Identification of Causal Effects with a Bunching Design*

Carolina Caetano
University of Georgia

Gregorio Caetano
University of Georgia

Leonard Goff
University of Calgary

Eric Nielsen
Federal Reserve Board

May 2025

Latest version can be found here.

Abstract

We show that causal effects can be nonparametrically identified using only bunching phenomena (i.e., in the absence of instrumental variables, panel data, or other strategies currently used for nonparametric causal identification). Specifically, if the treatment variable has bunching, we show that the selection bias can be identified. The main insight is the application of the change-of-variables theorem from integration theory, which allows us to write the selection bias as a ratio of the density of the treatment and the density of the selection function. Although the selection function cannot be identified, at the bunching point, the outcome differences reflect only the selection function and the idiosyncratic error. Thus, the density of the selection function can be recovered via deconvolution of the idiosyncratic errors from the distribution of the outcome at the bunching point. Our main result identifies the average causal response to the treatment among individuals who marginally select into the bunching point. We further show that, under additional smoothness assumptions on the endogeneity bias, the treatment effects away from the bunching point may also be identified. We propose estimators based on standard software packages and apply the method to estimate the effect of maternal smoking during pregnancy on birth weight.

1 Introduction

In this paper, we show that bunching in the distribution of a treatment variable can be used to identify causal effects in the absence of the tools currently used for causal identification. In particular, although the treatment may be endogenous, we do not rely on instrumental variables, regression discontinuity designs, panel data, functional form, or distributional assumptions.

The setting is a standard causal model where both the treatment and the outcome variables are observed. The treatment variable has a bunching point, and is continuously distributed near the bunching point. The example in our application is a useful benchmark, where the treatment is the

*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, Alfonso Flores-Lagunes, Stefan Hoderlein, Hugo Jales, Matthew Masten, Eric Mbakop, Whitney Newey, Joris Pinkse, Alexandre Poirier, Demian Pouzo, Karl Schurter, Chris Taber, as well as seminar participants at several institutions and conferences for valuable help and feedback. 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.

number of cigarettes a woman smokes during pregnancy (with 81% of the observations bunching at zero), and the outcome is the baby's birth weight. Caetano (2015) showed strong evidence of endogeneity in this application. To identify treatment effects, we need to get rid of selection bias, the part of the outcome variation that is due to confounders.

The key insight that makes the identification possible is that the change-of-variables theorem from integration theory can be used to write the magnitude of selection bias as the ratio of the probability density of the treatment variable and the probability density of the part of the outcome that is due to confounders. As a consequence, we do not need to observe the values of those confounders; it is instead sufficient to identify the distribution of the part of the outcome that varies with confounders. This is where bunching is useful: at the bunching point, the outcome variation is due only to confounders, since the treatment stays fixed. We use the distribution of the outcome at the bunching point to identify the selection bias.

We identify the average marginal effect at the bunching point among those near the bunching point (equivalently, we identify the rate of outcome response to a marginal increase in treatment at the bunching point, among the observations at the bunching point that are most similar to those near the bunching point). In our application, this quantity represents the expected rate of birth weight loss if a woman who is currently not a smoker but is very similar to the women who smoke very little were to start smoking. Alternatively, this quantity can be interpreted as the expected rate of birth weight gain if the women who currently smoke little were to quit smoking.

The approach relies on four conditions. First, the treatment effects must be sufficiently smooth near the bunching point. In our application, the condition on the treatment effects means that smoking is not so poisonous that a marginal amount of smoking could on average cause a discrete birth weight change. Second, the selection function must also be sufficiently smooth at the bunching point. In our application, this means that the mothers who smoke a marginal amount are comparable to the mothers who do not smoke but are indifferent between not smoking and smoking a positive amount. Third, the selection bias maintains the same sign in a neighborhood near the bunching point. In our application, it means that if smoking selection is negative among mothers who smoke little (i.e. smoking less is associated with higher untreated birth weights), then selecting into not smoking (i.e. selecting to not smoke as a corner solution) must be associated with even higher birth weights. Finally, the outcomes at the bunching point are determined by confounders, but we also allow additional (unconfounded) variation, which must then be deconvolved. By construction, the unconfounded variation is mean independent of the confounders, but the deconvolution step requires full independence at the bunching point (or alternatively the weaker subindependence condition in (Schennach, 2019)).

If the selection function is real analytic, then our results allow the identification of ATTs near the bunching point. If some bounds on the selection bias derivatives can be assumed, then the ATTs can be identified further away. Specifically, we can identify the effect among those who took a given treatment value versus a counterfactual where they take the bunching point treatment value. In our application, it is thus possible to identify the birth weight gains (or losses) if mothers who smoke a

given amount were to quit smoking.

We propose estimators that use well-known building blocks. We estimate expectations and derivatives near the bunching point with local linear estimators (Fan and Gijbels, 1992) and boundary densities using Pinkse and Schurter (2021)'s estimator, both of which achieve interior rates of convergence at the boundary. The deconvolution step follows standard nonparametric methods, equivalent to a standard kernel density estimator using a special kernel. All building blocks can be implemented by plugging in existing packaged software.

We also explore how controls may be used to study heterogeneous treatment effects as well as to weaken the identification assumptions (which may all then be required only conditional on controls). In particular, this allows the sign of selection bias near the bunching point to be different for different groups of observations. We discuss estimation in cases with discrete and continuous controls, as well as in the case where the vector of controls may be large, and include mixed discrete and continuous variables.

We apply our approach to the data on smoking and birth weight from Almond et al. (2005). We show that, after correcting for the selection bias, the effect of the first daily cigarette is an insignificant loss of about 8 grams in birth weight (less than 1/3 ounces). Smoking 5 cigarettes per day causes an insignificant loss of less than 1 ounce (compare this to the average weight of a full term newborn, which is 120 ounces). These estimates confirm and strengthen the qualitative point in Almond et al. (2005) that smoking is not an important cause of birth weight.

Bunching is a common phenomenon. It is often found at zero in non-negative variables, such as consumption goods,[1] financial variables,[2] time use,[3] and neighborhood characteristics.[4] Artificial constraints also can generate bunching, such as regulatory minimums[5] and maximums.[6] Bunching also occurs at interior points, often due to kinks or notches in budget sets, social norms and other restrictions.[7] Of course, small samples, coarse measurements, or attrition can make it impossible to implement this method in some of the examples above.

[1] E.g., number of tobacco products, alcoholic beverages, caffeinated drinks, sugary drinks, fast food meals, dining out meals, subscription services, supplements and vitamins, public transportation rides, books read, gym visits, doctor visits, trips, fuel usage amounts, expenditure on health, fitness, travel, vacations, education, childcare.

[2] E.g., credit access, bequests, savings, emergency fund levels, retirement account contributions, mortgage balance, credit card debt, student loan debt, income from investments, expenditure on ads, charitable donations, HSA and FSA balances, life insurance coverage, number of trades.

[3] Bunching is found for most time uses with few exceptions. Some examples: exercising, working, watching TV, using digital devices, doing homework, doing chores, volunteering, commuting.

[4] E.g., number of public transportation options or stops, retail stores, coffee shops, rental units, affordable housing units, vacant units, electric vehicle charging stations; length of biking lanes or walking paths; areas of green space, commercial districts, sport fields, parking lots.

[5] E.g., schooling time, wages, 401K contributions, coverage for auto insurance, nutritional standards for school meals, bank capital, bank deposit insurance, age started working, age started withdrawing from retirement accounts, age retired.

[6] E.g., contribution size in 401K, Roth IRA, HSA, FSA accounts, untaxed gifts, FHA loans, FDIC insurance, carbon emissions, liquor licenses, lot coverage, contributions to political campaigns, data usage, grades, absences from school, class size, commissions on sales.

[7] E.g. income at tax brackets, hours worked at overtime rules, multiples of 5, 40 hours per week, car speeds at ticket thresholds, financial reporting around profit targets, energy consumption around utility billing tiers, pricing below psychological points ($0.99), doctor visits at medical protocol numbers, hospital stay length at insurance payment thresholds.

The use of bunching phenomena for identification began with Saez (2010), followed by a large applied literature interested in the effect of a policy change (or a related structural parameter) at the threshold of a manipulable variable, which ends up with bunching at the change threshold. Theoretical treatment of these approaches may be found in Blomquist et al. (2021), Bertanha et al. (2023b), Goff (2023) and Lu et al. (2024). Our approach is more related to the literature initiated by Caetano (2015), where bunching for any reason on the treatment variable of a reduced form causal model allows the testing of the model's identification conditions (see Caetano et al. (2016), Caetano and Maheshri (2018), Caetano et al. (2021), and Khalil and Yıldız (2022)). Caetano et al. (2023) developed the first strategy for identification of treatment effects under endogeneity in this setting, followed by Caetano et al. (2024c) and Caetano et al. (2024d), where distributional and functional form assumptions are relaxed. Surveys of the bunching literature include Kleven (2016); Jales and Yu (2017); Blomquist et al. (2023), and Bertanha et al. (2023a). In this paper, we show that nonparametric identification with bunching can be attained.

Section 2 introduces our setting, in which a treatment is continuously distributed near a bunching point, and introduces a novel approach to identification of the average marginal treatment effect using the change-of-variables theorem from integration theory. We detail how the average marginal treatment effect at the bunching point can be identified in Section 3, and how treatment effects can then be identified away from the bunching point in Section 4. We turn to estimation in Section 5, and in Section 6 we present the application to the effects of smoking on birth weight. We conclude in Section 7. Appendices contain extensions for the use of controls, examples and generalizations referred to in the text, and proofs.

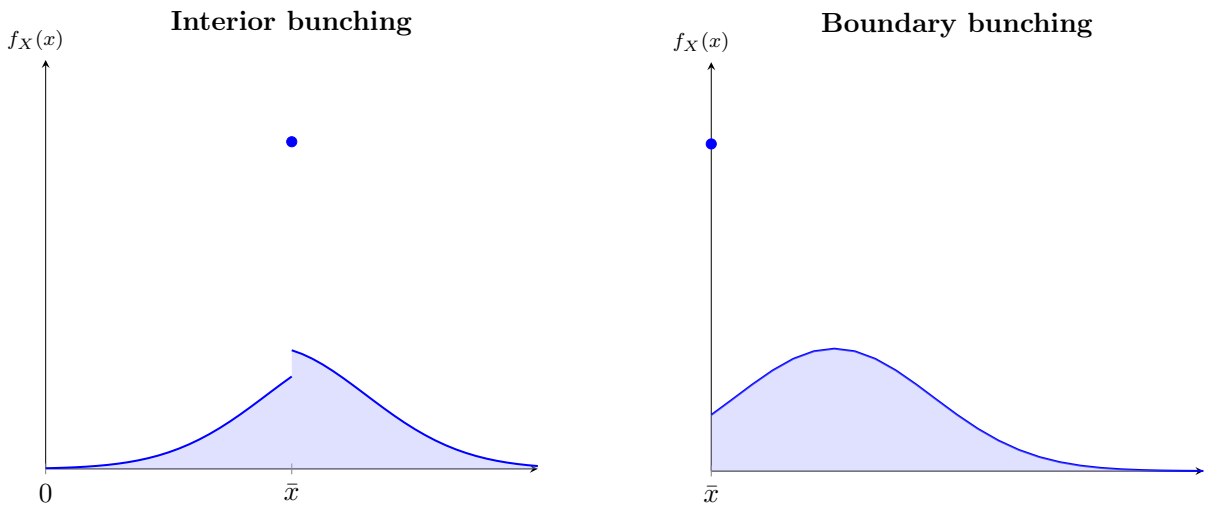
2 Identification near the bunching point

In this section, we set up the identification problem in the neighborhood of the bunching point and relate the identification of causal effects to the distribution of the selection bias.

Our setting is the standard potential outcomes framework, where observation i's outcome reacts to the value x of a multivalued scalar treatment variable through the potential outcome function $Y_i(x)$. We observe the treatment value X_i and the outcome $Y_i = Y_i(X_i)$. The support of X_i includes a nondegenerate interval, whose left boundary $\bar{x}$ exhibits bunching. In many relevant applications, X_i is continuously distributed on the positive real line with the bunching point $\bar{x} = 0$. The right panel of Figure 1 depicts this case, while the left panel depicts a case in which $\bar{x}$ is in the interior of the support of X_i. One can accommodate bunching on the right boundary of the support of X_i by simply redefining X_i as $\bar{x} - X_i$.

We will adopt the following notational conventions. For an arbitrary function $v \mapsto g(v)$, we denote the k-th derivative at $\tilde{v}$ as $g^{(k)}(\tilde{v})$. For the first derivative, we also use the notation $g'(\tilde{v}) = g^{(1)}(\tilde{v})$. For any function $g(x)$ let $g(\bar{x}^+) := \lim_{x \downarrow \bar{x}} g(x)$. To avoid unnecessarily strong differentiability assumptions, we define for derivatives $g'(\bar{x}^+) = \lim_{x \downarrow \bar{x}} (g(x) - g(\bar{x}^+))/(x - \bar{x})$ (rather than as $\lim_{x \downarrow \bar{x}} g'(x)$), which amounts to a right-derivative defined with respect to the limit $g(\bar{x}^+)$

Figure 1: Two examples of bunching



instead of $g(\bar{x})$. Higher order limit derivatives $g^{(k)}(\bar{x}^+)$ are defined analogously. For other composite functions $(g \circ h)(x) := g(h(x))$, we let $g(h(\bar{x}^+)) := (g \circ h)(\bar{x}^+)$. For an arbitrary random variable V_i, let $F_V(v) = \mathbb{P}(V_i \leq v)$ and $f_V(v) = F_V'(v)$ if the derivative exists. Analogously, for a set $S \subseteq \mathbb{R}$, $F_{V|S}(v) = \mathbb{P}(V_i \leq v|S)$ and $f_{V|S} = F_{V|S}'(v)$ if the derivative exists. For example, $f_{V|S}(x)$ denotes $\frac{f_V(v)}{P(S)}$ for any $v \in S$. Define the sign of v as $\text{sgn}(v) = \mathbf{1}(v \geq 0) - \mathbf{1}(v \leq 0)$. For a set $S \subseteq \mathbb{R}$, we say $g(S)$ to mean the image of the function g over S. For simplicity, we say the "support of V_i" when we mean the support of the distribution of V_i.

Remark: (Friction around $\bar{x}$) The examples in Figure 1 depict "perfect" bunching in the sense of a point mass in the distribution of X_i (to the extent that this can be visualized in a density plot). By contrast, interior bunching is often somewhat diffuse around the bunching point, due either to optimization frictions or X_i being measured with error. We abstract from this issue and assume that the researcher has a means of identifying the "bunched" observations $X_i = \bar{x}$. This is generally not problematic in settings with boundary bunching, and for interior bunching measurement error can sometimes be eliminated by having administrative data (see Goff 2023 for an example). For a general discussion of solutions in settings with optimization frictions, see Kleven (2016).

2.1 Parameters of interest and the identification problem

To define treatment effect parameters, we let the outcome $Y_i(\bar{x})$ that would occur if treatment were equal to $\bar{x}$ play the role of the "untreated" state. When bunching is at zero, this recovers the familiar notation $Y_i(0)$. Let $\text{ATT}(x)$ denote the average effect of moving from the bunching point to point

x, among those with $X_i = x$:

$$\text{ATT}(x) := \mathbb{E}[Y_i(x) - Y_i(\bar{x})|X_i = x] = \mathbb{E}[Y_i|X_i = x] - \mathbb{E}[Y_i(\bar{x})|X_i = x],$$

where we have used $Y_i = Y_i(X_i)$ in the rightmost expression. This expression highlights the challenge of identifying the causal quantity $\text{ATT}(x)$, since $\mathbb{E}[Y_i(\bar{x})|X_i = x]$ is counterfactual and not directly observed for $x \neq \bar{x}$.

In our empirical application, x measures cigarettes per day and the bunching point is $\bar{x} = 0$, so $\text{ATT}(x)$ measures the average birth weight loss (or gain) mothers who smoke x cigarettes incur for smoking that amount (versus a counterfactual where they do not smoke at all). There are many mothers in this application who smoke zero cigarettes, leading to a case of boundary bunching as in the right panel of Figure 1. Here, the $\text{ATT}(x)$ is the effect of smoking x cigarettes per day among those that smoked that amount, and $\mathbb{E}[Y_i(\bar{x})|X_i = x]$ is the birth weight if mothers who smoked x cigarettes per day were to quit.

Local effects of the treatment around a value x can be obtained by inspecting the derivative of the function $\text{ATT}(x)$. We define the *average marginal effect near the bunching point* $\text{AME}^+_{\bar{x}}$ as the right derivative of $\text{ATT}(x)$ as x approaches the bunching point from above, i.e.

$$\text{AME}^+_{\bar{x}} := \lim_{x \downarrow \bar{x}} \frac{\text{ATT}(x)}{x - \bar{x}} = \lim_{x \downarrow \bar{x}} \mathbb{E}\left[\frac{Y_i(x) - Y_i(\bar{x})}{x - \bar{x}} \middle| X_i = x\right]$$

When the bunching point is interior as in the left panel of Figure 1, or the bunching point is on the right boundary of the support of X_i one could define a similar $\text{AME}^-_{\bar{x}}$ parameter describing the left limit at $\bar{x}$. We use the right derivative to define the treatment effects of interest to fit our application in which bunching is at the left boundary of the support of X_i, i.e. $\bar{x} = 0$ with x a (weakly positive) number of cigarettes. In our application, the $\text{AME}^+_{\bar{x}}$ is the rate of birth weight loss (or gain) incurred by those who smoked just a little (versus the counterfactual where they would not have smoked at all), expressed on a per-cigarette basis.

If the individual potential outcome functions $Y_i(x)$ are differentiable and regularity conditions permitting the exchange of limits and expectations hold, one can interpret $\text{AME}^+_{\bar{x}}$ in terms of an average of the derivatives $Y_i'(x)$ of these dose-response functions, i.e. $\text{AME}^+_{\bar{x}} = \lim_{x \downarrow \bar{x}} \text{AME}(x)$ with $\text{AME}(x) := \mathbb{E}[Y_i'(x)|X_i = x]$. It is for this reason that we use the terminology of a *marginal effect* to refer to $Y_i'(x)$, in line with e.g. Hoderlein and Mammen (2007); Imbens and Newey (2009); Chiang and Sasaki (2019). Other authors use the term *partial effect* (see e.g. Sasaki 2015; Kato and Sasaki 2017), emphasizing the interpretation of $Y_i(x)$ as $g(x, U_i)$ for an underlying structural function g over heterogeneity U_i in potential outcomes, in which case $Y_i'(x)$ becomes the partial derivative of $g(x, U_i)$ with respect to x. A related quantity is the ACRT (average causal response on the treated function), studied in Callaway et al. (2024). Despite our use of the term *marginal effect*, our results do not require $Y_i'(x)$ to exist with probability one. Rather, we maintain weaker assumptions that are nevertheless sufficient to ensure that $\text{AME}^+_{\bar{x}}$ remains well-defined.

Like with $\text{ATT}(x)$, identifying average marginal effects is challenging because the regression

derivative of Y_i on X_i generally confounds the casual effect of treatment with a bias due to endogeneity:

$$\frac{d}{dx}\mathbb{E}[Y_i|X_i=x] = \frac{d}{dx}\mathbb{E}[Y_i(x)|X_i=x] = \overbrace{\frac{d}{dx}\mathbb{E}[Y_i(x)|X_i=x']\Big|_{x'=x}}^{\text{causal effect}} + \overbrace{\frac{d}{dx}\mathbb{E}[Y_i(x')|X_i=x]\Big|_{x'=x}}^{\text{selection bias}}$$

where the first term above is, under regularity conditions, equal to the average marginal effect at x: $\mathbb{E}[Y_i'(x)|X_i=x]$. An analogous decomposition for $\text{AME}_{\bar{x}}^{+}$ implies that:

$$\text{AME}_{\bar{x}}^{+} = \lim_{x\downarrow\bar{x}}\frac{d}{dx}\mathbb{E}[Y_i|X_i=x] - \lim_{x\downarrow\bar{x}}\frac{d}{dx}\mathbb{E}[Y_i(\bar{x})|X_i=x] \tag{1}$$

The first term above is the right limit of the derivative of the observed outcome and the second term reflects endogeneity: those with different values of X_i may have different mean values of $Y_i(\bar{x})$.

The following definitions will be useful in simplifying the exposition throughout our analysis:

$$m(x) := \mathbb{E}[Y_i|X_i=x] - \mathbb{E}[Y_i|X_i=\bar{x}^+], \tag{2}$$

which yields a comparison of observed outcomes at $X_i = x$ relative to the boundary as $x \downarrow \bar{x}$. Similarly, define:

$$s(x) := \mathbb{E}[Y_i(\bar{x})|X_i=x] - \mathbb{E}[Y_i(\bar{x})|X_i=\bar{x}^+], \tag{3}$$

which denotes the comparison of the counterfactual outcomes $Y_i(\bar{x})$ relative to the boundary as $x \downarrow \bar{x}$. Then, we write

$$\text{ATT}(x) = m(x) - s(x) \qquad \text{and} \qquad \text{AME}_{\bar{x}}^{+} = m'(\bar{x}^+) - s'(\bar{x}^+), \tag{4}$$

where the equalities follow from Proposition 2.1 below.

Note that the function m is identified directly from observables. In each case, the identification challenge therefore comes from the s term, which depends on unobservable counterfactuals. Intuitively, the equations in (4) depict the fundamental problem of causal inference, because the observable outcome variation among those with different treatment values, mapped by the function m, combines both the causal effect of x and selection bias, captured by the function s.

Both m and s are defined above "relative" to the limit at the bunching point, so that $m(\bar{x}^+) = s(\bar{x}^+) = 0$. Note as well that since only the first term in the definitions of m and x depends on x, $m'(x) = \frac{d}{dx}\mathbb{E}[Y_i|X_i=x]$ and $s'(x) = \frac{d}{dx}\mathbb{E}[Y_i(\bar{x})|X_i=x]$ whenever these derivatives exist.

2.2 Identification of the average marginal effect near the bunching point

The main insight of this paper is the observation that, in order to identify the derivative of the counterfactual function $\mathbb{E}[Y_i(\bar{x})|X_i=x]$ for values of x near the bunching point $\bar{x}$, it is sufficient to identify the *distribution* of the expected counterfactuals $E[Y_i(\bar{x})|X_i]$ for X_i near $\bar{x}$. We then show in Section 3 that bunching in X_i can make it possible to identify this distribution, even though the counterfactual outcome $Y_i(\bar{x})$ is never observed for those with $X_i \neq \bar{x}$. The second term in (1) can

then be estimated.

Assumption 1 (continuous support and outcome and treatment effect smoothness)**.** *The following hold:*

(i) $f_X(x)$ *exists and is continuous on an open interval* $(\bar{x}, \bar{x}+\varepsilon_1)$ *for some* $\varepsilon_1 > 0$*, and* $f_X(\bar{x}^+)$ *exists and is strictly positive.*

(ii) The function $x \mapsto \mathbb{E}[Y_i|X_i = x]$ *is differentiable on an interval* $(\bar{x}, \bar{x}+\varepsilon_2)$ *for some* $\varepsilon_2 > 0$*, and* $\mathbb{E}[Y_i|X_i = \bar{x}^+]$ *as well as* $\lim_{x\downarrow\bar{x}} \frac{d}{dx}\mathbb{E}[Y_i|X_i = x]$ *exist.*

(iii) $ATT'(\bar{x}^+)$ *exists, and* $ATT(\bar{x}^+) = 0$*.*

Item (i) of Assumption 1 states that X_i is continuously distributed with a density on an interval to the right of $\bar{x}$, though this density need not exist everywhere (e.g. there can be multiple bunching points provided that they are well-separated). Item (ii) says that a regression derivative exists and the regression function and its derivative have a right limit at the bunching point. Both parts of Assumption 1 are restrictions on the observable data that can in principle be verified empirically.

Item (iii) of Assumption 1 states that, on average the treatment effects are sufficiently smooth near the bunching point. This means that, at least on average among those with treatment levels near the bunching point, marginally small doses of the treatment should have only marginal effects. This rules out the case where $\bar{x}$ has a threshold effect (e.g. a sheepskin effect, for example). In the smoking example, this condition states that, among mothers who smoke very little, smoking is not so poisonous that a little amount can cause a stark decline in the baby's health.

Item (iii) of Assumption 1 is sufficient to guarantee that the $\text{AME}^+_{\bar{x}}$ is well defined. This is because, since $\text{ATT}(\bar{x}^+) = 0$, $\text{AME}^+_{\bar{x}} = \text{ATT}'(\bar{x}^+)$. A sufficient condition (though stronger than necessary) for item (iii) is that the $Y_i(x)$ are differentiable and uniformly bounded with probability one near the bunching point, which furthermore implies that $\text{AME}^+_{\bar{x}} = \mathbb{E}[Y_i'(\bar{x})|X_i = \bar{x}^+]$.

Next we make a similar assumption about counterfactuals:

Assumption 2 (counterfactuals smoothness)**.** *The function* $\mapsto \mathbb{E}[Y_i(\bar{x})|X_i = x]$ *is differentiable on an interval* $(\bar{x}, \bar{x}+\varepsilon_3)$ *for some* $\varepsilon_3 > 0$*, where* $\lim_{x\downarrow\bar{x}} \frac{d}{dx}\mathbb{E}[Y_i(\bar{x})|X_i = x]$ *exists and is different from zero.*

Assumption 2 requires that the selection function $s(x)$ be differentiable in a neighborhood above the bunching point. It also rules out the case in which there is no endogeneity when one approaches $\bar{x}$ from above, but in this case, no correction for endogeneity is necessary. Note that the presence of endogeneity needing a correction can be diagnosed using Caetano (2015)'s test.

The following proposition establishes the connection between the parameters of interest and the functions m and s presented in Equation (4). All proofs are found in Appendix E.

Proposition 2.1. *Under Assumptions 1-2,* $ATT(x) = m(x) - s(x)$ *and* $AME^+_{\bar{x}} = m'(\bar{x}^+) - s'(\bar{x}^+)$*.*

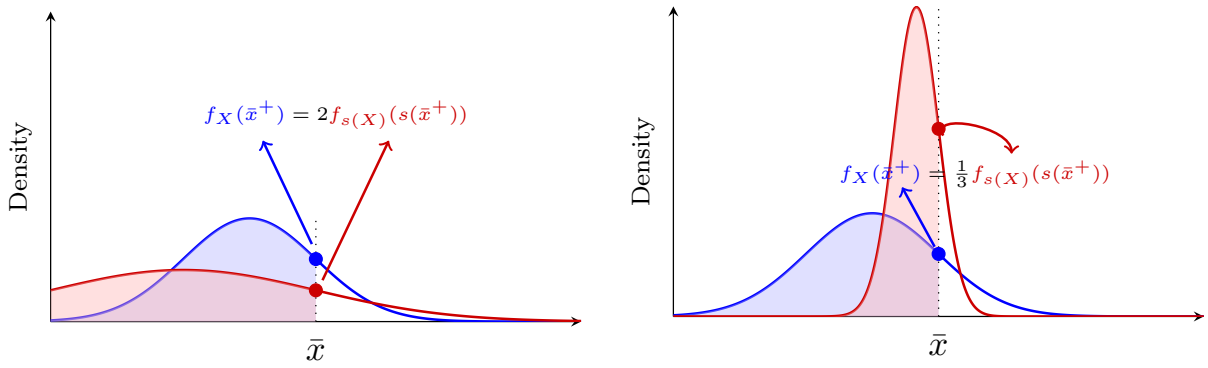
We now move to the first main result of this paper. Assumption 2 implies that the counterfactual function $s(x)$ is differentiable and locally monotonic for x in some neighborhood above the bunching point (without loss, we can take $I = (\bar{x}, \bar{x} + \varepsilon)$, for some $\varepsilon < \min\{\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4\}$, and note that ε need never be known). The local monotonicity and differentiability in I allows us to apply the well-known change-of-variables formula from integration theory.

Theorem 2.1 (Change of variables)**.** *If Assumptions 1 and 2 hold, then there exists $I = (\bar{x}, \bar{x} + \varepsilon)$ for some $\varepsilon > 0$ such that, for $x \in I$, $f_{s(X)|I}(s(x))$ exists and is non-zero for all $x \in I$, and*

$$\left|s'(x)\right| = \frac{f_{X|I}(x)}{f_{s(X)|I}(s(x))}. \tag{5}$$

Note the absolute value $|s'(x)|$ in (5): the RHS is always positive, but $s'(x)$ will be negative if there is negative selection. The textbook change-of-variables formula states that $f_{u(X)}(t) = f_X(u^{-1}(t))/|u'(u^{-1}(t))|$ for any $t \in u^{-1}(I)$, given any function $u(x)$ that is differentiable and strictly increasing on an interval I (and anagously for a strictly decreasing u). Then the claim follows with $u(X) = s(X)$ and $t = s(x)$. Though the change-of-variables formula is a standard tool (see e.g. Fremlin 2011 for a general formulation), a proof of Theorem 2.1 is provided in Appendix E, which also establishes the local monotonicity of s required for the change-of-variables result. Figure 2 provides a visual illustration of the change-of-variables theorem for scenarios with high and low selections for the case when $\bar{x} = 0$, as in our application.

Figure 2: Using the Change-of-Variables Theorem to Identify $s'(\bar{x}^+)$



High Selection: $s(X_i) = 2(X_i - \bar{x})$ Low Selection: $s(X_i) = \frac{1}{3}(X_i - \bar{x})$

Note: The blue curve is the density of $X_i \sim N(\bar{x} - 1, \bar{x} + 1)$ in both panels. The red curve is the density of $s(X_i)$ given selection taking the linear form $\mathbb{E}[Y_i(\bar{x})|X_i = x] = a + b \cdot (x - \bar{x})$, so that $s(x) = b \cdot (x - \bar{x})$ and $s'(x) = b$. The equations relating the heights of the solid dots on each panel highlight the proportionality of the densities at $x = \bar{x}^+$, matching $s'(\bar{x}^+) = b$. Note that for lower selection, the density of $s(X_i)$ at $\bar{x}^+$ is larger, increasing the denominator of the $s'(\bar{x}^+)$ formula. This illustration assumes $s(\bar{x}^+) = \bar{x}$, which is true when $\bar{x} = 0$, as in our application.

Let $\theta := \text{sgn}\left(\lim_{x \downarrow \bar{x}} \frac{d}{dx}\mathbb{E}[Y_i(\bar{x})|X_i = x]\right) = \text{sgn}(s'(\bar{x}^+))$ be the sign of the selection bias as one

approaches $\bar{x}$ from the right. As an application of Theorem 2.1, we have the following expression for $\text{AME}^{+}_{\bar{x}}$:

Theorem 2.2. *(Average marginal effect near the boundary) If Assumptions 1 and 2 hold, then*

$$AME^{+}_{\bar{x}} = m'(\bar{x}^{+}) - \theta \cdot \frac{f_{X|I}(\bar{x}^{+})}{f_{s(X)|I}(0)}.$$

3 Identification using bunching

The previous section established that identification of treatment effects may be possible if we can identify the sign of the selection bias at the boundary point, $\theta = \text{sgn}(s(\bar{x}^{+}))$, and the limit of the density of the selection bias variable, $f_{s(X)}(s(x))$, as $x \downarrow \bar{x}$. In this section, we show how information at the bunching point may be used to obtain these quantities.

We start by noting that the treatment variable describes two distinct concepts. First, it describes the dose taken by an individual, i.e. the number of cigarettes smoked in our application. Second, it tells us something about that individual, i.e. the fact that those with that treatment value are of the "type" that selected that amount. Thus, for example, in the parameter $\text{ATT}(x) = \mathbb{E}[Y_i(x) - Y_i(0)|X_i = x]$, the first x describes the dose, and the second describes the group that selected it. Here, we separate the notation of the two concepts: the dose is the treatment variable, denoted X_i, as before, and the selection variable is denoted X_i^*.

In most cases, the selection variable X_i^* is identical to X_i. Indeed, this is precisely what gives rise to endogeneity: if X_i^* is correlated with the potential outcomes $Y_i(x)$, then the correlation between $Y_i = Y_i(X_i)$ and $X_i = X_i^*$ reflects both the causal effect (i.e. the part that refers to the variation of the function $Y_i(x)$ with x for a given i), and the selection function (i.e. the part that refers to the variation of $Y_i(x)$ across the i with different values of X_i^*). Such is the case here as well, for values away from the bunching point. When $X_i^* > \bar{x}$, there is no constraint on the treatment value, and $X_i^* = X_i$. However, the bunching setting is interesting in that multiple values of the selection variable X_i^* occur simultaneously at the same treatment value $X_i = \bar{x}$.

A separation between "types" and their dosage is a common feature of the bunching literature. See, for example Kleven and Waseem (2013), Saez (2010), Blomquist et al. (2021), Bertanha et al. (2023b), Bertanha et al. (2023a), Caetano et al. (2023), Goff (2023), and Caetano et al. (2024c). The separation arises from constraints on individuals' choices that cause different types to all choose the common bunching point. The idea is that, at the bunching point, selection breaks away from the dose, and observations with diverse selection values have the exact same dosage amount. Thus, the bunching point affords the possibility of learning about the relationship between the selection and other variables without any confusion arising from the variation in the dose.

We write

$$X_i = \max\{X_i^*, \bar{x}\}, \tag{6}$$

which fits the case of bunching at the left boundary of the support of X_i. The right boundary case also fits this description, by redefining the treatment to be $\bar{x} - X_i$. Interior bunching resulting from

kinks in the budget function can also be adapted to fit Equation (6), as we describe in Example 3.2.

In general, one can think of X_i^* as an index of all observable and unobservable individual characteristics that determine the value of the treatment. In Section 3.4, we provide examples in which X_i^* has a specific economic interpretation in terms of structural model primitives or reduced form quantities, and we discuss the uniqueness of X_i^*. For the purposes of this section, we provide a brief intuitive discussion of X_i^* in the context of our empirical application to maternal smoking. In that setting, it is natural to think that although all non-smoking mothers share a value $X_i = 0$ of the treatment, they may differ in the intensity of their preference towards not smoking or other factors that influence their choice. For instance, if ρ_i is a parameter that governs the person's relative preference towards smoking, there may exist a value $\bar{\rho}$ and a smooth function h such that $X_i = h(\rho_i)$ when $\rho_i \geq \bar{\rho}$ (i.e. when the person's preference towards smoking is sufficiently high), and $X_i = 0$ otherwise, where $P(\rho < \bar{\rho}) > 0$ (i.e. some observations strictly prefer not to smoke).

Intuitively, X^* allows us to track observations at $X_i = \bar{x}$ based on how selected they are relative to the observations near the bunching point in the positive side. The key assumptions about X_i^* will be that the smoothness conditions from Section 2.2 can be extended to values of X_i^* around $\bar{x}$, so that those observations away from the bunching point with X_i near $\bar{x}$ are comparable to the observations with $X_i^* = \bar{x}$. This allows us to substitute the limit of the density of the selection above the bunching point in Theorem 2.2 into the density of the selection at the bunching point. The following sections then show that the sign of the selection bias is identified (Section 3.1), and the density may be obtained by a deconvolution from the density of the outcome at the bunching point (Sections 3.2 and 3.3). Finally, in Section 3.4 we discuss the nature of X_i^* when it arises from choice models, the invariance of the identification results to monotonic transformations of X_i^*, and how it may be artificially constructed under some conditions.

3.1 Identifying the sign of the endogeneity bias, θ.

We begin by extending the definition of $s(x)$ to X_i^* as $s(x) := \mathbb{E}[Y_i(\bar{x})|X_i^* = x] - \mathbb{E}[Y_i(\bar{x})|X_i^* = \bar{x}^+]$. Since $X_i = X_i^*$ when $X_i \geq \bar{x}$, this coincides with the function s defined in Section 2.2 for all $x > \bar{x}$.

Assumption 3. *For any $x < \bar{x}$: $sgn(s(x)) = -\lim_{x \downarrow \bar{x}} sgn(s(x))$.*

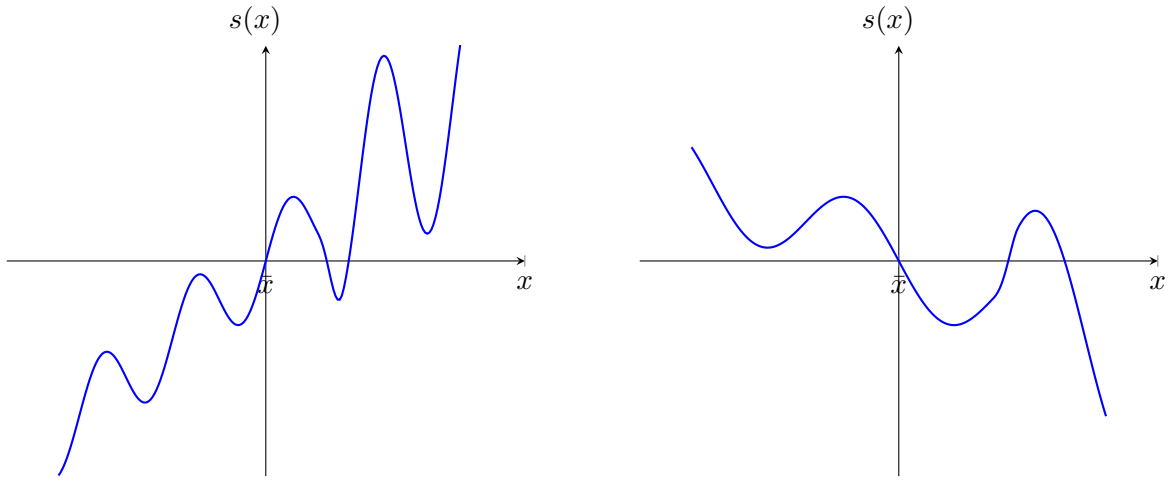
Assumption 3 states that if $s(x)$ is increasing in a positive neighborhood around the bunching point, then $s(x) < 0$ for all values of $x < \bar{x}$. Conversely, if $s(x)$ is decreasing in a positive neighborhood around the bunching point, then $s(x) < 0$ for all values of $x < \bar{x}$. Specifically, either $\mathbb{E}[Y_i(\bar{x})|X_i^* = x'] < \mathbb{E}[Y_i(\bar{x})|X_i^* = \bar{x}^+] < \mathbb{E}[Y_i(\bar{x})|X_i^* = x'']$ for all $x' < \bar{x}$ and all $x'' > \bar{x}$ in a neighborhood right above the bunching point, or the reverse ordering is true. Intuitively, the selection function $\mathbb{E}[Y_i(\bar{x})|X_i^* = x]$ maintains its tendency when comparing the values at the bunching point to the boundary as $x \downarrow \bar{x}$ and from there to the positive values.

In the smoking example, suppose that, among mothers smoking positive amounts, those who smoke more are negatively selected relative to those who smoke less (i.e., smoking more is associated with worse untreated outcomes). Then, Assumption 3 states that the nonsmoking mothers

would have even higher birth weights. This needs to hold only on average for every selection value ($\mathbb{E}[Y_i(\bar{x})|X_i^* = x] > \mathbb{E}[Y_i(\bar{x})|X_i^* = \bar{x}^+]$ for $x \leq 0$), so individual nonsmoking mothers could have lower birth weights than some smoking mothers.

Assumption 3 holds trivially if $\mathbb{E}[Y_i(\bar{x})|X_i^*]$ is monotonic, in which case $s(x)$ is also monotonic. However, the assumption is weaker than monotonicity. Figure 3 illustrates two examples of non-monotonic functions that satisfy Assumption 3, one case with $s'(\bar{x}) > 0$, and another with $s'(\bar{x}) < 0$, respectively. Note that $s(\bar{x}^+) = 0$ by definition, and Assumption 3 does not constrain the behavior of $s(x)$ for positive x, outside of a neighborhood of the bunching point.

Figure 3: Examples of functions s that satisfy Assumption 3 but are not monotonic.



Note: Each panel shows an example of a selection function $s(x)$ that satisfies Assumption 3 but is not monotonic in x. Note that in both cases $s(\bar{x}^+) = 0$, which is true by definition.

Lemma 1. *If Assumptions 1, 2 and 3 hold, then θ is identified as*

$$\theta = sgn\left(\mathbb{E}[Y_i|X_i = \bar{x}^+] - \mathbb{E}[Y_i|X_i = \bar{x}]\right).$$

Lemma 1 relates to Caetano (2015)'s test of exogeneity, which is based on the discontinuity of the outcome at the bunching point. If the sign of the discontinuity is not zero, then the test rejects the exogeneity of X_i. However, in our setting, we can do more than just test exogeneity, as the same discontinuity also allows us to sign the endogeneity within an interval of $\bar{x}$.

The intuition of Lemma 1 can be obtained from Figure 3. Suppose that we observe a positive discontinuity of the outcome at the bunching point. Then, at least some $\mathbb{E}[Y_i(\bar{x})|X_i^*]$ must be below $\mathbb{E}[Y_i(\bar{x})|X_i^* = \bar{x}^+]$. By Assumption 3, we know that actually all $\mathbb{E}[Y_i(\bar{x})|X_i^*]$ must be below $\mathbb{E}[Y_i(\bar{x})|X_i^* = \bar{x}^+]$ for $X_i < \bar{x}$, so $s(x) < 0$ for $x < 0$. We must therefore be in a situation akin to the left plot. It follows that $\mathbb{E}[Y_i(\bar{x})|X_i^*]$ for X_i^+ slightly above $\bar{x}$ are all above $\mathbb{E}[Y_i(\bar{x})|X_i^* = \bar{x}^+]$, and thus $s'(\bar{x}^+) > 0$.

3.2 Translating the identification problem into the bunching point

We next concern ourselves with the elimination of the unknown interval I from the formula of Theorem 2.2. More importantly, we need to substitute the density $f_{s(X)|I}$ with the density $f_{s(X^*)|X=\bar{x}}$, which may be identified, as we show in the following section.

Assumption 4. *The following hold:*

(i) $f_{X^*}(\bar{x})$ *exists and* $f_{X^*}(\bar{x}^+) = f_{X^*}(\bar{x})$.

(ii) $f_{s(X^*)}(s(x))$ *exists and is bounded for* $x \in (-\infty, \bar{x}]$.

(iii) The function $x \mapsto \mathbb{E}[Y_i(\bar{x})|X_i^* = x]$ *is continuous in* x *at* $\bar{x}$.

Item (i) of Assumption 4 together with Assumption 1 (i) state that X_i^* has a continuous density in $[\bar{x}, \bar{c} + \varepsilon_1)$, and $f_{X^*}(\bar{x})$ is positive. Item (ii) of Assumption 4 states that $s(X^*)$ has a density for X^* to the left of the bunching point. It is sufficient (but not necessary) for Assumption 4 (ii) that $x \mapsto \mathbb{E}[Y_i(\bar{x})|X_i^* = x]$ is monotonic and X^* has a density for $(-\infty, \bar{x}]$. More generally, if X^* has a density on $(-\infty, \bar{x}]$, then item (ii) holds provided that the set of x such that $s(x) = s$ has Lebesgue measure zero, for any $s \in s^{-1}((-\infty, \bar{x}])$. Thus item (ii) can be thought of as a consequence of X^* being continuously distributed on the left side of $\bar{x}$, requiring no restrictive assumptions on selection. We state conditions (i) and (ii) above because they are weaker.

Item (iii) of Assumption 4 implies that the observations with $X_i^* = \bar{x}$ are comparable to the observations with $X_i^* = \bar{x}^+$. In the smoking example, it is equivalent to saying that if the mothers who smoke very little (i.e. those with ρ_i slightly larger than $\bar{\rho}$) would stop smoking, their outcomes would be very similar to the outcomes of non-smoking mothers who are indifferent between not smoking and smoking a bit (i.e., those with $\rho_i = \bar{\rho}$). Note that we are not claiming that mothers near the bunching point are comparable to mothers at the bunching point, since the mothers at the bunching point can also include those who strictly prefer not to smoke (i.e. those with ρ_i much smaller than $\bar{\rho}$).

Assumption 4 (iii) implies that $\mathbb{E}[Y_i(\bar{x})|X_i^* = \bar{x}] = \mathbb{E}[Y_i(\bar{x})|X_i^* = \bar{x}^+]$ and thus $s(\bar{x}) = 0$. Moreover, it allows us to extend Assumption 2 to the interval $[\bar{x}, \bar{x} + \varepsilon_3)$, and note that, since $\lim_{x \downarrow \bar{x}} \frac{d}{dx}\mathbb{E}[Y_i(\bar{x})|X_i = x] \neq 0$, we have that $s'(\bar{x}) \neq 0$. We then obtain the following corollary of Theorem 2.2:

Corollary 3.1. *Given Assumptions 1-4:*

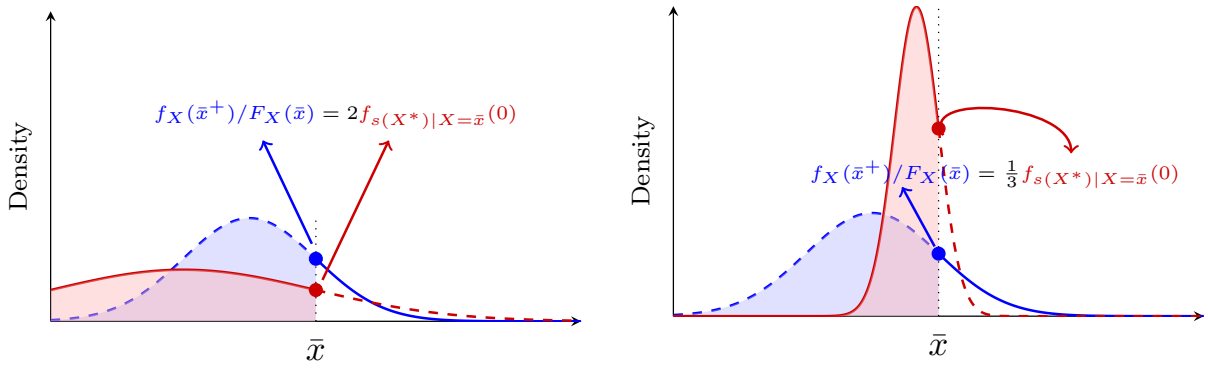
$$AME_{\bar{x}}^{+} = m'(\bar{x}^+) - \theta \cdot \frac{f_X(\bar{x}^+)/F_X(\bar{x})}{f_{s(X^*)|X=\bar{x}}(0)}. \tag{7}$$

Equation (7) shows that, under Assumptions 1-4, identifying $\text{AME}_{\bar{x}}^{+}$ reduces to the problem of identifying $f_{s(X^*)|X=\bar{x}}(0)$ and θ. Since θ is identified by Lemma 1 in Section 1, the only remaining piece is the density of $s(X_i^*)$ at the bunching point, which we tackle in the following section.

Figure 4 illustrates how we use the change-of-variables theorem in this corollary. This figure is identical to Figure 2, except that here we show as dashed lines the part of the densities that are not

observed, and the densities are conditional on $X_i = \bar{x}$. While $f_{X^*}(x)$ can be identified for $X_i^* \geq \bar{x}$, $f_{s(X^*)}$ can be identified for $X_i^* \leq \bar{x}$. Thus, the densities can be both identified only exactly at $X_i^* = \bar{x}$.

Figure 4: Using Corollary 3.1 to identify $s'(\bar{x})$



High Selection: $s(X_i^*) = 2(X_i^* - \bar{x})$ Low Selection: $s(X_i^*) = \frac{1}{3}(X_i^* - \bar{x})$

Note: This plot is identical to Figure 2, only it now refers to the selection variable X_i^*. Dashed lines then show the unobserved parts of the distributions. Note that only at $\bar{x}$ can we observe both densities. This illustration assumes $s(\bar{x}) = \bar{x}$, which is true when $\bar{x} = 0$, as in our application.

3.3 Identifying the distribution of $s(X^*)|X = \bar{x}$

Finally, we turn to the identification of $f_{s(X^*)|X=\bar{x}}$. Define the random variable

$$\epsilon_i = Y_i - \mathbb{E}[Y_i|X_i^*],$$

which is the unconfounded variation in the outcome, i.e. the idiosyncratic part of the outcome that remains after we eliminate the mean effect of treatment and the part determined by selection: $\mathbb{E}[Y_i|X_i^*] = \text{ATT}(X_i) + \mathbb{E}[Y_i(\bar{x})|X_i^*]$. Note that, for $X_i > 0$, ϵ_i is identified as $Y_i - \mathbb{E}[Y_i|X_i]$.

We can write $Y_i = y_0 + \text{ATT}(X_i) + s(X_i^*) + \epsilon_i$ with probability one, where the constant $y_0 := \mathbb{E}[Y_i(\bar{x})|X_i^* = \bar{x}]$ (using that $s(X_i^*) = \mathbb{E}[Y_i(\bar{x})|X_i^*] - y_0$ by Assumption 4). This shows that, when $X_i > \bar{x}$, we cannot disentangle the variations of $\text{ATT}(X_i)$ and $s(X_i^*)$. However, exactly at $X_i = \bar{x}$, $\text{ATT}(\bar{x}) = 0$, and thus

$$Y_i = y_0 + s(X_i^*) + \epsilon_i, \tag{8}$$

with probability one. Thus, the outcome variation at the bunching point reflects the variation of $s(X_i^*)$, unconfounded by the variation of the $\text{ATT}(X_i)$. Unfortunately, the distribution of the outcome at the bunching is a convolution of the density we want to identify, $f_{s(X^*)|X=\bar{x}}$ and the distribution of the idiosyncratic remainder, $y_0 + \epsilon_i$. The following conditions allow us to deconvolute these distributions.

Assumption 5. *The following hold:*

(i) $f_{Y|X=\bar{x}}$ *exists.*

(ii) $\epsilon_i|X_i^* = x \to_d \epsilon_i|X_i^* = \bar{x}$ *as* $x \downarrow \bar{x}$.

(iii) $\epsilon_i \perp\!\!\!\perp X_i^*|X_i = \bar{x}$.

Item (i) of Assumption 5 states that the outcome has a density at the bunching point. Item (ii) of Assumption 5 states that the idiosyncratic variation in the outcome near the bunching point is similarly distributed to the idiosyncratic variation in the outcome of those at the bunching point with $X_i^* = \bar{x}$. This weak continuity condition implies that $f_{\epsilon|X^*=\bar{x}} = f_{Y-\mathbb{E}[Y|X]|X=\bar{x}^+}$ is identified.

Note that $\epsilon_i = Y_i - \mathbb{E}[Y_i|X_i^*]$ is mean independent of X_i^* by construction. Item (iii) of Assumption 5 extends this into full independence, at least at the bunching point. In our application, this condition says that, among the non-smoking mothers, after removing the mean of the birth weight that is due to the selection variable X_i^*, the remainder is independent of the relative preference for smoking (or whatever else determines X_i^*). Item (iii) of Assumption 5 may be substituted with a weaker but less intuitive condition known as subindependence, which has long been used in the deconvolution literature (see discussion in e.g. Hamedani 2013), and was rigorously formalized in Schennach (2019). In our context, the subindependence condition translates to: for all $t \in \mathbb{R}$,

$$\mathbb{E}[e^{\mathbf{i}t(X_i^*+\epsilon_i)}|X_i = \bar{x}] = \mathbb{E}[e^{\mathbf{i}tX_i^*}|X_i = \bar{x}] \times \mathbb{E}[e^{\mathbf{i}t\epsilon_i}|X_i = \bar{x}],$$

where $\mathbf{i} = \sqrt{-1}$. Schennach (2019) shows that subindependence is no "stronger" than mean independence, in the sense that subindependence imposes the same number of restrictions on the data-generating process as mean independence does. Nevertheless, mean independence does not imply subindependence.

Lemma 2. *If Assumptions 2, 3, 4, and 5 hold, then* $f_{s(X^*)|X=\bar{x}}(0)$ *is identified as*

$$f_{s(X^*)|X=\bar{x}}(0) = \frac{1}{2\pi}\int \frac{\mathbb{E}[e^{\mathbf{i}\xi Y_i}|X_i = \bar{x}]}{\mathbb{E}[e^{\mathbf{i}\xi Y_i}|X_i = \bar{x}^+]}d\xi.$$

Proof. Conditional on $X_i = \bar{x}$, equation (8) holds, so:

$$\begin{aligned} F_{Y|X=\bar{x}}(y) &= \mathbb{P}(y_0 + s(X_i^*) + \epsilon_i \leq y|X_i = \bar{x}) \\ &= \int F_{s(X^*)|X=\bar{x},\epsilon=e}(y - e - y_0)dF_{\epsilon|X=\bar{x}}(e) \\ &= \int F_{s(X^*)|X=\bar{x}}(y - e - y_0)dF_{\epsilon|X=\bar{x}^+}(e), \end{aligned}$$

where the third equality follows from items (ii) and (iii) in Assumption 5 and Helly-Bray Theorem, due to the fact that $F_{s(X^*)|X=\bar{x}}(y - e - y_0)$ is bounded and continuous.

Then, by item (i) of Assumption 5, $F_{Y|X=\bar{x}}(y)$ is differentiable, and by item (ii) of Assumption 4, $F_{s(X^*)|X=\bar{x}}$ is differentiable. By the Dominated Convergence Theorem, we can write the convolution inverse problem

$$f_{Y|X=\bar{x}}(y) = \int f_{s(X^*)|X=\bar{x}}(y - e - y_0)dF_{\epsilon|X=\bar{x}^+}(e). \tag{9}$$

Equation (9) has a well-known closed form solution using the Fourier representation (see e.g. Schennach 2021).

$$f_{s(X^*)|X=\bar{x}}(v) = \frac{1}{2\pi}\int \frac{\mathbb{E}[e^{\mathbf{i}\xi Y_i}|X_i=\bar{x}]}{\mathbb{E}[e^{\mathbf{i}\xi(\epsilon_i+y_0)}|X_i=\bar{x}^+]}e^{-\mathbf{i}\xi v}d\xi = \frac{1}{2\pi}\int \frac{\mathbb{E}[e^{\mathbf{i}\xi Y_i}|X_i=\bar{x}]}{\mathbb{E}[e^{\mathbf{i}\xi Y_i}|X_i=\bar{x}^+]}e^{-\mathbf{i}\xi v}d\xi, \tag{10}$$

where the second equality follows because, for $X_i > \bar{x}$, $\epsilon_i = Y_i - \mathbb{E}[Y_i|X_i]$, and thus $\mathbb{E}[e^{\mathbf{i}\xi(\epsilon_i+y_0)}|X_i = x] = \mathbb{E}[e^{\mathbf{i}\xi(Y_i-\mathbb{E}[Y_i|X_i]+y_0)}|X_i = x] = e^{-\mathbf{i}\xi(\mathbb{E}[Y_i|X_i=x]-y_0)}\mathbb{E}[e^{\mathbf{i}\xi Y_i}|X_i = x]$. The equality then follows because $\mathbb{E}[Y_i|X_i = x]$ converges to y_0 as $x \downarrow 0$.

The result then follows if we evaluate the expression above at $v = 0$. □

Figure 5 illustrates the components of the deconvolution. The outcome distributions were specified as a sequence of normal distributions $f_{Y|X=x}$, depicted in solid blue. Since, for $X_i > 0$, $Y_i = \text{ATT}(X_i) + y_0 + \epsilon_i$, $f_{Y|X=x}$ converges to $f_{\epsilon|X=\bar{x}^+}$ as $x \downarrow \bar{x}$, the solid blue densities (which are observed) converge to the dotted blue density, which is equivalent to $f_{\epsilon|X=\bar{x}^+}$. This is how that distribution is obtained.

In the figure, we can also see the unobserved dashed red density of $s(X_i^*)$ at the bunching point, $f_{s(X^*)|X=\bar{x}}$, which is specified here as normal as well. The solid black line is the observed density of the outcome at the bunching point, which is the convolution of the red dashed density ($f_{s(X^*)|X=\bar{x}}$) and the blue dotted density ($f_{\epsilon|X=\bar{x}^+}$). Note that the images were produced using the actual convolution of the depicted densities, so the dimensions illustrate the real relationship between these densities.

Our final identification result is derived from the combination of Equation (7) with Lemmas 1 and 2:

Theorem 3.1. *Under Assumptions 1-5, $AME_{\bar{x}}^+$ is identified as:*

$$AME_{\bar{x}}^+ = \lim_{x\downarrow\bar{x}}\frac{d}{dx}\mathbb{E}[Y_i|X_i=x] - 2\pi\theta\left(\int \frac{\mathbb{E}[e^{\mathbf{i}\xi Y_i}|X_i=\bar{x}]}{\mathbb{E}[e^{\mathbf{i}\xi Y_i}|X_i=\bar{x}^+]}d\xi\right)^{-1}\cdot\frac{f_X(\bar{x}^+)}{F_X(\bar{x})}, \tag{11}$$

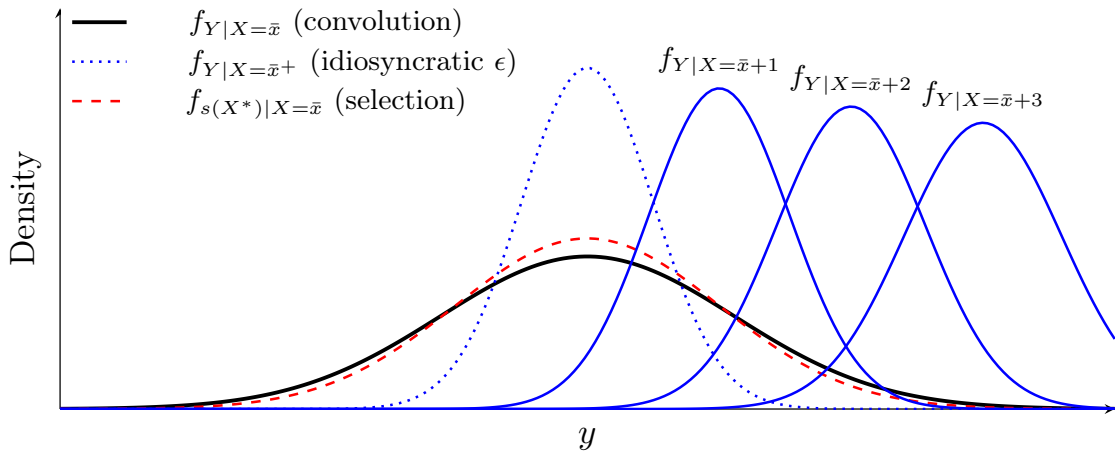
where $\theta = sgn\left(\mathbb{E}[Y_i|X_i=\bar{x}^+] - \mathbb{E}[Y_i|X_i=\bar{x}]\right)$.

The expression is familiar in that the treatment effect is calculated by correcting the outcome variation with a scaled inverse Mills Ratio term, as is usually seen in the censoring and sample selection literatures, where some of the model components are truncated or missing below a certain threshold.

Note that when there is no endogeneity at $X_i = \bar{x}$, $s'(\bar{x}) = 0$. Therefore, $\theta = 0$ and Equation (11) still holds. This means that the RHS of (11) can be used to identify $\text{AME}_{\bar{x}}^+$ while remaining agnostic about endogeneity.

In Appendix A, we show how identification may be obtained in the presence of controls. Specifically, all the assumptions required for identification may be done conditional on a vector of controls Z_i, so that the requirements may effectively be weaker. In particular, the treatment and selection

Figure 5: Components of the Deconvolution to Identify $f_{s(X^*)|X=\bar{x}}$



Note: The solid empirical distributions are observed. The solid blue lines depict the densities of the outcome, which approaches the density of $\epsilon|X_i = \bar{x}^+$ as $x \downarrow \bar{x}$, depicted by the dotted blue line. The dashed red line is the density of $s(X_i^*)|X_i = \bar{x}$, which is the object we want to identify. The solid black line is the observed density of $Y_i = X_i = \bar{x}$, which is the convolution of the dotted blue and the dashed red distributions. Plots were constructed with $\bar{x} = 0$, $Y_i|X_i = x \sim N(3500 + 625x, 312.5 + 20x)$, and $s(X_i^*)|X_i = \bar{x} \sim N(3500, 625)$. The density $f_{Y|X=\bar{x}}$ was produced by convoluting the distributions of $Y_i|X_i = \bar{x}^+$ and $s(X_i^*)|X_i = \bar{x}$.

effects may change direction for different subgroups, and we can identify the $\text{AME}_{\bar{x}}^+(Z_i)$ to study heterogeneous treatment effects. We propose estimators in the case with discrete controls (Section A.1), continuous controls (Section A.2), and when the vector of controls is large and with mixed continuous and discrete controls (Section A.3).

3.4 The selection variable X_i^*

While the existence of a selection variable X_i^* satisfying Equation (6) is without loss of generality, the identification result of Theorem 3.1 relies upon assumptions made about X_i^*. To guide researchers in assessing the plausibility of these assumptions, the examples below illustrate how X_i^* relates to well-defined quantities within empirically relevant choice models, both in settings with boundary bunching in Example 3.1, and in settings with interior bunching at a budget kink in Example 3.2. The selection variable X_i^* can in principle also be defined in the absence of any choice model being posited, given suitable assumptions regarding the distribution of $(X_i, \{Y_i(x)\}_{x \geq \bar{x}})$ alone. This is discussed in Example 3.3.

Example 3.1 (A parametric model of constrained choice)**.** *Consider the following parametric model where each individual chooses the number of units of a good or service to consume at price p, subject to the non-negativity constraints $x \geq 0$ and $r \geq 0$ as well as a budgetary constraint $W_i = px + r$, where W_i is the budget and r acts as the numeraire good. This model fits for instance the case of our application, where x refers to the choice of cigarettes smoked per day. Consider the family of*

utility functions:

$$V(x;\rho_i) = \begin{cases} (1+\rho_i)^\gamma \left(\frac{(1+x)^{1-\gamma}-1}{1-\gamma}\right) + (W_i - px), & \text{if } \gamma > 0,\ \gamma \neq 1 \\ (1+\rho_i)\log(1+x) + (W_i - px), & \text{if } \gamma = 1. \end{cases}$$

The parameter γ modulates the degree of concavity of the utility function. The individual-level parameter ρ_i can be interpreted as the preference for x relative to the numeraire good. For individuals with $\rho_i < -1$, x is seen as a "bad," while for individuals with $\rho_i > -1$, x is seen as a "good." In this particular family of utility functions, we have $X_i^* = a + b\rho_i$, where $a = (1/p^{(\frac{1}{\gamma})}) - 1$ and $b = 1/p^{(\frac{1}{\gamma})}$, so that X_i^* is a linear function of the primitive source of selection ρ_i.

If the price $p = 1$, then $a = 0$ and $b = 1$ and $X_i^* = \rho_i$, so the selection variable is exactly the preference parameter ρ_i. This is the case for instance in time use models, where x refers to the number of hours in the day spent on a given activity, such as watching TV (Caetano et al. 2023).[8] *In this example individuals with $-1 < \rho_i = X_i^* \leq 0$ are choosing $X_i = 0$ but still value watching TV positively, they just do not value it as highly as they value the alternative use of their time.*

In Proposition B.1 of Appendix B, we extend this example beyond the parametric utility of Example 3.1. Specifically, we show that in a large class of utility maximization models that feature a scalar preference parameter ρ_i, we can write $X_i^* = h(\rho_i)$ where h is a strictly increasing and differentiable function.

Then, in Proposition B.2, we abandon the utility maximization model and show that if $X_i^* = h(\rho_i)$ for any strictly increasing and differentiable h, the assumptions that establish Theorem 3.1 can be made directly on $\rho_i = h^{-1}(X_i^*)$, rather than on X_i^*. This means that any monotonic differentiable transformation of X_i^* yields the exact same identification results, This result is particularly useful in the context of a structural model of choice, where the researcher may be more comfortable making assumptions about the choice determinant ρ_i than about the more abstract object X_i^*. Proposition B.2 does not require that ρ_i is observable, nor that the function h be known to the econometrician.

In the following example, we discuss a class of models characterized by interior bunching at a kink in the decision-makers' choice sets. In such settings, there is no hard constraint that $X_i \geq \bar{x}$: rather, Equation (6) emerges from a discontinuous change in individuals' incentives at $X_i = \bar{x}$:

Example 3.2 (Interior bunching at a kink)**.** *Consider the treatment variable x as an individual's taxable income. Suppose that utility $u(x, t; A_i)$ is decreasing in t, and strictly quasi-concave in x, for each vector of individual characteristics A_i, which do not need to be observed by the researcher. Suppose that t as a function of x exhibits a convex kink at $\bar{x}$, so that costs increase faster with x when $x > \bar{x}$ than they do when $x < \bar{x}$. Goff (2023) shows that in this setting optimal choice can be*

[8] In time use choice models, r refers to the remaining activities adding up to $W_i = 24$ hours per day, so that $x + r = 24$. In this setting, the budget W_i is constant, and $p = 1$ because individuals trade-off x and r at the rate 1-to-1.

written as

$$\tilde{X}_i = \begin{cases} X_i(0) & \text{if } X_i(0) < \bar{x} \\ \bar{x} & \text{if } X_i(1) \leq \bar{x} \leq X_i(0) \\ X_i(1) & \text{if } X_i(1) > \bar{x} \end{cases}$$

where $X_i(0)$ and $X_i(1)$ are counterfactual choices that the individual would make if the budget function to the left of the kink applied globally, or if the budget function to the right of the kink applied globally, respectively.

If we consider only the observations with $X_i \geq \bar{x}$, then the mapping from Equation (6), $X_i = \max\{X_i^*, \bar{x}\}$, holds by defining $X_i^* = X_i(1)$. Thus, X_i^* could be interpreted as the counterfactual choice that would be made if the budget function to the right of the kink applied globally. Conversely, if we consider only the observations with $X_i \leq \bar{x}$, we would have $X_i = \min\{X_i^*, \bar{x}\}$ by defining $X_i^* = X_i(0)$, and X_i^* could be interpreted as the counterfactual choice that would be made if the budget function to the left of the kink applied globally.

We can also explicitly construct X_i^* without requiring any underlying choice model, referring only to potential outcomes and treatment values to the right of $\bar{x}$ as primitives. Specifically, if $s(x)$ is sufficiently smooth on the positive side, it can be extrapolated into the negative side, and thus X_i^* can be artificially constructed so as to satisfy the identification restrictions.

Example 3.3 (Bunching without choice model)**.** *Suppose that $s(x)$ is an analytic function for $x > \bar{x}$, and $\epsilon_i \perp\!\!\!\perp X_i | X_i > \bar{x}$. For the latter condition, the researcher can harness evidence on the basis of observable data (see e.g., the discussion of Figure 7 in Section 6). We show in Appendix C that, under these conditions, one can always define an X_i^* such that Equation (6) and $\epsilon_i \perp\!\!\!\perp X_i^* | X_i^* \leq \bar{x}$ hold. Specifically, the original probability space for $(X_i, \{Y_i(x)\}_{x \geq \bar{x}})$ can be replaced by a probability space for $(X_i^*, \{Y_i(x)\}_{x \geq \bar{x}})$ in which Equation (6) and $\epsilon_i \perp\!\!\!\perp X_i^* | X_i^* \leq \bar{x}$ hold.*

This type of construction of X_i^* is particularly useful in cases where X_i is not a choice variable. As discussed in the introduction, bunching has been observed in many examples where the treatment variable is not a clear function of individual choices. For instance, Caetano and Maheshri (2018) study of the effects of crime, and define X_i^* as an unknown index of unobserved factors that may lead some neighborhoods to have more crime than others. Neighborhoods often bunch at zero crimes per week, and yet some neighborhoods at $X_i = 0$ are only somewhat safe (i.e., they would have some crime every once in a while) while others are very safe (i.e., they would almost never have any crime), so they may have different values of the selection variable X_i^*. In any case, the amount of crime on a given week is not a choice variable of a specific individual.

4 Identification of causal effects away from the bunching point

Given identification of $\text{AME}_{\bar{x}}^{+}$ demonstrated in Section 3, we consider now if global effects $\text{ATT}(x)$ may be identified by extrapolating the information available near the bunching point. The extension

requires a sufficient degree of smoothness of the counterfactual function near $\bar{x}$, which is guaranteed by the following assumption.

Assumption 6. *The counterfactual function $x \mapsto \mathbb{E}[Y_i(\bar{x})|X_i = x]$ is real analytic on an interval $I = (\bar{x}, \bar{x} + \varepsilon_5)$ for some $\varepsilon_5 > 0$.*

Functions that are real analytic on I are infinitely differentiable functions for which the Taylor series around a given point $x \in I$ converges pointwise to the value of the function in a neighborhood of x. The class includes all functions that behave locally as, to give some examples: polynomial, exponential, trigonometric, hyperbolic, logarithmic, or inverse trigonometric functions, as well as composite, ratios, and roots of these. A sufficient condition for Assumption 6 is that the observable function $\mathbb{E}[Y_i|X_i = x]$ and the $\text{ATT}(x)$ function are both analytic on I. This may be an appealing argument if the researcher finds it plausible to assume that the dose response function $\text{ATT}(x)$ is sufficiently smooth, without needing to reason about properties of the selection function $s(x)$.

Recall that $\text{ATT}(x) = m(x) - s(x)$. The following theorem establishes the desired local extrapolation.

Theorem 4.1. *(Local ATTs) If Assumptions 1, 2, and 6 hold, then there exists an $\varepsilon > 0$ such that for all $x \in I_\varepsilon := (\bar{x}, \bar{x} + \varepsilon) \subset I$:*

$$ATT(x) = m(x) - \sum_{k=1}^{\infty} s^{(k)}(\bar{x}^+) \cdot \frac{(x - \bar{x})^k}{k!}, \tag{12}$$

where all the derivatives and limits in the equation above are well defined. Moreover, for any $K \geq 0$ there exists a value $\zeta_{\bar{x}}(x) \in (\bar{x}, x]$ such that,

$$R_k(x - \bar{x}) := \sum_{k=K+1}^{\infty} s^{(k)}(\bar{x}^+) \cdot \frac{(x - \bar{x})^k}{k!} = s^{(k)}(\zeta_{\bar{x}}(x)) \cdot \frac{(x - \bar{x})^{K+1}}{(K+1)!}.$$

The second part of Theorem 4.1 offers a practical strategy for finite approximations to Equation (12). For a suitably large K: $\text{ATT}(x) \approx m(x) - \sum_{k=1}^{K} s^{(k)}(\bar{x}^+) \cdot \frac{(x-\bar{x})^k}{k!}$. The error of the K-th degree approximation does not exceed $\sup_{x \in I_\varepsilon} |s^{(K+1)}(x)| \cdot \varepsilon^{K+1}/(K+1)!$. Since $(K+1)!$ has supra-exponential growth, even if ε and the high-order derivatives are very large, the approximation error decays quickly.

Equation (12) indicates that, if all the derivatives $s^{(k)}(\bar{x}^+)$ are identifiable, then the $\text{ATT}(x)$ is identifiable in a neighborhood of $\bar{x}$. This implies that extrapolations near the bunching point are possible, provided θ is identified and the $s^{(k)}(\bar{x}^+)$ are known for all $k \geq 1$. By differentiating Equation (5) with respect to x, we can see that this is indeed possible since the density $f_{s(X)|I}$ (and hence its derivatives) is identified on $s(I_\epsilon)$.

Corollary 4.1. *If Assumptions 1-6 hold, then $ATT(x)$ is identified for each $x \in I_\epsilon$.*

Corollary 4.1 guarantees the identification of the $\text{ATT}(x)$ in a neighborhood of the bunching point. In practice, finite approximations may be used to approximate the value. For example, a

first-degree approximation is simply

$$\text{ATT}_{\bar{x}}(x) \approx \mathbb{E}[Y_i|X_i = x] - \mathbb{E}[Y_i|X_i = \bar{x}^+] - 2\pi\theta \left(\int \frac{\mathbb{E}[e^{\mathbf{i}\xi Y_i}|X_i = \bar{x}]}{\mathbb{E}[e^{\mathbf{i}\xi Y_i}|X_i = \bar{x}^+]} d\xi \right)^{-1} \frac{f_X(\bar{x}^+)}{F_X(\bar{x})} (x - \bar{x}),$$

and a second-degree approximation adds the term

$$\left(s'(0) \frac{f'_X(\bar{x}^+)}{f_X(\bar{x}^+)} - s'(0)^2 \frac{f'_{s(X^*)|X=\bar{x}}(0)}{f_{s(X^*)|X=\bar{x}}(0)} \right) \frac{(x - \bar{x})^2}{2},$$

where $s'(0)$ is the correction term in Equation (11), and $f'_{s(X^*)|X=\bar{x}}(0)/f_{s(X^*)|X=\bar{x}}(0)$ is equal to $\left(\int \mathbb{E}[e^{\mathbf{i}\xi Y_i}|X_i = \bar{x}^+]^{-1}\mathbb{E}[e^{\mathbf{i}\xi Y_i}|X_i = \bar{x}]d\xi\right)^{-1} \left(\int \mathbf{i}\xi\mathbb{E}[e^{\mathbf{i}\xi Y_i}|X_i = \bar{x}^+]^{-1}\mathbb{E}[e^{\mathbf{i}\xi Y_i}|X_i = \bar{x}]d\xi\right)$. This expression may seem complex, but in practice, one would have already identified $f_{s(X^*)|X=\bar{x}}(0)$ for a first-degree approximation, and standard deconvolution packages automatically provide the first derivative $f'_{s(X^*)|X=\bar{x}}(0)$ at the same time.

One limitation of Theorem 4.1 and Corollary 4.1 is that the interval I_ε could in principle be quite small. A sufficient condition for the ATT to be defined far away from the bunching point is that the higher order derivatives decay suitably fast with k.

Corollary 4.2. *Suppose Assumptions 1-6 hold and* $\limsup_{k\to\infty} \left|\frac{s^{(k)}(\bar{x})}{k!}\right|^{1/k} < 1/M$, *then ATT*$(x)$ *is identified for all* $x \in [\bar{x}, \bar{x} + M]$.

Corollary 4.2 specifies "how far" one can extrapolate from the derivatives of $s^{(k)}(\bar{x})$ to obtain $\text{ATT}(x)$. Specifically, if the $|s^(k)(\bar{x})|$ are bounded by $M^{-k} \cdot k!$ uniformly over k for some M, then the $\text{ATT}(x)$ identification can be extrapolated as far as $\bar{x} + M$.

5 Estimation

For a sample $\{(Y_i, X_i)', i = 1, \ldots, n\}$, the average marginal effect at the bunching point may be estimated following Equation (11). Specifically, we use the following formulas:

$$\widehat{\text{AME}}^+_{\bar{x}} = \hat{m}'(\bar{x}^+) - \hat{\theta} \cdot \hat{f}_{s(X^*)|X=\bar{x}}(0)^{-1} \cdot \frac{\hat{f}_X(\bar{x}^+)}{\hat{F}_X(\bar{x})},$$

where $\hat{m}'(\bar{x}^+)$ is an estimator of $\lim_{x\downarrow\bar{x}} \frac{d}{dx}\mathbb{E}[Y_i|X_i = x]$, and $\hat{\theta} = \text{sgn}(\hat{\mathbb{E}}[Y_i|X_i = \bar{x}^+] - \hat{\mathbb{E}}[Y_i|X_i = \bar{x}])$.

A first-degree approximation following Corollary 4.1 uses the estimator

$$\widehat{\text{ATT}}(x) = \hat{E}[Y_i|X_i = x] - \hat{E}[Y_i|X_i = \bar{x}^+] - \hat{\theta} \cdot \hat{f}_{s(X^*)|X=\bar{x}}(0)^{-1} \cdot \frac{\hat{f}_X(\bar{x}^+)}{\hat{F}_X(\bar{x})} \cdot (x - \bar{x}),$$

and analogously for a second-degree approximation.

All the components of these formulas are standard objects frequently studied in econometrics. We discuss next how each component may be estimated.

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

$$\hat{F}_X(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i = \bar{x}), \text{ and } \hat{\mathbb{E}}[Y_i|X_i = \bar{x}] = \hat{F}_X(\bar{x})^{-1} \cdot \frac{1}{n}\sum_{i=1}^{n} Y_i \mathbf{1}(X_i = \bar{x}).$$

The terms $\hat{\mathbb{E}}[Y_i|X_i = \bar{x}^+]$ and $\hat{m}'(\bar{x}^+)$ 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 $X_i = \bar{x}$, using only observations such that $X_i > \bar{x}$, for its superior properties of bias reduction and variance control at the boundary over other methods.[9] The intercept coefficient of this regression is $\hat{\mathbb{E}}[Y_i|X_i = \bar{x}^+]$, and the slope coefficient is $\hat{m}'(\bar{x}^+)$. This may be executed using any package for local linear regression available in standard statistical software (R, STATA, etc.).

Explicitly, for a bandwidth $h_1 > 0$ and a kernel function k_1,[10] solve the problem

$$\hat{b}_0, \hat{b}_1 = arg\min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1(X_i - \bar{x}))^2 \cdot k_1\left(\frac{X_i - \bar{x}}{h_1}\right) \mathbf{1}(X_i > \bar{x}), \tag{13}$$

then $\hat{\mathbb{E}}[Y_i|X_i = \bar{x}^+] = \hat{b}_0$, and $\hat{m}'(\bar{x}^+) = \hat{b}_1$. This estimator has a closed-form expression, which is commonly found in nonparametric econometrics textbooks, e.g. Li and Racine (2007). Note that the term $\hat{E}[Y_i|X_i = x]$ is not a boundary quantity, but it may be estimated analogously, with a local linear regression of Y_i onto X_i at x, using only observations such that $X_i > \bar{x}$.

The term $\hat{f}_X(\bar{x}^+)$ 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),[11] 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 $L'_X(\bar{x}^+)$ as[12]

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

[9] 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.

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

[11] 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).

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

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

$$\hat{f}_X(\bar{x}^+) = \frac{\frac{1}{nh_2}\sum_{i=1}^n k_2\left(\frac{X_i-\bar{x}}{h_2}\right)}{\int_0^1 k_2(\nu)\exp(\hat{L}'_X(\bar{x}^+)\nu h_2)d\nu},$$

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 kernel recommended for boundary estimation in that paper) the denominator is equal to

$$\frac{3}{2}\cdot\frac{2+\hat{L}'_X(\bar{x}^+)^2h_2^2 - e^{\hat{L}'_X(\bar{x}^+)h_2}(2-2\hat{L}'_X(\bar{x}^+)h_2)}{\hat{L}'_X(\bar{x}^+)^3h_2^3}.$$

This estimator is available in packaged form in standard statistics software and can be implemented by simply restricting the sample to observations such that $X_i > \bar{x}$ and then using the package to estimate the density of X_i at $X_i = \bar{x}$. Incidentally, the same package also provides the estimator of the derivative $f'_X(\bar{x}^+)$, that can be used in the second-order approximation of the ATT(x) (see Section 4).

The final term $\hat{f}_{s(D^*)|X=0}(0)$ is a standard deconvolution estimator. We follow the estimator 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 $\hat{\mathbb{E}}[e^{\mathbf{i}\xi Y_i}|X_i = \bar{x}^+]$ as a local linear regression of $e^{\mathbf{i}\xi Y_i}$ onto X_i at $X_i = \bar{x}$ using only observations such that $X_i > \bar{x}$. To do this, for a matrix $\mathbf{x}$ with rows $(1,(X_i-\bar{x}))'$ and a diagonal matrix $\mathbf{k}$, with diagonal elements $k_3((X_i-\bar{x})/h_3)\mathbf{1}(X_i>\bar{x})$, where k_3 is the triangular kernel, and $\mathbf{e}_1 = (1,0)'$, define the vector $A(\xi) = (e^{\mathbf{i}\xi Y_1},\ldots,e^{\mathbf{i}\xi Y_n})'$, 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:

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

with

$$g(Y_i) = \frac{1}{\hat{F}_X(\bar{x})\cdot 2\pi}\int e^{\mathbf{i}\xi Y_i}\frac{\phi_K(h_4\xi)}{A(\xi)}d\xi,$$

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.[13] However, it is possible that the optimal

[13]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

bandwidths for $\widehat{\text{AME}}_{\bar{x}}^{+}$ 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, $\widehat{\text{AME}}_{\bar{x}}^{+}$ will likely be asymptotically biased. We leave these questions for future research.

Remark 5.1. *(Improving efficiency by parameterizing the outcome distributions) We can estimate $\hat{f}_{s(X^*)|X=\bar{x}}(0)$ more efficiently via deconvolution if either of the distributions $f_{Y|X=\bar{x}}$ or $f_{Y|X=\bar{x}^+}$ are assumed to be of a known parametric family. Since the outcome is observed both at and above the bunching point, this assumption may not be overly speculative, and it is directly testable using standard Kolmogorov-Smirnoff tests or Goldman and Kaplan (2018).*

Assuming a parametric distributional class for the outcome is helpful because it allows the relevant characteristic function to be estimated less noisily. For example, in the empirical analysis in Section 6, we assume that $Y_i|X_i = \bar{x}^+ \sim N(\mathbb{E}[Y_i|X_i = \bar{x}^+], \sigma^2)$, while we allow $f_{Y|X=\bar{x}}$ to be fully nonparametric. This normality assumption appears to be a good approximation, as can be seen in Figure 7.

We estimate the parameter σ^2 in two steps. First, we restrict the sample to observations with $X_i > \bar{x}$, and for each i, we predict $\hat{Y}_i$ via local linear regression estimated on the non-bunched sample. Then, we form the squared residual $\hat{\epsilon}_i^2 = (Y_i - \hat{Y}_i)^2$. Second, we fit a local linear model of $\hat{\epsilon}_i^2$ on X_i estimated on the $X_i > \bar{x}$ subsample, and we estimate $\hat{\sigma}^2$ as the predicted value of this regression at $X = \bar{x}$. We note that the variance of $\epsilon_i|X_i = \bar{x}^+$ is the same as the variance of $Y_i|X_i = \bar{x}^+$. We then deconvolve the distribution $N(0, \hat{\sigma}^2)$ from $\hat{F}_{Y|X=\bar{x}}$ using standard methods.[14]

6 Application: The Effect of Maternal Smoking on Birth Weight

In this section, we apply our method to estimate the marginal effect of smoking while pregnant on the baby's birth weight. This question holds significant importance in both economics and epidemiology, given that maternal smoking during pregnancy is recognized as a critical modifiable risk factor for low birth weight (Almond et al. 2005), which not only results in immediate societal costs but is also widely considered to be consequential for outcomes later in the child's life (Black et al. 2007).

Our application uses the data set from Almond et al. (2005), which is also used in Caetano (2015).

packages. There are many proposals for improvement of 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 $\hat{f}_X(0)_+$, 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).

[14] Specifically, we employ the "decon" package in R.

These data, from the U.S. National Center for Health Statistics, have available both maternal cigarettes smoked daily during pregnancy (our treatment) and birth weight in grams (our outcome) for over 430,000 mother-child pairs. These data also contain many additional covariates/controls which Almond et al. (2005) use to estimate a selection-on-observables effect of maternal smoking on birth weight of around -200 grams. Caetano (2015) uses these data to illustrate the discontinuity test, showing that selection-on-observables does not seem to be a valid assumption using Almond et al. (2005)'s very detailed specification.

We drop premature births (gestation $<$ 36 weeks) as well as birth weight outliers (a small subset of observations with very high, $>$6 kg, or very low, $<$1 kg, full-term birth weights) from the data.[15] In our analysis sample, about 81% of the mothers smoke zero cigarettes daily, while about 11% smoke between 1 and 10 cigarettes, with 99.95% of the sample smoking 40 cigarettes or less. Figure 6 shows $\mathbb{E}[Y_i|X_i = x]$, the average birth weight among mothers smoking different amounts in our sample. The evidence of a discontinuity in $\mathbb{E}[Y_i|X_i = x]$ at $x = 0$ is clear. While the average birth weight among mothers who smoke zero cigarettes is 3,499 grams, the average birth weight for mothers who smoke one cigarette is 3,338 grams, with the analogous quantity for mothers smoking 2–5 cigarettes ranges between 3,278-3,330. The rich controls in Almond et al. (2005) can account for only 55 out of the 161 gram (3,499-3,338) difference in birth weight between the children of mothers who smoke zero versus one cigarettes. Thus, there remains a lot of "room"—106 grams—for both the treatment effect of cigarettes and selection on unobservables to explain this birth weight difference.[16]

We estimate the marginal effect AME_0^+ of cigarette smoking on birth weight using the procedure outlined in Remark 5.1. In particular, as discussed in Remark 5.1, we increase efficiency by assuming that $Y_i|X_i = \bar{x}^+$ is normally distributed. This assumption is empirically well-supported in our setting; Figure 7 makes clear that the conditional distributions $Y_i|X_i = x$ are all very close to normal with nearly the same variance for $0 < x \leq 5$.[17] In addition to simplifying estimation, Figure 7 can also be interpreted as indirect evidence for the assumption that $\epsilon_i \perp\!\!\!\perp X_i^*|X_i = 0$. Specifically, because $f_{Y|X=x} = f_{\epsilon+\mathbb{E}[Y|X=x]|X}$, the distribution of $\epsilon_i|X_i = x$ is just a horizontal shift of the distribution of $Y_i|X_i = x$. Thus, if the $f_{Y|X=x}$ look like simple horizontal shifts for $0 < X_i \leq 5$, this is evidence of $\epsilon_i \perp\!\!\!\perp X_i|0 < X_i \leq 5$. This is thus indirect evidence that this pattern may continue for $X_i^* \leq 0$, though this cannot be verified.

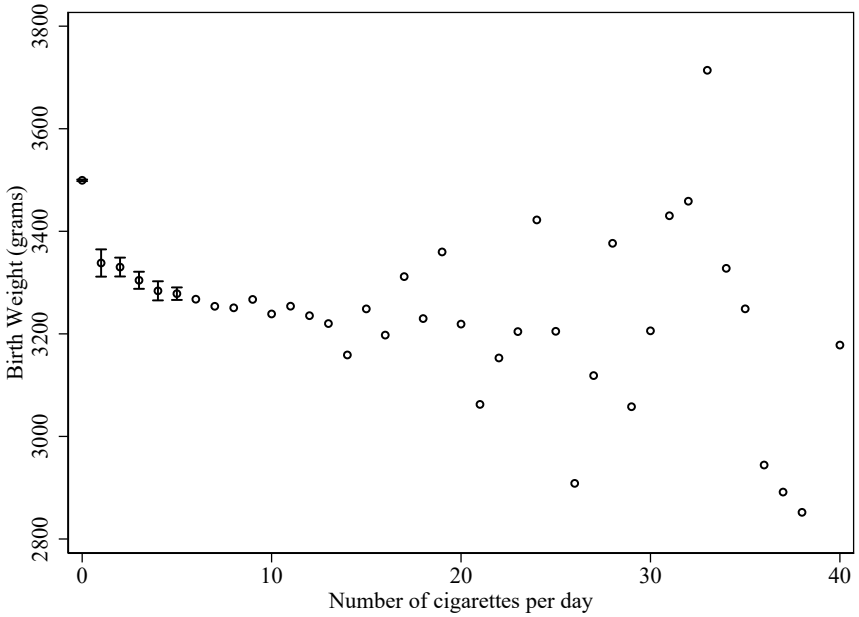
For context, we also show the conditional distribution $Y|X = 0$. While we estimate the variance of $Y_i|X_i = \bar{x}^+$ allowing for a trend in the analogous variances of $Y_i|X_i = x$ for $x > 0$, we obtain nearly identical results simply using the variance of $Y|X = 1$ as our estimator. This is not surprising given the evident homoskedasticity in Figure 7.

[15] Our estimates barely change when we extend the sample to include premature babies and outliers.

[16] The numbers cited in this paragraph are very similar to those reported in Caetano (2015). That paper removes neither premature nor "outlier" births from the analysis sample. Similarly, including premature and outlier births in the causal analysis yields similar estimates of AME_0^+ and $\text{ATT}(x)$. Indeed, the full-sample results tend to be smaller in magnitude than those reported in Table 1 and Figure 8, further supporting our qualitative claims.

[17] Figure 9 in Appendix D presents QQ-plots as additional graphical evidence that these conditional distributions are approximately normal.

Figure 6: Evidence of Bunching and Discontinuity



Note: The figure shows $\mathbb{E}[Y_i|X_i = x]$ for different values of x (along with the 95% confidence interval for $x \leq 5$).

Estimation requires that we select several different bandwidths. We use a bandwidth of 4 cigarettes for both $\hat{\mathbb{E}}[Y_i|X_i = 0^+]$ and $\hat{f}_X(0^+)$. However, we find very similar results using alternative, reasonable bandwidth choices for these objects. The selection of a bandwidth for $\hat{m}'(0^+)$ is more consequential for the standard error of our estimates, thus we report $\widehat{\text{AME}}_0^+$ for several bandwidth choices.

Table 1 presents our main results. We estimate the marginal effect of smoking on birth weight at zero to be around -8 grams. We estimate $s'(0^+)$, the endogeneity bias around zero, to also be around -8 grams. This endogeneity term is quite precisely estimated – most of the sampling variation in $\widehat{\text{AME}}_0^+$ comes from sampling variation in the estimated slope of $\mathbb{E}[Y_i|X_i = 0^+]$.

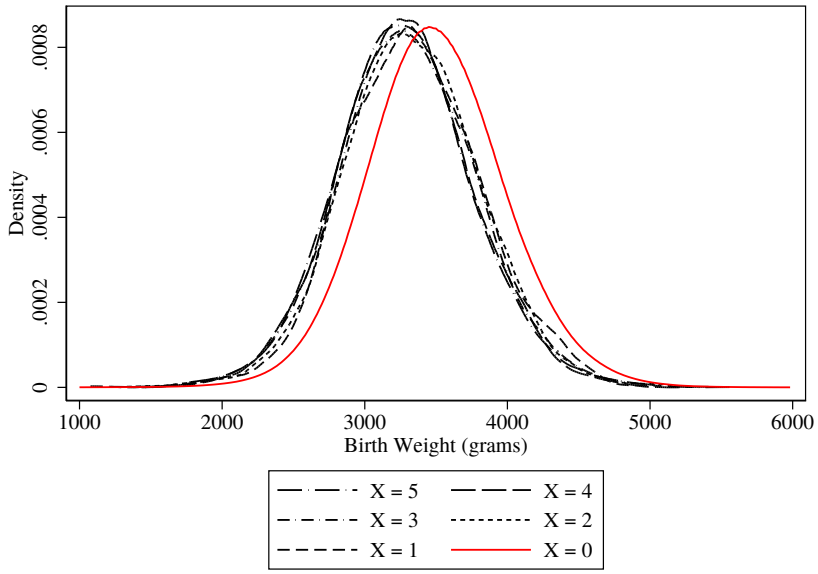
Table 1: Main Results: The Average Marginal Effect of Smoking Near Zero Cigarettes

	$\hat{m}'(0^+)$ per bandwidth				
	$h = 4$	$h = 5$	$h = 6$	$h = 7$	$h = 8$
$\widehat{\text{AME}}_0^+$	-8.15	-10.55	-8.61	-7.76	-7.21
	(8.10)	(5.59)	(3.22)	(2.73)	(2.38)

Note: X_i is measured in cigarettes per day, and Y_i is measured in grams. The bandwidths for $\hat{\mathbb{E}}[Y_i|X_i = 0^+]$ and $\hat{f}_X(0^+)$ are both set to 4. The estimate for $s'(0^+)$ used for each of the displayed bandwidths is -8.36. Standard errors based on 2,500 bootstrap iterations. Data taken from Almond et al. (2005).

As shown in Section 4, if we can make a local extrapolation, then we can further recover $\text{ATT}(x)$ for positive x. We use the first-degree ATT approximation formula in Section 5, and present the

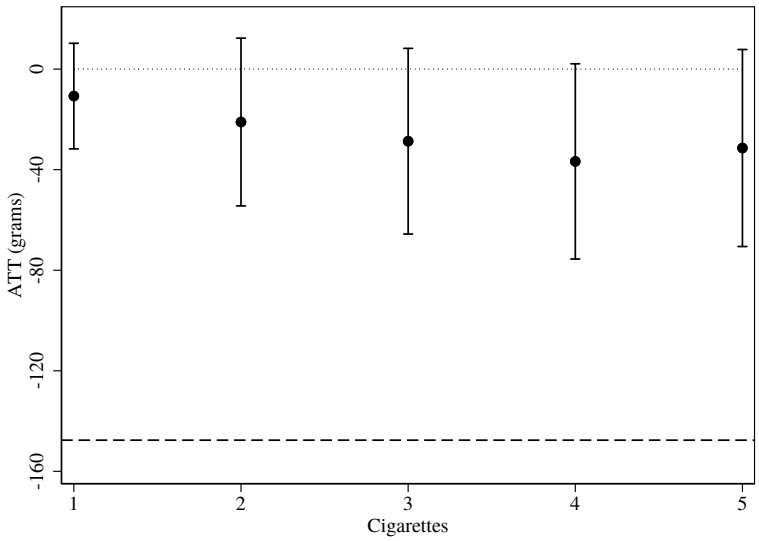
Figure 7: Birth Weight Distributions Conditional on Maternal Smoking



Note: The figure plots the kernel density estimates (Epanechnikov, bandwidth=100) of birth weight conditional on maternal cigarettes smoked per day (X).

estimates in Figure 8 for $x \leq 5$.

Figure 8: ATT Estimates on Birth Weight for Different Levels of Maternal Smoking



Note: This figure reports $\text{ATT}(x) = \mathbb{E}[Y_i(x) - Y_i(0) | X_i = x]$ for different values of x, represented in the horizontal axis. It uses the same bandwidths as in Table 1. The $\mathbb{E}[Y_i|X_i = x]$ estimates are smoothed using a local linear polynomial with triangular kernel and bandwidth of 6. 95% confidence intervals based on 2,500 bootstrap iterations shown. The dashed line at -147.6 g is the estimated discontinuity in birth weight at $X = 0$ (i.e., $\hat{\mathbb{E}}[Y_i|X_i = 0^+] - \hat{\mathbb{E}}[Y_i|X_i = 0]$). Data taken from Almond et al. (2005).

Our estimates show small negative ATT's. For instance, if mothers who currently smoke one

cigarette were to quit smoking, their babies would gain about 8 grams at birth. The effects of quitting smoking add up to a gain of only about 1 ounce for mothers who smoked 5 cigarettes per day. The estimates are not significant at standard levels, suggesting that the effect of maternal smoking on birth weight is small, and we can rule out effects larger than 3 ounces at standard levels of significance. For reference, in 2025 the average birth weight of a healthy full term baby in the U.S. is about 119 ounces (7lb and 6 ounces). Indeed, these estimates are much smaller than the effect implied by the discontinuity of the outcome at $X = 0$ shown in the dashed line of Figure 8. Our estimates support the qualitative point in Almond et al. (2005) that smoking seems to have only small effects on birth weight, although our findings suggest even smaller effects than in that paper.

7 Concluding remarks

When a treatment variable has bunching, 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 which does not make assumptions on functional forms or on shape of the distribution of the unobservables. Since the method does not rely on exclusion restrictions or special data structures, it provides a new avenue for the 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 selection function at the bunching point selection value. Intuitively, those who chose the bunching point as an interior solution (i.e., not as a corner solution) are comparable to those right above the bunching point. Besides this and other regularity conditions, the method also requires two other conditions that are hard to explain succinctly, but are implied if the selection equation is monotonic on the selection variable and if the idiosyncratic errors (which are by definition mean independent from the selection variable) are independent from the selection variable at the bunching point.

Identification is achieved by the comparison of the density of the treatment near the bunching point (observed on the positive side) and the distribution of the selection function at the bunching point (identifiable 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 magnitude of the endogeneity bias.

The approach results in the identification of the average marginal 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 expected outcome, and (3) the derivative of the expected outcome. The final term is the density of the selection variable at the bunching point, which is obtained through a deconvolution of the outcome near the bunching point from the outcome at the bunching point. All the terms in the identification equations can be estimated with off-the-shelf methods readily

available in package form in all standard statistical software.

We apply the method to the estimation of the effect of smoking during pregnancy on the baby's birth weight. Our results show that the effects are rather small, strengthening the qualitative results in the previous literature.

There is ample opportunity for further technical advancements that would enhance the applicability of this method. Of note, there seems to be a scarce supply of options for estimation of boundary derivatives in the literature. Even for local polynomial estimators (Fan et al., 1996), the optimal degree, kernel and bandwidths for estimation of boundary derivatives remain unknown. Deconvolution estimators are ubiquitous in other fields, but still rare in economics, with the notable exception of the measurement error literature (see e.g. Schennach 2021). The application of deconvolution to bunching is more aligned with the classical setting, where the distributions of the "recorded signal" and the "distortion" are identified, but the behavior of deconvolution estimators with boundary plugins is largely unexplored.

References

Almond, D., Chay, K. Y., and Lee, D. S. (2005). The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083.

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., Caetano, C., Jales, H., and Seegert, N. (2023a). Bunching estimation methods. *Handbook of Labor, Human Resources, and Population Economics (forthcoming). Springer.*

Bertanha, M., McCallum, A. H., and Seegert, N. (2023b). Better bunching, nicer notching. *Journal of Econometrics*, 237(2):105512.

Black, S. E., Devereux, P. J., and Salvanes, K. G. (2007). From the cradle to the labor market? the effect of birth weight on adult outcomes. *The Quarterly Journal of Economics*, 122(1):409–439.

Blomquist, S., Hausman, J. A., and Newey, W. K. (2023). The econometrics of nonlinear budget sets. *Annual Review of Economics*, 15:287–306.

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

Caetano, C., Caetano, G., and Nielsen, E. (2023). Correcting Endogeneity Bias in Models with Bunching. *Journal of Business & Economic Statistics*. Available here.

Caetano, C., Caetano, G., and Nielsen, E. (2024a). Are children spending too much time on enrichment activities? *Economics of Education Review*, 98:102503.

Caetano, C., Caetano, G., Nielsen, E., and Sanfelice, V. (2024b). The Effect of Maternal Labor Supply on Children: Evidence from Bunching. *Journal of Labor Economics*. Available here.

Caetano, C., Caetano, G., Nielsen, E., and Techio, O. (2024c). Partial Identification in Models with Bunching. *Working Paper*.

Caetano, C., Caetano, G., and Techio, O. (2024d). Identification of Causal Effects in Bunching Models with Endogeneity. *Working Paper*.

Caetano, C., Rothe, C., and Yıldız, N. (2016). A discontinuity test for identification in triangular nonseparable models. *Journal of econometrics*, 193(1):113–122.

Caetano, G. and Maheshri, V. (2018). Identifying dynamic spillovers of crime with a causal approach to model selection. *Quantitative Economics*, 9(1):343–394.

Callaway, B., Goodman-Bacon, A., and Sant'Anna, P. H. (2024). Difference-in-differences with a continuous treatment. Technical report, National Bureau of Economic Research.

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.

Chiang, H. D. and Sasaki, Y. (2019). Causal inference by quantile regression kink designs. *Journal of Econometrics*, 210(2):405–433.

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.

Fan, J. and Gijbels, I. (1992). Variable Bandwidth and Local Linear Regression Smoothers. *The Annals of Statistics*, 20(4):2008 – 2036.

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

Fan, J., Gijbels, I., Hu, T.-C., and Huang, L.-S. (1996). A study of variable bandwidth selection for local polynomial regression. *Statistica Sinica*, 6(1):113–127.

Fremlin, D. (2011). *Measure Theory, Volume 2*. Torres Fremlin.

Goff, L. (2023). Treatment effects in bunching designs: The impact of mandatory overtime pay on hours. Technical report, Working Paper.

Goff, L., Kédagni, D., and Wu, H. (2024). Testing identifying assumptions in parametric separable models: A conditional moment inequality approach.

Goldman, M. and Kaplan, D. M. (2018). Comparing distributions by multiple testing across quantiles or CDF values. *Journal of Econometrics*, 206(1):143–166.

Hamedani, G. G. (2013). Sub-independence: An expository perspective. *Communications in Statistics - Theory and Methods*, 42(20):3615–3638.

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.

Hoderlein, S. and Mammen, E. (2007). Identification of marginal effects in nonseparable models without monotonicity. *Econometrica*, 75(5):1513–1518.

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.

Imbens, G. W. and Newey, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512.

Jales, H. and Yu, Z. (2017). Identification and estimation using a density discontinuity approach. In *Regression Discontinuity Designs*, volume 38, pages 29–72. Emerald Group Publishing Limited.

Kato, R. and Sasaki, Y. (2017). On using linear quantile regressions for causal inference. *Econometric Theory*, 33(3):664–690.

Khalil, U. and Yıldız, N. (2022). A test of the selection on observables assumption using a discontinuously distributed covariate. *Journal of Econometrics*, 226(2):423–450.

Kleven, H. J. (2016). Bunching. *Annual Review of Economics*, 8:435–464.

Kleven, H. J. and Waseem, M. (2013). Using notches to uncover optimization frictions and structural elasticities: Theory and evidence from pakistan *. *The Quarterly Journal of Economics*, 128(2):669–723.

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.

Lu, Y., Wang, J., and Xie, H. (2024). Identifying causal effects under kink setting: Theory and evidence. *arXiv preprint arXiv:2404.09117*.

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.

Sasaki, Y. (2015). What do quantile regressions identify for general structural functions? *Econometric Theory*, 31(5):1102–1116.

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

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

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

Appendix

A Incorporating control variables

Suppose that in addition to X_i and Y_i, we observe a vector of control variables Z_i that are unaffected by treatment. Then, if Assumptions 1, 2, 3, 4, and 5 each hold conditional on Z_i,

$$\mathrm{AME}^{+}_{\bar{x}}(z) = m'(\bar{x}^{+}, z) - 2\pi \cdot \theta(z) \cdot \left(\int \frac{\mathbb{E}[e^{\mathbf{i}\xi Y_i} | X_i = \bar{x}, Z_i = z]}{\mathbb{E}[e^{\mathbf{i}\xi Y_i} | X_i = \bar{x}^{+}, Z_i = z]} d\xi \right)^{-1} \cdot \frac{f_{X|Z=z}(\bar{x}^{+})}{F_{X|Z=z}(\bar{x})},$$

where $\text{AME}^+_{\bar{x}}(z) := \lim_{x \downarrow \bar{x}} \mathbb{E}[(Y_i(x) - Y_i(\bar{x}))/(x - \bar{x}) | X_i = x, Z_i = z]$, $m'(\bar{x}^+, z) = \lim_{x \downarrow \bar{x}} \frac{d}{dx} \mathbb{E}[Y_i | X_i = x, Z_i = z]$, and $\theta(z) = \text{sgn}\left(\mathbb{E}[Y_i | X_i = \bar{x}^+, Z_i = z] - \mathbb{E}[Y_i | X_i = \bar{x}, Z_i = z]\right)$. Note then that we can identify the unconditional $\text{AME}^+_{\bar{x}}$ given that

$$\text{AME}^+_{\bar{x}} = \lim_{x \downarrow \bar{x}} \mathbb{E}\left[\mathbb{E}[(Y_i(x) - Y_i(\bar{x}))/(x - \bar{x}) | X_i = x, Z_i]\right] = \mathbb{E}\left[\text{AME}^+_{\bar{x}}(Z_i) | X_i = \bar{x}^+\right]$$

under suitable conditions to interchange the limit and the expectation.

A.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

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

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

The average marginal treatment effect estimator in this case is

$$\widehat{\text{AME}}^+_{\bar{x}} = \sum_{l=1}^{L} \hat{p}_{l,\bar{x}} \cdot \widehat{\text{AME}}^+_{\bar{x}}(z_l). \tag{15}$$

Where $\widehat{\text{AME}}^+_{\bar{x}}(z_l)$ is the estimator described in Section 5 applied to the subsample with $Z_i = z_l$. Note that it is not possible to estimate $\text{AME}^+_{\bar{x}}(z_l)$ when $\hat{p}_{l,\bar{x}} = 0$, but it is also not necessary to do so, since those treatment effects have weight equal to zero in the estimator formula.

A.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 $\text{AME}^+_{\bar{x}}(Z_i)$ 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.

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

perform the methods described for unconditional estimation, only weighting each observation by $k_\kappa(Z_t - Z_i) = K_\kappa(Z_t - Z_i)/\sum_{t=1}^T K_\kappa(Z_t - Z_i)$.[18]

Then,

$$\widehat{\text{AME}}^+_{\bar{x}} = \sum_{l=1}^{L} \hat{p}_{i,\bar{x}} \cdot \widehat{\text{AME}}^+_{\bar{x}}(Z_i).$$

As in the previous section, it is not necessary to compute $\widehat{\text{AME}}^+_{\bar{x}}(Z_i)$ when $\hat{p}_{i,\bar{x}} = 0$.

A.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.[19]

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 A.1.

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

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

Then, for those clusters with $\hat{p}_{c,\bar{x}} > 0$, estimate $\widehat{\text{AME}}^+_{\bar{x}}(\hat{\mathcal{C}}_c)$ 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,

$$\widehat{\text{AME}}^+_{\bar{x}} = \sum_{c=1}^{C} \hat{p}_{c,\bar{x}} \cdot \widehat{\text{AME}}^+_{\bar{x}}(\hat{\mathcal{C}}_c). \tag{17}$$

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

In general, if $\text{AME}^+_{\bar{x}}(z)$ is continuous in z, the ability of this estimator to approximate $\text{AME}^+_{\bar{x}}(z)$ 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

[18] Thus, $\hat{F}_{X|Z=Z_i}(\bar{x}) = \frac{1}{n_T}\sum_{t=1}^{n_T} \mathbf{1}(X_i = \bar{x}) k_\kappa(Z_t - Z_i)$, and $\hat{E}[Y_i|X_i = \bar{x}, Z_i] = \hat{F}_{X|Z=Z_i}(\bar{x})^{-1} \cdot \frac{1}{n_T}\sum_{t=1}^{n_T} Y_i \mathbf{1}(X_i = \bar{x}) k_\kappa(Z_t - Z_i)$, $\hat{\mathbb{E}}[Y_i|X_i = \bar{x}^+, Z_i]$. The densities $\hat{f}_{X|Z=Z_i}(\bar{x}^+)$ and $\hat{f}_{Y|X=\bar{x},Z_i}(Y_i)$ 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 by $k_\kappa(Z_t - Z_i)$. Finally, $\hat{\mathbb{E}}[Y_i|X_i = \bar{x}^+, Z_i]$ and $\hat{m}'(\bar{x}^+, Z_i)$ 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 $X_t > \bar{x}$ and weights $k_\kappa(Z_t - Z_i)$; and $\hat{\mathbb{E}}[\hat{f}_{Y|X=\bar{x},Z=Z_i}(Y_i)|X_i = \bar{x}, Z_i]$ is the intercept of the same procedure, only with $\hat{f}_{Y|X=\bar{x},Z=Z_i}(Y_i)$ instead of Y_i.

[19] See, e.g. Bonhomme and Manresa (2015); Bonhomme et al. (2017); Cheng et al. (2019); Cytrynbaum (2020); Caetano et al. (2023, 2024a,b).

be used (e.g. k-means, k-medoids, self-organizing maps, and spectral – see Hastie et al. (2009)). If feasible, we recommend using hierarchical clustering for its well-known stability.[20]

The clustering strategy requires the choice of the number of clusters, which modulates the bias-variance trade-off in the estimation of $\text{AME}^+_{\bar{x}}(z)$. 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 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 $\text{AME}^+_{\bar{x}}(z)$ but rather in $\widehat{\text{AME}}^+_{\bar{x}}$, which aggregates the information over all clusters. The trade-off is, in theory, much less important for $\widehat{\text{AME}}^+_{\bar{x}}$, 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).

B Theorem 3.1 with a generalized latent selection variable

In this section we first show that in settings where X_i is constrained to be greater than $\bar{x}$, a large class of choice models yields a constructive expression for the selection index X_i^* that determines X_i. In particular, we obtain a choice rule $X_i = \max\{h(\rho_i), \bar{x}\}$ for a strictly increasing and differentiable function h, and heterogeneity parameter ρ_i, which is equivalent to Eq. (6) with $X_i^* = h(\rho_i)$. Since such settings arise naturally due to physical positivity constraints (the quantity of a good consumed, the time spent on an activity, etc.), we take $\bar{x} = 0$ for simplicity. This is however without loss of generality since we can always redefine X_i as $X_i - \bar{x}$.

In this class of examples we suppose individuals have two scalar choice variables x and r, with individual i's utility denoted as $V(x, r; \theta_i)$ for a family of utility functions parametrized by θ. We suppose that individuals' choices are made subject to the budget constraint that $p \cdot x + r = W$, where p indicates the relative price of x versus r, and W a budget common to all individuals. This for example nests the class of Example 3.1 related to time use (e.g., hours spent on TV watching), where $p = 1$ and $W = 24$: all individuals have 24 hours in a day and can swap time one-for-one between watching TV x and spending time on other activities r. Note that the budget constraint can be relaxed to $p \cdot x + r \leq W$, provided that utility is strictly increasing in at least one of the two goods for all individuals.

Proposition B.1. *Suppose utility is a twice differentiable function of (x, r) with heterogeneity parameterized by a scalar θ, such that*

$$X_i = X(\theta_i) := \textit{argmax}_x\{V(x, r; \theta_i) \textit{ subject to } x \geq 0, r \geq 0, \quad p \cdot x + r = W\}$$

We can allow $V(x, r; \theta)$ to be defined only for positive values of x (and not e.g. on all of $\mathbb{R}^2$). For instance $V(\cdot, \cdot; \theta) : [0, 24] \times [0, 24] \mapsto \mathbb{R}$ in the TV setting: we do not need to conceive of the utility

[20] Hierarchical clustering requires the choice of a linkage method and a dissimilarity measure. We recommend using Ward's linkage and the Gower measure for mixed continuous and discrete controls.

that an individual would experience from watching negative TV, or spending more than 24 hours a day on other activities. Assume instead that $V(\cdot,\cdot;\theta)$ is defined on $[0,W]\times[0,W/p]$ for all $\theta\in\Theta$.

Suppose that θ regulates the relative preference for x over r in the sense that $V_x(x,r;\theta)/V_r(x,r;\theta)$ is strictly increasing in θ for all x,r. Define $MRS(x,\theta):=V_x(x,W-px;\theta)/V_r(x,W-px;\theta)$ and suppose that $MRS(x,\theta)$ is strictly decreasing in x and continuously differentiable on $[0,W]\times\Theta$. Suppose finally that $P(X_i=W)=0$ (nobody spends their entire budget on X_i). Then, with probability one $X_i=\max\{h(\rho_i),0\}$ for a strictly increasing and differentiable function h, where $\rho_i=MRS(0,\theta_i)$.

Proof. We can write the optimization problem as

$$X(\theta):=\text{argmax}_x\{V(x,W-px;\theta)\text{ subject to }x\geq 0,r\geq 0\}$$

Given convexity of V, we have by the first order condition that $MRS(X(\theta),\theta)=1$ if and only if there is an interior maximum $X(\theta)\in(0,W)$. In this case, differentiating with respect to θ:

$$MRS_x(X(\theta),\theta)\cdot X'(\theta)=-MRS_\theta(X(\theta),\theta)$$

where $X'(\theta)$ exists by the implicit function theorem. By assumption $MRS_\theta(x,\theta):=\frac{\partial}{\partial\theta}MRS(x,\theta)>0$ for all $x\in(0,\bar{W}),\theta\in\Theta$. Since $MRS_x(x,\theta):=\frac{\partial}{\partial x}MRS(x,\theta)<0$ for all $x\in(0,W),\theta\in\Theta$, we must then have that $X'(\theta)>0$ given that $X(\theta)\in(0,W)$: optimal choices are strictly increasing in θ for any θ such that $0<X(\theta)<W$. Let $\bar{\theta}=\lim_{x\downarrow 0}X^{-1}(x)$. If $\theta=\bar{\theta}$, then $X(\theta)=0$ by continuity and if $\theta<\bar{\theta}$, we must have $X(\theta(\rho_i))\notin(0,W)$.

Meanwhile, since $MRS_\theta(x,\theta)>0$, $MRS(x,\theta)$ is strictly increasing in θ for each $x\geq 0$, the function $\rho(\theta):=MRS(0,\theta)$ is also strictly increasing in θ. Let $\theta(\rho)$ be the inverse function of $\rho(\theta)$, which is also strictly increasing and differentiable in ρ (given that these properties hold for $\rho(\cdot)$).

Given the assumption that $P(X_i=W)=0$, we have combining cases that with probability one:

$$X_i=\begin{cases}0 & \text{if }\theta(\rho_i)\leq\bar{\theta}\\ X(\theta(\rho_i)) & \text{if }\theta(\rho_i)>\bar{\theta}\end{cases}=\max\{X(\theta(\rho_i)),0\}$$

The result of the Proposition then holds with the strictly increasing and differentiable function $h(\rho)=X(\theta(\rho))$.

□

The assumptions invoked in Proposition B.1 are natural if preferences are convex over (x,r), can be parameterized by a scalar θ_i, and utility is smooth in that scalar. While an assumption that the relative price p of x and r is homogeneous across individuals is relatively weak (within e.g. a given market), the restriction that budgets W are homogeneous across individuals i is much stronger. This can be relaxed by assuming that the demand for X does not depend on W, which is reasonable if preferences are approximately quasi-linear, r is a good, or W is suitably high. We can proceed without these assumptions by controlling for income in estimation, such that the assumptions of

Propositions and B.1 and Theorem 3.1 need only hold conditional on income. This strategy can be helpful in motivating Proposition B.1 more generally by also controlling for other proxies of preference heterogeneity, making the scalar heterogeneity assumption less restrictive. Appendix A discusses the use of controls in our approach.

B.1 A generalization of Theorem 3.1

Motivated by Proposition B.1, we now generalize Theorem 3.1 to the case where X_i^* is not the primitive source of selection, but instead there exists some (possibly unobserved) index of heterogeneity ρ_i such that $X_i^* = h(\rho_i)$. In this case we can make Assumptions 2-5 with ρ_i replacing X_i^*, while maintaining the constructive estimand of Theorem 3.1 for $\text{AME}_{\bar{x}}^+$, even if ρ_i is unobserved and the function h unknown. The index ρ_i will therefore not be unique. For example, in the case of Proposition B.1 we could make Assumptions 2-5 about $\rho_i := \theta_i$ or about $\rho_i := MRS(0, \theta_i)$, rather than $\rho_i := X_i^*$.

Proposition B.2. *Suppose that Assumption 1 holds and Assumptions 2-5 all hold not for X_i^*, but for a variable ρ_i such that $X_i^* = h(\rho_i)$ for a strictly increasing and differentiable function h. Then the expression for $AME_{\bar{x}}^+$ from Theorem 3.1 still holds, provided that $h'(\bar{\rho}) \neq 0$, where $\bar{\rho} = h^{-1}(\bar{x})$. The function h does not need to be known or identified, and hence ρ_i may be unobserved for all i.*

Proof. Rewriting Equation (6) as $X_i = \max\{h(\rho_i), h(\bar{\rho})\}$ where $\bar{\rho} = h^{-1}(\bar{x})$, we have, given Theorem 3.1, that

$$\widetilde{\text{AME}}_{\bar{\rho}}^{+} = \lim_{\rho \downarrow \bar{\rho}} \frac{d}{d\rho} \mathbb{E}[Y_i | \rho_i = \rho] - \frac{\text{sgn}\left(\mathbb{E}[Y_i|\rho_i = \bar{\rho}^+] - \mathbb{E}[Y_i|\rho_i = \bar{\rho}]\right) \cdot f_\rho(\bar{\rho}^+)}{F_\rho(\bar{\rho}) \cdot \frac{1}{2\pi} \int \frac{\mathbb{E}[e^{\mathbf{i}\xi Y_i}|\rho_i = \bar{\rho}]}{\mathbb{E}[e^{\mathbf{i}\xi Y_i}|\rho_i = \bar{\rho}^+]} d\xi}, \tag{18}$$

where $\widetilde{\text{AME}}_{\bar{\rho}}^{+} := \lim_{\rho \downarrow \bar{\rho}} \mathbb{E}[(\tilde{Y}_i(\rho) - \tilde{Y}_i(\bar{\rho}))/(\rho - \bar{\rho}) | \rho_i = \rho]$, and $\tilde{Y}_i(\rho)$ indicate potential outcomes with respect to ρ: $\tilde{Y}_i(\rho) = Y_i(h(\rho))$ for any $\rho \geq \bar{\rho}$. Equivalently, $Y_i(x) = \tilde{Y}_i(h^{-1}(x))$ for any $x \geq \bar{x}$.

First, notice that under the maintained assumptions

$$\begin{aligned}
\widetilde{\text{AME}}_{\bar{\rho}}^{+} &= \mathbb{E}[\tilde{Y}'(\bar{\rho}) | \rho_i = \bar{\rho}] = \mathbb{E}\left[\frac{d}{d\rho}\tilde{Y}(\bar{\rho}) \middle| \rho_i = \bar{\rho}\right] = \mathbb{E}\left[\frac{d}{d\rho} Y_i(h(\rho))\Big|_{\rho=\bar{\rho}} \middle| X_i^* = \bar{x}\right] \\
&= h'(\bar{\rho}) \cdot \mathbb{E}[Y_i'(\bar{x}) | X_i^* = \bar{x}] = h'(\bar{\rho}) \cdot \text{AME}_{\bar{x}}^+
\end{aligned}$$

using the chain rule and defining $X_i^* = h(\rho_i)$. Meanwhile, all of the terms on the RHS of (18) are invariant after replacing ρ_i by X_i^* and $\bar{\rho}$ by $\bar{x}$, except that similarly

$$\lim_{\rho \downarrow \bar{\rho}} \frac{d}{d\rho} \mathbb{E}[Y_i | \rho_i = \rho] = \lim_{\rho \downarrow \bar{\rho}} \frac{dx}{d\rho} \frac{d}{dx} \mathbb{E}[Y_i | X_i = x]\Big|_{x=\bar{\rho}} = h'(\bar{\rho}) \cdot \lim_{\rho \downarrow \bar{\rho}} \frac{d}{dx} \mathbb{E}[Y_i | X_i = x]$$

while also $f_\rho(\bar{\rho}^+) = h'(\bar{\rho}) \cdot f_X(\bar{x}^+)$. Equation (18) thus implies (11), provided that $h'(\bar{\rho}) \neq 0$. □

C Defining X_i^* by extending the support of treatment

For any $e \in [0,1]$, $x \geq \bar{x}$ and random variable A, define where $Q_A(a) := \inf\{a : F_A(a) \geq e\}$ and F_A is the CDF function of A. Given a conditioning variable X, we can write $A = Q_{A|X}(E)$ with probability one, for a random variable $E \sim \text{Unif}[0,1]$ that satisfies $E \perp\!\!\!\perp X$. See Lemmas 3 and 4 of Goff et al. (2024) for a proof of this property, which holds regardless of whether A is discrete or continuously distributed.

Consider in particular the conditional quantile function $g(x, x', e) = Q_{Y(x)|X=x'}(e)$. Then using that $Y_i = Y_i(X_i)$ we can write, with probability one:

$$Y_i = g(X_i, X_i, E_i) \tag{19}$$

where $E_i := F_{Y(X)|X}(Y) = F_{Y|X}(Y)$. Note that with Y continuously distributed conditional on X, then $E_i|X_i] = \text{Unif}[0,1]$ and thus $E_i \perp\!\!\!\perp X_i$.[21]

C.1 An endogeneity decomposition with nonseparable errors

For simplicity, we consider in this section a setting in which $\bar{x} = 0$, and assume only that X_i is defined on the positive side of the real line. Accordingly, consider any $x \geq 0$. By the fundamental theorem of calculus, we can write:

$$g(x,x,e) = \underbrace{g(0,0,e) + \int_0^x g_2(0,v,e)dv}_{:=s(x,e)} + \underbrace{\int_0^x g_1(x,v,e)dv}_{:=m(x,e)}$$

where g_1 and g_2 represent derivative functions of g, and the path of integration is from $(x, x') = (0,0)$ to $(0,x)$ and then from $(0,x)$ to (x,x).

We can write

$$s(e,x) = g(0,0,e) + \int_0^x g_2(0,v,e)dv = Q_{Y(0)|X=x}(e) \tag{20}$$

where the second term above $s(x,e)$ is a pure "endogeneity term", capturing how the distribution of the untreated potential outcome $Y_i(0)$ varies across groups with different treatment levels X_i. On the other hand $m(x,e) := \int_0^x g_1(x,v,e)dv$ is a pure causal term, summarizing how the distribution of $Y_i(x)$ changes with x with a fixed conditioning group $X_i = x$. Note that $m(0,e) = 0$ for all e.

By Equation (19), we have that with probability one:

$$Y = m(X, E_i) + s(X, E_i) \tag{21}$$

By totally differentiating $Q_{Y|X=x}(e) = Q_{Y(x)|X=x}(e)$ with respect to x, note that for any $x > 0$, it follows that we can identify

$$\frac{d}{dx} Q_{Y|X=x}(e) = m'(x,e) + s'(x,e) \tag{22}$$

[21] If Y were to exhibit mass points (e.g. Y is discrete), then we can define $E_i|Y_i, X_i \sim \text{Unif}[\lim_{y \uparrow Y_i} F_{Y|X}(y), F_{Y|X}(Y_i)]$ and (19) still holds with $E_i|X_i = \text{Unif}[0,1]$ (Cf. Lemma 4 of of Goff et al. 2024).

where we define $m'(x,e) := \frac{d}{dx}m(x,e) = g_1(x,x,e)$ and $s'(x,e) := \frac{d}{dx}s(x,e) = g_2(x,x,e)$. As above m' captures a causal effect and s' an endogeneity term.

Let $m'(x) := \int_0^1 m'(x,e) \cdot de$ and $s'(x) := \int_0^1 s'(x,e) \cdot de$. Under regularity conditions, the expectation analog of Equation (22) satisfies:

$$\frac{d}{dx}\mathbb{E}[Y_i|X_i = x] = m'(x) + s'(x) \tag{23}$$

and $m'(x) = \mathbb{E}[Y_i'(x)|X_i = x]$, where $Y_i'(x) = \frac{d}{dx}Y_i(x)$.

Recall that the parameter $\text{AME}_{\bar{x}}^+$ is well-defined given Assumption 1 from Section 2.2. But we can give it an economic interpretation without assuming X_i^* exists ex-ante, by supposing that a subset of the bunchers having $X_i = 0$ are "marginal", indicated by $M_i = 1$ (e.g. they satisfy a FOC at $X_i = 0$). Then suppose that all of the marginal bunchers bunch: $P(X_i = 0|M_i = 1) = 1$, and that the distribution of the slope $Y_i'(0^+)$ of the treatment response function $Y_i(x)$ as $x \downarrow 0$ is the same as for the non-bunchers that have very small values of X_i:

$$Y_i'(0^+)|X_i = x \xrightarrow{d} Y_i'(0^+)|M_i = 1$$

where convergence in distribution is as $x \downarrow 0$. Our parameter of interest β_m^* is then the average derivative effect of increasing x from zero among marginal bunchers.

To identify the function $f_{s(X)}(\cdot)$, we leverage the following additional assumptions:

Assumption 7. *The following hold:*

(i) As a function of both x and e on the domain $(0,\infty) \times [0,1]$, s is additively separable i.e. $s(x,e) = s(x) + \phi(e)$ for some functions $s(x)$ and $\phi(e)$.

(ii) $s(x)$ is an analytic function of x on $x > 0$.

(iii) $\phi(\cdot)$ is strictly increasing and there exists a continuous solution h satisfying the integral equation

$$\int_{-\infty}^{\infty} \phi^{-1}\left(y - s^*(x)\right) \cdot h(x) \cdot dx = F_{Y(0)}(y)$$

where $s^*(x)$ is an analytic continuation of $s(\cdot)$ to the real line.

Recall that the function s is defined as $s(x,e) = Q_{Y(0)|X=x}(e)$. To resolve an arbitrary normalization in the additively separable decomposition in Assumption 7, we let the x-dependent term $s(x)$ capture the conditional expectation of $Y_i(0)$ given $X_i = x$: $\mathbb{E}[Y_i(0)|X_i = x] = \int_0^1 Q_{Y(0)|X=x}(e) \cdot de = \int_0^1 s(x,e) = s(x)$. This implies that $\int_0^1 \phi(e) \cdot de = 0$.

To motivate item (i) of Assumption 7, observe the following

Proposition C.1. *$Q_{Y(0)|X=x}(e) = s(x) + \phi(e)$ for some functions s, ϕ iff $\{Y_i(0) - \mathbb{E}[Y_i(0)|X_i]\} \perp\!\!\!\perp X_i$*

Proof. In one direction, observe that if $Q_{Y(0)|X=x}(e) = s(x) + \phi(e)$, then $Y_i(0) - s(X_i) = Q_{Y(0)|X_i}(E_i) - s(X_i) = \phi(E_i) + s(X_i) - s(X_i) = \phi(E_i)$, and $E_i \perp\!\!\!\perp X_i$. To see the other direction, define $s(x) :=$

$\mathbb{E}[Y_i(0)|X_i = x]$ and $\mathcal{E}_i := Y_i(0) - s(X_i)$. Then by definition we can write $Q_{Y(0)|X_i}(E_i) = Y_i(0) = s(X_i) + \mathcal{E}_i = s(X_i) + Q_{\mathcal{E}_i|X_i}(\mathcal{E}_i)$. The condition $(Y_i(0) - \mathbb{E}[Y_i(0)|X_i]) \perp\!\!\!\perp X_i$ implies that $\mathcal{E}_i \perp\!\!\!\perp X_i$, so we can replace $Q_{\mathcal{E}_i|X_i}(\mathcal{E}_i)$ with the unconditional $Q_{\mathcal{E}_i}(\mathcal{E}_i)$. Note that for any $e \in [0,1]$, $Q_{Y(0)|X=x}(e) = s(x) + Q_{\mathcal{E}_i}(e)$ since $Q_{\mathcal{E}_i}(\cdot)$ is strictly increasing. Now define $\phi(e) = Q_{\mathcal{E}_i}(e)$. □

Thus item (i) of Assumption 7 is analagous to item (iii) of Assumption 5, but for positive X_i^*.

Discussion of Assumption 7: The assumption of additive separability (i) cannot be directly verified from the data, because $s(x,e) = Q_{Y(0)|X=x}(e)$ is only identified in the limit as $x \downarrow 0$ and not for multiple values of x. For $x > 0$, we can instead only identify $Q_{Y(x)|X=x}(e) = g(x,x,e) = m(x,e) + s(x,e)$. However, additive separability between x and e in $Q_{Y(x)|X=x}(e)$ for $x > 0$ may be construed as indirect evidence in favor of item (i). For example, if $Y_i(x) - Y_i(0)$ is the same for all i (homogeneous treatment effects), then additive separability of the observable quantile function $Q_{Y|X=x}(e)$ (i.e. $g(x,x,e)$) holds if and only if additive separability of $s(x,e)$ holds. We see this in Figure 7 in our application, where the distribution of $Y|X = x$ only differs substantially across different values of x by a location shift: it is approximately normal with nearly identical variance across x.

Item (iii) of Assumption 7 is high-level, but appears to be mild. The integral equation takes the form of a Fredholm Integral of the First Kind. As an analytic function, $s^*(x)$ is continuously differentiable, and if $s^*(x)$ is furthermore strictly monotonic with inverse $x(\cdot)$, then by a change of variables

$$\int_{-\infty}^{\infty} \phi^{-1}(y-t) \cdot \frac{h(x(t))}{s^{*\prime}(x(t))} \cdot dt = F_{Y(0)}(y)$$

which takes the form of a convolution equation, which typically has a solution.

The useful property of Assumption 7 for identification is that it will allow us to define a new probability space in which X is extended to take negative values, and such that $f_{s(X)}(\cdot)$ is identified through a deconvolution operation on that probability space.

Given the analytic function $s : \mathbb{R}^+ \to \mathbb{R}$ on the positive part of the real line, there exists a unique function $s^* : \mathbb{R}^+ \to \mathbb{R}$ such that $s^*(x) = s(x)$ for all $x \geq 0$, referred to as the *analytic continuation* of s. Now define $s^*(x,e) = s^*(x) + \phi(e)$.

Given our initial probability space with probability P defined over $(X_i, \{Y_i(x)\}_{x \geq 0})$, let us define a new probability measure P^* over random variables $(X_i^*, Y_i^*(0))$ using $s^*(x,e)$. In particular:

- We let the conditional distribution of $Y_i^*(0)$ given X_i^* be described by the quantile function $Q^*_{Y^*(0)|X^*=x}(e) = s^*(x,e)$ for all $e \in [0,1]$ and $x \in \mathbb{R}$.
- We let the marginal distribution of X^* according to P^* be described by CDF $F^*_{X^*}(x) = \int_{\infty}^{x} h(x)$ for $x < 0$ and $H(x) = F_X(x)$ if $x \geq 0$, where h is a solution to the equation in item (iii) of Assumption 7.

Under this construction, the following holds with probability one according to P^*:

$$Y_i^*(0) = Q^*_{Y^*(0)|X^*}(E_i^*) = s^*(X_i^*) + \phi(E_i^*) \tag{24}$$

where $E_i := F^*_{Y(0)|X^*}(Y_i(0))$ satisfies $E_i^* \perp\!\!\!\perp X_i^*$ with $E_i^* \sim \text{Unif}[0,1]$ according to P^*.

We note that the function $s^*(x,e)$ can be defined constructively from $s(x,e)$ provided that the radius of convergence of its Taylor series about $x = 0$ is infinite. In this case, the Taylor series of $s^*(x)$ converges and $s_e(x)$ can be expressed as $s^*(x) = \sum_{k=0}^{\infty} \frac{x^k}{k!} \cdot s^{(k)}(0)$, where for any k times differentiable function f we define $f^{(k)}(x) := \frac{d^k}{dx^k} f(x)$. However this stronger condition is not needed for what follows.

The following additional properties of the probability measure P^* will be key:

Proposition C.2. *Under Assumption 7, $\{Y_i^*(0) - \mathbb{E}^*[Y_i^*(0)|X_i^*]\} \perp\!\!\!\perp X_i^*$ according to P^*.*

Proof. Given (24), we have that with probability one according to P^*

$$Y_i^*(0) - \mathbb{E}^*[Y_i^*(0)|X_i^*] = \{s^*(X_i^*) + \phi(E_i^*)\} - s^*(X_i^*) = \phi(E_i^*)$$

Then, since $E_i^* \sim \text{Unif}[0,1]$ according to P^* and $Y_i^*(0) - \mathbb{E}^*[Y_i^*(0)|X_i^*]$ is a measurable function of E_i^*, it follows that $\{Y_i^*(0) - \mathbb{E}^*[Y_i^*(0)|X_i^*]\} \perp\!\!\!\perp X_i^*$ according to P^*. □

Proposition C.3. *Given Assumption 7, the distribution of $Y_i^*(0)|X_i^* \leq 0$ under P^* is the same as the distribution of $Y_i|X_i = 0$ under P.*

Proof.

$$\begin{aligned}
P^*(Y_i^*(0) \leq y | X_i^* \leq 0) &= \frac{1}{P^*(X_i^* \leq 0)} \cdot \int_{-\infty}^{0} dx \cdot h(x) \cdot P^*(Y_i^*(0) \leq y | X_i^* = x) \\
&= \frac{1}{P^*(X_i^* \leq 0)} \cdot \int_{-\infty}^{0} dx \cdot h(x) \cdot P^*(s^*(x) + \phi(E_i^*) \leq y | X_i^* = x) \\
&= \frac{1}{P^*(X_i^* \leq 0)} \cdot \int_{-\infty}^{0} dx \cdot h(x) \cdot P^*(E_i^* \leq \phi^{-1}(y - s^*(x))) \\
&= \frac{1}{P^*(X_i^* \leq 0)} \cdot \int_{-\infty}^{0} dx \cdot h(x) \cdot P^*(E_i^* \leq \phi^{-1}(y - s^*(x))) \\
&= \frac{1}{P(X_i = 0)} \cdot \int_{-\infty}^{0} dx \cdot h(x) \cdot \phi^{-1}(y - s^*(x))
\end{aligned}$$

using that $E_i^* \perp\!\!\!\perp X_i^*$ and that $E_i^* \sim \text{Unif}[0,1]$.

By the same steps and using that $h(x) = f_X(0)$ and $Q_{Y(0)|X=x}(e) = Q^*_{Y^*(0)|X^*=x}(e)$ for all

$x > 0$:

$$\begin{aligned} P(Y_i(0) \leq y | X_i > 0) &= \frac{1}{P(X_i > 0)} \cdot \int_0^\infty dx \cdot f_X(x) \cdot P(Y_i(0) \leq y | X_i = x) \\ &= \frac{1}{1 - P(X_i = 0)} \cdot \int_0^\infty dx \cdot h(x) \cdot P^*(s^*(x) + \phi(E_i^*) \leq y | X_i^* = x) \\ &= \frac{1}{1 - P(X_i = 0)} \cdot \int_0^\infty dx \cdot h(x) \cdot \phi^{-1}(y - s^*(x)) \end{aligned}$$

Using both of these results, we have by item (iii) of Assumption 7 that,

$$\begin{aligned} F_{Y(0)}(y) &= \int_{-\infty}^0 \phi^{-1}\left(y - s^*(x)\right) \cdot h(x) \cdot dx + \int_0^\infty \phi^{-1}\left(y - s^*(x)\right) \cdot h(x) \cdot dx \\ &= P(X_i = 0) \cdot P^*(Y_i^*(0) \leq y | X_i^* \leq 0) + (1 - P(X_i = 0)) \cdot P(Y_i(0) \leq y | X_i > 0) \end{aligned}$$

Meanwhile, by the law of iterated expectations we also have that

$$F_{Y(0)}(y) = P(X_i = 0) \cdot P(Y_i(0) \leq y | X_i = 0) + (1 - P(X_i = 0)) \cdot P(Y_i(0) \leq y | X_i > 0)$$

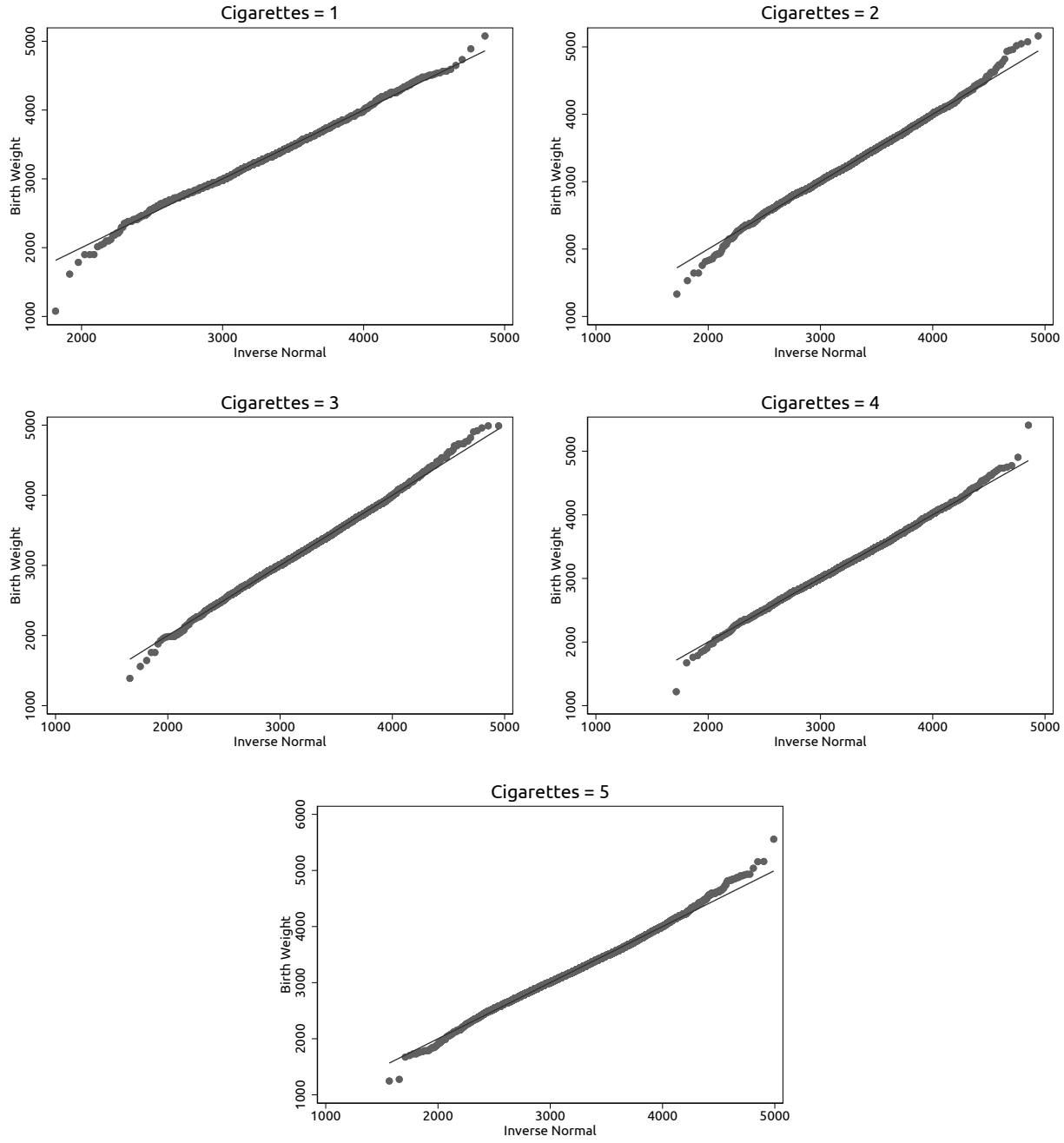
and therefore $P(Y_i \leq y | X_i = 0) = P(Y_i(0) \leq y | X_i = 0) = P^*(Y_i^*(0) \leq y | X_i^* \leq 0)$ for all y. □

By Proposition C.3 and Equation (24), we have that $Y_i | X_i = 0 \sim Y_i^*(0) | X_i^* \leq 0 \sim s^*(X^*) + \phi(E_i^*) | X_i^* \leq 0$. Thus the distribution of $s(X^*) + \phi(E_i^*)$ conditional on $X_i^* \leq 0$ is pinned down from the observable distribution of $Y_i | X_i = 0$.

Note furthermore that $Y_i | X_i = 0^+ \sim s(0^+) + \phi(E_i^*)$, so $\phi(E_i^*) \sim \{Y_i - \mathbb{E}[Y_i | X_i = 0^+]\} | X_i = 0^+$. Since $\phi(E_i^*) | X_i^* \leq 0 \sim \phi(E_i^*)$, we have that $\phi(E_i^*) | X_i^* \leq 0 \sim \{Y_i - \mathbb{E}[Y_i | X_i = 0^+]\} | X_i = 0^+$. It follows from Proposition C.2 that $\{(Y_i^*(0) - \mathbb{E}^*[Y_i^*(0) | X_i^*]) \perp\!\!\!\perp X_i^*\} | X_i^* \leq 0$, so we can then work out the distribution of $s(X_i^*)$ conditional on $X_i^* \leq 0$ via deconvolution as in Section 3.

D Supplemental Empirical Results

Figure 9: Birth Weight QQ Plots Conditional on Maternal Smoking



Note: Each panel plots the estimated QQ plot for birth weight conditional on a different value of maternal cigarettes smoked per day. The closer the plot is to the 45-degree line, the closer to normal is the conditional birth weight distribution. Data taken from Almond et al. (2005).

E Proofs

E.1 Proof of Proposition 2.1

Note that $\mathbb{E}[Y_i(\bar{x})|X_i = x] = \mathbb{E}[Y_i|X_i = x] - ATT(x)$. The limit $\mathbb{E}[Y_i|X_i = \bar{x}^+]$ exists by Assumption 1 (ii) and $ATT(\bar{x}^+)$ exists by Assumption 1 (iii). Therefore $\mathbb{E}[Y_i(\bar{x})|X_i = \bar{x}^+]$ exists as well, and by $ATT(\bar{x}^+) = 0$ must be equal to $\mathbb{E}[Y_i|X_i = \bar{x}^+]$. Therefore,

$$\begin{aligned}\text{ATT}(x) &= \mathbb{E}[Y_i|X_i = x] - \mathbb{E}[Y_i(\bar{x})|X_i = x] \\ &= (\mathbb{E}[Y_i|X_i = x] - \mathbb{E}[Y_i|X_i = \bar{x}^+]) - (\mathbb{E}[Y_i(\bar{x})|X_i = x]) + \mathbb{E}[Y_i(\bar{x})|X_i = \bar{x}^+]) := m(x) - s(x),\end{aligned}$$

which completes the result for the ATT(x).

Using again that $\mathbb{E}[Y_i(\bar{x})|X_i = \bar{x}^+] = \mathbb{E}[Y_i|X_i = \bar{x}^+]$:

$$\begin{aligned}\text{AME}^+_{\bar{x}} &:= \lim_{x \downarrow \bar{x}} \frac{\mathbb{E}[Y_i|X_i = x] - \mathbb{E}[Y_i(\bar{x})|X_i = x]}{x - \bar{x}} \\ &= \lim_{x \downarrow \bar{x}} \left\{ \frac{\mathbb{E}[Y_i|X_i = x] - \mathbb{E}[Y_i|X_i = \bar{x}^+]}{x - \bar{x}} - \frac{\mathbb{E}[Y_i(\bar{x})|X_i = x] + \mathbb{E}[Y_i(\bar{x})|X_i = \bar{x}^+]}{x - \bar{x}} \right\} \\ &= \lim_{x \downarrow \bar{x}} \frac{\mathbb{E}[Y_i|X_i = x] - \mathbb{E}[Y_i|X_i = \bar{x}^+]}{x - \bar{x}} - \lim_{x \downarrow \bar{x}} \frac{\mathbb{E}[Y_i(\bar{x})|X_i = x] + \mathbb{E}[Y_i(\bar{x})|X_i = \bar{x}^+]}{x - \bar{x}} \\ &= \lim_{x \downarrow \bar{x}} \frac{m(x) - m(\bar{x}^+)}{x - \bar{x}} - \lim_{x \downarrow \bar{x}} \frac{s(x) - s(\bar{x}^+)}{x - \bar{x}} := m'(\bar{x}^+) - s'(\bar{x}^+),\end{aligned}$$

where the third line follows from the two limits existings separately, and the final line by $m(\bar{x}^+) = s(\bar{x}^+) = 0$ (which follows immediately from their definitions), and then the definitions of $m'(\bar{x}^+), s'(\bar{x}^+)$.

That the two limits above exist follows from item (ii) from Assumption 1, Assumption 2 and the mean value theorem. Specifically, because $\mathbb{E}[Y_i|X_i = \bar{x}^+]$ exists and is finite and $\mathbb{E}[Y_i|X_i = x]$ is differentiable by item (ii) of Assumption 1, the mean-value theorem applies, and therefore

$$\frac{\mathbb{E}[Y_i|X_i = x] - \mathbb{E}[Y_i|X_i = \bar{x}^+]}{x - \bar{x}} = \frac{d}{dx}\mathbb{E}[Y_i|X_i = x]\Big|_{x=\zeta(x)}, \tag{25}$$

for some $\zeta(x) \in (\bar{x}, x)$. The limit of the RHS exists by item (ii) of Assumption 1, thus the limit of the LHS also exists. The argument for s is analagous using Assumption 2.

E.2 Proof of Theorem 2.1

We begin by establishing local monotonicity of $s(x)$ on an interval $I = (\bar{x}, \bar{x} + \delta)$ where $\delta > 0$. By Assumption 2 $\lim_{x \downarrow \bar{x}} s'(x) \neq 0$, or equivalently that for any $\epsilon > 0$, there exists a $\delta > 0$ such that $x \in (\bar{x}, \bar{x} + \delta)$ implies $|m'(x) - k| \leq \epsilon$, where we let $k := s'(\bar{x}^+)$. Now consider $\epsilon = |k|/2$ and define $I = (\bar{x}, \bar{x} + \delta)$ for the corresponding value of δ. Then $x \in I \implies |m'(x) - K| \leq K/2$, which implies that $m'(x)$ has the same (non-zero) sign as k does for all $x \in I$.

Thus, s admits an inverse function $u^{-1}(t)$ for all $t \in s(I)$. Given Assumption 2, $s'(x)$ exists for all $x \in (\bar{x}, \bar{x} + \min\{\delta, \varepsilon_3\})$, and so s^{-1} admits the derivative function $\frac{d}{dt}s^{-1}(t) = 1/s'(s^{-1}(t))$.

Suppose first that s is strictly increasing. Then, for any t:

$$P(s(X_i) \leq t | X_i \in I) = P(X_i \leq s^{-1}(t) | X_i \in I)$$

Since $f_{X|I}(x) = \frac{d}{dx} P(X_i \leq x | X_i \in I) = f_X(x)/P(X_i \in I)$ exists for all $x \in I$, this implies that $f_{s(X)|I}(t)$ exists for any $t \in s^{-1}(I)$ and is $\frac{d}{dt} P(X_i \leq s^{-1}(t) | X_i \in I) = f_{X|I}(s^{-1}(t))/s'(s^{-1}(t))$, using the chain rule. The case in which s is decreasing aside from the introduction of a minus sign. Combining both cases, we have that $|s'(x)| = f_{X|I}(x)/f_{s(X)|I}(s(x))$ for any $x \in I$.

E.3 Proof of Theorem 2.2

We have by Theorem 2.1 that $s'(x) = \text{sgn}(s'(x)) \cdot f_{X|I}(x)/f_{s(X)|I}(s(x))$ for any $x \in I$. The second term above is then $s'(\bar{x}^+)$, using that $\text{sgn}(s'(x))$ is constant for $x \in I$ (and hence equal to θ). The result then follows by Proposition 2.1, $s(\bar{x}^+) = 0$, and that $\lim_{x \downarrow \bar{x}} f_{s(X)|I}(s(x)) = \lim_{v \downarrow s(\bar{x}^+)} f_{s(X)|I}(v)$. To ease the notation of Theorem 2.2, we then simply define $f_{s(X)|I}(0)$ as the limit $\lim_{v \downarrow 0} f_{s(X)|I}(v)$. We note that the proof of Corollary 3.1 is self-contained and does not depend on this definition.

E.4 Proof of Lemma 1

Under Assumption 1, $\text{ATT}(\bar{x}^+) = 0$, so $\mathbb{E}[Y_i | X_i = \bar{x}^+] = \mathbb{E}[Y_i(\bar{x}) | X_i = \bar{x}^+]$. Denote $y_0 := \mathbb{E}[Y_i(\bar{x}) | X_i^* = \bar{x}]$. Now consider first $\mathbb{E}[Y_i | X_i = \bar{x}] = \mathbb{E}[Y_i(\bar{x}) | X_i^* \leq \bar{x}] = \mathbb{E}[\mathbb{E}[Y_i(\bar{x}) | X_i^*] | X_i^* \leq \bar{x}]$. The inner expectation is $\mathbb{E}[Y_i(\bar{x}) | X_i^*] = y_0 + s(X^*)$, and thus

$$\big(\mathbb{E}[Y_i | X_i = \bar{x}^+] - \mathbb{E}[Y_i | X_i = \bar{x}]\big) = -\mathbb{E}[s(X_i^*) | X_i^* \leq \bar{x}]$$

By item (ii) of Assumption 3, $\text{sgn}(s(x)) = -\theta$. Suppose that $\theta = 1$. Then $s(x) < 0$ for all $x \leq \bar{x}$, and $\text{sgn}\left(\mathbb{E}[Y_i | X_i = \bar{x}^+] - \mathbb{E}[Y_i | X_i = \bar{x}]\right) = 1$. If on the other hand $\theta = -1$, then $s(x) > 0$ for all $x \leq \bar{x}$, and $\text{sgn}\left(\mathbb{E}[Y_i | X_i = \bar{x}^+] - \mathbb{E}[Y_i | X_i = \bar{x}]\right) = -1$.

Finally suppose that, in violation of Assumption 2, $s'(\bar{x}) = 0$ so that $\theta = 0$. In this case, $s(x) = 0$ for all $x < \bar{x}$ by Assumption 3, and the expression of Lemma 1 still holds. This is useful in establishing that the expression in Theorem 3.1 holds even when there is no selection at $\bar{x}$.

E.5 Proof of Corollary 3.1

Assumptions 2 and 3 together allow us to extend Theorem 2.1 to include $\bar{x}$, i.e.:

$$\big|s'(x)\big| = \frac{f_{X^*|I^*}(x)}{f_{s(X^*)|I^*}(s(x))}. \tag{26}$$

for any $x \in I^*$, where $I^* = [\bar{x}, \bar{x} + \delta)$ now includes the bunching point $\bar{x}$ exactly.

Next, we show that we can transform this equation so that it no longer depends on $f_{s(X^*)|I^*}(s(x))$, but it depends instead on $f_{s(X^*)|X=x}(s(x))$. Note that the point $\bar{x}$ belongs to the intersection of I^*

and the set $(-\infty, \bar{x}]$, and with Assumption 4 and item (ii) of Assumption 4, this implies that

$$f_{s(X^*)|I^*}(s(x)) = \frac{P(X_i^* \leq \bar{x})}{P(X_i^* \in I^*)} \cdot f_{s(X^*)|X^* \leq \bar{x}}(s(x)).$$

Similarly, with item (i) of Assumption 4, $f_{X^*|I^*}(\bar{x}) = P(X_i^* \leq \bar{x})/P(X_i^* \in I^*) \cdot f_{X^*|X^* \leq \bar{x}}(\bar{x})$ and thus all together we have that

$$s'(x) = \theta \cdot \frac{f_{X^*}(x)/F_X(\bar{x})}{f_{s(X^*)|X=\bar{x}}(s(x))}, \tag{27}$$

noting that $f_{s(X^*)|X^* \leq \bar{x}} = f_{s(X^*)|X=\bar{x}}$ and $P(X^* \leq \bar{x}) = P(X = \bar{x}) = F_X(\bar{x})$, since $X = \bar{x}$ if and only if $X^* \leq \bar{x}$. This expression is useful because the quantity $P(X_i^* \in I^*)$ for the unknown interval I^* cancels out.

Evaluating at $x = \bar{x}$, we have that $s'(\bar{x}) = \theta \cdot f_{X^*|I^*}(\bar{x})/f_{s(X^*)|I^*}(s(\bar{x}))$ and thus

$$\text{AME}_{\bar{x}}^+ = m'(\bar{x}^+) - \theta \cdot \frac{f_X(\bar{x}^+)/F_X(\bar{x})}{f_{s(X^*)|X=\bar{x}}(s(\bar{x}))} = m'(\bar{x}^+) - \theta \cdot \frac{f_X(\bar{x}^+)/F_X(\bar{x})}{f_{s(X^*)|X=\bar{x}}(0)},$$

where we've used item (i) of Assumption 4 to obtain $f_{X^*}(\bar{x}) = f_{X^*}(\bar{x}^+) = f_X(\bar{x}^+)$, and $s(\bar{x}) = 0$.

E.6 Proof of Theorem 3.1

Combining Eq. (7) with Lemma 2, we have

$$\text{AME}_{\bar{x}}^+ = m'(\bar{x}^+) - 2\pi\theta \left(\int \frac{\mathbb{E}[e^{\mathbf{i}\xi Y_i}|X_i = \bar{x}]}{\mathbb{E}[e^{\mathbf{i}\xi Y_i}|X_i = \bar{x}^+]} d\xi \right)^{-1} \cdot \frac{f_X(\bar{x}^+)}{F_X(\bar{x})},$$

To see that $m'(\bar{x}^+) = \lim_{x \downarrow \bar{x}} \frac{d}{dx}\mathbb{E}[Y_i|X_i = x]$, take the limit of Eq. (25) in the proof of Proposition 2.1 as $x \downarrow \bar{x}$. The expression for θ follows from Lemma 1.

E.7 Proof of Theorem 4.1

Assumption 6 implies that s is also analytic on I. Let $\bar{I} = [\bar{x}, \bar{x} + \mathcal{I}]$, the closure of $(\bar{x}, \bar{x} + I)$ in $\mathbb{R}$. Taking a direct analytic continuation from $(\bar{x}, \bar{x} + I)$ to $\bar{I}$, we can take s to be analytic on all of $\bar{I}$ with $s(\bar{x}) = 0$. Then, there exists a $\varepsilon > 0$ such that $s(x) = \sum_{k=1}^{\infty} s^{(k)}(\bar{x}) \cdot \frac{(x-\bar{x})^k}{k!}$ for any $x \in [\bar{x}, \bar{x} + \varepsilon]$. Analyticity of s on $\bar{I}$ implies that $s^{(k)}(x)$ is continuous on $\bar{I}$ for each k and thus $s^{(k)}(\bar{x}) = s^{(k)}(\bar{x}^+)$. The expression for the remainder follows from the Taylor theorem.

E.8 Proof of Corollary 4.1

Differentiating Equation (27)

$$s^{(k)}(x) = \frac{\theta}{F_X(\bar{x})} \cdot \frac{d^{k-1}}{dx^{k-1}} \frac{f_{X^*}(x)}{f_{s(X^*)|X=\bar{x}}(s(x))},$$

for any $x \in I^*$, from the proof of Corollary 3.1. Working out the derivative and and evaluating at $\bar{x}$, we have

$$s^{(k)}(x) = \frac{\theta}{F_X(\bar{x})} \cdot \sum_{\ell=0}^{k-1} \binom{k-1}{\ell} f_{X^*}^{(k-1-\ell)}(\bar{x}) \cdot \frac{d^\ell}{dx^\ell} \{f_{s(X^*)|X=\bar{x}}(s(x))\}^{-1} \Big|_{x=\bar{x}}$$

The derivatives of $1/f_{s(X^*)|X=\bar{x}}(s(x))$ evaluated at $x = \bar{x}$ can be worked out recursively knowing $f_{s(X^*)|X=\bar{x}}(s(\bar{x}))$ and $f^{(\ell)}_{s(X^*)|X=\bar{x}}(s(\bar{x}))$ for each ℓ. These derivatives are identified since by Eq. (10), $f_{s(X^*)|X=\bar{x}}(v)$ is identified for every $v \in s((-\infty, \bar{x}))$, e.g.:

$$\frac{d}{dv} f_{s(X^*)|X=\bar{x}}(v) = \frac{1}{2\pi} \cdot \frac{d}{dv} \int \frac{\mathbb{E}[e^{\mathbf{i}\xi Y_i}|X_i = \bar{x}]}{\mathbb{E}[e^{\mathbf{i}\xi Y_i}|X_i = \bar{x}^+]} e^{-\mathbf{i}\xi v} d\xi$$

which can be differentiated again and again as needed. The power series converges for each $x \in I_\varepsilon$ by analyticity of $s(x)$.

E.9 Proof of Corollary 4.2

Given Equation (12):

$$\mathrm{ATT}(x) = m(x) - \sum_{k=1}^{\infty} s^{(k)}(\bar{x}^+) \cdot \frac{(x-\bar{x})^k}{k!}$$

The Cauchy–Hadamard formula yields the radius of convergence of the power series appearing on the RHS is $R = \left(\limsup_{k\to\infty} \left| \frac{s^{(k)}(c)}{k!} \right|^{1/k} \right)^{-1}$.