# Difference-in-Differences

## Mixtape Session

# Roadmap

- Imputation DiD
  - Imputation based robust estimator
  - 2SDiD

# Imputation

Some methods are more obviously imputations than others though.

I will consider something explicit imputation if it constructs counterfactual observations at the unit level, as opposed to implicit imputation (e.g., manual aggregation) which tends to directly estimate ATT measures such as CS.

We will discuss three explicit imputation methods: Borusyak, Jaravel and Spiess (2021) robust imputation estimator and Gardner (2021) two stage DiD

# Background

- First“new did” paper was Borusyak and Jaravel (2017) – a lot of what was simultaneously discovered elsewhere was in that paper
- We will discuss its successor – Borusyak, Jaravel and Spiess (2021)
- My interpretation: damning critique of OLS TWFE and a robust solution based on explicit imputation

# My Outline (versus their outline)

1. Discussion of their interpretation of “basic” DiD assumptions
2. Critique of TWFE OLS when strong assumptions don’t hold
3. Introduction of new assumptions
4. Robust efficient imputation estimator

# Broad view

- Under three standard DiD assumptions, TWFE OLS performs fine
- No anticipation creates some challenges for event studies that requires tweaks
- But one of them (treatment effect homogeneity) introduces major problems
- Remember: theirs was the first to bring attention to what happens when treatment effect heterogeneity occurs
- After detailed critique of TWFE OLS, they roll out a robust estimator
- BLUE like characteristics

# What are we after?

A key flavor of the new DiD papers is not merely to assume TWFE OLS recovers “reasonable” weighted averages of treatment effects, but to begin by explicitly naming the target parameter. Under what assumptions can we identify $\tau_w$?

Estimation target:

$$\tau_w = \sum_{it \in \Omega_1} w_{it} \tau_{it} = w_1' \tau$$

Weights need not add up to one. Weights could be $\frac{1}{N}$ for all $it \in \Omega_1$.We have a number of options.

# A1: Parallel trends

**Assumption 1: Parallel trends**. There exist non-stochastic $\alpha_i$ and $\beta_t$ such that:

$$Y_{it}(0) = \alpha_i + \beta_t + \varepsilon_{it}$$

with

$$E[\varepsilon_{it}] = 0$$

for all $it \in \Omega$. Can be extended (e.g., unit-specific trends). Only imposes restrictions on $Y(0)$, not treatment effects themselves. Notice how it is a TWFE assumption – it's actually the same data generating process as in baker.do.

# A2: No anticipation

- We saw this with SA, but I think it occurred slightly earlier with BJ (not sure)
- No anticipation effects means there are no treatment effects prior to the event date

$$Y_{it} = Y_{it}(0)$$

  for all $it \in \Omega_0$.
- I think this is probably ruling out “Ashenfelter’s dip”
- It’s also an extension of SUTVA if I’m not mistaken because SUTVA requires that your outcome is a function of your current treatment status not your future treatment status

# A2: No anticipation (continued)

- Notice how as an assumption, it literally imposes $\tau = 0$ for all pre-treatment periods.
- They argue that "some form of this assumption is necessary for DiD identification" because otherwise you don't have a reference period
- Even before Goodman-Bacon (2021), Sun and Abraham (2020) and Borusyak and Jaravel (2017), I had seen a million applied papers and only seen references to PT, not NA
- It's oftentimes treated as an implicit assumption that can be then tested using an event study, but they'll discuss that as that confuses estimation with identification

# A3: Restricted causal effects

This is the one that places restrictions on what treatment effects can and cannot be (i.e., homogenous treatment effects). Notice the very detailed expression:

**Assumption 3 (Restricted causal effects)**: $B\tau 0$ for a known $M \times N_1$ matrix $B$ of full row rank.

If we can assume something like homogenous treatment effects, then TWFE actually is best because its ability to *correctly* extrapolate will increase efficiency. But it's when A3 is not tenable or not really ex ante justified by theory that we should be worried. There's an A3' that is a slight modification.

# Critique of Common Practice

1. Under-identification in event studies
2. Negative weighting
3. Spurious identification of long-run casual effects

# Critique: Underidentification problem

We saw some of this earlier with SA, but mind you, there was simultaneous discoveries and a chronology. This result was in BJ, for whatever that is worth to you.

**Lemma 1**: If there are no never-treated units, the path of [pre-treatment lead population regression coefficients] is not point identified in the fully dynamic OLS specification. In particular. adding a linear trend to this path $\{\tau_h + k(h+1)\}$ for any $k \in R$ fits the data equally well with the fixed effects coefficients appropriately modified.

In english, it means you're going to have a multicollinearity problem even worse than you thought when estimating the fully dynamic event study model (i.e., dropping only one lead for all base comparisons)

# Underidentification of lead coefficients

## Under-identification problem

Formally the problem arises because a linear time trend $t$ and a linear term in the cohort $E_i$ (subsumed by the unit FEs) can perfectly reproduce a linear term in relative time $K_{it} = t - E_i$. Therefore a complete set of treatment leads and lags, which is equivalent to the FE of relative time, is collinear with the unit and period FEs.

Just one additional normalization is needed – drop $\tau_{-a} = 0$ and $\tau_{-1} = 0$. This will break the multicollinearity. We saw this in SA also. So multiple people saw this at the same time.

# Under-identification and theoretical justifications

- Imposing any $-a$ lead and $-1$ lead to equal zero is somewhat ad hoc. Why those two and not some other two?
- Recall with SA – it mattered which ones you dropped because otherwise leads were contaminated
- This is again about NA – if you chose $-a$ and $-1$, then you had some theoretical reason to assume NA held for them and not some other periods
- Researchers need an *a priori* reason to justify which leads they drop ideally
- I had a great one – Craigslist didn't announce or advertise or communicate intentions to enter markets before they did. NA was guaranteed
- You may need to scrutinize this.

# Negative weighting and violations of A3

Heterogeneous treatment effects creating problems *again*

- It's assumption 3 – homogeneity – that BJS (and really the first paper, Borusyak and Jaravel) showed was a problem for traditional event studies
- And we saw that earlier with Sun and Abraham
- What happens is that with heterogeneity, the weights on the treatment effects can become negative

# Negative weighting

Assume some simple static model with a single dummy for treatment. Then they lay out a second lemma

**Lemma 2**: If A1 and A2 hold, then the estimand of the static OLS specification satisfies $\tau^{static} = \sum_{it \in \Omega_1} w_{it}^{OLS} \tau_{it}$ for some weights $w_{it}^{OLS}$ that do not depend on the outcome realizations and add up to one $\sum_{it \in \Omega_1} = 1$.

The static OLS estimand cannot be interpreted as a “proper” weighted average, as some weights can be negative.

# Simple illustration

*Table:* TWFE dynamics

| $E(y_{it})$ | $i = A$ | $i = B$ |
|---|---|---|
| t=1 | $\alpha_A$ | $\alpha_B$ |
| t=2 | $\alpha_A + \beta_2 + \delta_{A2}$ | $\alpha_B + \beta_2$ |
| t=3 | $\alpha_A + \beta_3 + \delta_{A3}$ | $\alpha_B + \beta_3 + \delta_{B3}$ |
| Event date | $E_i = 2$ | $E_i = 3$ |

Static: $\delta = \delta_{A2} + \frac{1}{2}\delta_{B3} - \frac{1}{2}\delta_{A3}$.

Notice the negative weight on the furthest lag. This is what you get when A3 is not satisfied..

# Short-run bias of TWFE

- TWFE OLS has a severe short-run bias
- the long-run causal effect, corresponding to the early treated unit A and the late period 3, enters with a negative weight (-1/2)
- The larger the effects in the long-run, the smaller the coefficient will be
- It's caused by "forbidden comparisons" (late to early treated) – we saw this with Goodman-Bacon (2021)
- Forbidden comparisons create downward bias on long-run effects with treatment effect heterogeneity, *but not with treatment effect homogeneity* – so it really is an A3 violation

# Spurious Long-Run Causal Effects

More A3 problems, this time finding long-run effects where there are none. Basically, you need to impose a lot of pre-trend restrictions to get estimates of long-run population regression coefficients. Even then you can't get them all.

OLS estimates are fully driven by unwarranted extrapolations of treatment effects across observations and may not be trusted unless strong ex ante justifications for A3 exist

**Lemma 4**: Suppose there are no never-treated units and let $H = max_i E_i - min_i E_i$. Then for any non-negative weights $w_{it}$ defined over the set of observations with $K_{it} \geq \overline{H}$ (that are not identically zero), the weighted sum of causal effects $\sum_{it:K_{it} \geq \overline{H}} w_{it} \tau_{it}$ is not identified by A1 and A2.

# Modifications of general model

Modification of A1 to A1':

$$Y_{it}(0) = A'_{it}\lambda_i + X'_{it}\delta + \varepsilon_{it}$$

Assumption 4 is introduced (homoskedastic residuals). This is key, because they will be building an "efficient estimator" with BLUE like OLS properties.

Using A1' to A4, we get the "efficient estimator" which is for all linear unbiased estimates of $\delta_W$, the unique efficient estimator $\widehat{\delta^*_W}$ can be obtained with 3 steps

# Role of the untreated observations

*"At some level, all methods for causal inference can be viewed as imputation methods, although some more explicitly than others." – Imbens and Rubin (2015)*

*"The idea is to estimate the model of $Y_{it}^0$ using the untreated observations and extrapolate it to impute $Y_{it}^0$ for treated observations."*

# Steps

1. Estimate expected potential outcomes using OLS and only the untreated observations (this is similar to Gardner 2021)
2. Then calculate $\widehat{\delta}_{it} = Y_{it}^1 - \widehat{Y}_{it}^0$
3. Then estimate target parameters as weighted sums

$$\widehat{\delta}_W = \sum_{it} w_{it} \widehat{\delta}_{it}$$

# Why is this working?

- Think back to that original statement of the PT assumption – you're modeling $Y(0)_{it}$.
- That is, without treatment – so the potential outcomes do not depend on any treatment effect
- Hence where we get treatment heterogeneity
- We obtain consistent estimates of the fixed effects which are then used to extrapolate to the counterfactual units for all $Y(0)_{it \in \Omega_1}$
- I think this is a very cool trick personally, and as it is still OLS, it's computationally fast and flexible to unit-trends, triple diff, covariates and so forth (though remember what we said about covariates)

# Testing for parallel trends

- Perform pre-trend testing using untreated sample only
- This separation is preferable conceptually because it presents the conflation of using an identification assumption and validating it
- Traditional regression-based tests use the full sample, including the treated observations though
- Therefore it is not a test for A1 and A2; rather it is a joint test that is also sensitive to A3
- BJS test uses the untreated observations for which $Y_{it}^0$ is ok under A2

# Test

1. Choose an alternative model for $Y_{it}^0$ richer than A1

$$Y_{it}^0 = A_{it}'\lambda_i + X_{it}'\beta + w_{it}'\delta + \tilde{\varepsilon}_{it}$$

2. Estimate $\delta$ with $\widehat{\delta}$ using OLS on untreated units only
3. Test $\delta = 0$ using F-test or visually

# Comparisons to other estimators

Table 3: Efficiency and Bias of Alternative Estimators

| Horizon | Estimator | Baseline simulation | | More pre-periods | Heterosk. residuals | AR(1) residuals | Anticipation effects |
|---|---|---|---|---|---|---|---|
| | | Variance (1) | Coverage (2) | Variance (3) | Variance (4) | Variance (5) | Bias (6) |
| $h = 0$ | Imputation | 0.0099 | 0.942 | 0.0080 | 0.0347 | 0.0072 | -0.0569 |
| | DCDH | 0.0140 | 0.938 | 0.0140 | 0.0526 | 0.0070 | -0.0915 |
| | SA | 0.0115 | 0.938 | 0.0115 | 0.0404 | 0.0066 | -0.0753 |
| $h = 1$ | Imputation | 0.0145 | 0.936 | 0.0111 | 0.0532 | 0.0143 | -0.0719 |
| | DCDH | 0.0185 | 0.948 | 0.0185 | 0.0703 | 0.0151 | -0.0972 |
| | SA | 0.0177 | 0.948 | 0.0177 | 0.0643 | 0.0165 | -0.0812 |
| $h = 2$ | Imputation | 0.0222 | 0.956 | 0.0161 | 0.0813 | 0.0240 | -0.0886 |
| | DCDH | 0.0262 | 0.958 | 0.0262 | 0.0952 | 0.0257 | -0.1020 |
| | SA | 0.0317 | 0.950 | 0.0317 | 0.1108 | 0.0341 | -0.0850 |
| $h = 3$ | Imputation | 0.0366 | 0.928 | 0.0255 | 0.1379 | 0.0394 | -0.1101 |
| | DCDH | 0.0422 | 0.930 | 0.0422 | 0.1488 | 0.0446 | -0.1087 |
| | SA | 0.0479 | 0.952 | 0.0479 | 0.1659 | 0.0543 | -0.0932 |
| $h = 4$ | Imputation | 0.0800 | 0.942 | 0.0546 | 0.3197 | 0.0773 | -0.1487 |
| | DCDH | 0.0932 | 0.950 | 0.0932 | 0.3263 | 0.0903 | -0.1265 |
| | SA | 0.0932 | 0.954 | 0.0932 | 0.3263 | 0.0903 | -0.1265 |

*Notes:* See Section 4.6 for a detailed description of the data-generating processes and reported statistics.

# Returning to the minimum wage

- Now we can return to the minimum wage study from earlier (Clemens and Strain 2021)
- Recall that stacked regression had found large negative effects on employment when minimum wage increases were large, but not when they were small
- The authors also implemented the BJS imputation estimator
- One comment abt the following graphics: BJS procedure does not have a “base” period in the same sense as the regression models do because it is not contrasting each period relative to some omitted group
- Rather it is imputing counterfactuals, and therefore we can calculate each period’s effect

# BJS Results

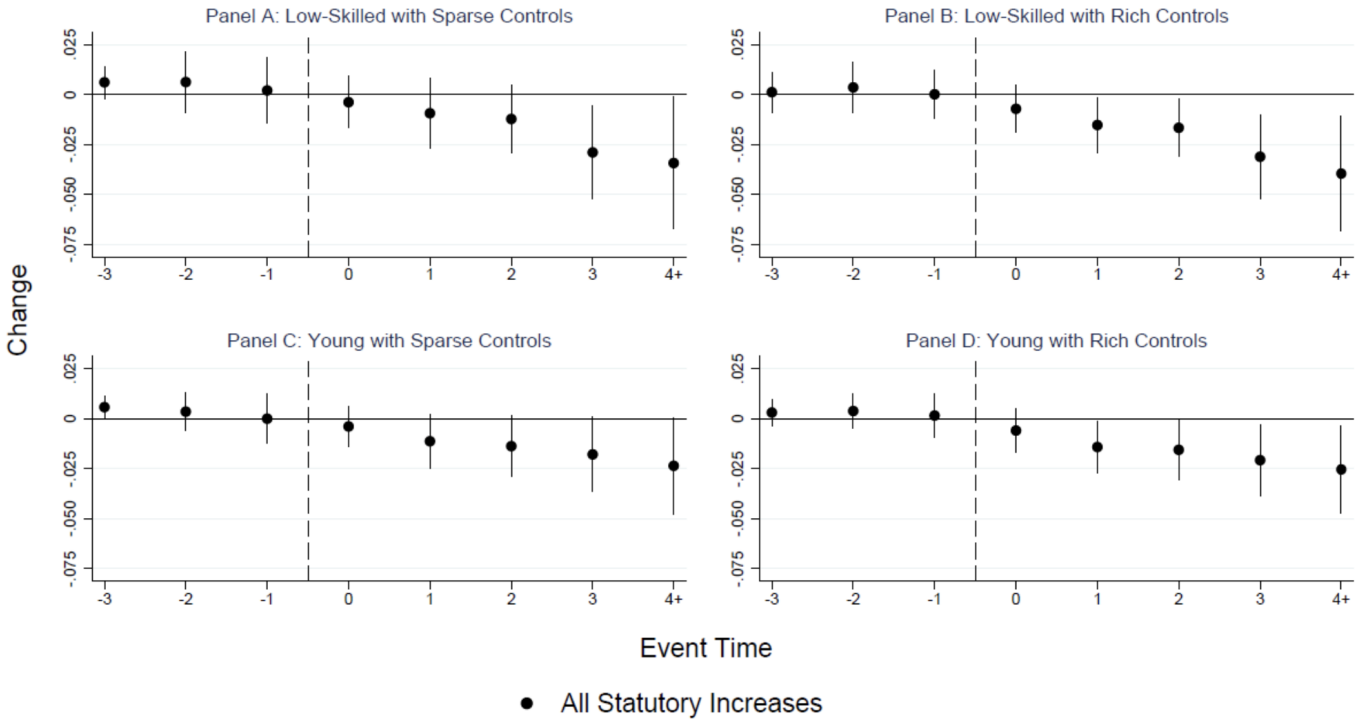


**Figure 11. Event Studies of Changes in Employment Following Statutory Minimum Wage Increases Using the BJS Imputation Estimator:** This figure displays coefficients obtained using the imputation estimator proposed by Borusyak, Jaravel and Spiess (2021) (BJS). For the BJS estimator, we code the first treatment year as the year in which a state's first statutory minimum wage increase took effect. Note that this appears graphically as "year 0" in the BJS figures, but corresponds with year 1 in the stacked event study figures. Panels A and B plot coefficients for low-skilled individuals defined as individuals ages 16–25 without a completed high school education. Panels C and D plot coefficients for young individuals defined as all individuals ages 16–21. The samples are from the ACS. Regressions with "sparse controls" include state and year fixed effects, as well as the log of annual average *per capita* income and the annual average state house price index used in our main regressions. Regressions with "rich controls" include all controls in the base controls regressions plus the three-year lag of log *per capita* income and the house price index, as well as a dummy variable for each education group and age. Error bars denote 95 percent confidence intervals around each estimated coefficient. Standard errors are clustered by state.

# BJS Results

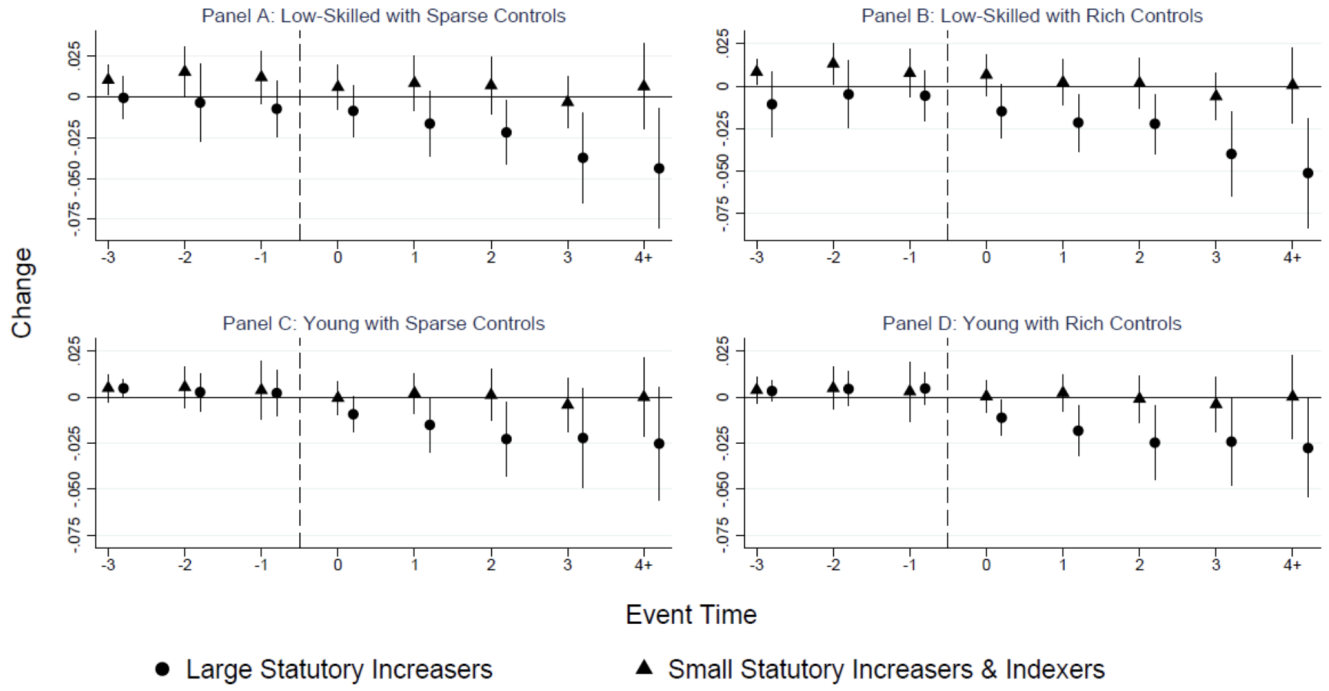


**Figure 12. Event Studies of Changes in Employment Following Large and Small Statutory Minimum Wage Increases Using the BJS Imputation Estimator:** This figure displays coefficients obtained using the imputation estimator proposed by Borusyak, Jaravel and Spiess (2021) (BJS). For the BJS estimator, we code the first treatment year as the year in which a state's first statutory minimum wage increase took effect. Note that this appears graphically as "year 0" in the BJS figures but corresponds with year 1 in the stacked event study figures. We compare estimates for large vs. small increases as defined in the main text. Panels A and B plot coefficients for low-skilled individuals defined as individuals ages 16–25 without a completed high school education. Panels C and D plot coefficients for young individuals defined as all individuals ages 16–21. The samples are from the ACS. Regressions with "sparse controls" include state and year fixed effects, as well as the log of annual average *per capita* income and the annual average state house price index used in our main regressions. Regressions with "rich controls" include all controls in the base controls regressions plus the three-year lag of log *per capita* income and the house price index, as well as a dummy variable for each education group and age. Error bars denote 95 percent confidence intervals around each estimated coefficient. Standard errors are clustered by state.

# Comments abt the minimum wage study

- Elasticity of employment with respect to minimum wage is -0.124 and -0.082 for those without high school and the young, respectively
- Differences by size of minimum wage increase:
  - → Large increases (around $2.90): own-wage elasticity is -1.01 for 16-25yo with less than HS and -0.41 for 16 to 21yo (large effects)
  - → Small increases (around $1.90): own-wage elasticity is 0.46 (i.e., no employment effects)
  - → Inflation-index increases (around $0.90): own-wage elasticity is 0.16 (no effect) and -0.17 (no effect)

# Concluding remarks about the minimum wage study

Clemens and Strain (2021) illustrates three things:

1. Sometimes theory may predict heterogenous effects which requires researchers explore such theoretically motivated heterogeneity
2. Since p-hacking is commonly associated with heterogeneity subsample analysis, we can partially protect against it through pre-registration
3. Robust DiD estimators should be used to double check for problems with TWFE when using DiD designs with differential timing

Reassuring that results are consistent across all models used. Do not count the minimum wage debate to be finished.

# 2SDiD

- I'd like to go back to a more traditional form of analysis by reviewing Gardner (2021)
- Like a few other papers, Gardner (2021) is both a diagnosis of the illness and a cure, and I'm putting his cure into an explicit imputation framework
- John Gardner is an assistant professor and applied econometrician at University of Mississippi – smart, cool, and former colleague of Brant Callaway of Callaway and Sant'Anna
- The cure will be nicely called two-stage difference-in-differences (2SDiD) – Nice name!

# Highlights

- Why does TWFE fail under differential timing? Violates strict exogeneity under heterogeneity
- The logic of the failure suggests an obvious, but previously unknown, solution which is the 2SDiD
- I'll explain 2SDiD, focus on the parallel trends implications, and show we can get a consistent and unbiased estimate of group and relative time fixed effects
- If you can get consistent and unbiased estimates of group and relative time fixed effects, then you can delete them and run normal analysis
- We'll work through some code

# Background

- By now, we all agree that TWFE just doesn't handle heterogeneity under differential timing very well
- We've seen in the Goodman-Bacon decomposition why – it's caused by TWFE implicitly calculating late to early 2x2s, which are a source of bias
- But some of you are coming straight from a panel econometrics course that maybe didn't use potential outcomes notation
- Isn't strict exogeneity enough for consistent estimates? What then does strict exogeneity have to do with heterogeneity and differential timing?
- Everything

# More background

> *“It seems natural that TWFE should identify the ATT” – Gardner (2021)*

It just seems like TWFE with a DiD will estimate the ATT with weights that we’ll find intuitive. Was this just a conjecture and was never true? Why isn’t this working?

# High level discussion

- TWFE identifies the ATT when the heterogeneous effects are distributed equally across all groups and periods, but since that is a knife-edge situation, it is likely that TWFE will not in our applications meet this special scenario
- In the two group case, that is what happens though which is why TWFE worked fine there
- Metaphorically, the two group case that we always used to pin our intuition of what DiD was doing was the exception not the rule
- Goodman-Bacon (2021) shows the problem is caused by late-to-early comparisons; Gardner (2021) will show that the problem is misspecification
- Think of these as different perspectives on the same problem

# Model misspecification

> *"Misspecified DiD regression models project heterogenous treatment effects onto group and period fixed effects rather than the treatment status itself"*

Spoiler: This analysis of the problem suggests solution – why don't we remove those?

# 2SDiD

"What's the name of that kid from Mexico?" – Ted Lasso
"Dani Rojas" – Nate the Great
"Great name" – Ted Lasso

- Two stage DiD is a great name because of its connection to that classic IV model 2SLS
- If you can link it to 2SLS in your mind, it may help you because it'll show you that Gardner's model is a two stage model
- First stage – estimate the group and relative time fixed effects using only the $D = 0$ observations
- Second stage – using predicted values based off those fixed effect coefficients, run your model off the transformed outcome
- Get the standard errors right just like 2SLS by taking the first stage into account

# More high level

- The second step recovers the average difference in outcomes between treated and untreated units after removing group and period fixed effects
- What I like about Gardner's method is its pleasant familiarity, its speed
- But note, it's not going to allow you to do the kind of heterogeneity analysis that CS allows for
- Some of the differences will be due to slightly different PT assumptions, and some will because 2SDID will be using all of the data for analysis, not just the baseline for calculating the DID estimates

# Notation

$i$: panel units
$t$: calendar time – think of real dates
$g \in \{0, 1, \ldots, G\}$ – groups
$p \in \{0, 1, \ldots, P\}$ – relative time or “periods”

Periods are successive. Group 0 – never treated. Group 1 – treated in period 1, 2, and on. Group 2 – treated in period 2, etc.

# Parameters

$$\beta_{gp} = E\left[Y^1_{gpit} - Y^0_{gpit} | g, p\right]$$

It's a group-time ATT but expressed in a more traditional econometric notation that you could easily find in Wooldridge or some such

# Modeling basics

Under parallel trends, mean outcomes will satisfy the following equation

$$E\left[Y_{gpit}|g,p,D_{gp}\right] = \lambda_g + \gamma_p + \beta_{gp}D_{gp}$$

In two-group, group and period effects are eliminated with dummies because TWFE uses dummies to demean across multiple dimensions. Then TWFE identifies ATT. But this does not hold when average effects vary across group and period. There are many ways to express a treatment effect's across group and time, but Gardner presented it as a weighted average of the coefficients for only that group-period situation:

$$E\left(\beta_{gp}|D_{gp}=1\right) = E\left(Y^1_{gpit} - Y^0_{gpit}|D_{gp}=1\right)$$

# Strict exogeneity violation

Rewriting the above we get:

$$
\begin{aligned}
E\left[Y_{gpit}|g,p,D_{gp}\right] &= \lambda_g + \gamma_p + E\left[\beta_{gp}|D_{gp}=1\right]D_{gp} \\
&\quad \left[\beta_{gp} - E(\beta_{gp}|D_{gp}=1)\right]D_{gp}
\end{aligned}
$$

The problem is there's this weird new error term and it isn't mean zero under heterogenous treatment effects spread across group and period. Unlike the two group case, the coefficient on $D_{gp}$ from TWFE doesn't identify the average $E(\beta_{gp}|D_{gp}=1)$

So let's see Gardner's solution, but note – his solution was suggested by the problem itself. Gardner is thoughtful and observant.

# DiD regression estimand

- So if TWFE isn't recovering $E(\beta_{gp}|D_{gp} = 1)$, then what is it recovering?
- He shows that under PT, the coefficient on $D_{gp}$ is:

$$\beta^* = \sum_{g=1}^{G} \sum_{p=g}^{P} w_{gp} \beta_{gp}$$

- So then – what are the weights $w_{gp}$?
- Groan – It's a huge mess, and I hate even showing it to you because I find the weights almost impossible to decipher, but maybe you'll have a better go at it than me

# Weights

$$w_{gp} = \frac{\left\{[1 - P(D_{gp} = 1|g)] - [P(D_{gp} = 1|p) - P(D_{gp} - 1)]\right\} P(g,p)}{\sum_{g=1}^{G} \sum_{p=g}^{P} \left\{[1 - P(D_{gp} = 1|g)] - [P(D_{gp} = 1|p) - P(D_{gp} = 1)]\right\} P(g,p)}$$

Terms:

- $P(D_{gp} = 1|p)$: share of units treated in period $p$
- $P(D_{gp} = 1|g)$: share of periods in which $g$ is treated
- $P(D_{gp} = 1)$: share of unit $\times$ time treated
- $P(g,p)$: population share of observation corresponding to group $g$ and period $p$

I thought about changing all those probabilities into means, but honestly, it really didn't help me at all. But Gardner notes that this is from theorem 1 of deChaisemartin and D'Haultfoeiller (2020) and his Appendix A

# Estimation

$$Y_{gpit} = \lambda_g + \gamma_p + \beta D_{gp} + \varepsilon_{gpit}$$

This specification assumes a conditional expectation function that is linear in group, period and treatment status. But when the model is misspecified, it will attribute some of the heterogeneity impacts of the treatment to group and period fixed effects. The longer the treatment, the greater $\overline{D}$ is, the more that group's treatment effects will be absorbed by group fixed effects. When misspecified, TWFE doesn't recover $E[\beta|D = 1]$.

# Statistical issues

- Common support: “as long as there are untreated and treated observations for each group and period, $\lambda_g$ and $\gamma_p$ are identified from the subpopulation of untreated groups and periods.”
- Identification: “the overall group $\times$ period ATT is identified from a comparison of mean outcomes between treated and untreated groups after removing group and period effects.”

# Estimation: First stage

First stage:

$$Y_{gpit} = \lambda_g + \gamma_p + \varepsilon_{gpit}$$

using only $D_{gp} = 0$, retaining the fixed effects. Collect the $\widehat{\lambda_g}$ and $\widehat{\gamma_p}$.

# Estimation: Second stage

Second stage:

$$\begin{aligned} \widehat{y}_{gpit} &= y_{gpit} - \widehat{\lambda_g} - \widehat{\gamma_p} \\ \widehat{y}_{gpit} &= \alpha + \beta D_{gp} + \psi_{gpit} \end{aligned}$$

Why does this work? Parallel trends assumption implies:

$$E(y_{gpit}|g,p,D_{gp}) - \lambda_g - \gamma_p = E\Big[\beta_{gp}|D_{gp}=1\Big]D_{gp} + \Big[\beta_{gp} - E(\beta_{gp}|D_{gp}=1)\Big]D_{gp}$$

But because

$$E\Big\{[\beta_{gp} - E(\beta_{gp}|D_{gp}=1)]D_{gp}|D_{gp}\Big\} = 0$$

# Estimand

Then this procedure will identify $E(\beta_{gp}|D_{gp}=1)$. Consistency and unbiasedness proofs.

This is $E(\beta_{gp}|D_{gp}=1) = \sum^{G}\sum^{P}\beta_{gp}P(g,p|D_{gp}=1)$. It will tend to put more weight, by definition, on groups earlier into their treatment. But this isn't the same as the negative weighting that BJS say occurs oof the long lags. It just means there are more of them.

Event studies are:

$$y_{gpit} = \lambda_g + \gamma_p + \sum_{r=-R}^{P}\beta_r D_{rgp} + \varepsilon_{gpit}$$

Just change the second stage with the transformed outcome.

# Inference

- Standard errors are wrong on the second stage because the dependent variable uses estimates obtained from the first stage.
- The asymptotic distribution of the second stage can be obtained by interpreting the two-stage procedure as a joint GMM