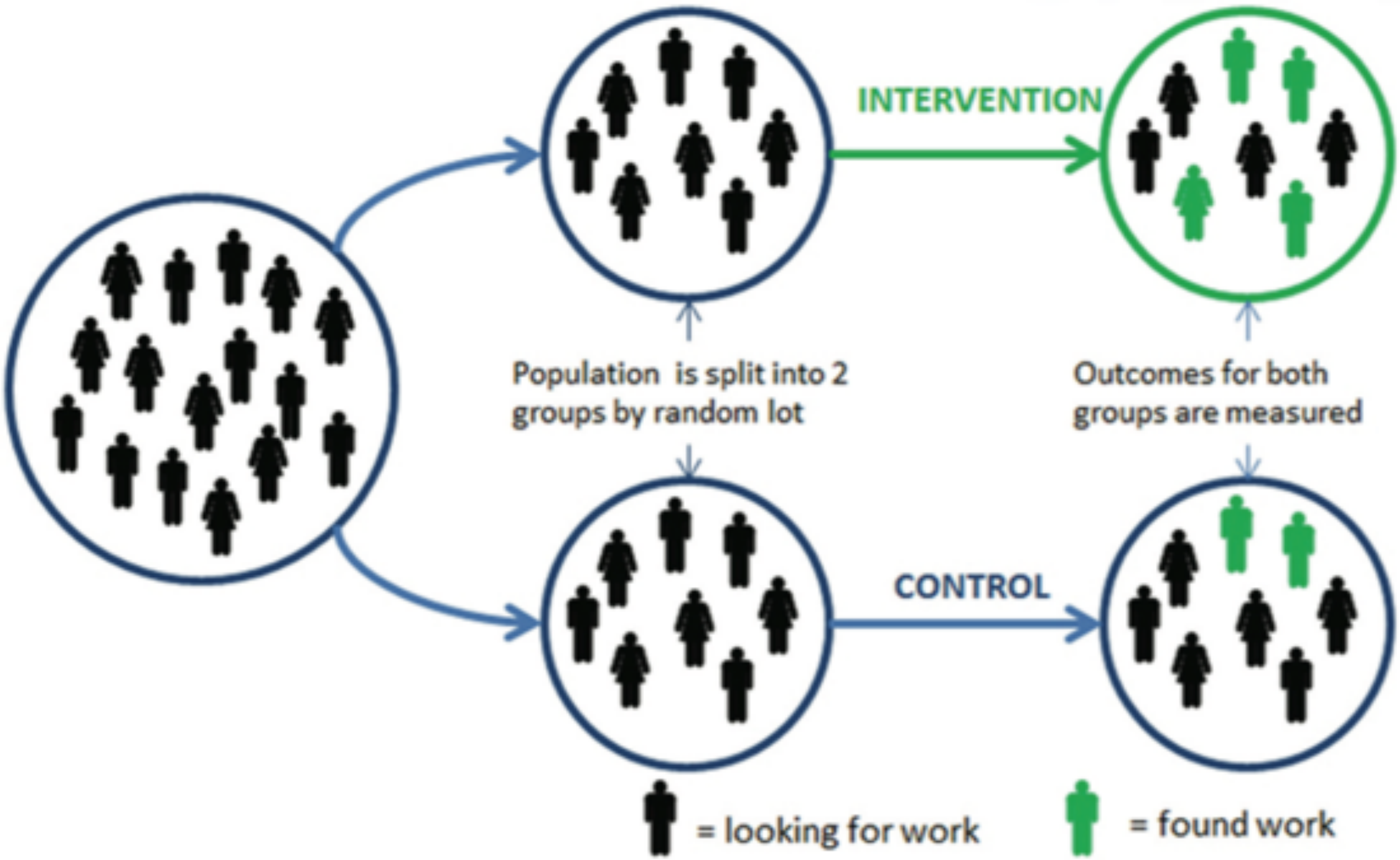


Figure 1. The basic design of a randomised controlled trial (RCT), illustrated with a test of a new 'back to work' programme.

Behavioural Insights Team, "Test, Learn, Adapt"

Formalizing via potential outcomes framework

For unit *i* (e.g. a country), outcome y_i (e.g. trade), and treatment d_i (e.g. membership in WTO), consider two **potential outcomes:**

y_{1i}: the amount of trade in country *i* if country *i* were a member of the WTO

y_{0i}: the amount of trade in country *i* if country *i* were **not** a member of the WTO

Alternative notation: $y_i(1)$, $y_i(0)$

Effect of treatment for unit i: $y_{1i} - y_{0i}$

Fundamental problem of causal inference (Holland 1986): we never observe both potential outcomes for any single unit => necessary to make assumptions and infer effects from comparisons across units.

Causal inference as a missing data problem

What we want:

Country	y_{0i}	y_{1i}	Effect
A	$1 billion	$1.2 billion	$.2 billion

What we have:

Country	y_{0i}	y_{1i}	Effect
A	$1 billion	?	?
B	?	$0.5 billion	?
C	?	$8 billion	?
D	$3 billion	?	?
E	?	$3.5 billion	?

What about simply comparing treated and untreated units?

Given a sample, we can always calculate

$$E[y_{1i}|d_i=1] - E[y_{0i}|d_i=0]$$

Under what assumptions will this tell us what we want to know?

If we want to report the difference in trade between WTO members and non-members, no further assumptions needed.

If we want to report the effect of WTO membership on trade for current members, i.e. "average treatment effect for the treated"

$$ATT = E[y_{1i}|d_i=1] - E[y_{0i}|d_i=1],$$

← Unobserved potential outcome

then we need an additional assumption, which we can state 3 ways:

$$E[y_{0i}|d_i=1] = E[y_{0i}|d_i=0]$$

$$E[y_{0i}|d_i] = E[y_{0i}]$$

$$y_{0i} \perp d_i$$

} "Unconfoundedness" or "ignorability"

The advantages of experiments

Consider the unconfoundedness assumption:

$$E[y_{0i}|d_i=1] = E[y_{0i}|d_i=0] \quad \text{i.e. } y_{0i} \perp d_i$$

- i.e., "control group offers valid counterfactual for treatment group"
- i.e., "countries that are **not** members of the WTO tell us what trade would be like on average in countries that **are** members of the WTO if those countries **were not** in the WTO"

When will unconfoundedness hold?

One case: when treatment (WTO membership) is randomly assigned.

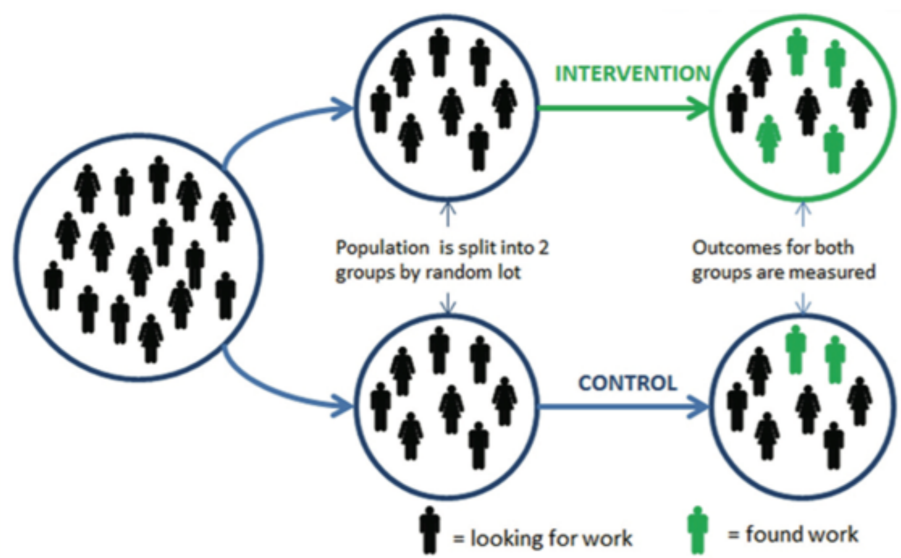


Figure 1. The basic design of a randomised controlled trial (RCT), illustrated with a test of a new 'back to work' programme.

Simulation 1: random assignment

Recipe:

(1) Generate both potential outcomes for a set of units according to

$$x_i \sim N(0,1)$$
$$y_{0i} \sim N(x_i, 1)$$
$$y_{1i} \sim N(x_i + \boldsymbol{\tau}, 1)$$
$$\boldsymbol{\tau} = 1$$

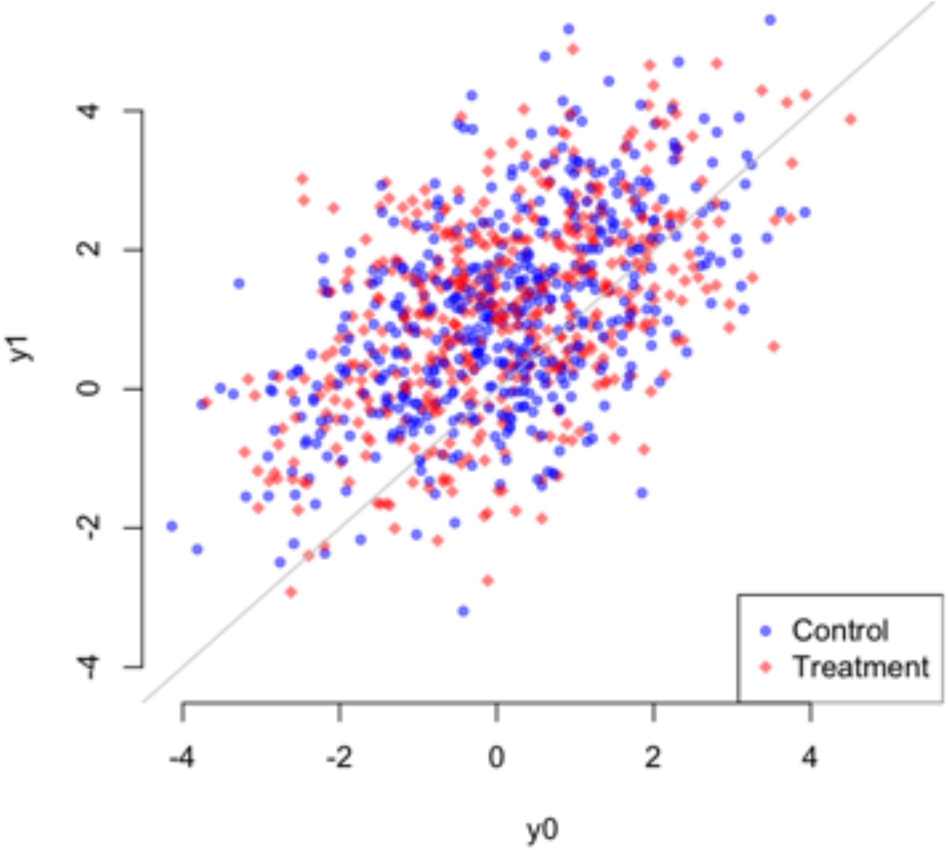
(2) Assign treatment (d) randomly

(3) Estimate ATT (effect of d on y) by

(3a) **Difference-in-means**: average difference in observed y between treated and control units

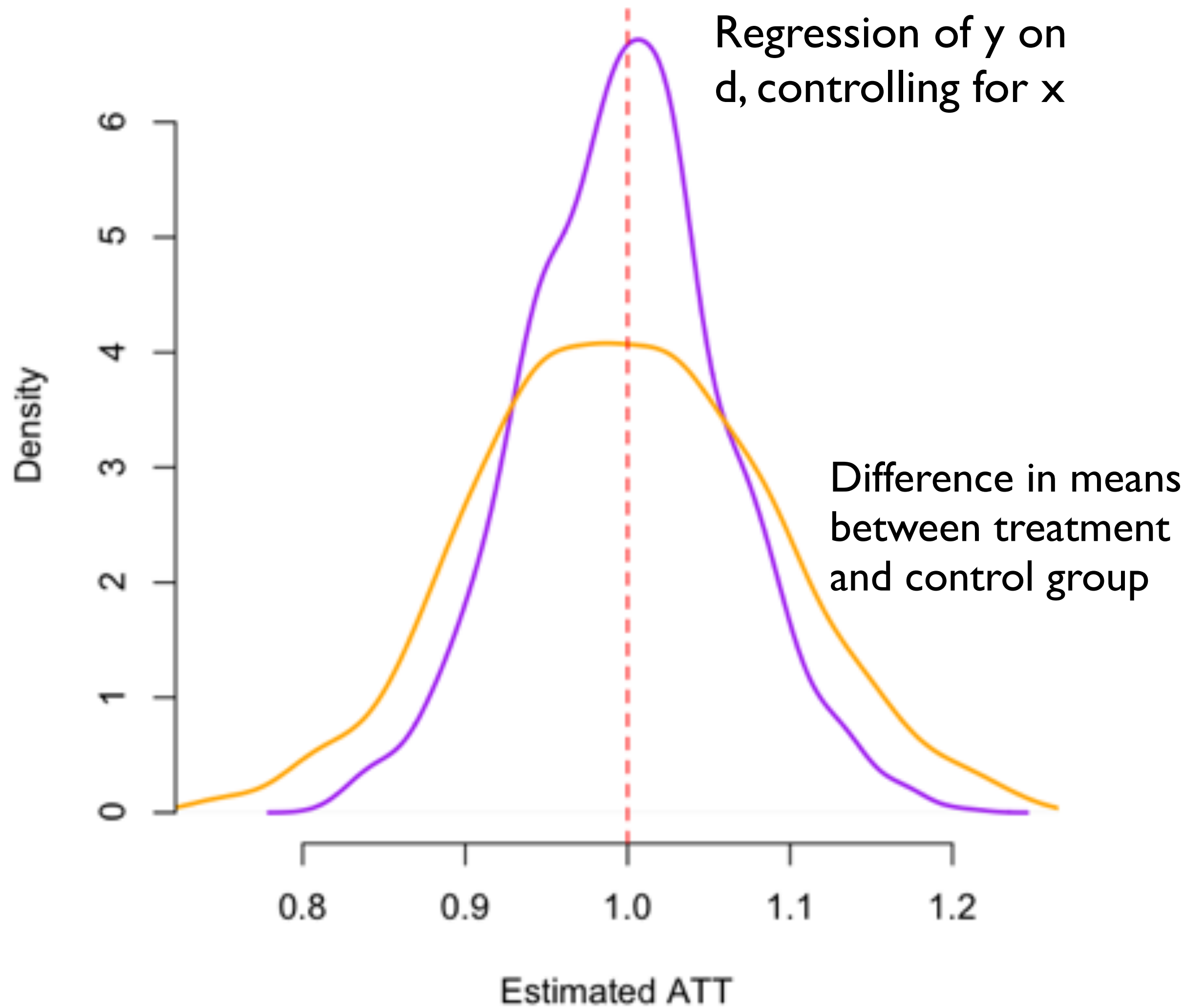
(3b) Regression of observed y on x and d

(4) Repeat from step 1



Is the unconfoundedness assumption met in this case?

Simulation 1 (random assignment): distribution of estimates across replications



Simulation 2: non-random assignment

Recipe:

(1) Generate both potential outcomes as in Simulation 1:

$$x_i \sim N(0,1)$$
$$y_{0i} \sim N(x_i,1)$$
$$y_{1i} \sim N(x_i + \boldsymbol{\tau}, 1)$$
$$\boldsymbol{\tau}=1$$

(2)* Assign treatment according to

$$\Pr(d_i=1) = 1/(1 + \exp(-x_i))$$

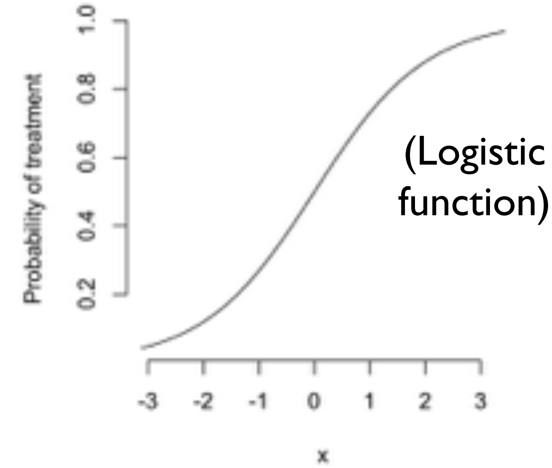
(3) Estimate ATT (effect of d on y) as in Simulation 1:

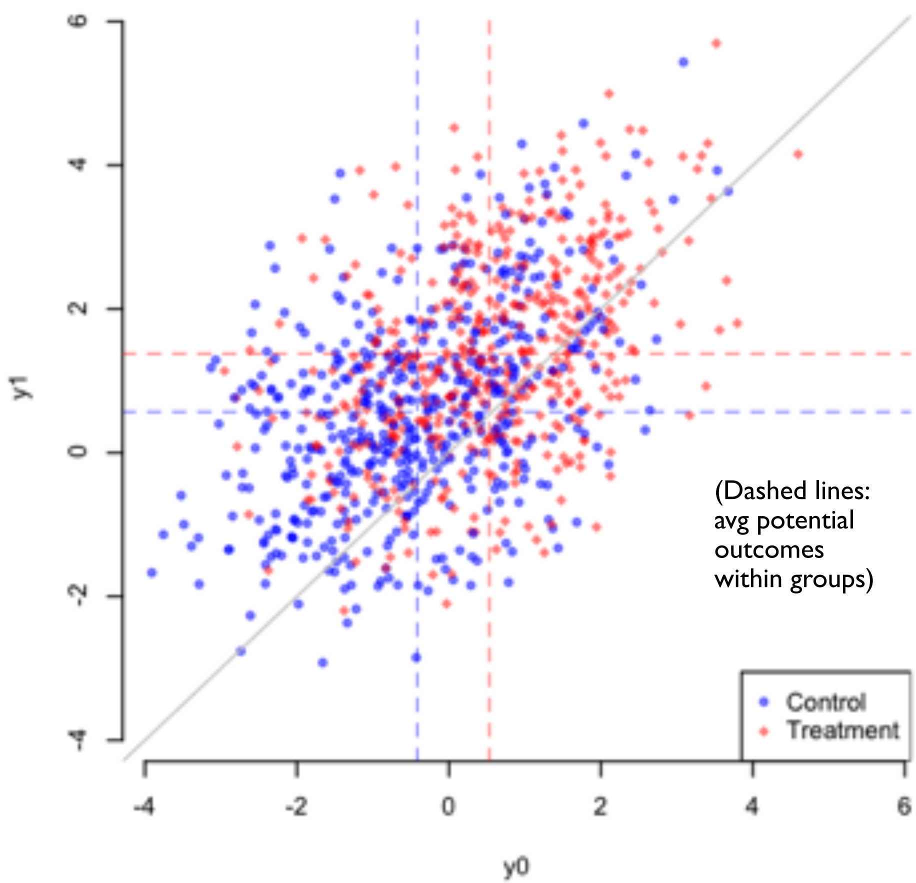
(3a) **Difference-in-means**: average difference in observed y between treated and control units

(3b) Regression of observed y on x and d

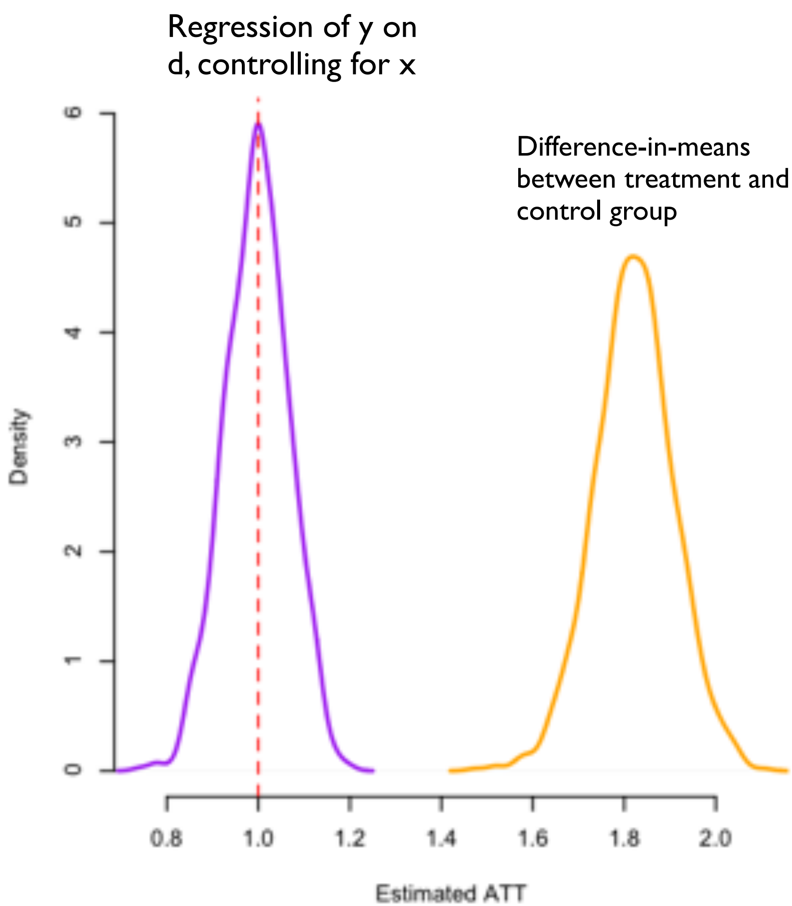
(4) Repeat from step 1

Is the unconfoundedness assumption met in this case?





Simulation 2 (non-random assignment): distribution of estimates across replications

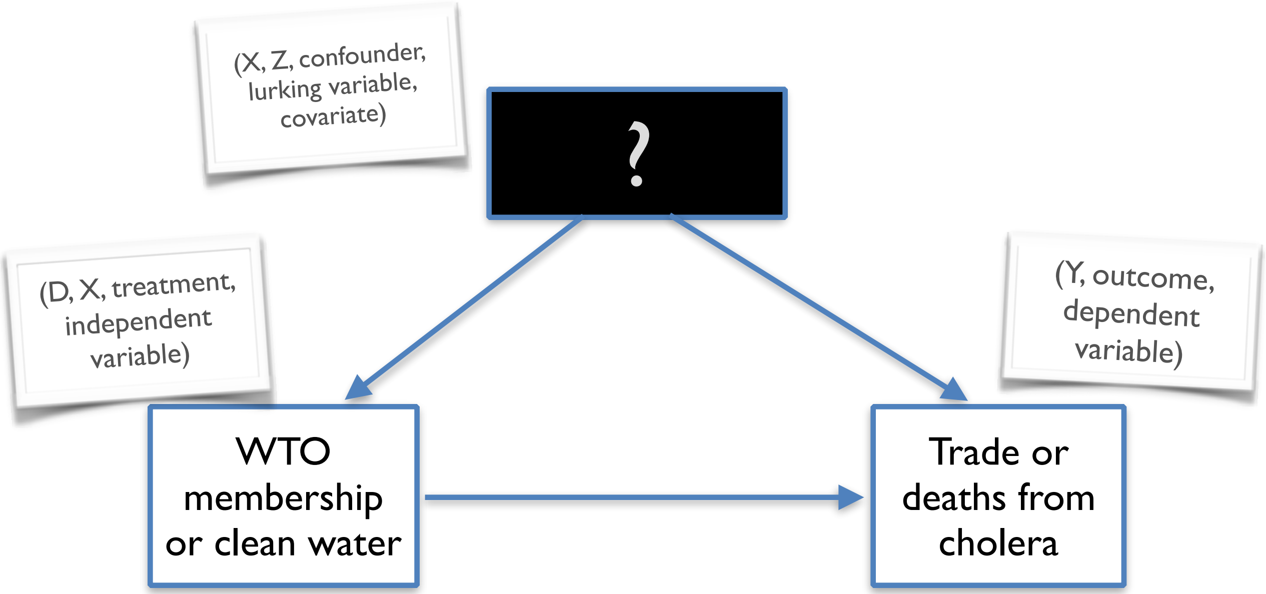


Difference-in-means now produces biased results. Why?

We call x a **covariate** or **confounder**.

What are some possible confounders in

- the WTO example?
- the cholera example?



What about when we don't observe an important covariate/confounder? (From here we assume x not observed.)

(What covariates are likely to be unobserved in the WTO example? the cholera example?)

Our options:

- run an experiment (when you can)
- instrumental variables (when there is an instrument)
- diff-in-diff and other panel methods (when confounding variables are time-invariant)

Other options: change your question and/or setting (e.g. RDD); sensitivity analysis/bounds

Simulation 3: random assignment with baseline (pre-treatment) outcomes

Recipe:

(1)* Same data generating process (DGP) as above, but adding a baseline outcome and time trend:

$$x_i \sim N(0, 1)$$
$$y_{i,pre} \sim N(x_i, 1)$$
$$y_{0i,post} \sim N(x_i + \lambda, 1)$$
$$y_{1i,post} \sim N(x_i + \lambda + \boldsymbol{\tau}, 1)$$
$$\boldsymbol{\tau}=1$$
$$\lambda=0.5$$

(2) Assign treatment randomly (as in Simulation 1)

(3)* Four ways of estimating ATT:

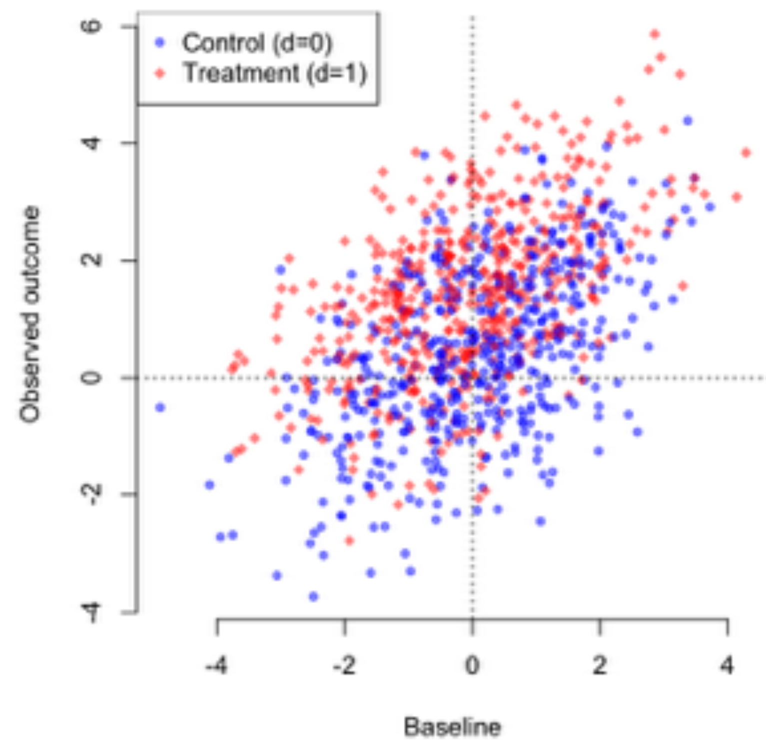
(3a) **Difference-in-means**: average difference in observed y between treated and control units

(3b) **Regression** of observed y on baseline outcome ($y_{i,pre}$) and d

(3c)* **Before-and-after**: average change over time ($E[y_{i,post} - y_{i,pre}]$) in treatment group

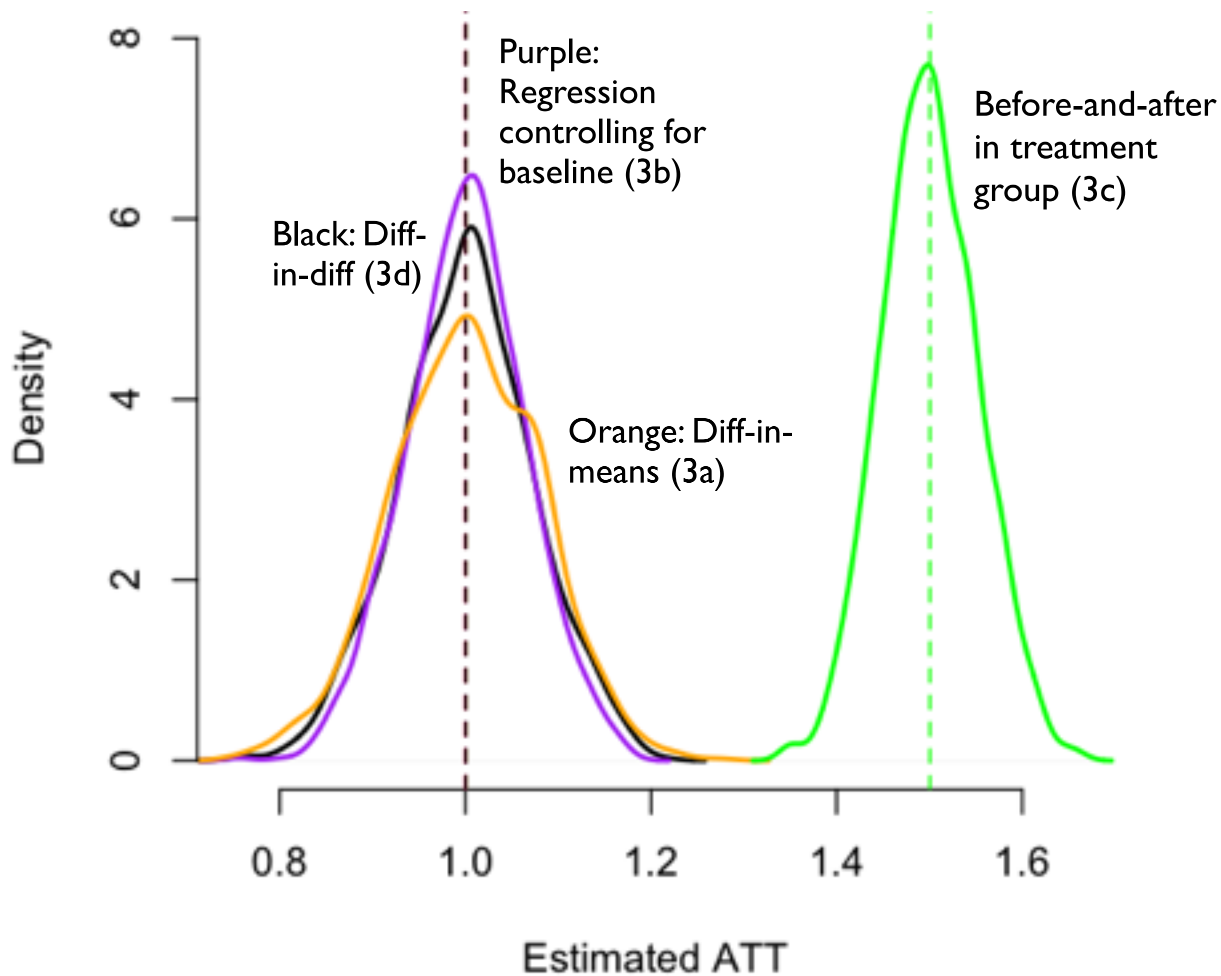
(3d)* **Diff-in-diff:** Difference in before-and-after between treated and control units

(4) Repeat from step 1



Simulation 3 (random assignment with baseline outcomes): distribution of estimates

Do these results make sense?



Simulation 4: non-random assignment with pre-treatment outcomes

Recipe:

(1) Same data-generating process (DGP) as Simulation 3:

$$x_i \sim N(0,1)$$
$$y_{i,pre} \sim N(x_i,1)$$
$$y_{0i,post} \sim N(x_i + \lambda,1)$$
$$y_{1i,post} \sim N(x_i + \lambda + \boldsymbol{\tau},1)$$
$$\boldsymbol{\tau} = 1$$
$$\lambda = 0.5$$

(2)* Assign treatment as in Simulation 2:

$$\Pr(d_i=1) = 1/(1 + \exp(-x_i))$$

(3) Same four ways of estimating ATT as in Simulation 3:

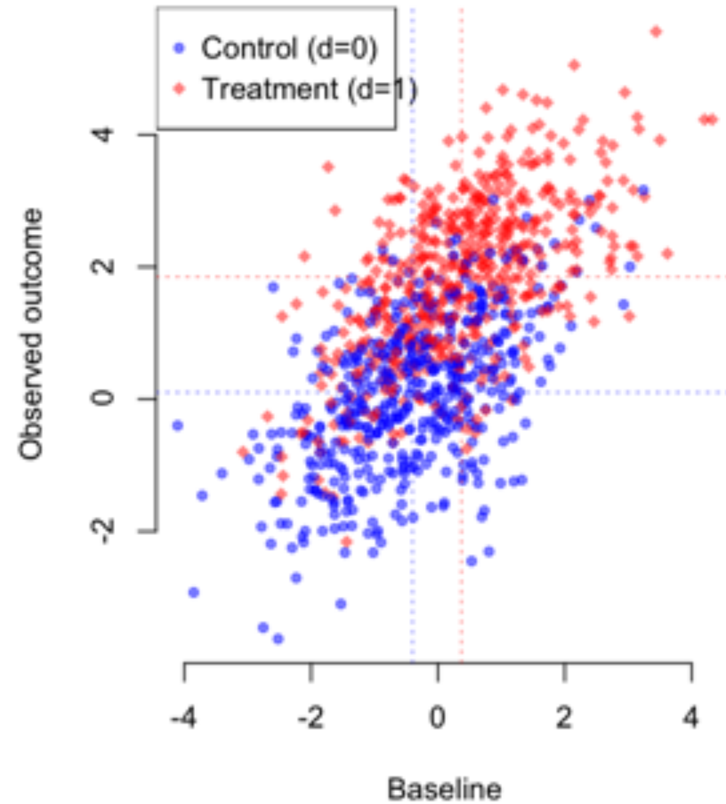
(3a) **Difference-in-means**: average difference in observed y between treated and control units

(3b) **Regression** of observed y on baseline outcome ($y_{i,pre}$) and d

(3c) **Before-and-after**: average change over time ($E[y_{i,post} - y_{i,pre}]$) in treatment group

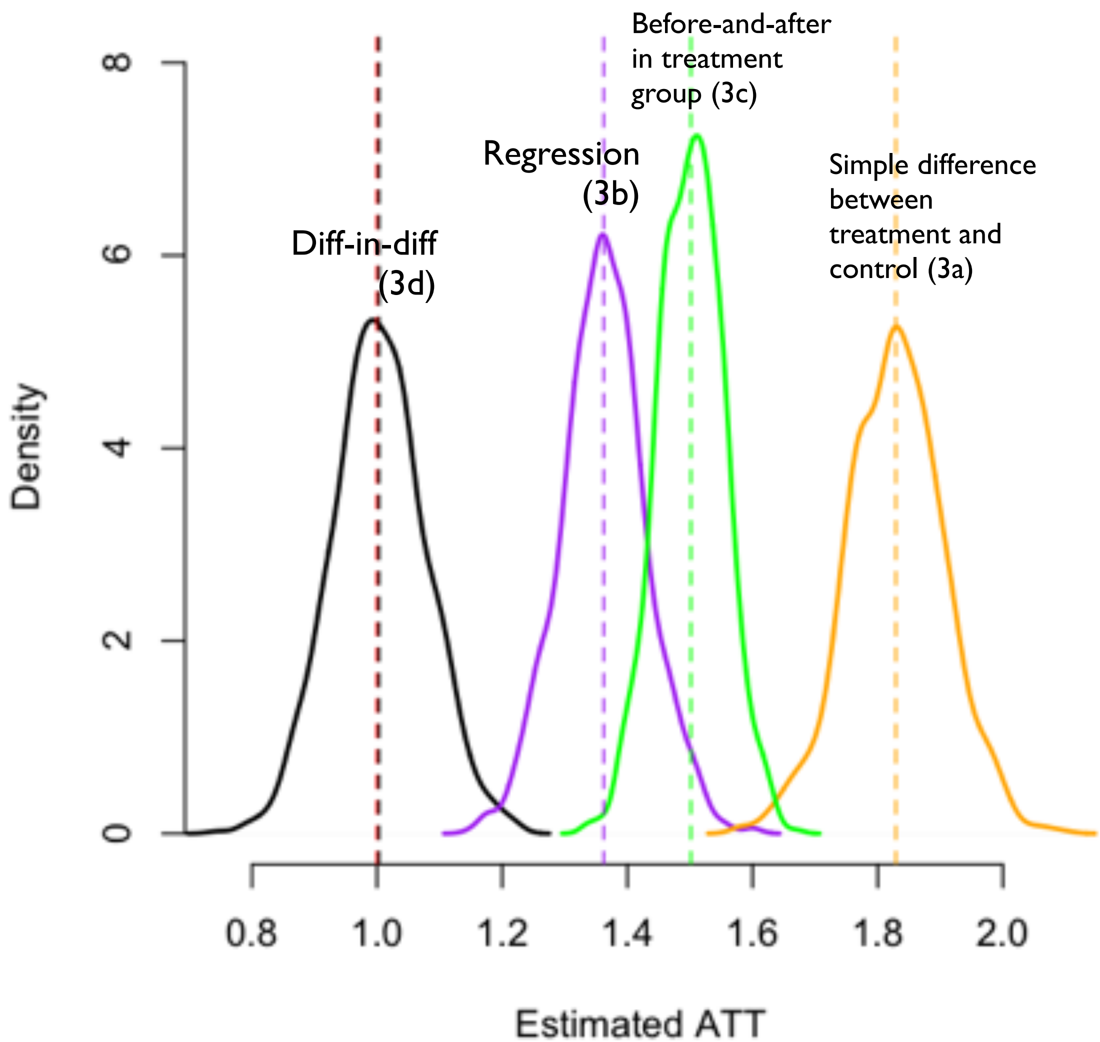
(3d) **Diff-in-diff:** Difference in before-and-after between treated and control units

(4) Repeat from step 1



Simulation 4 (non-random assignment with baseline outcomes): distribution of estimates

Do these results make sense?



Simulation 5: non-random assignment with pre-treatment outcomes (v2)

Recipe:

(1)* Same DGP as Simulation 3 except **time trend depends on x**:

$$x_i \sim N(0,1)$$
$$y_{i,pre} \sim N(x_i,1)$$
$$y_{0i,post} \sim N(x_i*(1+\lambda),1)$$
$$y_{1i,post} \sim N(x_i*(1+\lambda)+\boldsymbol{\tau},1)$$
$$\boldsymbol{\tau} = 1$$
$$\lambda=0.5$$

(2) Assign treatment as in Simulations 2 & 4:

$$\Pr(d_i=1) = 1/(1+\exp(-x_i))$$

(3) Same four ways of estimating ATT as in Simulations 3 & 4:

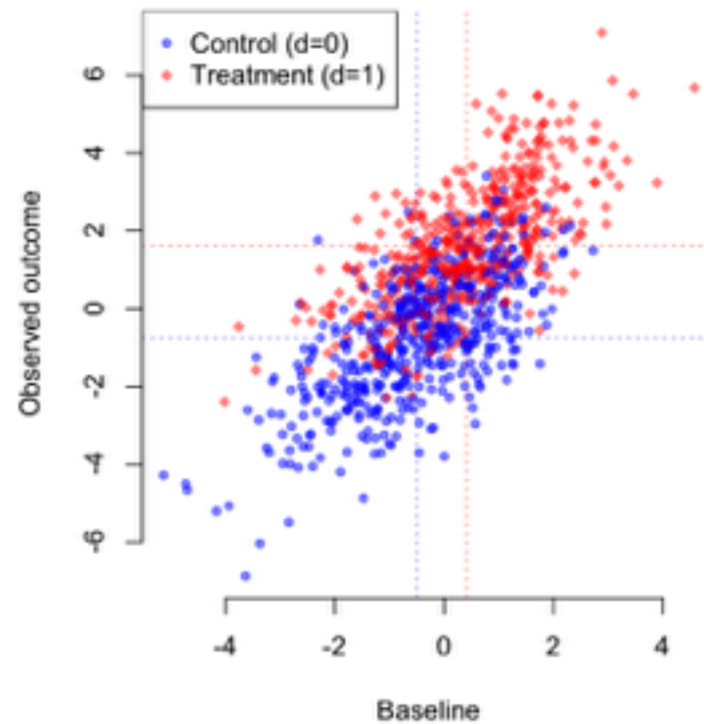
(3a) **Difference-in-means**: average difference in observed y between treated and control units

(3b) **Regression** of observed y on baseline outcome ($y_{i,pre}$) and d

(3c) **Before-and-after**: average change over time ($E[y_{i,post} - y_{i,pre}]$) in treatment group

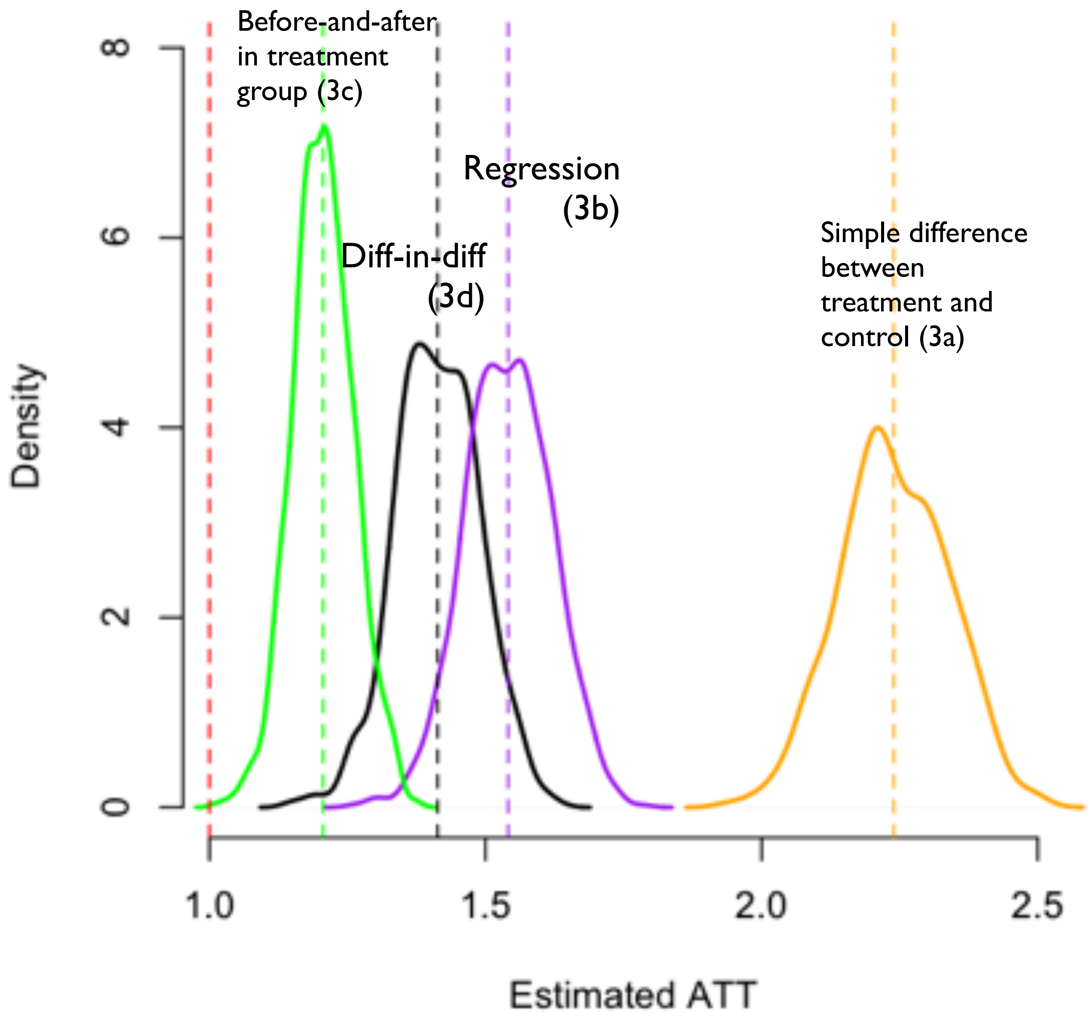
(3d) **Diff-in-diff:** Difference in before-and-after between treated and control units

(4) Repeat from step 1



Simulation 5 (non-random assignment with baseline outcomes, v2): distribution of estimates

Why does diff-in-diff fail now?

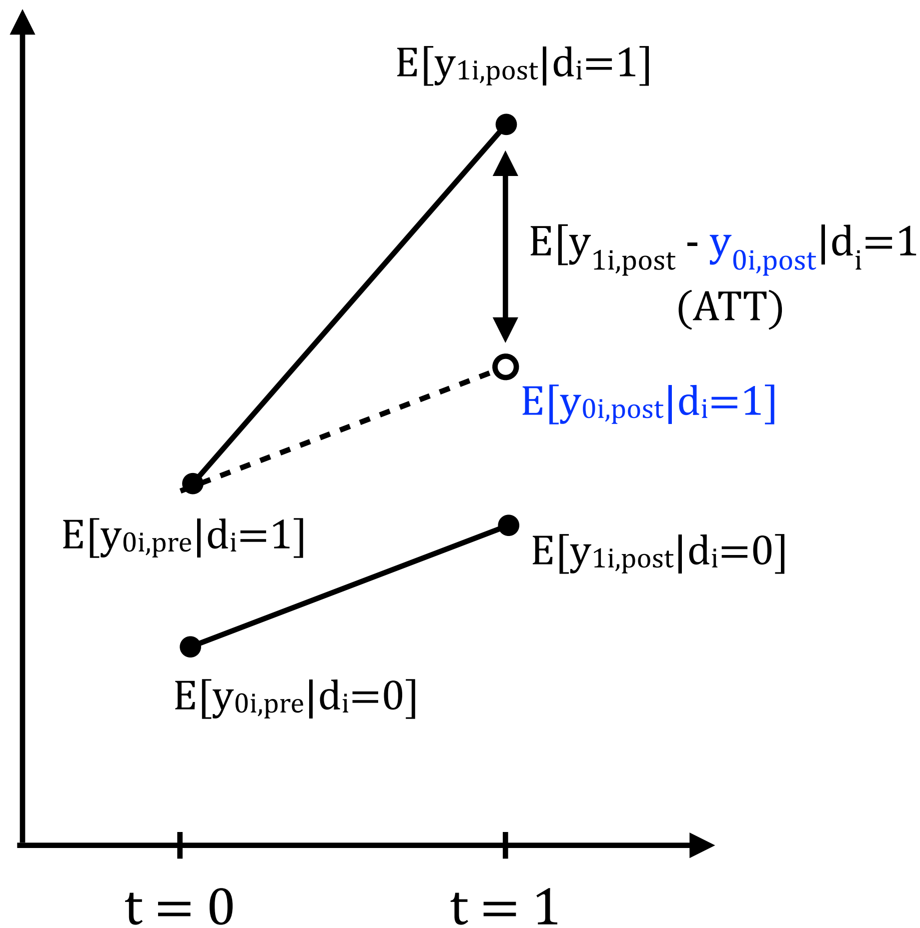


Why and when diff-in-diff works

Informally:

- Diff-in-diff is **potentially useful** when
  - binary treatment vs control
  - treatment and control group differ even in the absence of treatment (e.g. in the pre-treatment period)
- Diff-in-diff **works** when the baseline difference between the treatment and control group is constant over time (parallel trends assumption).

Parallel trends assumption



Parallel trends assumption:

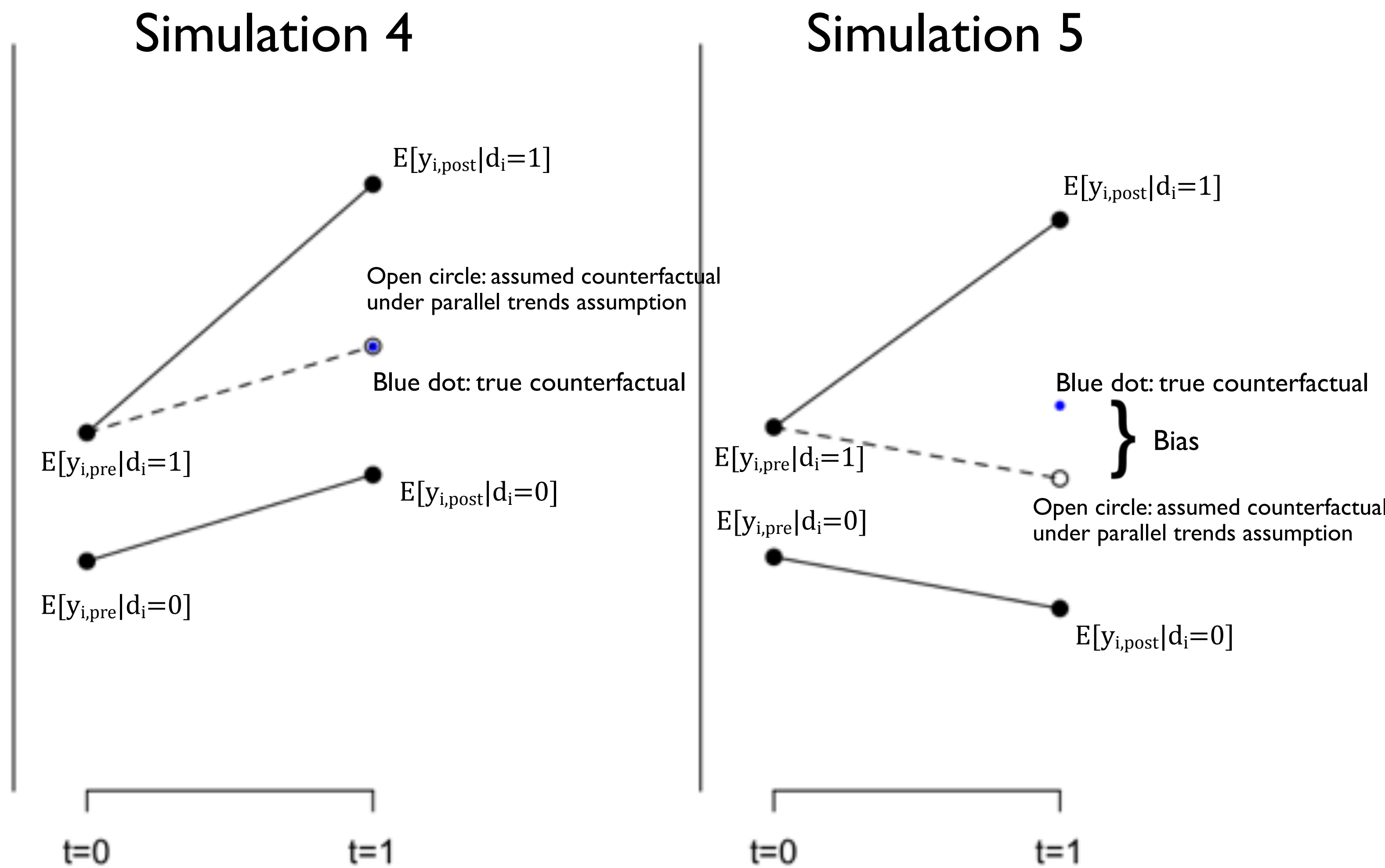
$$E[y_{0i,post} - y_{0i,pre}|d_i=1] = E[y_{0i,post} - y_{0i,pre}|d_i=0]$$

Change over time in potential outcome for treated

Change over time in potential outcome for control

The parallel trends assumption cannot be directly tested

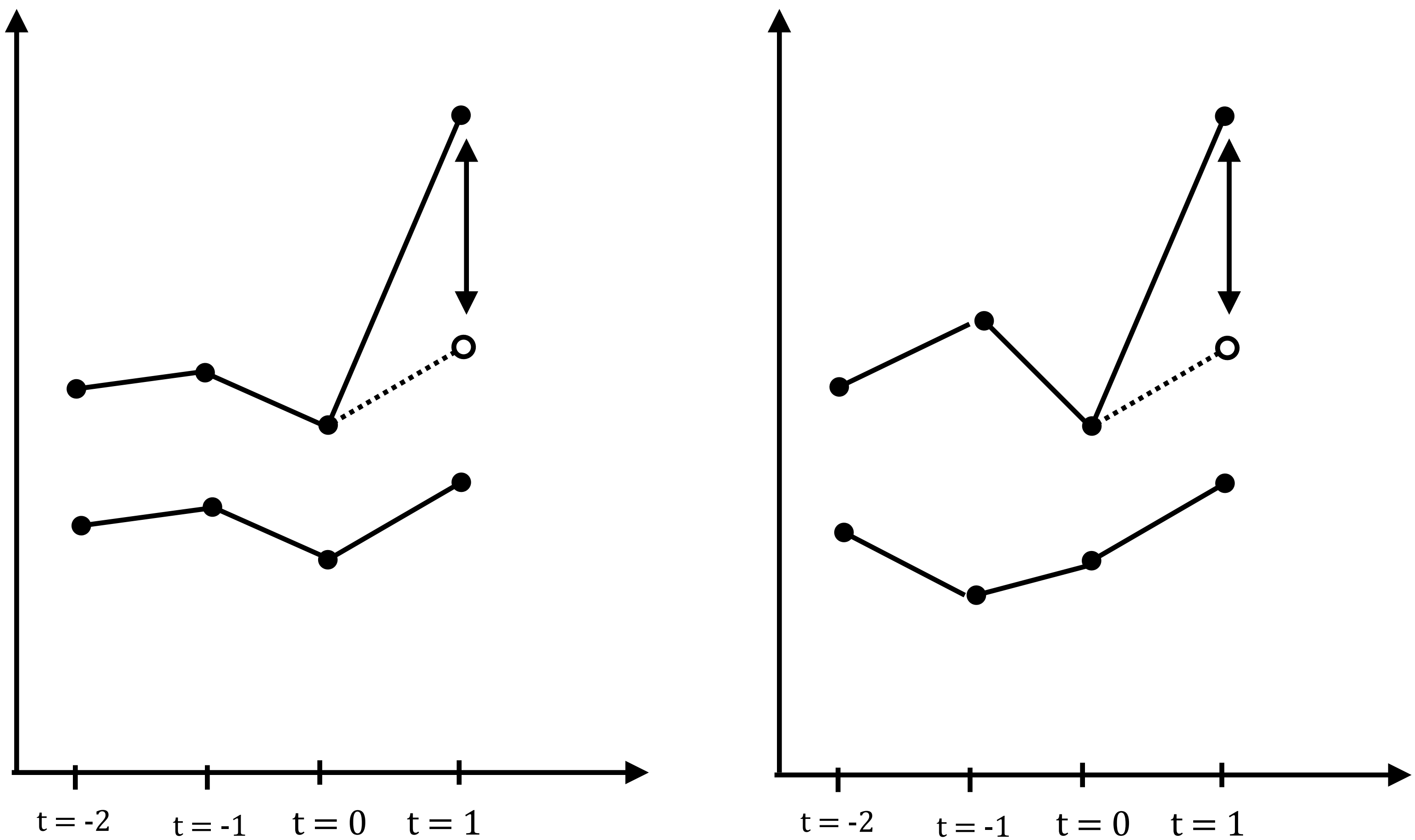
Consider simulations 4 and 5, where we observe the potential outcomes.



But we can check if trends are parallel in other periods

Parallel trends assumption looks good

Parallel trends assumption looks bad



Applying and implementing the diff-in-diff

Research question: Did the 2001 flood make its victims more supportive of the SPD government (due e.g. to its vigorous response)?

Figure 3: The Elbe Valley: Before the Flood 2001 and During the Flood 2002

Applying and implementing the diff-in-diff

The units are (SMD) electoral districts in Germany.

- What is the treatment?
- What is the outcome? What are the pre- and post-treatment periods?
- Name some possible confounding variables.
- What might be wrong with a simple difference-in-means?
- What is the parallel trends assumption behind the diff-in-diff in this case? Why might it not be satisfied?

Estimating the diff-in-diff: group means version

Simply calculate mean vote share for SPD in pre- and post-treatment period for flooded and non-flooded districts; subtract to get diff-in-diff.

```
. import delimited 1998_2002
(35 vars, 598 obs)

.
. **** TSCS versions
. * group means version
. mean spd_z_vs, over(postperiod flooded)

Mean estimation                   Number of obs    =      598

        Over: postperiod flooded
    _subpop_1: 0 0
    _subpop_2: 0 1
    _subpop_3: 1 0
    _subpop_4: 1 1
```

Over	Mean	Std. Err.	[95% Conf.	Interval]
spd_z_vs				
_subpop_1	41.70632	.4744889	40.77445	42.63819
_subpop_2	33.02612	1.116933	30.83253	35.21972
_subpop_3	38.82595	.5270443	37.79086	39.86104
_subpop_4	37.28977	1.109351	35.11107	39.46848

spd_z_vs: SPD vote share in district

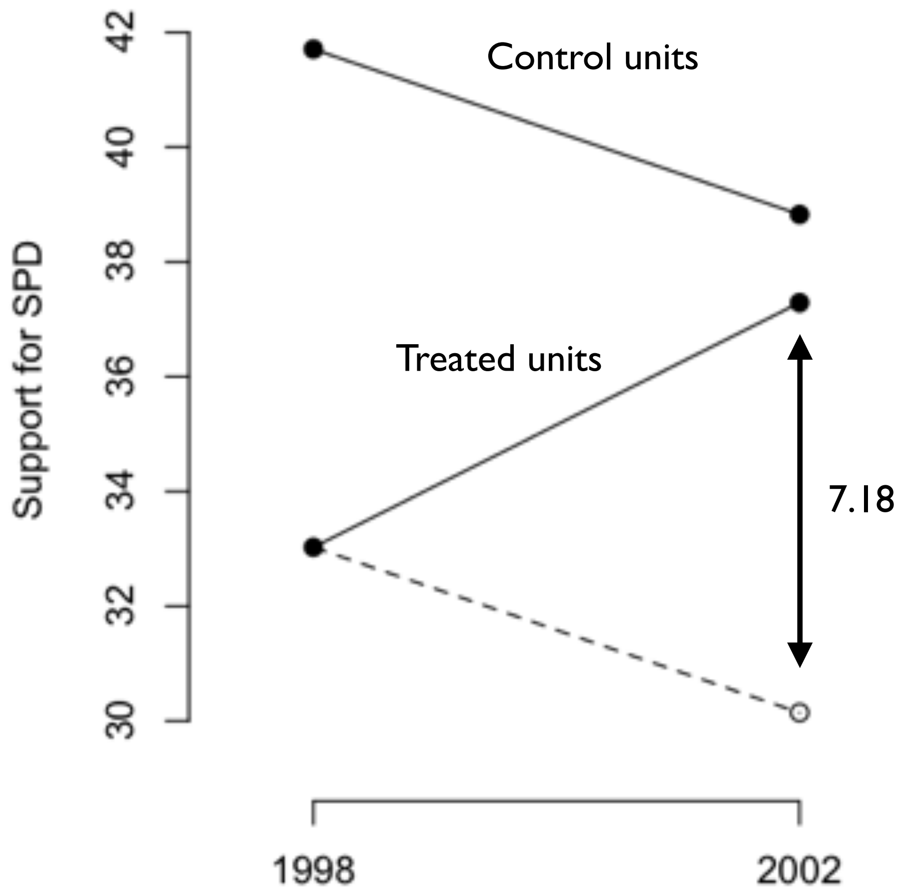
postperiod: 1 if 2002, 0 if 1998

flooded: 1 if district was flooded in 2001, 0 if not

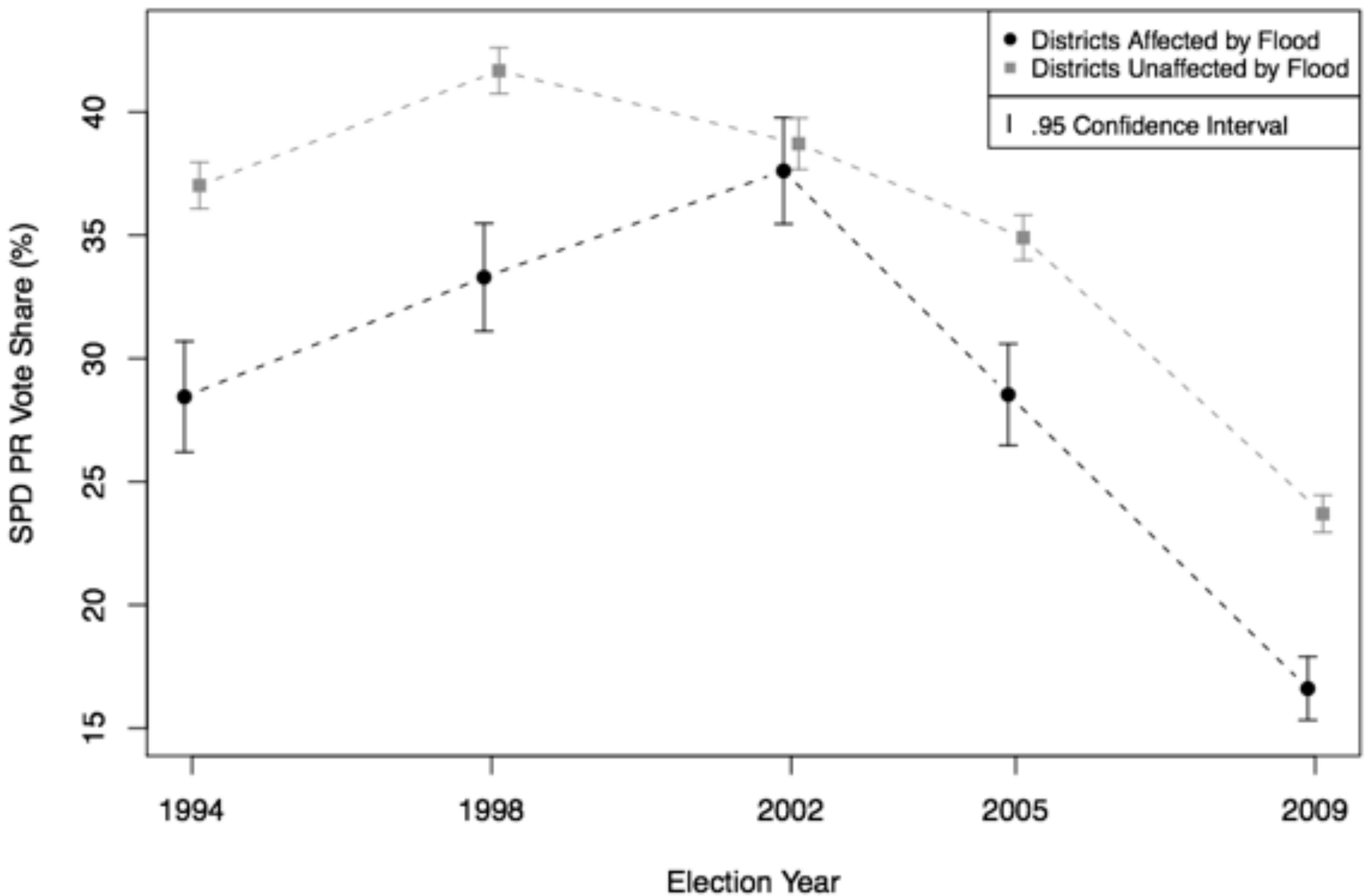
= (37.3-33.02)-(38.8-41.7)

= 7.18

Plotting the diff-in-diff



Assessing the parallel trends assumption



Estimating the diff-in-diff: interactions version

Convenient way to estimate the same thing in a regression:

```
. * interactions version, with clustering by district
. gen postflood = flooded*postperiod

. regress spd_z_vs flooded postperiod postflood, cl(wkr)

Linear regression                                      Number of obs =     598
                                                       F(  3,    298) =   99.02
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0666
                                                       Root MSE      =  8.0548

                                  (Std. Err. adjusted for 299 clusters in wkr)
```

spd_z_vs	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
flooded	-8.680194	1.200359	-7.23	0.000	-11.04245	-6.317939
postperiod	-2.880367	.2281177	-12.63	0.000	-3.329293	-2.431441
postflood	7.144014	.4685778	15.25	0.000	6.221874	8.066155
_cons	41.70632	.4755999	87.69	0.000	40.77036	42.64228

wkr: id for electoral district

Here, clustering standard errors because districts appear more than once. (How much data do we have if pre- and post- are separated by 20 minutes?)

See MHE section 8.1 and 8.2 for more on clustering.

Panel vs repeated cross-section

Everything so far applies to both

- repeated cross-sectional datasets (i.e. datasets where the specific units being surveyed change from time period to time period)
- panel datasets (i.e. datasets where the same units appear in each period)

If we have a panel, we can use other approaches that often yield more precise estimates.

Estimating the diff-in-diff: LSDV version

Least squares dummy variable model: Regress outcome on treatment and year, including a dummy for each each unit.

```
. xi: regress spd_z_vs postperiod postflood i.wkr, cl(wkr)
i.wkr             _Iwkr_1-299         (naturally coded; _Iwkr_1 omitted)

Linear regression                                      Number of obs =     598
                                                       F(  1,    298) =       .
                                                       Prob > F      =       .
                                                       R-squared     =  0.9528
                                                       Root MSE      =  2.5629

                                  (Std. Err. adjusted for 299 clusters in wkr)
```

spd_z_vs	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
postperiod	-2.880367	.3226071	-8.93	0.000	-3.515244	-2.24549
postflood	7.144014	.6626691	10.78	0.000	5.83991	8.448118
_Iwkr_2	-2.633802	2.84e-12	-9.3e+11	0.000	-2.633802	-2.633802
_Iwkr_3	-2.668777	2.84e-12	-9.4e+11	0.000	-2.668777	-2.668777
_Iwkr_4	-1.818636	2.84e-12	-6.4e+11	0.000	-1.818636	-1.818636
_Iwkr_5	.6821861	2.84e-12	2.4e+11	0.000	.6821861	.6821861
_Iwkr_6	.3288879	2.84e-12	1.2e+11	0.000	.3288879	.3288879
			……			
_Iwkr_296	3.327847	2.84e-12	1.2e+12	0.000	3.327847	3.327847
_Iwkr_297	3.345711	2.84e-12	1.2e+12	0.000	3.345711	3.345711
_Iwkr_298	3.293018	2.84e-12	1.2e+12	0.000	3.293018	3.293018
_Iwkr_299	4.427282	2.84e-12	1.6e+12	0.000	4.427282	4.427282
_cons	47.04176	.1613036	291.63	0.000	46.72432	47.3592

Intuition for the LSDV version

Parallel trends assumption required that difference between treatment and control groups is constant over time in the absence of treatment.

In interaction version, treatment and control groups got their own intercepts.

In LSDV version, all units get their own intercept.

(Note: Parallel trends assumption could apply at the group level even if it does not apply at the individual level.)

Estimating the diff-in-diff: `areg` version

Stata's areg command lets us run LSDV while suppressing the coefficients on the dummy variables:

```
. areg spd_z_vs postperiod postflood, cl(wkr) absorb(wkr) /* exactly the same as LSDV*/

Linear regression, absorbing indicators          Number of obs   =        598
                                                 F(   2,    298) =      66.99
                                                 Prob > F        =     0.0000
                                                 R-squared       =     0.9528
                                                 Adj R-squared   =     0.9050
                                                 Root MSE        =     2.5629

                                  (Std. Err. adjusted for 299 clusters in wkr)
------------------------------------------------------------------------------
             |               Robust
    spd_z_vs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  postperiod |  -2.880367   .3226071    -8.93   0.000    -3.515244    -2.24549
   postflood |   7.144014   .6626691    10.78   0.000      5.83991    8.448118
       _cons |   40.86443   .1483389   275.48   0.000      40.5725    41.15635
-------------+----------------------------------------------------------------
         wkr |   absorbed                                     (299 categories)
```

Estimating the diff-in-diff: fixed effects version

(We'll talk more about fixed effects next week.)

```
. xtset wkr postperiod /* wkr: election district; postperiod: after */
       panel variable:  wkr (strongly balanced)
        time variable:  postperiod, 0 to 1
                delta:  1 unit

. xtreg spd_z_vs postperiod postflood, cl(wkr) fe

Fixed-effects (within) regression               Number of obs      =       598
Group variable: wkr                             Number of groups   =       299

R-sq:  within  = 0.4150                         Obs per group: min =         2
       between = 0.0360                                        avg =       2.0
       overall = 0.0022                                        max =         2

                                                F(2,298)           =    134.20
corr(u_i, Xb)  = -0.1781                        Prob > F           =    0.0000

                                 (Std. Err. adjusted for 299 clusters in wkr)
------------------------------------------------------------------------------
             |               Robust
    spd_z_vs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  postperiod |  -2.880367   .2279259   -12.64   0.000    -3.328915   -2.431819
   postflood |   7.144014   .4681839    15.26   0.000     6.222649     8.06538
       _cons |   40.86443   .1048033   389.92   0.000     40.65818    41.07067
-------------+----------------------------------------------------------------
     sigma_u |  8.2468683
     sigma_e |  2.5628706
         rho |  .91192838   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Estimating the diff-in-diff: first-differences version

```
. drop postflood

. keep spd_z_vs flooded wkr postperiod

. reshape wide spd_z_vs, i(wkr) j(postperiod)
(note: j = 0 1)

Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                      598   ->     299
Number of variables                   4   ->       4
j variable (2 values)        postperiod   ->   (dropped)
xij variables:
                               spd_z_vs   ->   spd_z_vs0 spd_z_vs1
-----------------------------------------------------------------------------

. gen change_spd_vs = spd_z_vs1 - spd_z_vs0

. regress change_spd_vs flooded

      Source |       SS       df       MS              Number of obs =     299
-------------+------------------------------           F(  1,   297) =  101.74
       Model |  1336.51922     1  1336.51922           Prob > F      =  0.0000
    Residual |  3901.57368   297  13.1366117           R-squared     =  0.2552
-------------+------------------------------           Adj R-squared =  0.2526
       Total |   5238.0929   298   17.577493           Root MSE      =  3.6244

------------------------------------------------------------------------------
change_spd~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     flooded |   7.144014    .708266    10.09   0.000     5.750159     8.53787
       _cons |  -2.880367   .2205768   -13.06   0.000    -3.314458   -2.446276
------------------------------------------------------------------------------
```

Intuition: testing whether, at the district level, SPD vote share increased more 1998-2002 in flooded districts than others.

Two useful ways of thinking about the diff-in-diff

$$(E[y_{1i,post}|d_i=1] - E[y_{0i,pre}|d_i=1]) - (E[y_{0i,post}|d_i=0] - E[y_{0i,pre}|d_i=0])$$

(Before-and-after in treatment group) - (Before-and-after in control group)

"We are doing before-and-after in the treatment group, but we subtract the before-and-after in control group because we think that things might have changed over time in the treatment group even in the absence of the treatment."

$$(E[y_{1i,post}|d_i=1] - E[y_{0i,post}|d_i=0]) - (E[y_{0i,pre}|d_i=1]) - E[y_{0i,pre}|d_i=0])$$

(Treatment-control difference after) - (Treatment-control difference before)

"We are comparing the treatment and control group, but we subtract the difference between treatment and control before treatment because we think the two groups might have differed in levels even in the absence of the treatment."

Next week

Homework: Apply these techniques to Snow's cholera diff-in-diff.

Next week: From randomized experiments to fixed effects: different route to same techniques, with broader application.