Bunching Designs: A guide to practice

Marinho Bertanha[1], Carolina Caetano[2], Hugo Jales[3], and Nathan Seegert[4]

[1]University of Notre Dame
[2]University of Georgia
[3]Syracuse University
[4]University of Utah

October 16, 2022

Abstract

Abstract: We review the recent developments in the bunching literature, both when bunching is presented in the outcome variable and when it is presented in the treatment variable. We discuss issues related to identification, estimation, practical considerations, and suggest directions for future work.

JEL Codes: C14, H20, C24, C52
Keywords: Bunching, Notches, Kinks, Discontinuity, Endogeneity

1 Introduction

In a seminal paper, Saez (2010) suggested using discontinuities in the worker's budget constraint as a source of identification of worker's responsiveness to the tax rate. This branch of the literature has grown considerably over the last few years, finding applications in finance (Collier, Ellis, and Keys, 2021; Ewens, Xiao, and Xu, 2021a), labor (Cengiz, Dube, Lindner, and Zipperer, 2019; Jales, 2018), environmental economics (Ghanem, Shen, and Zhang, 2020), and many other settings.

In this chapter, we review the basic setting in which Bunching designs might be useful. We discuss identification issues and the recent work that point out the limits of (point) identification of elasticities under more general non-parametric settings, and how to construct bounds for the parameters of interest in these instances. We also discuss practical implementation issues and suggest directions for future work.

In our discussion of the estimation of taxable earnings elasticities, we outline a general setup that nests models where the budget line is continuous and its slope changes at a known point K (i.e., kink) or where the budget line has a jump discontinuity at a known point K (i.e., notch) but the slope is constant otherwise. In the kink case, the slope may either decrease (concave kink, e.g., tax rate goes up) or increase (convex kink, e.g., tax rate goes down). In the notch case, the jump discontinuity may either be negative (negative notch, e.g., lump-sum tax) or positive (positive notch, e.g., subsidy). Most of the applied work so far dealt with concave kinks and negative notches, although there are important exceptions (for examples, see Bajari, Hong, Park, and Town (2017); Kuhn and Yu (2021)).

The original branch of this literature dealt with a class of problems in which the variable that displayed bunching was considered an outcome of interest; which could be potentially affected by discontinuous incentives (typically, but not always, a tax schedule with discontinuities or kinks). Recently, another branch of the literature developed, in which researchers are asking what can be learned when the variable that displays bunching is a treatment variable, as opposed to an outcome. In these settings, it turns out that

bunching behavior can be used to test and correct for endogeneity in standard reduced-form treatment effect models (Caetano, 2015; Caetano, Caetano, and Nielsen, 2020; Caetano, Caetano, Fe, and Nielsen, 2021).

This paper is organized as follows: In Section 2 we look at notches, in Section 3 we study the case of kinks. In Section 4, we dive into the issues of non-parametric identification, semi-parametric identification, bounds, and practical implementation issues; in Section 5 we discuss how bunching in the treatment variable can be leveraged to test and correct for endogeneity, and Section 6 concludes.

2 Notches

2.1 Negative Notches

Assume that a worker faces the following utility function:[1]

$$U(Y;N) = Y - T(Y) - \frac{N}{1+1/\varepsilon}\left(\frac{Y}{N}\right)^{1+1/\varepsilon},$$

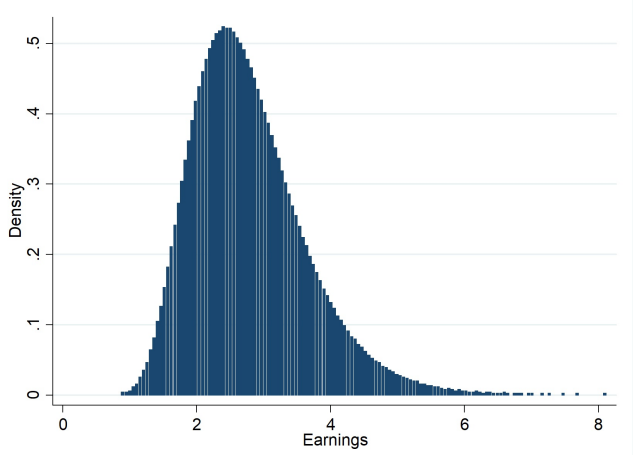
where Y is earnings, $T(Y)$ is a tax liability, N is an ability term, and ε, for reasons we explain below, is a particular type of elasticity. To begin, assume that the worker faces a proportional tax, so that $T(Y) = tY$, where t is the marginal tax rate. In this case, taking first-order conditions, we can solve for the worker's optimal level of earnings, given his ability parameter, as:

$$Y = (1-t)^{\varepsilon} N. \tag{1}$$

Taking logs, we arrive at a $\log(Y) = \varepsilon \log(1-t) + \log(N)$, which makes it clear the interpretation of ε as a elasticity of earnings with respect to the net-of-tax rate. It allows for a simple measure of how much earnings react (if at all) to the net-of-tax rate.

[1] This specific form for the utility function is convenient because it yields a constant elasticity of earnings with respect to the net-of-tax rate given by ε.

Figure 1: Earnings Distribution under a proportional and continuous tax rate



It is useful to write the worker's value function, that is, the utility attained at the optimum level of earnings. It is given by:

$$V_0(N) = (1-t)^{1+\varepsilon} N \left(\frac{1/\varepsilon}{1+1/\varepsilon} \right).$$

For reasons that will become clear later, it is useful to plot the implied distribution of log earnings that arise from particular values of the elasticity, tax rate, and a specified distribution of earnings. For example, if earnings are lognormal, and the elasticity ε is zero, then the distribution of earnings will coincide with the distribution of ability. As ε rises, individuals become more sensitive to the tax rate, and react by lowering earnings (working less, buying leisure) and, as a result, the distribution of earnings becomes a left-shifted version of the distribution of skill. Figure 1 displays one example of such a distribution.

Suppose now, instead, that the worker faces a discontinuous tax schedule, such as the one analyzed by Kleven and Waseem (2013). In this instance, the worker's tax burden increases discontinuously once the worker crosses a particular level of earnings, which we will denote by K. In this instance, the worker faces a tax liability function of the form $T(Y) = tY + \mathbb{1}\{Y > K\}\Delta$, where Δ is the discrete and positive increase in the worker's tax liability once he/she crosses the earnings threshold K. Note that once the worker

crosses the threshold earnings K his tax liability will increase discontinuously. This is the so-called notch.

For the workers that would, in the absence of the notch, choose earnings below the notch point, the solution is the same as before. The workers that would choose earnings above the notch point, however, might be better off staying right at the notch, so as to avoid placing themselves in the higher tax interval. There is one worker, at a particular level of skill, that is indifferent between placing himself at the notch, or behaving just as he would do without the notch. This is the marginal buncher. Every worker with skill level above him will prefer not to bunch. The optimal level of earnings in this case becomes:

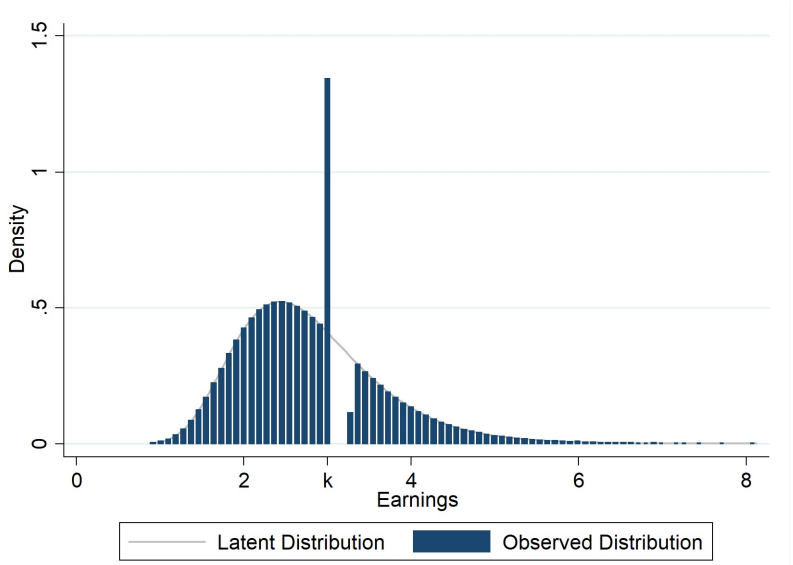
$$Y(N) = \begin{cases} N(1-t)^{\varepsilon} & \text{, if } \quad N < N^* \\ K & \text{, if } \quad N \in (N^*, \overline{N}) \\ N(1-t)^{\varepsilon} & \text{, if } \quad N > \overline{N}, \end{cases} \tag{2}$$

where $N^* = (1-t)^{-\varepsilon}K$ which is the level of skill associated with the individual that would optimally choose the threshold level even in the absence of the discontinuity, and $\overline{N}$ is the skill level associated with the marginal individual.

Note that in this case, a continuous distribution of skill N, the resulting distribution of earnings will be a mixed random variable, with a point mass at K and a continuous part everywhere else. Figure 2 provides one illustration of such distribution.

It is useful to note that, for a continuous and fixed distribution of latent earnings, the higher the elasticity parameter, the more individuals are willing to change their taxable earnings as a result of a tax change. In other words, individuals' behavior is more sensitive to the tax. Thus, it is intuitive to think that, the larger the elasticity parameter, for any given and fixed distribution of skill, the more bunching will be present. Thus, larger bunching would be, in a way, indicative of larger elasticities and perhaps bunching behavior could be used to identify these elasticities. This is the essence of bunching for the identification of elasticities.

Figure 2: Bunching at a tax notch



There are two objects that one might aim to identify from data under the presence of a discontinuous tax schedule: The distribution of earnings that would prevail if the discontinuity were to be removed (the latent distribution); and the elasticity parameter. If one is able to identify the latent distribution, then it should be straightforward to identify the elasticity. The central challenge is that the observed distribution of earnings only coincides with the latent distribution in certain parts of its domain.

In the first wave of this literature (Saez, 2010; Kleven and Waseem, 2013; Chetty, Friedman, Olsen, and Pistaferri, 2011), several variations of a similar procedure were proposed with the aim to estimate the latent distribution of earnings using data from the observed distribution in the presence of a notch. These procedures essentially estimate a flexible parametric model using only the data that is believed to be unaffected by the notch. Below we clarify the conditions required for these procedures to work:

Assumption 1. *In an interval A containing $[K, Y(\overline{N})]$ for which $Pr[Y \notin [K, Y(\overline{N})] | Y \in A] > 0$, where $Y(\overline{N})$ is the earnings chosen by the marginal buncher, the latent density of earnings is equal to $m(Y; \beta_0)$ for a known function m and unknown vector of parameters $\beta_0 \in \mathbb{R}^k$.*

This assumption essentially states two important requirements: The earnings distribution belongs to a parametric class over its most important part, which is the interval $(K, Y(\overline{N}))$ that includes all individuals that displayed a behavioral response to the notch; and this parametric form extends at least a little beyond this interval,[2] which makes it possible to learn the shape in the distorted area by extrapolating the behavior observed on the neighborhood of the bunching point. One special case of this assumption is when the interval A is the whole real line. In this instance, Assumption 1 reduces to a parametric functional form on the latent earnings distribution.[3]

Under this assumption and standard regularity conditions, one can hope to estimate the parameters that characterize the behavior of the density of earnings by leveraging the information contained in the set $A \setminus [K, Y(\overline{N})]$. To do that, one must first (i) estimate the vector of parameters β in the area unaffected by the notch; and (ii) use the estimated parameters to predict how the density would be in the affected range in the absence of the notch.

Although one can use flexible methods that resemble non-parametric estimators for such a task, it should be stressed that the identification of the latent density in the interval affected by the notch is essentially parametric. That is, there must be a belief that one can properly extrapolate the observed behavior in the unaffected area of the distribution towards the affected part. We will get back to this discussion in the following sections.

One important question is then whether or not it is possible to identify the elasticity in this setting under only a continuity assumption. It turns out that the answer to this question is subtle. It is indeed possible to identify the elasticity in the case of a notch (both when there is an increase in the tax and also when there is a decrease in the tax);

[2]Note that the upper-end of this interval, $Y(\overline{N})$, depends on the value of the elasticity ε, an object that is unknown to the econometrician. In other words, the excluded range of the observed domain of the earnings distribution is not known in advance. Without a prior restriction on the parameter space that specifies an upper bound for the value of the elasticity ε, the excluded interval to the right of the notch can be large. We return to this point in the next sections.

[3]The first paper we are aware of that makes use of such an assumption in a related setting is Meyer and Wise (1983).

but the answer is in general negative in the case of a (concave) kink.

The key to the identification, in this case, is that – in the absence of frictions – there is a gap in the distribution of Y between $Y(\underline{N})$ and K. The magnitude of this gap is informative of the elasticity.[4] To see how that works, let $Y(\overline{N})$ be the level of earnings of the marginal buncher. For the individual to be indifferent between bunching at K and choosing the optimum level as given by his first-order condition, it must be the case that:

$$V_0(\overline{N}) - \Delta = U(K; \overline{N}). \tag{3}$$

Denoting by $\overline{Y} = (1-t)^{\varepsilon}\overline{N}$, we can write the terms of the equation above as:

$$V_0(\overline{N}) = (1-t)\overline{Y}\left(\frac{1/\varepsilon}{1+1/\varepsilon}\right).$$

It is useful to note that the only unknown objects on the left-hand side of equation (3) are the level of earnings associated with the marginal buncher and the elasticity ε. The remaining objects are the known value of the size of the notch Δ and the marginal tax rate t. As for the right-hand side, we have that:

$$U(K; \overline{N}) = (1-t)K\left(1 - \frac{1}{1+1/\varepsilon}(K/\overline{Y})^{1/\varepsilon}\right).$$

Once again, it is useful to note that the right-hand side of equation (3) is a function of known objects, such as K and t, and two unknowns: The level of earnings of the marginal buncher, and the elasticity ε. Thus, equation (3) can be used to identify the elasticity as long as the level of earnings of the marginal buncher is identified. This level, $\overline{Y}$, is identified from the upper boundary of the gap in the distribution of observed earnings. Thus, the gap in the distribution of earnings identifies the elasticity of earnings to the net-of-tax rate.[5]

[4] Note that although the object is identified under general continuity assumptions, the source of identification is different from the one that drives the estimates used in Kleven and Waseem (2013), since they rely on the bunching mass instead.

[5] Under the isoelastic functional form assumed for the utility function, there is no closed form expression

For more details, see Theorem (1) in Bertanha, McCallum, and Seegert (2021).

2.2 Positive Notches

The discussion in Section 2.1 is about negative notches, simply referred to as notches. A closely related yet distinct setting is the case of a positive notch; when agents are given a lump-sum subsidy as income crosses a threshold. Although this is rarely the case in the context of taxation, this setting is quite common in the context of pay-for-performance compensation packages, in which workers are given bonuses whenever they reach performance targets (see Kuhn and Yu (2021) for an example). We will, however, continue to cast the problem in the standard setting of taxation, to ease the comparison with the other sections in which we discuss negative notches and kinks.

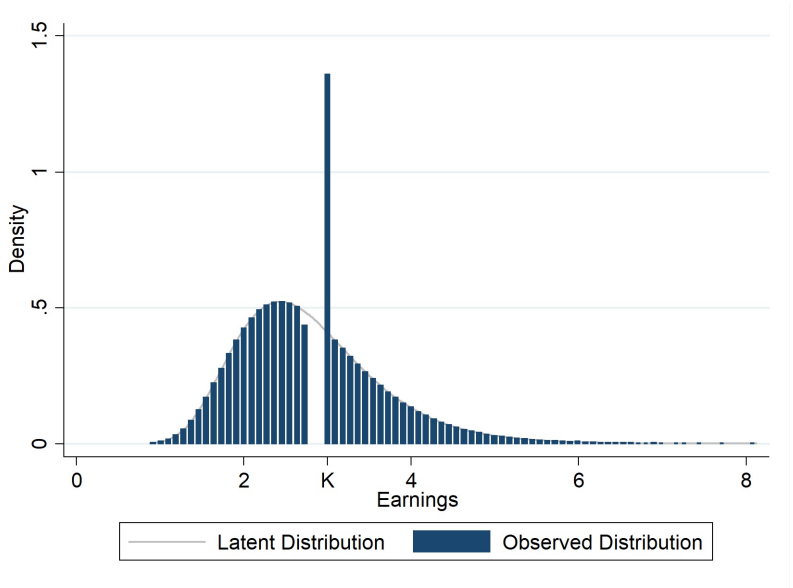
In this instance, the worker faces a tax liability function of the form $T(Y) = tY + \mathbb{1}\{Y \geq K\}\Delta$, where $\Delta < 0$ (now a negative number) is the discrete decline in the worker's tax liability once he/she crosses the earnings threshold K. In other words, the worker earns a bonus whenever his earnings cross a certain point. Note that once the worker crosses the threshold earnings K his tax liability will decrease discontinuously. This is the so-called positive notch.

The solution to the worker's optimization problem in the presence of a positive notch has three regimes in terms of agent type N, similar to the regimes in Equation 2. The difference, however, is that the gap presents itself in the distribution of earnings below K, as opposed to above it. One example of such setting is displayed in Figure 3.

The underlying relationship between the worker's heterogeneity N and the chosen level of income Y is given by:

for the elasticity as a function of the observed objects (notch location K, notch size Δ, marginal tax rate t, and the upper end of the gap $\overline{Y}$), but the elasticity can be obtained from knowledge of the other objects by means of standard numerical procedures.

Figure 3: Earnings distribution under a positive notch



$$Y = \begin{cases} N(1-t)^{\varepsilon} & \text{, if } \quad 0 < N < \underline{N} \\ K & \text{, if } \quad \underline{N} \leq N \leq \overline{N} \\ N(1-t)^{\varepsilon} & \text{, if } \quad \overline{N} < N. \end{cases} \tag{4}$$

The distribution of Y is continuous below $Y(\underline{N}) = \underline{N}(1-t)^{\varepsilon}$, there is a region with zero mass, or a gap, between $Y(\underline{N})$ and K, there is bunching at $Y = K$, and then Y is again continuous above $Y = K$. The distribution of Y is observed, so the lower limit of the gap, i.e., $Y(\underline{N})$, is identified. That determines a relationship between $\underline{N}$ and ε, which can be used to identify ε along the same lines as in the case of a negative notch described in the previous section. This yields non-parametric identification of the elasticity, which is similar to the case of negative notches demonstrated by Bertanha, McCallum, and Seegert (2021).

It is worth noting that, although these arguments are correct to suggest that the elasticity is identified from the gap, the practical usefulness of such result is debatable. In a modified, more realistic scenario, a fraction of workers could be unaware of the tax code, or could face costs of adjustments or some frictions that prevent them from reacting to the tax schedule. In these instances, the distribution will display both bunching and also some

"missing mass", but no gap. Thus, in practice, other strategies to identify this elasticity might be useful. We discuss these approaches in the context of kinks in the following sections (see Sections 3.4, 3.5, and 3.6).

3 Identification Strategies for Elasticities Using Kinks

This section reviews econometric strategies to identify the elasticity parameter using bunching in a distribution coming from kink in a constraint. Consider an observable distribution of a dependent variable Y. The literature focuses on models where this distribution is generated by three components: (1) a distribution F_N of unobserved agent heterogeneity N; (2) model parameters, which typically include an elasticity ε; and (3) a nonlinear budget constraint. Throughout this section we use capital F to the denote the cumulative distribution function (CDF) of a variable; lower-case f denotes the probability density function (PDF) of a variable.

The backbone of existing identification strategies is to first derive a relationship between observed and unobserved components of the model; and second, to solve for the part of the unobserved components that is of interest. In the income tax example, the observed components are: (1) parameters of the budget constraint (slopes, intercepts, points of change) and (2) the distribution of income F_Y. The unobserved components are (A) the distribution of unobserved heterogeneity F_N and (B) primitives of the utility function (i.e., elasticity parameter). Once we have the relationship between these four components, we seek to solve for (B) in terms of (1) and (2) while imposing minimal assumptions on (A). The fact that we have agents maximizing utility subject to a nonlinear budget constraint leads to a distribution of income Y with distinct features. In the case of a kink, the observed distribution of Y includes a mass point at K, referred to as bunching.

The elasticity of taxable income with respect to the net-of-tax rate may only be identified after imposing a set of assumptions—particularly on the unobserved distribution

F_N. This is a point of considerable confusion in the literature that we hope to clarify by defining the assumptions that were implicit in the early bunching literature. Bertanha, McCallum, and Seegert (2021) show that it is possible to identify the elasticity by using assumptions on F_N of various levels of strength. This section begins with the general model following Saez (2010), provides a brief intuition for the lack of identification result without structure, defines the structure used in the original bunching estimators (Saez, 2010; Chetty et al., 2011), discusses (i) the partial identification results from Blomquist, Newey, Kumar, and Liang (2021) and Bertanha, McCallum, and Seegert (2021), (ii) point identification with additional data following Coles, Patel, Seegert, and Smith (2022), and (iii) point identification with semi-parametric assumptions following Bertanha, McCallum, and Seegert (2021), and then concludes by discussing practical issues and extensions and future work.

3.1 Model

Consider the example from Saez (2010). The endogenous variable is taxable income Y, which depends on the continuously distributed disutility of labor N, the elasticity of taxable income with respect to the net-of-tax rate ε, tax rates $t_0 < t_1$, and a concave kink point K. The utility function is linear in consumption and non-linear in labor and has a constant elasticity of substitution.

The observable distribution of Y that results from individual utility maximization is given by

$$Y = \begin{cases} (1-t_0)^{\varepsilon} N, & \text{if } N < \underline{N}\left(k, \varepsilon, s_0\right) \\ K, & \text{if } \underline{N}\left(k, \varepsilon, s_0\right) \leq N \leq \overline{N}\left(k, \varepsilon, s_1\right) \\ (1-t_1)^{\varepsilon} N, & \text{if } N > \overline{N}\left(k, \varepsilon, s_1\right), \end{cases} \tag{5}$$

where $\underline{N} = (1-t_0)^{-\varepsilon} K$ and $\overline{N} = (1-t_1)^{-\varepsilon} K$.

Equation (5) maps the unobserved variable N to observed variable Y and is a function

of (i) the kink point K; (ii) the slopes of the budget constraint on the left, $S_0 = (1 - t_0)$, and on the right, $S_1 = (1 - t_1)$, of the kink point; and (iii) the elasticity ε. The equation allows us to write the mixed continuous-discrete distribution of income Y, i.e., F_Y, as a function of the continuous distribution of N, i.e., F_N.

The researcher observes F_Y, K, S_0, and S_1, but does not observe F_N and ε. The goal is to solve for ε given the observed objects. The distribution of Y is distinct in that there is a mass of individuals that report taxable income at the kink point K. The literature defines this mass as the amount of bunching B, that is,

$$B \equiv \mathbb{P}\left(Y_i = K\right) = \mathbb{P}\left(\underline{N}\left(K, \varepsilon, S_0\right) \leq N_i \leq \overline{N}\left(K, \varepsilon, S_1\right)\right) \tag{6}$$

$$= F_N\left(\overline{N}\left(K, \varepsilon, S_1\right)\right) - F_N\left(\underline{N}\left(K, \varepsilon, S_0\right)\right). \tag{7}$$

We can also conceptualize the amount of bunching as the difference in the cumulative distribution functions of a counterfactual distribution F_{Y_0}. The variable counterfactual income Y_0 represents the income agents would have chosen in the absence of a tax change at the kink, i.e., when $t_0 = t_1$. In this case,

$$B = \int_K^{K+\Delta Y} f_{Y_0}(u)\, du = F_{Y_0}(K + \Delta Y) - F_{Y_0}(K), \tag{8}$$

where ΔY represents the range of values that change to the bunching point when the tax changes. Equations (5), (7), and (8) are the basis for all estimators in the literature. We briefly describe some of these estimators and compare their strengths and weaknesses in sections 3.3–3.6. Before doing that, though, we briefly discuss when an elasticity can be identified.

3.2 It is Impossible to Identify the Elasticity without Structure

The goal of this section is to show that it is impossible to identify the elasticity if we do not assume anything beyond continuity for the counterfactual distribution of income. Bunching estimators proposed by Saez (2010) and Chetty et al. (2011) make implicit assumptions about that distribution. The main issue is that the elasticity value that one identifies with the help of these implicit assumptions is generally very sensitive to small changes in the assumptions.[6]

It is useful to first rewrite the main equations of the previous section in logs. Lower-case variables denote the natural log of upper-case variables. We have, $y = \varepsilon s_0 + n$ if $n < \underline{n}$, $y = \varepsilon s_1 + n$ if $n > \overline{n}$, or $y = k$ otherwise; $B = \mathbb{P}[y = k] = F_n(\overline{n}) - F_n(\underline{n})$ and $B = \int_k^{k+\Delta y} f_{y_0}(u)\, du = F_{y_0}(k + \Delta y) - F_{y_0}(k)$.

Figure 4 illustrates the problem of identification for the quasi-linear utility model with constant elasticity. The distribution in Panel a is identified from the data, but the distribution in Panel b is not fully identified. Panel b shows the distribution of income that agents would have reported in the counterfactual world of no tax change at the kink. The nature of the problem allows us to learn some aspects of f_{y_0} given our knowledge of f_y (Bertanha et al., 2021). First, we know that the portion of f_{y_0} to the left of k is identical to the portion of f_y to the left of k, i.e., $f_{y_0}(y_0) = f_y(y_0)$ for $y_0 < k$. Second, the portion of f_{y_0} to the right of $k + \varepsilon(s_0 - s_1)$, where ε is unknown, is identical to the portion of f_y to the right of k, i.e., $f_{y_0}(y_0) = f_y(y_0 - \varepsilon(s_0 - s_1))$ for $y_0 > k$. Finally, the area under f_{y_0} over the bunching interval $[k, k + \varepsilon(s_0 - s_1)]$ equals the bunching mass B, where B is observed as part of the distribution of y. We know B but we do not know ϕ, that is, the shape of f_{y_0} in that interval. The only thing we know about ϕ are its values at the boundaries of the interval. In other terms, $\phi(k) = f_{y_0}(k) = f_y(k^-)$ and $\phi(k + \varepsilon(s_0 - s_1)) = f_{y_0}(k + \varepsilon(s_0 - s_1)) = f_y(k^+)$, where $f_y(k^+)$ and $f_y(k^-)$ are side limits of

[6] We discuss assumptions on the distribution of Y_0 but these are equivalent to assumptions on the distribution of N.

the PDF of y.

Current bunching estimators assume a shape for ϕ in order to identify the elasticity. One example of an assumption is the trapezoidal approximation that Saez (2010) uses, which is displayed in Figure 4c.[7] In this case, the formula for the area of the trapezoid relates quantities that are identified (B, $f_y(k^-)$, $f_y(k^+)$) to the elasticity ε, which is then identified. The main issue with this and related approaches is that the value of ε one solves for is highly dependent on the assumption on ϕ.

A priori we want to be general and allow for any possible counterfactual distribution of income that is continuous. That means that ϕ could have any possible shape, as long as it is continuous over that interval. With this level of generality, it is impossible to identify the elasticity (Blomquist and Newey, 2017). If ϕ is peaked extremely high, we obtain area B by integrating over a very small interval $[k, k + \varepsilon(s_0 - s_1)]$, which translates into a small elasticity; if ϕ is flat, we obtain the same area B by integrating over larger interval $[k, k + \varepsilon(s_0 - s_1)]$, which translates into a bigger elasticity. The bottom line is that we may obtain any elasticity we want by using different shapes of ϕ.

It is necessary to have either more data variation or structure to identify an elasticity from this model with a convex kink. A series of papers derive methods using different combinations of additional data and structure (Saez, 2010; Chetty et al., 2011; Blomquist et al., 2021; Coles et al., 2022; Bertanha et al., 2021; Marx, 2022; Hamilton, 2018). We briefly discuss some of them and their strengths and weaknesses.

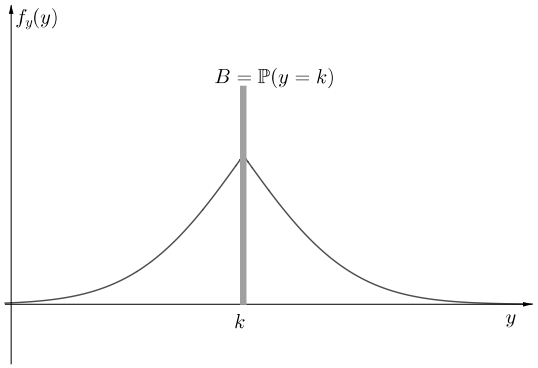
3.3 Original Bunching Estimators

The original bunching estimators focused on the size of bunching to recover an elasticity. These estimators build on the fact that the elasticity increases with the amount of bunching, ceteris paribus. The previous section demonstrates, however, that the elasticity can also increase with the shape of the unobserved counterfactual distribution, ceteris

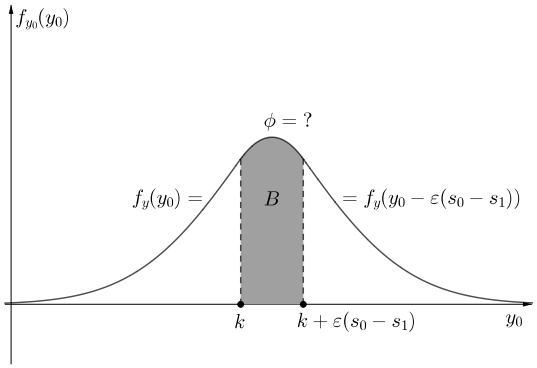
[7] Different from the discussion here, Saez (2010) works with variables in levels instead of logs. See Section 3.3 below.

Figure 4: The Identification Problem

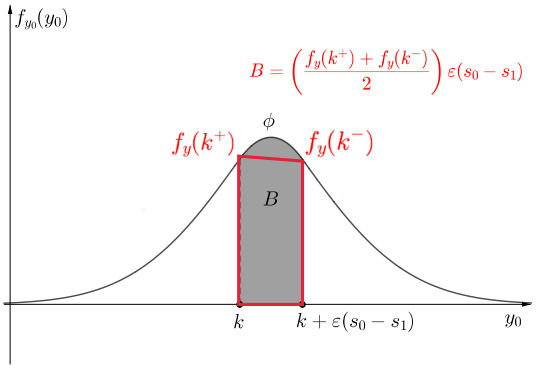
(a) Distribution of Observed Income y



(b) Distribution of Counterfactual Income y_0



(c) Trapezoidal Approximation



Notes: Panel a shows the probability density function (PDF) of the observed distribution of income y, which has a bunching mass B at the kink k. Panel b displays the counterfactual PDF of income y_0 in the absence of tax changes. All variables are in logs. The unobserved PDF of y_0 relates to the observed PDF of y as explained in the main text; however, it is impossible to back out the segment ϕ and ε without further assumptions on the PDF of y_0. Panel c illustrates an example of assumption on ϕ: the trapezoidal approximation. The formula for the area of the trapezoid relates B and other observed quantities to the elasticity, allowing for identification.

paribus. It is therefore critical to understand the assumptions being made about the counterfactual distribution and the sensitivity of the estimates to those assumptions in evaluating different methods.

Consider the estimator given in Equation (6) of Chetty et al. (2011). We obtain this estimator by taking our Equation (8) and adding the assumption that the PDF f_{Y_0} is constant inside the bunching interval $[K, K + \Delta Y]$; specifically, $F_{Y_0}(u) = a + f_{Y_0}(K)u$ for some scalar a and $u \in [K, K + \Delta Y]$. This assumption restricts the class of the distributions of N such that it is equivalent to assuming that N is uniformly distributed in that region (Bertanha et al., 2021). With this assumption Equation (8) can be written as

$$\begin{aligned} B =& \int_K^{K+\Delta Y} f_{Y_0}(u)\, du = F_{Y_0}(K + \Delta Y) - F_{Y_0}(K) \\ =& f_{Y_0}(K)\Delta Y = f_{Y_0}(K)K \left[\left(\frac{1 - t_0}{1 - t_1} \right)^{\varepsilon} - 1 \right]. \end{aligned}$$

With the additional approximation for small tax changes that $[(1 - t_0)/(1 - t_1)]^{\varepsilon} - 1 \cong \varepsilon \ln[(1 - t_0)/(1 - t_1)]$, Chetty et al. (2011) finds the elasticity is

$$\varepsilon \cong \frac{B/f_{Y_0}(K)}{K \ln\left(\frac{1-t_0}{1-t_1}\right)}. \tag{9}$$

Subsequent work, therefore, that use Equation (9) are implicitly restricting the distribution of N to be uniform in that region. This assumption has come under criticism and in some cases the elasticity estimate has been shown to be sensitive to this assumption (Blomquist et al., 2021; Coles et al., 2022).

Alternatively, the estimator in Equation (5) of Saez (2010) is less restrictive than the estimator in Chetty et al. (2011) by using a trapezoidal approximation. The trapezoidal approximation assumes the PDF is affine, which allows the counterfactual distribution f_{Y_0} to have a non-zero slope in the bunching interval. Specifically, the use of the trapezoidal

approximation allows the amount of bunching in Equation (8) to be written as

$$B = \int_{K}^{K+\Delta Y} f_{Y_0}(u)\ du \cong \left(\frac{f_{Y_0}(K+\Delta Y) + f_{Y_0}(K)}{2} \right) \Delta Y, \tag{10}$$

$$\cong \frac{1}{2} \left(f_Y(K^+) \left(\frac{1-t_1}{1-t_0} \right)^{\varepsilon} + f_Y(K^-) \right) K \left[\left(\frac{1-t_0}{1-t_1} \right)^{\varepsilon} - 1 \right]. \tag{11}$$

Saez (2010) uses Equation (11) to solve for the elasticity as an implicit function of the side limits of f_Y, $(f_Y(K^-), f_Y(K^+))$, the tax rates, the kink point and the bunching mass. While the affine assumption is more flexible than the uniform assumption, it relies on a linear approximation being good within a small interval. The problem, however, is that without a priori knowledge of the elasticity, the size of the interval is unknown and could be large (Bertanha et al., 2021).

These original bunching estimators pioneered methods that allowed researchers to investigate behavioral responses previously unstudied due to a lack of experimental or quasi-experimental data, panel data, instrumental variables, etc. These estimators have also sparked an innovative literature to estimate elasticities with different and weaker assumptions, given the reasonable objections to the stark assumptions in these estimators. These new methods provide researchers with a set of tools to estimate elasticities with weaker assumptions and visual tests of appropriateness of the assumptions.

3.4 Partial Identification—Bounding the Elasticity

Blomquist and Newey (2017) and Bertanha, McCallum, and Seegert (2018) provide methods to partially identify the elasticity using nonparametric restrictions on the distribution of N. More recently, Goff (2022) develops partial identification methods using bi-log concavity shape constraints. The main advantage of these bounds is that they require the least amount of structure and no additional data to partially identify the elasticity. The bounds of Blomquist and Newey (2017) partially identify the elasticity by assuming the PDF of N is monotone. The bounds of Bertanha et al. (2018) partially

identify the elasticity by assuming a shape restriction that the slope of the PDF of N is bounded by a maximum slope M. The bounds of Bertanha et al. (2018) have four valuable properties (1) the bounds have closed form solutions, (2) a positive elasticity is assured for any distribution with bunching, (3) it nests and is easily comparable to estimates using a trapezoidal approximation, and (4) it is easily computed using the Stata package `bunching`.[8]

Figure 1 by Bertanha et al. (2021) provides visual intuition for how the bounds work (this figure comes from Bertanha, McCallum, and Seegert (2022)). Panel b in that figure provides the counterfactual distribution of income in the absence of the kink, where the gray region is unobserved as those individuals bunch in the observed distribution. Panels c and d provide potential distributions for different shapes ϕ. The bounds increase as the maximum slope M gets larger. With a larger slope the distance between $k - \varepsilon s_0$ and $k - \varepsilon s_1$ is larger for the upper bound (panel c) and smaller for the lower bound (panel d). Figure 3 by Bertanha et al. (2021) shows the bounds implemente using real data. The figures display how the bounds change as the maximum slope M changes. One feature of this graph is that it provides the trapezoidal approximation estimate in the red-dashed line. This figure then shows the sensitivity of the estimate to different assumptions on the shape of the distribution ϕ.

These bounds provide an important sensitivity check for researchers identifying an elasticity with stricter assumptions.

3.5 Point Identification with Additional Data

A series of papers following Coles et al. (2022) develop methods that rely on additional data to identify the elasticity of interest. These methods use the additional data to estimate a counterfactual distribution. With a counterfactual distribution, these methods estimate the change in Y, ΔY, using Equation (8) (see for example Equation 7 of Coles

[8] See Bertanha, McCallum, Payne, and Seegert (2022).

et al. (2022)). The elasticity then is calculated using the formula[9]

$$\varepsilon = \frac{\Delta Y/K}{ln\left(\frac{1-t_0}{1-t_1}\right)}. \tag{12}$$

The strength of these methods is that they simply restrict the counterfactual distribution on its own using shape or functional form restrictions; they also leverage additional data to estimate the counterfactual. The ability of these methods to provide policy relevant estimates depends on the strength of the estimation of the counterfactual distribution. In most of these methods the counterfactual distribution is estimated using a control group that is unaffected by the kink, at least locally. For example, Coles et al. (2022) use the distribution of firms with different levels of net-operating-losses (NOLs) as a control group because these firms experience the kink point at different levels of Y. Specifically, a firm with \$10,000 in NOLs experiences a 0% tax rate until they earn \$10,000 of income and then 15% for each dollar after (until the next kink). Firms with \$11,000 in NOLs provide a natural control group. These firms are unaffected by a kink at \$10,000 because their kink is at \$11,000. Firms with \$11,000 in NOLs, therefore, can be used to estimate what the distribution for firms with \$10,000 in NOLs would have looked like in the absence of the kink. Other papers use distributions across years or subsamples when the nonlinear budget constraint changes, see for example Hamilton (2018) and Gelber, Jones, and Sacks (2020). For example, Gelber et al. (2020) uses the earnings density of 72-year-olds to estimate a counterfactual distribution for 70- and 71-year-olds that face a nonlinear budget constraint due to the Earnings Test in social security.

The weakness of these methods is that there may not always be a plausible control group to estimate the counterfactual distribution. For example, changes across years may be less convincing due to other potential policy and economic changes.

[9] See Equation (4) of Coles et al. (2022).

3.6 Semi-parametric Point Identification

Bertanha, McCallum, and Seegert (2021) develop two different methods that rely on additional structure on the distribution of N plus covariates in the dataset. The key insight of these methods is to connect the bunching methods to censored regression models.[10] The first method is a truncated Tobit model with covariates. The second method is a censored quantile regression. These methods rely on different identifying assumptions and therefore provide complementary evidence. Two additional advantages of these methods are that (1) they are easily implemented using existing statistical software for censored regression models (or the custom built Stata package `bunching`[11]) and (2) there are visual tests of the appropriateness of the identification assumptions.

The first method is a Tobit elasticity estimator that adds structure to the problem by restricting the unobserved distribution F_N. With the assumption that the conditional distribution N given covariates X is normal, the elasticity could be easily identified with a Tobit model—but this assumption is unnecessarily strong. Bertanha, McCallum, and Seegert (2021) provide sufficient conditions on the joint distribution of (N, X) for consistency of the Tobit elasticity estimator (Lemma 2 of Bertanha, McCallum, and Seegert (2021)). These are semi-parametric restrictions on the distribution of (N, X) that do not require normality of the conditional distribution N given covariates X (see simulation examples by Bertanha, McCallum, and Seegert (2022)). In short, these restrictions are: (i) the distribution of N is a mixture of normals averaged over the non-parametric distribution of X; and (ii) the Tobit best-fit distribution for Y matches the observed distribution of Y.

The restriction (i) becomes weaker with more variation in covariates because the class of distributions of N becomes richer the bigger is the class of distributions of X. For small elasticities, one can show that restriction (ii) is implied by linear restrictions on the first two moments of the distribution of (N, X). Restriction (ii) is easy to verify in practice.

[10] The methods in Bertanha, McCallum, and Seegert (2021) build on a large literature of censoring models, see surveys by Maddala (1983), Amemiya (1984), and Greene (2005).

[11] See Bertanha et al. (2022).

Researchers simply need to compare the Tobit best-fit distribution of Y to the observed distribution of Y as a visual test of the appropriateness of this method. Figure 6 by Bertanha et al. (2021) demonstrates this visual test in a stark example of the Tobit model's robustness to a lack of normality from simulations. This figure graphs the observed distribution in tan bars and the black line gives the Tobit model fit. Despite the observed data being asymmetric and including spikes (and generally not looking normally distributed) the Tobit model fits the data well, satisfying restriction (ii), and the Tobit elasticity estimate is very close to the true elasticity. The additional advantage of this estimator is that standard quasi-MLE asymptotic inference applies.

Bunching is naturally a method that relies on data close to the kink. Bertanha, McCallum, and Seegert (2021) recommend estimating truncated Tobit models for decreasing intervals of Y values centered at the kink—and their Stata package `bunching` graphs the estimates for different window sizes in its default setting. The advantage of truncation is that it only requires the Tobit identification assumptions to hold in a small interval around the kink. In practice, using smaller windows tends to improve the distribution fit but also tends to increase the variance as it relies on less data. Again, the researcher can compare the Tobit best-fit distribution of Y and the observed distribution of Y as a visual test of whether the identification assumption is likely to hold in their specific context. Figure 2 by Bertanha et al. (2021) demonstrates this visual test using the example from Bertanha et al. (2021). To demonstrate how the truncation works, this example uses a misspecified model, as it does not include covariates. In panel d, the Tobit best fit line using 100% of the data (black line) does not fit the simulated data (tan bars) and does not recover the true elasticity estimate. Panels e shows the Tobit best fit lines using 40% of the data, and the fit is better using the truncated sample. Panel f graphs the elasticity estimate for different amounts of the data used for estimation and shows that the estimate gets closer to the true elasticity of 1 as the percent of the data decreases and the Tobit model fit increases. The figures are all produced using the Stata package `bunching`.

The second method is a quantile elasticity estimator that restricts the conditional quantile of N given covariates X. The approach estimates the conditional quantile of Y given X, allowing for different intercepts before and after the kink. If there is sufficient variation in the covariates above and below the kink, then the elasticity is identified as a function of the two intercepts. This only requires a parametric restriction on a conditional quantile of the distribution of N given X, that is, a semi-parametric restriction on the distribution of (N, X). A weakness of this method is that it may be computationally difficult and it requires sufficient variation in covariates (e.g., continuous rather than discrete X). Bertanha, McCallum, and Seegert (2021) provides practical steps to estimate the elasticity building on Chernozhukov and Hong (2002).

Researchers looking to estimate policy-relevant elasticities now have a suite of estimators that rely on identifying assumptions of various degrees of strength. The estimators developed by Bertanha, McCallum, and Seegert (2021) are (1) easy to implement, (2) transparent about the identifying assumptions (some have visual ways of being tested), and (3) are consistent for the true elasticity under assumptions that are weaker than those implicitly made by the original bunching estimators.

3.7 Concave Kinks

Most of the discussion thus far is on convex kinks, simply referred to as kinks. A possible extension of these methods is to the case of concave kinks, that is, when the marginal tax rate decreases as income crosses a threshold.[12]

In the case of concave kink, we have that $t_0 > t_1$ and $I_1 = I_0 + (1 - t_0)K$ in the budget constraint,

$$C = \mathbb{I}\{Y \leq K\}[I_0 + (1 - t_0)Y] + \mathbb{I}\{Y > K\}\left[I_1 + (1 - t_1)\left(Y - K\right)\right]. \tag{13}$$

[12]For an example of such a setting, Bajari et al. (2017) study a very similar problem in the context of hospitals' reimbursement schemes.

The shape of an agent's indifference curve depends on the agent's type N. As N increases, the point of the tangency of the highest indifference curve shifts to the right, moving from just one point with $Y < K$, to two points $Y < K$ and $Y' > K$, and finally to one point $Y > K$. The threshold level $\underline{N}$ is found when the optimal utility of picking $Y < K$ is equal to the utility of picking $Y' > K$. Said differently, if optimal income for agent type N is $Y(N)$, there exists $\underline{N}$ such that

$$\begin{aligned} &I_0 + (1 - t_0)Y(\underline{N}) - \frac{\underline{N}}{1 + 1/\varepsilon}\left(\frac{Y(\underline{N})}{\underline{N}}\right)^{1+1/\varepsilon} \\ &= I_1 + (1 - t_1)Y(\underline{N}^+) - \frac{\underline{N}}{1 + 1/\varepsilon}\left(\frac{Y(\underline{N}^+)}{\underline{N}}\right)^{1+1/\varepsilon}, \end{aligned} \tag{14}$$

where $Y(\underline{N}^+)$ is the side limit of $Y(N)$ as $N \downarrow \underline{N}$, $I_0 + (1 - t_0)Y(\underline{N})$ is consumption when income equals $Y(\underline{N}) < K$, $I_1 + (1 - t_1)Y(\underline{N}^+)$ is consumption when income equals $Y(\underline{N}^+) > K$, and $I_1 = I_0 + (1 - t_0)K$. to ensure continuity of the budget line.

The solution for Y has two regimes in terms of agent type N, as opposed to three regimes as in Equation (5) above. Specifically, the income agents report depends on whether they are below or above a certain threshold $\underline{N}$,

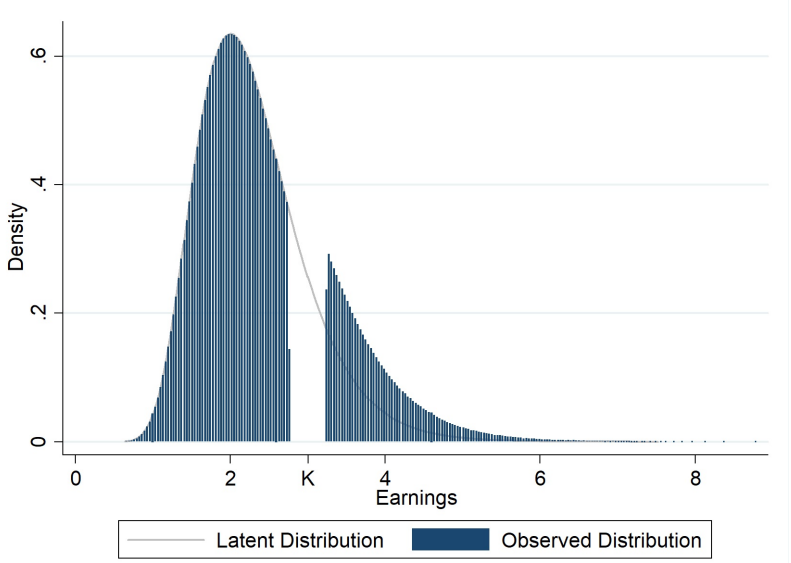
$$Y = \begin{cases} N(1 - t_0)^{\varepsilon} & \text{, if } \ 0 < N \leq \underline{N} \\ N(1 - t_1)^{\varepsilon} & \text{, if } \ \underline{N} < N. \end{cases} \tag{15}$$

Equations (14) and (15) provide a closed-form solution for the threshold level $\underline{N}$, in terms of ε, K, t_0, and t_1:

$$\underline{N} = (1 + \varepsilon)K(1 - t_0)/\left[(1 - t_0)^{\varepsilon+1} - (1 - t_1)^{\varepsilon+1}\right]. \tag{16}$$

The distribution of Y is continuous except for an interval around the kink point where the distribution of Y has zero mass (see Figure 5 for an example). The lower and upper bounds of that gap can be written as nonlinear functions of observable quantities and the

Figure 5: Earnings distribution under a concave kink



elasticity. Specifically, $\underline{Y} = \underline{N}(1-t_0)^{\varepsilon}$ and $\overline{Y} = \underline{N}(1-t_1)^{\varepsilon}$.[13]

We can substitute the closed form solution for $\underline{N}$ from Equation (16) into $\underline{Y}$ and $\overline{Y}$ to find two conditions relating observables and the unknown elasticity. Note, $\underline{Y}$ and $\overline{Y}$ are observable in the data—they are the beginning and end of the gap, that is, the region of zero mass in the distribution of Y. Specifically,

$$\underline{Y} = (1-t_0)^{1+\varepsilon}(1+\varepsilon)K/\left((1-t_0)^{\varepsilon+1} - (1-t_1)^{\varepsilon+1}\right) \tag{17}$$

$$\overline{Y} = (1-t_1)^{\varepsilon}(1+\varepsilon)K(1-t_0)/\left((1-t_0)^{\varepsilon+1} - (1-t_1)^{\varepsilon+1}\right). \tag{18}$$

The elasticity ε can then be solved for using the numerical methods. Unlike the case of a convex kink, it is possible to non-parametrically identify the elasticity without having to assume more than just continuity on the distribution of N.

[13]Note that $\overline{Y} > \underline{Y}$ because $t_0 > t_1$.

4 Practical Issues and Future work

4.1 Practical Issues

There are several important practical considerations when estimating elasticities using distributions shaped by nonlinear budget constraints. The first practical consideration is that the model of the observed distribution necessarily abstracts from additional factors that may shape the observed distribution. In some cases, the elasticity estimates are not sensitive to these additional factors. In other cases, the elasticity estimates are sensitive and these additional factors need to be taken into account.

The additional factor that has received the most attention in the literature is what is referred to as adjustment costs or optimizing frictions. Optimizing frictions limit how precisely agents can adjust Y. These frictions are then used to explain why in some contexts the observed distribution of Y has increased mass not only right at the kink but also around it. If the observed distribution of Y has diffuse bunching, then not accounting for optimizing frictions could bias the elasticity estimate.

There are two different styles of proposed solutions to incorporate optimizing frictions. The first set of solutions try to recover the distribution in the absence of optimizing frictions and then proceed to estimate the elasticity using the methods discussed above. The best solution to this problem is a proper deconvolution theory that models and decomposes the distribution. In the absence of a comprehensive theory for optimizing errors, less general but practical solutions have been proposed. For example, Bertanha, McCallum, and Seegert (2021) develop a filtering procedure where optimizing frictions are modeled as an additive error the optimal income variable. Their filtering procedure works well when 1) the support of the error distribution is small, finite, and known by the researcher and 2) agents that bunch are more affected by the frictions than agents that do not. The Stata package `bunching` has an option that performs this filtering procedure. A similar filtering procedure that uses the bulge in the cumulative distribution function to

filter out the error is proposed by Alvero and Xiao (2020). The second set of solutions model frictions as being part of their models. For example, Gelber, Jones, and Sacks (2013) uses policy-induced changes in the magnitude of kinks to estimate adjustment costs. Similarly, Bertanha, McCallum, and Seegert (2021) suggest that their Tobit model may be extended to include optimizing frictions following censoring models that include measurement error.

The second practical consideration is that empirical methods often rely on tuning parameters, e.g., bandwidth choices. In practice, these tuning parameters can be as important as identifying assumptions. Coles et al. (2022) provide a comparison of elasticity estimates across methods with different tuning parameters. In their setting, they find the elasticity estimates to be considerably sensitive to choices such as the bunching window that is required in the original bunching methods. These findings suggest a large benefit to methods with relatively few tuning parameters or methods that have been shown to be less sensitive to these tuning parameters.

4.2 Extensions and Future Work

There is considerable scope for additional work that uses observed distributions to identify model primitives. This section has provided some foundations for how to identify model primitives in examples with nonlinear budget constraints. Future work can build off of these in at least three ways; build different models that underlie observed distributions, extend the features of the model, and use the model to estimate different parameters.

4.2.1 Build different models that underlie observed distributions

The insights of the original bunching models can be applied to different agents and contexts. The original bunching estimators are developed in the context of individual federal income taxes. Subsequent work has demonstrated how to extend this to different contexts. For example, Einav, Finkelstein, and Schrimpf (2017) develop models in the

context of prescription drug insurance for the elderly in Medicare Part D. An important insight from this paper is that different models can lead to different implications despite these different models fitting the basic distribution. As another example, Coles, Patel, Seegert, and Smith (2022) develop a model where firms maximize profits and respond to non-linear incentives in the corporate tax schedule.

Each context may have idiosyncratic features that are important to model. For example, work in progress by Agostini, Bertanha, Bernier, Bilicka, He, Koumanakos, Lichard, Massenz, Palguta, Patel, Perrault, Riedel, Seegert, Todtenhaupt, and Zudel (2022) develop methods to focus on a kink in the corporate tax schedule at zero. This work focuses on the kink at zero because it is common to corporate tax schedules around the world. The empirical hurdle is that most bunching methods log the endogenous variable (taxable income in this example) and therefore are ill-suited to investigating a kink at zero. This work overcomes this hurdle by developing a transformation method and a two-step approach.

The addition of a model and structure in these cases provide researchers with additional tools to estimate model parameters and perform policy experiments. We find extensions in this style to be extremely fruitful for policy-relevant research.

4.2.2 Extend the features of the model

Future work can add important features to the model. Three areas that have received attention are the extensive margin, dynamic effects, and decomposing elasticities into different components. Gelber, Jones, Sacks, and Song (2021) extend the basic model to include fixed costs of having positive earnings. Their setting is the nonlinear incentives in Social Security created by the Annual Earnings Test that effectively creates a kink with a marginal tax rate above the exempt amount. The inclusion of these fixed costs creates the possibility of an extensive margin response where individuals respond to the kink by reducing their earnings to zero. Gelber et al. (2021) find the extensive margin is

empirically important as their employment elasticity is relatively large, 0.49 in the full sample, and for several reasons is likely a lower bound.

Le Maire and Schjerning (2013) and Marx (2022) extend the model to consider dynamic effects. Le Maire and Schjerning (2013) derive a bunching formula from a dynamic model of income shifting in the context of self-employed workers in Denmark. Their model allows them to estimate that 50-70% of observed bunching is due to income shifting. To put this in context, the elasticity of taxable income estimate from the static model is between 0.43 and 0.53 and with the dynamic model is between 0.14 and 0.20. Marx (2022) similarly extends the static model to a dynamic setting to show how serial dependence in choice variables can bias static-model estimates. This work also considers extensive margin responses, heterogeneous treatment effects, and long-run effects.

Hamilton (2018), Le Maire and Schjerning (2013), and Coles et al. (2022) extend the model to decompose bunching into different components to help understand how agents respond to incentives. Hamilton (2018) separately considers the components of taxable income and finds that a two thirds of the response is due to changes in gross income and one third of the response is due to changes in deductions. Coles et al. (2022) uses panel data and a basic assumption of how revenues and costs co-move to decompose their elasticity of corporate taxable income estimate into economic responses and tax-motivated accounting transactions. They find that in response to a 10% increase in the expected marginal tax rate firms decrease taxable income by 6.1% from accounting transactions (e.g., revenue and expense timing) and 3.0% from economic responses (e.g., scaling operations).

4.2.3 Estimate different parameters from the model

Future work can extend the methods described here to estimate parameters other than elasticities of endogenous variables. Current work is focused in two areas; (1) estimating the change in incentives at thresholds, and (2) estimating policy relevant parameters that require weaker assumptions than elasticities.

In many contexts, the researcher may be interested in the nonlinear incentive structure. For example, Burgstahler and Dichev (1997) note that firms avoid negative earnings and therefore there is excess mass just above zero. Ewens, Xiao, and Xu (2021b) and Bertanha, Seegert, and Yang (2022) are developing methods to estimate the nonlinear incentives that exist for firms (or managers) that induce this level of excess mass. The key hurdles in this work include how to (1) integrate frictions into the model, (2) differentiate kinks, notches, or both, (3) account for agent responsiveness, and (4) identify the change in incentives in light of the impossibility result discussed in Section 3.2.

In some contexts estimating an elasticity may not be necessary to provide policy-relevant insights. For example, researchers may be interested in the tax revenue implications for a given policy change, where an elasticity estimate may be sufficient but not necessary. In work in progress, Moore (2022) shows how the bunching mass can be used as a sufficient statistic for the revenue effect of behavioral responses to small changes of the threshold without having to make assumptions necessary to identify an elasticity. Future work can both apply the methods in Moore (2022) to provide policy-relevant estimates and extend these methods to different types of policy changes.

4.2.4 Discussion

Overall, there remains many parameters of interest that can be estimated using observed distributions and nonlinear budget constraints, incentives, or tax schedules. To do so, researchers should seek out methods with the weakest restrictions that also incorporate key features.

5 Bunching for Causal Inference

A new branch of the literature focuses on how bunching can be leveraged to test or correct for endogeneity in reduced-form causal models. The methods leverage the insight, first

brought up in Caetano (2015), that the bunched observations tend to be discontinuously different in comparison with the observations near the bunching point. Thus, for instance, consider the variable "average number of cigarettes per day among pregnant women," which has a bunching of 80% of the sample at zero. Figure 6 shows that mothers who do not smoke in pregnancy have a discontinuously higher education, and are discontinuously more likely to be married in comparison with mothers who smoke any positive amount. Since the discontinuous patterns in these figures are the standard among all the observable mother, father and pregnancy characteristics which are correlated with smoking, it is expected that a similar pattern exists among the unobservable variables that are correlated with smoking.

Mother's demographic characteristics.

Figure 6: Education (years)

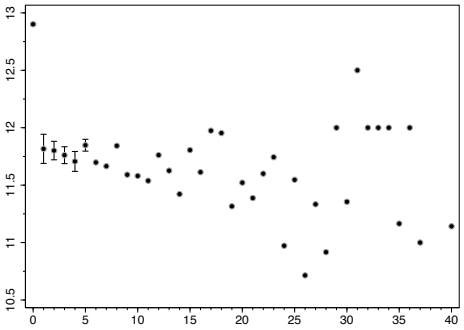
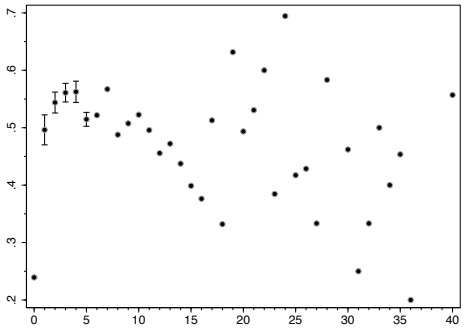


Figure 7: Age



Figures 6 and 7: Dots represent average values referring to the pregnant mothers for each level of daily cigarette consumption. The vertical lines represent the 95% confidence interval of the mean. Source: Caetano (2015), p. 1592.

5.1 Testing Identification Assumptions

Define $Y_i(x)$ as the potential outcome of observation i under treatment level x, and let X_i denote observation i's actual treatment. Suppose that $Y_i(x)$ is differentiable. For a given vector of controls Z_i, we say that

$$X_i \text{ is exogenous } \iff X_i \perp\!\!\!\perp Y_i(x)|Z_i.$$

If X_i is exogenous, then the average conditional marginal treatment effect function $E[Y_i'(x)|X_i = x, Z_i]$ can be identified.[14] We would like to test the exogeneity of X_i.

Caetano (2015) showed that, if the distribution of X_i has bunching at $X_i = \bar{x}$,[15] it is possible to test the exogeneity of X_i. When we compare the outcome of observations at the bunching point and those around it, the treatment itself is very similar. Therefore, there cannot be more than a marginal difference in the outcome that is due to treatment variation. Any discontinuity in the outcome (conditional on controls) at the bunching point must be due to one of two reasons. First, the treatment effect may be discontinuous at the bunching point. Second, there may be discontinuous selection on unobservables, that is, the distribution of the unobservable confounders is discontinuous at the bunching point (and therefore X_i is endogenous). If we can assume that $Y_i(x)$ is continuous in x at $\bar{x}$ with probability one conditional on Z_i, then the first possibility is ruled out. We can then test the exogeneity of X_i by checking if $E[Y_i|X_i = x, Z_i] =$ is continuous in x at $\bar{x}$. If X_i is exogenous, then $E[Y_i|X_i = x, Z_i] = E[Y_i(x)|X_i = x, Z_i] = = E[Y_i|Z_i]$ must be continuous.

This test makes sense when X_i has bunching because, as discussed above, confounders tend to be discontinuously different at the bunching point. Specifically, if X_i is endogenous and has bunching at $\bar{x}$, then $Y_i(x)|X_i = x, Z_i$ will usually be discontinuous in x at $\bar{x}$. Therefore, if X_i is endogenous, $E[Y_i|X_i = x, Z_i]$ will usually be discontinuous in x at $\bar{x}$.

To implement this test nonparametrically, we need to concern ourselves with the dimension of Z_i. Caetano (2015) proposes aggregating the discontinuities, and testing instead if an average of the discontinuities is discontinuous. A convenient aggregation

[14] In this case, $E[Y_i'(x)|X_i = x, Z_i] = dE[Y_i|X_i = x, Z_i]/dx$. Estimation depends on assumptions on $E[Y_i|X_i = x, Z_i]$. For example, if $E[Y_i|X_i = x, Z_i]$ is assumed to be linear on X_i and Z_i, the marginal treatment effect estimator is the coefficient of X_i on a regression of Y_i on X_i and Z_i.

[15] Bunching is not formally defined in most of the literature. A necessary condition for bunching is that $0 < P(X_i = \bar{x}) < 1$. A sufficient condition is to add to the previous condition the requirement that, in a neighborhood of $\bar{x}$, X_i is continuously distributed with a continuous density. The specific strength of the requirements implicit in the definition of bunching vary across the different methods discussed in this section.

explored in that paper yields the following testing quantity:

$$\theta = \lim_{x \downarrow \bar{x}} E\left[E[Y_i|X_i = \bar{x}, Z_i] - Y_i|X_i = x\right].$$

This is particularly convenient when there is a large amount of bunching at $\bar{x}$. In this case, estimation can be done in a two step process: (1) estimate $E[Y_i|X_i = \bar{x}, Z_i]$ nonparametrically,[16] and (2) do a local linear regression of $\hat{E}[Y_i|X_i = \bar{x}, Z_i] - Y_i$ onto X_i at $\bar{x}$, using only observations such that $X_i > \bar{x}$. The approach just described is known as the Discontinuity Test. See empirical implementations in Caetano (2015), Rozenas, Schutte, and Zhukov (2017), Erhardt (2017), Pang (2018), Bleemer (2018), and Bleemer (2020).

More recent papers in this literature have implemented a much simpler approach, first introduced in Caetano and Maheshri (2018), and studied in Caetano et al. (2021), which is known as the Dummy Test. This test is suitable to cases where some semi-parametric assumptions on $Y_i(x)$ are made, and estimation takes these into account. For example, suppose that $Y_i(x) = \beta x + U_i$, where U_i is not observed, and we intend to estimate β as the coefficient of X_i in a regression of Y_i onto X_i and Z_i. In this case, the main identification assumption is that $E[U_i|X_i, Z_i] = Z_i'\alpha$, which implies that $E[Y_i|X_i, Z_i] = \beta X_i + Z_i'\alpha$. The setting in this example includes difference-in-differences approaches which are estimated using linear regressions with fixed effects, which is a popular empirical design for causal inference. The Dummy Test consists of adding the dummy $1(X_i = \bar{x})$ to the regression (i.e. regress Y_i onto X_i, Z_i and $1(X_i = \bar{x})$) and implementing a simple t-test that the coefficient of $1(X_i = \bar{x})$ is significant.

The Dummy Test operates under the same principles as the Discontinuity Test, in leveraging the idea that, if X_i has bunching and is endogenous, the distribution of $U_i|X_i = x, Z_i$ is likely to be discontinuous in x at $\bar{x}$. This would then generate a discontinuity in $E[Y_i|X_i = x, Z_i]$ at the bunching point, which can be detected by including

[16] Or, in practice, as nonparametrically as possible. Machine learning strategies or a kitchen sink regression may be used. Since this is not a boundary expectation, virtually any nonparametric regression technique may be considered.

the dummy $1(X_i = \bar{x})$ in the regression. While the Discontinuity Test tests exclusively the exogeneity of X_i, the Dummy Test is a joint test of the exogeneity of X_i and the assumed functional form, and is generally more powerful. Implementations of the Dummy test in applied work can be seen in Caetano and Maheshri (2018), Ferreira, Ferreira, and Mariano (2018), Lavetti and Schmutte (2018), De Vito, Jacob, and Müller (2019), Caetano, Kinsler, and Teng (2019), Kaneko and Noguchi (2020), Caetano, Caetano, and Nielsen (2021), Jürges and Khanam (2021), Hussein (2021), and Fe and Sanfelice (2022). A similar dummy strategy can be implemented in other semi-parametric models that are popular in empirical research, including nonlinear regression models estimated with GMM, probits and discrete-choice models. In all these cases, the main identification assumptions can be tested by including $1(X_i = \bar{x})$ in the set of controls and performing a simple t-test of the significance of its coefficient.

The same ideas are leveraged in Khalil and Yıldız (2019) to build a test of the exogeneity of X_i in a model where the treatment variable does not have bunching (and may, in fact, be binary), but one of the control variables does. Furthermore, Caetano, Rothe, and Yıldız (2016) showed that bunching on the treatment variable can be used to test the validity of the instrumental variable in a triangular model.

5.2 Identifying Treatment Effects

In the general model discussed in the previous section, all observations at the bunching point have the same treatment. Therefore, after controlling for observables, any variation in the outcomes of those observations must be due to the unobservables. Bunching provides a glimpse into the endogeneity variation unencumbered by treatment variation. With some additional structure, it may be possible to use the variation at the bunching point to correct for endogeneity.

The available strategies in the literature focus on bunching at a corner of the distribution of the treatment, which is the most common form of bunching. For simplicity

of notation, we suppose that $\bar{x} = 0$ and is the lower extreme, so that $X_i \geq 0$. The main structural restriction in the methods discussed below is that observations at the bunching point can be ordered. This is parameterized by a latent variable X_i^*, as

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

If X_i is a choice variable, it is useful (but not necessary) to understand X_i^* as the optimal choice arising from an unconstrained optimization problem. For example, in the maternal smoking example, X_i is the average daily number of cigarettes, and X_i^* is the optimal number as resulting from the maximization of her utility. If the maximization yields a positive number of cigarettes (or zero), then she smokes that number, and thus $X_i = X_i^*$. If the maximization yields a negative number, then she will smoke $X_i = 0$.

Caetano et al. (2020) develop a correction in the model

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

where

$$E[U_i|X_i^*, Z_i] = \delta X_i^* + Z_i'\alpha.$$

Then,

$$E[Y_i|X_i, Z_i] = \beta X_i + Z_i'\alpha + \delta(X_i + E[X_i^*|X_i^* \leq 0, Z_i]1(X_i = 0)).$$

If it is possible to estimate $E[X_i^*|X_i^* \leq 0, Z_i]$, then $X_i + \hat{E}[X_i^*|X_i^* \leq 0, Z_i]1(X_i = 0)$ can be added to the regression as a correction term. The same type of strategy can be used in other, more general, models also considered in Caetano et al. (2020).

Caetano et al. (2020) propose identifying $E[X_i^*|X_i^* \leq 0, Z_i]$ using models on the shape of the conditional distribution of X_i^*. For example, if $X_i^*|Z_i \sim \mathcal{N}(\mu(Z_i), \sigma^2(Z_i))$, for arbitrary functions μ and σ, then $E[X_i^*|X_i^* \leq 0, Z_i] = \mu(Z_i) - \sigma(Z_i)^2 \cdot \lambda(-\mu(Z_i)/\sigma(Z_i))$, where $\lambda(\cdot)$ is the inverse Mills ratio. If $P(X_i = 0|Z_i) \geq 0.5$ and $X_i^*|Z_i$ is symmetric in the

tails, then $E[X_i^*|X_i^* \leq 0, Z_i] = F_i^{-1}(1 - F_i(0)) - E[X_i|X_i \geq F_i^{-1}(1 - F_i(0)), Z_i]$, where $F_i(x) = P(X_i \leq x|Z_i)$.

All the quantities in the two cases above are identifiable, and may be nonparametrically estimated with standard methods, but Caetano et al. (2020) propose a simpler empirical strategy in two steps: (1) discretize the Z_i using hyerarchical clustering (Hastie, Tibshirani, and Friedman (2009)). Let the cluster of observation i be denoted C_i. (2) Do the estimation within each cluster under the assumption that, for all i such that $C_i = c$, $E[X_i^*|X_i^* \leq 0, Z_i] = E[X_i^*|X_i^* \leq 0, C_i = c]$. In the normality case, the second step is equivalent to running a Tobit regression on a constant within each cluster c. The estimate of the constant is $\hat{\mu}(Z_i)$ for all i such that $C_i = c$, and the estimate of the standard deviation is $\hat{\sigma}(Z_i)$. Analogously, in the tail symmetry case, for all i such that $C_i = c$, one would estimate $F_i(0)$ as the probability of bunching in cluster c, $F_i^{-1}(q)$ as the quantile q of X_i in cluster c, and $E[X_i|X_i \geq a, Z_i]$ as the mean of the $X_i \geq a$ in cluster c. Implementations of this method in applied research can be seen in Caetano et al. (2021) and Caetano, Caetano, Nielsen, and Sanfelice (2021).

Caetano, Caetano, and Nielsen (2022b) considers partial identification strategies when the assumptions on the distribution of $X_i^*|Z_i$ are relaxed. They show that a bound on β can be obtained under no distributional assumption. An opposite sharp bound can be obtained under mild assumptions such as that, for $x \leq 0$, the density of $X_i^*|Z_i$, $f_{X^*|Z}(x)$, has no peaks larger than the right limit of the density of $X_i|Z_i$ at the bunching point. Bounds can be narrowed if assumptions on $f_{X^*|Z}(x)$ for $x \leq 0$ are strengthened, such as assuming concavity or convexity, both of which are testable conditions, or that $f_{X^*|Z}$ belongs to families such as bi-log concave or log concave.

Recently, Caetano, Caetano, and Nielsen (2022a) showed that it is possible to obtain nonparametric identification of treatment effects using bunching. Specifically, they show that $E[Y_i'(0)|X_i^* = 0]$, the average marginal treatment effect at the bunching point for the population with $X_i^* = 0$, is identifiable if (1) the treatment effects are continuously

differentiable at the bunching point; (2) the endogenous selection as a function of X_i^* is continuously differentiable at the bunching point; (3) the endogenous selection bias is monotonic on X_i^* for $X_i^* \leq 0$; and (4) the distribution of the idiosyncratic variation conditional on X_i is continuous at the bunching point. Identification is obtained by the use of the change of variables theorem, which relates the bias of endogeneity for $X_i^* = 0$ to the ratio of $\lim_{x \downarrow 0} f_{X|Z}(x)$, and the density of the selection bias term evaluated at $X_i^* = 0$. The latter term is identified because, at the bunching point, the treatment does not vary, and therefore any variation in outcome is due only to the variation of the selection and the idiosyncratic term. The density of the selection can be deconvoluted from the density of the idiosyncratic term using the observations near the bunching point.

This leg of the bunching literature, and especially the advancements in Caetano et al. (2022a), show that the potential of bunching as a source of identification is still far from exhausted, and there is likely much more to be discovered.

6 Conclusion

In this chapter, we provide a thorough review of modern bunching methods. We study the limits of non-parametric identification of taxable earnings elasticity under continuity assumptions in the settings of kinks and notches. We also discuss what can be identified when point identification is not feasible under general continuity conditions, such as in the case of standard, convex kinks. We provide practical guidance for the applied econometrician, discuss how to implement these procedures using canned packages in Stata, and suggest directions for future work.

We also provide the first review of another important and growing branch of this literature that deals with bunching in the treatment variable in standard reduced-form causal models. We discuss how these settings allow testing for endogeneity, and how one can go beyond, into implementing endogeneity corrections.

References

Agostini, C., M. Bertanha, G. Bernier, K. Bilicka, Y. He, E. Koumanakos, T. Lichard, G. Massenz, J. Palguta, E. Patel, L. Perrault, N. Riedel, N. Seegert, M. Todtenhaupt, and B. Zudel (2022). The elasticity of corporate taxable income across countries. Working Paper.

Alvero, A. and K. Xiao (2020). Fuzzy bunching. *Available at SSRN 3611447*.

Amemiya, T. (1984). Tobit Models: A Survey. *Journal of Econometrics 24*(1-2), 3–61.

Bajari, P., H. Hong, M. Park, and R. Town (2017). Estimating price sensitivity of economic agents using discontinuity in nonlinear contracts. *Quantitative Economics 8*(2), 397–433.

Bertanha, M., A. McCallum, A. Payne, and N. Seegert (2022). Bunching estimation of elasticities using Stata. Working Paper.

Bertanha, M., A. H. McCallum, and N. Seegert (2018, March). Better Bunching, Nicer Notching. Working paper.

Bertanha, M., A. H. McCallum, and N. Seegert (2021). Better bunching, nicer botching. Working paper.

Bertanha, M., A. H. McCallum, and N. Seegert (2022). Better bunching, nicer botching. Working paper.

Bertanha, M., N. Seegert, and M.-J. Yang (2022). Estimating incentives using observed distributions of earnings. *Work in progress*.

Bleemer, Z. (2018). The effect of selective public research university enrollment: Evidence from California. Working paper.

Bleemer, Z. (2020). Top percent policies and the return to postsecondary selectivity. Working paper.

Blomquist, S. and W. Newey (2017, September). The Bunching Estimator Cannot Identify the Taxable Income Elasticity. Working Paper 40/17, Cemmap.

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

Burgstahler, D. and I. Dichev (1997). Earnings management to avoid earnings decreases and losses. *Journal of Accounting and Economics 24*(1), 99–126.

Caetano, C. (2015). A test of exogeneity without instrumental variables in models with bunching. *Econometrica 83*(4), 1581–1600.

Caetano, C., G. Caetano, H. Fe, and E. Nielsen (2021). A dummy test of identification in models with bunching. Working paper.

Caetano, C., G. Caetano, and E. Nielsen (2020). Correcting for endogeneity in models with bunching. Working paper.

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

Caetano, C., G. Caetano, and E. Nielsen (2022a). Identification and estimation of average marginal treatment effects with a bunching design. Working paper.

Caetano, C., G. Caetano, and E. Nielsen (2022b). Partial identification of treatment effects using bunching. Working paper.

Caetano, C., G. Caetano, E. Nielsen, and V. Sanfelice (2021). The effect of maternal labor supply on children's skills. Working paper.

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

Caetano, G., J. Kinsler, and H. Teng (2019). Towards causal estimates of children's time allocation on skill development. *Journal of Applied Econometrics 34*(4), 588–605.

Caetano, G. and V. Maheshri (2018). Identifying Dynamic Spillovers of Crime with a Causal Approach to Model Selection. *Quantitative Economics 9*(1), 343–394.

Cengiz, D., A. Dube, A. Lindner, and B. Zipperer (2019). The effect of minimum wages on low-wage jobs. *The Quarterly Journal of Economics 134*(3), 1405–1454.

Chernozhukov, V. and H. Hong (2002). Three-step Censored Quantile Regression and Extramarital Affairs. *Journal of the American Statistical Association 97*(459), 872–882.

Chetty, R., J. N. Friedman, T. Olsen, and L. Pistaferri (2011). Adjustment costs, firm responses, and micro vs. macro labor supply elasticities: Evidence from Danish tax records. *The Quarterly Journal of Economics 126*(2), 749–804.

Coles, J. L., E. Patel, N. Seegert, and M. Smith (2022). How do firms respond to corporate taxes? *Journal of Accounting Research 60*(3), 965–1006.

Collier, B. L., C. Ellis, and B. J. Keys (2021). The cost of consumer collateral: Evidence from bunching. Working paper.

De Vito, A., M. Jacob, and M. A. Müller (2019). Avoiding taxes to fix the tax code. Working paper.

Einav, L., A. Finkelstein, and P. Schrimpf (2017). Bunching at the kink: Implications for spending responses to health insurance contracts. *Journal of Public Economics 146*, 27–40.

Erhardt, E. C. (2017). Microfinance beyond self-employment: Evidence for firms in bulgaria. *Labour economics 47*, 75–95.

Ewens, M., K. Xiao, and T. Xu (2021a). Regulatory costs of being public: Evidence from bunching estimation. Working paper.

Ewens, M., K. Xiao, and T. Xu (2021b). Regulatory costs of being public: Evidence from bunching estimation. Technical report, National Bureau of Economic Research.

Fe, H. and V. Sanfelice (2022). How bad is crime for business? Evidence from consumer behavior. *Journal of Urban Economics*, forthcoming.

Ferreira, D., M. A. Ferreira, and B. Mariano (2018). Creditor control rights and board independence. *The Journal of Finance 73*(5), 2385–2423.

Gelber, A. M., D. Jones, and D. W. Sacks (2013). Earnings adjustment frictions: Evidence from the social security earnings test. Working Paper.

Gelber, A. M., D. Jones, and D. W. Sacks (2020). Estimating adjustment frictions using nonlinear budget sets: Method and evidence from the earnings test. *American Economic Journal: Applied Economics 12*(1), 1–31.

Gelber, A. M., D. Jones, D. W. Sacks, and J. Song (2021). Using nonlinear budget sets to estimate extensive margin responses: Method and evidence from the earnings test. *American Economic Journal: Applied Economics 13*(4), 150–93.

Ghanem, D., S. Shen, and J. Zhang (2020). A censored maximum likelihood approach to quantifying manipulation in china's air pollution data. *Journal of the Association of Environmental and Resource Economists 7*(5), 965–1003.

Goff, L. (2022). Treatment effects in bunching designs: The impact of the federal overtime rule on hours. *arXiv preprint arXiv:2205.10310*.

Greene, W. H. (2005). Censored Data and Truncated Distributions. In T. Mills and K. Patterson (Eds.), *Palgrave Handbook of Econometrics*, Volume 1 of *5*, Chapter 20, pp. 695–736. London: Palgrave Macmillan.

Hamilton, S. (2018). Optimal deductibility: Theory, and evidence from a bunching decomposition. *The Tax and Transfer Policy Institute-Working paper 14*.

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

Hussein, S. M. (2021). *Educational Time Use and the Cognitive Development of Children: Evidence from the Longitudinal Study of Australian Children*. Ph. D. thesis, The Australian National University (Australia).

Jales, H. (2018). Estimating the effects of the minimum wage in a developing country: A density discontinuity design approach. *Journal of Applied Econometrics 33*(1), 29–51.

Jürges, H. and R. Khanam (2021). Adolescents' time allocation and skill production. *Economics of Education Review 85*, 102178.

Kaneko, S. and H. Noguchi (2020). Impacts of natural disaster on changes in parental and children's time allocation: Evidence from the great east japan earthquake. Working paper.

Khalil, U. and N. Yıldız (2019). A test of selection on observables assumption using a discontinuously distributed covariate. Working paper.

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

Kuhn, P. J. and L. Yu (2021). Kinks as goals: Accelerating commissions and the performance of sales teams. Technical report, National Bureau of Economic Research.

Lavetti, K. and I. M. Schmutte (2018). Estimating compensating wage differentials with endogenous job mobility. Working paper.

Le Maire, D. and B. Schjerning (2013). Tax bunching, income shifting and self-employment. *Journal of Public Economics 107*, 1–18.

Maddala, G. S. (1983). *Limited-dependent and Qualitative Variables in Econometrics*. Econometric Society Monographs. Cambridge University Press.

Marx, B. M. (2022). Dynamic bunching estimation with panel data. Working Paper.

Meyer, R. H. and D. A. Wise (1983). Discontinuous distributions and missing persons: The minimum wage and unemployed youth. *Econometrica 51*(6), 1677–1698.

Moore, D. T. (2022). Evaluating tax reforms without elasticities: What bunching can identify. Working Paper.

Pang, J. (2018). *The Effect of Urban Transportation Systems on Employment Outcomes and Traffic Congestion*. Ph. D. thesis, Syracuse University.

Rozenas, A., S. Schutte, and Y. Zhukov (2017). The political legacy of violence: The long-term impact of stalin's repression in ukraine. *The Journal of Politics 79*(4), 1147–1161.

Saez, E. (2010). Do taxpayers bunch at kink points? *American economic Journal: economic policy 2*(3), 180–212.