Identification and Estimation of Average Marginal Treatment Effects with a Bunching Design*

Carolina Caetano
University of Georgia

Gregorio Caetano
University of Georgia

Eric Nielsen
Federal Reserve Board

April 2022

Abstract

We show that bunching on the treatment variable can be used for identification of the average marginal treatment effect near the bunching point. The approach requires no functional form or distributional assumptions, no exclusion restrictions, and no special data structures (e.g. panel data) and instead relies on continuity conditions that are similar to those imposed in Regression Discontinuity Designs with continuous treatment, and a monotonicity condition. Adding some types of parametric structures to the endogeneity bias allows the identification of all average marginal treatment effects. We provide estimators which can be implemented with off-the-shelf packaged software, and we apply the method to estimate the effects of watching television on children's cognitive and non-cognitive skills.

1 Introduction

In this paper, we introduce a new design for the identification of treatment effects in the presence of endogeneity. The design leverages the phenomenon of bunching at one extreme of the treatment variable to identify the average marginal treatment effect near the bunching point, which we clarify below. The approach requires no other special phenomena except for the bunching, and it imposes no exclusion restrictions, no functional form or distributional assumptions, nor does it need special data structures, such as panel, pooled cross section, etc. Therefore, it opens a reasonable path for identification of treatment effects when none of the established methods (e.g. instrumental variables, Regression Discontinuity Design, and difference-in-differences) are available.

The list of treatment variables with bunching at one extreme of their distribution is extensive. We refer the reader to Caetano (2015) and Caetano et al. (2020), where numerous examples are provided, as well as to the papers that have applied these methods. Such bunching is very common when the treatment is a choice which must not be negative, as observations often bunch at zero. For example, although one may smoke any number of cigarettes, a disproportional share of people do

*We thank Stéphane Bonhomme, Guido Imbens and Elie Tamer for pushing us towards developing the ideas of an earlier paper in a new direction, which led to this paper. We also thank David Card, Len Goff, Stefan Hoderlein, Hugo Jales, Joris Pinkse, Demian Pouzo, and Karl Schurter. The analysis and conclusions set forth here are those of the authors and do not indicate concurrence by other members of the research staff, the Board of Governors, or the Federal Reserve System.

not smoke at all. Analogously, a disproportional share of children do not watch TV or participate in extra-curricular activities, and a disproportional share of mothers spend no time on paid work. Other settings may also result in this type of bunching, e.g. minimum wages, minimum schooling laws, minimum working age laws, and minimum 401K contribution constraints.

To clarify the method, we focus on the case where the treatment is bunched at the lower extreme of the support, say at zero. The first requirement of the method is that there exists an **unobservable** variable that indexes the endogenous variation in the outcome among variables at the bunching point. This variable must be such that observations indexed with marginally small negative values of that variable are comparable to observations with marginally small, positive treatment values. This variable needs no structural interpretation, but it is helpful to think of it as a desired treatment in a constrained choice problem. If the treatment cannot be negative, then those who desire a positive treatment actually choose their desired, positive level, but those who desire a negative treatment choose zero. The desired choice indexes the endogeneity at the bunching point, so that those who desired a marginally negative amount are comparable to those who chose a marginally positive amount. In this sense, this bears some resemblance to the Regression Discontinuity Design, because we make such types of threshold comparability assumptions. It is important to note that this is not a censoring model. The treatment is the observed actual treatment, not the desired choice absent the constraint.

In this setting, exactly at the bunching point, there is no variation in the treatment, but there is variation in the indexing variable, with some observations indexed with marginally negative values and others with very negative values (that is, some observations would have chosen a marginally negative treatment if allowed while others would have chosen a very negative amount.) The distribution of the outcome at the bunching point reveals the convolution of the distribution of the endogenous variation indexed by the unobservable variable and additional random noise from the idiosyncratic error. Importantly, the distribution of the outcome at the bunching point is not affected by variation in the treatment, since all observations there receive the same treatment, zero. At the same time, the distribution of the outcome among those who chose a marginally small positive treatment reveals only the random noise from the idiosyncratic error (it is not affected by either variation in the treatment nor by endogeneity, since both the treatment and the indexing variable are constant there, equal to a marginally small positive amount). This distribution can then be used to deconvolute the distribution of the outcome at the bunching point to reveal the distribution of the endogenous variation in the outcome indexed by the unobservable variable. (In the constrained choice example, the deconvolution reveals the distribution of how the desired negative choices affect the outcome).

An important insight is that although we cannot identify the actual values of the indexing variable, and thus we can never observe directly how it affects the outcome, its effect on the outcome can be obtained by the comparison of the distribution of the indexing variable and the distribution of the resulting endogenous variation in the outcome. This is true because both have the same unique source of randomness: the variation of the indexing variable. It is therefore possible to write one

distribution as a change of variable of the other, and thus the ratio reveals exactly the endogeneity bias. However, we do not have the distribution of the indexing variable at the bunching point, since it is not observed. What we have is (a) the distribution of the indexing variable for positive values of the treatment, since for positive treatments the endogeneity can always be indexed by the treatment, and (b) the distribution of the endogenous variation in the outcome for negative values of the indexing variable, obtained from the deconvolution explained in the previous paragraph. The only place where we may then compare both is exactly at the boundary. That is, we can compare the density of the treatment right above zero with the density of the endogenous variation in the outcome right below zero, thus revealing the endogeneity bias of those exactly at zero, which is all we need in order to identify the average marginal treatment effect among those at the bunching point who are the most indifferent between choosing zero or a marginally marginally positive.

It must be clear from this explanation that continuity requirements will be necessary. Specifically, the assumptions are (1) that the treatment and indexing variables are continuously distributed near the bunching point; (2) that the average marginal treatment effect must be continuous at the bunching point; (3) that the endogeneity bias as a function of the indexing variable must be continuous at the bunching point, that is, observations with marginally small positive treatment must be comparable with observations with marginally small negative value of the indexing variable; (4) that the distribution of the idiosyncratic errors must be continuous at the bunching point, i.e. that the remaining variation after accounting for the factors that cause endogeneity bias must be the same at the bunching point as right above. This condition implies a type of local independence between the idiosyncratic errors and the indexing variable at the bunching point. Additionally, (5) we require that, at the bunching point, as the value of the indexing variable decreases, the expected outcome either increases or decreases monotonically (in the constrained choice example, as the desired choice becomes more negative, the expected outcomes either increase or decrease monotonically.) Note that this requirement is on the expectation of the outcomes, not on the individual observations. We use this requirement to obtain the sign of the endogeneity bias at the bunching point, which is necessary for identification. The assumptions of our method are reminiscent of the "validity" assumptions on regression discontinuity designs (RDD). Conditions (1)-(3) are identical if one understands the indexing variable below the bunching point and the treatment above the bunching point as the running variable. Conditions (4) and (5) are not made in the standard RDD, due to the fact that the treatment is binary in standard RDDs. However, Dong et al. (2021) considers the identification of marginal treatment effects with a continuous treatment in the RDD, and there an assumption related to condition (4) is made, as well as a monotonicity assumption which is hard to compare to condition (5) because in our context there is no first stage regression.

The method may also be implemented by first conditioning on covariates and then aggregating the conditional marginal treatment effects. In this case, all requirements are substituted by conditional requirements, which weakens the assumptions. Specifically, conditioning on covariates means that the endogeneity is now indexed by additional variables, thus weakening the continuity

conditions, and that, as the indexing variable becomes more negative, the expected outcome may increase for some values of the controls and decrease for others, thus weakening the monotonicity condition.

The strategy outlined so far allows the identification of the average marginal treatment effect at the bunching point among those that are the most similar to those who chose a marginally positive amount. Precisely, this approach identifies the treatment effect at the bunching point among the subpopulation with index variable exactly at zero. If we understand the treatment as resulting from a constrained optimization problem, then this is the average marginal treatment effect among those who chose the bunching point but for whom the restriction was not binding, that is, those most indifferent between choosing the bunching point or a marginally positive amount, and thus would be exactly the group that would react to a marginal policy. By the continuity conditions, another interpretation is that this method identifies the average marginal effect among those that chose a marginally positive treatment, which is why we refer to this, for simplicity, as the average marginal effect near the bunching point. The identification is local only in that the average effect of a marginal increase in the treatment is nonparametrically identifiable exclusively at that point and nowhere else. However, we do identify the true average marginal treatment effect, so the effect is not local in the sense commonly associated with treatment effects, i.e. it is not local to compliers for a specific source of random variation. Moreover, we show that if the endogeneity bias is known up to a scalar unknown parameter (e.g. when the endogeneity can be specified as a linear function of the index variable), then the average marginal treatment effects may be nonparametrically identified everywhere.

The exploration of bunching for identification started with Saez (2010), which gave rise to a large empirical literature that leverages bunching on the outcome variable to identify elasticities in structural models. Theoretical treatments of identification in this context are given by Blomquist et al. (2021), Bertanha et al. (2021) and Goff (2020), who relax some of the strong parametric requirements in the typical models in favor of shape restrictions which yield partial identification results. The first instance of exploration of bunching on the treatment variable for identification in reduced form models is in Caetano (2015), which shows that it is possible to nonparametrically identify the presence of endogeneity (but does not address it), see also Caetano et al. (2021b). Caetano et al. (2020) shows that treatment effects may be identified in models with separable endogeneity by making semi-parametric or symmetry assumptions on the distribution of confounders, and partially identified with shape restrictions such as convexity. Overall, the entire theoretical bunching literature is characterized by the valuable endeavor to achieve identification without the use of exclusion restrictions or special data structures, which is, of course, a very challenging proposition. The trade-off has been the adoption of functional form and distributional assumptions. In this paper, we introduce the first strategy leveraging bunching phenomena which escapes this trade-off.

We develop an estimator of the average marginal treatment effect near the bunching point which relies on well known off-the-shelf estimators as building blocks. The approach requires the estimation of standard boundary quantities, including (1) the expected outcome, and the expected derivative of

the outcome at the boundary of the bunching point, estimated with a standard local linear estimator, which is well known to have excellent boundary properties, and (2) the density of the treatment at the boundary of the bunching point, which is estimated using the approach in Pinkse and Schurter (2021), which provides the same rates of convergence at the boundary of the bunching point as can be achieved in interior points, and guarantees a non-negative density estimate, both extremely desirable properties in our case. Our estimator also requires the estimation of a deconvolution of the density of the outcome at the bunching point, where the error in the deconvolution is derived from the distribution of the outcome right above the bunching point. The deconvolution estimator is therefore standard, with the denominator estimated with local linear regressions.

We apply our method to the estimation of the average marginal treatment effect of watching television (TV) on a child's cognitive and non-cognitive skills using time diary data from the Panel Study of Income Dynamics Child Development Supplement (PSID-CDS). Children who watch different amounts of TV are likely to have different skills for reasons other than TV time *per se*, so TV time is likely endogenous. However, 5% of children do not watch any TV at all, a discontinuously high proportion relative to those who spend little time watching TV. We estimate the average effect of a marginal increase in the amount of time watching TV among the children who currently do not watch TV, but are the most indifferent towards watching some TV. Our results show that this marginal effect is positive for cognitive skills but negative for non-cognitive skills.

The rest of the paper is organized as follows. Section 2 presents the model and explains how identification of average marginal treatment effects may be achieved by shutting down the treatment variation (as opposed to shutting down the variation of unobservables as is typically done with exclusion restrictions). Section 3 then shows how that strategy may be carried out in the presence of bunching on the treatment variable, thus demonstrating how identification may be achieved near the bunching point, and also beyond if the endogeneity follows a parametric structure. Section 4 provides estimators. The application to the effects of TV watching on children's cognitive and non-cognitive skills is in Section 5, and we conclude in Section 6. The appendix develops an example of a choice model under non-negativity constraints which provides further intuition about the indexing unobservable variable, and relates it to the TV application.

2 Setup and preliminaries

Throughout the paper, probability spaces are self-evident and thus not explicitly defined. Assume that functions are measurable wherever this is needed, and that moments exist wherever written.

For a scalar treatment intensity x, consider the potential outcome random function $Y_i(x)$, which may vary per individual i. Let Y_i and X_i be the observed outcome and treatment variables, respectively, then

$$Y_i = Y_i(X_i).$$

Suppose the function $Y_i(\cdot)$ is differentiable, and let $Y_i'(x)$ be its derivative at x. The average marginal treatment effect at $X_i = x$ is $\beta(x) := \mathbb{E}[Y_i'(x)|X_i = x]$.

In this section, we provide two results which are the basis of our identification strategy. The first result and its corollary state that the observed outcome can be additively decomposed into a component that determines the treatment effects, a component that encapsulates all the endogeneity, and an idiosyncratic error.

Theorem 2.1. *Let X_i be a continuously distributed random variable, and suppose that $Y_i(x)$ is continuously differentiable with probability one, $\mathbb{E}[Y_i(x)|X_i = x']$ is continuously differentiable with respect to x and x', and $\mathbb{E}[|Y_i'(x)|\,|X_i = x'] < \infty$ for all x and x'. Then, with probability one, the observed outcome $Y_i = Y_i(X_i)$ can be uniquely decomposed as*

$$Y_i = B(X_i) + u(X_i) + \epsilon_i,$$

where $\beta(x) = B'(x)$ is the average marginal treatment effect, $B(0) = 0$, and $\epsilon_i = Y_i - \mathbb{E}[Y_i|X_i]$.

Proof. Define the function $G(x, x') = \mathbb{E}[Y_i(x)|X_i = x']$, $G_1(x, x') = \frac{\partial}{\partial x}G(x, x')$ and $G_2(x, x') = \frac{\partial}{\partial x'}G(x, x')$. First, we show that

$$\beta(x) = G_1(x, x).$$

We prove the stronger result that $G_1(x, x') = \mathbb{E}[Y_i'(x)|X_i = x']$.

$$G_1(x, x') = \frac{d}{dx}\mathbb{E}[Y_i(x)|X = x'] = \lim_{h\downarrow 0}\mathbb{E}\left[h^{-1}(Y_i(x + h) - Y_i(x))|X_i = x'\right] = \lim_{h\downarrow 0}\mathbb{E}[Y_i'(l(x, h))|X_i = x'],$$

for a $l(x, h) \in (x, x + h)$ by the Mean Value Theorem. Since $\mathbb{E}[|Y_i'(l(x, h))|\,|X_i = x'] < \infty$, by the Dominated Convergence theorem,

$$G_1(x, x') = \mathbb{E}\left[\lim_{h\downarrow 0} Y_i'(l(x, h))|X_i = x'\right] = \mathbb{E}[Y_i'(x)|X_i = x'],$$

where the last equality follows by the continuous differentiability of $Y_i'(x)$.

Next, we prove the decomposition. If $x \geq 0$, define

$$\begin{aligned} B(x) &= \int_0^x G_1(\nu, \nu)d\nu = \int_0^x \beta(\nu)d\nu, \\ u(x) &= \int_0^x G_2(\nu, \nu)d\nu + G(0, 0). \end{aligned} \tag{1}$$

Then, by the Second Fundamental Theorem of Calculus, $B(0) = 0$ and, since β is continuous, by the First Fundamental Theorem of Calculus, $B'(x) = \beta(x)$. Since $\frac{d}{dx}G(x, x) = G_1(x, x) + G_2(x, x)$, by the Second Fundamental Theorem of Calculus again, $G(x, x) = \int_0^x (G_1(\nu, \nu)+G_2(\nu, \nu))d\nu+G(0, 0) = B(x) + u(x)$.

Since $Y_i = G(X_i, X_i) + \epsilon_i$, then for $X_i \geq 0$,

$$Y_i = B(X_i) + u(X_i) + \epsilon_i.$$

The decomposition is unique because $B(x)$ is unique. This is because if a function φ satisfies $\varphi'(x) = \beta(x)$, then $\varphi(0) = 0 \implies \varphi(x) = B(x)$. Then, for a given pair Y_i and X_i, ϵ_i and $B(X_i)$ are

unique and, therefore, so is $u(X_i)$.

If $x < 0$, define $B(x) = -\int_x^0 \beta(\nu)d\nu$ and $u(x) = -\int_x^0 G_2(\nu,\nu)d\nu + G(0,0)$, then, by analogous arguments, the unique decomposition also holds for $X_i < 0$. □

The following corollary shows that the function u determines the endogeneity bias.

Corollary 2.1. *Under the assumptions of Theorem 2.1, u is differentiable, and*

$$\frac{d}{dx}\mathbb{E}[Y_i|X_i = x] = \beta(x) + u'(x).$$

Proof. This corollary is trivial, except for the differentiability of u, which follows from the continuity of G_2 and the First Fundamental Theorem of Calculus. Moreover, $u'(x) = G_2(x,x)$. □

This corollary specifies the fundamental problem of causality. Specifically, when we vary x, the differences in the outcome reflect the treatment effect plus the effect of the endogenous treatment selection. It is generally impossible to obtain variation in $B(x)$ without also obtaining variation in $u(x)$. Most identification approaches solve this problem either by assumption (e.g. assuming $u'(x) = 0$, or, more realistically, that $u'(X_i) = 0$ conditional on controls), by exploring variation in the outcome along alternative dimensions where the corresponding endogeneity term has derivative equal to zero (e.g. instrumental variable, difference-in-differences), or by identifying $u(x)$ with a combination of assumptions including two equation models, exclusion restrictions, functional forms, and distribution assumptions (e.g. control functions).

Here, we propose an alternative strategy that considers an instance in which we may obtain variation in $u(X_i)$ without variation of $B(X_i)$. The main issue is that, in the situation we have in mind, the $u(X_i)$ themselves are never observed. However, the following result shows that if the distribution of $u(X_i)$ can be recovered, this may be enough for the identification of $u'(x)$.

Theorem 2.2. *Suppose that X_i admits a density function, $f_X(x)$, and that $u(\cdot)$ is differentiable, monotonic, and $\mathbb{P}(u'(X) = 0) = 0$. Then, $u(X_i)$ admits a density function, $f_{u(X)}(u(x))$, and if $f_{u(X)}(u(x)) > 0$, the endogeneity bias satisfies*

$$u'(x) = sgn(u'(x))\frac{f_X(x)}{f_{u(X)}(u(x))}, \tag{2}$$

where $sgn(\cdot)$ is the sign function.

Proof. Since the conditions imply that u is injective with probability one, this theorem is a well known consequence of any of several existing integration by substitution results (see, e.g. Fremlin (2010), Theorem 263D, which also covers the multivariate case). □

To understand the relation in equation (2), suppose e.g. that $u(x) = \delta x$, for some $\delta > 1$. Then $f_{\delta X}(\delta x)$ is a compression of $f_X(x)$. To see why, let $V = \delta X$ and $v = \delta x$, then $f_V(v) = f_X(x)$, and so f_X gives the same weight to 1, for example, that f_V gives to $1/\delta$. It is a change to a smaller scale, much like compressing the horizontal dimension of a map while keeping the same total area

by stretching the vertical dimension. Figure 1 depicts this relation, noting that $f_{u(X)}$ may not only compress or dilate f_X, but may also shift the entire distribution.

Figure 1: Identifying the endogeneity bias from the comparison of densities

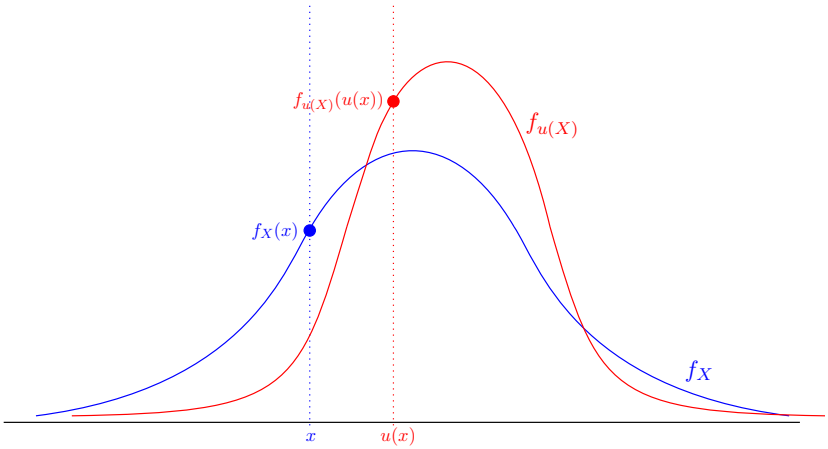


Figure 1 depicts the relation between the density of X_i and the density of $u(X_i)$. The ratio of the heights of the blue and red dots determines the endogeneity bias at $X_i = x$.

As discussed in the proof of Theorem 2.2, the mathematical relation described in equation (2) is well known, and we do not claim it as ours. Rather, we claim the idea that, considering the decomposition from Theorem 2.1, and Corollary 2.1, the endogeneity bias can be obtained from the comparison of the densities of X_i and $u(X_i)$. Equation (2) also makes evident what the difficulties are. Besides the obviously hard identification of $f_{u(X)}$, the sign of $u'(x)$ must also be identified. If those two problems were solved for many values of x, then it might be possible to identify $u'(x)$ implicitly through equation (2). More likely, in order to identify $u'(x)$ at a specific value of x, one must also know the value of $u(x)$.

In the next section we show that bunching in the treatment variable X_i provides the perfect opportunity to solve all three problems exactly at the bunching point.

3 Identification

For ease of exposition, we focus on the case where X_i has bunching at the lower extreme of the support of its distribution, at $X_i = 0$. The results hold if the lower extreme is at any other value, and analogous results hold when bunching is at the upper extreme of the support. We believe that similar arguments may be used for identification at interior bunching points, but we leave the investigation of that case for the future.

We define a variable X_i^* that is equal to the treatment when it is positive, but which also varies when the treatment is zero:

$$X_i = \max\{X_i^*, 0\}. \tag{3}$$

X_i^* is not observed when it is negative. The characterization of bunched variables as a restriction of an unobservable latent variable is standard, and can be seen in different forms in all the bunching papers cited in the Introduction.

If X_i is chosen through the maximization of a utility function with a non-negativity constraint, which fits almost all the applications with bunching at one corner of the distribution of the treatment we have seen, this type of structure arises naturally, as we demonstrate in Appendix A. In that case, X_i^* may be thought of as the optimal choice under no constraint, or a "desired choice," which can be negative. For those unaccustomed to bunching models, it may be difficult to conceive of a negative choice in many applications. However, we show in Appendix A that bunching at $X_i = 0$ without the existence of a negative desired choice implies that there must also be bunching of preferences. Besides being implausible and difficult to motivate, preference bunching has implications which are unlikely to hold, such as that a relaxation of the budget must imply that all previously bunched observations then choose a positive amount of X_i.[1] More likely, the preferences among those bunched at $X_i = 0$ differ, and just as X_i indexes the preference differences among the unbunched, there must be a way to index the preference differences among those bunched. That index is X_i^*.

Although the utility optimization framework we discuss above is useful, note that we make no structural assumption on X_i^*. In reality X_i^* may be any variable that indexes the endogeneity at $X_i = 0$. The specific conditions on X_i^* are explicitly enumerated below in Assumption 1. Intuitively, these conditions mean that, (1) whatever the nature of X_i^* for very negative values, in a neighborhood of zero, X_i^* must behave similarly to X_i. That is, the potential outcomes of observations with X_i^* slightly negative must be comparable to the potential outcomes of observations with X_i slightly positive; and (2) when $X_i = 0$, X_i^* orders observations by the expected potential outcome. This means that as X_i^* becomes more negative, the expected potential outcome either increases or decreases.

We suppose that $Y_i(x)$ is a differentiable potential outcome which reacts to the treatment x, which is given by the variable X_i, and thus the argument of the potential outcome cannot be negative. We stress this point so there is no confusion: the actual treatment is X_i, which is observed – this is not a censoring model. However, we allow $Y_i(x)$ to be correlated with X_i^* both when $X_i > 0$ (and thus $X_i^* = X_i$) as well as when $X_i = 0$ (and thus $X_i^* \leq 0$). In the rest of this section, we show how

$$\beta(0) = \mathbb{E}[Y_i'(0)|X_i^* = 0]$$

can be identified. This is the average marginal treatment effect among those who are indexed exactly at zero, that is, those at the bunching point for whom the restriction was not binding. In the utility maximization framework discussed above, those are the bunched observations that are the most indifferent between choosing $X_i = 0$ or a marginally positive amount, and therefore are

[1] For example, in our application, X_i is the amount of time per week a child spends watching TV. If watching TV is a good, and all children at $X_i = 0$ made the optimal unconstrained choice, then if an additional hour in the day were available (e.g. due to daylight savings, or because in the weekend the effective number of available hours for leisure increases), then all children must watch a positive amount of TV. This is clearly not true both conceptually (some families do not even own a TV), as well as in the data (there is bunching at zero on the weekends).

the subpopulation that would react to a marginal policy.

The first result is a generalization of Theorem 2.1 which nests the original model when $X_i^* > 0$, but also accounts for the variation of X_i^* when it is negative.

Theorem 3.1. *Let X_i^* be an unobservable variable which is continuously distributed in $(-\infty, h)$, for some $h > 0$. Let $X_i = \max\{X_i^*, 0\}$ with probability equal to one. Finally, suppose that $Y_i = Y_i(X_i)$, where $Y_i(x)$ is continuously differentiable with probability one, $\mathbb{E}[Y_i(x)|X_i^* = x']$ is continuously differentiable with respect to x and x', and $\mathbb{E}[|Y_i'(x)|\,|X_i^* = x'] < \infty$ for all $x \in [0, h)$, and $x' \in (-\infty, h)$, for some $h > 0$. Then, with probability one, for $X_i \in [0, h)$, the observed outcome $Y_i = Y_i(X_i)$ can be uniquely decomposed as*

$$Y_i = B(X_i) + u(X_i^*) + \epsilon_i,$$

where for $x \in (0, h)$, $\beta(x) = B'(x)$ is the average marginal treatment effect, $B(0) = 0$, and $\epsilon_i = Y_i - \mathbb{E}[Y_i|X_i^*]$.

Proof. When $X_i > 0$, $X_i = X_i^*$ with probability one, and therefore the proof is identical to the proof of Theorem 2.1, only caring to make all statements local to $X_i \in [0, h)$. Moreover, the definition of $B(X_i)$, $u(X_i)$ and ϵ_i are unchanged.

Here, we prove the case where $X_i = 0$. Redefine G as

$$G(x, x') = \mathbb{E}[Y_i(x)|X_i^* = x'],$$

which does not affect the previous part of this proof. Define $B(0) = 0$, and, for $x \leq 0$, define

$$u(x) = -\int_x^0 G_2(0, \nu)d\nu + G(0, 0).$$

Then, for $x \leq 0$, by the Second Fundamental Theorem of Calculus, $G(0, x) = -\int_x^0 G_2(0, \nu)d\nu + G(0, 0) = B(0) + u(0)$.

When $X_i = 0$, $Y_i = G(0, X_i^*) + \epsilon_i$, and thus, with probability one,

$$Y_i = B(X_i) + u(X_i^*) + \epsilon_i.$$

□

The following corollary specifies the fundamental problem of causality in the bunching setting.

Corollary 3.1. *Under the assumptions of Theorem 3.1, u is differentiable and, for $x > 0$,*

$$\frac{d}{dx}\mathbb{E}[Y_i|X_i = x] = \beta(x) + u'(x). \tag{4}$$

Moreover,

$$\lim_{x \downarrow 0} \frac{d}{dx}\mathbb{E}[Y_i|X_i = x] = \beta(0) + u'(0). \tag{5}$$

Proof. The first part follows from Corollary 2.1. For the second part, note that the continuity of G_1 guarantees that $\lim_{x\downarrow 0}\beta(x) = \beta(0)$. Next, note that by the continuity of G_2 and the First Fundamental Theorem of Calculus, $u'(x) = G_2(x,x) \to G_2(0,0)$. Finally, we need to show that u is differentiable at zero and $u'(0) = G_2(0,0)$. First, $u(x)$ is continuous at $x = 0$ (it is left-continuous by definition, and right continuous because $u(x) \to G(0,0) = u(0)$ as $x \downarrow 0$.) By the definition of $u(x)$ for $x \leq 0$ and again by the First Fundamental Theorem of Calculus, as $x \uparrow 0$, $u'(x) \to G_2(0,0)$. Since $u(x)$ is continuous at zero, the proof is complete. □

In the bunching environment, for $X_i = 0$, $u(X_i^*)$ varies while $B(X_i) = B(0) = 0$ does not. Since we do not observe X_i^*, we cannot discover $u'(X_i^*)$ directly. However, we can use the following modification of Theorem 2.2 to discover $u'(0)$.

Theorem 3.2. *Suppose that $X_i = \max\{X_i^*, 0\}$, and that, for $X_i = 0$, X_i^* admits a density function $f_{X^*|X=0}(x)$, such that $f_{X^*|X=0}(0) > 0$. Suppose also that $u(\cdot)$ is differentiable and weakly monotonic, and that $\mathbb{P}(u'(X_i^*) = 0|X_i = 0) = 0$. Then, with probability one, $u(X_i^*)$ admits a density $f_{u(X^*)|X=0}(u(x))$, and the endogeneity bias at $X_i^* = 0$ satisfies*

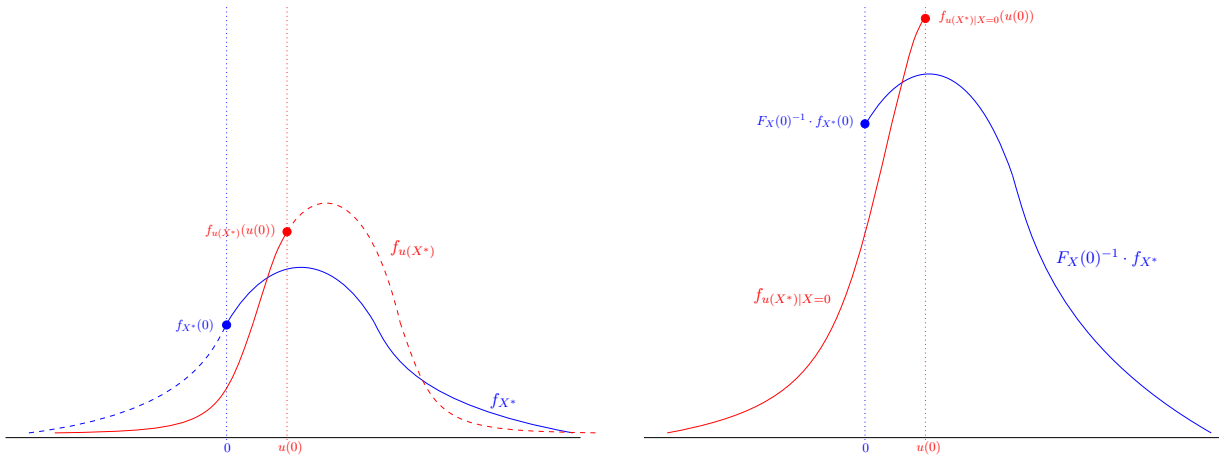
$$u'(0) = sgn(u'(0))\frac{f_{X^*|X=0}(0)}{f_{u(X^*)|X=0}(u(0))}.^{2} \tag{6}$$

Proof. This theorem is a direct modification of Theorem 2.2. By the same results and arguments, $f_{X^*|X=0}(x) = f_{u(X^*)|X=0}(u(x)) \cdot |u'(x)|$. Since $f_{X^*|X=0}(0) > 0$, $f_{u(X^*)|X=0}(u(0)) > 0$, and thus the result follows. □

The idea for the identification of $u'(0)$ can be seen in the left panel of Figure 2, where we assume that $u(x)$ is increasing. Supposing that the densities $f_X(x)$ and $f_{u(X)}(u(x))$ exist everywhere, then the figure depicts the same relationship as in Figure 1. However, we can only hope to identify $f_{X^*}(x)$ for $x \geq 0$, and $f_{u(X^*)|X=0}(v)$ for $v \leq u(0)$. Therefore, only at $x = 0$ we may hope to obtain both quantities needed for the application of Theorem 3.2.

[2]Note that, although we do not need it for our results, equation (6) holds not only at $X_i^* = 0$, but also at any value x such that $f_{u(X^*)|X=0}(u(x)) > 0$.

Figure 2: Identifying the endogeneity bias from the comparison of densities



The left panel of Figure 2 depicts the relation between the density of X_i^* and the density of $u(X_i^*)$ if u is increasing. The ratio of the heights of the blue and red dots determines the endogeneity bias at $X_i = x$. The right panel depicts what we can hope to identify with bunching: $F_X(0)^{-1} \cdot f_{X^*}(x)$, for $X \geq 0$, and $f_{u(X^*)|X=0}(v)$, for $v \leq 0$, where $F_X(0)$ is the probability of bunching at $x = 0$.

The following assumption collects all the requirements of the method. Those who prefer structural equation models with explicitly defined unobservables to the potential outcomes framework used here can read Assumption 4 in Appendix B instead.

Assumption 1. *Suppose that X_i^*, $X_i := \max\{X_i^*, 0\}$, $Y_i(x)$ and $Y_i = Y_i(X_i)$ are random variables that satisfy the following conditions.*

1. *X_i^* has a density $f_{X^*}(x) \leq C \leq \infty$ for $x \leq h$, for some $C, h > 0$, which is continuous at zero and positive in a neighborhood of zero.*

2. *$Y_i(x)$ is continuously differentiable in x with probability one, $\mathbb{E}[Y_i(x)|X_i^* = x']$ is continuously differentiable in x and x', and $\mathbb{E}[|Y_i'(x)|\,|X_i^* = x'] < \infty$ for all $x \in [0, h)$ and $x' \in (-\infty, h)$, for some $h > 0$.*

3. *$\mathbb{E}[Y_i(0)|X_i^* = x]$ is monotonic for $x \leq 0$, and $\mathbb{P}\big(\frac{d}{dX_i^*}\mathbb{E}[Y_i(0)|X_i^*] = 0 \big| X_i = 0\big) \in \{0, 1\}$.*

4. *If the probability in the previous item is zero, then additionally either (a) i: $Y_i(0)|X_i^* = x \to_d Y_i(0)|X_i^* = 0$ as $x \downarrow 0$; ii: $Y_i(0) - \mathbb{E}[Y_i(0)|X_i^*] \perp\!\!\!\perp X_i^*|X_i = 0$, and iii: $f_{\mathbb{E}[Y(0)|X^*]|X=0}(y) \leq C < \infty$, for all y and some $C > 0$; or the weaker condition (b) suppose that $Y_i(0)|X_i = 0$ admits a density $f_{Y(0)|X=0}$, and that $F_{Y(0)-\mathbb{E}[Y(0)|X^*]|X=x}$ converges pointwise as $x \downarrow 0$. Denote the limit distribution as $F_{Y(0)-\mathbb{E}[Y(0)|X^*]|X=0^+}$. Define W_i, a random variable derived from the deconvolution of $dF_{Y(0)-\mathbb{E}[Y(0)|X^*]|X=0^+}$ from $f_{Y(0)|X=0}$. Then, $f_{W|X=0}(\mathbb{E}[Y_i(0)|X_i^* = 0]) = f_{\mathbb{E}[Y(0)|X^*]|X=0}(\mathbb{E}[Y_i(0)|X_i^* = 0])$ hold.*[3]

[3] $f_{\mathbb{E}[Y(0)|X^*]|X=0}$ is the density of $\mathbb{E}[Y_i(0)|X_i^*]$ conditional on $X_i = 0$. We show that it exists in either case in the proof of Theorem 3.3 below.

Assumption 1 is similar in nature to some of the conditions of the Regression Discontinuity Design (RDD), which may help clarify it. We discuss the requirements next.

1. Assumption 1 (1) requires that X_i^* be continuously distributed in a neighborhood of zero, which is analogous to the requirement that the running variable be continuously distributed near the threshold in the RDD. One can therefore think of X^* as a sort of "running variable" in our setting.

2. Assumption 1 (2) requires that the treatment effects be continuous in a neighborhood of zero. This is a similar requirement to the continuity of the treatment effects at the threshold in the RDD.

   Assumption 1 (2) also requires that the effects of the unobservable confounders on the expected potential outcome is continuous at zero. This is also a requirement in the RDD, although here the continuity requirement extends to all the negative side of the running variable (because we want to apply a decomposition Theorem like 2.1). An intuitive way to understand this condition is to think that those with marginally close values of X_i^* are comparable (in terms of confounders.) In particular, those with a marginally negative X_i^* are comparable to those with marginally positive X_i.

3. Assumption 1 (3) refers specifically to $X_i = 0$. The monotonicity requirement means that, as X_i^* decreases from zero, the aggregated influence of the confounders must always be in the same direction – either increasing or decreasing the potential outcome at zero treatment. It is important to emphasize that the requirement is an aggregated one: confounders may affect individuals in different ways, and there may be individuals in the sample, say i and j such that, if we were to decrease their levels of X_i^* and X_j^* from zero, respectively, $Y_i(x)$ would increase and $Y_j(x)$ would decrease.

   There are no monotonicity requirements in the RDD, except for treatment monotonicity to the variation in the running variable, which is not necessary here. However, the standard RDD has a binary treatment. As far as we know, Dong et al. (2021) is the only available identification approach in the RDD with a continuous treatment, although in that paper the objective is not the identification of average marginal treatment effects, but rather the effect of the variation in the treatment quantile.[4] Dong et al. (2021) impose a condition on the monotonicity of the quantiles of the treatment on the unobservables. This condition is somewhat related, in specifying that the selection is monotonic, but drawing a direct parallel is hard.

   The probability condition means that either there is no endogeneity at $X_i = 0$ (thus $\mathbb{E}[Y_i(0)|X_i^*]$ is constant for $X_i^* \leq 0$ with probability one), or otherwise, if there is endogeneity, then $\mathbb{E}[Y_i(0)|X_i^*]$ must vary almost everywhere for $X_i^* \leq 0$. When there is endogeneity, this is used in order to apply Theorem 3.2.

[4] To the extent of our knowledge, the only available approach for identification of marginal treatment effects of a multivalued treatment in the RDD is Caetano et al. (2021a), which identifies average marginal treatment effects among compliers in a multivalued but discrete treatment RDD setting using covariate separability conditions.

4. When there is endogeneity at $X_i = 0$, Assumption 1 (4) is used to obtain the density $f_{u(X^*)|X=0}(u(0))$ from the distribution of the outcome at $X_i = 0$. Because of Theorem 3.1, $Y_i = u(X_i^*) + \epsilon_i$, and we must therefore deconvolute the distribution of the idiosyncratic errors from the distribution of the outcomes.

   The stronger condition (4a) yields a more intuitive proof. The first requirement, *i*, is the continuity of the distribution of $Y_i(0)$ as X_i^* approaches zero from above. As in the previous item, note that this is conditional on X_i^*, not on X_i, so it means that those observations with X_i marginally positive must not only have similar means, but also a similar distribution to the observations with X_i^* marginally negative. The second requirement, *ii*, is the independence of $Y_i(0) - \mathbb{E}[Y_i(0)|X_i^*]$ and the selection index X_i^* among those who chose zero treatment. Intuitively, it means that X_i^* is a sufficient index of the endogeneity, so that conditional on it, any remaining variation on the outcome among the observations at $X_i = 0$ is independent of the index. Note that the mean independence holds automatically, and thus the requirement is the extension of this to the other moments of the distribution. Technically, the results hold under only **subindependence**, a much weaker condition which is well known in the deconvolution literature (see e.g. Schennach 2019) and is different, but not stronger than, mean independence.

   Assumption 1 (4b) is the weakest condition for identification through the deconvolution. It yields the exact same identification formula for $f_{u(X^*)|X=0}$ as (4a), and thus the identification strategy and resulting estimators are unchanged. However, identification through this condition is not intuitive, though we do try to clarify it in Footnote 7. Condition (4b) is essentially a local independence condition which is reminiscent of the rank similarity condition in Dong et al. (2021) in the context of RDD with a continuous treatment. It means that once we clean up $dF_{Y_i(0)-\mathbb{E}[Y_i(0)|X_i^*]|X=0^+}$ from $f_{Y_i(0)|X=0}$, whatever remains in excess of $\mathbb{E}[Y_i(0)|X_i^*]$ may be dependent of X_i^*, but exactly at $\mathbb{E}[Y_i(0)|X_i^* = 0]$ that dependency must not exist. In other words, those with $X_i = 0$ with endogeneity equal to the endogeneity right at the threshold, $\mathbb{E}[Y_i(0)|X_i^* = 0]$, also have an idiosyncratic error $Y_i(0) - \mathbb{E}[Y_i(0)|X_i^*]$ which varies similarly to the errors of those right above the threshold.

Theorem 3.3. *If Assumptions 1 holds, then $\beta(0)$ is identifiable.*

Proof. Assumptions 1 (1) and (2) imply that Theorem 3.1 holds. This allows us to write the decomposition for all X_i and X_i^* smaller than some $h > 0$, and to apply Corollary 3.1 to derive $\beta(0)$ as a function of an identifiable quantity and $u'(0)$, through equation (5).

When there is endogeneity at $X_i = 0$, i.e. $u(X_i^*) = \mathbb{E}[Y_i(0)|X_i^*]$ varies with positive probability when $X_i^* \leq 0$, Assumptions 1 (1)-(3) allow us to apply Theorem 3.2 to establish the existence of $f_{u(X^*)|X=0}$. In this case, we can write $u'(0) = \text{sgn}(u'(0)) \cdot f_{X^*|X=0}(0)/f_{u(X^*)|X=0}(u(0))$. The following lemmas show how $u'(0)$ can be identified in this case.

Lemma 1. *If Assumption 1 (1) holds, then $f_{X^*|X=0}(0)$ is identifiable.*

Proof. By Assumption 1 (1), the limit $f_X(0)_+ := \lim_{x \downarrow 0} f_X(x)$ exists and is identifiable. Moreover, $F_X(0) := \mathbb{P}(X_i = 0) > 0$. Therefore,

$$f_{X^*|X=0}(0) = \lim_{x \uparrow 0} \frac{f_{X^*}(x)}{F_X(0)} = \lim_{x \downarrow 0} \frac{f_X(x)}{F_X(0)} = \frac{f_X(0)_+}{F_X(0)}$$

is identifiable. □

Lemma 2. *If Assumption 1 holds and there is endogeneity at $X_i = 0$, then $f_{u(X^*)}(u(0))$ is identifiable.*

Proof. By Assumptions 1 (1) and (2), we can apply Theorem 3.1 to find that, for $X_i = 0$, $Y_i = u(X_i^*) + \epsilon_i$. This means that the variation of Y_i at zero informs us about the variation of $u(X_i^*)$. However, this variation is convoluted with the variation of ϵ_i. We must therefore deconvolute the distribution of $u(X_i^*)$ from the sum.

First, we prove this using Assumption 1 (4a). For any y, u and e, define the cumulative distribution functions $F_{Y|X=0}(y) = \mathbb{P}(Y_i \leq y | X_i = 0)$, $F_{u(X_i^*)|X=0}(u) = \mathbb{P}(u(X_i^*) \leq u | X_i = 0)$, $F_{\epsilon|X=0}(e) = \mathbb{P}(\epsilon_i \leq e | X_i = 0)$, and $F_{\epsilon|X=0^+}(e) = \lim_{x \downarrow 0} F_{\epsilon|X=x}(e)$, where the latter is well defined by Assumption 1 (4a) i. Then,

$$\begin{aligned} F_{Y|X=0}(y) &= \mathbb{P}(u(X_i^*) + \epsilon_i \leq y | X_i = 0) = \int F_{u(X_i^*)|X=0}(y - e) dF_{\epsilon|X=0}(e) \\ &= \int F_{u(X_i^*)|X=0}(y - e) dF_{\epsilon|X=0^+}(e), \end{aligned}$$

which follows by conditions (4a) i and ii.[5]

Then, by Assumption 1 (4a) iii and the Dominated Convergence Theorem, $Y_i | X_i = 0$ has a density function, $f_{Y|X=0}$ and we can write the convolution inverse problem[6]

$$f_{Y|X=0}(y) = \int f_{u(X^*)|X=0}(y - e) dF_{\epsilon|X=0^+}(e). \tag{7}$$

This problem has a well known closed form solution using the Fourier representation (see e.g. Schennach 2021):

$$f_{u(X^*)|X=0}(u(0)) = \frac{1}{2\pi} \int \frac{\mathbb{E}[\exp(\mathbf{i}\xi Y_i) | X_i = 0]}{\mathbb{E}[\exp(\mathbf{i}\xi(Y_i - u(0))) | X_i = 0^+]} \exp(-\mathbf{i}\xi u(0)) d\xi,$$

where $\mathbf{i} = \sqrt{-1}$, and $\mathbb{E}[\exp(\mathbf{i}\xi(Y_i - u(0))) | X_i = 0^+] = \lim_{x \downarrow 0} \mathbb{E}[\exp(\mathbf{i}\xi(Y_i - u(0))) | X_i = x]$.

Next, we show that $u(0)$ is identifiable. Since B and u are differentiable (and thus continuous)

[5] Precisely, the second equality follows by condition (4a) ii, and the third equality follows by condition (4a) i and Helly-Bray Theorem, due to the fact that $F_{u(X^*)|X=0}(y - e)$ is bounded and continuous.

[6] Suppose u is non-decreasing in $(-h, 0]$, then $\int f_{u(X^*)|X=0}(y-e) dF_{\epsilon|X=0^+}(e) = \lim_{h \downarrow 0} \int h^{-1}(F_{u(X_i^*)|X=0}(y-e) - F_{u(X_i^*)|X=0}(y-e-h)) dF_{\epsilon|X=0^+}(e) = \lim_{h \downarrow 0} \int f_{u(X_i^*)|X=0}(l(y-e,h)) dF_{\epsilon|X=0^+}(e)$, where the second equality follows because for each h, there exists $l(y-e,h) \in (y-e-h, y-e)$ such that the terms inside the integrals are equal by the Mean Value Theorem. Then, the result follows by the Dominated Convergence Theorem applied to the sequence $f_{u(X_i^*)|X=0}(l(y-e,h))$ as $h \downarrow 0$. The case where u is non-increasing is analogous.

at zero,

$$u(0) = \lim_{x \downarrow 0}(B(x) + u(x)) = \lim_{x \downarrow 0} \mathbb{E}[Y_i | X_i = x] =: \mathbb{E}[Y_i | X_i = 0^+], \tag{8}$$

which is identifiable by Assumption 1 (1).

Therefore,

$$f_{u(X^*)|X=0}(u(0)) = \frac{1}{2\pi} \int \frac{\mathbb{E}[\exp(\mathbf{i}\xi Y_i)|X_i = 0]}{\mathbb{E}[\exp(\mathbf{i}\xi(Y_i - \mathbb{E}[Y_i|X_i = 0^+]))|X_i = 0^+]} \exp(-\mathbf{i}\xi \mathbb{E}[Y_i|X_i = 0^+]) d\xi \tag{9}$$

is identifiable by Assumption 1 (1).

To prove this using Assumption 1 (4b) instead of (4a), note that $f_{Y|X=0}$ and $F_{\epsilon|X=0^+} = F_{Y_i(0)-\mathbb{E}[Y_i(0)|X_i^*]|X=0^+}$ exist, and therefore the equation

$$f_{Y|X=0}(y) = \int f_{W|X=0}(y - e) dF_{\epsilon|X=0^+}(e)$$

defines the distribution $f_{W|X=0}$, which is identifiable through the right-hand side of equation (9) by Assumption 1 (1). By Assumption 1 (4b), the result follows.[7] □

Lemma 3. *If Assumptions 1 (1)-(3) hold, then sgn$(u'(0))$ is identifiable.*

Proof. By Assumptions 1 (1) and (2), the decomposition from Theorem 3.1 holds, and thus the discontinuity in the outcome at zero satisfies

$$\Delta := \mathbb{E}[Y_i|X_i = 0^+] - \mathbb{E}[Y_i|X_i = 0] = u(0) - \mathbb{E}[u(X_i^*)|X_i^* \leq 0], \tag{10}$$

where $\mathbb{E}[Y_i|X_i = 0^+] = \lim_{x \downarrow 0} \mathbb{E}[Y_i|X_i = x]$. By Assumption 1 (1), Δ is identifiable. Next, we show that $\text{sgn}(u'(0)) = \text{sgn}(\Delta)$.

If $u'(0) > 0$, by Assumptions 1 (1)-(3), in that case u must be non-decreasing, and varying with positive probability, which implies that $\Delta > 0$. Conversely, if $\Delta > 0$, then u must be non-decreasing and varying, which implies that $u'(0) \geq 0$. Since by Theorem 3.2, $|u'(0)| = f_{X^*|X=0}(0)/f_{u(X^*)|X=0}(u(0)) > 0$, and both densities are positive, $u'(0) > 0$. The case $u'(0) < 0$ is analogous.

When $u'(0) = 0$, by Assumptions 1 (1)-(3), if there is endogeneity at $X_i = 0$, Theorem 3.2 holds, and thus $|u'(0)| > 0$, which is absurd. This rules out the possibility that $u'(0) = 0$ when there is endogeneity at $X_i = 0$. It must then be the case that there is no endogeneity at $X_i = 0$, and therefore $u(x)$ is constant almost everywhere for $x \leq 0$. By the continuities of $u(x)$ and $f_{X^*}(x)$ in a neighborhood of zero, $\mathbb{E}[Y_i|X_i = 0] = u(0)$, which implies that $\Delta = 0$. Conversely, if $\Delta = 0$, since by Assumption 1 (1), $\mathbb{P}(X^* < 0) > 0$ and by Assumption 1 (3), u is monotonic, we must have

[7] The idea is that if we generate ξ_i and W_i from $dF_{\epsilon|X_0^+}$ and $f_{W|X=0}$, respectively, and $\tilde{Y}_i = W_i + \xi_i$, then $f_{\tilde{Y}|X=0} \sim f_{Y|X=0}$. Thus, cleaning up the distribution of $\epsilon|X = 0^+$ from $Y|X = 0$ is equivalent to finding the distribution of W. Condition (4b) then means that those who are drawn with $W_i = u(0)$ in the simulated distribution are similar (that is, present the same variation) as those who were drawn with $u(X_i^*) = u(0)$ (i.e., those who are at the threshold $X_i^* = 0$, and vary $dF_{\epsilon|X^*=0}$).

$u(x)$ constant almost everywhere in $(-\infty, 0]$. In this case, by the continuities of $u'(x)$ and $f_{X^*}(x)$ in a neighborhood of zero, $u'(0) = 0$. □

The application of the three lemmas allows us to write, when there is endogeneity at $X_i = 0$,

$$\beta(0) = \lim_{x \downarrow 0} \frac{d}{dx}\mathbb{E}[Y_i|X_i = x] - \text{sgn}(\Delta) \cdot \frac{f_X(0)_+}{F_X(0) \cdot \frac{1}{2\pi} \int \frac{\mathbb{E}[e^{\mathbf{i}\xi Y_i}|X_i=0]}{\mathbb{E}[e^{\mathbf{i}\xi(Y_i - \mathbb{E}[Y_i|X_i=0^+])}|X_i=0^+]} e^{-\mathbf{i}\xi\mathbb{E}[Y_i|X_i=0^+]} d\xi}. \quad (11)$$

When there is no endogeneity at $X_i = 0$, we showed in the proof of Lemma 3 that $u'(0) = 0$, and that, in that case, $\Delta = 0$. Therefore, equation (11) can be used to identify $\beta(0)$ in all cases. □

3.1 Introducing covariates

If a vector of variables which may be used as controls is observed, then it is possible to identify the average marginal treatment effects with weaker assumptions.

Let the vector of observable controls be denoted by Z_i. Define the conditional average marginal treatment effect at $X_i = x$ as $\beta(x, Z_i) = \mathbb{E}[Y_i'(x)|X_i = x, Z_i]$, and

$$X_i = \max\{X_i^*, 0\}, \quad \mathbb{P}(X_i^* < 0|Z_i = z) > 0. \quad (12)$$

All the components of the analysis in the previous section may then be redefined analogously to reflect the conditioning on Z_i, e.g. $u(x, Z_i)$, $\epsilon_i = Y_i - \mathbb{E}[Y_i|X_i^*, Z_i]$, and $\Delta(z) = \mathbb{E}[Y_i|X_i = 0^+, Z_i = z] - \mathbb{E}[Y_i|X_i = 0, Z_i = z]$.

Assumption 2. *Suppose that X_i^*, $X_i = \max\{X_i^*, 0\}$, $Y_i(x)$ and Z_i are random variables that satisfy:*

1. *X_i^* has a conditional density $f_{X^*|Z=z}(x) \leq C < \infty$ for $x \leq h$, for some $C, h > 0$, which is continuous at zero and positive a neighborhood of zero.*

2. *$Y_i(x)$ is continuously differentiable in x with probability one, $\mathbb{E}[Y_i(x)|X_i^* = x', Z_i = z]$ is continuously differentiable in x and x', and $\mathbb{E}[|Y_i'(x)|X_i^* = x', Z_i = z] < \infty$ for all $x \in [0, h)$ and $x' \in (-\infty, h)$, for some $h > 0$.*

3. *$\mathbb{E}[Y_i(0)|X_i^* = x, Z_i = z]$ is monotonic for $x \leq 0$, and $\mathbb{P}\big(\frac{d}{dx}\mathbb{E}[Y_i(0)|X_i^* = x, Z_i = z] = 0\big) \in \{0, 1\}$.*

4. *If the probability in the previous item is zero, then additionally either (a) i: $Y_i(0)|X_i^* = x, Z_i = z \to_d Y_i(0)|X_i^* = 0, Z_i = z$ as $x \downarrow 0$; ii: $Y_i(0) - \mathbb{E}[Y_i(0)|X_i^*, Z_i = z] \perp\!\!\!\perp X_i^*|X_i = 0, Z_i = z$; and iii: $f_{\mathbb{E}[Y(0)|X^*,Z=z]|X=0}(y) \leq C < \infty$, for all y and some $C > 0$; or the weaker condition (b) suppose that $Y_i(0)|X_i = 0, Z_i = z$ admits a density $f_{Y(0)|X=0,Z=z}$, and that $F_{Y(0)-\mathbb{E}[Y(0)|X^*,Z=z]|X=x,Z=z}$ converges pointwise as $x \downarrow 0$. Denote the limit distribution as $F_{Y(0)-\mathbb{E}[Y(0)|X^*,Z=z]|X=0^+,Z=z}$. Define W_i, a random variable derived from the deconvolution of $dF_{Y(0)-\mathbb{E}[Y(0)|X^*,Z=z]|X=0^+,Z=z}$ from $f_{Y(0)|X=0,Z=z}$. Then $f_{W|X=0,Z=z}(\mathbb{E}[Y_i|X_i = 0^+, Z_i = z]) = f_{\mathbb{E}[Y(0)|X^*,Z=z]|X=0,Z=z}(\mathbb{E}[Y_i|X_i = 0^+, Z_i = z])$.*

Assumption 2 is weaker than Assumption 1 in two ways. First, the endogeneity at $X_i = 0$ is now indexed by (X_i^*, Z_i), which is a multidimensional vector. This means that all continuity requirements and the local independence assumption are weaker. Second, $\mathbb{E}[Y_i(0)|X = x, Z = z]$ may be increasing in x for some z, and decreasing for others.

Theorem 3.4. *If Assumption 2 holds, then $\beta(0, z)$ is identifiable.*

Proof. All the identification arguments in the previous section may be repeated conditionally, and so the following equations hold:

$$\lim_{x \downarrow 0} \mathbb{E}[Y_i|X_i = x, Z_i = z] = \beta(0, z) + u'(0, z)$$

and

$$u'(0, z) = \text{sgn}(\Delta(z)) \cdot \frac{f_{X|Z=z}(0)_+}{F_{X|Z=z}(0) \cdot \frac{1}{2\pi} \int \frac{\mathbb{E}[e^{\mathbf{i}\xi Y_i}|X_i=0, Z_i=z]}{\mathbb{E}[e^{\mathbf{i}\xi(Y_i - \mathbb{E}[Y_i|X_i=0^+, Z_i=z])}|X_i=0^+, Z_i=z]} e^{-\mathbf{i}\xi \mathbb{E}[Y_i|X_i=0^+, Z_i=z]} d\xi}.$$

□

The identification of $\beta(0)$ must take some issues of aggregation into account, since conditional on Z_i there may be some instances where there is no bunching.

Theorem 3.5. *Let $\mathcal{Z}$ be the set of values of the support of the distribution of Z_i for which $\mathbb{P}(X_i = 0|Z_i) > 0$. Then, if Assumption 2 holds on $\mathcal{Z}$ with probability one, then $\beta(0)$ is identifiable.*

Proof. We must first point out that the values of Z_i for which there is no bunching do not matter in the identification of the marginal treatment effect at the bunching point. Therefore, although our method cannot be applied for those Z_i, the limitation is irrelevant.

$$\beta(0) = \mathbb{E}[\beta(0, Z_i)|X_i = 0] = \mathbb{E}[\beta(0, Z_i)|X_i = 0, Z_i \in \mathcal{Z}]\mathbb{P}(Z_i \in \mathcal{Z}|X_i = 0).$$

For almost all $Z_i \in \mathcal{Z}$, by Theorem 3.4, $\beta(0, Z_i)$ is identified. Since $\mathbb{P}(Z_i \in \mathcal{Z}|X_i = 0)$ is also identified, it follows that so is $\beta(0)$.

□

3.2 Identification of average marginal treatment effects beyond the bunching point

From equation (4), the identification of treatment effects $\beta(x)$ for $x > 0$ depends on the identification of $u'(x)$, which is nonparametrically identified in our setting only at $x = 0$. However, if it is possible to assume that u satisfies some parametric structures, then it may be possible to identify all the marginal treatment effects. It is interesting that moving from local to global identification of all marginal treatment effects requires only assumptions on u without the need for additional requirements on either $B(x)$ (beyond differentiability) or ϵ_i.

Assumption 3. *$u'(x) = g(x;\theta)$, where g is known, but the scalar θ is not. Suppose also that $g(0;\theta)$ is invertible in θ.*

Theorem 3.6. *If Assumptions 1 and 3 hold, then $\beta(x)$ is identified for all x.*

Proof. Assumption 3 implies that

$$\theta = g^{-1}(0; u'(0)).$$

From Assumption 1, $u'(0)$ is identifiable, and therefore so is θ. Then, from equation (4),

$$\beta(x) = \frac{d}{dx}\mathbb{E}[Y_i|X_i = x] - g(x;\theta)$$

is identifiable. □

This type of result is particularly useful in models with controls because it makes it possible to extend the typical exogeneity conditions seen in applied work to allow for a semiparametric endogeneity structure while still achieving nonparametric identification of the treatment effects. The following example illustrates this point.

Example 3.1. *(Linear endogeneity) Suppose that*

$$u(x,z) = \delta_0(z) + \delta(z)x,$$

then if Assumption 2 holds for $Z_i = z$, $u'(0,z) = \delta(z)$, and thus the treatment effect $\beta(x,z) = \frac{d}{dx}\mathbb{E}[Y_i|X_i = x, Z_i = z] - \delta(z)$ is identifiable for all x.

Since the analysis above requires no additional assumptions on $B(x)$ or ϵ_i beyond differentiability of $B(x)$ and what is implied by Assumption 2, this is far more general than most models implemented in applied work. Consider, for example, a more restrictive model closer to what is often used in applications,

$$Y_i = \beta_i X_i + Z_i'\gamma_i + \delta_i\eta_i + \varepsilon_i,$$

where $\beta_i, \gamma_i, \delta_i, \varepsilon_i \perp\!\!\!\perp X_i|Z_i$. We are interested in the average treatment effect $\beta = \mathbb{E}[\beta_i]$. This model allows for endogeneity from a possible correlation between X_i and the unobservable η_i. This is one of the models considered in Caetano et al. (2020), Caetano et al. (2021c) and Caetano et al. (2021d), for example, where it is assumed that $X_i^* = \pi_i Z_i + \eta_i$, and $\pi_i \perp\!\!\!\perp X_i|Z_i$. In those papers, identification is obtained from semiparametric or shape restrictions on the distribution of $\eta_i|Z_i$.

However, identification may be obtained without the need to specify any distribution by observing that, in this model, $B(x,z) = \beta(z)x$, and $u(x,z) = z'\alpha(z) + \delta(z)x$, where $\beta(z) = \mathbb{E}[\beta_i|Z_i = z]$, $\alpha(z) = \mathbb{E}[\gamma_i - \pi\delta_i + \epsilon_i|Z_i = z]$, and $\delta(z) = \mathbb{E}[\delta_i|Z_i = z]$. Then, if Assumptions 2(1) and 2(5) hold (since all the other assumptions are satisfied automatically given the model), $\delta(z) = u'(0,z)$ is identified, and therefore so is $\beta(z)$. In fact, we do not need derivatives to identify the treatment effects, only differencing:

$$\beta(z) = \mathbb{E}[Y_i|X_i = x+1, Z_i = z] - \mathbb{E}[Y_i|X_i = x, Z_i = z] - \delta(z).$$

Then, if Assumption 2 holds almost surely for all Z_i, the average treatment effects are identifiable:

$$\beta = \mathbb{E}[Y_i|X_i = x+1] - \mathbb{E}[Y_i|X_i = x] - \mathbb{E}[\delta(Z_i)].$$

4 Estimation

For a sample $\{(Y_i, X_i)', i = 1, \ldots, n\}$, the average marginal treatment effect at the bunching point may be estimated with the following formula

$$\hat{\beta}(0) = d_X\hat{\mathbb{E}}[Y_i|X_i = 0^+] - \text{sgn}(\hat{\mathbb{E}}[Y_i|X_i = 0^+] - \hat{\mathbb{E}}[Y_i|X_i = 0]) \cdot \frac{\hat{f}_X(0)_+}{\hat{F}_X(0) \cdot \hat{f}_{u(X^*)|X=0}(u(0))},$$

where $d_X\hat{\mathbb{E}}[Y_i|X_i = 0^+]$ is an estimator of $\lim_{x\downarrow 0} \frac{d}{dx}\mathbb{E}[Y_i|X_i = x]$, and

$$\hat{f}_{u(X^*)|X=0}(u(0)) = \frac{1}{2\pi}\int \frac{\hat{\mathbb{E}}[e^{\mathbf{i}\xi Y_i}|X_i = 0]}{\hat{\mathbb{E}}[e^{\mathbf{i}\xi(Y_i - \hat{\mathbb{E}}[Y_i|X_i=0^+])}|X_i = 0^+]} e^{-\mathbf{i}\xi\hat{\mathbb{E}}[Y_i|X_i=0^+]} d\xi.$$

All the components of this formula are standard objects frequently studied in econometrics. The following list contains instructions for estimating all components.

1. The terms $\hat{F}_X(0)$ and $\hat{\mathbb{E}}[Y_i|X_i = 0]$ may be estimated with simple averages:

   $$\hat{F}_X(0) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i = 0),$$

   and

   $$\hat{E}[Y_i|X_i = 0] = \hat{F}_X(0)^{-1} \cdot \frac{1}{n}\sum_{i=1}^{n} Y_i\mathbf{1}(X_i = 0).$$

2. The terms $\hat{\mathbb{E}}[Y_i|X_i = 0^+]$ and $d_X\hat{\mathbb{E}}[Y_i|X_i = 0^+]$ are standard non-parametric regression boundary quantities. Estimation of these objects has been extensively researched in the statistics literature on local polynomial estimators, and in the Regression Discontinuity Design and Regression Kink Design literatures in economics. In line with classical methods in this literature and with the vast majority of applications in boundary regression estimation, we propose using a local linear regression of Y_i onto X_i at $X_i = 0$, using only observations such that $X_i > 0$, for its superior properties of bias reduction and variance control at the boundary over other methods.[8] The intercept coefficient of this regression is $\hat{\mathbb{E}}[Y_i|X_i = 0^+]$, and the slope coefficient is $d_X\hat{\mathbb{E}}[Y_i|X_i = 0^+]$. This may be executed using any package for local linear regression available in standard statistical packages (R, STATA, etc.).

[8] See Ruppert and Wand (1994) and Fan and Gijbels (2018), and also Cheruiyot (2020). See Imbens and Wager (2019) and citations therein for recent proposals which may be superior to local linear estimators.

Explicitly, for a bandwidth $h_1 > 0$ and a kernel function k_1,[9] solve the problem

$$\hat{b}_0, \hat{b}_1 = arg \min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 k_1(X_i/h_1) \mathbf{1}(X_i > 0), \tag{13}$$

then $\hat{\mathbb{E}}[Y_i|X_i = 0^+] = \hat{b}_0$, and $d_X \hat{\mathbb{E}}[Y_i|X_i = 0^+] = \hat{b}_1$. This estimator has a closed form expression, which is commonly found on nonparametric econometrics textbooks, e.g. Li and Racine (2007).

3. The term $\hat{f}_X(0)_+$ is a boundary density. As in the case of nonparametric boundary regression discussed in the previous item, the tendency for higher bias in this scenario necessitates the use of corrective methods, such as the use of local polynomial estimators. We recommend the approach recently proposed in Pinkse and Schurter (2021),[10] which has two important properties which are of great value in our case and which are not found in other estimators currently available. First, this estimator achieves the same rates of bias convergence at the boundary that is normally achieved in interior points. Second, the density estimator is never negative, a situation which would be complicated to address in our case. Additionally, the estimators have simple closed-form expressions, requiring only the choice of a bandwidth tuning parameter, h_2.

   Following Pinkse and Schurter (2021), let $L_X(x) = \log f_X(x)$. We begin by estimating $L'_X(0)$ as[11]

$$\hat{L}'_X(0) = -\frac{\sum_{i=1}^{n} (1 - 2X_i/h_2) \mathbf{1}(0 \leq X_i \leq h_2)}{\sum_{i=1}^{n} X_i (1 - X_i/h_2) \mathbf{1}(0 \leq X_i \leq h_2)}.$$

   This, then, allows us to estimate the density at the boundary as

$$\hat{f}_X(0)_+ = \frac{\frac{1}{nh_2} \sum_{i=1}^{n} k_2 (X_i/h_2)}{\int_0^1 k_2(\nu) \exp(\hat{L}'_X(0) \nu h_2) d\nu},$$

   which can be calculated for many standard positive kernel functions k_2. For example, as in Example 5 of Pinkse and Schurter (2021), when k_2 is the Epanechnikov kernel $k_2(\nu) = 3/4(1 - \nu^2)$ (which is the choice recommended by that paper) the denominator is equal to

$$\frac{3}{2} \cdot \frac{2 + \hat{L}'_X(0)^2 h_2^2 - \exp(\hat{L}'_X(0) h_2)(2 - 2\hat{L}'_X(0) h_2)}{\hat{L}'_X(0)^3 h_2^3}.$$

   This estimator is available in packaged form in standard statistics software and can be implemented by simply restricting the sample to observations such that $X_i > 0$ and then using the package to estimate the density of X_i at $X_i = 0$.

4. The final term $\hat{f}_{u(X^*)|X=0}(u(0))$ is a standard deconvolution estimator. We follow the esti-

[9] The triangular kernel $k_1(\nu) = (1 - |\nu|)$, where $|\nu| \leq 1$ is recommended for boundary regressions such as this (Cheng et al. 1997).

[10] Other estimators of boundary densities include Hjort and Jones (1996), Loader (1996), Cheng et al. (1997), Zhang and Karunamuni (1998), Bouezmarni and Rombouts (2010) and Cattaneo et al. (2020).

[11] This estimator is derived from applying Example 1 with $z = 0$ to equation (2) in Pinkse and Schurter (2021).

mator described in Schennach (2021), which is the focus of an extensive literature, although there are many alternative proposals which are also referenced therein.

We first write $\hat{\mathbb{E}}[e^{\mathbf{i}\xi(Y_i-\hat{\mathbb{E}}[Y_i|X_i=0^+])}|X_i=0^+]$ as a local linear regression of $e^{\mathbf{i}\xi(Y_i-\hat{\mathbb{E}}[Y_i|X_i=0^+])}$ onto X_i at $X_i=0$ using only observations such that $X_i>0$. To do this, for a matrix $\mathbf{x}$ with rows $(1,X_i)'$ and a diagonal matrix $\mathbf{k}$, with diagonal elements $k_3(X_i/h_3)\mathbf{1}(X_i>0)$, where k_3 is the triangular kernel, and $\mathbf{e}_1=(1,0)'$, define the vector $A(\xi)=(e^{\mathbf{i}\xi(Y_1-\hat{\mathbb{E}}[Y_i|X=0^+]},\dots$ $\dots,e^{\mathbf{i}\xi(Y_n-\hat{\mathbb{E}}[Y_i|X=0^+]})'$, and program the function

$$\hat{\phi}(\xi)=\mathbf{e}_1(\mathbf{x}'\mathbf{k}\mathbf{x})^{-1}\mathbf{x}'\mathbf{k}A(\xi).$$

This is then imputed into a standard convolution estimator, such as for example:

$$\hat{f}_{u(X^*)|X=0}(u(0))=\frac{1}{nh_4}\sum_{i=1}^{n}g(Y_i)\mathbf{1}(X_i=0),$$

with

$$g(Y_i)=\frac{1}{\hat{F}_X(0)\cdot 2\pi}\int e^{\mathbf{i}\xi(Y_i-\hat{\mathbb{E}}[Y_i|X_i=0^+])}\frac{\phi_K(h_4\xi)}{A(\xi)}d\xi,$$

where $\phi_{k_4}(h_4\xi)=\int k_4(\nu)e^{\mathbf{i}h_4\xi\nu}d\nu$ is the Fourier transform of the kernel k_4 evaluated at $h_4\xi$.

The nonparametric estimators just described require the choice of the bandwidth tuning parameters, h_1,h_2,h_3 and h_4, which modulate the bias-variance trade-off. This choice is rather important, and the subject of a great deal of interest in the nonparametrics estimation literature. At this stage, our recommendation is that if an optimal method for bandwidth selection exists for the specific estimator used at a given step, then it should be used.[12] However, it is possible that the optimal bandwidths for $\hat{\beta}(0)$ are not the optimal bandwidths for each of the separate components.

Additionally, there is an interest in the use of bias correction techniques for inference in the Regression Discontinuity Design literature which may have relevance in this context as well (e.g. Calonico et al. (2014), Noack and Rothe (2019), He and Bartalotti (2020), Armstrong and Kolesár (2020) and citations therein). This is because, if optimal bandwidths are used, $\hat{\beta}(0)$ will be asymptotically biased. We leave these interesting questions for future research.

[12] For the selection of h_1 and h_3, Ruppert et al. (1995) propose an optimal bandwidth estimator for the local linear regression, and this or similar approaches for bandwidth selection are usually offered in standard local linear regression packages. There are many proposals for improvement on bandwidth selection in the Regression Discontinuity Design literature which may be adapted to this context, see, e.g. Imbens and Kalyanaraman (2012), Arai and Ichimura (2016), Arai and Ichimura (2018) and Calonico et al. (2020). For h_2, the optimal bandwidth is $h=(72/(nf_X(0)^+\beta_2^2))^{1/5}$, which may be calculated following Example 6 in Pinkse and Schurter (2021). β_2 can be estimated using a pilot estimate of $\hat{f}_X(0)_+$ which is estimated with a large bandwidth (over-smoothing), and both these terms are then added to the formula of the optimal bandwidth. Nevertheless, although theoretically sound, this method is not yet studied, and thus in practice we recommend testing several bandwidths around this benchmark and looking for robustness of the results. For h_4, consider the several approaches studied in Delaigle and Gijbels (2004).

4.1 Estimation with discrete controls

Estimation with controls depends on the nature of Z_i. If Z_i has a finite support, i.e. $Z_i \in \{z_1, \ldots, z_L\}$ with $\mathbb{P}(Z_i = z_l) > 0$, for all $l = 1, \ldots, L$, then the exact procedures described for the unconditional case may be performed separately for each z_l. That is, for all z_l, calculate

$$\hat{p}_{l,0} = \hat{\mathbb{P}}(Z_i = z_l | X_i = 0) = \hat{F}_X(0)^{-1} \cdot \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i = 0, Z_i = z_l), \text{ for } l = 1, \ldots, L. \tag{14}$$

Then, for all z_l such that $\hat{p}_{l,0} > 0$, restrict the sample to observations such that $Z_i = z_l$, and estimate $\hat{\beta}(0, z_l)$ just as described in the unconditional case using the new, restricted, data.

The average marginal treatment effect estimator in this case is

$$\hat{\beta}(0) = \sum_{l=1}^{L} \hat{\beta}(0, z_l) \hat{p}_{l,0}. \tag{15}$$

Note that it is not possible to estimate $\hat{\beta}(0, z_l)$ when $\hat{p}_{l,0} = 0$, but it is also not necessary to do so, since those treatment effects have weight equal to zero in the estimator formula.

4.2 Estimation with continuous controls

When Z_i is continuously distributed, one may apply a smoothing technique to the estimators described above, so as to use information coming from values of the control around Z_i to perform the estimation. A simple strategy to estimate $\hat{\beta}(0, Z_i)$ is as follows: let $Z_i = (Z_{1i}, \ldots, Z_{Mi})'$, and for bandwidths $\kappa_1, \ldots, \kappa_M$, and kernel functions $K_1, \ldots, K_M$, restrict the sample to observations such that $-\kappa_1 < Z_{1j} < \kappa_1, \ldots, -\kappa_M < Z_{Mj} < \kappa_M$. Index the resulting dataset by t, suppose it has n_T observations, and define

$$K_\kappa(Z_t - Z_i) := \frac{1}{\kappa_1 \cdots \kappa_M} K_1\left(\frac{Z_{1t} - Z_{1i}}{\kappa_1}\right) \cdots K_M\left(\frac{Z_{Mt} - Z_{Mi}}{\kappa_M}\right).$$

Then, for each value Z_i such that the restricted sample has bunching, i.e.

$$\hat{p}_{i,0} = \frac{1}{n_T} \sum_{t=1}^{n_T} \mathbf{1}(X_t = 0) > 0,$$

perform the methods described for unconditional estimation, only weighting each observation by $k_\kappa(Z_t - Z_i) = K_\kappa(Z_t - Z_i) / \sum_{t=1}^{T} K_\kappa(Z_t - Z_i)$.[13]

[13] Thus, $\hat{F}_{X|Z=Z_i}(0) = \frac{1}{n_T} \sum_{t=1}^{n_T} \mathbf{1}(X_i = 0) k_\kappa(Z_t - Z_i)$, and $\hat{E}[Y_i | X_i = 0, Z_i] = \hat{F}_{X|Z=Z_i}(0)^{-1} \cdot \frac{1}{n_T} \sum_{t=1}^{n_T} Y_i \mathbf{1}(X_i = 0) k_\kappa(Z_t - Z_i)$, $\hat{\mathbb{E}}[Y_i | X_i = 0^+, Z_i]$. The densities $\hat{f}_{X|Z=Z_i}(0)_+$ and $\hat{f}_{Y|X=0,Z_i}(Y_i)$ are implemented in the same way using the restricted sample, substituting i by t and n by n_T in the formulas, and multiplying terms inside sums indexed by t $k_\kappa(Z_t - Z_i)$. Finally, $\hat{\mathbb{E}}[Y_i | X_i = 0^+, Z_i]$ and $d_X \hat{\mathbb{E}}[Y_i | X_i = 0^+, Z_i]$ are respectively the intercept and slope coefficients of a local linear regression of Y_t onto X_t at zero, using only observations such that $X_t > 0$ and weights $k_\kappa(Z_t - Z_i)$, and $\hat{\mathbb{E}}[\hat{f}_{Y|X=0,Z=Z_i}(Y_i) | X_i = 0^+, Z_i]$ is the intercept of the same procedure, only with $\hat{f}_{Y|X=0,Z=Z_i}(Y_i)$ instead of Y_i.

Then,

$$\hat{\beta}(0) = \sum_{i=1}^{n} \hat{\beta}(0, Z_i)\hat{p}_{i,0}.$$

As in the previous section, it is not necessary to estimate $\hat{\beta}(0, Z_i)$ when $\hat{p}_{i,0} = 0$.

4.3 Estimation with mixed or large dimensional controls

In practice, most control lists include a mixture of discrete and continuous variables, and may include a large number of terms. In such cases, smoothing is either impractical or impossible. We have had success with a discretization technique which implements clustering methods, which are popular in machine learning and have been recently adopted in economics.[14]

Let $\{\hat{\mathcal{C}}_1, \ldots, \hat{\mathcal{C}}_C\}$ be a finite partition of the observations into groups, which we call clusters, and let $\hat{C}_i = (\mathbf{1}(Z_i \in \hat{\mathcal{C}}_1), \ldots, \mathbf{1}(Z_i \in \hat{\mathcal{C}}_C))'$ be the cluster indicators. We propose substituting Z_i with $\hat{C}_i$, which has finite support. This then transforms the estimation procedure into a discrete controls case, which can be implemented exactly as described in Section 4.1.

Explicitly, for each cluster $\mathcal{C}_c$, calculate

$$\hat{p}_{c,0} = \hat{F}_X(0)^{-1} \cdot \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i = 0, Z_i \in \mathcal{C}_c), \text{ for } c = 1, \ldots, C. \tag{16}$$

Then, for those clusters with $\hat{p}_{c,0} > 0$, estimate $\hat{\beta}(0, \hat{\mathcal{C}}_c)$ separately using a new dataset composed only of observations within cluster $\mathcal{C}_c$ (i.e. i such that $Z_i \in \mathcal{C}_c$). For this, follow the exact procedures described in the unconditional case. The average marginal treatment effect estimator is, then,

$$\hat{\beta}(0) = \sum_{c=1}^{C} \hat{\beta}(0, \mathcal{C}_c)\hat{p}_{c,0}. \tag{17}$$

As in the previous sections, it is not necessary to estimate $\hat{\beta}(0, \mathcal{C}_c)$ when $\hat{p}_{c,0} = 0$.

In general, if $\beta(0, z)$ is continuous in z, the ability of this estimator to approximate $\beta(0)$ depends on how much information about Z_i is given by the cluster indicator vector $\hat{C}_C$. Thus, it is desirable to choose a clustering method that minimizes the within-cluster variation in the values of Z_i. All unsupervised clustering methods in the statistical learning literature could in principle be used (e.g. k-means, k-medoids, self-organizing maps, and spectral – see Hastie et al. (2009)). In our application, we show results using hierarchical clustering for its well-known stability.[15]

The clustering strategy requires the choice of the number of clusters, which modulates the bias-variance trade-off in the estimation of $\beta(0, z)$. The more clusters are used, the more similar are the Z_i within each cluster, and thus the smaller the bias and the larger the variance. Although there

[14] See, e.g. Bonhomme and Manresa (2015); Bonhomme et al. (2017); Cheng et al. (2019); Cytrynbaum (2020); Caetano et al. (2020, 2021c,d).

[15] Hierarchical clustering requires the choice of a linkage method and a dissimilarity measure. The results we report in our application use the Ward's linkage and the Gower measure, which are recommended for mixed continuous and discrete controls.

must exist an optimal number of clusters, there are as yet no established methods to aid with this decision.

Nevertheless, note that we are not directly interested in $\hat{\beta}(0,z)$ but rather in $\hat{\beta}(0)$, which aggregates the information over all clusters. The trade-off is, in theory, much less important for $\hat{\beta}(0)$, and thus one should err on the side of having a larger number of clusters, with an eye for instability which could be created by pathological clusters (e.g. clusters with bunching but with too few observations near the bunching point, or clusters where every observation is bunched).

5 Application: the effect of watching TV on skills

In this section, we apply our method to estimate the marginal effect of time spent watching TV on children's skills, for those children currently choosing zero TV time.[16] We use the 1997, 2002 and 2007 Waves of the Child Development Supplement from the Panel Study of Income Dynamics (CDS-PSID).[17]

Table 1 presents summary statistics for our sample broken down also by grade range in order to provide more context. Our outcome variable Y_i is either cognitive or non-cognitive skill, and our treatment variable X_i represents how many hours per week the child watches TV.[18] For context, we also show summary statistics for an exhaustive list of other activities that make up the day. We additionally show an extensive list of covariates that a researcher may wish to control for, in order to weaken the identifying assumptions, as discussed in Section 3.1.

We have a pooled sample of 4,389 children ranging from 5 to 18 years of age, with an average age of just over 11. While about half of these children are attending elementary school, the other half is evenly split between middle and high school. The amount of time spent watching TV is on average around 2 hours per day, and this average is fairly constant across grade ranges.

Importantly, about 5% of the sample is bunched at zero hours watching TV, and this proportion is a bit higher in high school than in earlier grades. The bunching is evident in Figure 3, which shows the unconditional c.d.f (left panel) and the empirical distribution (right panel) of X_i.

The figure also shows that a very small proportion of children spend a marginal amount of TV time above zero, suggesting that something special might be happening at zero TV time. Indeed, Appendix A presents a model of constrained optimization that helps explain why bunching occurs in this context. In that model, bunching may happen because some children are at a corner solution: they would like to choose a quantity of TV time below zero, but they are constrained to choose only non-negative amounts. In this scenario, the group at zero TV time is particularly heterogeneous,

[16] See for instance Zavodny (2006), Gentzkow and Shapiro (2008) and Munasib and Bhattacharya (2010) for recent empirical papers in this literature. These papers are well aware that watching television may be endogenous. Zavodny (2006) tackles endogeneity using fixed effects, Munasib and Bhattacharya (2010) uses IV, while Gentzkow and Shapiro (2008) uses the timing of the roll-out of children's programming to different local markets to obtain causal estimates.

[17] This application uses the same sample as Caetano et al. (2021c), which estimates the effect of enrichment activities on skills. See that paper for further details about how the sample was created.

[18] For concreteness, the summary statistics table shows X as well as the other time activities in terms of hours per day, instead of hours per week.

Table 1: Summary Statistics

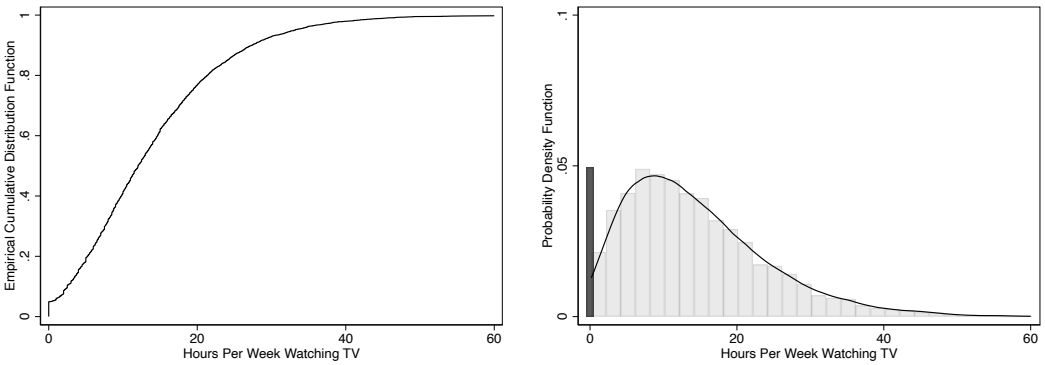
Outcome Variables	All	Grades K-5	Grades 6-8	Grades 9-12
Cognitive Skill	0.07 (0.93)	-0.35 (0.93)	0.30 (0.76)	0.56 (0.72)
Non-cognitive Skill	0.01 (0.95)	-0.02 (0.95)	-0.03 (0.98)	0.11 (0.92)
Activities (hours per day)				
Watch TV	1.99 (1.46)	1.92 (1.30)	2.12 (1.51)	1.97 (1.67)
Sleep	9.65 (1.34)	10.09 (1.13)	9.47 (1.29)	9.07 (1.48)
Class	4.40 (1.59)	4.29 (1.61)	4.64 (1.39)	4.35 (1.70)
Homework	0.55 (0.73)	0.42 (0.51)	0.60 (0.69)	0.75 (1.00)
Extra-Curricular Activities	0.37 (0.68)	0.30 (0.51)	0.40 (0.71)	0.47 (0.89)
Active Leisure	2.23 (1.60)	2.62 (1.63)	1.89 (1.42)	1.89 (1.56)
Other Passive Leisure	0.73 (1.10)	0.43 (0.71)	0.89 (1.12)	1.10 (1.46)
Duties/Chores	3.80 (1.47)	3.64 (1.16)	3.70 (1.33)	4.19 (1.94)
Other	0.26 (0.99)	0.29 (0.97)	0.27 (1.07)	0.22 (0.93)
Covariates				
Watch TV = 0	0.05 (0.22)	0.03 (0.17)	0.04 (0.20)	0.09 (0.28)
Child is Male	0.51 (0.50)	0.53 (0.50)	0.46 (0.50)	0.50 (0.50)
Child is White	0.48 (0.50)	0.49 (0.50)	0.47 (0.50)	0.47 (0.50)
Child is Black	0.40 (0.49)	0.39 (0.49)	0.41 (0.49)	0.41 (0.49)
Child is Hispanic	0.08 (0.26)	0.08 (0.27)	0.07 (0.26)	0.08 (0.27)
Child is Another Race	0.05 (0.21)	0.05 (0.21)	0.05 (0.21)	0.04 (0.20)
Child is in Grade PreK-5	0.47 (0.50)	1.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Child is in Grade 6-8	0.27 (0.44)	0.00 (0.00)	1.00 (0.00)	0.00 (0.00)
Child is in Grade 9-12	0.27 (0.44)	0.00 (0.00)	0.00 (0.00)	1.00 (0.00)
1997 Wave	0.30 (0.46)	0.52 (0.50)	0.21 (0.41)	0.00 (0.00)
2002 Wave	0.44 (0.50)	0.44 (0.50)	0.37 (0.48)	0.49 (0.50)
2007 Wave	0.27 (0.44)	0.05 (0.21)	0.42 (0.49)	0.51 (0.50)
Child's Father is Alive	0.97 (0.16)	0.98 (0.14)	0.97 (0.17)	0.96 (0.18)
Child's Mother is Alive	0.99 (0.07)	1.00 (0.06)	0.99 (0.08)	0.99 (0.09)
Child is Home Schooled	0.01 (0.11)	0.01 (0.11)	0.01 (0.11)	0.01 (0.10)
Child is in Private School	0.08 (0.27)	0.08 (0.27)	0.08 (0.27)	0.07 (0.25)
Household Income (in $1,000s)	73.62 (82.21)	66.63 (67.65)	72.79 (70.87)	86.77 (109.95)
Age (years)	11.29 (3.64)	8.10 (2.13)	12.38 (0.95)	15.82 (1.19)
Observations	4,396	2,060	1,167	1,169

Activity categories are exhaustive. The 1997, 2002 and 2007 CDS Waves are pooled.

since they all choose the same amount of TV time (i.e., $X_i = 0$), but the constraint may be binding to different degrees for different children, in the sense that their unconstrained choices would have differed (i.e. X_i^* differs across observations such that $X_i = 0$). Intuitively, different children at $X_i = 0$ would have wanted to "borrow against TV time" in different amounts in order to increase their time spent on other activities. Thus, in the notation of Section 3, our identification strategy relies on variation around $X_i = 0$, since under Assumption 1 we can isolate the variation in Y_i that is entirely due to X_i^* and not due to X_i or ϵ_i.

To see how Assumption 1 is plausible in the context of this application, note that Assumption 1(1) and (3) are reasonable if one interprets X as a choice and X^* as the desired choice, which relates

Figure 3: Unconditional Distribution of X_i



Note: The left panel shows the cumulative distribution function of X_i. The right panel shows the kernel density estimate along with the histogram for $X_i > 0$ (bandwidth equals to 2). The darker bar is the proportion of observations with $X_i = 0$.

to the preference for TV time relative to other activities (see Appendix A for further details). See also the left panel from Figure 3 as direct evidence that $X^* = X$ is continuous for $X > 0$. Assumption 1(2) also makes sense, as a child spending one minute a week watching TV is likely to have a similar skill than if they had chosen 0 minutes instead. The other requirements from Assumption 1 are standard and technical.

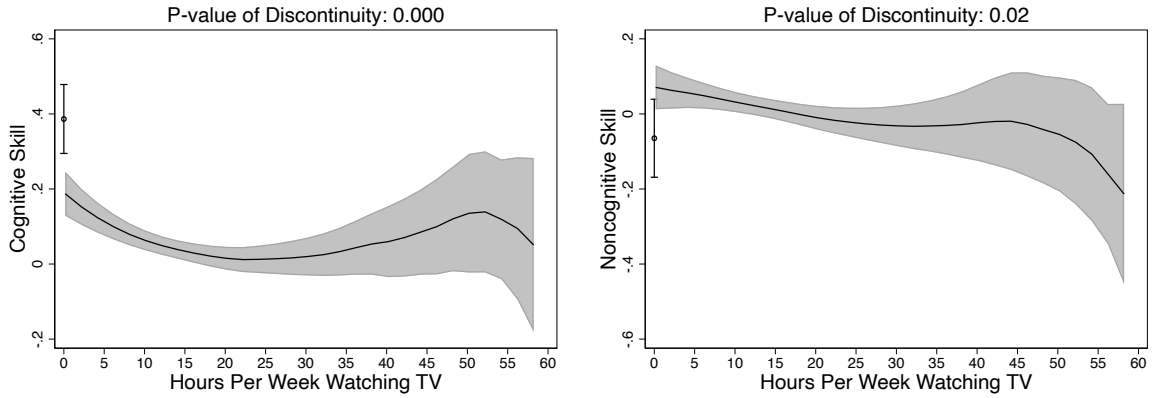
Next, we estimate $\Delta = \mathbb{E}[Y_i|X_i = 0^+] - \mathbb{E}[Y_i|X_i = 0]$ for each outcome variable. Specifically, we run a local linear regression of Y_i on X_i using only observations such that $X_i > 0$, and compare its prediction at $X_i = 0$ with the average of Y_i at $X_i = 0$. For context, each panel of Figure 4 shows the fit of the local linear polynomial for all values of X_i and the average of Y_i at $X_i = 0$, along with their corresponding 90% confidence intervals. In the header of each panel we also show the p-value of the test for whether there is a discontinuity at $X_i = 0$. The discontinuity is clearly negative for cognitive skills and positive for non-cognitive skills. Although this evidence is not local (i.e., $X_i = 0$ corresponds to $\infty < X_i^* \leq 0$), Lemma 3 implies that the sign of the bias of endogeneity ($\text{sgn}(u'(0))$ is the same as the sign of the discontinuity Δ.

Finally, we show the main results in Table 2. For simplicity in the exposition, we present the results for $h_3 = 0.5$ (the bandwidth used for Y_i, measured in standard deviations) and use the same bandwidth for X_i in all the other methods, $h := h_1 = h_2 = h_4$ (measured in hours per week) of varying size.[19] We have considered different bandwidths near the ones shown and the findings are similar. We also chose the kernels k_1, k_2 and k_3 following the discussion in Section 4.[20] The results show that a marginal increase in TV time from zero to one hour per week would increase the children's cognitive skills, while at the same time reducing their non-cognitive skills. The marginal endogenous variation is of similar magnitude as the effect, but in opposite direction, as expected given the discontinuities shown in Figure 4.

[19]We use a local linear regression in steps 2 and 4, and the method from Pinkse and Schurter (2021) in steps 3 and 4. See Section 4 for a detailed description of each step.

[20]Specifically, k_1 is the triangular kernel and k_2 and k_3 are the Epanechnikov kernel.

Figure 4: Discontinuity in Y_i at $X_i = 0$ due to Confounders



Note: Each panel shows a plot of the local linear estimator (bandwidth equals to 10) of each outcome variable Y_i onto X_i, estimated using only observations with $X_i > 0$, where X_i represents the hours spent watching TV in a typical week. We also present the average Y_i for $X_i = 0$ and show 90% confidence intervals everywhere. The discontinuity at $X_i = 0$ (along with the p-value on top) shows evidence of the sign of $u'(0)$: negative for cognitive skills, and positive for non-cognitive skills.

Table 2: Main Results

	Bandwidth (in hours per week)		
Cognitive Skill	$h = 5$	$h = 7$	$h = 10$
$\beta(0)$	0.178	0.259	0.338
	(0.122)	(0.113)	(0.106)
$u'(0)$	-0.232	-0.295	-0.357
	(0.118)	(0.105)	(0.097)
Non-Cognitive Skill	$h = 5$	$h = 7$	$h = 10$
$\beta(0)$	-0.215	-0.277	-0.331
	(0.103)	(0.098)	(0.092)
$u'(0)$	0.212	0.269	0.326
	(0.099)	(0.087)	(0.085)

Note: X is measured as hours per week, and Y is measured in standard deviation units. Following the discussion in Section 4, k_1 is the triangular kernel and k_2 and k_3 are the Epanechnikov kernel. While $h_3 = 0.5$, $h = h_1 = h_2 = h_4$ is shown in each column of the table.

6 Concluding remarks

When the treatment variable has bunching at the extreme of its distribution, this paper presents a new design for identification of the average marginal treatment effects at the bunching point. This is the first identification approach leveraging bunching phenomena which does not make assumptions on functional forms nor on the distribution of the unobservables. Since the method does not rely on exclusion restrictions or special data structures, it provides a new avenue for identification of treatment effects when well established methods are not applicable.

The approach requires that the treatment be continuously distributed near the bunching point, and it relies on the continuity of the treatment effects at the bunching point, the continuity of the

function that indexes the endogeneity above and below the bunching point cutoff across the bunching point, and the continuity of the distribution of the idiosyncratic error at the bunching point, plus a local independence assumption between the idiosyncratic error and the indexing variable near the threshold. These conditions are reminiscent of the Regression Discontinuity Design conditions for continuous treatment if one understands the treatment extended into the negative side as the running variable. The method requires that the function that indexes the endogeneity among the bunched observations is monotonic, a condition which may be substituted by local monotonicity in a negative neighborhood of the bunching point if the sign of the endogeneity bias is known (that is, ≥ 0 or ≤ 0, so it is not necessary to know if it is strictly positive or negative) in some other way, e.g. through economic reasoning.

Identification is achieved by the comparison of the distribution of the treatment (observed on the positive side) and the distribution of the function that indexes the endogeneity (observed on the negative side, thanks to a deconvolution of the distribution of the outcome at the bunching point to eliminate the noise from the idiosyncratic error). The ratio of these is exactly the endogeneity bias.

The approach results in the identification of the average marginal treatment effect as a closed form expression of identifiable quantities which are fairly standard well known quantities in the econometrics literature, including the limits as the treatment approaches the bunching point of (1) the density of the treatment, (2) the outcome, (3) the derivative of the outcome, and (4) the deconvolution of the density of the outcome minus the limit of the expected outcome when the treatment approaches the bunching point from the distribution of the outcome when the treatment is at the bunching point. All these terms can be estimated with off-the-shelf methods readily available in packaged form for all standard statistics software.

We also apply the method to the estimation of the effect of time watching TV on children's cognitive and non-cognitive skill, revealing that while children who do not watch TV would gain cognitive skills from watching TV for some positive time, they would lose non-cognitive skills. Measured in standard deviations, the offsetting effect is roughly the same, so that all that is gained cognitively is lost non-cognitively, with perhaps more of a loss than a gain.

Appendix

A Treatment as an optimal constrained choice

Consider the following model, where person i chooses X_i and R_i to maximize their utility function:

$$\begin{aligned}(X_i, R_i) =& arg\max_{x,r} V(x, r; \rho_i) \\ s.t. \qquad & x + r = 24 \\ & x \geq 0 \\ & r \geq 0.\end{aligned} \tag{18}$$

Here we design the model so as to reflect our application, where x stands for TV time, r stands for remaining activities in the day, and time is measured continuously as hours per day. The parameter ρ_i represents the preference parameter. In this set up, we are modelling the remaining activity as a "numeraire good" to keep our focus on x, but a generalization of this model to several activities is straightforward.

Suppose V is a strictly concave differentiable function of (x, r), with partial derivatives $V_x(x, r, \rho_i) = \frac{\partial}{\partial x}V(x, r; \rho_i)$ and $V_r(x, r, \rho_i) = \frac{\partial}{\partial r}V(x, r; \rho_i)$ which are differentiable with respect to ρ. Moreover, suppose that $\frac{\partial}{\partial \rho}V_x(x, r; \rho) > 0$ and $\frac{\partial}{\partial \rho}V_r(x, r; \rho) < 0$, so that the parameter ρ regulates the relative preference for x over r. The relative preferences are drawn from a distribution F_ρ, so that each observation i has a preference ρ_i.

The person will choose the optimal X_i and $R_i = 24 - X$, where:

$$\begin{aligned}V_x(X_i, R_i; \rho_i) = V_r(X_i, R_i; \rho_i) \qquad & \text{if} \qquad X_i, R_i > 0 \\ V_x(X_i, R_i; \rho_i) \leq V_r(X_i, R_i; \rho_i) \qquad & \text{if} \qquad X_i = 0 \\ V_x(X_i, R_i; \rho_i) \geq V_r(X_i, R_i; \rho_i) \qquad & \text{if} \qquad X_i = 24\end{aligned} \tag{19}$$

Since we do not observe anybody choosing $X_i = 24$ in the data, for simplicity we consider only the first two cases. There will be a threshold $\bar{\rho}$ such that

$$\begin{aligned}X_i = 0, \quad & \text{if } \rho_i \leq \bar{\rho} \\ X_i > 0, \quad & \text{if } \rho_i > \bar{\rho}.\end{aligned}$$

Thus, $F_\rho(\bar{\rho})$ is the probability of bunching at $X_i = 0$. This probability might be larger than zero, as in our application, if there are individuals whose preference for TV relative to other activities is sufficiently low, $\rho_i \leq \bar{\rho}$.

Note that in this context X_i^* from Section 3 can be written as $X_i^* = b(\rho_i - \bar{\rho})$ for a continuous function b such that $b(0) = 0$. Indeed, there is a direct connection between ρ_i, the preference towards x relative to r, and X_i^*. In the context of a constrained choice model, X_i^* can be interpreted as the "desired" choice of x under no non-negativity constraint, while X_i can be interpreted as the actual

choice of x. For identification in this paper, we explore equation (3), which is restated here for convenience:

$$X_i = \max\{X_i^*, 0\}, \qquad 0 < \mathbb{P}(X_i^* < 0) < 1.$$

so that $X_i = X_i^*$ if $X_i^* \geq 0$ and $X_i = 0$ if $X_i^* \leq 0$, with some individuals with $X_i^* < 0$. In the context of the choice model from equations (18) and (19), this means that the non-negativity constraint is binding for some individuals, so that $\rho_i < \bar{\rho}$ and $V_x(X_i, R_i; \rho_i) < V_r(X_i, R_i; \rho_i)$ for some i. If one were to think that this is not the case, then one would have to assume that for some reason there is a mass point exactly at $\bar{\rho}$ in the otherwise continuous distribution of ρ_i, which is difficult to conceive.

To gather intuition about why X_i^* could be negative for some people, it may be helpful to consider the idea that some people dislike watching TV so much that, if they could, they would have preferred to "borrow" one hour from watching TV (going from $x = 0$ to $x = -1$) in order to use this extra hour on another activity, r. Of course, it is difficult to conceive this idea more concretely because x cannot be negative ever in the real world. However, there is one concrete situation that might aid this intuition. Due to the daylight saving time, some countries set back their clocks by one hour in the Fall in order to return to standard time. It is as if that one day had 25 hours instead of 24 hours. If all individuals at $X_i = 0$ were not constrained by non-negativity constraints, i.e. $V_x(X_i, R_i; \rho_i) = V_r(X_i, R_i; \rho_i)\ \forall i$, then all individuals who typically choose $X_i = 0$ in a regular 24 hour day would choose $X_i > 0$ in that 25 hour day.[21] This does not happen in practice: in the context of our application, some individuals at $X_i = 0$ do not even own a TV, and some others who do own a TV choose to use the full extra hour some other way. Thus, it is plausible that $X_i^* < 0$ for some individuals in the regular 24 hour day, so some individuals are not exactly indifferent between a marginal increase in X_i and a marginal increase in R_i at $X_i = 0$.

B Structural Equation Model

This section aims to clarify the conditions of the method for those who prefer structural equation models with explicitly defined unobservables. We translate all the conditions of Assumption 1 to that notation (making slightly stronger assumptions for simplicity).

$$Y_i = g(X_i, U_i).$$

Assumption 4. *Suppose that Y_i, X_i, X_i^* and U_i satisfy $X_i = \max\{X_i^*, 0\}$, and $Y_i = g(X_i, U_i)$. Suppose also that*

1. *X_i^* has a density $f_{X^*}(x) \leq C < \infty$ for $x \leq h$, for some $C, h > 0$, which is continuous at zero and positive in a neighborhood of zero.*

[21] This follows from the concavity of $V(\cdot, \cdot; \rho_i)$. As R_i increases to R_i' due to the extra hour, it must be that $V_x(X_i, R_i'; \rho_i) > V_r(X_i, R_i'; \rho_i)$, since $V_x(X_i, R_i; \rho_i) = V_r(X_i, R_i; \rho_i)$ in the regular day. It is easier to see this when V is twice differentiable, since then $\frac{\partial V_x(X_i, R_i; \rho_i)}{\partial r} > \frac{\partial V_r(X_i, R_i; \rho_i)}{\partial r}$ is implied by concavity.

2. *$g(x,u)$ is continuously differentiable in x for all u and $x \in [0,h)$, and $f_{U|X^*=x}(u)$ exists and is continuously differentiable in x for all u and $x \in (-\infty,h)$, for some $h>0$. Moreover, $\mathbb{E}[|g_x(x,U_i)|\,|X_i^*=x'] < \infty$ for all (x,x') in the support of the distribution of (X_i,X_i^*).*

3. *For $x \leq 0$, $\mathbb{E}[g(0,U_i)|X_i^*=x]$ is monotonic, and $g(0,u)$ is either constant in u with probability one, or $\mathbb{P}\left(\frac{d}{dX_i^*}\mathbb{E}[g(0,U_i)|X_i^*]=0\right)=0$.*

4. *Define $\epsilon_i = g(0,U_i) - \mathbb{E}[g(0,U_i)|X_i^*]$. Either (a) i: $\epsilon_i|X_i^*=x \to_d \epsilon_i|X_i^*=0$ as $x \downarrow 0$; ii: $\epsilon_i \perp\!\!\!\perp X_i^*|X_i=0$; and iii: $f_{\mathbb{E}[g(0,U)|X^*]|X=0}(y) \leq C < \infty$, for all y and some $C>0$; or the weaker condition (b) suppose that $g(0,U_i)|X_i=0$ admits a density $f_{g(0,U)|X=0}$ and that $F_{\epsilon|X=x}$ converges pointwise as $x \downarrow 0$. Denote the limit distribution as $F_{\epsilon|X=0^+}$. Define W_i a random variable derived from the deconvolution of $dF_{\epsilon|X=0^+}$ from $f_{g(0,U)|X=0}$. Then $f_{W|X=0}(\mathbb{E}[g(0,U_i)|X_i=0^+]) = f_{\mathbb{E}[g(0,U_i)|X=0}(\mathbb{E}[g(0,U_i)|X_i=0^+])$.*

References

Arai, Y. and Ichimura, H. (2016). Optimal bandwidth selection for the fuzzy regression discontinuity estimator. *Economics Letters*, 141:103–106.

Arai, Y. and Ichimura, H. (2018). Simultaneous selection of optimal bandwidths for the sharp regression discontinuity estimator. *Quantitative Economics*, 9(1):441–482.

Armstrong, T. B. and Kolesár, M. (2020). Simple and honest confidence intervals in nonparametric regression. *Quantitative Economics*, 11(1):1–39.

Bertanha, M., McCallum, A. H., and Seegert, N. (2021). Better bunching, nicer notching. *arXiv preprint arXiv:2101.01170*.

Blomquist, S., Newey, W. K., Kumar, A., and Liang, C.-Y. (2021). On bunching and identification of the taxable income elasticity. *Journal of Political Economy*, 129(8):2320–2343.

Bonhomme, S., Lamadon, T., and Manresa, E. (2017). Discretizing Unobserved Heterogeneity. Working Paper.

Bonhomme, S. and Manresa, E. (2015). Grouped Patterns of Heterogeneity in Panel Data. *Econometrica*, 83(3):1147–1184.

Bouezmarni, T. and Rombouts, J. V. (2010). Nonparametric density estimation for multivariate bounded data. *Journal of Statistical Planning and Inference*, 140(1):139–152.

Caetano, C. (2015). A Test of Exogeneity Without Instrumental Variables in Models With Bunching. *Econometrica*, 83(4):1581–1600.

Caetano, C., Caetano, G., and Escanciano, J. C. (2021a). Regression discontinuity design with multivalued treatments. *arXiv preprint arXiv:2007.00185*. Available here.

Caetano, C., Caetano, G., Fe, H., and Nielsen, E. (2021b). A Dummy Test of Identification in Linear and Panel Models with Bunching. *Working Paper*. Available here.

Caetano, C., Caetano, G., and Nielsen, E. (2020). Correcting for Endogeneity in Models with Bunching. Working Paper 2020-80, Federal Reserve Board. Available here.

Caetano, C., Caetano, G., and Nielsen, E. (2021c). Should children do more enrichment activities? Leveraging bunching to correct for endogeneity. *Working Paper*. Available here.

Caetano, C., Caetano, G., Nielsen, E., and Sanfelice, V. (2021d). The Effect of Maternal Labor Supply on Children's Skills. *Working Paper*. Available here.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2020). Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. *The Econometrics Journal*, 23(2):192–210.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

Cattaneo, M. D., Jansson, M., and Ma, X. (2020). Simple local polynomial density estimators. *Journal of the American Statistical Association*, 115(531):1449–1455.

Cheng, M.-Y., Fan, J., and Marron, J. S. (1997). On automatic boundary corrections. *The Annals of Statistics*, 25(4):1691–1708.

Cheng, X., Schorfheide, F., and Shao, P. (2019). Clustering for multi-dimensional heterogeneity.

Cheruiyot, L. R. (2020). Local linear regression estimator on the boundary correction in nonparametric regression estimation. *Journal of Statistical Theory and Applications*, 19(3):460–471.

Cytrynbaum, M. (2020). Blocked clusterwise regression. *arXiv preprint arXiv:2001.11130*.

Delaigle, A. and Gijbels, I. (2004). Practical bandwidth selection in deconvolution kernel density estimation. *Computational statistics & data analysis*, 45(2):249–267.

Dong, Y., Lee, Y.-Y., and Gou, M. (2021). Regression discontinuity designs with a continuous treatment. *Journal of the American Statistical Association*, pages 1–14.

Fan, J. and Gijbels, I. (2018). *Local polynomial modelling and its applications*. Routledge.

Gentzkow, M. and Shapiro, J. M. (2008). Preschool television viewing and adolescent test scores: Historical evidence from the Coleman study. *The Quarterly Journal of Economics*, 123(1):279–323.

Goff, L. (2020). Treatment effects in bunching designs: The impact of the federal overtime rule on hours. Technical report, Working Paper.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

He, Y. and Bartalotti, O. (2020). Wild bootstrap for fuzzy regression discontinuity designs: obtaining robust bias-corrected confidence intervals. *The Econometrics Journal*, 23(2):211–231.

Hjort, N. L. and Jones, M. C. (1996). Locally parametric nonparametric density estimation. *The Annals of Statistics*, pages 1619–1647.

Imbens, G. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of economic studies*, 79(3):933–959.

Imbens, G. and Wager, S. (2019). Optimized regression discontinuity designs. *Review of Economics and Statistics*, 101(2):264–278.

Li, Q. and Racine, J. S. (2007). *Nonparametric econometrics: theory and practice*. Princeton University Press.

Loader, C. R. (1996). Local likelihood density estimation. *The Annals of Statistics*, 24(4):1602–1618.

Munasib, A. and Bhattacharya, S. (2010). Is the 'idiot's box' raising idiocy? Early and middle childhood television watching and child cognitive outcome. *Economics of Education Review*, 29(5):873 – 883.

Noack, C. and Rothe, C. (2019). Bias-aware inference in fuzzy regression discontinuity designs. *arXiv preprint arXiv:1906.04631*.

Pinkse, J. and Schurter, K. (2021). Estimates of derivatives of (log) densities and related objects. *Econometric Theory*, pages 1–36.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90(432):1257–1270.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The annals of statistics*, pages 1346–1370.

Saez, E. (2010). Do Taxpayers Bunch at Kink Points? *American Economic Journal: Economic Policy*, 2(3):180–212.

Schennach, S. (2021). Measurement systems. *Journal of Economic Literature*.

Schennach, S. M. (2019). Convolution without independence. *Journal of econometrics*, 211(1):308–318.

Zavodny, M. (2006). Does watching television rot your mind? Estimates of the effect on test scores. *Economics of Education Review*, 25(5):565 – 573.

Zhang, S. and Karunamuni, R. J. (1998). On kernel density estimation near endpoints. *Journal of statistical Planning and inference*, 70(2):301–316.