Low Dose or No Dose? Continuous Treatment Difference-in-Differences with Unknown Controls

Elird Haxhiu [*] Thomas Helgerman [*]

haxhiu@umich.edu tehelg@umich.edu

September 23, 2022

Abstract

This paper studies difference-in-differences research designs where all units receive a continuous treatment, or dose, so there is no group that is ex ante unexposed. We present a framework to identify and estimate average treatment effect and causal response parameters when the continuous treatment takes effect only after some cutoff value. In applied settings, this parameter is usually unknown and hence neglected from econometric analysis. Under a range of data-generating processes, we illustrate the bias from Two-Way Fixed-Effects (TWFE) estimators when treatment is defined as (i) the full dose or (ii) an indicator for units with doses above some researcher-specified value or percentile, such as the median. For large jumps or sharp discontinuities at the cutoff value, researchers should instead jointly estimate the threshold along with treatment effect parameters using existing methods. This restores identification and produces correct standard errors but fails when parametric assumptions do not hold or the dose response function is flat around the true cutoff. In these cases, we argue that researchers should instead target binned average treatment effects and document an intuitive bias-variance tradeoff in recategorizing low dose units as controls in estimation. We then exploit this trade-off to derive the MSE-optimal estimator, show that it depends on the unknown cutoff, and propose a minimax constraint and partial identification procedure to make progress on inference.

JEL codes: C14, C23, C24.

Key words: Difference-in-Differences, Parallel Trends, Threshold Estimation, Dose Response curves

[*] Department of Economics, University of Michigan, 611 Tappan Ave, Ann Arbor, MI, 48109. The authors thank Florian Gunsilius, Emir Murathanoglu, Charles Brown, Melvin Stephens, and participants at the Michigan Labor Lunch seminar for their helpful discussions and suggestions!

Contents

1 Introduction

Without additional assumptions, it is generally impossible to infer the effects of a treatment by comparing participants to non-participants or comparing the participants over time. For valid inference, unit comparisons must restrict selection into treatment, which is generally difficult to justify in the absence of an instrument. On the other hand, time comparisons (e.g. interrupted time series methods) cannot allow for contemporaneous trends. Difference-in-differences (DD) research designs combine these two estimators to infer causal effects by subtracting the change in outcomes over time for non-participants from the change for participants. This is valid whenever the average change in outcomes for participants in the absence of treatment (a counterfactual moment) is well proxied by the average change in outcomes for those who actually go untreated, known as a Parallel Trends assumption (PTA). Panel data thus enables identification of a causal effect without ruling out selection into treatment levels or contemporaneous trends.

A common extension in practice admits a continuously distributed treatment (or dose) variable. In contrast to the binary case, all units are exposed to some level of the dose $D_i \geq 0$. The first documented use of this method by Card (1992) measured a state's exposure to a higher federal minimum wage by the share of teenage workers who fall below the new legislated minimum. The typical estimating equation follows via analogy to a standard DD estimator, and specifies

$$Y_{it} = \alpha_i + \theta_t + \beta_{\text{TWFE}} \cdot D_i P_t + e_{it} \tag{1}$$

where Y_{it} is some outcome of interest, α_i and θ_t are unit and time fixed-effects respectively, P_t indicates the periods that units are exposed, and e_{it} is some error term. The main coefficient of interest in the Two-Way Fixed-Effects (TWFE) regression above is β_{TWFE}, which represents an aggregation of the "effect" of the continuous treatment. Researchers have typically interpreted this as the weighted average dose response function (DRF), in line with the "average derivative" interpretation of regression coefficients (Yitzhaki 1996, Angrist & Pishke 2008).

Recent work by Callaway, Goodman-Bacon, and Sant-Anna (2021), hereafter referred to as CGS, illuminates exactly what this "effect" is and what assumptions are required to identify it. They show that whenever researchers have access to pure control units (which we define as those units *ex ante* known to be unexposed to the treatment with dose $D_i = 0$) only minor modifications to the Parallel Trends assumption are sufficient to identify average treatment effect on the treated (ATT) parameters, defined for each positive dose $d > 0$. However, standard assumptions are not strong enough to successfully identify the average effect of marginally increasing the dose at some level. Intuitively, this is because comparisons across treated units at different doses must additionally restrict selection into treatment, which is not typical in standard DD settings. In the case of a continuous dose, the TWFE estimator in (1) amounts to differencing adjacent ATT estimates across the dose distribution, which only reveals a causal response under strong restrictions on counterfactual heterogeneity (what CGS call strong parallel trends).

Thus, without strong homogeneity assumptions *and whenever a pure control group exists*, researchers can only identify ATT parameters. But what about cases when researchers do not have access to these units? To understand current strategies, we conducted a metastudy of 44 American Economic Review (AER) papers since 2000 estimating continuous difference-in-differences designs. We found that 70% of the papers in our metastudy estimate a full dose regression like (1). Without pure control units, this strategy relies exclusively on comparisons across doses, which implicitly assumes that comparing "high" treatment groups with "low" treatment groups reveals the causal response to treatment, with attuenuation bias at worst. The remaining 30% of papers binarize the dose variable and estimate a traditional DD effect. This approach bins all "intensely" exposed units together and compares them with everyone else; the hope is that even if some control units receive a small dose, as long as the effect is monotonic this will reveal an attenuated version of the ATT. This approach is not cost free, however, as it throws away all dose information.

Both approaches implicitly assume that units with "low enough" doses do not actually experience treatment to make progress on causal inference. We can formalize this by stating that $\exists\ d_c > 0$ s.t.

$$\text{Treat}_i := \mu_i(D_i) \cdot 1\{D_i \geq d_c\} \tag{2}$$

where μ_i is a unit's dose response function (DRF), which we define in section 2, and $1\{\cdot\}$ is the indicator function. However, even under this assumption, both methods are biased. The main culprit is the fact that the true cutoff value beyond which treatment takes effect (d_c) is typically unknown to researchers and hence neglected from econometric analysis. We characterize the exact form of the bias under a range of data-generating processes in section 3.

For large jumps or sharp discontinuities in the marginal dose response at the cutoff value, we argue that researchers should instead jointly estimate the threshold along with treatment effect parameters using methods first developed by Hanson (1991, 1996). In this case, a parametric assumption on the shape of the dose response function is sufficient to identify the threshold and DRF parameters (Fong et al. 2010, 2017). These procedures restore identification and produce correct standard errors but involve non-standard inference due to non-identified parameters under a null hypothesis of no treatment effect. We show how to apply them to continuous treatment DD models, and under the parametric restriction of a linear dose effect beyond the unknown cutoff.

Although parametric methods offer one solution to the problem of an unknown cutoff value, they fail when parametric assumptions do not hold (by definition) or when the dose response function is flat around the true cutoff. In the former case, we only identify the least-squares best fit cutoff value, which is not equal to d_c. In the latter, while identification is not violated under the correct parametric restriction, cutoff estimation is unreliable without a large data set. In any case, if the researcher is unwilling to make a parametric restriction on the form of the dose response function, a standard "guess and binarize" approach leads to further issues. As a difference-in-differences estimator can only identify average treatment effects for treated units, changing who is considered "treated" changes the underlying estimand. Researchers targeting ATT-type parameters must acknowledge a major part of the data-generating process is unknown. We argue they should instead target an alternative estimand, the binned average treatment effect on the treated (BATT)

$$\text{BATT}(r) := E[\mu_i(D_i)|D_i \geq r] \tag{3}$$

for some percentile r. We define BATTs carefully in terms of potential outcomes in section 2, and connect them to other common estimands. For example, BATT(d_c) = ATT by definition.

One motivation to target BATTs also suggests an intuitive method to optimally estimate them: an approximately flat dose response function around the percentile used to define the BATT suggests a bias-variance trade-off when recategorizing low-dose units as controls. Doing so introduces bias into estimation (in expectation) because we are possibly misclassifying some treated units as controls. However, it reduces variance by bolstering a researcher-defined control group. This holds whenever the treated units are held fixed, which the BATT estimands accomplish, and leads to a "donut" estimator that is mean-squared error (MSE) optimal via refinements to a naive binarization. The MSE criterion function that defines the estimator also depends on the unknown d_c, which we overcome with both a minimax constraint and a simple method for partial identification.

This paper contributes to three literatures. Recent technical advances to difference-in-differences dealing with staggered adoption, heterogeneous treatement effects, pre-testing, and functional form specification are increasingly well understood (Goodman-Bacon, 2022; Sun and Abraham, 2022; de Chaisemartin and D'Haultfoeuille, 2020; Wooldridge, 2022; Roth, 2021; Callaway and Sant-Anna, 2021; Roth and Sant-Anna, 2021). However, identification issues related to continuous treatments are actively being litigated among econometricians and applied researchers alike (Callaway et al., 2021; de Chaisemartin et al., 2022; Sun and Shapiro, 2022; Butts, 2022).

The approach in Butts (2022) to estimate treatment effects at specific locations with geocoded data is most similar to ours. In studying treatment explicitly as continuous distance, it notes that researchers must know the threshold distance beyond (or below) which treatment effects begin, and proposes a non-parametric method to estimate a treatment effect curve (or what we call the dose response function) with large data. de Chaisemartin et al. (2022) identify causal responses with additional information on the treatment variable. If the dose changes over time and there exists a group of no-movers or a group of quasi-movers, then the average derivative among the treated is identified. No-movers are related to pure controls, as we have defined them, whereas quasi-movers are similar to the group identified by our "donut" estimator. Observing a changing dose is useful in assessing the existence of these groups; this is impossible to do in our setting, so we impose suitable restrictions on treatment heterogeneity to infer them. Sun and Shapiro (2022) provide impossibility results on identifying causal responses in TWFE regressions absent such additional information. They show that with the existence of a pure control, a modified instrumented difference-in-differences approach is sufficient to target an average of causal responses among treated units. We contribute to this work by providing a consistent framework to identify and estimate meaningful estimands without pure controls or observed alternative control groups, a case that is common in continuous difference-in-differences designs.

A second contribution is characterizing the bias with current practice that ignores the unknown cutoff value and providing solutions under parametric restrictions. Researchers typically assume attenuation bias at worst with current methods, but we show alternatively that bias of unknown sign is pervasive. In principle, our bias propositions in section 3 could be used in conjunction with hypothesized values for the true cutoff to bias-adjust estimates in current work given some priors. However, it is practically more useful to jointly identify and estimate the unknown cutoff value with relevant treatment effect parameters. We demonstrate how to use the methods developed in

Hanson (1991, 1996) and subsequent work with an intuitive parametric example.

Finally, our "donut" estimator of the BATT is derived under minimal adjustments to traditional Parallel Trends assumptions on untreated potential outcomes. The MSE problem can be intuitively characterized as a marginal decision of "how far up" the dose distribution to go when constructing a control group one unit at a time (subject to the existence of a true cutoff). This is similar to optimal bandwidth estimation in regression discontinuity designs (Calonico et al., 2019).

The issue of an unknown threshold of some *continuous* variable determining the *discrete* status of units in a sample is more pervasive than estimating treatment effects with unknown controls (as we highlight in this paper). For example, studies estimating treatment effects (or heterogeneous responses) related to size must determine what cutoff separates "large" units versus "small" units. Ivanov et al. (2022) show that, contrary to prior research, cuts to state corporate income taxes led to increases in corporate leverage, with these effects concentrated for small private firms. They define a firm to be small if its assets are below the sample median, which is a common approach in practice. Our method could be extended to account for an unknown latent cutoff value in the firm asset distribution that determines which firms are small enough to generate a heterogenous response that leads to higher leverage in response to tax cuts. The "donut estimator" would drop high asset firms below the median in defining a "small" firm cutoff, with MSE-gains for an objective function targeting differences in treatment effects by firm size.

2 Framework

This section outlines our framework, defining potential outcomes under a continuous treatment which only takes effect after some unknown threshold. We define treatment effect and causal response estimands, state the assumptions needed to identify them, and end with a discussion of how our framework connects with other papers in the difference-in-differences literature.

2.1 Potential Outcomes

Suppose there are two time periods $t \in \{\tau - 1, \tau\}$ with treatment at $t = \tau$, N units $i \in \{1, ..., N\}$, and a continuously distributed treatment, or dose, $D_i \in [d_L, d_U] \subset \mathbb{R}_+$. A true cutoff for treatment $d_c \in (d_L, d_U)$ defines the treated group $T_i = 1\{D_i \geq d_c\}$, while $P_t = 1\{t = \tau\}$ defines the treatment period. With potential outcomes $Y_{it}(D_i, T_i)$ we write observed outcomes each period

$$\begin{aligned}
Y_{it} &= (1-P_t)Y_{it}(D_i,0) + P_t\left[T_iY_{it}(D_i,1) + (1-T_i)Y_{it}(D_i,0)\right] \\
&= Y_{it}(D_i,0) + P_tT_i[\underbrace{Y_{it}(D_i,1) - Y_{it}(D_i,0)}_{:=\mu_i(D_i)}] \\
&= Y_{it}(D_i,0) + \mu_i(D_i)\cdot P_tT_i
\end{aligned}$$

where $\mu_i(D_i)$ is the unit-specific dose response function (hereafter the DRF).

2.2 Assumptions

We use the assumptions in this section to build our results in the paper. The first three involve basic restrictions on a two-period dataset of observed units (panel or repeated cross section) as well as no-anticipation in treatment effects prior to exposure (common in the difference-in-differences literature). Assumptions [A4] and [A5] then define weak and strong parallel trends in our setting, respectively, while [A6] and [A7] introduce homogeneity (of dose responses) or exogeneity (of dose assignments) that we argue are equivalent to strong parallel trends but easier to state.

A1 Random sampling: Data $\{Y_{i,\tau}, Y_{i,\tau-1}, D_i\}_{i=1}^N$ is iid

A2 Dose distribution: $D_i \sim F_D(d)$ over compact $\text{supp}\{D_i\} := [d_L, d_U] \subset \mathbb{R}_+$

A3 No anticipation: $Y_{i,\tau-1} = Y_{i,\tau-1}(D_i,0)$ and $Y_{i,\tau} = T_iY_{i\tau}(D_i,1) + (1-T_i)Y_{i\tau}(D_i,0)\ \forall i$

A4 W-PTA(d, d_c): Weak parallel trends *across* dose distribution given true cutoff

$$E[\Delta Y_{it}(D_i,0)|T_i=1, D_i=d] \quad = \quad E[\Delta Y_{it}(D_i,0)|T_i=0, D_i=d']$$

$\forall (d,d') \in [d_L, d_c) \times (d_c, d_U]$, or the equivalent parametric version that $Y_{it}(D_i,0) = \alpha_i + \theta_t + \varepsilon_{it}$ whenever $t = \tau - 1$ or $T_i = 0$, where $\alpha_i, \theta_t, \varepsilon_{it}$ denote unit, time, and idiosyncratic effects.

A5 S-PTA(d): Strong parallel trends

$$E[Y_{it}(D_i,1) - Y_{i,t-1}(D_i,0)] - E[Y_{it}(D_i,1) - Y_{i,t-1}(D_i,0)|D_i=d] = 0 \quad \forall d \in (d_L, d_U)$$

A6 Dose response function (DRF) homogeneity, and functional form structure

$$\mu_i(D_i) = \mu(D_i) \quad \forall i$$

A6.1 Binary $\mu(D_i) = \beta_1$

A6.2 Jump linear $\mu(D_i) = \beta_1 + \beta_2(D_i - d_c)$

A6.2′ Connected linear $\mu(D_i) = \beta_2(D_i - d_c)$

A6.3 Smoothness $\mu''(D_i) \neq 0$

A7 Random dose assignment w.r.t. DRFs: $D_i \perp \mu_i | \alpha_i \Rightarrow \mu(D_i|d) = \mu(D_i)$

2.3 Estimands

The estimands are conditional averages of DRFs, or the slopes of those averages

$$\begin{aligned} \mu(a|d) &:= E[Y_{it}(a,1) - Y_{it}(a,0)|D_i = d] = E[\mu_i(a)|D_i = d] \\ \mu'(a|d) &:= E\left[\frac{\partial}{\partial a}\mu_i(a)\middle| D_i = d\right] \end{aligned}$$

across the dose distribution.[1] For each dose $d \in (d_L, d_U)$, we target a discrete set of estimands indexed by a sequence of doses $\ell_j \in \{\ell_1, ..., \ell_J\} \subset (d_L, d_U)$

$$\begin{aligned} \overrightarrow{\mu}_d &:= \{\mu(\ell_1|d), ..., \mu(\ell_J|d)\} \quad \forall d \\ \overrightarrow{\mu}'_d &:= \{\mu'(\ell_1|d), ..., \mu'(\ell_J|d)\} \quad \forall d \end{aligned}$$

Additionally, we define the binned average treatment effect on the treated (BATT) estimands

$$\begin{aligned} \text{BATT}(r) &:= E[Y_{it}(D_i,1) - Y_{it}(D_i,0)|D_i \geq r] = E[\mu_i(D_i)|D_i \geq r] \\ \text{BATT}(r,k) &:= E[Y_{it}(D_i,1) - Y_{it}(D_i,0)|D_i \in (r,k)] \end{aligned}$$

which are meaningful estimands even when the true cutoff d_c is unknown.

[1] Dominated convergence theorem issue, our ACRs not same as CGS (2021). Highlight this better.

2.4 Discussion

Remark 1: Assumptions [A1]-[A3] with [A4] deliver the familiar TWFE specification

$$\begin{aligned} Y_{it} &= Y_{it}(D_i, 0) + \mu_i(D_i) \cdot P_t T_i \\ &= \alpha_i + \theta_t + \mu_i(D_i) \cdot P_t T_i + \varepsilon_{it} \end{aligned}$$

under additional restrictions on the DRFs.

Remark 2: With two time periods, TWFE is equivalent to a bivariate regression on first differences

$$\begin{aligned} Y_{it} &= \alpha_i + \theta_t + \mu_i(D_i) \cdot P_t T_i + \varepsilon_{it} \\ \Delta Y_{it} &= \Delta\theta_t + \mu_i(D_i) \cdot T_i + \Delta\varepsilon_{it} \end{aligned}$$

which we use throughout the paper (as in CGS).

Remark 3: The error in the TWFE regression above is structural since we fully specify the data-generating process for *individual* potential outcomes, rather than targeting some "best" linear approximation of the *population* conditional expectation function of outcomes. Practitioners often interpret these regressions in a population sense. In these cases, we can interpret the error term as including both a structural part and a specification part, so that

$$\begin{aligned} \Delta Y_{it} = E[\Delta Y_{it}|D_i] + \epsilon_i &= E\left[\Delta\theta_t + \mu_i(D_i) \cdot T_i + \Delta\varepsilon_{it} | D_i = d\right] + \epsilon_i \\ &= \Delta\theta_t + E\left[\mu_i(D_i)|D_i\right] \cdot T_i + \Delta\varepsilon_{it} + \epsilon_i \end{aligned}$$

where $\epsilon_i := T_i\left(\mu_i(D_i) - E[\mu_i(D_i)|D_i]\right)$ is mean independent of D_i.

Remark 4: Under W-PTA [A4], assumptions [A6] or [A7] are equivalent to S-PTA [A5]. When untreated potential outcomes evolve equivalently, restricting counterfactual treatment effect heterogeneity is the same as randomly assigning doses among units or imposing homogeneous dose responses across units. All three are equivalent to assuming away selection bias when trying to infer causal *responses* from the difference of estimated causal *effects*. This is a clarification of one of the main results in CGS (2021): it is difficult to infer to ACRs from dose comparisons without strong assumptions. S-PTA is a difficult assumption to state and understand, whereas homogeneity of effects or exogeneity of assignment clarify just how strong it is in practice.

3 Current Practice

This section presents our metastudy of continuous treatment difference-in-differences to characterize current practice: (i) estimate full dose regression or, (ii) create binary treatment variable with a chosen cutoff (like the median). Using our framework, we show what happens when practitioners do not know true cutoff d_c prior to estimation, sometimes also guessing a value d_r.

3.1 Metastudy

Using Google Scholar, we query all papers published in the American Economic Review since 2000 which contain the keywords "difference-in-difference" and "continuous" in the manuscript. We find 178 total papers that satisfy these requirements and after checking each one, retain only those that estimate some type of continuous treatment difference-in-difference parameter that we have defined. There is clear pattern in applied work: 70% estimate a full dose regression like (1) while 30% binarize and compare units above and below some researcher-defined cutoff. Many papers estimating a dose regression also binarize as a robustness check. The next sections describe the bias in these approaches when the cutoff d_c is ignored (all proofs are in Appendix A).

3.2 Full dose regression

The most common approach in our metastudy involves estimating the equation

$$\Delta Y_{it} = b_0 + b_1 \cdot D_i + e_{it}^{\text{DOSE}}$$

via least squares, which only uses dose comparisons for causal inference on the effect of treatment (no pure control units assumed by design). Propositions [P1] and [P2] illustrate the bias in $\hat{b}_1$.

P1 Suppose that assumptions A1-A3 and A6 (DRF homogeneity) hold. Then the least squares estimate of the slope parameter of a dose regression satisfies

$$\begin{aligned} \widehat{b}_1^{\text{DOSE}} &\rightarrow^p \beta_1 \cdot \frac{(\overline{D} - \underline{D})\text{Var}(T_i)}{\text{Var}(D_i)} && \text{under A6.1 (binary DGP)} \\ &\rightarrow^p \beta_2 \cdot \left[1 - F_D(d_c) + \frac{\heartsuit(T_i, \overline{D}, \underline{D}, d_c)}{\text{Var}(D_i)}\right] && \text{under A6.2' (connected linear)} \\ &\rightarrow^p \text{plim}(\widehat{b}_1^{\text{DOSE}})_{\text{A6.1}} + \text{plim}(\widehat{b}_1^{\text{DOSE}})_{\text{A6.2'}} && \text{under A6.2 (jump linear)} \end{aligned}$$

where $\overline{D} := E[D_i | D_i \geq d_c]$ and $\underline{D} := E[D_i | D_i < d_c]$ and

$$\heartsuit(T_i, \overline{D}, \underline{D}, d_c) := \left[\overline{D}^2 + \underline{D}^2\right] \cdot \text{Cov}(T_i, T_i^2) + \left[\overline{D}\underline{D} - \underline{D}^2 + d_c(\underline{D} - \overline{D})\right] \cdot \text{Var}(T_i)$$

P2 [CGS results] Suppose that assumptions A1 and A2 hold. Then the least squares estimate

$$\begin{aligned}
\widehat{b}_1^{\text{DOSE}} &\rightarrow^p \int_{d_L}^{d_U} w(\ell) \cdot \frac{\partial E[\Delta Y_{it} | D_i = \ell]}{\partial \ell} d\ell \\
&\rightarrow^p \int_{d_L}^{d_U} w(\ell) \cdot \left[\mu'(\ell|\ell) + \frac{\partial}{\partial h}\mu(\ell|h)|_{h=\ell}\right] d\ell \quad \text{under A3, A4 (W-PTA), and A6.3 (smoothness)} \\
&\rightarrow^p \int_{d_L}^{d_U} w(\ell) \cdot \mu'(\ell) d\ell \quad \text{under A3, A5 (S-PTA), and A6.3 (smoothness)}
\end{aligned}$$

where the weights are all non-negative and integrate to 1:

$$w(\ell) := \frac{(E[D_i | D_i \geq \ell] - E[D_i]) \cdot (1 - F_D(\ell)))}{\text{Var}(D_i)}$$

What do practitioners have in mind with this approach? Comparing "high" versus "low" treated units should reveal the causal response to treatment, perhaps with some attenuation bias. However, the main results in CGS (2021) endure: these comparisons are fraught without much stronger homogeneity restrictions on treatment effects. Our propositions reveal how this bias manifests when the cutoff separating the treated and control units is omitted from analysis.

3.3 Guess and binarize

The remaining 30% of papers in our metastudy estimate the equation

$$\Delta Y_{it} = b_0 + b_1 \cdot 1\{D_i \geq d_r\} + e_{it}^{\text{BIN}}$$

where $d_r \in (d_L, d_U)$ is researcher guess of cutoff. The following results illustrate the bias in $\hat{b}_1$.

P3 Suppose that assumptions A1-A3, A4 (W-PTA), and A6 (DRF homogeneity) hold. Then the least squares estimate of the slope parameter satisfies

$$
\begin{aligned}
\widehat{b}_1^{\text{BIN}} &\to^p \begin{cases} \beta_1 \cdot \frac{1-F_D(d_c)}{1-F_D(d_r)} & \text{if } d_r < d_c \\ \beta_1 \cdot \left[1 - \frac{F_D(d_r)-F_D(d_c)}{F_D(d_r)}\right] & \text{if } d_r \geq d_c \end{cases} && \text{under A6.1 (binary DGP)} \\
&\to^p \begin{cases} \beta_2 \cdot (\overline{D}_r - d_c)\frac{1-F_D(d_c)}{1-F_D(d_r)} & \text{if } d_r < d_c \\ \beta_2 \cdot \left[\overline{D}_r - \frac{F_D(d_r)-F_D(d_c)}{F_D(d_r)} \underline{D}_r^{T_i=1} - d_c \frac{F_D(d_c)}{F_D(d_r)}\right] & \text{if } d_r \geq d_c \end{cases} && \text{under A6.2' (connected linear)} \\
&\to^p \begin{cases} (\beta_1 + \beta_2) \cdot (\overline{D}_r - d_c)\frac{1-F_D(d_c)}{1-F_D(d_r)} & \text{if } d_r < d_c \\ \text{plim}(\widehat{b}_1^{\text{BIN}})_{\text{A6.1}} + \text{plim}(\widehat{b}_1^{\text{BIN}})_{\text{A6.2'}} & \text{if } d_r \geq d_c \end{cases} && \text{under A6.2 (jump linear)}
\end{aligned}
$$

P4 Suppose that assumptions A1-A3 and A7 (mean independence) hold. Then

$$
\widehat{b}_1^{\text{BIN}} \to^p \begin{cases} \beta_r \cdot \frac{1-F_D(d_c)}{1-F_D(d_r)} & \text{if } d_r < d_c \\ \beta_r - \beta_{cr} \cdot \frac{F_D(d_r)-F_D(d_c)}{F_D(d_r)} & \text{if } d_r \geq d_c \end{cases} \qquad \text{under A6.3 (smoothness of the DRF)}
$$

where we additionally define the estimands

$$
\begin{aligned}
\beta_r &:= \mu(D_i | D_i \geq d_r) \\
\beta_{cr} &:= \mu(D_i | d_c < D_i < d_r)
\end{aligned}
$$

What do practitioners have in mind with this approach? Binning "intensely" exposed units and comparing them to *everyone* below is one way to deal with CGS (2021) issues. Note that it is a costly solution, as it throws away all dose information. (This is one motivation for our general theory in the following section, which essentially uses the dose information to build an "optimal" control sample given some percentile-defined treatment group, or BATT.) Moreover, our propositions show that bias of unknown sign characterizes these estimates as well.

All four propositions show that standard TWFE estimates contain bias when the true, unknown cutoff value beyond which treatment starts d_c is ignored from econometric analysis. One clear way to solve this problem is confront it directly and estimate the cutoff along with any desired treatment effect parameters. This approach imposes a parametric form on the dose response function, but involves non-standard econometrics for inference, as the cutoff parameter is undefined under any null hypothesis of zero treatment effects at the threshold by definition, which implies test statistics exhibit non-standard limiting distributions. Hanson (1996, 2000) show how to simulate valid p-values, and we illustrate this approach in our setting by assuming a "partial dose" design where only units above $d_c > d_L$ experience a constant effect β_2 given their dose D_i.

3.4 Parametric Solution

This approach involves estimating treatment effects and the threshold d_c simultaneously

$$(\widehat{\beta}, \widehat{d_c}) := \underset{\beta, d}{\text{argmin}} \sum_{i=1}^{N} [\Delta Y_{it} - \beta_0 - \mu_i(D_i; \beta_1, \beta_2) \cdot 1\{D_i \geq d_c\}]^2$$

using methods in Hanson (1996). We specify $\mu_i(D_i; \beta_1, \beta_2) = \beta_1 + \beta_2 D_i$ as a linear "partial dose" approach, allowing for a possible jump in treatment effects beyond the cutoff (β_1) as well as a constant multiple effect given their dose (β_2). To do!

4 General theory

One weakness of the parametric approach is that it involves strong restrictions on μ to implement in practice. Of course, one option is to non-parametrically trace out the dose response function, as in Butts (2022) where treatment is defined as distance to some location. This requires large amounts of data across the full support of the dose distribution, which is not always available. Even if the dose response function μ is known, this method struggles when μ is flat around true cutoff d_c (without access to increasing amounts of data). Moreover, the estimand is hard to interpret without knowing the true cutoff since it is impossible to define the treated group. We suggest that researchers instead target a set of BATT(r) estimands, which are attractive from a meta-study perspective as they are consistent across different papers without "pure control" units.

A flat dose response function around the cutoff also implies an intuitive bias-variance trade-off, which we exploit to derive a Mean Square Error (MSE)-optimal "donut" estimator of the slightly different (but stable and meaningful) BATT estimands. Our approach is a compromise between the old way of doing things (use the dose for causal inference, which CGS have shown is bad) and binarizations that throw away all dose information (which we show are all biased when d_c is unknown). Instead of the ATT, which is unidentified without pure controls, our "donut" estimator targets BATT(r) and only uses dose information to create a better (MSE-optimal) control group (rather than using it for causal inference, which is fraught without strong assumptions). This is superior to binarizing, which uses all units below the researcher-defined percentile r as controls.

4.1 "Donut" Estimator

The following result is the strongest in the paper, under the weakest assumptions. It also illustrates the gains to targeting alternative estimand with unknown cutoff. If you guess low, then you are guaranteed to be unbiased. If you guess high, then you are guaranteed attenuation bias at worst.

[T1]

Suppose that assumptions [A1]-[A3], and [A4] (W-PTA) hold. Then

$$E\left[\widehat{b}_1^{\text{BIN}}\right] = \begin{cases} \text{BATT}(r) & \text{if } d_r < d_c \\ \text{BATT}(r) - \text{BATT}(d_c, d_r) \cdot P(D_i \geq d_c | D_i < d_r) & \text{if } d_r \geq d_c \end{cases}$$

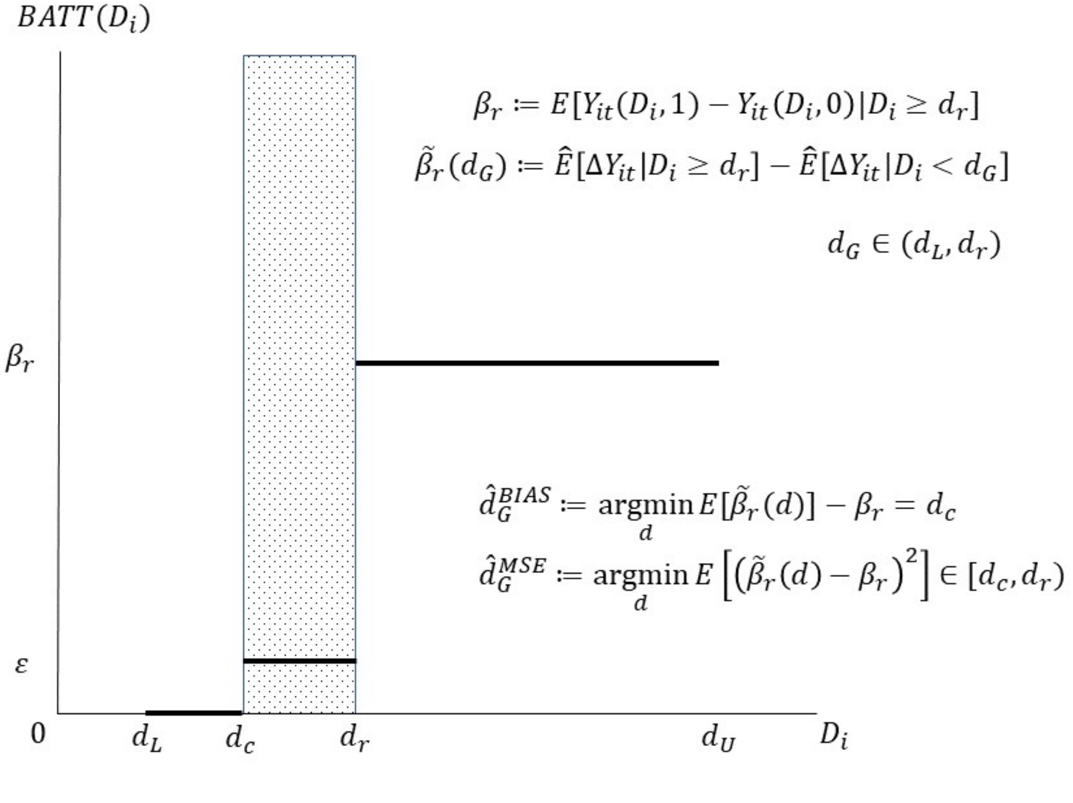
Remark 5: Both Theorem 1 and Proposition 4 show how estimates based on binarizing at a percentile behave under different assumptions (W-PTA only for T1, random dose assignment for P4). Need to reconcile... they should both be targeting BATTs same way I think? (Discuss with Thomas)

The “donut” estimator of the BATT amounts to de-biasing using Theorem 1 under a suitable MSE criterion function! If d_G denotes the cutoff defining control units (choice variable), $\widehat{\beta}_r^{\text{DON}}$ denotes the donut estimator, and $\beta_r = \text{BATT}(r)$, we can use Theorem 1 and some tinkering to write

$$\begin{aligned}
\text{Bias}(\widehat{\beta}_r^{\text{DON}}, \beta_r) &= E\left[\widehat{\beta}_r^{\text{DON}}\right] - \beta_r = -\frac{F_D(d_G) - F_D(d_c)}{F_D(d_c)} \cdot \int_{d_c}^{d_G} \frac{\Delta Y_{it}(D_i, 1) f_D(k) dk}{F_D(d_G) - F_D(d_c)} \\
\text{Var}(\widehat{\beta}_r^{\text{DON}}) &= \frac{1}{n} \frac{\sigma^2(d_c, d_g, d_r) + \sigma^2_{\triangle\varepsilon}}{\frac{(1-F(d_r))F(d_r)}{[1-F(d_r)+F(d_g)]^2}} \\
\sigma^2(d_c, d_g, d_r) &= \int_{d_l}^{d_u} [\triangle\theta_t + \mu(k)\mathbb{1}(k \geq d_c) - \hat{b}_0^{MSE} - \hat{b}_1^{MSE}\mathbb{1}(k \geq d_r)]^2 \mathbb{1}(k \notin (d_g, d_r)) f(k) \text{d}k
\end{aligned}$$

which defines the optimal donut estimator of BATT(r)

$$\begin{aligned}
\widehat{d}_G &:= \underset{d}{\text{argmin}}\, \text{MSE}(\widehat{\beta}_r^{\text{DON}}, \beta_r) \\
\frac{\partial}{\partial d_G}\text{Bias}(\widehat{\beta}_r^{\text{DON}}, \beta_r)^2 &= -\frac{\partial}{\partial d_G}\text{Var}(\widehat{\beta}_r^{\text{DON}})
\end{aligned}$$


$BATT(D_i)$
$\beta_r := E[Y_{it}(D_i,1) - Y_{it}(D_i,0)|D_i \geq d_r]$
$\tilde{\beta}_r(d_G) := \hat{E}[\Delta Y_{it}|D_i \geq d_r] - \hat{E}[\Delta Y_{it}|D_i < d_G]$
$d_G \in (d_L, d_r)$
β_r
$\hat{d}_G^{BIAS} := \underset{d}{\text{argmin}}\, E[\tilde{\beta}_r(d)] - \beta_r = d_c$
$\hat{d}_G^{MSE} := \underset{d}{\text{argmin}}\, E\left[\left(\tilde{\beta}_r(d) - \beta_r\right)^2\right] \in [d_c, d_r)$
ε
0
d_L
d_c
d_r
d_U
D_i

The donut estimator is simple to define, and simple to estimate. In contrast, choosing the MSE-optimal cutoff for control group is challenging because the unknown parameter d_c appears in the problem. We propose two solutions to deal with this: a minimax loss procedure, and a partial identification approach, taking minimum and maximum estimated treatment effects to bound TE parameters by solving the MSE problem across a range of cutoffs.

4.2 Minimax Procedure

4.3 Partial Identification Approach

5 Application

Ideas: Federal minimum wage, Vietnam draft lottery?

Ideally: some setting where we expect μ is flat somewhere around the true cutoff! This is where our adjustments should make the most (first order) difference relative to current methods.

Mel suggestion: deep dive into metastudy rather than doing one replication

6 Conclusion

Focusing on ATTs as the relevant building block for causal estimands, rather than ACRs, is critical to our approach. It facilitates a clean definition of BATT, a meaningful estimand, when the ATT is unidentified otherwise. Intuition for when this should matter: low treatment effects among low-dose units. When binarizing, do you think that marginal units below your chosen cutoff experience *at least some* treatment effect? If no, then you know the unknowable d_c! If yes, then you can clearly stand to gain precision in treatment effect estimation by dropping those units, which is what the "donut" estimator does. Our next step is to solve the MSE problem given the unknown cutoff, illustrate the method with an application, and formalize a randomization inference test to insure against $d_r < d_c$, which is important for MSE gains to exist by Theorem 1.

7 References

Angrist, Joshua and Jorn-Steffen Pischke. 2008. Mostly Harmless Econometrics. Book.

Butts, Kyle. 2022. Difference-in-Differences with Geocoded Microdata. Journal of Urban Economics.

Callaway, Brantly and Pedro HC Sant-Anna. 2021. Difference-in-differences with Multiple Time Periods. Working Paper.

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro HC Sant'Anna. 2021. Differences-in-differences with a continuous treatment. Working paper.

Calonico, Sebastian, Matias Cattaneo, and Max Farrell. 2019 Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. The Econometrics Journal.

de Chaisemartin, Clément and Xavier D'Haultfoeuille. 2020. Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. American Economic Review.

de Chaisemartin, Clément, Xavier D'Haultfoeuille, Felix Pasquier, & Gonzalo Vasquez-Bare. 2022. Difference-in-Differences Estimators for Treatments Continuously Distributed at Every Period.

Fong, Youyi, Chongzhi Di, and Sallie Permar. Change Point Testing in Logistic Regression Models with Interaction Term. Stat Med.

Fong, Youyi, Ying Huang, Peter Gilbert, and Sallie Permar.chngpt: Threshold Regression Model Estimation and Inference. BMC Bioinformatics.

Goodman-Bacon, Andrew. 2020. Difference-in-differences with variation in treatment timing. Journal of Econometrics.

Hanson, Bruce. 1991. Inference When a Nuisance Parameter Is Not Identified Under the Null Hypothesis. Working Paper.

Hanson, Bruce. 1996. Inference When a Nuisance Parameter Is Not Identified Under the Null Hypothesis. Econometrica.

Ivanov, Ivan, Luke Pettit, and Toni Whited. 2022. Taxes Depress Corporate Borrowing: Evidence from Private Firms. Working Paper.

Roth, Jonathan and Pedro HC Sant-Anna. 2021. When is Parallel Trends Sensitive to Functional Form? Working Paper.

Roth, Jonathan. 2021. Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends. Working Paper.

Sun, Liyang and Jesse Shapiro. 2022. A Linear Panel Model with Heterogenous Coefficients and Variation in Exposure. Working Paper.

Sun, Liyang and Sarah Abraham. 2022. Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects. Journal of Econometrics.

Wooldridge, Jeffrey. 2021. Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators. Working Paper.

8 Appendix A: Proofs

[P1]

By the weak law of large numbers and Slutsky's theorem

$$\begin{aligned}
\widehat{b}_1^{\text{DOSE}} &= \frac{\widehat{\text{Cov}}(D_i, \Delta Y_{it})}{\widehat{\text{Var}}(D_i)} \\
&\to^p \frac{\text{Cov}(D_i, \Delta Y_{it})}{\text{Var}(D_i)} \\
&= \frac{\text{Cov}(D_i, \Delta\theta_t + \mu(D_i)T_i + \Delta\varepsilon_{it})}{\text{Var}(D_i)} \\
&= \frac{\text{Cov}(D_i, \mu(D_i)T_i)}{\text{Var}(D_i)} \\
&= \frac{E\left[\text{Cov}(D_i, \mu(D_i)T_i|T_i)\right] + \text{Cov}\left(E[D_i|T_i], E[\mu(D_i)T_i|T_i]\right)}{\text{Var}(D_i)} \\
&= \frac{\heartsuit_{\mu,D} \cdot (1 - F_d(d_c)) + \text{Cov}(\overline{D}T_i, \square_{\mu,D}) - \text{Cov}(\underline{D}T_i, \square_{\mu,D})}{\text{Var}(D_i)}
\end{aligned}$$

where we use $E[D_i|T_i] = \overline{D}T_i + \underline{D}(1 - T_i)$ and $\text{Cov}(D_i, \mu(D_i)T_i|T_i = 0) = 0$ and define

$$\begin{aligned}
\heartsuit_{\mu,D} &:= \text{Cov}(D_i, \mu(D_i)T_i|T_i = 1) \\
\square_{\mu,D} &:= E[\mu(D_i)T_i|T_i]
\end{aligned}$$

Different assumptions about the DRF yield different expressions. Under a binary DGP

$$\begin{aligned}
\heartsuit_{\mu,D}^{\text{A6.1}} &:= \text{Cov}(D_i, \beta_1 T_i|T_i = 1) = 0 \\
\square_{\mu,D}^{\text{A6.1}} &:= E[\beta_1 T_i|T_i] = \beta_1 T_i
\end{aligned}$$

Substituting into our expression for the probability limit proves the first part

$$\begin{aligned}
\widehat{b}_1^{\text{DOSE}} &\to^p \frac{0 \cdot (1 - F_d(d_c)) + \text{Cov}(\overline{D}T_i, \beta_1 T_i) - \text{Cov}(\underline{D}T_i, \beta_1 T_i)}{\text{Var}(D_i)} \\
&= \frac{\overline{D}\beta_1\text{Var}(T_i) - \underline{D}\beta_1\text{Var}(T_i)}{\text{Var}(D_i)}
\end{aligned}$$

$$= \quad \beta_1 \cdot \frac{(\overline{D} - \underline{D})\mathrm{Var}(T_i)}{\mathrm{Var}(D_i)}$$

Under a connected linear DGP, we have

$$\begin{aligned}
\heartsuit^{\text{A6.2}'}_{\mu,D} &:= \mathrm{Cov}(D_i, \beta_2(D_i - d_c)T_i | T_i = 1) = \beta_2 \mathrm{Var}(D_i) \\
\square^{\text{A6.2}'}_{\mu,D} &:= E[\beta_2(D_i - d_c)T_i | T_i] = E[\beta_2 D_i T_i | T_i] - E[\beta_2 d_c T_i | T_i] \\
&= \beta_2 T_i \cdot E[D_i | T_i] - \beta_2 d_c T_i \\
&= \beta_2 T_i \cdot (\overline{D} T_i + \underline{D}(1 - T_i)) - \beta_2 d_c T_i \\
&= \beta_2 T_i (\overline{D} T_i + \underline{D}(1 - T_i) - d_c)
\end{aligned}$$

Using these quantities, we can compute the two covariance terms

$$\begin{aligned}
\mathrm{Cov}(\overline{D} T_i, \square^{\text{A6.2}'}_{\mu,D}) &= \mathrm{Cov}(\overline{D} T_i, \beta_2 T_i(\overline{D} T_i + \underline{D}(1 - T_i) - d_c)) \\
&= \mathrm{Cov}(\overline{D} T_i, \beta_2 \overline{D} T_i^2) + \mathrm{Cov}(\overline{D} T_i, \beta_2 T_i \underline{D}(1 - T_i)) - \mathrm{Cov}(\overline{D} T_i, \beta_2 T_i d_c) \\
&= \beta_2 \overline{D}^2 \mathrm{Cov}(T_i, T_i^2) + \beta_2 \overline{D}\underline{D} \underbrace{\mathrm{Cov}(T_i, T_i(1 - T_i))}_{= \mathrm{Var}(T_i) - \mathrm{Cov}(T_i, T_i^2)} - \beta_2 \overline{D} d_c \mathrm{Var}(T_i) \\
&= \beta_2(\overline{D}^2 - \overline{D}\underline{D})\mathrm{Cov}(T_i, T_i^2) + \beta_2(\overline{D}\underline{D} - \overline{D} d_c)\mathrm{Var}(T_i) =: A \\
\mathrm{Cov}(\underline{D} T_i, \square^{\text{A6.2}'}_{\mu,D}) &= \beta_2 \underline{D}\overline{D} \mathrm{Cov}(T_i, T_i^2) + \beta_2 \underline{D}^2 \mathrm{Cov}(T_i, T_i(1 - T_i)) - \beta_2 \underline{D} d_c \mathrm{Var}(T_i) \\
&= \beta_2(\underline{D}\overline{D} - \underline{D}^2)\mathrm{Cov}(T_i, T_i^2) + \beta_2(\underline{D}^2 - \underline{D} d_c)\mathrm{Var}(T_i) =: B
\end{aligned}$$

and take their difference

$$\begin{aligned}
A - B &= \beta_2 \left[(\overline{D}^2 + \underline{D}^2)\mathrm{Cov}(T_i, T_i^2) + (\overline{D}\underline{D} - \underline{D}^2 + d_c(\underline{D} - \overline{D}))\mathrm{Var}(T_i) \right] \\
&= \beta_2 \cdot \heartsuit(T_i, \overline{D}, \underline{D}, d_c)
\end{aligned}$$

which implies

$$\begin{aligned}
\widehat{b}_1^{\text{DOSE}} &\to^p \frac{\beta_2 \mathrm{Var}(D_i) \cdot (1 - F_d(d_c)) + \beta_2 \cdot \heartsuit(T_i, \overline{D}, \underline{D}, d_c)}{\mathrm{Var}(D_i)} \\
&= \beta_2 \cdot \left[1 - F_D(d_c) + \frac{\heartsuit(T_i, \overline{D}, \underline{D}, d_c)}{\mathrm{Var}(D_i)} \right]
\end{aligned}$$

Under a jump linear DGP, we have

$$\begin{aligned}
\heartsuit^{\text{A6.2}}_{\mu,D} &:= \text{Cov}(D_i, (\beta_1 + \beta_2(D_i - d_c))T_i|T_i = 1) \\
&= \text{Cov}(D_i, \beta_1 T_i|T_i = 1) + \text{Cov}(D_i, \beta_2(D_i - d_c))T_i|T_i = 1) \\
&= 0 + \beta_2 \text{Var}(D_i) \\
\square^{\text{A6.2}}_{\mu,D} &:= E[(\beta_1 + \beta_2(D_i - d_c))T_i|T_i] \\
&= E[\beta_1 T_i|T_i] + E[\beta_2(D_i - d_c)T_i|T_i] \\
&= \beta_1 T_i + \beta_2 T_i(\overline{D}T_i + \underline{D}(1 - T_i) - d_c)
\end{aligned}$$

which we use to compute the covariance terms

$$\begin{aligned}
\text{Cov}(\overline{D}T_i, \square^{\text{A6.2}}_{\mu,D}) &= \text{Cov}(\overline{D}T_i, \beta_1 T_i + \beta_2 T_i(\overline{D}T_i + \underline{D}(1 - T_i) - d_c)) \\
&= \beta_1 \overline{D}\text{Var}(T_i) + \text{Cov}(\overline{D}T_i, \beta_2 T_i(\overline{D}T_i + \underline{D}(1 - T_i) - d_c)) \\
&= \beta_1 \overline{D}\text{Var}(T_i) + \beta_2(\overline{D}^2 - \overline{D}\underline{D})\text{Cov}(T_i, T_i^2) + \beta_2(\overline{D}\underline{D} - \overline{D}d_c)\text{Var}(T_i) \\
\text{Cov}(\underline{D}T_i, \square^{\text{A6.2}}_{\mu,D}) &= \text{Cov}(\underline{D}T_i, \beta_1 T_i + \beta_2 T_i(\overline{D}T_i + \underline{D}(1 - T_i) - d_c)) \\
&= \beta_1 \underline{D}\text{Var}(T_i) + \text{Cov}(\underline{D}T_i, \beta_2 T_i(\overline{D}T_i + \underline{D}(1 - T_i) - d_c)) \\
&= \beta_1 \underline{D}\text{Var}(T_i) + \beta_2(\underline{D}\overline{D} - \underline{D}^2)\text{Cov}(T_i, T_i^2) + \beta_2(\underline{D}^2 - \underline{D}d_c)\text{Var}(T_i)
\end{aligned}$$

and the probability limit is the sum of the previous two

$$\begin{aligned}
\widehat{b}_1^{\text{DOSE}} &\to^p \frac{\beta_2 \text{Var}(D_i) \cdot (1 - F_d(d_c)) + \beta_1(\overline{D} - \underline{D})\text{Var}(T_i) + \beta_2 \cdot \heartsuit(T_i, \overline{D}, \underline{D}, d_c)}{\text{Var}(D_i)} \\
&= \beta_2 \cdot \left[1 - F_D(d_c) + \frac{\heartsuit(T_i, \overline{D}, \underline{D}, d_c)}{\text{Var}(D_i)}\right] + \beta_1 \cdot \frac{(\overline{D} - \underline{D})\text{Var}(T_i)}{\text{Var}(D_i)}
\end{aligned}$$

[P2]

All results follow immediately from CGS (2022) propositions 4, 5, and 10.

[P3]

By the weak law of large numbers and Slutsky's theorem

$$\widehat{b}_1^{\text{BIN}} = \frac{\widehat{\text{Cov}}(R_i, \Delta Y_{it})}{\widehat{\text{Var}}(R_i)}$$

$$\begin{aligned}
&\to^p \quad \frac{\text{Cov}(R_i, \Delta Y_{it})}{\text{Var}(R_i)} \\
&= \quad E[\Delta Y_{it}|R_i = 1] - E[\Delta Y_{it}|R_i = 0] \\
&= \quad E[\mu(D_i)T_i|R_i = 1] - E[\mu(D_i)T_i|R_i = 0]
\end{aligned}$$

Suppose that $d_r < d_c$. Then we have

$$\begin{aligned}
\widehat{b}_1^{\text{BIN}} \quad &\to^p \quad E[\mu(D_i)T_i|R_i = 1] - E[\mu(D_i)T_i|R_i = 0] \\
&= \quad E[\mu(D_i)T_i|R_i = 1, T_i = 0]P(T_i = 0|R_i = 1) + E[\mu(D_i)T_i|R_i = 1, T_i = 1]P(T_i = 1|R_i = 1) - 0 \\
&= \quad E[\mu(D_i)T_i|R_i = 1, T_i = 1]P(D_i \geq d_c|D_i \geq d_r) \\
&= \quad \frac{1 - F_D(d_c)}{1 - F_D(d_r)} \cdot E[\mu(D_i)T_i|R_i = 1, T_i = 1] \\
&= \quad \frac{1 - F_D(d_c)}{1 - F_D(d_r)} \cdot \begin{cases} \beta_1 & \text{if A6.1 holds} \\ \beta_2(\overline{D}_r - d_c) & \text{if A6.2}' \text{ holds} \\ (\beta_1 + \beta_2) \cdot (\overline{D}_r - d_c) & \text{if A6.2 holds} \end{cases}
\end{aligned}$$

Alternatively, assume that $d_r \geq d_c$. Then

$$\begin{aligned}
\widehat{b}_1^{\text{BIN}} \quad &\to^p \quad E[\mu(D_i)T_i|R_i = 1] - E[\mu(D_i)T_i|R_i = 0] \\
&= \quad E[\mu(D_i)T_i|R_i = 1] - \\
&\quad\quad E[\mu(D_i)T_i|R_i = 0, T_i = 1]P(T_i = 1|R_i = 0) - E[\mu(D_i)T_i|R_i = 0, T_i = 0]P(T_i = 0|R_i = 0) \\
&= \quad E[\mu(D_i)T_i|R_i = 1] - E[\mu(D_i)T_i|R_i = 0, T_i = 1]P(T_i = 1|R_i = 0) - 0 \\
&= \quad E[\mu(D_i)T_i|R_i = 1] - E[\mu(D_i)T_i|R_i = 0, T_i = 1] \cdot \frac{F_D(d_r) - F_D(d_c)}{F_D(d_r)} \\
&= \quad \begin{cases} \beta_1 \cdot \left[1 - \frac{F_D(d_r) - F_D(d_c)}{F_D(d_r)}\right] & \text{if A6.1 holds} \\ \beta_2 \cdot \left[\overline{D}_r - \frac{F_D(d_r) - F_D(d_c)}{F_D(d_r)} \underline{D}_r^{T_i=1} - d_c \frac{F_D(d_c)}{F_d(d_r)}\right] & \text{if A6.2}' \text{ holds} \\ \text{plim}(\widehat{b}_1^{\text{BIN}})_{\text{A6.1}} + \text{plim}(\widehat{b}_1^{\text{BIN}})_{\text{A6.2}'} & \text{if A6.2 holds} \end{cases}
\end{aligned}$$

[P4]

Suppose that $d_r < d_c$. Then we have

$$\widehat{b}_1^{\text{BIN}} \quad \to^p \quad E[\mu(D_i)T_i|R_i = 1] - E[\mu(D_i)T_i|R_i = 0]$$

$$\begin{aligned}
&= E[\mu(D_i)T_i|R_i = 1, T_i = 0]P(T_i = 0|R_i = 1) + E[\mu(D_i)T_i|R_i = 1, T_i = 1]P(T_i = 1|R_i = 1) - 0 \\
&= \mu(D_i|D_i \geq d_r)P(D_i \geq d_c|D_i \geq d_r) \\
&= \mu(D_i|D_i \geq d_r) \cdot \frac{1 - F_D(d_c)}{1 - F_D(d_r)}
\end{aligned}$$

If instead $d_r \geq d_c$

$$\begin{aligned}
\widehat{b}_1^{\text{BIN}} &\to^p E[\mu(D_i)T_i|R_i = 1] - E[\mu(D_i)T_i|R_i = 0] \\
&= E[\mu(D_i)T_i|R_i = 1] - \\
&\quad E[\mu(D_i)T_i|R_i = 0, T_i = 1]P(T_i = 1|R_i = 0) - E[\mu(D_i)T_i|R_i = 0, T_i = 0]P(T_i = 0|R_i = 0) \\
&= \mu(D_i|D_i \geq d_r) - \mu(D_i|d_c < D_i < d_r) \cdot P(D_i < d_c|D_i < d_r) \\
&= \mu(D_i|D_i \geq d_r) - \mu(D_i|d_c < D_i < d_r) \cdot \frac{F_D(d_r) - F_D(d_c)}{F_D(d_r)}
\end{aligned}$$

[T1]

Consider a researcher guess d_g such that $d_g \geq d_r \geq d_c$. $\hat{b}_0^{MSE}, \hat{b}_1^{MSE}$ are estimates from the bivariate regression

$$\triangle Y_i = \beta_0 + \beta_1 \mathbb{1}(d_i \geq d_r)$$

Where we have dropped all observations where $d_i \in (d_g, d_r)$. Standard results give us that

$$\begin{aligned}
\hat{b}_0^{MSE} &= \mathbb{E}[\triangle Y_i|d_i \leq d_r] + o_p(1) = \mathbb{E}[\triangle Y_i|d_i \leq d_c]\frac{F(d_c)}{F(d_r)} + \mathbb{E}[\triangle Y_i|d_c \leq d_i \leq d_r]\frac{F(d_r) - F(d_c)}{F(d_r)} + o_p(1) \\
&= \triangle\theta_t + \mathbb{E}[\mu(d_i)|d_c \leq d_i \leq d_r]\frac{F(d_r) - F(d_c)}{F(d_r)} + o_p(1) \\
\hat{b}_1^{MSE} &= \mathbb{E}[\triangle Y_i|d_i \geq d_r] - \mathbb{E}[\triangle Y_i|d_i \leq d_r] + o_p(1)
\end{aligned}$$

$$\text{Bias}(\hat{b}_1^{MSE}, \beta_r) = -\frac{F(d_g) - F(d_c)}{F(d_g)} \int_{d_c}^{d_g} \triangle Y_{it}(k, 1)\frac{f(k)}{F(d_g) - F(d_c)}\text{d}k$$

$$\begin{aligned}
\text{Var}(\hat{b}_1^{MSE}) &= V(\sigma^2(d_c, d_g, d_r), d_g, d_r) \\
V(\sigma^2(d_c, d_g, d_r), d_g, d_r) &= \frac{1}{n}\frac{\sigma^2(d_c, d_g, d_r) + \sigma^2_{\triangle\varepsilon}}{\frac{(1-F(d_r))F(d_r)}{[1-F(d_r)+F(d_g)]^2}}
\end{aligned}$$

$$\sigma^2(d_c, d_g, d_r) = \int_{d_l}^{d_u} [\triangle\theta_t + \mu(k)\mathbb{1}(k \geq d_c) - \hat{b}_0^{MSE} - \hat{b}_1^{MSE}\mathbb{1}(k \geq d_r)]^2 \mathbb{1}(k \notin (d_g, d_r)) f(k) \mathrm{d}k$$

Proof.

$$\frac{\mathrm{d}V(\sigma^2(d_c, d_g, d_r), d_g, d_r)}{\mathrm{d}d_c} = V_1\sigma_1^2$$

It is easy to show that $V_1 > 0$, so it must be the case that sign $< \frac{\mathrm{d}V(\sigma^2(d_c,d_g,d_r),d_g,d_r)}{\mathrm{d}d_c} >=$ sign $< \sigma_1 >$ We split $\sigma^2(\cdot)$ into three components:

$$\begin{aligned}\sigma^2(d_c, d_g, d_r) = &\int_{d_l}^{d_c} [\triangle\theta_t + \mu(k)\mathbb{1}(k \geq d_c) - \hat{b}_0^{MSE} - \hat{b}_1^{MSE}\mathbb{1}(k \geq d_r)]^2 \mathbb{1}(k \notin (d_g, d_r)) f(k) \mathrm{d}k \\ &+ \int_{d_c}^{d_g} [\triangle\theta_t + \mu(k)\mathbb{1}(k \geq d_c) - \hat{b}_0^{MSE} - \hat{b}_1^{MSE}\mathbb{1}(k \geq d_r)]^2 \mathbb{1}(k \notin (d_g, d_r)) f(k) \mathrm{d}k \\ &+ \int_{d_r}^{d_u} [\triangle\theta_t + \mu(k)\mathbb{1}(k \geq d_c) - \hat{b}_0^{MSE} - \hat{b}_1^{MSE}\mathbb{1}(k \geq d_r)]^2 \mathbb{1}(k \notin (d_g, d_r)) f(k) \mathrm{d}k\end{aligned}$$

Simplifying each component,

$$\begin{aligned}\sigma^2(d_c, d_g, d_r) = &\int_{d_l}^{d_c} [-\mathbb{E}[\mu(d_i)|d_c \leq d_i \leq d_r]\frac{F(d_r) - F(d_c)}{F(d_r)}]^2 f(k)\mathrm{d}k && g_1 \\ &+ \int_{d_c}^{d_g} [\mu(k) - \mathbb{E}[\mu(d_i)|d_c \leq d_i \leq d_r]\frac{F(d_r) - F(d_c)}{F(d_r)}]^2 f(k)\mathrm{d}k && g_2 \\ &+ \int_{d_r}^{d_u} [\triangle\theta_t + \mu(k) - \hat{b}_0^{MSE} - \hat{b}_1^{MSE}]^2 f(k)\mathrm{d}k && g_3\end{aligned}$$

Now we take derivatives. Since (3) does not depend on d_c, it is clear that $\mathrm{d}g_1/\mathrm{d}d_c = 0$.

$$\frac{\mathrm{d}g_2}{\mathrm{d}d_c} =$$