Causal Inference II

Mixtape Session

MIXTAPE SESSIONS

Roadmap

Differential timing

- Introduction
- TWFE Estimator
- Simulation

Aggregating building blocks

- CS

Two Regression Models

$$\begin{aligned} Y_{ist} &= \alpha_0 + \alpha_1 Treat_{is} + \alpha_2 Post_t + \delta(Treat_{is} \times Post_t) + \varepsilon_{ist} && \text{(1)} \\ Y_{ist} &= \beta_0 + \delta D_{ist} + \tau_t + \sigma_s + \varepsilon_{ist} && \text{(2)} \end{aligned}$$

First equation is used for simple designs when everyone is treated at once; second equation was used when different groups were treated at different times ("differential timing")

First equation works; second one only sometimes works

Twoway fixed effects

- When working with panel data, the so-called TWFE estimator is the workhorse estimator
- It's easy to implement, handles time-varying treatments, has a relatively straightforward interpretation under constant treatment effects, standard errors are easy to calculate and understand
- Interpretation is more complicated with heterogenous treatment effects

Difference-in-differences

- This is the TWFE specification for diff-in-diff

$$Y_{ist} = \alpha + \delta D_{st} + \sigma_s + \tau_t + \varepsilon_{ist}$$

- The hope was that $\widehat{\delta}$ equaled a “reasonably weighted average” over all underlying treatment effects and therefore was the ATT
- But let’s find out what it actually is doing

K^2 distinct DDs

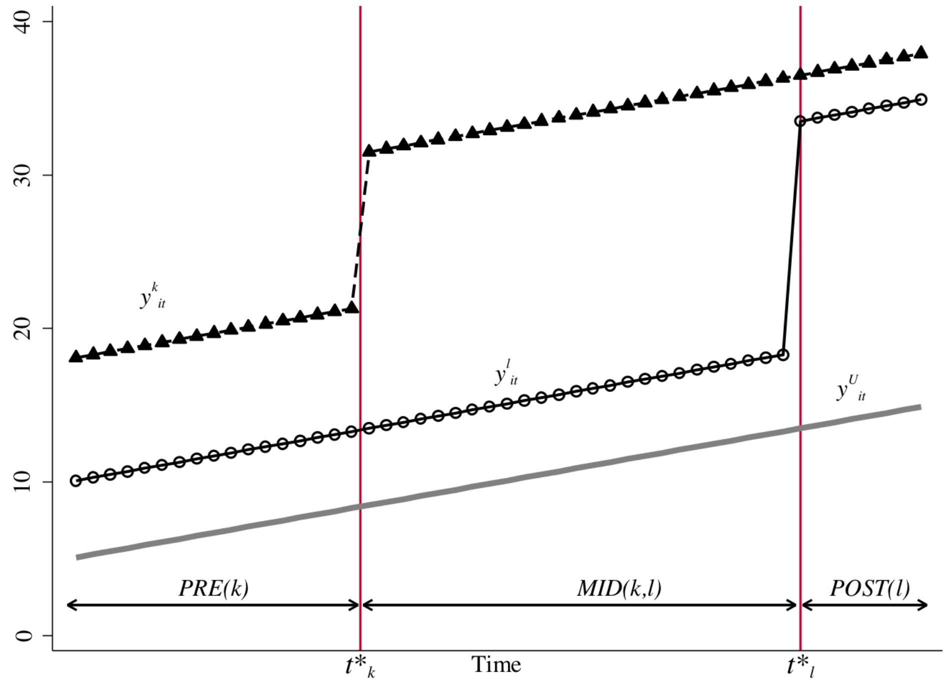
Let’s look at 3 timing groups (a, b and c) and one untreated group (U). With 3 timing groups, there are 9 2x2 DDs. Here they are:

a to b	b to a	c to a
a to c	b to c	c to b
a to U	b to U	c to U

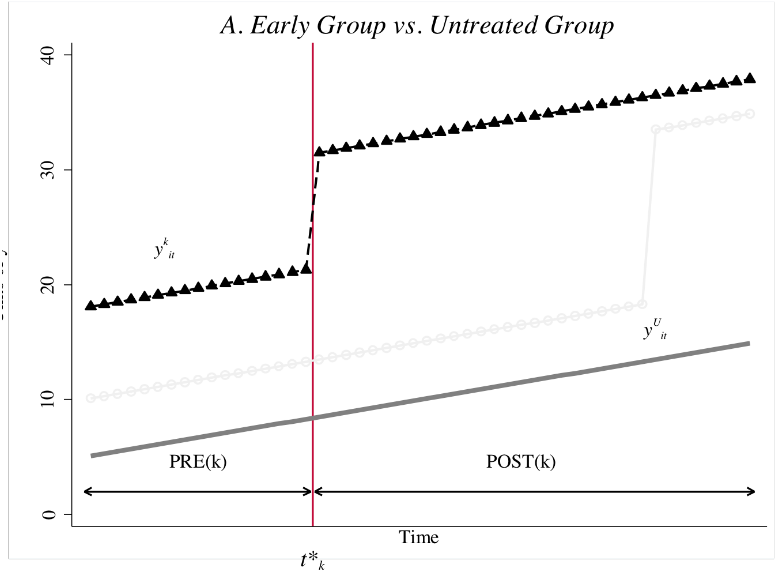
Let’s return to a simpler example with only two groups – a k group treated at t_k^* and an l treated at t_l^* plus an never-treated group called the U untreated group

Terms and notation

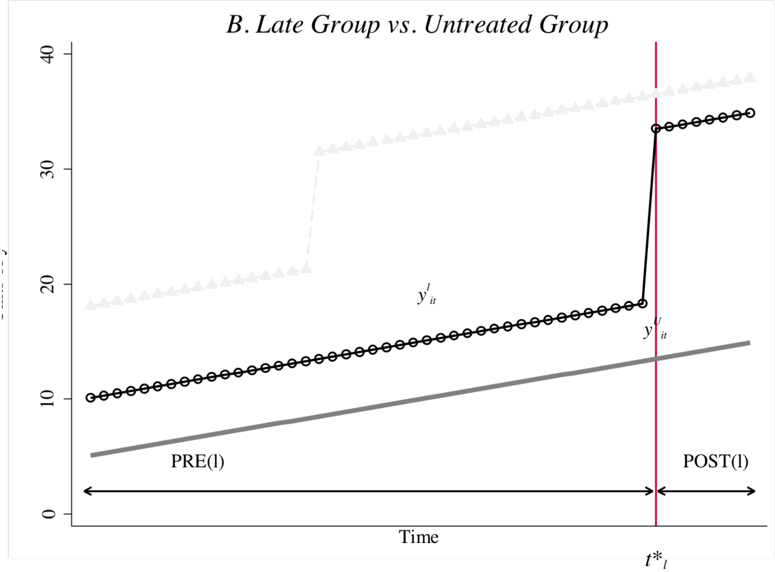
- Let there be two treatment groups (k,l) and one untreated group (U)
- k,l define the groups based on when they receive treatment (differently in time) with k receiving it earlier than l
- Denote $\overline{D}_k$ as the share of time each group spends in treatment status
- Denote $\widehat{\delta}^{2x2}_{jb}$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


40
30
20
10
0
y^k_{it}
y^l_{it}
y^U_{it}
PRE(k)
MID(k,l)
POST(l)
t^*_k
Time
t^*_l

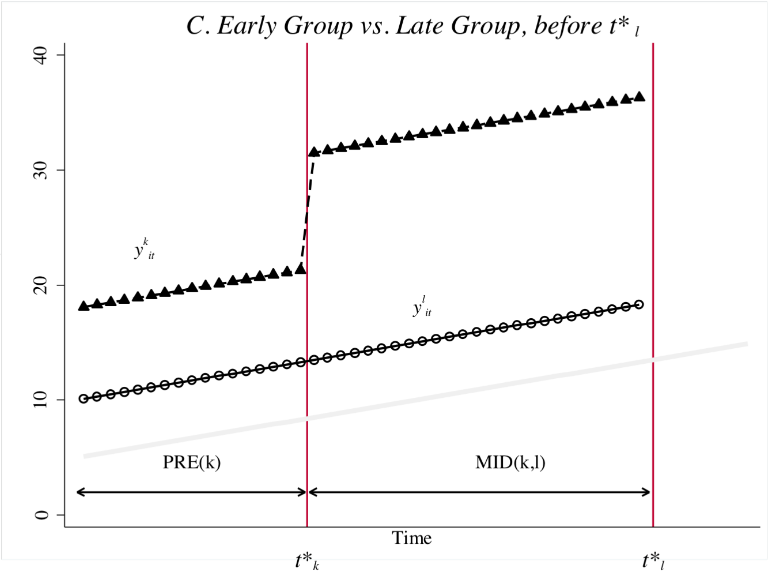
$$\widehat{\delta}^{2x2}_{kU} = \left(\overline{y}_k^{post(k)} - \overline{y}_k^{pre(k)}\right) - \left(\overline{y}_U^{post(k)} - \overline{y}_U^{pre(k)}\right)$$



$$\widehat{\delta}^{2x2}_{lU} = \left(\overline{y}_l^{post(l)} - \overline{y}_l^{pre(l)}\right) - \left(\overline{y}_U^{post(l)} - \overline{y}_U^{pre(l)}\right)$$

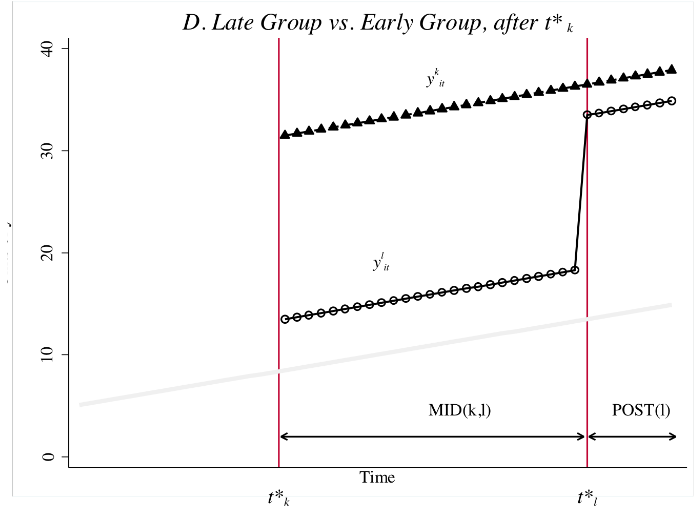

B. Late Group vs. Untreated Group

$$\delta_{kl}^{2x2,k} = \left(\overline{y}_k^{MID(k,l)} - \overline{y}_k^{Pre(k,l)}\right) - \left(\overline{y}_l^{MID(k,l)} - \overline{y}_l^{PRE(k,l)}\right)$$



C. Early Group vs. Late Group, before $t*_l$

$$\delta_{lk}^{2x2,l} = \left(\overline{y}_l^{POST(k,l)} - \overline{y}_l^{MID(k,l)}\right) - \left(\overline{y}_k^{POST(k,l)} - \overline{y}_k^{MID(k,l)}\right)$$



D. Late Group vs. Early Group, after $t*_k$

Bacon decomposition

$$Y_{ist} = \beta_0 + \delta D_{ist} + \tau_t + \sigma_s + \varepsilon_{ist}$$

TWFE estimate of $\widehat{\delta}$ is equal to a weighted average over all group 2x2 (of which there are 4 in this example)

$$\widehat{\delta}^{TWFE} = \sum_{k \neq U} s_{kU} \widehat{\delta}_{kU}^{2x2} + \sum_{k \neq U} \sum_{l > k} s_{kl} \left[\mu_{kl} \widehat{\delta}_{kl}^{2x2,k} + (1 - \mu_{kl}) \widehat{\delta}_{lk}^{2x2,l} \right]$$

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

Third, the Weights

$$\begin{aligned}
s_{ku} &= \frac{n_k n_u \overline{D}_k(1-\overline{D}_k)}{\widehat{Var}(\tilde{D}_{it})} \\
s_{kl} &= \frac{n_k n_l(\overline{D}_k-\overline{D}_l)(1-(\overline{D}_k-\overline{D}_l))}{\widehat{Var}(\tilde{D}_{it})} \\
\mu_{kl} &= \frac{1-\overline{D}_k}{1-(\overline{D}_k-\overline{D}_l)}
\end{aligned}$$

where n refer to the panel group shares, $\overline{D}_k(1-\overline{D}_k)$, as well as $(\overline{D}_k-\overline{D}_l)(1-(\overline{D}_k-\overline{D}_l))$ expressions refer to variance of treatment, and the final equation is the same for two timing groups.

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 s_{ku} weights.
  - → $\overline{D} = 0.1$. Then $0.1 \times 0.9 = 0.09$
  - → $\overline{D} = 0.4$. Then $0.4 \times 0.6 = 0.24$
  - → $\overline{D} = 0.5$. Then $0.5 \times 0.5 = 0.25$
  - → $\overline{D} = 0.6$. 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*

More weights discussion

- But what about the "treated on treated" weights (i.e., $\overline{D}_k - \overline{D}_l$)
- Same principle as before - when the difference between treatment variance is close to 0.5, those 2x2s are given the greatest weight
- For instance, say $t_k^* = 0.15$ and $t_l^* = 0.67$. Then $\overline{D}_k - \overline{D}_l = 0.52$. And thus $0.52 \times 0.48 = 0.2496$.

Summarizing TWFE centralities

- Groups in the middle of the panel weight up their respective 2x2s via the variance weighting
- Decomposition highlights the strange role of panel length when using TWFE
- Different choices about panel length change both the 2x2 and the weights based on variance of treatment

Back to TWFE

$$Y_{ist} = \beta_0 + \delta D_{ist} + \tau_t + \sigma_s + \varepsilon_{ist}$$

- So we know that the estimate is a weighted average over all "four averages and three subtractions" but is that good or bad?
- It's good if it's unbiased; it's bad if it isn't, and the decomposition doesn't tell us which unless we replace realized outcomes with potential outcomes
- Bacon shows that TWFE estimate of δ needs two assumptions for unbiasedness:
  1. variance weighted parallel trends are zero and
  2. 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 (not just weighted average of DiDs)

Moving from 2x2s to causal effects and bias terms

Let's start breaking down these estimators into their corresponding estimation objects expressed in causal effects and biases

$$\begin{aligned}
\widehat{\delta}^{2x2}_{kU} &= ATT_kPost + \Delta Y^0_k(Post(k), Pre(k)) - \Delta Y^0_U(Post(k), Pre) \\
\widehat{\delta}^{2x2}_{kl} &= ATT_k(MID) + \Delta Y^0_k(MID, Pre) - \Delta Y^0_l(MID, Pre)
\end{aligned}$$

These look the same because you're always comparing the treated unit with an untreated unit (though in the second case it's just that they haven't been treated *yet*).

The dangerous 2x2

But what about the 2x2 that compared the late groups to the already-treated earlier groups? With a lot of substitutions we get:

$$\begin{aligned}\widehat{\delta}_{lk}^{2x2} &= ATT_{l,Post(l)} + \underbrace{\Delta Y_l^0(Post(l), MID) - \Delta Y_k^0(Post(l), MID)}_{\text{Parallel trends bias}} \\ &\quad - \underbrace{(ATT_k(Post) - ATT_k(Mid))}_{\text{Heterogeneity bias!}}\end{aligned}$$

Substitute all this stuff into the decomposition formula

$$\widehat{\delta}^{TWFE} = \sum_{k \neq U} s_{kU} \widehat{\delta}^{2x2}_{kU} + \sum_{k \neq U} \sum_{l > k} s_{kl} \left[\mu_{kl} \widehat{\delta}^{2x2,k}_{kl} + (1 - \mu_{kl}) \widehat{\delta}^{2x2,l}_{kl} \right]$$

where we will make these substitutions

$$\begin{aligned}
\widehat{\delta}^{2x2}_{kU} &= ATT_k(Post) + \Delta Y^0_k(Post, Pre) - \Delta Y^0_U(Post, Pre) \\
\widehat{\delta}^{2x2,k}_{kl} &= ATT_k(Mid) + \Delta Y^0_k(Mid, Pre) - \Delta Y^0_l(Mid, Pre) \\
\widehat{\delta}^{2x2,l}_{lk} &= ATT_l Post(l) + \Delta Y^0_l(Post(l), MID) - \Delta Y^0_k(Post(l), MID) \\
&\quad - (ATT_k(Post) - ATT_k(Mid))
\end{aligned}$$

Notice all those potential sources of biases!

Potential Outcome Notation

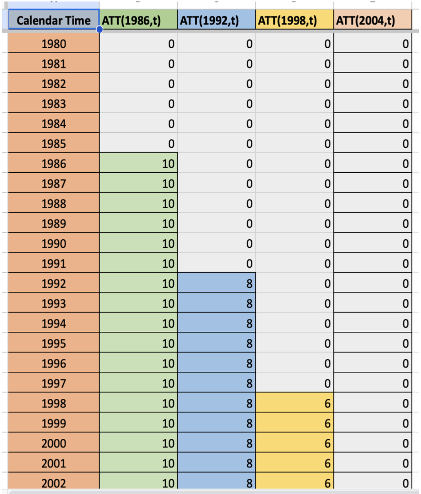
$$p\, lim\, \widehat{\delta}^{TWFE}_{n\to\infty} \quad = \quad VWATT + VWPT - \Delta ATT$$

- Notice the number of assumptions needed *even* to estimate this very strange weighted ATT (which is a function of how you drew the panel in the first place).
- With dynamics, it attenuates the estimate (bias) and can even reverse sign depending on the magnitudes of what is otherwise effects in the sign in a reinforcing direction!
- Model can flip signs (does not satisfy a “no sign flip property”)

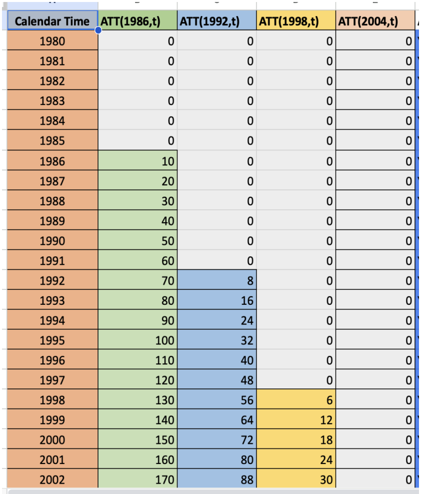
Simulated data

- 1000 firms, 40 states, 25 firms per states, 1980 to 2009 or 30 years, 30,000 observations, four groups
- I'll impose "unit level parallel trends", which is much stronger than we need (we only need average parallel trends)
- Also no anticipation of treatment effects until treatment occurs but does *not* guarantee homogenous treatment effects
- Two types of situations: constant versus dynamic treatment effects

Constant vs Dynamic Treatment Effects

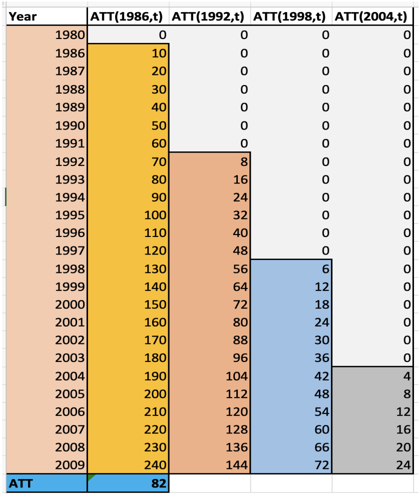


Calendar Time	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0
1981	0	0	0	0
1982	0	0	0	0
1983	0	0	0	0
1984	0	0	0	0
1985	0	0	0	0
1986	10	0	0	0
1987	10	0	0	0
1988	10	0	0	0
1989	10	0	0	0
1990	10	0	0	0
1991	10	0	0	0
1992	10	8	0	0
1993	10	8	0	0
1994	10	8	0	0
1995	10	8	0	0
1996	10	8	0	0
1997	10	8	0	0
1998	10	8	6	0
1999	10	8	6	0
2000	10	8	6	0
2001	10	8	6	0
2002	10	8	6	0



Calendar Time	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0
1981	0	0	0	0
1982	0	0	0	0
1983	0	0	0	0
1984	0	0	0	0
1985	0	0	0	0
1986	10	0	0	0
1987	20	0	0	0
1988	30	0	0	0
1989	40	0	0	0
1990	50	0	0	0
1991	60	0	0	0
1992	70	8	0	0
1993	80	16	0	0
1994	90	24	0	0
1995	100	32	0	0
1996	110	40	0	0
1997	120	48	0	0
1998	130	56	6	0
1999	140	64	12	0
2000	150	72	18	0
2001	160	80	24	0
2002	170	88	30	0

Group-time ATT



Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0
1986	10	0	0	0
1987	20	0	0	0
1988	30	0	0	0
1989	40	0	0	0
1990	50	0	0	0
1991	60	0	0	0
1992	70	8	0	0
1993	80	16	0	0
1994	90	24	0	0
1995	100	32	0	0
1996	110	40	0	0
1997	120	48	0	0
1998	130	56	6	0
1999	140	64	12	0
2000	150	72	18	0
2001	160	80	24	0
2002	170	88	30	0
2003	180	96	36	0
2004	190	104	42	4
2005	200	112	48	8
2006	210	120	54	12
2007	220	128	60	16
2008	230	136	66	20
2009	240	144	72	24
ATT	82			

- Heterogenous treatment effects across time and across groups
- Cells are called "group-time ATT" (Callaway and Sant'anna 2020) or "cohort ATT" (Sun and Abraham 2020)
- ATT is weighted average of all cells and $+82$ with uniform weights $1/60$

Estimation

Estimate the following equation using OLS:

$$Y_{ist} = \alpha_i + \gamma_t + \delta D_{it} + \varepsilon_{ist}$$

Table: Estimating ATT with different models

	Truth	**(TWFE)**	**(CS)**	**(SA)**	**(BJS)**
$\widehat{ATT}$	82	-6.69***			

The sign flipped. Why? Because of *extreme* dynamics (i.e., $-\Delta ATT$)

Bacon decomposition

Table: Bacon Decomposition (TWFE $= -6.69$)

DD Comparison	**Weight**	**Avg DD Est**
Earlier T vs. Later C	0.500	51.800
Later T vs. Earlier C	0.500	-65.180

T $=$ Treatment; C$=$ Comparison

$(0.5 * 51.8) + (0.5 * -65.180) = -6.69$

While large weight on the "late to early 2x2" is *suggestive* of an issue, these would appear even if we had constant treatment effects

Roadmap

Differential timing

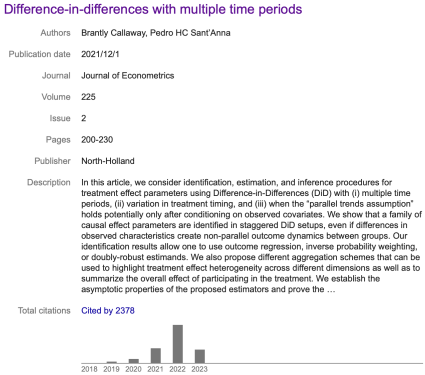
- Introduction
- TWFE Estimator
- Simulation

Aggregating building blocks

- CS

Callaway and Sant’Anna 2020

CS is a DiD estimator used for estimating and then summarizing smaller ATT parameters under differential timing and conditional parallel trends into more policy relevant ATT parameters (either dynamic or static)



Difference-in-differences with multiple time periods

Authors	Brantly Callaway, Pedro HC Sant’Anna
Publication date	2021/12/1
Journal	Journal of Econometrics
Volume	225
Issue	2
Pages	200-230
Publisher	North-Holland
Description	In this article, we consider identification, estimation, and inference procedures for treatment effect parameters using Difference-in-Differences (DiD) with (i) multiple time periods, (ii) variation in treatment timing, and (iii) when the “parallel trends assumption” holds potentially only after conditioning on observed covariates. We show that a family of causal effect parameters are identified in staggered DiD setups, even if differences in observed characteristics create non-parallel outcome dynamics between groups. Our identification results allow one to use outcome regression, inverse probability weighting, or doubly-robust estimands. We also propose different aggregation schemes that can be used to highlight treatment effect heterogeneity across different dimensions as well as to summarize the overall effect of participating in the treatment. We establish the asymptotic properties of the proposed estimators and prove the …
Total citations	Cited by 2378

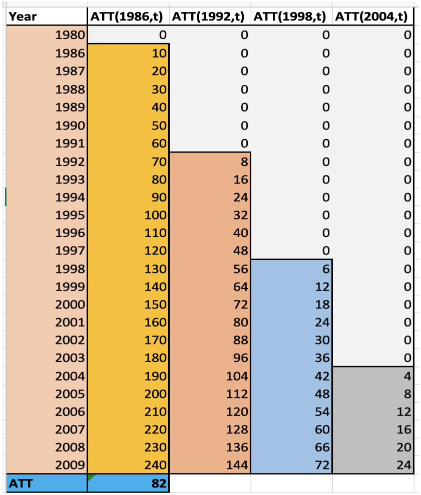
When is CS used

Just some examples of when you'd want to consider it:

1. When treatment effects differ depending on when it was adopted
2. When treatment effects change over time
3. When shortrun treatment effects are different than longrun effects
4. When treatment effect dynamics differ if people are first treated in a recession relative to expansion years

CS estimates the ATT by identifying smaller causal effects and aggregating them using non-negative weights

Group-time ATT



Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0
1986	10	0	0	0
1987	20	0	0	0
1988	30	0	0	0
1989	40	0	0	0
1990	50	0	0	0
1991	60	0	0	0
1992	70	8	0	0
1993	80	16	0	0
1994	90	24	0	0
1995	100	32	0	0
1996	110	40	0	0
1997	120	48	0	0
1998	130	56	6	0
1999	140	64	12	0
2000	150	72	18	0
2001	160	80	24	0
2002	170	88	30	0
2003	180	96	36	0
2004	190	104	42	4
2005	200	112	48	8
2006	210	120	54	12
2007	220	128	60	16
2008	230	136	66	20
2009	240	144	72	24
ATT	82			

Each cell contains that group's ATT(g,t)

$$ATT(g,t) = E[Y_t^1 - Y_t^0 | G_g = 1]$$

CS identifies all feasible ATT(g,t)

Group-time ATT

Group-time ATT is the ATT for a specific group and time

- Groups are basically cohorts of units treated at the same time
- Group-time ATT estimates are simple (weighted) differences in means
- Does not directly restrict heterogeneity with respect to observed covariates, timing or the evolution of treatment effects over time
- Allows us ways to choose our aggregations
- Inference is the bootstrap

Notation

- T periods going from $t = 1, \dots, T$
- Units are either treated ($D_t = 1$) or untreated ($D_t = 0$) but once treated cannot revert to untreated state
- G_g signifies a group and is binary. Equals one if individual units are treated at time period t.
- C is also binary and indicates a control group unit equalling one if "never treated" (can be relaxed though to "not yet treated")
  - → Recall the problem with TWFE on using treatment units as controls
- Generalized propensity score enters into the estimator as a weight:

$$\widehat{p(X)} = Pr(G_g = 1 | X, G_g + C = 1)$$

CS Estimator (the IPW version)

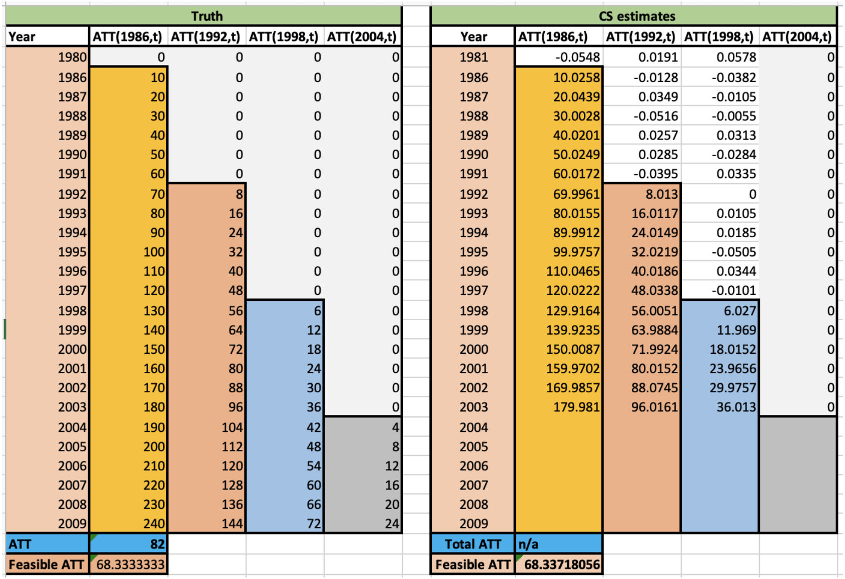
$$ATT(g,t) = E\left[\left(\frac{G_g}{E[G_g]} - \frac{\frac{\hat{p}(X)C}{1-\hat{p}(X)}}{E\left[\frac{\hat{p}(X)C}{1-\hat{p}(X)}\right]}\right)(Y_t - Y_{g-1})\right)\right]$$

This is the inverse probability weighting estimator. Alternatively, there is an outcome regression approach and a doubly robust. Sant'Anna recommends DR. CS uses the never-treated or the not-yet-treated as controls but never the already-treated

Aggregated vs single year/group ATT

- The method they propose is really just identifying very narrow ATT per group time.
- But we are often interested in more aggregate parameters, like the ATT across all groups and all times
- They present two alternative methods for building “interesting parameters”
- Inference from a bootstrap

Group-time ATT



	Truth			
Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1980	0	0	0	0
1986	10	0	0	0
1987	20	0	0	0
1988	30	0	0	0
1989	40	0	0	0
1990	50	0	0	0
1991	60	0	0	0
1992	70	8	0	0
1993	80	16	0	0
1994	90	24	0	0
1995	100	32	0	0
1996	110	40	0	0
1997	120	48	0	0
1998	130	56	6	0
1999	140	64	12	0
2000	150	72	18	0
2001	160	80	24	0
2002	170	88	30	0
2003	180	96	36	0
2004	190	104	42	4
2005	200	112	48	8
2006	210	120	54	12
2007	220	128	60	16
2008	230	136	66	20
2009	240	144	72	24
ATT	82			
Feasible ATT	68.3333333			

	CS estimates			
Year	ATT(1986,t)	ATT(1992,t)	ATT(1998,t)	ATT(2004,t)
1981	-0.0548	0.0191	0.0578	0
1986	10.0258	-0.0128	-0.0382	0
1987	20.0439	0.0349	-0.0105	0
1988	30.0028	-0.0516	-0.0055	0
1989	40.0201	0.0257	0.0313	0
1990	50.0249	0.0285	-0.0284	0
1991	60.0172	-0.0395	0.0335	0
1992	69.9961	8.013	0	0
1993	80.0155	16.0117	0.0105	0
1994	89.9912	24.0149	0.0185	0
1995	99.9757	32.0219	-0.0505	0
1996	110.0465	40.0186	0.0344	0
1997	120.0222	48.0338	-0.0101	0
1998	129.9164	56.0051	6.027	0
1999	139.9235	63.9884	11.969	0
2000	150.0087	71.9924	18.0152	0
2001	159.9702	80.0152	23.9656	0
2002	169.9857	88.0745	29.9757	0
2003	179.981	96.0161	36.013	0
2004				
2005				
2006				
2007				
2008				
2009				
Total ATT	n/a			
Feasible ATT	68.33718056			

Question: Why didn’t CS estimate all ATT(g,t)? What is “feasible ATT”?

Reporting results

Table: Estimating ATT using only pre-2004 data

	(Truth)	**(TWFE)**	**(CS)**	**(SA)**	**(BJS)**
$\widehat{Feasible\ ATT}$	68.33	26.81 ***	68.34***		

TWFE is no longer negative, interestingly, once we eliminate the last group (giving us a never-treated group), but is still suffering from attenuation bias.

Conclusion

- The previous methods are fairly comparable, but note, these models all assume parallel trends or conditional parallel trends
- Question is which *comparison group* is more sensible to use – the never-treated or the not-yet-treated?
- There is no single answer to that – the more the treatment was quasi-random, the more the never-treated is appealing
- But maybe the not-yet-treated is better as why didn't the never-treated adopt the treatment?
- Don't get lost in the decisions and forget the importance of *designing*, focus on treatment assignment mechanism, checking imbalance