Differences in Differences

Compass-Lexecon

Raquel Carrasco

Universidad Carlos III de Madrid

May-June 2024

Outline Session 4

- DD with multiple time periods and differential timing
  - Problems with TWFE Regressions
    - Goodman-Bacon (2021) decomposition

DD with staggered timing

- Canonical DD: units are treated at the same time
- But when the treatment timing differs between units, what do we compare with what?
- If we have staggered treatment adoption, it is tempting to use variations of the following TWFE specification:

$$Y_{it} = \alpha_i + \alpha_t + \beta D_{it} + \varepsilon_{it},$$

  where D_{it} is an indicator for unit i being treated by period t.
- Does β recover any interesting causal parameter of interest?
  - When the treatment effect is constant across time and units, β is the ATT.
  - But, if treatment effect is heterogeneous, β does not recover an easy-to-interpret parameter:
    - weighted average of ATT's, but some weights on treatment effects for some units and time periods can be negative and probably this does not make sense
  - Goodman-Bacon (2021) provides the most popular explanation.

DD with staggered timing

- The intuition for this problematic interpretation is that the TWFE specification combines two sources of comparisons:
  - Clean comparisons: DD's between treated and not-yet-treated units
  - Forbidden comparisons: DD's between two sets of already-treated units (who began treatment at different times)
- These forbidden comparisons can lead to negative weights: the "control group" is already treated, so we run into problems if their treatment effects change over time.

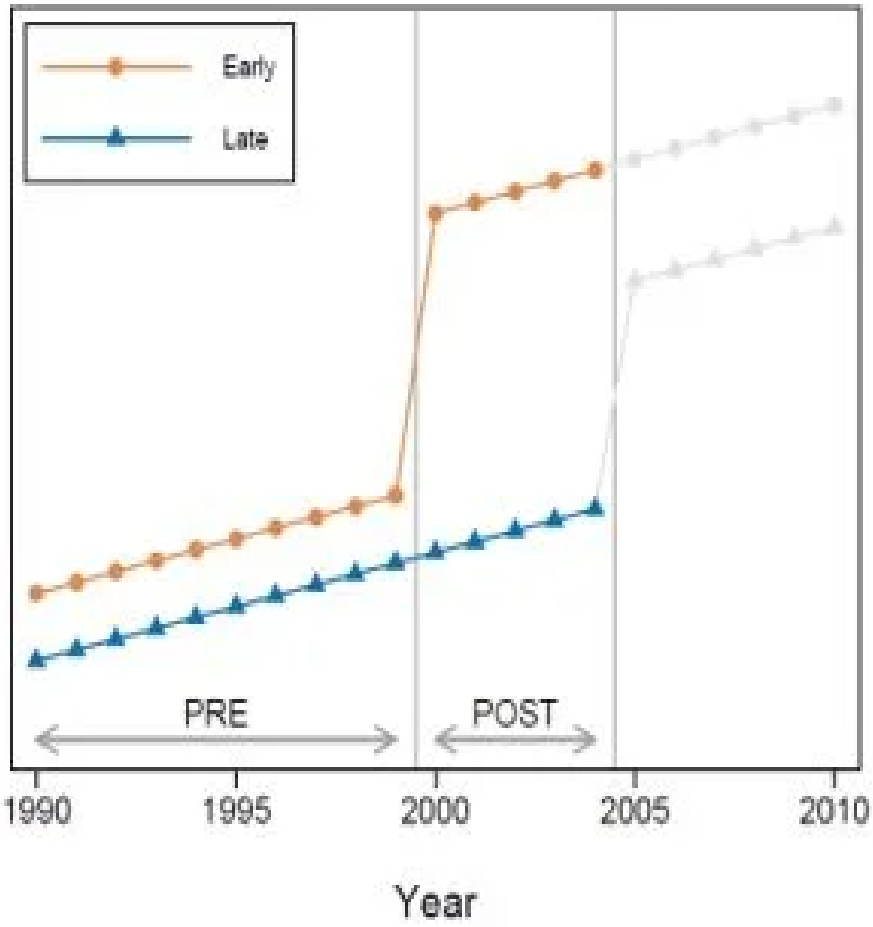
DD with staggered timing

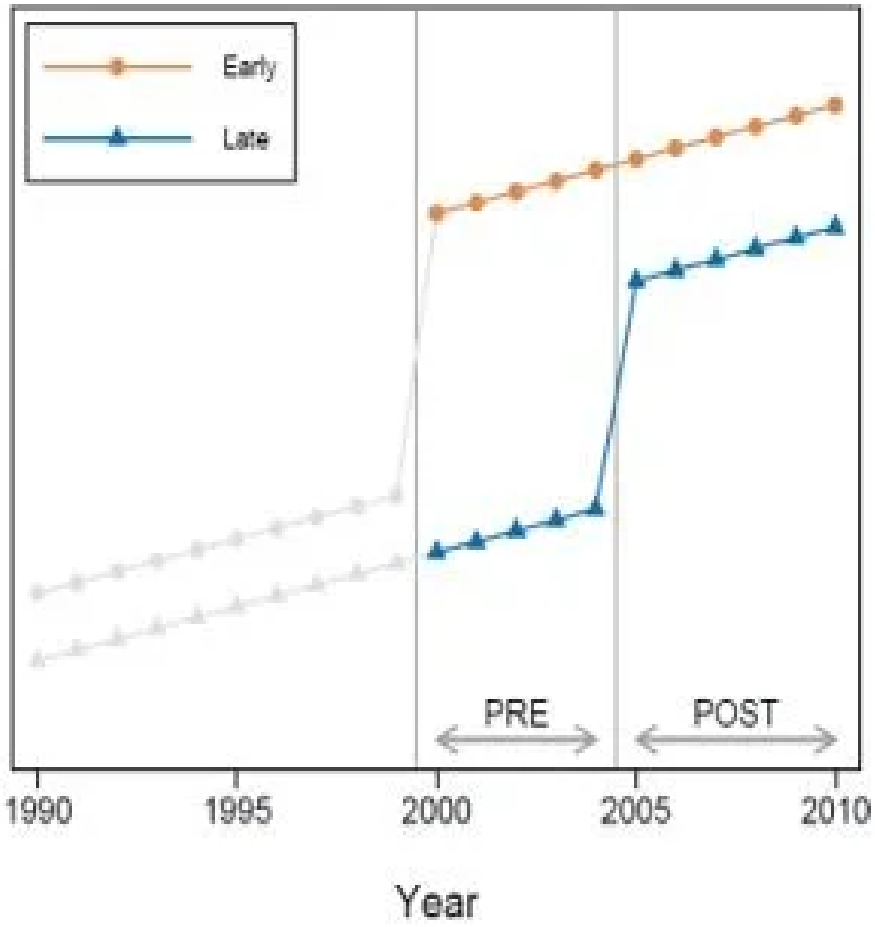
- Goodman-Bacon (2021) shows that any TWFE estimate of DD can be decomposed into a weighted average of all possible two-by-two DD estimators that can be constructed.
- Consider a (hypothetical) data set comprising two types of countries: one group that made primary education free in 2000, and another group that made primary education free in 2005. Call the first group the "early" adopters and the second group the "late" adopters.
- Suppose you have data on primary school completion rates every year from 1990 to 2010. Such a data set allows for two distinct DD comparisons:

DD with staggered timing

- First, we could focus on the period from 1990 to 2004 (the left panel in the figure below):
  - In that time frame, the late adopter countries are "never treated" in the sense that they do not implement free primary education – so they can be used as a comparison group to estimate the impact of free primary on test scores in the early adopter countries.
- However, we can also construct a second DD estimate of the treatment effect of free primary school by focusing on the years 2001 to 2010.
  - During that period, the treatment status of the early adopters never changes – they remain treated throughout – so they can be used as a comparison group to estimate the impact of free primary on test scores in the late adopting countries.

DD with staggered timing





In this example, the effect of treatment is constant within each group

DD with staggered timing

- Goodman-Bacon shows that any TWFE estimate of DD is a weighted average of:
  (1) Comparisons between (relatively) early adopters and later adopters over the periods when the later adopters are not yet treated
  (2) Comparisons between early adopters and later adopters over the periods when the early adopters are treated – so that they can be used as a comparison group for the later adopters, and
  (3) Comparisons between different timing groups (e.g., early adopters or later adopters) and the never-treated group, if there is one.
- This decomposition has several important implications, in particular TWFE can be severely biased.

DD with staggered timing

- Goodman-Bacon (2021) points to several challenges for estimation and interpretation:
  - Late adopters are a control group for early adopters
  - But early adopters are also a control group for late adopters
  - Heterogeneous treatment effects may lead to severe bias
- The decomposition shows:
  - DD estimator is a (strange) weighted average of 2×2 comparisons (sometimes with "negative weights")
  - But even if the weights are non-negative, they might not give us the most intuitive parameter
- Stata command to implement the Goodman-Bacon's DD decomposition: `bacondecomp`
- Goodman-Bacon decomposition as diagnosis of the problem

The Bacon Decomposition: Bias in the "standard" TWFE estimators

- A 2x2 is a simple difference over time between a treated group, k, and an untreated group, u:

$$\beta_{ku}^{2x2} = [E(Y_k \mid post) - E(Y_k \mid pre)] - [E(Y_u \mid post) - E(Y_u \mid pre)]$$

- Non-parallel trend bias:

$$\begin{aligned} \beta_{ku}^{2x2} &= \Big[E\left(Y_k^1 \mid post\right) - E(Y_k^0 \mid post\Big)\Big] \\ &+ \underbrace{\Big[E\left(Y_k^0 \mid post\right) - E(Y_k^0 \mid pre\Big)\Big] - \Big[E\left(Y_u^0 \mid post\right) - E(Y_u^0 \mid pre\Big)}_{\text{Non Parallel Trend bias}} \end{aligned}$$

- With parallel trends assumption, we get unbiased estimate of the ATT.
- But this is only the case for the simple 2x2.

The Bacon Decomposition: Bias in the "standard" TWFE estimators

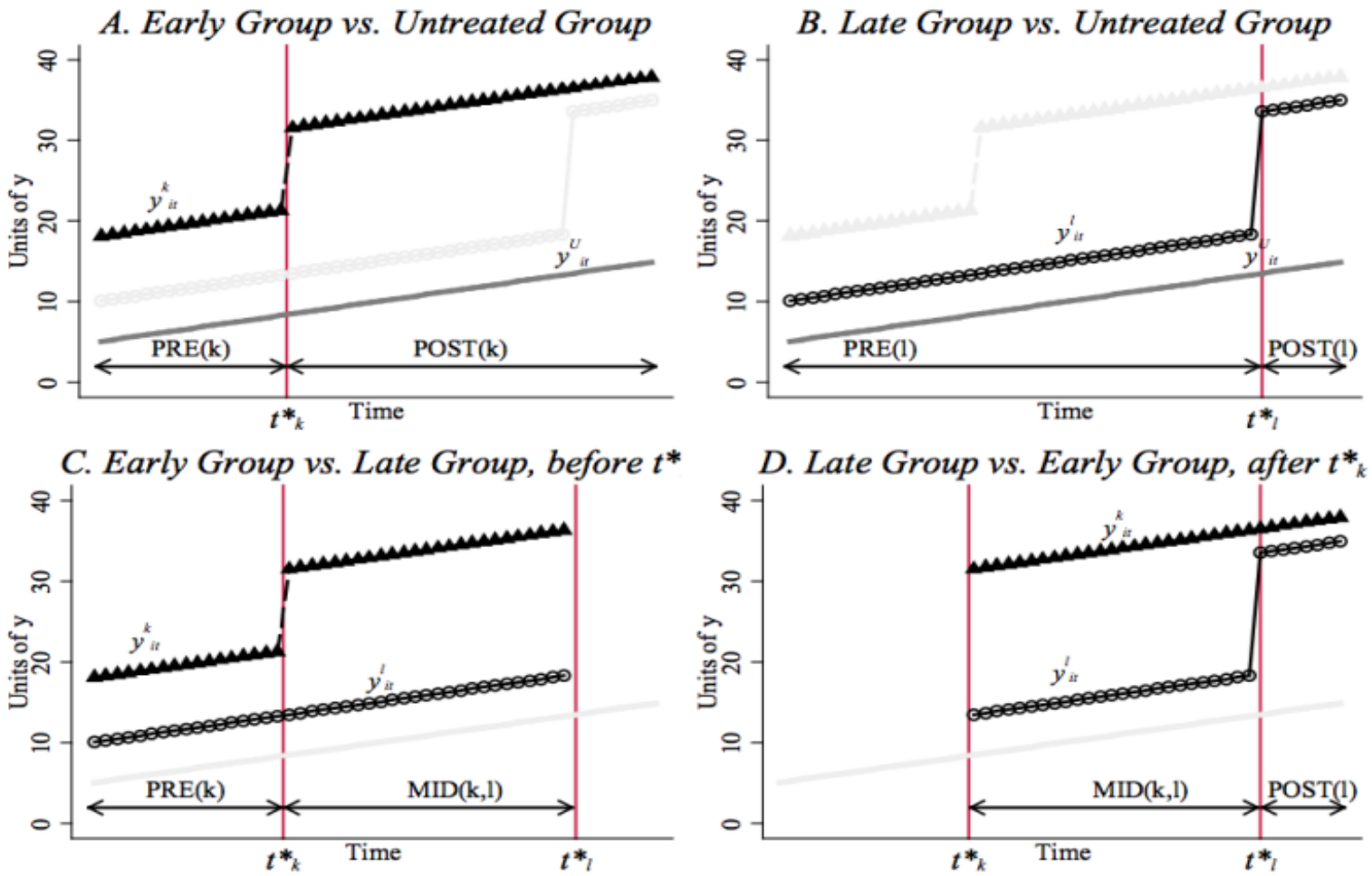
- Parallel trends is not enough for TWFE to be unbiased when treatment adoption is described by differential timing:
  - TWFE with differential timing use already-treated groups as controls.
- Assume a balanced panel dataset with T periods (t) and N cross-sectional units (i)
- Suppose two treatment groups (k, l) and one untreated group (u)

The Bacon Decomposition: Bias in the "standard" TWFE estimators

- k, l define the groups based on when they receive treatment (differently in time) with l receiving it later than k
  - k group treated at $t_i = t_k^*$ and l group treated at $t_i = t_l^*$, $t_k^* < t_l^*$
- Each timing group's sample share is $n_k = \sum_i 1(t_i = t_k^*)/N$, and the share of time it spends treated is $\overline{D}_k = \sum_t 1(t > t_k^*)/T$
- Denote $\widehat{\beta}_{ab}^{2x2}$ as the canonical 2x2 DD estimator for groups a and b
- How many 2x2 combinations are there? When there's three groups, there are four 2x2s

The Bacon Decomposition: Bias in the "standard" TWFE estimators

Figure 2. The Four Simple (2x2) Difference-in-Differences Estimates from the Three Group Case



The Bacon Decomposition: Bias in the "standard" TWFE estimators

- Panels A and B show that if we consider only one of the two treatment groups, the TWFE estimator reduces to the canonical case comparing a treated to an untreated group:
- Treated vs. Untreated I:

$$\widehat{\beta}_{ku}^{2x2} = \left(\overline{Y}_{k}^{post(k)} - \overline{Y}_{k}^{pre(k)}\right) - \left(\overline{Y}_{u}^{post(k)} - \overline{Y}_{u}^{pre(k)}\right)$$

- Treated vs. Untreated II:

$$\widehat{\beta}_{lu}^{2x2} = \left(\overline{Y}_{l}^{post(l)} - \overline{Y}_{l}^{pre(l)}\right) - \left(\overline{Y}_{u}^{post(l)} - \overline{Y}_{u}^{pre(l)}\right)$$

The Bacon Decomposition: Bias in the "standard" TWFE estimators

- If instead there were no untreated units, the TWFE estimator would be identified by the differential treatment timing between groups k and l.
- Treated at Different Times I (k treated):
  - Before t_l^*, the early units act as the treatment group because their treatment status changes, and later units act as controls during their pre-period. We compare outcomes between the window when treatment status varies, $mid(k,l)$, and timing group k pre-period, $pre(k)$

$$\widehat{\beta}_{kl}^{2x2,k} = \left(\overline{Y}_k^{mid(k,l)} - \overline{Y}_k^{pre(k)}\right) - \left(\overline{Y}_l^{mid(k,l)} - \overline{Y}_l^{pre(k)}\right)$$

The Bacon Decomposition: Bias in the "standard" TWFE estimators

- Treated at Different Times II (late vs. early after the early group has been treated):
  - The opposite situation, shown in panel D, arises after t_l^* when the later group changes treatment status but the early group does not. Later units act as the treatment group, early units act as controls, and we compare average outcomes between the periods $post(l)$ and $mid(k,l)$:

$$\widehat{\beta}_{lk}^{2x2,l} = \left(\overline{Y}_l^{post(l)} - \overline{Y}_l^{mid(k,l)}\right) - \left(\overline{Y}_k^{post(l)} - \overline{Y}_k^{mid(k,l)}\right)$$

  - The already-treated units in timing group k can serve as controls even though they are treated because treatment status does not change.

The Bacon Decomposition

$$Y_{it} = \alpha_i + \alpha_t + \beta D_{it} + \varepsilon_{it}$$

- Goodman-Bacon (2021) shows that the TWFE model is a weighted average of all the 2x2 DDs

$$\widehat{\beta}^{DD} = \sum_{k \neq u} s_{ku} \widehat{\beta}^{2x2}_{ku} + \sum_{k \neq u} \sum_{l > k} s_{kl} \left[\mu_{kl} \widehat{\beta}^{2x2,k}_{kl} + (1 - \mu_{kl}) \widehat{\beta}^{2x2,l}_{lk} \right]$$

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

- s_{ku} : weight of treated vs. untreated group
- s_{kl} : weight of early vs. late adopters
- μ_{kl} : relative weight of comparison early-late vs. late-early

The Bacon Decomposition

- The weights:

$$s_{ku} = \frac{n_k n_u \overline{D}_k (1 - \overline{D}_k)}{Var(\widetilde{D}_{it})}$$

$$s_{kl} = \frac{n_k n_l (\overline{D}_k - \overline{D}_l)(1 - (\overline{D}_k - \overline{D}_l))}{Var(\widetilde{D}_{it})}$$

$$\mu_{kl} = \frac{(1 - \overline{D}_k)}{1 - (\overline{D}_k - \overline{D}_l)},$$

  where n refer to sample sizes, $\overline{D}_k(1 - \overline{D}_k)$, $(\overline{D}_k - \overline{D}_l)(1 - (\overline{D}_k - \overline{D}_l))$ expressions refer to the variance of the treatment, μ_{kl} share of time spent under treatment early vs. late and $Var(\widetilde{D}_{it})$ is the overall variance in treatment (conditional on FE).

- It's all about weights¡

The Bacon Decomposition

- Let's think about the s_{ku} weights

$$s_{ku} = \frac{n_k n_u \overline{D}_k (1 - \overline{D}_k)}{Var(\widetilde{D}_{it})}$$

- More units in a group, the bigger its 2x2 weight is
- Moreover, the weights depend on the share of time each group spends in treatment status, $\overline{D}$ (i.e. $\overline{D}_k = \sum_t 1(t > t_k^*)/T$)
  - $\overline{D}(1 - \overline{D})$ is maximized at $\overline{D}^* = 0.5$ (then $0.5x0.5 = 0.25$)
- This means the weight is maximized for groups treated in middle of the panel, that is, when treatment occurs closer to the middle of the time window
  - Units that are treated early (more than 50%) or late (less than 50%) receive very little weight in the estimation

The Bacon Decomposition

- Interpreting the weights is less clear for earlier-later comparisons
- Same principle as before - when the difference $(\overline{D}_k - \overline{D}_l) = 0.5$
  - the earlier group is under treatment for $\overline{D}_k$ periods
  - the late group is under treatment for $\overline{D}_l$ periods
  - Example 1: $\overline{D}_k = 67\%, \overline{D}_l = 15\%$
    - $(\overline{D}_k - \overline{D}_l)(1 - (\overline{D}_k - \overline{D}_l)) = (0.67 - 0.15)(1 - 0.52) = (0.52)x(0.48) = 0.2496$
    - This is close to the maximum $0.5x0.5 = 0.25$
  - The DD estimator gives the greatest weight to groups whose treatment periods are 50% of the sample period apart...
  - Example 2: $\overline{D}_k = 70\%, \overline{D}_l = 65\%$
    - $(\overline{D}_k - \overline{D}_l)(1 - (\overline{D}_k - \overline{D}_l)) = (0.05)x(0.95) = 0.0475$

The Bacon Decomposition

- DD depends on the weights for three groups
  - Treated vs. untreated
  - Early vs. late
  - Late vs. early
- Greater weight will be given to pairs with
  - big groups (i.e. many observations, n)
  - groups that are treated closer to the middle of the sampling period
  - and treated groups whose treatment periods are half the sample period apart
- Highlights the strange role of panel length:
  - Given the weighting function, panel length alone can change the DD estimates, even when each $2x2$ DD does not change.
- Are these the weights we want to give??

The Bacon Decomposition

- What parameter are we estimating?
- Decompose these estimators into their corresponding estimation objects expressed in causal effects and biases:
  - each 2x2 DD equals an ATT plus bias from differential trends:
- ATT for timing group k:

$$\widehat{\beta}_{ku}^{2x2} = ATT_k Post(k) + \underbrace{\Delta Y_k^0(post(k), pre(k)) - \Delta Y_u^0(post(k), pre(k))}_{\text{PT bias}}$$

- Early vs. Late: the late adopters are a counterfactual for early adopters

$$\widehat{\beta}_{kl}^{2x2,k} = ATT_k Mid(k,l) + \underbrace{\Delta Y_k^0(Mid(k,l), pre(k)) - \Delta Y_l^0(Mid(k,l), pre(k}_{\text{PT bias}}$$

- Trends need to be parallel until both groups are treated

The Bacon Decomposition

- But what about the 2x2 that compared the late groups to the already-treated earlier groups?

$$\widehat{\beta}_{lk}^{2x2,l} = ATT_l Post(l) + \underbrace{\Delta Y_l^0(Post(l), Mid(k,l)) - \Delta Y_k^0(Post(l), Mid(k,l)}_{\text{PT bias}}$$

$$- \underbrace{(ATT_k(Post(l)) - ATT_k(Mid(k,l))}_{\text{Heterogeneity bias}}$$

- Trends need to be parallel from the time the early adopter has been treated
- But that is not enough. The treatment effect for the early adopter needs to be constant over time.
  - A heterogeneity bias if the ATT for k is dynamic. If not, then it just zeroes out.
- Notice all those potential sources of biases!

The Bacon Decomposition

- The problem with $\widehat{\beta}_{lk}^{2x2,l}$ is that changes in the early-treated outcomes my reflect changes in their treatment effects over time.
- Then, the resultant DD estimator could reflect differences in treatment effects over time between different treatment cohorts:

$$\widehat{\beta}_{lk}^{2x2,l} = \left(\overline{Y}_{l}^{post(l)} - \overline{Y}_{l}^{mid(k,l)}\right) - \left(\overline{Y}_{k}^{post(l)} - \underbrace{\overline{Y}_{k}^{mid(k,l)}}\right)$$

This already includes treatment¡

Details Bacon Decomposition: heterogeneity bias

$$\widehat{\beta}_{kl}^{2x2,I} = \underbrace{\left(\overline{Y}_I(post(I)) - \overline{Y}_I(mid)\right)}_{(a)} - \underbrace{\left(\overline{Y}_k(post(I)) - \overline{Y}_k(mid)\right)}_{(b)}$$

- The problem is that $\overline{Y}_k(mid)$ is an observed outcome with treatment

Let's decompose each component in $\widehat{\beta}_{kl}^{2x2,I}$:

$$\begin{aligned}
(a) \quad : \quad & \left(\overline{Y}_I(post(I)) - \overline{Y}_I(mid)\right) = \text{ by } + - \overline{Y}_I^0(post(I)) \\
= \quad & \left(\overline{Y}_I(post(I)) - \overline{Y}_I^0(post(I)\right) \\
& + \left(\overline{Y}_I^0(post(I)) - \overline{Y}_I(mid)\right) \\
= \quad & ATT_I(post(I)) + \underbrace{\left(\overline{Y}_I^0(post(I)) - \overline{Y}_I(mid)\right)}_{\text{Diff. betw. the two periods in absence of treat.}}
\end{aligned}$$

Details Bacon Decomposition: heterogeneity bias

- The problem is in part (b):

$$
\begin{aligned}
(b): \quad & \left(\overline{Y}_k(post(l)) - \overline{Y}_k(mid)\right) = \text{ by } +- \overline{Y}^0_k(post(l)) \\
= \quad & \left(\overline{Y}_k(post(l)) - \overline{Y}^0_k(post(l)\right) \\
& + \left(\overline{Y}^0_k(post(l)) - \overline{Y}_k(mid)\right) \\
= \quad & ATT_k(post(l)) + \left(\overline{Y}^0_k(post(l)) - \overline{Y}_k(mid)\right)
\end{aligned}
$$

- $\overline{Y}_k(mid)$ is an outcome that already "includes" treatment. **It can be writen:**

$$
\begin{aligned}
\overline{Y}_k(mid) &= \overline{Y}^0_k(mid) + (\overline{Y}_k(mid) - \overline{Y}^0_k(mid)) \\
&= \overline{Y}^0_k(mid) + ATT_k(mid)
\end{aligned}
$$

Therefore:

$$
\begin{aligned}
\left(\overline{Y}_k(post(l)) - \overline{Y}_k(mid)\right) &= ATT_k(post(l)) - ATT_k(mid) \\
&\quad + \underbrace{(\overline{Y}^0_k(post(l)) - \overline{Y}^0_k(mid))}_{\text{Diff. in absence of treatment}}
\end{aligned}
$$

Details Bacon Decomposition: heterogeneity bias

- All potential biases in :

$$\widehat{\beta}_{kl}^{2x2,l} = ATT_l(post(l)) + \underbrace{[ATT_k(mid) - ATT_k(post(l))]}_{\text{Het. bias}} + \underbrace{\left(\overline{Y}_l^0(post(l)) - \overline{Y}_l(mid)\right)}_{\text{PT bias}} - \underbrace{(\overline{Y}_k^0(post(l)) - \overline{Y}_k^0(mid))}_{\text{PT bias}}$$

- The heterogeneity bias disappears when the treatment effect for k is constant, that is, when the change in the outcome for k in the absence of the treatment would be the same in the two periods, mid and $post(l)$:

$$\left(\overline{Y}_k(mid) - \overline{Y}_k^0(mid)\right) = \left(\overline{Y}_k(post(l)) - \overline{Y}_k^0(post(l)\right)$$

The Bacon Decomposition:

- Substituting previous equations into the DD decomposition theorem expresses the probability limit of the TWFE estimator as:

$$p \lim_{N \to \infty} \widehat{\beta}^{DD} = \beta^{DD} = VWATT + VWCT + \Delta ATT$$

- $VWATT$: This term is the interpretable causal parameter that TWFE can estimate, which it is called the "variance-weighted average treatment effect on the treated"
- $VWCT$: This term, which it is called "variance-weighted common trends"
- ΔATT : This term is the change in a treatment effects (within groups) over time
- To identify $VWATT$, we need to assume (and justify) why $VWCT = \Delta ATT = 0$
- The key of this decomposition is that it clarifies the comparisons and weights involved in the TWFE estimator.

The Bacon Decomposition:

- Variance weighted ATT ($VWATT$) : The ATT is a positively weighted average of each of the 2x2 DD ATT Estimates, where the weights depend on group shares and treatment variance:

$$\begin{aligned} VWATT &= \sum_{k \neq u} \sigma_{ku} ATT_k(Post(k)) \\ &\quad + \sum_{k \neq u} \sum_{l > k} \sigma_{kl} \left[\mu_{kl} ATT_k(Mid(k,l)) + (1 - \mu_{kl}) ATT_l(Post(l)) \right] \end{aligned}$$

  where σ is like s only population terms not samples
- Weights sum to one.
- Note, if all the ATT are identical, then the weighting is irrelevant:
  - The $VWATT = ATT$ if the ATTs are the same for each pair
- But otherwise, it's basically weighting each of the individual sets of ATT, where weights depend on group size and variance (it places more weight on groups with more variance in the treatment)
- But the VWATT can be far away from the ATT if some groups carry a heavy weight

The Bacon Decomposition:

- Variance weighted common trends ($VWCT$): It is an average of the difference in counterfactual trends between pairs of timing groups and different time periods using the weights from the decomposition theorem.

$$
\begin{aligned}
VWCT \quad = \quad & \sum_{k \neq u} \sigma_{ku} \left[\Delta Y_k^0(Post(k), pre(k)) - \Delta Y_u^0(Post(k), pre(k)) \right] \\
& + \sum_{k \neq u} \sum_{l > k} \sigma_{kl} [\mu_{kl} \left\{ \Delta Y_k^0(Mid(k,l)), \Pr e(k)) - \Delta Y_l^0(Mid(k,l)), \right. \\
& + (1 - \mu_{kl}) \left\{ \Delta Y_l^0(Post(l), Mid) - \Delta Y_k^0(Post(l), Mid(k,l) \right\}]
\end{aligned}
$$

- This is new. That's a lot of parallel trends we need equalling zero, and this was only with two treatment groups!

The Bacon Decomposition:

- Heterogeneity within-time bias (ΔATT): This term equals a weighted sum of the change in treatment effects within each timing group's before and after a later treatment time:

$$\Delta ATT = \sum_{k \neq u} \sum_{l > k} \sigma_{kl}(1 - \mu_{kl}) \left[ATT_k(Post(l) - ATT_k(Mid(k, l)) \right]$$

- If the ATT is constant over time, then this difference is zero, but what if the ATT is not constant?
  - There is no problem when there is heterogeneity across groups, the root of the problem comes from the heterogeneity within groups over time.
  - ΔATT is the source of the negative weights: this bias has arbitrary sign

The Bacon Decomposition:

- ΔATT quantifies the extent to which the changes in early-treated outcomes are contaminated by changes in treatment effects over time for this group.
- To the extent that this occurs, these outcome trend are inappropriate counterfactuals for the later treated.
- Therefore, we need to look for the adequate counterfactual:
  - We need to substract the difference $[ATT_k(Post(l) - ATT_k(Mid(k, l))]$

The Bacon Decomposition:

- When the PT assumptions are satisfied $(VWCT = 0)$ and when the treatment effects are constant $(\Delta ATT = 0)$, but the treatment effects varies across groups:

$$p \lim_{N \to \infty} \widehat{\beta}^{DD} = \beta^{DD} = VWATT$$

- Even in this case, $VWATT$ differs from just the sample average ATT, because OLS applies weights on each ATT estimate that generally differ from the sample shares.

Lessons from Goodman-Bacon (2021)

- DD, while seemingly intuitive and transparent, is actually not that easy
- Differential treatment timing adds a layer of complexity
- We can use the Goodman-Bacon decomposition to report the weights placed on the different TWFE estimates from each 2-period, 2-group estimate. This allows us to evaluate how much weight is being placed on "forbidden" comparisons of already-treated units and how removing the comparisons would change the estimate.
- Stata commands: `bacondecomposition` or `ddtiming`.
- Several recent papers have proposed alternative estimators that more sensibly aggregate heterogeneous treatment effects in settings with staggered treatment timing:
  - Callaway and Sant'Anna (2021) (CS) (Stata command: `csdid`)
  - Sun and Abraham (2021) (SA) (Stata command: `eventstudyinteract`)