Scott Cunningham's Codechella: DiD

Erich Denk

7/16/2021

Here are my notes from attending Scott Cunningham's "Codechella" session covering new developments in Difference-in-Differences estimators. I will note that some things I have lifted directly from his slides, while other points are my own notes. All of Scott's resources are posted on his github page. I went ahead and omitted the R code in the output, but it can easily be added if wanted.

Difference-in-Differences Intro

- A research designed when non-random treatment is applied to one or more groups
- A group of units do not receive the units at the same time (either never, or not yet)
- Observations before and after each group
- Research differences before and after, then differences the differences - hence the name

History

- An old, conceptually intuitive research design. Early attempts at using date back to several health policy debates.
- In econ, particularly labor economics, Orley Ashenfelter (1978), Bob LaLonde (1986) and Card and Krueger (1994).
- The single most popular method in economics at this point. Even more popular than RD. 25% of NBER working papers.
- DiD is popular because economists interest in large potentially impactful policies. However, a bunch of papers are showing the standard methods are biased. But, lots of new papers suggesting solutions.

Potential Outcomes Review

- Sometimes called the Rubin-Neyman model. Potential outcomes are thoguht experiements about worlds that never existed, but which *could have*
- Peter Hull and Pedro Sant'Anna insist that the potential outcomes notation should be different

An example: Aliens come and orbit earth, see sick people in hospitals and concluse the "hospitals" are hurting people. Motivated by anger and compassion they kill the doctors to save the patients. An instance of making causality synonymous with correlations.

Another example: "If a doctor puts a patient on a ventilator (D), will her symptoms change." Granger causality is another example of this. Every morning the rooster crows and the sun rises. Did the rooster cause the sun to rise, or did the sun cause the rooster to crow. Classic fallacy and excellent West Wing episode: *post hoc ergo propter hoc* "after this, therefore, because of this".

Final example: Sailor moves rudder back and forth but she remains on a straight line. There isn't even an observable correlation! Wind blows, she perfectly counters by turning the rudder. Systematic process erases the observable correlations. People rarely are behaving randomly. The movement of the rudder is endogenous. Rudder isn't broken.

Potential Outcomes Notation

"Potential outcomes are hypothetical states of the world (ex-ante) but historical outcomes are ex-post realizations. Major philosophical move here."

The individual treatment effect δ_i equals $Y_i^1 - Y_i^0$. The outcome with and without the treatment. But we only observe one of these. We are left to look at the averages (ATE).

$$\begin{aligned} E[\delta_i] &= E[Y_i^1 - Y_i^0] \\ &= E[Y_i^1] - E[Y_i^0] \end{aligned}$$

The *switching equation* is defined by the individual's observed health outcomes, Y, is determined by the treatment assignment, D_i, and corresponding potential outcomes

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

Twoway Fixed Effects

When workign with panel data, the so-called "twoway fixed effects" (TWFE) estimator is the workhorse estimator. It is easy to run, a verions of OLS and many people are interested in only in mean effects. Often times the outcome variable depends on unobserved factors which are also correlated with our explanatory variable of interest.

Traditionally we use it for estimating constant treatment effects with unobserved time-invariant heterogeneity. It is a linear model, so you'll be estimating conditional mean treatment effects - if you want the median, you can't use this. Once you have dynamic treatment effects and differential timing, we have a lot of issues.

Fixed Effects Regressions

There are a lot of things FE cannot help. We can think of FE as a regression with panel dummies. But, by putting in the dummies, it is equivalent to demeaning the data. A feature of panel data. Running a regression with the time-demeaned variables is numerically equivalent to a regression of y_{it} on x_{it} and unit specific dummy variables.

Even better, the regression with the time demeaned variables is consistent for β even when $Cov[x_{it}, c_i] \neq 0$

The identification assumptions are that regressors are strictly exogenous conditional on the unobserved effect. That allows x_{it} to be arbitrarilly related to c_i

Conclusion

- We reviewed TWFE because it is commonly used with individual level panel data and difference-in-differences
- Their main value is how they control for unobserved heterogeneity through a simple demeaning
- What we will see in this seminar, though, is that strict exogeneity actually imposed not just parallel trends, but also treatment effect homogeneity under differential timing

Covariates

Think of the John Snow cholera experiment. Policy cahnge moving dumping down the Thames.

Sample Averages:

$$\widehat{\delta}_{kU}^{2\times 2} = \left(\overline{y}_k^{\text{post}(k)} - \overline{y}_k^{\text{pre}(k)}\right) - \left(\overline{y}_U^{\text{post}(k)} - \overline{y}_U^{\text{pre}(k)}\right)$$

The two by two is not the causal parameter. However, we don't get ATT until we move to the Rubin framework.

$$\widehat{\delta}_{kU}^{2\times 2} = \underbrace{\left(E\big[Y_k^1 \mid \text{Post}\big] - E\big[Y_k^0 \mid \text{Pre}\big]\right) - \left(E\big[Y_U^0 \mid \text{Post}\big] - E\big[Y_U^0 \mid \text{Pre}\big]\right)}_{\text{Switching equation}}$$
$$+ \underbrace{E\big[Y_k^0 \mid \text{Post}\big] - E\big[Y_k^0 \mid \text{Post}\big]}_{\text{Adding zero}}$$

Then we can add 0, and rearrange things.

$$\widehat{\delta}_{kU}^{2\times 2} = \underbrace{E\big[Y_k^1 \mid \text{Post}\big] - E\big[Y_k^0 \mid \text{Post}\big]}_{\text{ATT}}$$
$$+ \underbrace{\Big[E\big[Y_k^0 \mid \text{Post}\big] - E\big[Y_k^0 \mid \text{Pre}\big]\Big] - \Big[E\big[Y_U^0 \mid \text{Post}\big] - E\big[Y_U^0 \mid \text{Pre}\big]\Big]}_{\text{Non-parallel trends bias in } 2\times 2 \text{ case}}$$

DiD doesn't have to be some exotic estimator. At the end of the day it only has to be a group of differences in means. The ATT is the first line! If our second and third terms are equal, the second line drops out and we have ATT by itself. If not, our pretrend assumption is violated.

Bias in our go-to estimators

- Good reasons to use TWFE:
  - It estimates the ATT under parallel trends
  - It's easy to calculate the standard errors
  - It's easy to include multiple periods
  - We can study treatments with different treatment intensity. (e.g., varying increases in the minimum wage for different states)

Think about Card and Krueger. OLS basically implicitly imputes the counterfactual for the treatment. Under parallel trends, OLS estimates the ATT for the two group case. Calculating standard errors is easy, multiple time periods is easy. But including covariates and time varying treatment ("differential timing") will introduce problems.

Alberto Abadie (2005)

- Abadie is really used best for longitudinal data or repeated cross sections where treatment occurs at one point in time.
- Abadie modeled differential selection based on covariates
- DD type estimator but not TWFE.
- Still need treatment and comparison groups, before and after, but with conditional parallel trends.

- High level: Look for natural experiments and use econometrics to clean things up.
- No randomization. Remember, DD doesn't require randomization = it requires parallel trends. Treatment is selecting on observable covariates.
- We may control for X because treatement is only conditional on X.
- In TWFE when you control for baseline X, it gets absorbed by the unit FE.

Three step method for Abadie:

1. Compute each unit's "after minus before"
2. Then estimate a propensity score which you'll use to weight each unit
3. Finally, compare weighted changes in "after minus before" for treatment versus comparison groups

You can have heterogeneous treatment effects, but not differential timing

Assumptions

1. Conditional parallel trends
2. Common support - Only those where propensity scores overlap. Range of propensity score that contains treatment and control group units.

Define the ATT parameter of interest

$$ATT = E[Y_t^1 - Y_t^0 | D_t = 1]$$

Abadie's estimator

$$E[\frac{Yt - Yb}{Pr(D_t = 1)} \times \frac{D_t - Pr(D = 1|X_b)}{1 - Pr(D = 1|X_b)}]$$

Propensity scores

- Usually there's almost no guidance that I've seen in how to estimate the propensity score except to say use logit or probit
- Dehejia and Wahba (2002) anyway
- Not so here – this is semi-parametric in the sense that you have to use a series of polynomials based on the X controls
- Weirdly, you can use OLS linear probability models (which I've never seen) or something called series logit estimation

Stata command is called absdid

You need treatment (varname), X variables, the order in which the variables occur, and the exact estimator (LPM or logit)

Use the LaLonde NSW job trainings program data?

Concluding remarks

LaLonde longitudinal data where you have a baseline and a follow-up Repeated cross-sections or panels Controls will cause the estimates to vary based on the type of approximation you use (logit for instance vs LPM) and the order in which the polynomials are used

Sant'Anna and Zhao (2020) Doubly Robust

- They combine regression and weighting estimators into one specification and are consistent so long as:
  - The regression specification for the outcome is correctly specified
  - The propensity score specification is correctly specified
  - DR is a class of estimators that possess this property
  - You're basically controlling for X twice: with a linear regression, with a propensity score, to cover your bas
- Bridging gaps while simultaneously moving the ball forward
- Basic assumptions for DD with covariates.
- TWFE assumptions for DD with covariates
- Estimation alternatives to TWFE

Maybe you are unsure whether the propensity score was properly specificed. Two strikes instead of one.

DD *always* estimates the ATT becasue it's only the treatment effect for the treatment group in the post-treatment period

Basic assumptions of DiD 1. Assume panel data or repeated cross-sectional data 2. Conditional parallel trends: If you were putting covariates into your DD regression, then you were assuming conditional parallel trends 3. Common support or overlap

Assumption 4

- The implications of that TWFE regression with assumptions 1-3 gave us those previous expressions which then require placing further restrictions on treatment effects and trends when estimating with TWFE.
- TWFE Assumption 4: Homogenous treatment effects in X

This is because when you difference out those previous equations, you need θX to cancel to leave you with δ which implies homogeneity in X.

Assumption 5 and 6

For D = 0, 1, we need "no X-specific trends in both groups":

$$E[Y_1 - Y_0|D = d, X] = E[Y_1 - Y_0|D = d]$$

Intuition: Sant'Anna and Zhao (2020) say in footnote 4 "[this] follows from analogous arguments" which is the previous slides' manipulation of terms. Key is to remember these are time-varying covariates so they don't cancel out within treatment category, so you need the trends in X to cancel out.

Without these six, in general TWFE will not identify ATT. Unclear how off it'll be, but it will be biased is the point.

What if you claim you need X for conditional parallel trends? You have three options:

1. Outcome regression a la Heckman 1997 (Assumptions 1-3)
2. Inverse probability weighting (Abadie 2005) Assumptions 1-6
3. TWFE (Everybody!) Need assumptions 1-6

Problem is options 1 and 2 need the models to be correctly specified. Doubly robust combines them to give us insurance. That's the basic idea. Gives you two chances to be wrong.

To a kid with a hammer, everything is a nail

Use the right tool (oven) for the job (making lasagna), not the same tool (hammer) regardless of the job (making lasagna)

One of the main things I learned from this paper was again biases in TWFE with covariates – Mixtape and MHE don't cover this. This method only needed three assumptions not the six for TWFE.

Like everything Pedro does, there is code for this but it's only in R – `DRDID`

But it's one of the main options in Callaway and Sant'anna under differential timing, and therefore it's crucial we understand this. But you still have to have specified correctly either at least the outcome model or propensity score model.

Differential Timing

- We covered mostly the simple two group case
- In the two group case, we can estimate the ATT under parallel trends using OLS with unit and time fixed effects
- If we have covariates, then we can use TWFE under restrictive assumptions, or we have other options (OR, IPW, DR)
- For this next part, similar to how we did with Sant'Anna and Zhou (2020), we will decompose TWFE to understand what it needs for unbiasedness under differential timing
- All of this is from Goodman-Bacon (2021, forthcoming) though the expression of the weights is from 2018 for personal preference
- Goodman-Bacon (2021, forthcoming) shows that parallel trends is not enough for TWFE to be unbiased when treatment adoption is described by differential timing
- TWFE with differential timing uses treated groups as controls – not all estimators do – and this can introduce bias

Decomposition

- TWFE estimates a parameter that is a weighted average over all 2x2 in your sample
- TWFE assigns weights that are a function of sample sizes of each "group" and the variance of the treatment dummies for those groups
- TWFE needs two assumptions: that the variance weighted parallel trends are zero (far more parallel trends iow) and no dynamic treatment effects (not the case with 2x2)
- Under those assumptions, TWFE estimator estimates the variance weighted ATT as a weighted average of all possible ATTs ## Cheng and Hoekstra Castle Doctrine
- Cheng and Hoekstra (2013) are interested in whether expansions to "castle doctrine statutes" at the state level increase or decrease gun violence.
- Prior to these expansions, English common law principle required "duty to retreat" before using lethal force against an assailant except when the assailant is an intruder in the home
  - The home is one's "castle" – hence, "castle doctrine"
  - When intruders threatened the victim in the home, the duty to retreat was waived and lethal force in self-defense was allowed

Bacon Decomposition

TWFE estimate yields a weighted combination of each groups' respective 2x2 (of which there are 4 in this example)

- Let there be two treatment groups (k,l) and one untreated group (U)
- k,l define the groups based on when they receive treatment (differently in time) with k receiving it earlier than l
- Denote $\overline{D}_k$ as the share of time each group spends in treatment status
- Denote δ_{jb}^{2x2} as the canonical 2 × 2 DD estimator for groups j and b where j is the treatment group and b is the comparison group

We will get k^2 2 x 2s with k timing groups.

TWFE estimates yields a weighted combination of each groups' respective 2x2 (of which there are 4 in this example)

$$\delta^{DD} = \sum_{k \neq U} s_{kU} \delta_{kU}^{2x2} + \sum_{k \neq U} \sum_{I > k} s_{kl} [\mu_{kl} \delta_{kl}^{2x2,k} + (1 - \mu_{kl}) \delta_{lk}^{2x2,l}]$$

where the first 2x2 combines the k compared to U and the I to U (combined to make the equation shorter)

Weights discussion

- Two things to note:
  - More units in a group, the bigger its 2x2 weight is
  - Group treatment variance weights up or down a group's 2x2
- Think about what causes the treatment variance to be as big as possible. Let's think about the sku weights.
  - D=0.1. Then$0.1 \times 0.9 = 0.09$
  - D=0.4. Then$0.4 \times 0.6 = 0.24$
  - D=0.5. Then$0.5 \times 0.5 = 0.25$
  - D=0.6. Then$0.6 \times 0.4 = 0.24$
- This means the weight on treatment variance is maximized for *groups treated in middle of the panel*

Weighted Group-Time ATT

Callaway and Sant'Anna 2020

CS considers identification, aggregation, estimation and inference procedures for ATT in DD designs with: 1. multiple time periods 2. variation in treatment timing (i.e., differential timing) 3. parallel trends only holds after conditioning on observables

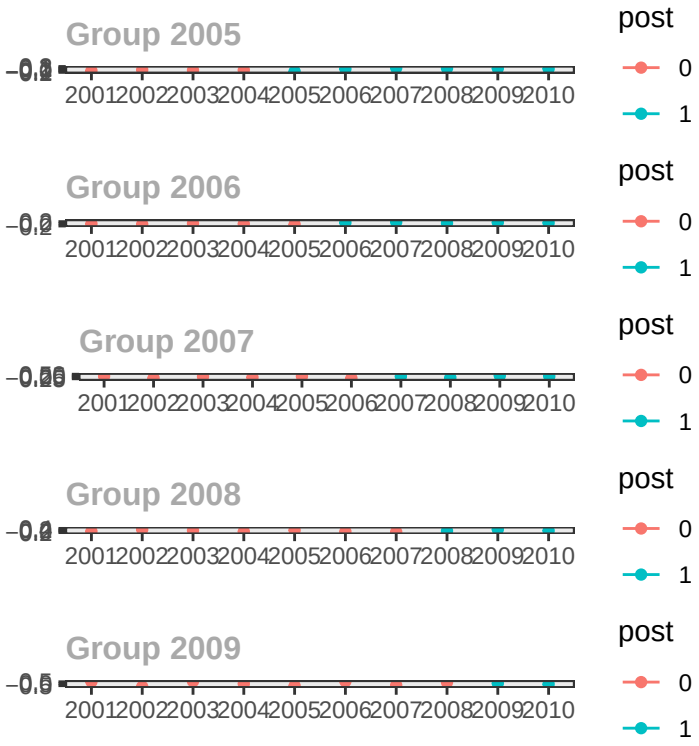
Group-time ATT is the parameter of interest in CS

We can apply their R code to the Castle doctrine paper.

```
##
## Call:
## aggte(MP = atts, type = "group")
##
```

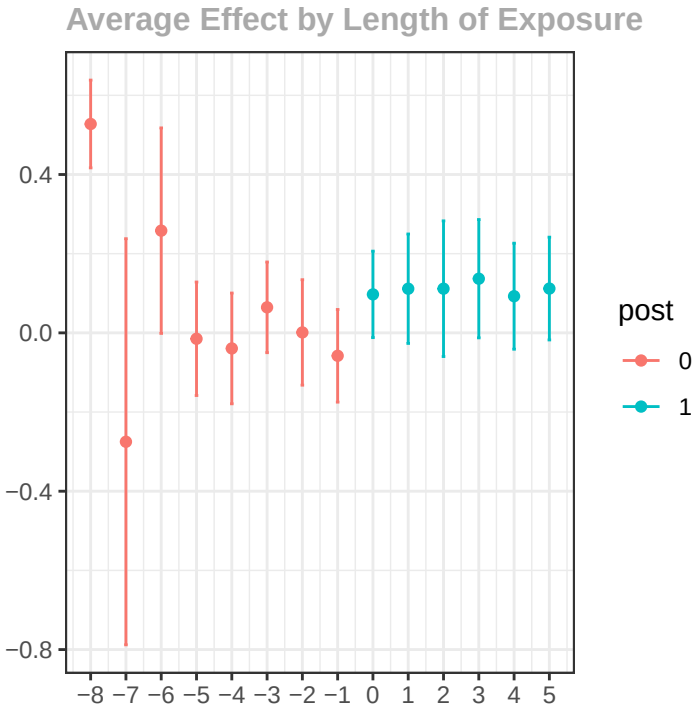
```
## Reference: Callaway, Brantly and Pedro H.C. Sant'Anna.  "Difference-in-Differences with Multiple Time
##
##
## Overall ATT:
##     ATT Std. Error     [95%  Conf. Int.]
##  0.1084     0.0376    0.0348      0.1821 *
##
##
## Group Effects:
##  group     ATT Std. Error [95% Simult.  Conf. Band]
##   2005  0.0931     0.0334        0.0154      0.1708 *
##   2006  0.1099     0.0542       -0.0160      0.2359
##   2007  0.1284     0.0544        0.0020      0.2548 *
##   2008  0.1221     0.0584       -0.0134      0.2577
##   2009 -0.0028     0.0385       -0.0922      0.0866
## ---
## Signif. codes: '*' confidence band does not cover 0
##
## Control Group:  Never Treated,  Anticipation Periods:  0
## Estimation Method:  Doubly Robust


##
## Call:
## att_gt(yname = "l_homicide", tname = "year", idname = "sid",
##     gname = "effyear", xformla = NULL, data = castle, panel = TRUE,
##     control_group = "nevertreated", bstrap = TRUE, biters = 1000,
##     clustervars = "sid", est_method = "dr", print_details = FALSE)
##
## Reference: Callaway, Brantly and Pedro H.C. Sant'Anna.  "Difference-in-Differences with Multiple Time
##
## Group-Time Average Treatment Effects:
##  Group Time ATT(g,t) Std. Error [95% Simult.  Conf. Band]
##   2005 2001  -0.0593     0.0435       -0.1791      0.0604
##   2005 2002   0.0171     0.0445       -0.1053      0.1395
##   2005 2003  -0.0139     0.0333       -0.1055      0.0777
##   2005 2004   0.0006     0.0337       -0.0921      0.0933
##   2005 2005  -0.1203     0.0395       -0.2288     -0.0117 *
##   2005 2006   0.0990     0.0335        0.0068      0.1912 *
##   2005 2007   0.1769     0.0443        0.0551      0.2987 *
##   2005 2008   0.1496     0.0462        0.0225      0.2767 *
##   2005 2009   0.1413     0.0418        0.0263      0.2562 *
##   2005 2010   0.1119     0.0515       -0.0298      0.2537
##   2006 2001   0.0024     0.0771       -0.2095      0.2144
##   2006 2002  -0.0397     0.0694       -0.2305      0.1510
##   2006 2003   0.0417     0.0612       -0.1267      0.2101
##   2006 2004  -0.0050     0.0706       -0.1992      0.1891
##   2006 2005  -0.0556     0.0622       -0.2268      0.1155
##   2006 2006   0.1080     0.0513       -0.0332      0.2492
##   2006 2007   0.1603     0.0605       -0.0060      0.3265
##   2006 2008   0.0638     0.0854       -0.1711      0.2987
##   2006 2009   0.1288     0.0764       -0.0813      0.3390
##   2006 2010   0.0888     0.0588       -0.0729      0.2506
##   2007 2001   0.1764     0.1456       -0.2241      0.5770
##   2007 2002  -0.1351     0.0812       -0.3585      0.0882
```

```
##   2007 2003   0.1037     0.1726       -0.3709      0.5783
##   2007 2004  -0.0251     0.0832       -0.2539      0.2036
##   2007 2005   0.1507     0.0864       -0.0870      0.3884
##   2007 2006  -0.1618     0.0969       -0.4284      0.1048
##   2007 2007   0.1454     0.1551       -0.2813      0.5721
##   2007 2008  -0.0624     0.1388       -0.4442      0.3194
##   2007 2009   0.2710     0.1010       -0.0068      0.5489
##   2007 2010   0.1596     0.0932       -0.0967      0.4158
##   2008 2001  -0.0304     0.0937       -0.2880      0.2272
##   2008 2002   0.2458     0.0876        0.0050      0.4867 *
##   2008 2003   0.1110     0.1027       -0.1716      0.3935
##   2008 2004  -0.0577     0.0357       -0.1559      0.0405
##   2008 2005   0.1414     0.0403        0.0306      0.2522 *
##   2008 2006  -0.0591     0.0486       -0.1926      0.0745
##   2008 2007  -0.1035     0.0840       -0.3345      0.1275
##   2008 2008   0.0368     0.0556       -0.1161      0.1898
##   2008 2009   0.2588     0.1044       -0.0284      0.5460
##   2008 2010   0.0707     0.0561       -0.0835      0.2250
##   2009 2001   0.5276     0.0435        0.4079      0.6474 *
##   2009 2002  -0.7645     0.0445       -0.8869     -0.6420 *
##   2009 2003   0.6098     0.0333        0.5182      0.7014 *
##   2009 2004  -0.0113     0.0337       -0.1040      0.0814
##   2009 2005  -0.5490     0.0395       -0.6576     -0.4405 *
##   2009 2006   0.6128     0.0340        0.5193      0.7062 *
##   2009 2007  -0.3821     0.0353       -0.4793     -0.2849 *
##   2009 2008   0.3607     0.0559        0.2071      0.5143 *
##   2009 2009   0.1026     0.0405       -0.0089      0.2141
##   2009 2010  -0.1082     0.0443       -0.2300      0.0135
## ---
## Signif. codes: '*' confidence band does not cover 0
##
## Control Group:  Never Treated,  Anticipation Periods:  0
## Estimation Method:  Doubly Robust
```




```
##
## Call:
## aggte(MP = atts, type = "dynamic")
##
## Reference: Callaway, Brantly and Pedro H.C. Sant'Anna.  "Difference-in-Differences with Multiple Time
##
##
## Overall ATT:
##     ATT Std. Error     [95%  Conf. Int.]
##  0.1103     0.0347    0.0423      0.1783 *
##
##
## Dynamic Effects:
##  event time     ATT Std. Error [95% Simult.  Conf. Band]
##          -8  0.5276     0.0426        0.4168      0.6384 *
##          -7 -0.2751     0.1971       -0.7880      0.2379
##          -6  0.2582     0.0997       -0.0014      0.5177
##          -5 -0.0149     0.0551       -0.1582      0.1284
##          -4 -0.0393     0.0537       -0.1792      0.1006
##          -3  0.0645     0.0440       -0.0500      0.1790
##          -2  0.0011     0.0512       -0.1321      0.1343
##          -1 -0.0579     0.0450       -0.1749      0.0591
##           0  0.0972     0.0419       -0.0119      0.2064
##           1  0.1115     0.0531       -0.0266      0.2497
##           2  0.1116     0.0658       -0.0597      0.2829
##           3  0.1368     0.0574       -0.0126      0.2862
##           4  0.0926     0.0515       -0.0413      0.2265
##           5  0.1119     0.0499       -0.0178      0.2417
## ---
## Signif. codes: '*' confidence band does not cover 0
```

```
##
## Control Group:  Never Treated,  Anticipation Periods:  0
## Estimation Method:  Doubly Robust
```



Chaisemartin and D'Haultoeuiflle (2020)

Main takeaways: - TWFE can give you non-intelligible weights. Hard to interpret. - Can choose a alternative estimand that bypass these issues.

Some code implementing this in the dCdH decomposition.

```
## OLS estimation, Dep. Var.: l_homicide
## Observations: 550
## Fixed-effects: sid: 50,  year: 11
## Standard-errors: Clustered (sid)
##         Estimate Std. Error t value Pr(>|t|)
## treated   0.0788   0.058094  1.3564 0.181183
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.176387     Adj. R2: 0.899525
##                  Within R2: 0.01263
```

```
##                       type  weight  avg_est
## 1 Earlier vs Later Treated 0.06648 -0.02022
## 2 Later vs Earlier Treated 0.03904  0.07752
## 3     Treated vs Untreated 0.89447  0.08622
```

```
## [1] "Under the common trends assumption, "
## [1] "beta estimates a weighted sum of 97 ATTs."
## [1] "97 ATT receive a positive weight, and 0 receive a negative weight."
```

```
## [1] "The sum of the positive weights is equal to 1.000000."
## [1] "The sum of the negative weights is equal to 0.000000."
## [1] "beta is compatible with a DGP where the average of those ATT is equal to 0,"
## [1] "while their standard deviation is equal to 0.325135."
## [1] "All the weights are positive, so beta cannot be of a different sign than all those ATT."
```

```
## [1] 1
```

```
## [1] 0
```

Sun and Abraham

- SA is a decomposition of the population regression coefficeints on event study leads and lags with differential timing estimated with TWFE
- They show that the population coefficient is "contaminated by information from other leads and lags
- SA presents an alternative estimator similar to CS.
- Problems seem to occur with DD when we introduce treatment effect heterogeneity
- Under treatment effect heterogeneity, spurious non-zero positive lead coefficients even without a pre-trend.
- Exacerbate by the TWFE related weights related weights as under some scenarios, the weights sum to zero and"cancel out" the treatment effects from other periods
- They present a 3-step TWFE based alternative estimator which addresses the problems that they find
- When treatment occurs at the same time, we say they are part of the same cohort, e
- If we bin the data, then a lead or lag l will appear in the bin g so sometimes they use g instead of l or $l \in g$
- Building block is the "cohort-specific ATT" or $CATT_{e,l}$ – same thing as CS group-time ATT
- Estimate $CATT_{e,l}$ with population regression coefficient μ_l

I'm skipping the notation in my notes here, but Scott's slides are available here.

Identifying Assumptions

Assumption 1: Parallel trends in baseline outcomes: $E[Y^{\infty}_{i,t} - Y^{\infty} + i, s | E_i = e]$ is the same for all $e \in supp(E_i)$ and for all s, t and is equal to $E[Y^{\infty} - Y^{\infty}]$

Interesting SA comment: Never-treated units are likely to differ from ever-treated units in many ways; think of a Roy model. What does it imply that they chose not to get treated? It may imply net negative treatment effects and that could mean they may not share the same evolution of baseline outcomes as the treatment groups. If you think they are unlikely to satisfy this assumption, then drop them. Almost like a synthetic control approach.

Assumption 2: No anticipator behavior in pre-treatment periods: There is a set of pre-treatment periods such that $E[Y^{e}_{i,e+l} - Y^{\infty}_{i,e+l} | E_i = e] = 0$ for all possible leads.

Basically means that potential outcomes prior to treatment at baseline by on average the same. This means there is no pre-trends, essentially. This is most plausible if the full treatment paths are not known to the units (e.g., Craigslist opening erotic services without announcement)

Assumption 3: Treatment effect homogeneity: For each relative time period l, the $CATT_{e,l}$ doesn't depend on the cohort and is equal to $CATT_l$.

Assumption 3 requires each cohort experience the same path of treatment effects. Treatment effects need to be the same across cohorts in every relative period for homogeneity to hold, whereas for heterogeneity to occur, treatment effects just need to differ across cohorts in one relative time period. Doesn't preclude dynamic treatment effects, though. It just imposes that cohorts share the same treatment path.

Again, I got a little behind and have skipped ahead beyond the discussion of the weights here.

Conclusions

- Bacon shows the TWFE coefficient on the static parameter is "contaminated" by other periods leads and lags
- Three strong assumptions needed for TWFE to be unbiased: parallel trends, no anticipation, and treatment homogeneity
- Three step interaction-weighted estimator is an alternative Doesn't restrict to treatment profile homogeneity
- Callaway and Sant'Anna (2020) and Sun and Abraham (2020) use different controls, but under certain situations (no covariates, never treated) they are the same ("nested")
- Software in R and Stata exist

Borusyak, Jaravel and Spiess (2021)

Explicit imputation estimator - They provided analysis of TWFE flaws under heterogeneity as well as event study analysis - This is a paper that shows the problems with TWFE under heterogeneity, but then writes out a solution that uses imputation

Static model

Themes $y_{it} = \alpha_i + \gamma t + \delta D_i + \epsilon_i$ Contribution: Define target parameters and assumptions. Proposes a more formal disciplined approach of choosing the weighted average of treatment effects

Basics

- Potential outcomes without treatment will follow a parallel trend (but one with a bit more structure)
- No anticipatory effects
- Treatment effects follow some model that restricts heterogeneity a priori for economic reasons

Event study contributions

1. Can't identify point estimates of leads in event study design Seperate out the testable assumptions about pre-trends from dynamic treatment effects under these assumptions
2. Implicit homogeneity assumption in event study may lead to estimates putting negative weight on long-run lags under differential timing

- When we have long lags, regression is using extrapolation based on forbidden regressions which negatively weights
- This is fine with homogenous treatment effects and in fact is an argument for TWFE, but not with heterogeneity

3. Spurious identification of longrun effects can happen under heterogeneity with staggered rollouts

Again, this paper is telling us that the presence of heterogeneous treatment effects is largely what makes analysis so challenging

Imputation estimator

- "The most efficient linear unbiased estimator of any pre- specified weighted average of treatment effects under ho- moskedasticity"
- Separate assumption from estimation; incorporate the former Estimate a flexible high-dimensional regression Aggregate the coefficients
- All other unbiased linear estimators are less efficient
- Avoids pre-test problems pointed out by Roth (2018) (just wasn't able to work it in unfortunately)

Again, more slides for this paper, he was moving quickly as time was running out.

Basic Assumptions Going Forward

Everything is kinda broken and needs fixed

Differential timing with heterogeneity - Bacon, Callaway and Sant'anna, etc. Covariates - Abadie, Santa'anna and Zhao Fuzzy - de Chaisemartin and D'Haultfoeuille

We are at the conclusion of the waves of papers and the software is now widely available. Just need to make the initial investment.

- Simple 2x2 has its own problems when estimated using TWFE if you include covariates
- Stronger assumptions needed to include covariates, and bias can be large
- Don't control for covariates that could be affected by the outcome
- Why pay more for the same car? Actually fewer assumptions in some cases
- Main problem in differential timing is heterogeneity and the use of already-treated units as controls
- If you use TWFE for differential timing, report the Bacon decomposition and report the number of never-treated units
- If you are estimating event studies using TWFE, remember to drop two leads to address multiple forms of collinearity (SA; BJS)
- If you have differential timing, consider going directly to one of the robust estimators we discussed
- CS has additional benefits like examining heterogenous responses by timing – this is part of the value of defining target parameters as weighted averages