
COLLEGIO CARLO ALBERTO
2025
2x2 = (Ẏk,t1 − Ẏk,t-1) −
− (Ẏu,t1 − Ẏut-1)

Roadmap

- Continuous DiD
  - Dx2 and DxT
  - Identification
  - Selection bias
  - Interpreting TWFE

Continuous DiD

- It is very common for people to estimate difference-in-differences panel models where the treatment, D, is multi-valued or continuous, not binary:

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

- Examples include minimum wage papers, my JHR on abortion clinic closures causing increased travel distance, vaccinations, price elasticity of demand etc.
- Variation is in “treatment intensity” and researchers typically use TWFE for estimation, or perhaps count models like Poisson

Praise for OLS and Continuous Treatments

"The two-period regression estimator can be easily modified to allow for continuous, or at least non-binary, treatments." (Wooldridge 2005)

"A second advantage of regression DiD is that it facilitates the study of policies other than those that can be described by a dummy." (Angrist and Pischke 2008)

New Continuous DiD

1. But new work suggests that the TWFE approach to continuous is problematic in light of unrestricted heterogenous treatment effects (Baker, et al. 2025)
2. What of the 2x2 and 2xT will be relevant for dosage designs (Dx2 and DxT)?
3. What is the target parameter, what new assumptions, what estimation methods, what control group?

Continuous Literature in Causal Inference

- Continuous treatments in instrumental variables (Angrist and Imbens 1995; Angrist, Graddy and Imbens 2000)
- Continuous instruments (Imbens and Angrist 1994; Heckman and Vytacil 2005)
- But the work on continuous diff-in-diff is newer (de Chaisemartin, et al. 2024; de Chaisemartin, et al. 2025; Callaway, Goodman-Bacon and Sant'Anna 2025)
- We will primarily focus on Callaway, Goodman-Bacon and Sant'Anna (2025) for the sake of time and focus

Average causal response functions

- **ATT**: Average effect for a binary treatment on a sub-population of treated units after they were treated
- **Dose**: Treatment is not binary but rather multi-valued or continuous. Represents either ATTs for groups at dosages, or movements along a dosage curve – not the same thing as it turns out

Parameters

Average treated on the treated

$$ATT(d|d) = E[Y_{it}^{d} - Y_{it}^{0}|D_{it} = d]$$

while the treatment, D, can be any amount, d, that amount is technically a particular dose. We raised the minimum wage, but we raised it to a particular wage.

Parameters

Average treated on the treated

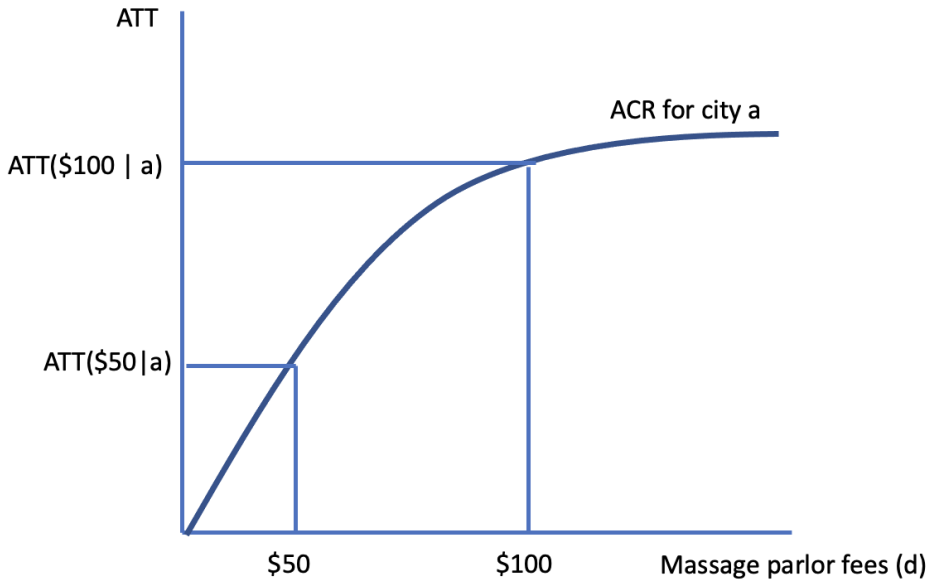
$$ATT(d|d) = E[Y_{it}^{d} - Y_{it}^{0}|D_{it} = d]$$

This is “the ATT of d for the groups that chose d dosage” which uses as its comparison no dose.

Angrist and Imbens 1995

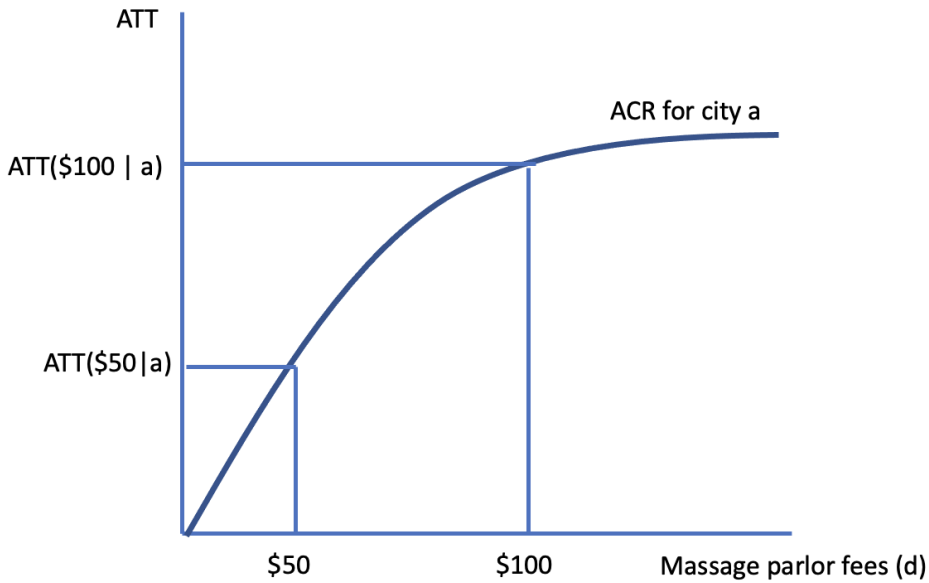
*"We refer to the parameter β as the **average causal response (ACR)**. This parameter captures a weighed average causal responses to a unit change in treatment, for those whose treatment status is affected by the instrument. …"*

ATT for a given dose



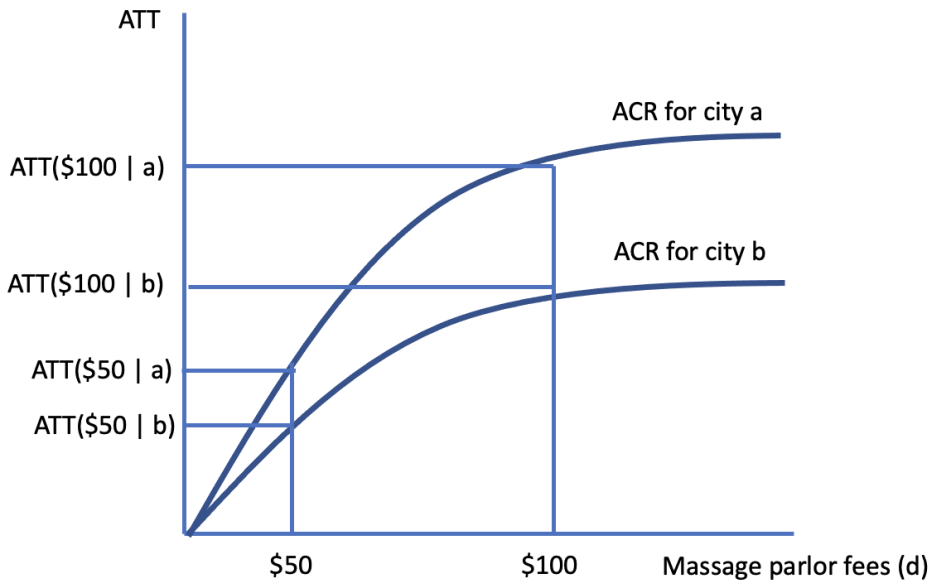
What is the effect of setting fees to $100 versus nothing at all? It's ATT($100|a) for this city.

ATT for a given dose



Assume city a did choose $d = \$100$. Then ATT($50|a) just means that that is its ATT *had* it chosen the lower level. The curve, in other words, is tracing out all average causal response for this city.

ATT for a given dose



What if everyone has different responses? In other words, city a has the higher curve than city b. Then there are several comparisons possible. What is the effect of $50 on outcomes for cities that actually chose $50 versus those than actually chose $100?

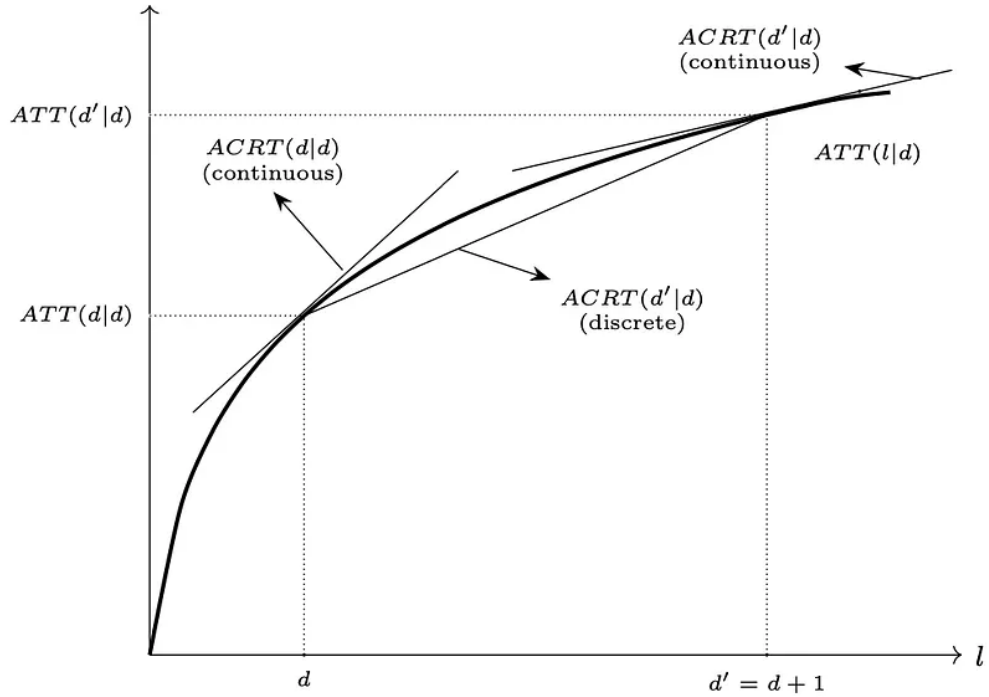
Parameters

Average treated on the treated

$$ATT(d|d) = E[Y_{it}^{d} - Y_{it}^{0}|D_{it} = d]$$

Notice that you are comparing any dose d to no treatment at all – sort of an extensive margin causal response, but that isn't the only causal concept we have. Elasticities are causal, demand curves are causal, but they aren't based on comparisons to nothing – they are intensive margin comparisons, local comparisons, adjacencies. Zero isn't the only counterfactual in other words.

Figure 2: Causal Parameters in a Continuous Difference-in-Differences Design



Notes: The figure plots $ATT(\cdot|d)$ (the average effect of experiencing each dose among units that actually experienced dose d). We highlight causal parameters for two doses, d and d'. $ATT(d|d)$ and $ATT(d'|d)$ are average treatment effect on the treated parameters and refer to the height of the curve. $ACRT(d|d)$ and $ACRT(d'|d)$ are average causal response parameters and refer to the slope of the curve. We show them for a continuous dose, when the $ACRT$ is a tangent line, and for a discrete dose when $ACRT$ is a line connecting two discrete points on $ATT(D|d)$.

What is the ACRT?

- ACRT is the causal effect of dose $D = d_j$ vs a different dose $D = D_{j-1}$ for group d
  - → Easiest example is the demand function: at $p = \$10$, I buy 10 units, but at $p = \$11$, I buy 5 units.
  - → Causal effect of that one dollar increase is -5 units
  - → Demand curves are pairs of potential outcomes and treatments and equilibrium "selects" one of them
- Discrete/multi-valued treatment is linear difference between two ATTs for the same city
- Continuous treatment is the derivative of the function itself

Definition of the ACRT

$$ACRT(d|d') = \left.\frac{\partial ATT(l|d')}{\partial l}\right|_{l=d}$$

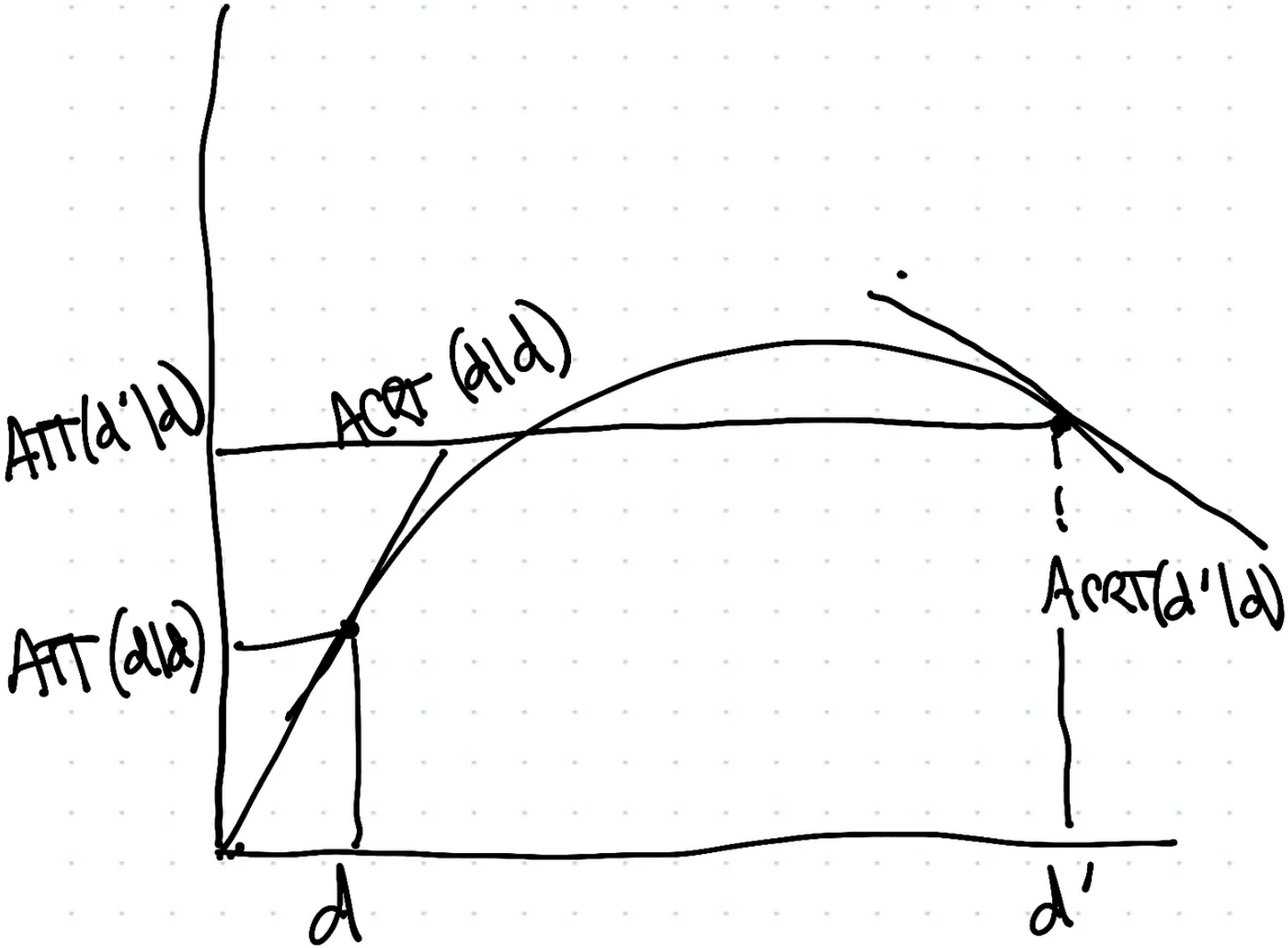
Derivation of ACRT

Average causal response parameters for absolutely continuous treatments are defined as

$$ACRT(d|d') = \frac{\partial ATT(l|d')}{\partial l}\bigg|_{l=d} = \frac{\partial \mathbb{E}[Y_{t=2}(l)|D=d']}{\partial l}\bigg|_{l=d} \text{ and } ACR(d) = \frac{\partial ATE(d)}{\partial d} = \frac{\partial \mathbb{E}[Y_{t=2}(d)]}{\partial d}.$$

$ACRT(d|d)$ equals the derivative of the $t = 2$ average potential outcome for units that received dose d evaluated at d. This is equivalent to the derivative of $ATT(l|d)$ with respect to l, evaluated at $l = d$. For discrete treatments, average causal responses are defined in a similar way but with slightly

Heterogeneities



Assumptions

The authors lay out 5 assumptions, but I'm going to focus on 4. They are:

1. Random sampling
2. Continuous (2a) and Multi-Valued Treatment (2b)
3. No Anticipation and Observed Outcomes
4. Parallel trends

Identifying $ATT(d|d)$

We can estimate the $ATT(d|d)$ using the simple DiD equation:

$$E[\Delta Y_{it}|D_i = d] - E[\Delta Y_{it}|D_i = 0]$$

No anticipation and parallel trends converts this comparison of before and after into the $ATT(d|d)$

$ATT(d|d)$ is using as its counterfactual the “no treatment”, note. Treatment is a dosage compared to zero iow.

Identifying ACRT

$$\begin{aligned} ATT(b|b) - ATT(a|a) &= (E[\Delta Y_{it}|D_i = a] - E[\Delta Y_{it}|D_i = 0]) \\ &\quad -(E[\Delta Y_{it}|D_i = b] - E[\Delta Y_{it}|D_i = 0]) \\ &= E[\Delta Y_{it}|D_i = a] - E[\Delta Y_{it}|D_i = b] \end{aligned}$$

Comparing high and low dose groups.

Identifying ACRT

$$
\begin{aligned}
ATT(d_j|d_j) - ATT(d_{j-1}|d_{j-1}) &= \\
(ATT(d_j|d_j) - ATT(d_{j-1}|d_j)) + (ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})) &= \\
(ACRT(d_j|d_j)) + (ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})) &=
\end{aligned}
$$

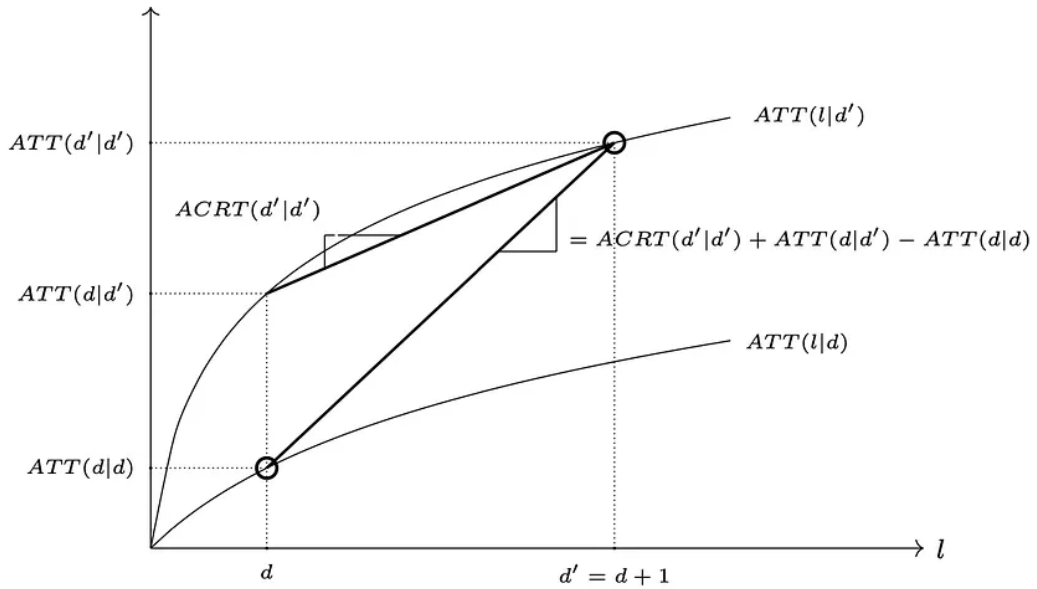
Part in blue is the movement along the average causal response function, the ACRT, and is causal. The part in red is selection bias.

Identifying ACRT

$$
\begin{aligned}
ATT(d_j|d_j) - ATT(d_{j-1}|d_{j-1}) &= \\
(ATT(d_j|d_j) - ATT(d_{j-1}|d_j)) + (ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})) &= \\
(ACRT(d_j|d_j)) + (ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})) &=
\end{aligned}
$$

Notice parallel trends allows to identify ATT terms but we need additional assumptions for this red part to vanish. We must assume that the ATT for cities that chose d_j and cities that chose d_{j-1} are the same had they both chose d_{j-1}.

Figure 3: Non-identification of Average Causal Response with Treatment Effect Heterogeneity, Two Discrete Doses



Notes: The figure shows that comparing adjacent $ATT(d|d)$ estimates equals an $ACRT$ parameter (the slope of the higher-dose group's ATT function) and selection bias (the difference between the two groups' ATT functions at the lower dose).

Theorem 3.2. *Under Assumptions 1 to 4, causal response parameters are not identified. Specifically,*

(a) Under Assumption 2(a), for $d \in \mathcal{D}_+^c$,

$$\frac{\partial \mathbb{E}[\Delta Y|D=d]}{\partial d} = \frac{\partial ATT(d|d)}{\partial d} = ACRT(d|d) + \underbrace{\frac{\partial ATT(d|l)}{\partial l}\Big|_{l=d}}_{\textit{selection bias}};$$

(b) For $(h,l) \in \mathcal{D} \times \mathcal{D}$,

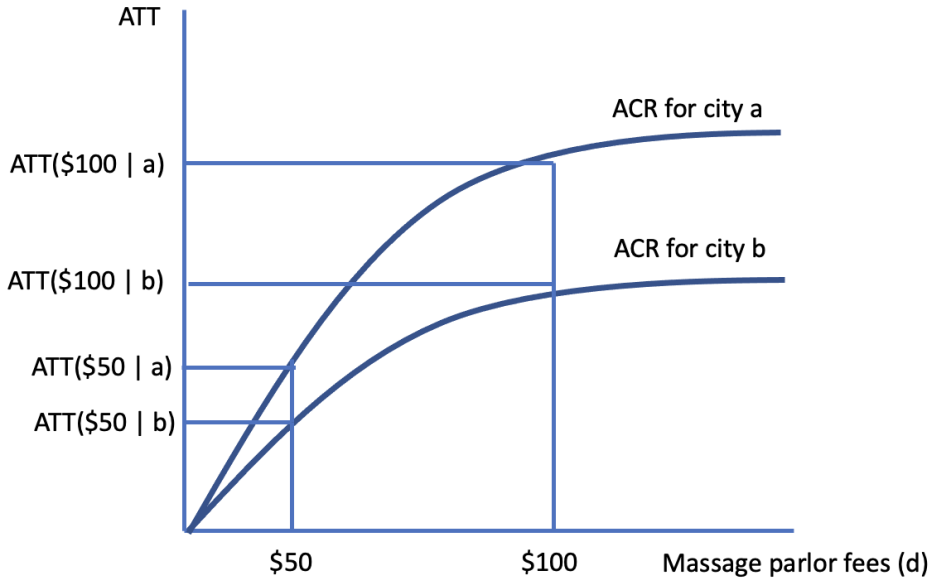
$$\mathbb{E}[\Delta Y|D=h] - \mathbb{E}[\Delta Y|D=l] = ATT(h|h) - ATT(l|l)$$

$$= \underbrace{\mathbb{E}[Y_{t=2}(h) - Y_{t=2}(l)|D=h]}_{\textit{causal response}} + \underbrace{\Big(ATT(l|h) - ATT(l|l)\Big)}_{\textit{selection bias}}.$$

When Assumption 2(b) holds, taking $h = d_j$ *and* $l = d_{j-1}$ *implies that*

$$\mathbb{E}[\Delta Y|D=d_j] - \mathbb{E}[\Delta Y|D=d_{j-1}] = ACRT(d_j|d_j) + \underbrace{ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1})}_{\textit{selection bias}}.$$

Causality and selection bias



Draw the ACRT for top curve and the selection bias from estimation under assumptions 1 to 4.

Interpreting this

- Unrestricted heterogenous treatment effects (across dosage levels and across units with difference dose response functions) is not itself the problem
- If we randomized dosages, then $ATT(d_{j-1}|d_j) - ATT(d_{j-1}|d_{j-1}) = 0$
- Why? Because then there is no selection on gains from dosages, and average causal response functions are the same for all dosage groups
- So then when is this a problem? Sorting on gains

Interpreting this

- When estimating treatment effects using continuous DiD, you will need to make one of two assumptions
  1. Strong parallel trends: Average change in $E[Y^0]$ for the entire sample is the same as the d group
  2. Parallel trends plus homogenous treatment effect functions
- Roy model like sorting on gains typically lead to violations of the second condition insofar as there is heterogenous returns to dosages across units
- So the question you have to ask yourself is do you think that cities are “optimally setting the minimum wage” around some given minimum wage?

Stronger assumption

- I'm really not so sure I think that when it comes to state legislation that I think a Roy model is likely responsible for the equilibrium
- Solving constrained optimization problems is hard and unlikely is it the case that Florida's ATT and Georgia's ATT are terribly different from one another had both chosen the same minimum wage (but that is the bias)
- Authors introduce a fifth assumption that will eliminate selection bias, but at the price of restricting heterogeneity

Discussion of strong parallel trends

We discuss an alternative but typically stronger assumption, which we call strong parallel trends, that says that the path of outcomes for lower-dose units must reflect how higher-dose units' outcomes would have changed had they instead experienced the lower dose. Thus, strong parallel trends restricts treatment effect heterogeneity and justifies comparing dose groups. Absent this type of condition, comparisons across dose groups include causal responses but are "contaminated" by an additional term involving possibly different treatment effects of the same dose for different dose groups—we refer to this additional term as selection bias.

A5: Strong parallel trends

Assumption 5 (Strong Parallel Trends). *For all* $d \in \mathcal{D}$,

$$\mathbb{E}[Y_{t=2}(d) - Y_{t=1}(0)] = \mathbb{E}[Y_{t=2}(d) - Y_{t=1}(0)|D = d].$$

Randomization and strong parallel trends

- Randomized dosages guarantees that the ACRT are the same across all dosage groups
- In this situation, strong parallel trends holds because all dosages have the same ATE and ACRT
- Roy like sorting on dosage may be the biggest challenge you'll face – schooling stops, family size may not satisfy strong parallel trends

Interpreting TWFE results

We next use the identification results to evaluate the most common way that practitioners estimate continuous DiD designs, which is to run a TWFE regression that includes time fixed effects (θ_t), unit fixed effects (η_i), and the interaction of a dummy for the post-treatment period ($Post_t$) with a variable that measures unit i's dose or treatment intensity, D_i:

$$Y_{i,t} = \theta_t + \eta_i + \beta^{twfe} D_i \cdot Post_t + v_{i,t}. \tag{1.1}$$

This TWFE specification is clearly motivated by DiD setups with two periods and two treatment groups, though many prominent textbooks recommend using it in more general setups (e.g., Cameron and Trivedi, 2005, Angrist and Pischke, 2008, and Wooldridge, 2010). There are several ways to interpret β^{twfe}, each corresponding to a different type of causal parameter. We decompose it in terms of level effects, scaled level effects, causal responses, and scaled high-versus-low (2×2) effects. Each decomposition is a weighted integral of dose-specific causal parameters, and none provide a clear causal and policy-relevant interpretation of β^{twfe}, at least not when treatment effects are allowed to vary across doses and/or groups.

Our impression is that empirical researchers typically interpret β^{twfe} in three main (and related) ways, implicitly relying on different building blocks. First, β^{twfe} is often directly interpreted as a causal response parameter; that is, how much the outcome causally increases on average when the treatment increases by one unit. This is the causal version of how regression coefficients are often taught to be interpreted in introductory econometrics classes. Second, it is common to pick a representative value for d, to report $d \times \beta^{twfe}$, and interpret this quantity as $ATT(d)$. This is the main interpretation provided in Acemoglu and Finkelstein (2008): "Given that the average hospital has a

38 percent Medicare share prior to PPS, this estimate [i.e., of β^{twfe}, here equal to 1.129] suggests that in its first 3 years, the introduction of PPS was associated with an increase in the depreciation share of about 0.42 ($\approx 1.129 \times 0.38$) for the average hospital." Rearranging this expression shows that under this interpretation $\beta^{twfe} = ATT(d|d)/d$, which relates β^{twfe} to a scaled level effect. Third, it is common to take two different representative values of the dose, d_1 and d_2—a common choice is the 25th percentiles and 75th percentiles of the dose—and interpret β^{twfe} as the average causal response of moving from dose d_1 to dose d_2 scaled by the distance between d_1 and d_2; this is a scaled 2×2 effect. We aim to assess whether such types of interpretations are justified and under which conditions.

Interpreting TWFE

Theorem 3.4. *Under Assumptions 1, 2(a), 3, and 4, β^{twfe} can be decomposed in the following ways:*

(a) Causal Response Decomposition:

$$\beta^{twfe} = \int_{d_L}^{d_U} w_1^{acr}(l) \left(ACRT(l|l) + \underbrace{\frac{\partial ATT(l|h)}{\partial h}\Big|_{h=l}}_{\text{selection bias}} \right) dl + w_0^{acr} \frac{ATT(d_L|d_L)}{d_L}$$

where the weights are always positive and integrate to 1.

[10]The decompositions in the main text integrate over all possible doses. In Appendix SC.2 in the Supplementary Appendix, we additionally consider scaled level and scaled 2×2 decompositions for particular, fixed values of the dose. There we show that, even under strong parallel trends, β^{twfe} can be (possibly much) different from these parameters when there is treatment effect heterogeneity due to (i) different weighting schemes (similar to the differences that we point out in this section) and (ii) β^{twfe} being dependent on causal responses at other doses.

(b) Levels Decomposition:

$$\beta^{twfe} = \int_{d_L}^{d_U} w_1^{lev}(l) ATT(l|l)\, dl,$$

where $w_1^{lev}(l) \lesseqgtr 0$ *for* $l \lesseqgtr \mathbb{E}[D]$, *and* $\int_{d_L}^{d_U} w_1^{lev}(l)\, dl + w_0^{lev} = 0$.

(c) Scaled Levels Decomposition:

$$\beta^{twfe} = \int_{d_L}^{d_U} w^{s}(l) \frac{ATT(l|l)}{l}\, dl,$$

where $w^{s}(l) \lesseqgtr 0$ *for* $l \lesseqgtr \mathbb{E}[D]$, *and* $\int_{d_L}^{d_U} w^{s}(l)\, dl = 1$.

(d) Scaled 2×2 *Decomposition*

$$\beta^{twfe} = \int_{d_L}^{d_U} \int_{\mathcal{D}, h>l} w_1^{2\times 2}(l,h) \left(\underbrace{\frac{\mathbb{E}[Y_{t=2}(h) - Y_{t=2}(l)|D=h]}{h-l}}_{\text{causal response}} + \underbrace{\frac{ATT(l|h) - ATT(l|l)}{h-l}}_{\text{selection bias}} \right) dh\, dl$$
$$+ \int_{d_L}^{d_U} w_0^{2\times 2}(h) \frac{ATT(l|l)}{l}\, dl,$$

where the weights $w_1^{2\times 2}$ *and* $w_0^{2\times 2}$ *are always positive and integrate to 1.*

If one imposes Assumption 5 instead of Assumption 4, then the selection bias terms from Part (a) and Part (d) become zero, and the remainder of the decompositions remain true, except one needs to replace $ACRT(l|l)$ *with* $ACR(l)$ *in Part (a),* $ATT(l|l)$ *with* $ATE(l)$ *in Parts (b), (c) and (d), and* $\mathbb{E}[Y_{t=2}(h) - Y_{t=2}(l)|D=h]$ *with* $\mathbb{E}[Y_{t=2}(h) - Y_{t=2}(l)]$ *in Part (d).*

Table 1: TWFE Decomposition Weights

Decomposition	$D > 0$ Weights	$D = 0$ Weights
Causal response	$w_1^{\text{acr}}(l) = \frac{(\mathbb{E}[D\|D \geq l] - \mathbb{E}[D])\mathbb{P}(D \geq l)}{\text{Var}(D)}$	$w_0^{\text{acr}} = \frac{(\mathbb{E}[D\|D > 0] - \mathbb{E}[D])\mathbb{P}(D > 0)d_L}{\text{Var}(D)}$
Levels	$w_1^{\text{lev}}(l) = \frac{(l - \mathbb{E}[D])}{\text{Var}(D)} f_D(l)$	$w_0^{lev} = -\frac{\mathbb{E}[D]\mathbb{P}(D = 0)}{\text{Var}(D)}$
Scaled levels	$w^{\text{s}}(l) = l\frac{(l - \mathbb{E}[D])}{\text{Var}(D)} f_D(l)$	
Scaled 2×2	$w_1^{2\times2}(l, h) = \frac{(h - l)^2 f_D(h) f_D(l)}{\text{Var}(D)}$	$w_0^{2\times2}(h) = \frac{h^2 f_D(h)\mathbb{P}(D = 0)}{\text{Var}(D)}$

Notes: The table provides the formulas for the weights used in the decompositions of β^{twfe} provided in this section.

Understanding Decomposition Results

- The pattern from decomposition shows distinct impacts of parameter types.
- **Level-effect parameters** (parts b and c):
  - → β_{twfe} is not influenced by selection bias.
  - → Includes negative weights.
- **Comparative doses parameters** (parts a and d):
  - → β_{twfe} carries positive weights.
  - → Encounters selection bias under parallel trends.

Addressing Selection Bias and Weighting Schemes

- Parametric linearity restrictions may overlook weighting scheme issues inherent in TWFE regression.
- These restrictions do not resolve selection bias problems.
- Next, we explore:
  - → Alternative estimators to TWFE that adjust the weighting scheme.
  - → These alternatives do not rely on the stringent linearity assumption.
  - → Selection bias issues persist and require different solutions.