Identification of Treatment Effects with Bunching Designs

Carolina Caetano

University of Georgia

Bunching Designs

- Recent line of identification strategies that leverage bunching phenomenon.
- Original approaches successful in Public Finance following Saez (2010), focusing on identification of the taxable income's elasticity to changes in the marginal tax rate.
- Large applied literature leveraging **bunching on outcome** variable in structural models. Few methodological studies: Blomquist and Newey (2017), Bertanha, McCallum and Seegert (2022) and Goff (2022).
- Line started with Caetano (2015) leverages **bunching on treatment** variable to build tests of identification. More recent advancements following Caetano, Caetano and Nielsen (2023) focus on identification in models with endogeneity.

This Talk

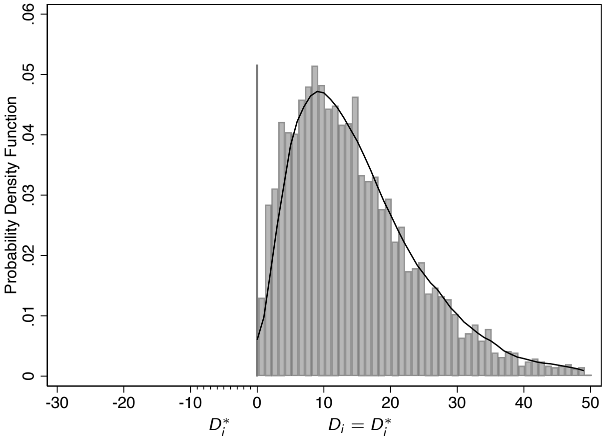
We will discuss methods that leverage corner bunching on the treatment variable when treatment is endogenous.

Methods use bunching and a menu of assumptions to achieve identification.

- Do not require instrumental variables or panel data. Can be used when these are not available, or when question cannot be answered with these.
- Allow the study of heterogeneity along any dimension for which there are bunched and non-bunched observations.
- Not vulnerable to weak identification, not data hungry and very stable.
- Can be combined with panel data.

Bunching

Bunching: Concentration of observations at a point of an otherwise locally continuous distribution.



Hours per week children watch TV on CDS-PSID

Histogram cells: 1 hour, Bandwidth: 4 hours

Some treatment variables with corner bunching

- **Consumption goods:** number of tobacco products, alcoholic beverages, caffeinated drinks, sugary drinks, fast food meals, dining out meals, subscription services, supplements and vitamins, public transportation rides, books read, gym visits, doctor visits, trips, fuel usage amounts, expenditure on health, fitness, travel, vacations, education, childcare.
- **Financial variables:** credit access, bequests, savings, emergency fund levels, retirement account contributions, mortgage balance, credit card debt, student loan debt, income from investments, expenditure on ads, charitable donations, HSA and FSA balances, life insurance coverage, number of trades.
- **Time use:** few exceptions, like sleeping, eating and using the bathroom. Some examples: exercising, working, watching TV, using digital devices, doing homework, doing chores, volunteering, commuting.
- **Neighborhood characteristics:** number of public transportation options or stops, retail stores, coffee shops, rental units, affordable housing units, vacant units, electric vehicle charging stations; length of biking lanes or walking paths; areas of green space, comercial districts, sport fields, parking lots.
- **Variables with regulatory minimums:** schooling time, wages, 401K contributions, coverage for auto insurance, nutritional standards for school meals, bank capital, bank deposit insurance, age started working, age started withdrawing from retirement accounts, age retired.
- **Variables with regulatory maximums:** 401K, Roth IRA, HSA, FSA contributions, untaxed gifts, FHA loans, FDIC insurance, carbon emissions, liquor licenses, lot coverage, contributions to political campaigns, data usage, grades, absences from school, class size, commissions on sales.

Fundamental Problem of Causality

- D_i: treatment variable
- $Y_i(d)$: potential outcome for treatment level $D_i = d$

Then

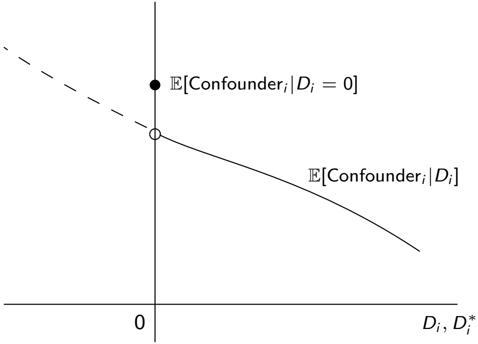
$$\mathbb{E}[Y_i|D_i] - \mathbb{E}[Y_i|D_i = 0] = \underbrace{\mathbb{E}[Y_i(D_i) - Y_i(0)|D_i]}_{\text{ATT}(D_i)} + \underbrace{\mathbb{E}[Y_i(0)|D_i] - \mathbb{E}[Y_i(0)|D_i = 0]}_{\text{Selection Bias}}$$

- $\text{ATT}(D_i)$: Average Treatment Effect of D_i units on those treated D_i units

Nature of Selection under Constraints

We conceptually separate the treatment variable and the selection variable:

- D_i: treatment variable
- D_i^*: selection variable



At the bunching point, treatment does not vary, but selection does:

$$D_i^* = D_i + D_i^* 1(D_i^* \leq 0), \text{ assume } \mathbb{P}(D_i^* < 0) > 0$$

Linear Model with Endogeneity and No Controls

$$Y_i(d) = \beta d + U_i,$$

where

$$\mathbb{E}[U_i | D_i^*] = \alpha + \delta D_i^*.$$

Then, we can write

$$Y_i = \alpha + \beta D_i + \delta D_i^* + \varepsilon_i, \text{ where } \mathbb{E}[\varepsilon_i | D_i^*] = 0.$$

Linear Model with Endogeneity and No Controls

$$Y_i(d) = \beta d + U_i,$$

where

$$\mathbb{E}[U_i | D_i^*] = \alpha + \delta D_i^*.$$

Then, we can write

$$Y_i = \alpha + \beta D_i + \delta D_i^* + \varepsilon_i, \text{ where } \mathbb{E}[\varepsilon_i | D_i^*] = 0.$$

The separation between the selection variable and the treatment at the bunching point causes a discontinuity in the expected outcome at the bunching point:

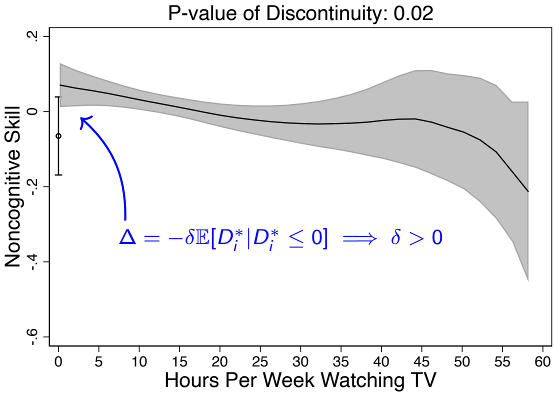
$$\mathbb{E}[Y_i | D_i] = \alpha + (\beta + \delta) D_i + \delta \mathbb{E}[D_i^* | D_i^* \leq 0] 1(D_i = 0)$$

Sign of Selection Bias is Identified

$$\mathbb{E}[Y_i|D_i] = \alpha + (\beta+\delta)D_i + \delta\mathbb{E}[D_i^*|D_i^* \leq 0]1(D_i = 0)$$

So,

$$\Delta := \underbrace{\lim_{d\downarrow 0}\mathbb{E}[Y_i|D_i = d] - \mathbb{E}[Y_i|D_i = 0]}_{\text{Outcome Discontinuity at } D_i=0} = -\delta\mathbb{E}[D_i^*|D_i^* \leq 0]$$



Simple Bound for the Treatment Effect

$$\mathbb{E}[Y_i|D_i] = \alpha + (\beta+\delta)D_i + \delta\mathbb{E}[D_i^*|D_i^* \leq 0]1(D_i = 0)$$

$$\Delta = -\delta\mathbb{E}[D_i^*|D_i^* \leq 0]$$

1. $(\beta+\delta)$ and Δ obtained from the regression of Y_i onto D_i and $1(D_i = 0)$, since

$$\mathbb{E}[Y_i|D_i] = \alpha + (\beta+\delta)D_i - \Delta 1(D_i = 0).$$

Simple Bound for the Treatment Effect

$$\mathbb{E}[Y_i|D_i] = \alpha + (\beta+\delta)D_i + \delta\mathbb{E}[D_i^*|D_i^* \leq 0]1(D_i = 0)$$

$$\Delta = -\delta\mathbb{E}[D_i^*|D_i^* \leq 0]$$

1. $(\beta+\delta)$ and Δ obtained from the regression of Y_i onto D_i and $1(D_i = 0)$, since
$$\mathbb{E}[Y_i|D_i] = \alpha + (\beta+\delta)D_i - \Delta 1(D_i = 0).$$

2. Since $\mathbb{E}[D_i^*|D_i^* \leq 0] < 0$, δ has the same sign as Δ, so
   - If $\Delta > 0$, then $\delta > 0$, and $\beta = (\beta+\delta) - \delta < (\beta+\delta) = -0.38$
   - If $\Delta < 0$, then $\delta < 0$, and $\beta = (\beta+\delta) - \delta > (\beta+\delta)$
   - If $\Delta = 0$, then $\delta = 0$, and $\beta = (\beta+\delta)$

Simple Bound for the Treatment Effect

$$\mathbb{E}[Y_i|D_i] = \alpha + (\beta+\delta)D_i + \delta\mathbb{E}[D_i^*|D_i^* \le 0]1(D_i = 0)$$

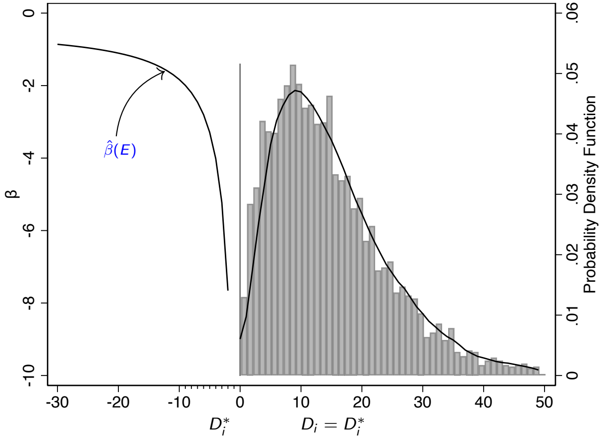
$$\Delta = -\delta\mathbb{E}[D_i^*|D_i^* \le 0]$$

1. $(\beta+\delta)$ and Δ obtained from the regression of Y_i onto D_i and $1(D_i = 0)$, since
$$\mathbb{E}[Y_i|D_i] = \alpha + (\beta+\delta)D_i - \Delta 1(D_i = 0).$$
2. Since $\mathbb{E}[D_i^*|D_i^* \le 0] < 0$, δ has the same sign as Δ, so
   - If $\Delta > 0$, then $\delta > 0$, and $\beta = (\beta+\delta) - \delta < (\beta+\delta) = -0.38$
   - If $\Delta < 0$, then $\delta < 0$, and $\beta = (\beta+\delta) - \delta > (\beta+\delta)$
   - If $\Delta = 0$, then $\delta = 0$, and $\beta = (\beta+\delta)$
3. For each value E of $\mathbb{E}[D_i^*|D_i^* \le 0]$, the function of possible treatment effects is
$$\beta(E) = (\beta+\delta) + \frac{\Delta}{E}$$

Path of Treatment Effects for Values of $\mathbb{E}[D_i^*|D_i^* \leq 0]$

$$\beta(E) = (\beta + \delta) + \frac{\Delta}{E}$$

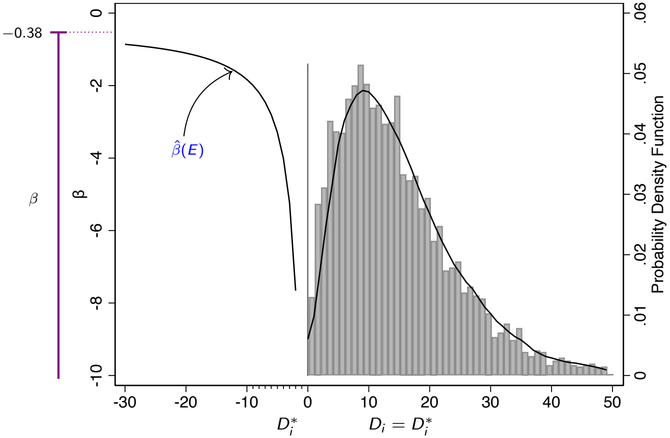
Figure: Non-cognitive Skills



Path of Treatment Effects for Values of $\mathbb{E}[D_i^*|D_i^* \leq 0]$

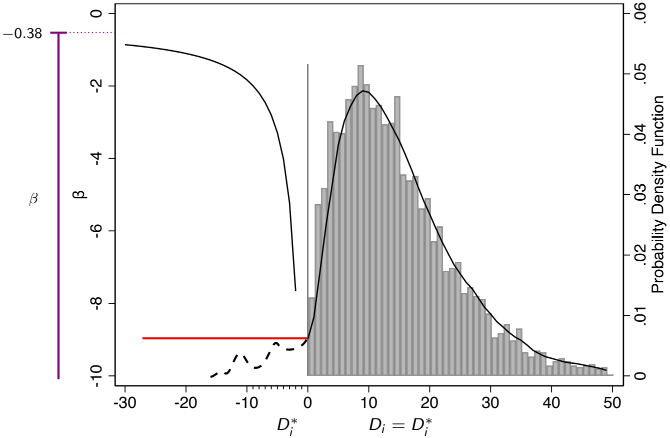
$$\beta(E) = (\beta + \delta) + \frac{\Delta}{E}$$

Figure: Non-cognitive Skills



Opposite Bound: No High Peaks

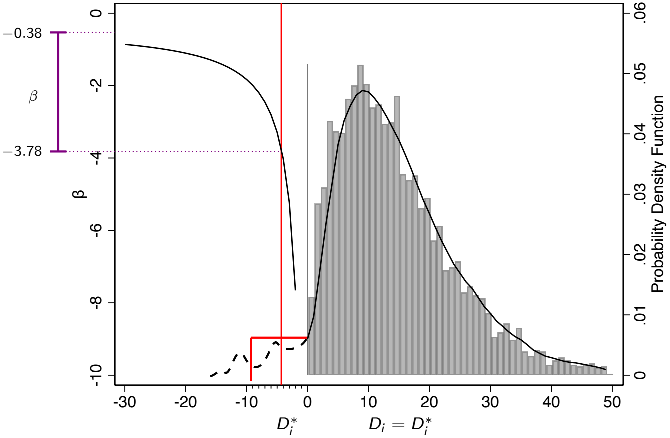
If $f_{D^*}(d) \leq f_D(0^+) := \lim_{d \downarrow 0} f_D(d)$ for all $d \leq 0$,



Opposite Bound: No High Peaks

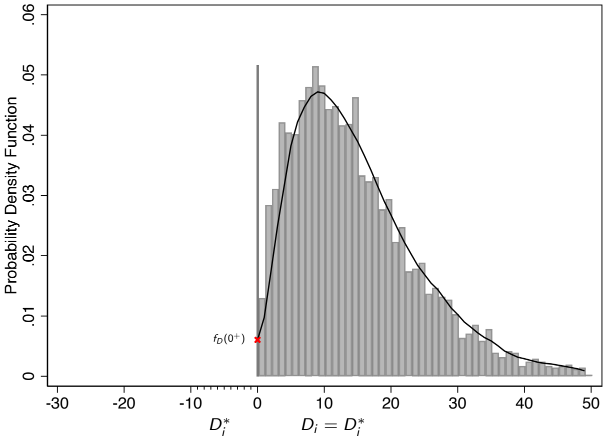
If $f_{D^*}(d) \leq f_D(0^+) := \lim_{d \downarrow 0} f_D(d)$ for all $d \leq 0$,

then $\mathbb{E}[D_i^* | D_i^* \leq 0] \leq -\frac{1}{2} \frac{\mathbb{P}(D_i = 0)}{f_D(0^+)}$ ← Expectation of Uniform Distribution



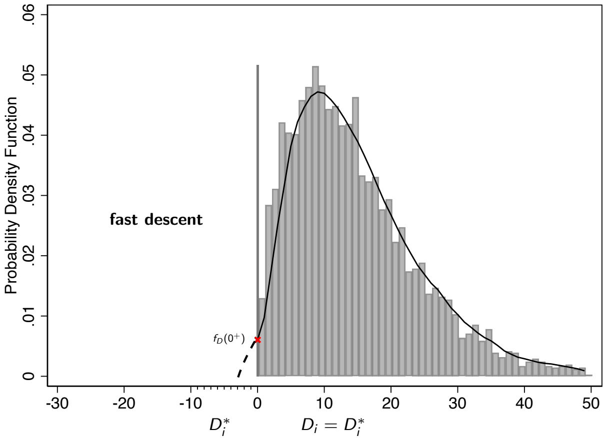
Slow and Fast Descent Densities

Figure: TV hours per week



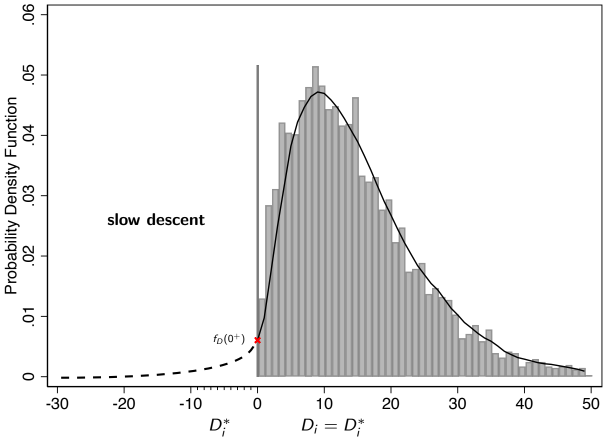
Slow and Fast Descent Densities

Figure: TV hours per week

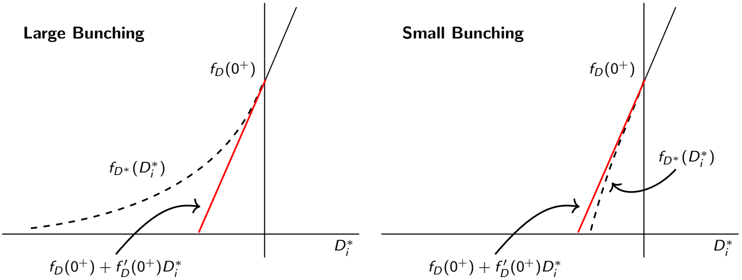


Slow and Fast Descent Densities

Figure: TV hours per week

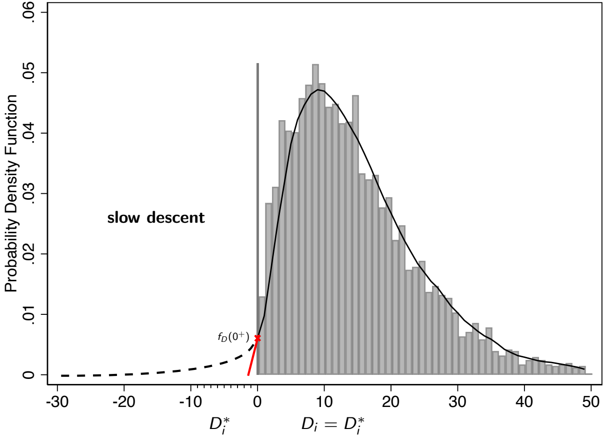


Amount of Bunching and Distribution Categories



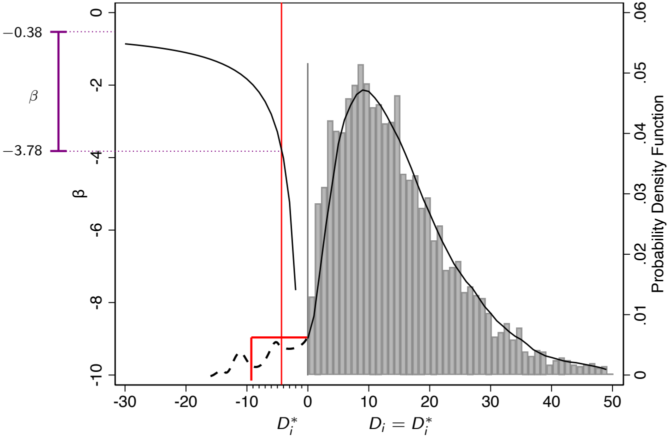
Large Bunching: $\mathbb{P}(D_i = 0) > \frac{1}{2}\frac{f_D(0^+)^2}{f_D'(0^+)}$. If "no high peaks" holds, then Slow Descent

Our Application: Slow Descent



Better Opposite Bound: Large Bunching Convex Class

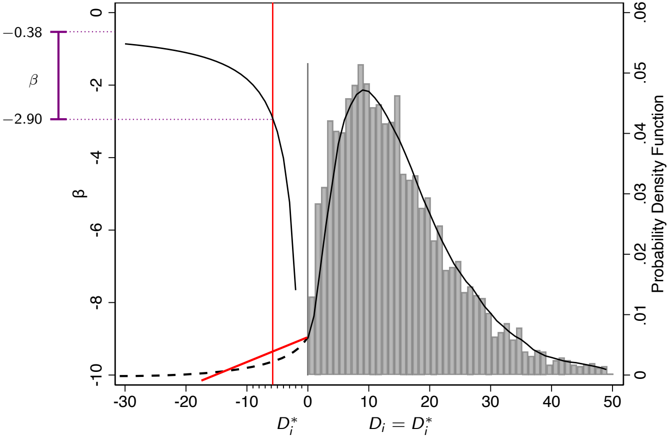
$f_{D^*}(d)$ no high peaks: $\mathbb{E}[D^*|D^* \leq 0] \leq -\frac{1}{2}\frac{\mathbb{P}(D_i = 0)}{f_D(0^+)}$ ← Uniform Distribution



Better Opposite Bound: Large Bunching Convex Class

$f_{D^*}(d)$ no high peaks: $\mathbb{E}[D^*|D^* \leq 0] \leq -\frac{1}{2}\frac{\mathbb{P}(D_i = 0)}{f_D(0^+)}$ ← Uniform Distribution

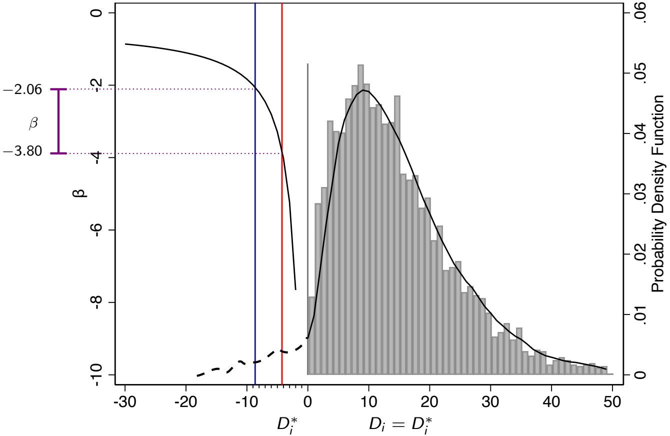
$f_{D^*}(d)$ convex: $\mathbb{E}[D^*|D^* \leq 0] \leq -\frac{2}{3}\frac{\mathbb{P}(D_i = 0)}{f_D(0^+)}$ ← Triangular Distribution



Non-Nested Bounds: Bi-Log Concave Class

$f_{D^*}(d)$ is Bi-log concave ($\implies$ $f_{D^*}(d)$ has locally bounded derivatives).

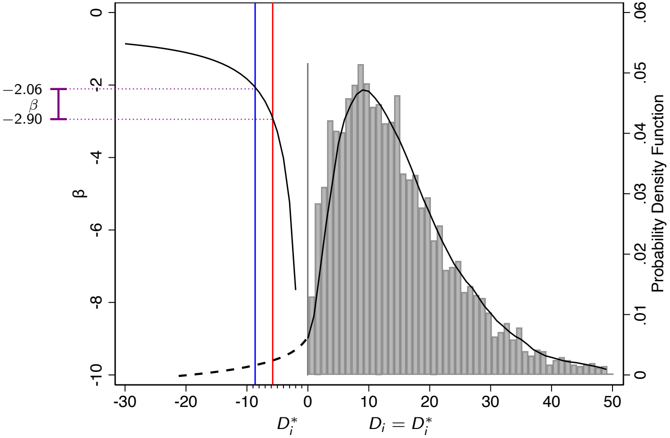
$$-\frac{\mathbb{P}(D_i = 0)}{f_D(0^+)} \leq \mathbb{E}[D_i^* | D_i^* \leq 0] \leq \frac{\mathbb{P}(D_i > 0)}{f_D(0^+)} \left(1 + \frac{\log(\mathbb{P}(D_i > 0))}{\mathbb{P}(D_i = 0)}\right).$$



Combining Non-Nested Classes

$f_{D^*}(d)$ is convex and Bi-log concave (convex of bounded variation)

$$-\frac{\mathbb{P}(D_i = 0)}{f_D(0^+)} \leq \mathbb{E}[D_i^* | D_i^* \leq 0] \leq -\frac{2}{3}\frac{\mathbb{P}(D_i = 0)}{f_D(0^+)}$$



Identification of the Value of β

$$\mathbb{E}[Y_i|D_i] = \alpha + (\beta + \delta)D_i + \delta\mathbb{E}[D_i^*|D_i^* \leq 0]1(D_i = 0)$$

For $D_i > 0$,

$$\mathbb{E}[Y_i|D_i = d + 1] - \mathbb{E}[Y_i|D_i = d] = \beta + \delta$$

Identification of the Value of β

$$\mathbb{E}[Y_i|D_i] = \alpha + (\beta+\delta)D_i + \delta\mathbb{E}[D_i^*|D_i^* \leq 0]1(D_i = 0)$$

For $D_i > 0$,

$$\mathbb{E}[Y_i|D_i = d+1] - \mathbb{E}[Y_i|D_i = d] = \beta + \delta$$

The Selection Bias can be written as

$$\delta = \frac{\Delta}{\mathbb{E}[D_i^*|D_i^* \leq 0]} = \frac{\lim_{d\downarrow 0}\mathbb{E}[Y_i|D_i = d] - \mathbb{E}[Y_i|D_i = 0]}{\lim_{d\downarrow 0}\mathbb{E}[D_i^*|D_i = d] - \mathbb{E}[D_i^*|D_i = 0]}$$

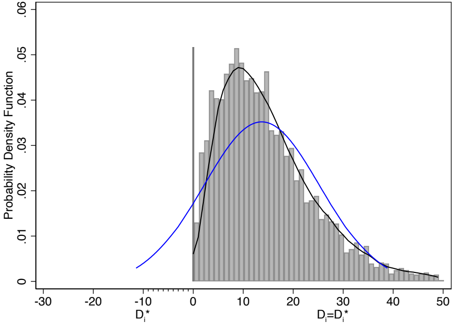
- IV parallel: we want to identify the "effect" of the selection variable D_i^* on Y_i, and D_i is a local IV around the bunching point.
- We are missing one component, as we do not observe the average selection among the bunched observations.

Identification of $\mathbb{E}[D^*|D^* \leq 0]$: Gaussian Family

$$D_i^* \sim \mathcal{N}(\mu, \sigma^2) \implies \mathbb{E}[D_i^*|D_i^* \leq 0] = \mu - \sigma\lambda(\mu/\sigma),$$

where $\lambda(x) = \phi(x)/\Phi(x)$ is the inverse Mills Ratio.

- Estimate μ and σ by running a Tobit regression of D_i onto a constant. The constant is $\hat{\mu}$, and the variance is $\hat{\sigma}^2$. Our estimate: -3.07.



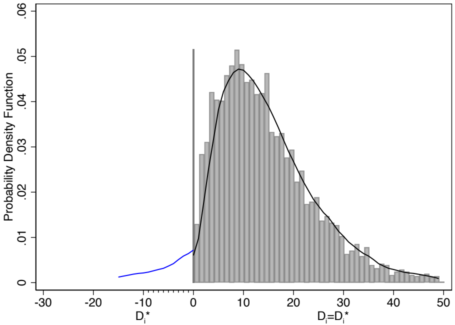
- Parametric distributions can be tested (e.g. Kolmogorov-Smirnoff test and Goldman and Kaplan (2018)).

Identification of $\mathbb{E}[D^*|D^* \leq 0]$: Tail Symmetric Family

$D_i^* \sim$ Symmetric below 0 and above quantile $1 - \mathbb{P}(D_i = 0)$ (call it q_0).

$$\mathbb{E}[D_i^*|D_i^* \leq 0] = q_0 - \mathbb{E}[D_i|D_i \geq q_0]$$

- Estimate q_0 as the $(1 - \mathbb{P}(D_i = 0))$-th quantile of D_i and $\mathbb{E}[D_i|D_i \geq q_0]$ as the average of the D_i above q_0. Our estimate: -2.11.



- Tail symmetry not testable, but full symmetry is testable between 0 and q_0.

A Taste of Nonparametric Identification of $\mathbb{E}[D^*|D^* \leq 0]$

We will set up the problem of identification of $\mathbb{E}[D_i^*|D_i^* \leq 0]$ as a problem of identification in a nonparametric censoring model. Let X_i be a vector of covariates.

Since $D_i^* = D_i + D_i^* 1(D_i = 0)$, $\mathbb{E}[D_i^*] = \mathbb{E}[D_i] + \mathbb{E}[D_i^*|D_i^* \leq 0]\mathbb{P}(D_i = 0)$. So,

$$\mathbb{E}[D_i^*|D_i^* \leq 0] = \frac{\mathbb{E}[D_i^*] - \mathbb{E}[D_i]}{\mathbb{P}(D_i = 0)} = \frac{\mathbb{E}[\mathbb{E}[D_i^*|X_i]] - \mathbb{E}[D_i]}{\mathbb{P}(D_i = 0)}.$$

All pieces of this equation are identified, except for $\mathbb{E}[D_i^*|X_i]$.

A Taste of Nonparametric Identification of $\mathbb{E}[D^*|D^* \leq 0]$

We will set up the problem of identification of $\mathbb{E}[D_i^*|D_i^* \leq 0]$ as a problem of identification in a nonparametric censoring model. Let X_i be a vector of covariates.

Since $D_i^* = D_i + D_i^* 1(D_i = 0)$, $\mathbb{E}[D_i^*] = \mathbb{E}[D_i] + \mathbb{E}[D_i^*|D_i^* \leq 0]\mathbb{P}(D_i = 0)$. So,

$$\mathbb{E}[D_i^*|D_i^* \leq 0] = \frac{\mathbb{E}[D_i^*] - \mathbb{E}[D_i]}{\mathbb{P}(D_i = 0)} = \frac{\mathbb{E}[\mathbb{E}[D_i^*|X_i]] - \mathbb{E}[D_i]}{\mathbb{P}(D_i = 0)}.$$

All pieces of this equation are identified, except for $\mathbb{E}[D_i^*|X_i]$.

Define $m(X_i) = \mathbb{E}[D_i^*|X_i]$. Then,

$$\begin{aligned} D_i^* &= m(X_i) + \eta_i, \text{ where } \mathbb{E}[\eta_i|X_i] = 0, \\ D_i &= \max\{D_i^*, 0\}, \end{aligned}$$

which is a standard nonparametric censoring model.

We can identify $m(X_i)$ using any of the techniques in the censoring literature, e.g. Powell (1984), Powell (1986); Horowitz (1986); Chen and Khan (2000); Newey (2001); Lewbel and Linton (2002); Chen, Dahl and Khan (2005).

Implementation Through Control Function

The method may be implemented by estimating all the pieces and plugging them into the identification formulas.

A more convenient approach is to consider the equation

$$\mathbb{E}[Y_i|D_i] = \alpha + (\beta + \delta)D_i + \delta\mathbb{E}[D_i^*|D_i^* \leq 0]1(D_i = 0),$$

and reorganize it

$$\mathbb{E}[Y_i|D_i] = \alpha + \beta D_i + \delta \underbrace{[D_i + \mathbb{E}[D_i^*|D_i^* \leq 0]1(D_i = 0)]}_{\text{Control Function}}.$$

Implementation Through Control Function

The method may be implemented by estimating all the pieces and plugging them into the identification formulas.

A more convenient approach is to consider the equation

$$\mathbb{E}[Y_i|D_i] = \alpha + (\beta + \delta)D_i + \delta\mathbb{E}[D_i^*|D_i^* \leq 0]1(D_i = 0),$$

and reorganize it

$$\mathbb{E}[Y_i|D_i] = \alpha + \beta D_i + \delta \underbrace{[D_i + \mathbb{E}[D_i^*|D_i^* \leq 0]1(D_i = 0)]}_{\text{Control Function}}.$$

For any consistent estimator $\hat{E}[D_i^*|D_i^* \leq 0]$, build the control function

$$D_i + \hat{\mathbb{E}}[D_i^*|D_i^* \leq 0]1(D_i = 0)$$

and add it to the regression to control for the endogeneity of D_i.

Implementation Through Control Function

The method may be implemented by estimating all the pieces and plugging them into the identification formulas.

A more convenient approach is to consider the equation

$$\mathbb{E}[Y_i|D_i] = \alpha + (\beta + \delta)D_i + \delta\mathbb{E}[D_i^*|D_i^* \leq 0]1(D_i = 0),$$

and reorganize it

$$\mathbb{E}[Y_i|D_i] = \alpha + \beta D_i + \delta \underbrace{[D_i + \mathbb{E}[D_i^*|D_i^* \leq 0]1(D_i = 0)]}_{\text{Control Function}}.$$

For any consistent estimator $\hat{E}[D_i^*|D_i^* \leq 0]$, build the control function

$$D_i + \hat{\mathbb{E}}[D_i^*|D_i^* \leq 0]1(D_i = 0)$$

and add it to the regression to control for the endogeneity of D_i.

Same approach can be used for all bounds of β. Simply substitute the expectation bound estimator in place of $\hat{\mathbb{E}}[D_i^*|D_i^* \leq 0]$ in the control function formula and run the same regression.

Implementation Through Control Function

The method may be implemented by estimating all the pieces and plugging them into the identification formulas.

A more convenient approach is to consider the equation

$$\mathbb{E}[Y_i|D_i] = \alpha + (\beta + \delta)D_i + \delta\mathbb{E}[D_i^*|D_i^* \leq 0]1(D_i = 0),$$

and reorganize it

$$\mathbb{E}[Y_i|D_i] = \alpha + \beta D_i + \delta \underbrace{[D_i + \mathbb{E}[D_i^*|D_i^* \leq 0]1(D_i = 0)]}_{\text{Control Function}}.$$

For any consistent estimator $\hat{E}[D_i^*|D_i^* \leq 0]$, build the control function

$$D_i + \hat{\mathbb{E}}[D_i^*|D_i^* \leq 0]1(D_i = 0)$$

and add it to the regression to control for the endogeneity of D_i.

Same approach can be used for all bounds of β. Simply substitute the expectation bound estimator in place of $\hat{\mathbb{E}}[D_i^*|D_i^* \leq 0]$ in the control function formula and run the same regression.

The standard errors need to be corrected for the fact that we use $\hat{E}[D_i^*|D_i^* \leq 0]$ instead of $E[D_i^*|D_i^* \leq 0]$, but we showed that the bootstrap works.

Including Controls in the Model

Suppose that X_i is a vector of controls,

$$Y_i(d) = \beta(X_i)d + U_i, \text{ where } \mathbb{E}[U_i|D_i^*, X_i] = \alpha(X_i) + \delta(X_i)D_i^*.$$

Including Controls in the Model

Suppose that X_i is a vector of controls,

$$Y_i(d) = \beta(X_i)d + U_i, \text{ where } \mathbb{E}[U_i|D_i^*, X_i] = \alpha(X_i) + \delta(X_i)D_i^*.$$

Then, we can write

$$Y_i = \alpha(X_i) + \beta(X_i)D_i + \delta(X_i)D_i^* + \varepsilon_i, \text{ where } \mathbb{E}[\varepsilon_i|D_i^*, X_i] = 0.$$

Therefore,

$$\mathbb{E}[Y_i|D_i, X_i] = \alpha(X_i) + \beta(X_i)D_i + \delta(X_i)\underbrace{[D_i + \mathbb{E}[D_i^*|D_i^* \leq 0, X_i]1(D_i = 0)]}_{\text{Control Function}}.$$

Including Controls in the Model

Suppose that X_i is a vector of controls,

$$Y_i(d) = \beta(X_i)d + U_i, \text{ where } \mathbb{E}[U_i|D_i^*, X_i] = \alpha(X_i) + \delta(X_i)D_i^*.$$

Then, we can write

$$Y_i = \alpha(X_i) + \beta(X_i)D_i + \delta(X_i)D_i^* + \varepsilon_i, \text{ where } \mathbb{E}[\varepsilon_i|D_i^*, X_i] = 0.$$

Therefore,

$$\mathbb{E}[Y_i|D_i, X_i] = \alpha(X_i) + \beta(X_i)D_i + \delta(X_i)\underbrace{[D_i + \mathbb{E}[D_i^*|D_i^* \leq 0, X_i]1(D_i = 0)]}_{\text{Control Function}}.$$

Say $\beta(X_i) = \beta_0 + \beta_1 X_{i1}$, $\alpha(X_i) = \alpha_0 + X_i'\alpha_1$, and $\delta(X_i) = \delta$. Then we estimate

$$\mathbb{E}[Y_i|D_i, X_i = x] = \alpha_0 + \beta_0 D_i + \beta_1 X_{1i} D_i + X_i'\alpha_1 + \delta[D_i + \mathbb{E}[D_i^*|D_i^* \leq 0, X_i = x]1(D_i = 0)],$$

which allows the study of heterogeneous treatment effects in X_{1i}, and the linear inclusion of covariates, including fixed effects.

Discrete Controls

If $X_i \in \{x_1, \dots, x_K\}$. For a specific value of the controls, say x_k, the model reverts to the previous model:

$$\mathbb{E}[Y_i|D_i, X_i = x_k] = \alpha_k + \beta_k D_i + \delta_k[D_i + \mathbb{E}[D_i^*|D_i^* \leq 0, X_i = x_k]1(D_i = 0)].$$

$\Longrightarrow$ if the controls are discrete, all the procedures can be done within each subgroup separately, as if there were no controls.

Discrete Controls

If $X_i \in \{x_1, \dots, x_K\}$. For a specific value of the controls, say x_k, the model reverts to the previous model:

$$\mathbb{E}[Y_i|D_i, X_i = x_k] = \alpha_k + \beta_k D_i + \delta_k[D_i + \mathbb{E}[D_i^*|D_i^* \leq 0, X_i = x_k]1(D_i = 0)].$$

$\Longrightarrow$ if the controls are discrete, all the procedures can be done within each subgroup separately, as if there were no controls.

Alternatively, let

- $C_i = (1(X_i = x_1), \dots, 1(X_i = x_K))'$
- $\alpha = (\alpha(x_1), \dots, \alpha(x_K))'$, $\beta = (\beta(x_1), \dots, \beta(x_K))'$, $\delta = (\delta(x_1), \dots, \delta(x_K))'$
- $E_i = (\mathbb{E}[D_i^*|D_i^* \leq 0, X_i = x_1], \dots, \mathbb{E}[D_i^*|D_i^* \leq 0, X_i = x_K])' C_i$

Then the following model is equivalent:

$$\mathbb{E}[Y_i|D_i, C_i] = C_i'\alpha + D_i C_i'\beta + (D_i + E_i 1(D_i = 0))C_i'\delta,$$

which can be estimated with a single regression.

Control Discretization Through Clustering

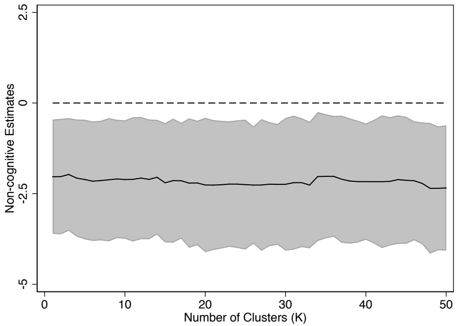
Clustering methods (e.g. hierarchical clustering) distribute the X_i into K sets: $\mathcal{C}_1, \dots, \mathcal{C}_K$, where the X_i are as similar as possible.

Define $C_i = (1(X_i \in \mathcal{C}_1), \dots, 1(X_i \in \mathcal{C}_K))'$ and calculate

$$\hat{\mathbb{E}}[D^*|D^* \leq 0, X_i] = \hat{\mathbb{E}}[D^*|D^* \leq 0, C_i]$$

and proceed as in the discrete control case.

Figure: Non-Cognitive Estimates per Cluster Number



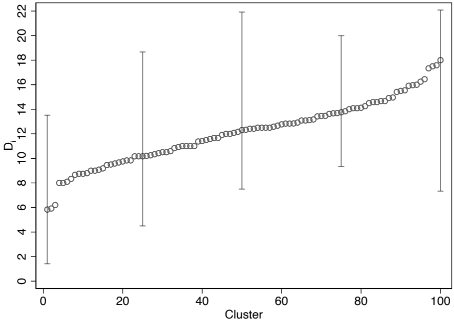
Relaxing Linearity Using Controls

Model with controls:

$$Y_i = \alpha(X_i) + \beta(X_i)D_i + \delta(X_i)D_i^* + \varepsilon_i, \text{ where } \mathbb{E}[\varepsilon_i | D_i^*, X_i] = 0$$

- Allows treatment effects and selection effects to change with X_i, including to change the sign.
- Because D_i and X_i are correlated, the use of covariates weakens the linearity assumptions indirectly.

Figure: Average TV Hours per Cluster



Relaxing the Constant Selection Effects Assumption

Constant selection effects are required only locally around the bunching point:

$$\mathbb{E}[U_i|D_i^*] = \alpha + \delta D_i^*, \text{ for } D_i^* \leq \zeta,$$

for some $\zeta > 0$.

In this case, the control function model holds locally around the bunching point:

$$\mathbb{E}[Y_i|D_i] = \alpha + \beta D_i + \delta[D_i + \mathbb{E}[D_i^*|D_i^* \leq 0]1(D_i = 0)], \text{ for } D_i^* \leq \zeta.$$

Then,

$$\beta = \lim_{d \downarrow 0} \frac{d}{dD_i}\mathbb{E}[Y_i|D_i]\Big|_{D_i=d} - \frac{\Delta}{\mathbb{E}[D_i^*|D_i^* \leq 0]}.$$

Estimation may be done in two steps:

1. Estimate $\hat{\delta} = \hat{\Delta}/\hat{E}[D_i^*|D_i^* \leq 0]$.
2. Estimate a local linear regression of $Y_i - \hat{\delta}D_i$ onto D_i at $D_i = 0$ using only observations such that $D_i > 0$. The coeffcient of the constant is $\hat{\alpha}$, and the slope coefficient is $\hat{\beta}$.

Relaxing the Constant Treatment Effects Assumption

In the general setting, we can write

$$\mathbb{E}[Y_i | D_i] = \underbrace{\mathbb{E}[Y_i(D_i) - Y_i(0) | D_i]}_{\text{ATT}_i(D_i)} + \mathbb{E}[Y_i(0) | D_i^*],$$

so without loss of generality we can write

$$Y_i(d) = \text{ATT}(D_i) + U_i, \text{ where } \mathbb{E}[U_i | D_i^*] = \mathbb{E}[Y_i(0) | D_i^*].$$

Relaxing the Constant Treatment Effects Assumption

In the general setting, we can write

$$\mathbb{E}[Y_i|D_i] = \underbrace{\mathbb{E}[Y_i(D_i) - Y_i(0)|D_i]}_{\text{ATT}_i(D_i)} + \mathbb{E}[Y_i(0)|D_i^*],$$

so without loss of generality we can write

$$Y_i(d) = \text{ATT}(D_i) + U_i, \text{ where } \mathbb{E}[U_i|D_i^*] = \mathbb{E}[Y_i(0)|D_i^*].$$

Heterogeneous treatment effects and nonlinearities in the treatment effects are not a problem. Supposing that the constant selection effects assumption holds:

$$\mathbb{E}[Y_i(0)|D_i^*] = \alpha + \delta D_i^*,$$

then

$$\mathbb{E}[Y_i|D_i] = \alpha + ATT(D_i) + \delta \underbrace{[D_i + \mathbb{E}[D_i^*|D_i^* \leq 0]1(D_i = 0)]}_{\text{Control Function}}.$$

$\text{ATT}(D_i)$ can be estimated parametrically with GMM or nonparametrically with a partially linear method.

Relaxing Constant Treatment and Selection Effects

Assume local constant selection effects around the bunching point:

$$\mathbb{E}[U_i|D_i^*] = \alpha + \delta D_i^*, \text{ for } D_i^* \leq \zeta,$$

for some $\zeta > 0$.

Then

$$\mathbb{E}[Y_i|D_i] = \alpha + ATT(D_i) + \delta[D_i + \mathbb{E}[D_i^*|D_i^* \leq 0]1(D_i = 0)], \text{ for } D_i^* \leq \zeta.$$

With some continuous differentiability conditions,

$$\lim_{d\downarrow 0} \frac{d}{dD_i} ATT(D_i)\Big|_{D_i=d} = \lim_{d\downarrow 0} \mathbb{E}[Y_i'(0)|D_i = d]$$

← Extensive Margin Effect for the Marginal Observations

This quantity can be identified:

$$\lim_{d\downarrow 0} \mathbb{E}[Y_i'(0)|D_i = d] = \lim_{d\downarrow 0} \frac{d}{dD_i} \mathbb{E}[Y_i|D_i]\Big|_{D_i=d} - \frac{\Delta}{\mathbb{E}[D_i^*|D_i^* \leq 0]}.$$

Relaxing All Linearity Restrictions

Without loss of generality, we can write

$$Y_i = \text{ATT}(D_i) + \mathbb{E}[Y_i | D_i^*] + \varepsilon_i, \text{ where } \mathbb{E}[\varepsilon_i | D_i^*] = 0.$$

Relaxing All Linearity Restrictions

Without loss of generality, we can write

$$Y_i = \text{ATT}(D_i) + \mathbb{E}[Y_i|D_i^*] + \varepsilon_i, \text{ where } \mathbb{E}[\varepsilon_i|D_i^*] = 0.$$

Assume:

1. D_i^* continuously distributed around $D_i^* = 0$
2. $\mathbb{E}[Y_i(0)|D_i^*]$ monotonic and continuously differentiable near the bunching point
3. ε_i and D_i^* are independent at the bunching point

Then

$$\lim_{d\downarrow 0} \mathbb{E}[Y_i'(0)|D_i = d] = \text{sgn}(\Delta) \cdot \frac{\lim_{d\downarrow 0} f_D(d)}{\mathbb{P}(D_i = 0) \cdot f_{\mathbb{E}[Y(0)|D^*]}(\lim_{d\downarrow 0} \mathbb{E}[Y_i|D_i = 0])}.$$

For $D_i = 0$, $Y_i = \mathbb{E}[Y_i(0)|D_i^*] + \varepsilon_i$ is a convolution.

The density $f_{\mathbb{E}[Y(0)|D^*]}$ can be deconvoluted using the distribution of Y_i at the bunching point, and the distribution of Y_i near the bunching point.

Final Remarks

- Find link to Stata codes used to implement the approach on my website.
- For applied researchers interested on these methods: our JOLE paper on the effect of maternal labor supply on children's skills showcases many of these techniques in actual empirical work.
- We developed several tests of the assumptions, and many other informal robustness checks for the assumptions. See appendix of the paper that introduces the endogeneity correction (JBES) and applied papers.
- Bunching can also be used to test identification in other types of models. We can test parallel trends in two-way fixed effects models without the need of using pre-trends. Simply add a dummy of the bunching point, or a dummy of the bunching point interacted with one of the periods.
- Econometricians: huge amount of unanswered questions, requiring all types of technical skills. Many opportunities for specialists in nonparametric identification, boundary estimation, quantiles, measurement error, deconvolutions, censoring, etc.

Thank You!

Estimation Details in General Models

$$\lim_{d \downarrow 0} \mathbb{E}[Y_i'(0)|D_i = d] = \lim_{d \downarrow 0} \frac{d}{D_i} \mathbb{E}[Y_i|D_i = d] + \frac{\Delta}{\mathbb{E}[D_i^*|D_i^* \leq 0]}$$

More efficient but less convenient:

1. $\lim_{d \downarrow 0} \frac{d}{dD_i} \mathbb{E}[Y_i|D_i = d]$ is estimated with a specialized boundary derivative estimator, e.g. Dai, Tong and Genton (2016).
2. $\Delta = \mathbb{E}[Y_i|D_i = 0^+] - \mathbb{E}[Y_i|D_i = 0]$, estimate $\mathbb{E}[Y_i|D_i = 0^+]$ with local linear regression, and $\mathbb{E}[Y_i|D_i = 0]$ by simple average at bunching point.

Estimation - Limits of Densities

We use Pinkse and Schurter (2021)'s estimator:

$$\hat{f}_D(0^+) = \frac{1}{D}\frac{1}{nh}\sum_{i=1}^{n} k\left(D_i/h\right),$$

$$D = \frac{3}{2}\cdot\frac{2+\hat{L}'_D(0)^2h^2 - e^{\hat{L}'_D(0)h}(2-2\hat{L}'_D(0)h)}{\hat{L}'_D(0)^3h^3}$$

$$\hat{L}'_D(0) = \frac{\frac{1}{nh}\sum_{i=1}^{n}(1-2D_i/h)1(0\leq D_i\leq h)}{\frac{1}{nh}\sum_{i=1}^{n}D_i(1-D_i/h)1(0\leq D_i\leq h)},$$

Relaxing All Linearity Restrictions - More Details

$$Y_i = \text{ATT}(D_i) + \mathbb{E}[Y_i|D_i^*] + \varepsilon_i, \text{ where } \mathbb{E}[\varepsilon_i|D_i^*] = 0.$$

Differentiating near the bunching point, we get

$$\mathbb{E}[Y_i'(0)|D_i^* = 0] = \lim_{d\downarrow 0} \frac{d}{dD_i}\mathbb{E}[Y_i|D_i]\Big|_{D_i=d} - \lim_{d\downarrow 0} \frac{d}{dD_i}\mathbb{E}[Y_i(0)|D_i]\Big|_{D_i=d}.$$

Assume continuous selection effects at the bunching point, then

$$\lim_{d\downarrow 0} \frac{d}{dD_i}\mathbb{E}[Y_i(0)|D_i]\Big|_{D_i=d} = \frac{d}{dD_i}\mathbb{E}[Y_i(0)|D_i^*]\Big|_{D_i^*=0}.$$

Assume that $\mathbb{E}[Y_i(0)|D_i^*]$ is monotonic at the bunching point. Then, by the change in variables theorem,

$$\frac{d}{dD_i}\mathbb{E}[Y_i(0)|D_i^*]\Big|_{D_i^*=0} = \text{sgn}(\Delta) \cdot \frac{f_{D^*}(0)}{f_{\mathbb{E}[Y(0)|D^*]}(\mathbb{E}[Y_i(0)|D_i^* = 0])}.$$

Assume that $f_{D^*}(d)$ is continuous at $d = 0$, then $f_{D^*}(0) = \lim_{d\downarrow 0} f_D(d)$.

For $D_i^* \leq 0$, $Y_i = \mathbb{E}[Y_i(0)|D_i^*] + \varepsilon_i$. Assume that ε_i and D_i^* are independent at the bunching point, then we can deconvolute $f_{\mathbb{E}[Y(0)|D^*]}(\mathbb{E}[Y_i(0)|D_i^* = 0])$ using observations at the bunching point.