Bunching Estimation Methods

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

August 7, 2023

Abstract

Abstract: We review the recent developments in the bunching literature, both when bunching appears in the outcome variable and when it appears 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 this chapter, we review the usefulness of bunching designs for the identification of causal effects of interest. We consider how bunching may aid in the identification of the model parameters, as well as the limits of identification given the current state of the literature. We also discuss practical issues of implementation of these methods, provide guidance to practitioners, and suggest directions for future work.

We note that the term "bunching" is not formally defined in most of the literature, although the intuitive understanding of the term is clear. A necessary condition for bunching is that observations of a certain variable are concentrated at a point, i.e., that a certain value occurs with positive probability. A sufficient condition is to add to the previous condition the requirement that, in a neighborhood of the concentration value, the variable is continuously distributed with a continuous density. The specific strength of the requirements implicit in the definition of bunching varies across the different methods discussed in this chapter and are brought up whenever relevant.

Even for the strongest definition, bunching is a common phenomenon, frequently found due to natural restrictions, artificial constraints, or incentives to concentration. The number of examples is very large, and extensive lists can be found in most methodological papers cited in this chapter. Overall, to find bunching, often one needs only to start investigating variables, by plotting frequencies and cumulative distributions.

The bunching methods currently available are mainly divided into two branches. The first branch leverages bunching in the outcome variable of a model, where the bunching results from a discontinuous change in a schedule of incentives, for example, a change in marginal tax rates between brackets. The seminal paper in this literature, in fact, the first paper to leverage bunching at all, is Saez (2010), which suggested using discontinuities in the worker's budget constraint as a source of identification of worker's responsiveness to the tax rate. These discontinuities lead to bunching behavior, that is, heterogeneous workers all choose the same amount of labor supply at these discontinuities, and we see mass points in the distribution of earned income. Following this framework, a large literature has emerged in which bunching is used to identify the causal effect of changing incentives on the response of agents, which is typically summarized by an elasticity parameter. Some interesting examples of applications include Collier et al (2021) and Ewens et al (2021a) in finance, Jales (2018), Cengiz et al (2019), Goff (2022) in labor, and Ghanem et al (2020) in environmental economics. Many other examples are mentioned in this chapter.

The second branch of the literature focuses on models with bunching in the treatment variable, where the researcher is interested in the causal effects of such treatment. The seminal paper in this literature is Caetano (2015), which examined the effects of smoking during pregnancy on the child's birth weight. The natural non-negativity constraint in the amount smoked accounts for a pronounced bunching of around 80% of the sample at zero. In this setting, bunching makes it possible to test whether the controls available in the sample are sufficient to guarantee the exogeneity of the smoking variable. Succeeding literature expanded testing potential into selection-on-unobservable models and into the development of methods to correct estimators when the treatment is endogenous but instrumental variables and panel data are unavailable. This branch of the literature has grown considerably over the last few years, finding applications in economics, finance, and political science. See, for example, the applications in Ferreira et al (2018); Caetano and Maheshri (2018); Caetano et al (2019); Fe and Sanfelice (2022); Caetano et al (2023), among several others mentioned in this chapter.

For bunching in the outcome variable, we discuss identification issues and the recent work that points out the limits of (point) identification of elasticity parameters 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 follow the literature and consider the utility maximization problem of heterogeneous agents choosing consumption and labor supply. For a fixed wage, the budget set gives all feasible combinations of consumption and earned income. The setup is general as it considers two types of discontinuous changes in taxes. In the case of a kink, the budget frontier line is continuous except for a discontinuous change in slope at a known point K, where the marginal income tax rate changes. In the case of a notch, the budget line has a jump discontinuity at a known point K, but the slope is constant otherwise; this occurs when there is a lump-sum tax for income higher than K.

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 et al (2017); Kuhn and Yu (2021)).

For bunching on the treatment, it is not necessary to specify the structural problem that determines the choice of treatment. We discuss the general setting and why bunching in the treatment variable in a model allows one to test the exogeneity of the treatment variable, and discuss the available test choices and implementation. We focus our attention in Caetano (2015)'s discontinuity test of the exogeneity of the treatment variable on a selection-on-observables model, and in Caetano et al (2021a)'s dummy test of identification in linear and two-way fixed effects models. However, identification testing has also been studied in selection-on-observable models when there is bunching in a control variable, in discrete-choice models including those with choice-level unobservables, and in triangular models with instrumental variables, and we provide the appropriate references.

We also discuss how further structure in the model allows the use of bunching to identify treatment effects in the presence of endogeneity without the use of exclusion restrictions or panel data. We focus most of our attention in a simple linear model, where the ideas can be easily understood, but we note that it is straightforward to apply the same strategies in more general models, including models with non-parametric correlated random effects. Importantly, since the bunching correction strategy is not prone to weak identification, it makes it possible to study heterogeneity of treatment effects along different dimensions, which is not often possible with other identification techniques.

We also discuss how the structure in the above models can be relaxed in exchange for partial identification of the treatment effects, which is often sufficient to justify empirical claims. In fact, both the use of bounds as well as the usefulness of the exploration of heterogeneous treatment effects along policy-relevant dimensions are well illustrated in Caetano et al (2023), which is currently the best guide for how to apply these methods in empirical work.

We end with a discussion of the direction of this branch of the literature, where some recent findings opened a promising new frontier where bunching might be used as the sole source of identification of non-parametric causal effects, as is the case with instrumental variables and regression discontinuity designs.

The rest of the chapter is organized as follows. Section 2 covers bunching in the outcome variable, while Section 3 covers bunching in the treatment variable. For bunching in outcomes, Section 2.1 sets up the canonical utility maximization model; Section 2.2 reviews the original methods and assumptions; Section 2.3 presents the modern methods; Section 2.4 discusses practical issues; and Section 2.5 describes extensions and future directions for research. For bunching in the treatment variable, Section 3.1 explains how to use bunching to test identifying assumptions, Section 3.2 concerns the identification of treatment effects, and Section 3.3 describes recent developments and future directions for research. We conclude in Section 4, and acknowledge others' contributions to this chapter in Section 5.

2 Bunching in the Outcome Variable and the Income Tax Example

2.1 Model

In this section, we lay down the utility maximization model first proposed by the public finance literature on bunching (Saez, 2010; Chetty et al, 2011; Kleven and Waseem, 2013). These methods have been extensively used in applied research in taxation (Kopczuk and Munroe, 2015), health care (Einav et al, 2017a), labor (Garicano et al, 2016; Goff, 2022), environmental regulations (Sallee and Slemrod, 2012; Ghanem et al, 2019), education (Dee et al, 2019), and energy demand (Ito, 2014). The exercise introduces the key underpinnings for the rest of Section 2 and gives researchers a starting point for generalizations.

Assume that a worker has earnings Y, tax liability $T(Y)$, and ability term N. Workers differ in their ability term N but are otherwise identical. The population of workers is characterized by a continuous distribution of N. The preferences over consumption C and earned income Y of each worker are characterized by

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

where $C = Y - T(Y)$ and ε is the elasticity parameter, which is constant across workers.

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:

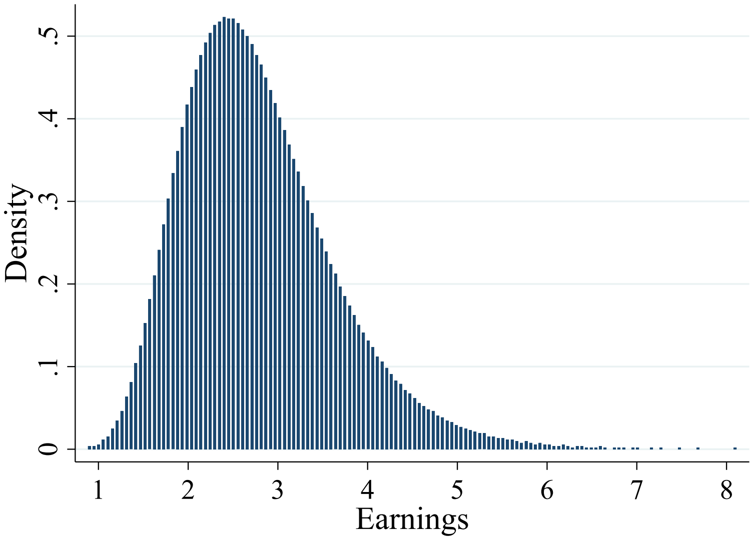
$$Y = (1 - t)^{\varepsilon} N.$$

To see why this is the solution, note that the marginal benefit of an increase in earnings is given by the net-of-tax rate $(1 - t)$, whereas the marginal cost is $(Y/N)^{1/\varepsilon}$. Equating them, we obtain $Y = (1-t)^{\varepsilon} N$.

Taking logs, we arrive at a $\log(Y) = \varepsilon \log(1-t) + \log(N)$, which makes it clear the interpretation of ε as the elasticity of earnings with respect to the net-of-tax rate. Knowledge of ε implies knowledge of the causal effect of the tax rate on earnings.

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 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). As a result, the distribution of log earnings becomes a left-shifted version of the distribution of ability. Figure 1 displays one example of an earnings distribution for particular values of the parameters under the assumption that ability N is log-normal.

Figure 1: Earnings Distribution under a Proportional and Continuous Tax Rate



Notes: Picture generated using $Y = (1-t)^{\varepsilon} N$, where we set $t = 0.1$, $\varepsilon = 0.2$, and assume that N is distributed as log-normal with mean 1 and variance 0.09.

Suppose now, instead, that the worker faces a discontinuous tax schedule. Two common types of discontinuities are kinks, defined as changes in the slope of $T(Y)$, and notches, defined as jump discontinuities in $T(Y)$. We discuss each of these cases in turn. The general tax liability function with one discontinuity at $Y = K$ is

$$T(Y) = \mathbb{1}\{Y \leq K\}t_1 Y + \mathbb{1}\{Y > K\}(\Delta + t_2 Y). \tag{1}$$

When the discontinuity is a kink, the marginal tax rate changes at K, such that the worker's tax liability is $t_1 Y$ for $Y \leq K$ and $t_1 K + t_2(Y - K)$ for $Y > K$. In this case, we set $\Delta = (t_1 - t_2)K$ in equation (1) so the budget line is continuous except for a slope discontinuity at $Y = K$. When the discontinuity is a notch, the worker's tax burden changes discontinuously at K by Δ, and $t_1 = t_2$. The most common case of a kink is the one where $t_1 < t_2$, e.g., Saez (2010); for a notch, it is $\Delta > 0$, e.g., Kleven and Waseem (2013). For now we assume the most common cases but later in this chapter we also discuss extensions to the less common cases of a kink with $t_1 > t_2$ and a notch with $\Delta < 0$ (Sections 2.5.2 and 2.5.4, respectively).

Regardless of the nature of the discontinuity (kink or notch), the optimal solution to the utility maximization problem has the following form:

$$Y(N) = \begin{cases} N(1-t_1)^{\varepsilon} & \text{, if} \quad N < \underline{N} \\ K & \text{, if} \quad N \in [\underline{N}, \overline{N}] \\ N(1-t_2)^{\varepsilon} & \text{, if} \quad N > \overline{N}, \end{cases} \tag{2}$$

where $\underline{N} = K(1-t_1)^{-\varepsilon}$ is the lowest level of ability among bunching individuals, and the expression for $\overline{N}$ depends on the nature of the discontinuity and is given below for kinks and notches.

Consider first the typical case of a kink, that is, $t_1 < t_2$ and $\Delta = (t_1 - t_2)K$. For workers that would, in the absence of the tax rate increase, choose earnings below the kink point, the solution is the same as already discussed above: $Y = N(1-t_1)^{\varepsilon}$. However, the workers that would choose earnings above the kink point, now choose a lower income because they face a higher marginal tax rate. Those workers bunch at K because they would optimally choose $Y < K$ if the higher tax rate t_2 also prevailed below the kink; but in reality the tax rate is t_1 below the kink, so they obtain higher utility by choosing $Y = K$. Among the workers that bunch, the one with the highest ability is the one who would choose exactly $Y = K$ if the higher tax rate t_2 prevailed below the kink point. Every worker with ability level higher than that will choose $Y > K$. Therefore, $\overline{N} = K(1-t_2)^{-\varepsilon}$ in equation (2) above.

Next, consider the typical case of a notch where the worker's tax liability discontinuously increases after the threshold K; that is, the case where $\Delta > 0$ and $t_1 = t_2$. Again, for the workers that would, in the absence of the notch, choose earnings below the notch point, the solution is the same with or without the notch: $Y = N(1-t_1)^{\varepsilon}$. However, the workers that would choose earnings above the notch point might be better off staying right at the notch to avoid the lump-sum tax. There is a worker at a particular ability level N^I who is indifferent between placing themselves at the notch or behaving just as they would do without the notch. The value of N^I is implicitly determined by the following indifference condition,

$$K(1-t_1) - \frac{N^I}{1+1/\varepsilon}\left(\frac{K}{N^I}\right)^{1+1/\varepsilon} = N^I(1-t_1)^{\varepsilon+1} - \Delta - \frac{N^I}{1+1/\varepsilon}\left(\frac{N^I(1-t_1)^{\varepsilon}}{N^I}\right)^{1+1/\varepsilon}, \quad (3)$$

that is, the utility of ability type N^I at income $Y = K$ must equal the utility of optimally choosing $Y^I = N^I(1-t_1)^{\varepsilon} > K$. Every worker with an ability level above N^I will prefer not to bunch. Therefore, $\overline{N} = N^I$ in equation (2) above. Note that, unlike the kink case, $Y(N)$ is a discontinuous function of N in the notch case because $t_1 = t_2$ and $\underline{N} < \overline{N}$ by equation (2).

Starting with a continuous distribution of ability N with full support, equation (2) determines the distribution of Y. The resulting distribution of earnings is a mixed continuous-discrete random variable: there is a mass point at $Y = K$, which equals

$\mathbb{P}[N \in [\underline{N}, \overline{N}]] > 0$, but the distribution is continuous elsewhere. The literature defines this mass as the bunching mass B, that is, the proportion of individuals that report taxable income at the discontinuity point K,

$$B \equiv \mathbb{P}\left[Y = K\right] = \mathbb{P}\left[\underline{N} \leq N \leq \overline{N}\right]. \tag{4}$$

Figure 2 illustrates how the earnings distribution looks like in the cases of a kink and a notch, both using the same log-normal distribution of N. Note that the discontinuity of $Y(N)$ as a function of N in the case of a notch translates into a gap of missing mass in the distribution of Y. The gap interval equals $[K, Y^I]$, where $Y^I = N^I(1-t_1)^\varepsilon$ is the optimal income of the agent that is indifferent between bunching or not, by equation (3).

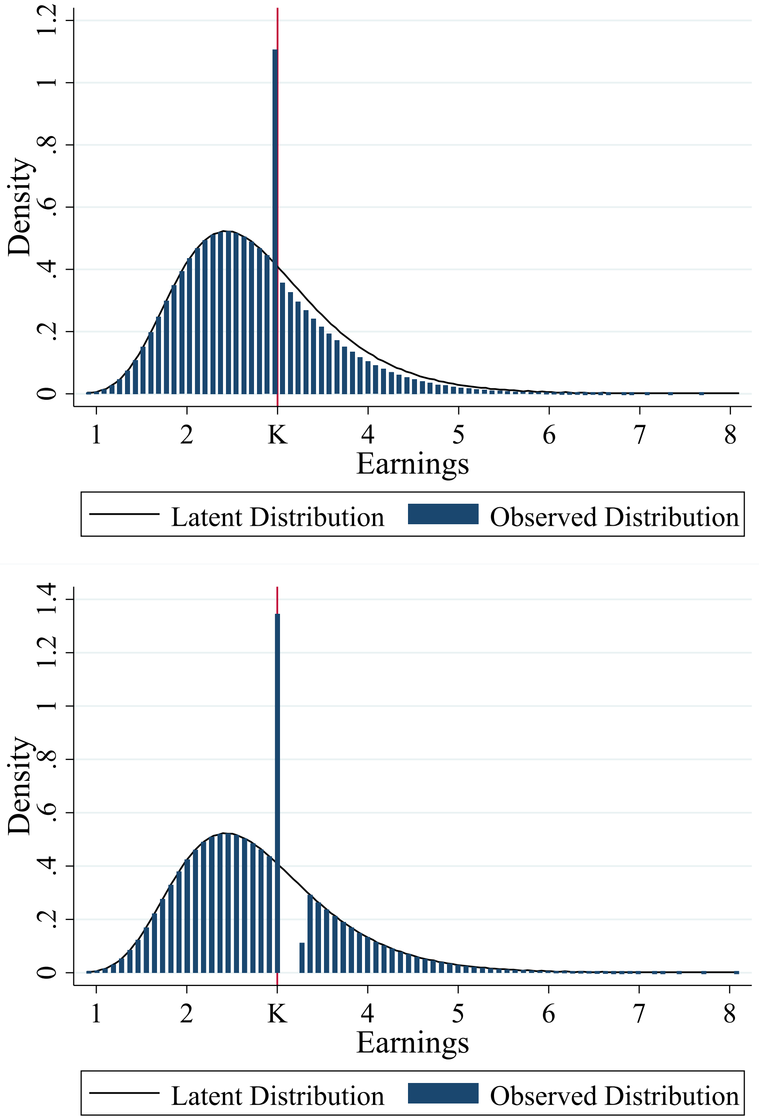
We can also conceptualize the counterfactual distribution of income Y_0 that we would observe in the absence of a kink or a notch. The solid contour lines in Figure 2 display the counterfactual distribution of income corresponding to each case. The range of counterfactual income values for those agents that otherwise choose to bunch takes the form of $[K + \Delta Y(\varepsilon)]$, where the expression for $\Delta Y(\varepsilon)$ depends on ε and whether the discontinuity is a kink or a notch. The range corresponds to values of $Y_0 = N(1-t_1)^\varepsilon$ for $N \in [\underline{N}, \overline{N}]$. In the kink case, agents with counterfactual income between $\underline{N}(1-t_1)^\varepsilon = K$ and $\overline{N}(1-t_1)^\varepsilon = K\left[(1-t_1)/(1-t_2)\right]^\varepsilon$ will bunch at $Y = K$ in the presence of the kink. Likewise, in the notch case, agents with counterfactual income between $\underline{N}(1-t_1)^\varepsilon = K$ and $\overline{N}(1-t_1)^\varepsilon = N^I(1-t_1)^\varepsilon = Y^I$ will bunch at $Y = K$ in the presence of the notch. The bunching mass can be written as,

$$B = \mathbb{P}\left[K \leq Y_0 \leq K + \Delta Y(\varepsilon)\right]. \tag{5}$$

Equations (2), (4), and (5) are the basis for all estimators in the literature.

All empirical analyses build on this type of framework to explain observed bunching in the distribution as a result of optimal behavior from a discontinuous change in incentives.

Figure 2: Bunching Resulting From a Kink or a Notch in the Budget Constraint



Notes: Plot generated using $Y_0 = (1-t_1)^{\varepsilon} N$ for the counterfactual income and equation (2) for the observed income in the case of a kink (top panel) or a notch (bottom panel). We set $t_1 = 0.1$, $t_2 = 0.2$, $\Delta = 0.05$, $K = 3$, $\varepsilon = 0.2$, and assume that N is distributed as log-normal with mean 1 and variance 0.09.

The key insight from this framework is that the higher the elasticity parameter, the more workers are willing to change their taxable earnings due to a tax change. In other words, workers' behavior is more sensitive to the tax. Thus, it is intuitive to think that the larger the elasticity parameter, the more bunching will be present for any given and fixed distribution of ability. This basic intuition led researchers to rely on the observation of bunching to identify the elasticity parameter.

The problem of identification of ε is as follows. Equation (2) maps the distribution of N to the distribution of Y and is a function of ε, t_1, t_2, and Δ. The researcher observes the distribution of income Y, the point K, the tax variables (t_1, t_2, Δ), but does not observe ε nor the distribution of N. To point-identify ε is to solve for a unique ε that is consistent with the observed distribution of Y and the tax values (t_1, t_2, Δ), regardless of the distribution of N. Likewise, to partially-identify ε is to find all values of ε that are consistent with the observed quantities regardless of the unobserved distribution of N. We may also describe the problem of identification in terms of Y_0 in the place of N. First, we write down the map that takes the distribution of Y_0 to the distribution of Y, where the map is again a function of ε, t_1, t_2, and Δ. Then, the identification problem consists of solving for ε that is consistent with the observed distribution of Y and the tax values (t_1, t_2, Δ), regardless of the counterfactual distribution of Y_0.

2.2 Original Bunching Methods

This section presents the original bunching identification methods developed for kinks and notches (Saez, 2010; Chetty et al, 2011; Kleven and Waseem, 2013) and discusses the assumptions required, in light of the critique by Blomquist and Newey (2017) and Bertanha et al (2018).

To solve the identification problem for ε, a natural first step is to restrict the class of possible distributions of ability N or counterfactual income Y_0. All original bunching methods rely on such restrictions in an implicit way. This is a point of considerable confusion in the literature that we hope to clarify in this section. Before we talk about these restrictions, let f_Y and f_{Y_0} denote the probability density functions (PDF) of Y and Y_0. It is important to look at the relationship between these PDFs. From the discussion in the previous section, in the kink case,

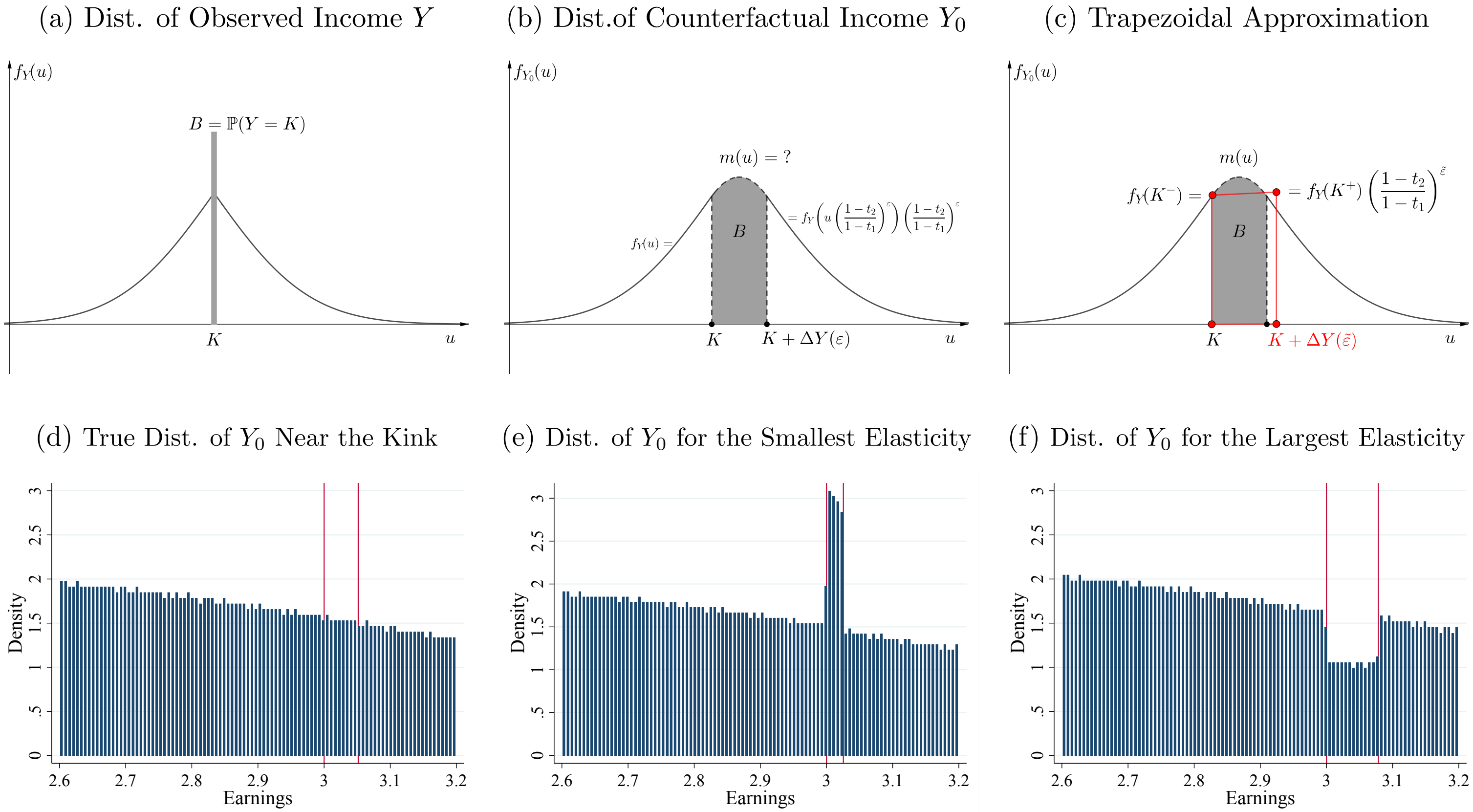
$$f_Y(y) = \begin{cases} f_{Y_0}(y), & \text{if } \ y < K \\ \int_K^{K+\Delta Y(\varepsilon)} f_{Y_0}(u) \ du, & \text{if } \ y = K \\ f_{Y_0}\left(y\left(\frac{1-t_1}{1-t_2}\right)^{\varepsilon}\right)\left(\frac{1-t_1}{1-t_2}\right)^{\varepsilon}, & \text{if } \ y > K, \end{cases} \tag{6}$$

and in the notch case,

$$f_Y(y) = \begin{cases} f_{Y_0}(y), & \text{if } \ y \neq K \\ \int_K^{K+\Delta Y(\varepsilon)} f_{Y_0}(u) \ du, & \text{if } \ y = K. \end{cases} \tag{7}$$

Panels (a) and (b) of Figure 3 illustrate the relationship in equation (6). The following assumption gives a general characterization of the type of restrictions made in the literature.

Figure 3: The Identification Problem in the Kink Case



(a) Dist. of Observed Income Y

(b) Dist.of Counterfactual Income Y_0

(c) Trapezoidal Approximation

(d) True Dist. of Y_0 Near the Kink

(e) Dist. of Y_0 for the Smallest Elasticity

(f) Dist. of Y_0 for the Largest Elasticity

Notes: Panels (a), (b), and (c) illustrate the observed and counterfactual distributions. Panels (d), (e), and (f) use simulated data on the counterfactual distributions. Panel (a) illustrates the probability density function (PDF) of observed income Y with mass point B at the kink K. Panel (b) shows the PDF of counterfactual income Y_0 in the absence of the kink. The relationship between PDFs in (a) and (b) is given in equation (6), but there is no information in f_Y about m other than the area B. Panel (c) illustrates the trapezoidal approximation to m, which in this illustration overestimates the true elasticity, $\tilde{\varepsilon} > \varepsilon$. Panel (d) takes the distribution of counterfactual income of Figure 1 and zooms in at the bunching interval. The red lines in Panels (d),(e), and (f) denote K and $K + \Delta Y(\varepsilon)$, respectively. Panels (e) and (f) present alternative distributions of counterfactual income that are consistent with the observed distribution of income but have different values of the elasticity. For a given maximum slope restriction on that distribution (Section 2.3.1), Panel (e) shows the highest PDF of Y_0, which corresponds to the lower bound on the elasticity. Similarly, Panel (f) shows the lowest PDF of Y_0, which corresponds to the upper bound on the elasticity.

Assumption 1. *There exists an interval A containing $(K, K + \Delta Y(\varepsilon))$, a known function m, and a unknown vector of parameters $\beta_0 \in \mathbb{R}^k$ such that $f_{Y_0}(y) = m(y; \beta_0)$ for every $y \in A$. Moreover, replacing f_{Y_0} with m in equation (6) in the kink case, or equation (7) in the notch case, sets up a system of equations that solve uniquely for β_0 and ε.*

This assumption essentially states two important requirements. First, the counterfactual earnings distribution belongs to a parametric class over its most important part, which is the interval $(K, K + \Delta Y(\varepsilon))$ that consists of individuals that will bunch at $Y = K$ in the presence of a kink or a notch; $m(\cdot; \beta_0)$ corresponds to $m(\cdot)$ in Panel (b) of Figure 3. Second, this parametric form applies to A which extends a little beyond the bunching interval $(K, K + \Delta Y(\varepsilon))$; this is done because it is outside of the bunching interval where we could learn something about the shape of f_{Y_0} through the observation of the shape of f_Y outside of the discontinuity region. The interval A is long enough and the function m is rich enough such that the extension of A over $(K, K + \Delta Y(\varepsilon))$ allows for the identification of ε and β_0. One special case of this assumption is when the interval A is the whole real line. In this instance, we have a parametric functional form on the entire distribution of counterfactual earnings; see Meyer and Wise (1983) for an early example of this assumption. We don't need to go that far but simply need A to be big enough to pin down β_0 and ε through equations (6) or (7). We discuss this procedure in the context of the first wave of the literature below.

2.2.1 Original Kink Methods

The original kink bunching methods focused on the size of the bunching mass to recover an elasticity. These methods build on the fact that the elasticity increases with the amount of bunching, *ceteris paribus*. This can be seen in equation (5), where bunching increases with $\Delta Y(\varepsilon)$. The previous section demonstrates, however, that the elasticity can also increase with a change in the shape of the unobserved counterfactual distribution, *ceteris 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.

The method proposed by Saez (2010) originates in the following approximation of B from equation (5),

$$B = \int_{K}^{K+\Delta Y(\varepsilon)} f_{Y_0}(u)\ du \approx \left(\frac{f_{Y_0}(K+\Delta Y(\varepsilon)) + f_{Y_0}(K)}{2}\right)\Delta Y(\varepsilon)$$
$$= \frac{1}{2}\left(f_Y(K^+)\left(\frac{1-t_2}{1-t_1}\right)^{\varepsilon} + f_Y(K^-)\right) K\left[\left(\frac{1-t_1}{1-t_2}\right)^{\varepsilon} - 1\right], \quad (8)$$

where $f_Y(K^+) = \lim_{u\downarrow K} f_Y(u)$, $f_Y(K^-) = \lim_{u\uparrow K} f_Y(u)$, $f_{Y_0}(K) = f_Y(K^-)$, $f_{Y_0}(K+\Delta Y(\varepsilon)) = f_Y(K^+)\left(\frac{1-t_2}{1-t_1}\right)^{\varepsilon}$, and $\Delta Y(\varepsilon) = K\left[\left(\frac{1-t_1}{1-t_2}\right)^{\varepsilon} - 1\right]$.

Equation (8) above corresponds to equations (4) and (5) in Saez (2010). The approximation to the integral is called the trapezoidal approximation. Saez (2010) uses equation (8) to solve for the elasticity as an implicit function of the known quantities K, t_1, t_2, and the quantities estimated from the data $f_Y(K^-)$, $f_Y(K^+)$, and B. Although many think of this as a flexible procedure, it is important to clarify that the trapezoidal approximation implicitly assumes that the PDF of Y_0 is an affine function over the bunching interval and to the right of the kink. In terms of Assumption 1, $A = [K, K+\Delta Y(\varepsilon)]$ and $m(y;\beta_0) = \beta_{0,1} + \beta_{0,2}y$, where $\beta_0 = (\beta_{0,1}, \beta_{0,2})'$. The trapezoidal approximation is illustrated in Panel (c) of Figure 3.

The equation for B given in equation (6) in Chetty et al (2011) comes from a similar derivation, where f_{Y_0} is approximated with the uniform density inside the bunching interval and the counterfactual distribution immediately to the right of the kink; in terms of Assumption 1, set $A = [K, K+\Delta Y(\varepsilon)]$ and $m(y;\beta_0) = \beta_0$. Equation (5) can then be written as,

$$B = \int_{K}^{K+\Delta Y(\varepsilon)} f_{Y_0}(u)\ du \approx f_{Y_0}(K)\Delta Y(\varepsilon) = f_{Y_0}(K)K\left[\left(\frac{1-t_1}{1-t_2}\right)^{\varepsilon} - 1\right]. \quad (9)$$

Therefore, subsequent work that use equation (9) are implicitly restricting the distribution of Y_0 to be uniform in the bunching region. For small tax changes, $[(1-t_1)/(1-t_2)]^{\varepsilon} - 1 \approx \varepsilon \log[(1-t_1)/(1-t_2)]$. Substituting this into equation (9) gives

$$\varepsilon \approx \frac{B/f_{Y_0}(K)}{K \ln\left(\frac{1-t_1}{1-t_2}\right)} = \frac{B/f_Y(K^-)}{K \ln\left(\frac{1-t_1}{1-t_2}\right)}.$$

The affine assumption implicit in (8) is slightly weaker than the uniform assumption in (9) because it allows the counterfactual distribution f_{Y_0} to have a non-zero slope in the bunching interval. One can argue that these assumptions are only approximations over a "small" interval. The problem, however, is that without a priori knowledge of the elasticity, the size of the interval and the quality of the approximation are both unknown. Unfortunately, the elasticity is sensitive to the shape of f_{Y_0}, and thus mistakes on the distributional assumption can lead to substantial bias in the estimator (Blomquist and Newey, 2017; Bertanha et al, 2018; Coles et al, 2022), as we discuss in Section 2.2.3.

2.2.2 Original Notch Methods

Kleven and Waseem (2013) propose an method for the case of a notch that also makes the uniform assumption on f_{Y_0} over the bunching interval and the counterfactual distribution immediately to the right of the notch. Equation 5 in that paper results from taking the indifference condition in equation (3) above and substituting $N^I = Y^I(1-t_1)^{-\varepsilon}$:

$$\frac{1}{1+\Delta Y(\varepsilon)/K}\left[1+\frac{\Delta/K}{1-t_1}\right] - \frac{1}{1+1/\varepsilon}\left[\frac{1}{1+\Delta Y(\varepsilon)/K}\right]^{1+\frac{1}{\varepsilon}} - \frac{1}{1+\varepsilon} = 0.$$

Kleven and Waseem (2013) combine this equation with knowledge of Δ, K, and t_1, plus an estimate for $\Delta Y(\varepsilon)$ to numerically solve for the elasticity. They use the bunching mass and the uniform assumption on f_{Y_0} in equation (5) to obtain the estimate for $\Delta Y(\varepsilon)$,

$$\Delta Y(\varepsilon) = \frac{B}{f_{Y_0}(K)} = \frac{B}{f_Y(K^-)}.$$

Kleven and Waseem (2013) follow the previous literature on kinks and thus implicitly impose a parametric assumption on f_{Y_0} to obtain $\Delta Y(\varepsilon)$. 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 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.

2.2.3 Discussion of the Original Method's Assumptions

The fundamental empirical hurdle with bunching is that the counterfactual distribution is unobserved. In the case of notches, the interval over which agents at the notch would have been in the absence of the notch is given by the gap in the distribution. The same interval exists with kinks but is unobserved because everyone above the kink changes their behavior in reaction to the kink and fills in that gap. In the case with kinks, the interval length itself is unknown and could be large because it depends on the unknown elasticity that we wish to identify. The original bunching methods rely on trapezoidal or uniform approximations to the unobserved counterfactual distribution over the relevant interval. If it is true that the PDF of the unobserved distribution is an affine function over that interval, which is a parametric assumption, then approximations used in the original bunching methods become exact.

However, Blomquist and Newey (2017) demonstrate that these parametric assumptions may be too strong in many settings and that the estimates are very sensitive to small changes in the quality of the approximations or, equivalently, to small deviations in the unobserved PDF relative to the affine function. They also show that retrieving the elasticity without any assumption on the unobserved PDF beyond continuity is impossible. This does not mean that researchers must live with strong parametric conditions. The modern bunching methods discussed in Section 2.3 use flexible semiparametric assumptions instead, which applied researchers may find more natural.

In this section, we discuss the assumptions in the original bunching methods and explain how the estimates are affected by deviations from these assumptions. Figure 3 illustrates the problem of identification of ε. The distribution in Panel (a) is identified from the data, but the distribution in Panel (b) is not fully identified. The nature of the problem allows us to learn some aspects of f_{Y_0} given our knowledge of f_Y. 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. Second, the portion of f_{Y_0} to the right of $K + \Delta Y(\varepsilon)$, where ε is unknown, relates to the portion of f_Y to the right of K, up to scaling factors. Finally, the area under f_{Y_0} over the bunching interval $[K, K + \Delta Y(\varepsilon)]$ equals the bunching mass B, where B is observed because the distribution of Y is observed. We know B, but we do not know m, that is, the shape of f_{Y_0} in that interval. The only thing we know about m are the area underneath m and its values at the boundaries of the interval, up to a scaling factor. In other terms, $m(K) = f_{Y_0}(K) = f_Y(K^-)$ and $m(K + \Delta Y(\varepsilon)) = f_{Y_0}(K + \Delta Y(\varepsilon)) = f_Y(K^+)\left(\frac{1-t_2}{1-t_1}\right)^{\varepsilon}$.

A natural question to ask is whether it is possible to identify the elasticity without imposing restrictions on m other than continuity. The answer to this question is no and was demonstrated by Blomquist and Newey (2017). Their result can be explained graphically in terms of Figure 3. If m has an extremely high peak, we obtain area B by integrating over a very small interval $[K, K + \Delta Y(\varepsilon)]$, which translates into a small elasticity as in panel e of Figure 3; if m has a valley, we obtain the same area B by integrating over a much larger interval $[K, K + \Delta Y(\varepsilon)]$, which translates into a bigger elasticity as in panel f of Figure 3. The bottom line is that we may obtain any elasticity we want by using different shapes of m.

The same identification problem affects the original notch method because it relies on the same assumption that f_{Y_0} is uniform. However, Bertanha et al (2018) showed that there exists an alternative way to identify elasticities from notches that do not require further assumptions on m other than continuity. The superior identification potential in the notch case is unfortunately undermined by the difficulties brought upon by friction errors. We discuss the identification with notches and practical challenges in Section 2.5.1 below.

Despite the limitations, the original bunching methods were pioneers in the investigation of behavioral responses previously unstudied due to a lack of experimental or quasi-experimental data, panel data, instrumental variables, etc. Following these influential methods, a large body of empirical work developed in a variety of areas, including finance (Collier et al, 2021; Ewens et al, 2021a), labor (Jales, 2018; Cengiz et al, 2019; Goff, 2022), sports (Allen et al, 2017), electricity markets Ito (2014), real estate (Kopczuk and Munroe, 2015), and many others. As discussed above, additional structure (for example, more assumptions on the distribution of ability beyond simply continuity) or data variation (for example, credible control groups, different time periods, or individual characteristics that predict ability) is necessary to identify the elasticity. These needs have sparked a recent literature that aims to identify elasticities with different and weaker assumptions. We discuss these modern methods in Section 2.3 below.

2.3 Modern Methods

In our view, the modern bunching field can be seen as a wide expanse of methods aiming to leverage bunching while minimizing assumptions in specific contexts. This section can be read as a practical menu, where different bunching methods may be chosen according to which assumptions are admissible in the specific situation and the type of data available. In many cases, it is possible to provide different estimates for the same application using assumptions of varying strength. Several of these methods are readily implementable through the Stata package `bunching`, which is discussed in detail in Bertanha et al (2022b) and is available at Boston College's Statistical Software Components (SSC).

Most of the new developments in this literature leverage bunching that results from kinks in the budget constraint. For the most part, we follow this literature in this section and thus focus the exposition on the kink case. However, the solutions are general and could be applied to contexts with notches or other models with bunching; we discuss extensions to other models in Section 2.5.

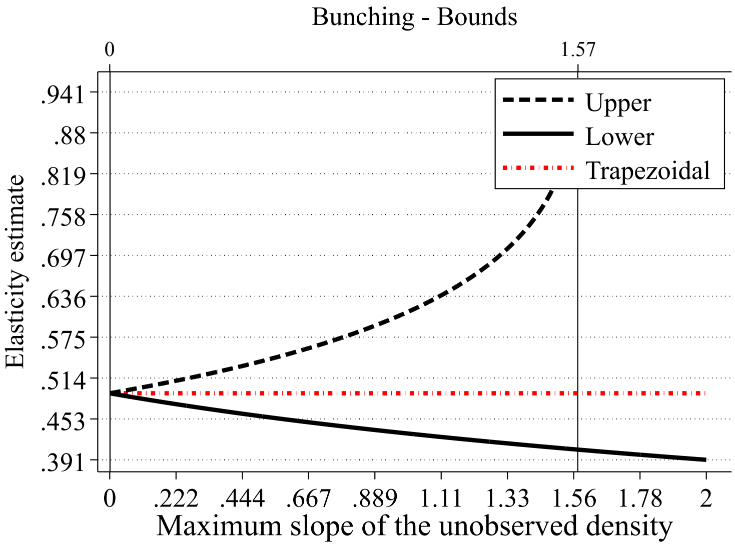
2.3.1 Bounds for the Elasticity with the Weakest Assumptions

In this section, we discuss partial identification of the elasticity under mild shape assumptions on the unobserved distribution of ability when there is a kink in the budget constraint. Blomquist and Newey (2017) showed that if the PDF of ability is monotone, then it is possible to obtain bounds on the elasticity. Bertanha et al (2018) proposed different bounds under the assumption that the magnitude of the slope of the PDF of log ability ($n = \log(N)$) is bounded by a known constant $M > 0$. More recently, Goff (2022) showed that bounds on the elasticity could also be obtained by assuming that the distribution of ability belongs to the large class of bi-log concave distributions. All of these approaches bypass the need for parametric assumptions on the distribution of ability.

The bounds in Bertanha et al (2018) and Goff (2022) have three attractive properties: closed-form expressions, yield a positive lower bound when the distribution of Y displays bunching, and nest the estimate based on the trapezoidal approximation. We focus our attention here on Bertanha et al (2018)'s bounds, which are easily computed in Stata using the command `bunchbounds`. An application of these bounds appear in Section 5 of Bertanha et al (2022c) using data from U.S. tax returns and various choices of M.

Figure 3 provides visual intuition for the bounds. Panel (d) in that figure displays the counterfactual distribution of income from Figure 1 and zooms in at the bunching interval (red lines). Panels (e) and (f) provide alternative counterfactual distributions of income that are consistent with the observed data on Y but have different shapes of m over the bunching interval. For a given value of the maximum slope M, these are the highest and lowest possible m functions, which correspond to the lowest and highest elasticity values, respectively. Figure 4 computes the bounds for various choices of M using the sample data included in the Stata package `bunching`; the true elasticity used to generate that data is 0.5. The red-dashed line shows the trapezoidal approximation estimate, which is always at the intersection of bounds at the lowest value of M. When the true slope of the distribution happens to be equal to that lowest value, the bounds are the tightest and the trapezoidal approximation estimate will retrieve the correct elasticity. When the true slope is large, it is possible to rationalize a much wider range of elasticities and the trapezoidal estimate may be far off. The bounds method provides an important sensitivity check for researchers identifying an elasticity with stricter assumptions. For example, Kostøl and Myhre (2021) utilize these bounds to examine the robustness of labor supply responses to incentives in the Norwegian welfare system (see their Appendix A, page 8).

Figure 4: Bounds for Various Choices of Maximum Slope



Notes: This figure plots the lower and upper bounds for various choices of the maximum slope parameter M. The graph was produced by the Stata command `bunchbounds` using the included sample data where the true elasticity is 0.5.

2.3.2 Point Identification when Plausible Control Groups are Available

Coles et al (2022) develop a method of using a control group to estimate the counterfactual distribution. To implement this method, researchers need both richer data and to gauge the plausibility of the control groups in the context of their specific empirical setting.

For example, Hungerman and Ottoni-Wilhelm (2021) study a kink created by a tax deduction in Indiana for donations to universities. For a control group, they use the distribution of donations in other states that do not have that kink.

Hungerman and Ottoni-Wilhelm (2021) contrasts the control group method with methods in Saez (2010), Kleven and Waseem (2013), and Chetty et al (2011). Briefly, Saez (2010) uses the density on either side of the kink and extrapolates a linear line between them as the estimate of the counterfactual distribution (see the discussion in Section 2.2.1 of this trapezoidal approximation). Kleven and Waseem (2013) and Chetty et al (2011) use a polynomial estimation using additional data on both sides of the kink. An advantage of these methods is that round number bunching can be addressed by including dummy variables for round numbers.

Finally, the control group method estimates a polynomial with additional data from a group that does not experience a kink but is otherwise similar to the treated group. The estimates rely on data unaffected by the kink and over the relevant range for the estimation, which includes both the area around the kink where there is bunching and the area in the distribution where the individuals who bunch came from. Hungerman and Ottoni-Wilhelm (2021) provide several checks of the identifying assumption by (1) gathering qualitative data to check the balance between the treatment and control groups, (2) estimating bounds based on potential heterogeneity across states, and (3) providing several placebo tests.

The change in the running variable $\Delta Y(\varepsilon)$ is identified using the control group, treatment group, and the relationship between them. Specifically, from equation (5), $B = \int_{K}^{K+\Delta Y(\varepsilon)} f_{Y_0}(u)\, du$, where f_{Y_0} is identified from the control group, B is identified from the treatment group, and we solve for $\Delta Y(\varepsilon)$. The estimate of the change in the running variable can then be used to calculate the elasticity, which is valid for small or large price-change kinks (Hungerman and Ottoni-Wilhelm, 2021). Because the control group method estimates the change in the running variable, Coles et al (2022) reports it in their main results and demonstrates how to calculate the elasticity with different assumptions on the kink size.

The control group can face kinks in their distribution as long as they are not affected by a kink local to the focal kink in the treatment group. 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. Therefore, firms with \$11,000 in NOLs 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. Similarly, Gelber et al (2020) use 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. Alternatively, policy changes can be used to generate control groups (Hamilton, 2018), though bunching may be persistent across time, and other dynamic effects may need to be accounted for (Marx, 2022).

2.3.3 Point Estimates with Covariates or Flexible Distribution Assumptions

Bertanha et al (2022c) propose two methods that achieve point identification of the elasticity with substantially more flexible distributional assumptions than those adopted by the original bunching methods. They apply these methods to data on U.S. tax returns from self-employed individuals and find that elasticity estimates for self-employed and not married individuals are robust across methods.

The key insight of these methods is to connect bunching to censored regression models. The first method is a truncated Tobit model. The second method is a censored quantile regression. These methods rely on different identifying assumptions and therefore provide complementary evidence. They could also be easily implemented using existing statistical software for censored regression models. In particular, the truncated Tobit model can be implemented using the custom-built Stata command `bunchtobit`, which is part of the package `bunching` (Bertanha et al, 2022b); an additional advantage of this method is that it provides visual diagnostics of the appropriateness of the distribution assumptions that it makes.

The first method adds structure to the problem by restricting the unobserved distribution of ability. With the assumption that the conditional distribution of log ability n given covariates X is normal, the elasticity could be estimated with a Tobit model — but this assumption is unnecessarily strong. Lemma 1 by Bertanha et al (2022c) provides sufficient conditions on the joint distribution of (n, X) for consistency of the Tobit elasticity estimator. These are semi-parametric assumptions on the distribution of (n, X) that do not require normality of the conditional distribution n given covariates X. Figure 5 displays an example of such distribution using data simulated from Experiment 2 in Section 4.2.1 by Bertanha et al (2022c). In short, these assumptions 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 unconditional distribution for log income y matches the observed distribution of y.

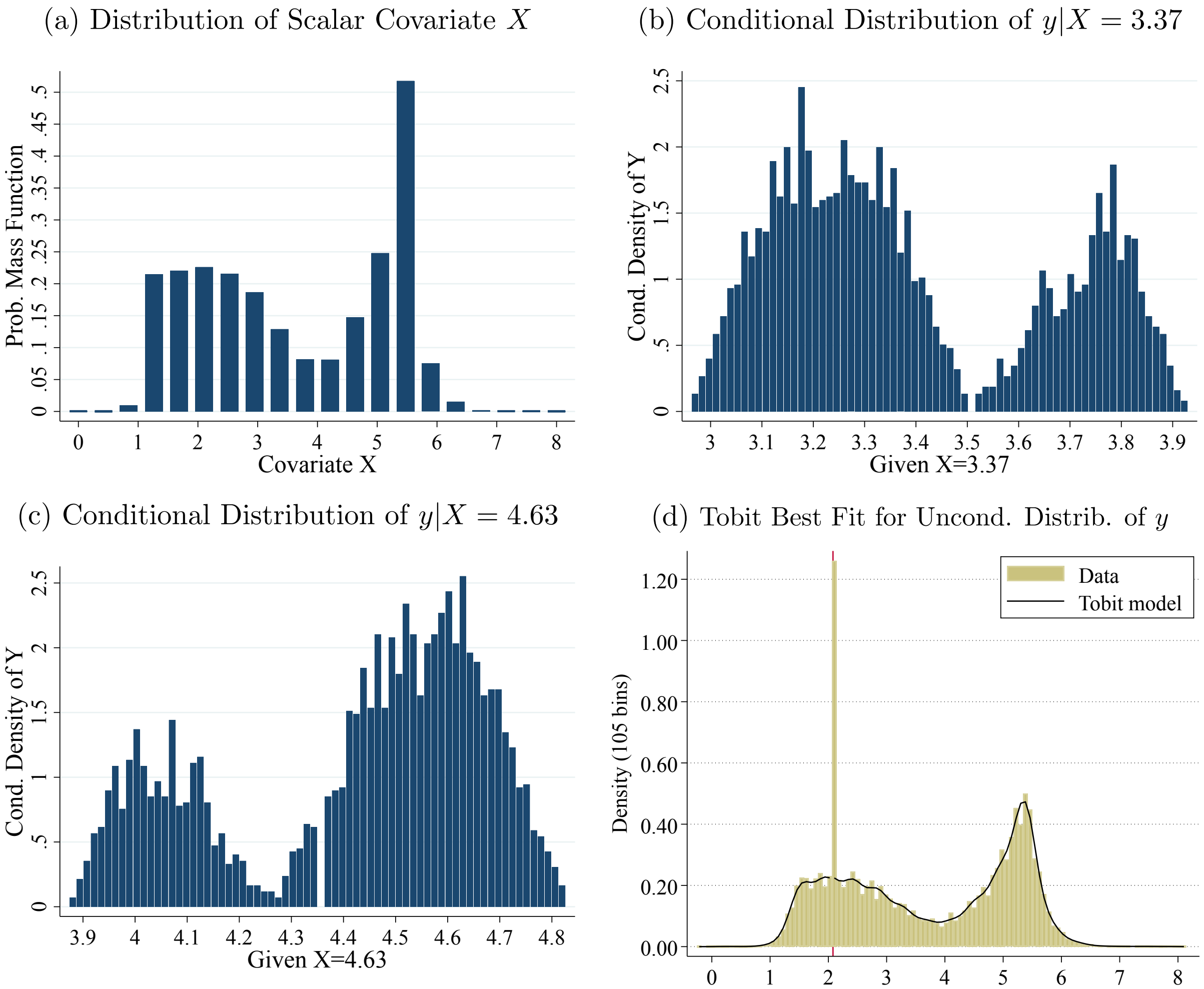
Assumption (i) becomes weaker with more variation in covariates because the class of distributions of n becomes richer the bigger the class of distributions of X. For small elasticities, one can show that assumption (ii) is implied by a linear assumption on the first two moments of the distribution of (n, X). Assumption (ii) is easy to verify in practice. Researchers simply need to compare the Tobit best-fit distribution of y to the observed distribution of y as a visual diagnostic of the appropriateness of Assumption (ii). Figure 5 demonstrates this visual diagnostic in a stark example of the Tobit model's robustness to a lack of normality. Panel (d) of 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 assumption (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, which makes it readily implementable in the vast majority of statistical software.

The original bunching methods relied solely on data near a kink or notch point to estimate an elasticity using local assumptions. Likewise, Bertanha et al (2022c) recommend estimating Tobit models using truncated samples of y at decreasing windows centered at the kink. The Stata command `bunchtobit` 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 decrease estimation precision as it relies on less data. The researcher can compare the Tobit best-fit distribution of y and the observed distribution of y as a visual diagnostic of whether the identification assumption is likely to hold in their specific context.

Figure 6 demonstrates this visual diagnostic using data simulated from Experiment 1 in Section 4.2.1 by Bertanha et al (2022c). To demonstrate how the truncation works, this example shows the Tobit best-fit distribution both with and without covariates; more truncation always improves fit, but fitting the distribution requires less truncation when we include the right covariates, that is, those that predict the distribution of y. The distribution of log ability is again asymmetric and has a pointed peak, which is far from Gaussian. The first row of panels in Figure 6 corresponds to the misspecified Tobit model, that is, the one that omits the correct covariate. 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

Figure 5: Consistency of the Tobit Estimator with Non-Normal Distributions



(a) Distribution of Scalar Covariate X

(b) Conditional Distribution of $y|X = 3.37$

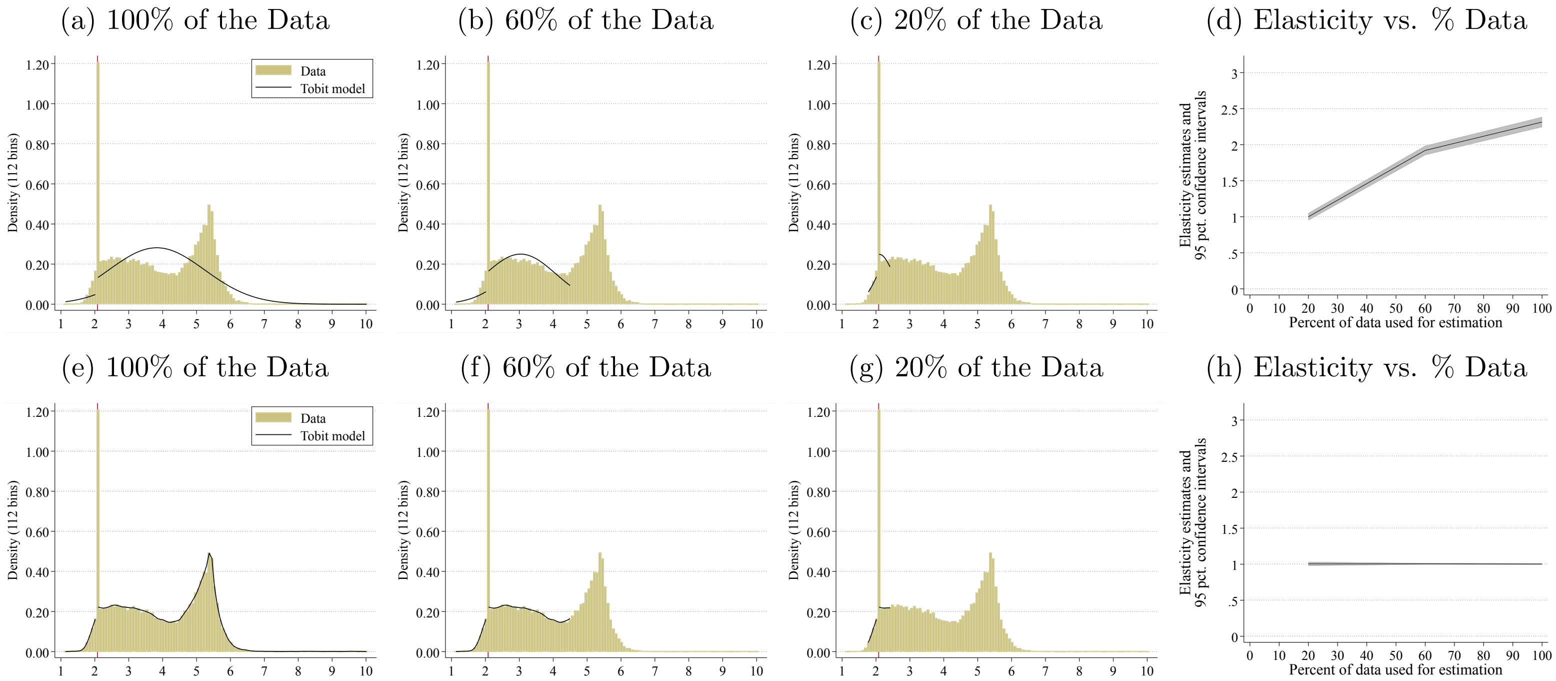
(c) Conditional Distribution of $y|X = 4.63$

(d) Tobit Best Fit for Uncond. Distrib. of y

Notes: The figure illustrates 50,000 observations simulated from the data-generating process in Experiment 2 by Bertanha et al (2022c). We refer the reader to Section 4.2.1 of that paper for the details. The variable log ability n is approximately a mixture of two Skewed Generalized Error Distributions, the kink point is at $k = 2.0794$, and $\varepsilon = 1$. Panel (a) shows the distribution of the discrete scalar covariate X. Panels (b) and (c) display the lack of normality in the conditional distribution of log income y given X for two values of X. Panel (d) shows the histogram of simulated data for log income y and the best-fit Tobit distribution that correctly includes covariate X. The last panel was produced using the Stata command `bunchtobit`.

elasticity estimate. The fit gets better with the smallest truncation window that uses 20% of the data, in which case the estimator recovers the true elasticity. The second row of panels in Figure 6 refer to the correctly specified Tobit model, that is, the one that includes the correct covariate. Even with a lack of normality, we see that it is possible to achieve a perfect fit and retrieve the correct elasticity without any truncation once we include the right covariate.

Figure 6: Truncated Tobit Method under a Non-Normal Distribution of Ability



(a) 100% of the Data

(b) 60% of the Data

(c) 20% of the Data

(d) Elasticity vs. % Data

(e) 100% of the Data

(f) 60% of the Data

(g) 20% of the Data

(h) Elasticity vs. % Data

Notes: The figure illustrates 50,000 observations simulated from the data-generating process in Experiment 1 by Bertanha et al (2022c). We refer the reader to Section 4.2.1 of that paper for the details. Log ability n is distributed as a mixture of two Skewed Generalized Error Distributions, the log-kink point is at $k = 2.0794$, and $\varepsilon = 1$. Panels (a)–(d) correspond to the fitting of a Tobit model that is incorrectly specified by omitting the covariate X. Panels (a)–(c) display the histogram of log income y and the best-fit Tobit distributions for three truncation windows. Panel (d) shows the elasticity estimate vs. the amount of data used in each truncated sample and 95% confidence intervals. Finally, Panels (e)–(h) repeat the exercise for a Tobit model that is correctly specified by including the covariate X. This figure was produced using the Stata command `bunchtobit`.

The second method relies on censored quantile regressions to identify the elasticity. The assumption is that the conditional quantile of log ability n given covariates X is a known parametric function of X with unknown parameters. 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 the conditional quantile of n given X, that is, a semi-parametric restriction on the distribution of (n, X). Section B.4 of the supplemental appendix by Bertanha et al (2022c) provides practical steps to estimate the elasticity building on Chernozhukov and Hong (2002).

2.4 Practical Issues on Identification and Estimation

There are three important practical considerations with bunching estimators. The first is that bunching estimators often rely on tuning and identification parameters. For example, in the original bunching estimators, researchers needed to choose the upper and lower bounds of the bunching region over which bunching was calculated, the income window to use to estimate the flexible polynomial for the counterfactual distribution, and the order of the polynomial in the counterfactual distribution. Coles et al (2022) find that the elasticity estimates using the control group method described in Section 2.3.2 were less sensitive to these parameter choices than the original bunching estimators in the context of firms bunching at kink points in the corporate tax schedule. Another approach is to favor methods with fewer tuning parameters. For example, Goff (2022) provides a different approach to bounding an elasticity that avoids tuning parameters. Researchers should test the sensitivity of their estimates to these parameters and consider using data-driven methods, such as cross-validation, to select them if the estimates are sensitive.

The second practical consideration is the existence of multiple kinks and notches in the budget constraint. When the budget constraint only has kinks and no notches, inference can be applied to each kink separately, as implemented in Goff (2022) and Agostini et al (2022). The reason that inference can be applied in this way is that the range of individuals that bunch at a kink point is unaffected by the kinks that precede it. The same is not true when the budget constraint has kinks and notches. In the case with notches, the range of individuals that bunch at a kink point may be affected by a preceding notch. It should be noted that, in some contexts, it is reasonable to assume that the elasticity is the same across all kink points. In this case, intersections of bounds at each kink point could be used to produce narrower bounds for the elasticity (Bertanha et al (2022c), Corollary 1). In other cases, the elasticity likely differs, and comparing elasticities across kinks is informative.

The third practical consideration is that the model of the observed income distribution necessarily abstracts from additional factors that may shape it. In some cases, the elasticity estimates are sensitive to these factors, and so these must 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. For example, adjustments to income might be lumpy, or there might be uncertainty about how income will come in. As a result of these frictions, the observed distribution of Y has increased mass at the kink and around it. If the observed distribution of Y has diffuse bunching, then not accounting for optimizing frictions could bias the elasticity estimate.

In order to incorporate optimizing frictions into the model, two different strategies are usually considered. The first strategy is to recover the income distribution if there are no optimizing frictions and then estimate the elasticity using the methods discussed in the previous sections. An ideal way to do this would be to model the distribution of income as a convolution of the distribution of income without friction errors and the distribution of optimizing errors and then deconvolute the two distributions. Dube et al (2018) use deconvolution to study heterogeneous parameters across firms using bunching in wages at round numbers. They are able to provide a distribution of parameters under a range of assumptions. They show that round-number bunching suggests that employers face optimization frictions and have some monopsony power. This remains an open area of research and Cattaneo et al (2018) are currently working on this problem. Without a comprehensive theory for optimizing errors, this approach remains infeasible. Less general but practical solutions have been proposed in their place. For example, Bertanha et al (2022c) develop a filtering procedure where optimizing frictions are modeled as an additive error in the optimal income variable. The 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 strategy is to incorporate a comprehensive theory of friction into the model. For example, Gelber et al (2013) uses policy-induced changes in the magnitude of kinks to estimate adjustment costs. Bertanha et al (2022c) propose an extension of their Tobit model that includes optimizing frictions, similar to how measurement error is incorporated into censoring models.

2.5 Extensions and Future Work

This section provides a description of preliminary work and extensions to the basic model.

2.5.1 Models with Notches in the Absence of Measurement Error

In some settings, bunching is sharp because agents do not experience adjustment costs or frictions, and there is minimal measurement error. In this case, Bertanha et al (2018) showed how to non-parametrically identify the elasticity using notches, that is, without having to assume more than the continuity of the distribution of ability N. As shown in Section 2.1, the key to the identification of elasticity in the notch case is that – in the absence of frictions – there is a gap in the distribution of Y, between K and $Y^I = Y(N^I) = Y(\overline{N})$, which is depicted in Panel (b) of Figure 2. The magnitude of this gap is informative of the elasticity.

In equation (2), we have $Y^I = N^I(1 - t_1)^{\varepsilon}$. The left-hand side is observed from the gap in the data and the right-hand side is a function of ε and observed quantities. Thus, we can solve for ε, and the gap in the distribution of earnings non-parametrically identifies the elasticity. Under the isoelastic functional form assumed for the utility function, there is no closed-form expression for the elasticity as a function of the observed objects but the elasticity can be obtained from knowledge of the other objects by means of standard numerical procedures. For more details, see Theorem 1 in Bertanha et al (2022c) (a result that first appeared in the 2018 version of that paper). Note that the source of identification is different from the one in Kleven and Waseem (2013), which uses the bunching mass instead.

The identification strategy in the case of notches is of limited practical use in many empirical settings. The issue is that in most applications, we can expect a portion of the workers to be unaware of details of the tax code, or they could face cost adjustments or other frictions that prevent them from reacting to the tax schedule. In these instances, the income distribution will display both bunching and some "missing mass," but no gap. The lack of a sharp observable dominated region makes estimation based on the notch identification strategy difficult in practice. Given the prevalence of notches examples, this is an area where more research is needed.

2.5.2 Models with Kinks From Decreasing Tax Rates

In some settings, agents face marginal incentives that decrease at a kink point. This would be the case if the marginal tax rate decreased for incomes greater than a given threshold. Bajari et al (2017) and Einav et al (2017b) study a similar problem in the context of hospitals' health insurance schemes. Using the notation from equation (1), we have $t_1 > t_2$ and $\Delta = (t_1 - t_2)K$.

In this case, the solution for Y has two regimes in terms of agent type N, as opposed to three regimes as in equation (2). Specifically, the income agents report depends on whether they are below or above a certain threshold $\underline{N}$,

$$Y = \begin{cases} N(1-t_1)^{\varepsilon} & \text{, if } \; 0 < N \leq \underline{N} \\ N(1-t_2)^{\varepsilon} & \text{, if } \; \underline{N} < N, \end{cases} \tag{10}$$

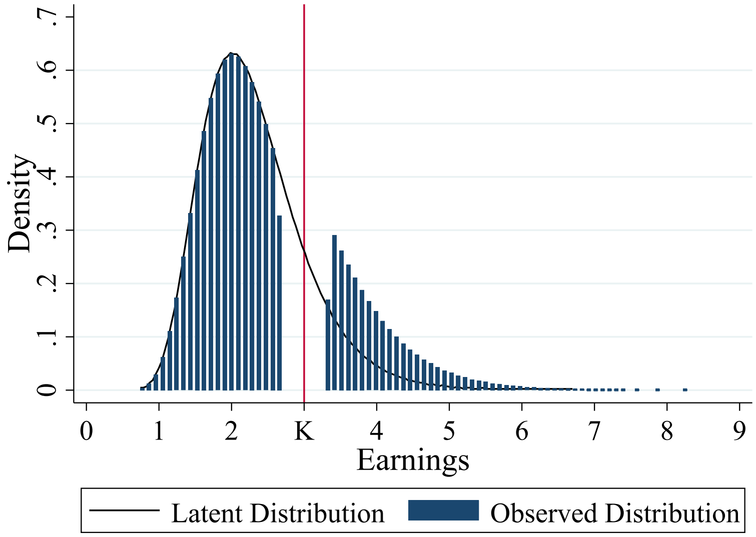
where

$$\underline{N} = (1+\varepsilon)K(t_2 - t_1)/\left[(1-t_1)^{\varepsilon+1} - (1-t_2)^{\varepsilon+1}\right]. \tag{11}$$

The distribution of Y is continuous except for an interval around the kink point where the distribution of Y has zero mass (see Figure 7 for an example). The lower and upper bounds of that gap can be written as nonlinear functions of observable quantities and the elasticity. Specifically, $\underline{Y} = \underline{N}(1-t_1)^{\varepsilon}$ and $\overline{Y} = \underline{N}(1-t_2)^{\varepsilon}$. Note that $\overline{Y} > \underline{Y}$ because $t_1 > t_2$.

Note, $\underline{Y}$ and $\overline{Y}$ are observable in the data—they are the beginning and the end of the gap, that is, the region of zero mass in the distribution of Y. At the same time, we can substitute the closed form solution for $\underline{N}$ from equation (11) into $\underline{Y}$ and $\overline{Y}$ to find two conditions relating observables and the unknown elasticity. Taking the difference in logs of $\underline{Y}$ and $\overline{Y}$ and diving it by the difference of logs of $1 - t_1$ and $1 - t_2$ yields the elasticity (Section A.5 by Bertanha et al (2022c)). Unlike the case with a kink and increasing tax rates, it is possible to non-parametrically identify the elasticity without having to assume more than just continuity on the distribution of N. The difference is the gap in the distribution, which allows for this non-parametric identification. Note, however, that this identification strategy requires that there exist no measurement error or other frictions (see

Figure 7: Earnings Under a Kink With Decreasing Tax Rates



Notes: Picture generated using $Y_0 = (1 - t_1)^{\varepsilon} N$ for the counterfactual income and equation (10) for the observed income in the case of a kink from increasing tax rates. We set $t_1 = 0.35$, $t_2 = 0.11$, $K = 3$, $\varepsilon = 0.5$, and assume that N is distributed as log-normal with mean 1 and variance 0.09.

Section 2.4 for a discussion of these frictions).

To further develop the intuition as to why the gap identifies the elasticity when there is a decrease in marginal tax rates, it is useful to think of a slightly different setting. Suppose that the policymaker was to introduce two different tax regimes based on skill levels. Individuals with a skill level above a certain threshold $\underline{N}$ would face a lower tax rate, whereas individuals with a skill level below $\underline{N}$ would face a higher tax rate.

If you look, however, at the solution to the worker's problem with a negative kink on tax rates, it is essentially the same setting. That is, a discontinuous drop in marginal tax rates for income above a certain threshold will act *as if* there is a threshold level of skill $\underline{N}$ that divides the population into two distinct groups that face different tax rates.

If we had access to data on *both* skill levels N and optimal income choices Y, then it would be natural to use a Regression Discontinuity Design (RDD) strategy to obtain the causal effect here. To begin, let's write the equation for potential outcomes under two distinct proportional tax regimes (note that these equations are the logs of the optimal choices for income under two distinct tax rates, according to the model; see Figure 8 below):

$$\log(Y_t(N)) = \log(N) + \varepsilon \log(1-t)$$

That is, if an individual that was to be faced with a marginal tax rate t would have log income according to the equation above. Since individuals with skill above $\underline{N}$ face tax rate t_2 and individuals with skill level below $\underline{N}$ face a different (higher) tax rate t_1, we have that the observed income choices of individuals in this setting, as a function of skill N, are given by:

$$\log(Y(N)) = \begin{cases} \log(N) + \varepsilon \log(1-t_1) & \text{, if} \quad N \leq \underline{N} \\ \log(N) + \varepsilon \log(1-t_2) & \text{, if} \quad N > \underline{N}, \end{cases} \tag{12}$$

Or, written differently,

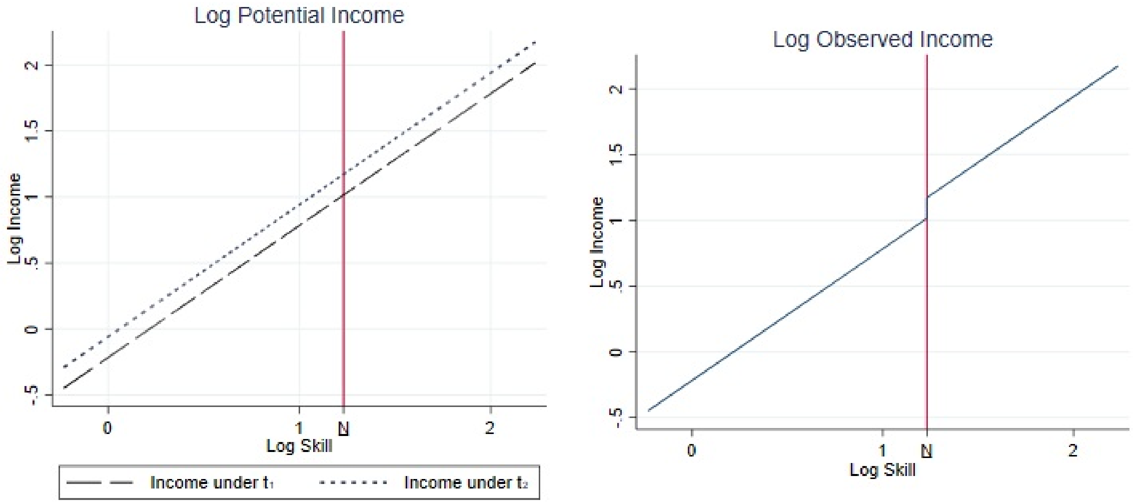
$$\log(Y(N)) = \log(N) + \varepsilon \log(1-t_1) + \epsilon \mathbb{1}\{N > \underline{N}\} \log(1-t_2)/(1-t_1).$$

Note that above $\underline{N}$, the potential income curve at marginal tax rate t_1 is observed, whereas below $\underline{N}$, the potential income curve at a different marginal tax rate t_2 is observed. The *vertical difference* in height between these two curves at each and every level of ability $\log(N)$ is precisely the causal effect of the change in taxes – from t_1 to t_2 – on the log of income of individuals with skill level N. This suggests that at $\underline{N}$, an RDD strategy could potentially be used. Taking lateral limits above and below $\underline{N}$, the sharp RDD estimand is then:

$$\lim_{n \downarrow \underline{N}} E[\log(Y)|N=n] - \lim_{n \uparrow \underline{N}} E[\log(Y)|N=n] = \varepsilon(\log(1-t_2) - \log(1-t_1)).$$

That is, the (sharp) RDD estimand identifies the product of the elasticity by the difference in marginal tax rates. Note also that, without optimization frictions or

Figure 8: Income as a function of ability under distinct tax rates



Notes: Picture generated using $\log(Y(N)) = \log(N) + \varepsilon \log(1-t)$ for the potential income under tax rates $t_1 = 0.35$ and $t_2 = 0.11$, and elasticity $\varepsilon = 0.5$.

measurement errors in the outcome, the conditional mean operator is not needed, since income will be a deterministic function of ability, given the elasticity and tax parameters.

Since this difference between tax rates is known (it is essentially the size of the "first-stage" RDD coefficient of marginal tax rates as the dependent variable and skill N as the running variable), we can easily obtain the elasticity as:

$$\varepsilon = \frac{\lim_{n \downarrow \underline{N}} E[\log(Y)|N=n] - \lim_{n \uparrow \underline{N}} E[\log(Y)|N=n]}{\log(1-t_2) - \log(1-t_1)}.$$

If we had access to data on the pair of skill level N and income choice Y, we could use a standard RDD strategy to recover ε, by looking at the discontinuity of income at the threshold of skill N under which marginal tax rates change discontinuously.

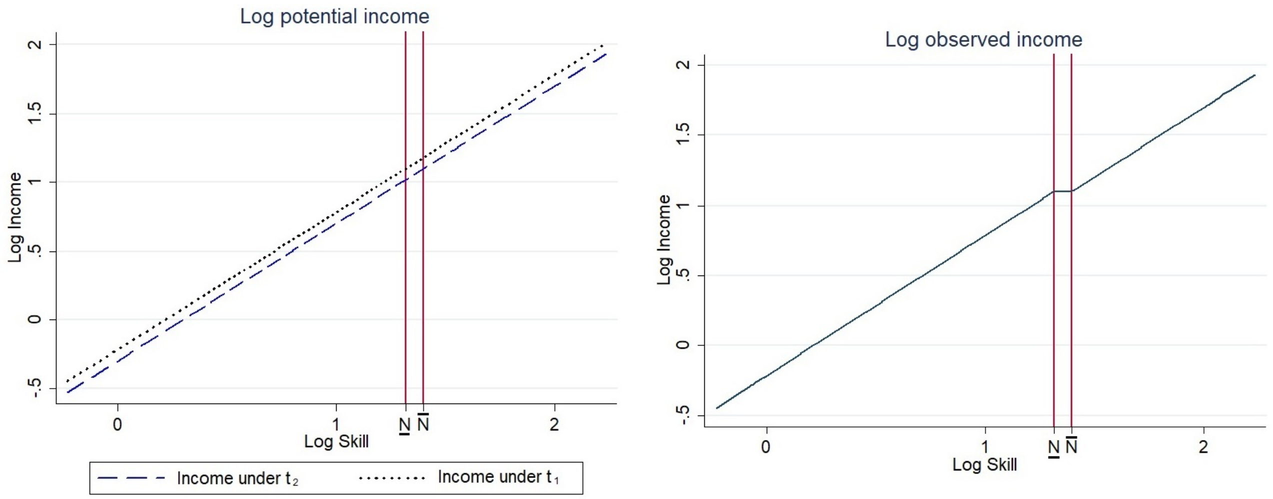
It is useful to note that the magnitude of such discontinuity in the conditional mean of income given skill around $\underline{N}$ is actually *identical to the gap in the distribution of the outcome variable* (income). Thus, even when we do not have the ideal data to run this RDD, the elasticity is still identified when there are no frictions. This comes from the fact that the gap is identical to the intention to treat estimand from the RDD and the denominator is given by the known size of the change in marginal tax rates. Thus, we can recover the elasticity since both the numerator and the denominator of the Wald ratio are identified, even with only access to data on income and no data on the "running variable" (ability) (see Jales and Yu (2017) for a related discussion).

Intuitively, we would like to compare individuals with arbitrarily close skill levels – so close that they are comparable to one another–, but subject to different marginal tax rates. In doing so, their difference in income could be attributed to differences in the responses to the tax rates that they face. A RDD strategy would search for discontinuities in the tax rate faced by individuals with skill levels around $\underline{N}$. Individuals to the left of $\underline{N}$ are subject to larger taxes, whereas individuals to the right are subject to lower taxes. In the vicinity of $\underline{N}$, skill levels are roughly the same, so differences in income are mostly coming from differences in marginal tax rates.

Without data on skill, using only data on income, we can argue that with a continuous distribution of skill, the individual with the highest income that is faced with the higher tax rate t_2 is going to have a skill level that is arbitrarily close to the skill level of the individual with the lowest income among those that are subject to the lower tax rate t_1. Given that there is essentially no difference between their skill levels, the difference in their income must be coming entirely from the fact that the former is making choices under a higher marginal tax and the latter is making choices under a lower marginal tax. That is the variation that the gap exploits to identify the elasticity.

This discussion also helps to explain why even a zero mean noise creates problems for the identification that uses the gap. Adding random noise in Equation 12 to represent optimization frictions would not create any issue for an RDD strategy since the RDD estimand would be the discontinuity in the *conditional expectation* of $\log(Y)$ around $\underline{N}$. However, if one attempts to identify the same object without data on N – that is, by looking at the gap in the distribution of income–then any standard form of measurement error would hide the gap from the observed distribution of income.

Figure 9: Income as a function of ability under distinct tax rates – The case of the Kink



Notes: Picture generated using $\log(Y(N)) = \log(N) + \varepsilon \log(1-t)$ for the potential income under tax rates $t_1 = 0.35$ and $t_2 = 0.45$, and elasticity $\varepsilon = 0.5$.

2.5.3 The Kink Case, Revisited

When the taxes increase above the threshold, however, the setting is slightly different. The equations that characterize the potential income under both marginal tax rates are still the same as before. That is, for any level of tax t, we have that $\log(Y_t(N)) = \log(N) + \varepsilon \log(1-t)$. However, the graph of the relationship between skill N and income Y displays a key difference, as one can see in Figure 9.

$$\log(Y(N)) = \begin{cases} \log(N) + \varepsilon \log(1-t_1) & \text{, if} \quad N < \underline{N} \\ \log(K) & \text{, if} \quad \underline{N} < N < \overline{N} \\ \log(N) + \varepsilon \log(1-t_2) & \text{, if} \quad N > \underline{N}, \end{cases} \tag{13}$$

Note that even if we had access to data on the pair of skill levels N and income choices Y, we would still need a parametric assumption (such as linearity) on the curves to identify the vertical difference between them. This setting, in fact, would look quite a lot like a donut type of RDD, in which the data around the threshold (here, the bunching mass) is excluded from the analysis (Dowd, 2021) and functional form extrapolation arguments are usually required.

This is the case because, in contrast with the previous example, there is not a single neighborhood of values of ability N in which we observe the individuals' behavior under two different tax regimes. When the tax falls after a certain value of income, then there is a skill level $\underline{N}$ such that individuals to the left of it – but arbitrarily close to it – behave as if they face proportional tax t_2, whereas individuals to the left – but again arbitrarily close to it – behave as if they face a proportional tax t_1.

When taxes rise after a certain income level, individuals to the right of a certain skill level $\underline{N}$ will behave as if they face a proportional tax t_2, and individuals above a *different* level of skill $\overline{N}$ will behave as if they face a different proportional tax level t_1. A group of individuals with skill levels between these values will simply set their income to K and, by doing so, will not behave in the way they would behave according to either one of these different taxes – neither t_1 nor t_2–, if the tax code were to be continuous.

It is useful to note that the gap allows us to see individuals with virtually the same skill making optimal choices – interior optimal choices, the same that would prevail under a proportional tax regime – under two different but continuous tax rates, whereas the bunching does almost the opposite. It allows us to see individuals with the same *income* but with very different skill levels. It is clear that the first kind of contrast is more informative of the elasticity than the second. This also helps to explain why (non-parametric) identification is possible in the presence of gaps but not when there is only bunching.

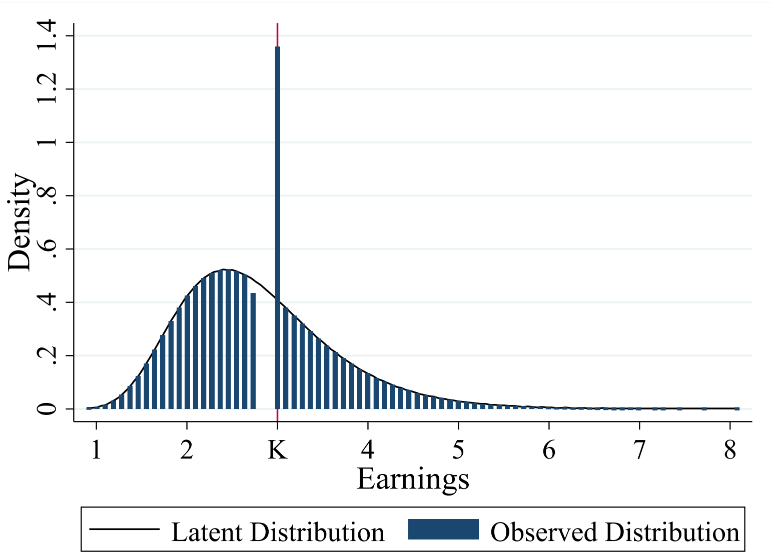
2.5.4 Models with Notches From Lump-sum Subsidies

In some settings, agents face a lump-sum subsidy at a notch point. This would be the case if agents received a benefit from crossing 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 setting, the worker faces a tax-liability function of the form $T(Y) = tY + \mathbb{1}\{Y \geq K\}\Delta$, where $\Delta < 0$ is the discrete decline 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 decrease discontinuously, creating a notch.

The solution to the worker's optimization problem has three regimes in terms of agent type N, as in equation (2). The difference, however, is that the gap presents itself in the distribution of earnings below K, as opposed to above it. Figure 10 displays an example of such a setting.

Figure 10: Earnings Under a Notch With Lump-Sum Subsidy



Notes: Picture generated using $Y_0 = (1 - t_1)^\varepsilon N$ for the counterfactual income and equation (2) for the observed income in the case of a notch from a lump-sum tax benefit. We set $t_1 = t_2 = 0.1$, $K = 3$, $\Delta = -0.05$, $\varepsilon = 0.2$, and assume that N is distributed as log-normal with mean 1 and variance 0.09.

The underlying relationship between the worker's ability N and the chosen level of income Y is given by equation (2), except that now $\underline{N} = N^I$ and $\overline{N} = K(1 - t)^\varepsilon$, where N^I is the lowest level of ability that bunches. The ability level N^I is determined by an indifference condition, similar to equation (3), and corresponds to income level $Y^I = N^I(1-t)^\varepsilon$.

The distribution of Y is continuous below Y^I, there is a region with zero mass, or a gap, between Y^I 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, Y^I, is identified. Using $Y^I = N^I(1-t)^\varepsilon$ and the implicit relationship between N^I and ε allows for non-parametric identification of ε along the same lines of Section 2.5.1. It is worth noting that, although the elasticity is identified from the gap, the practical usefulness of such a result is curtailed if there are adjustment costs and frictions. See the discussion in Section 2.4.

2.5.5 Models with Firms

In some settings, the agents are not individuals but firms (or something else), and thus the model needs to be adapted. Coles et al (2022) and Agostini et al (2022) both consider this type of setting. Here we show how the basic bunching model can be extended to firms by discussing the setting in Agostini et al (2022). We discuss the model in Agostini et al (2022) to demonstrate how the basic bunching model can be extended to firms.

Firms differ from individuals in that they often have negative income in a given year. Therefore, the model for individuals in Section 2.1 that implicitly assumes income is always positive is a bad approximation for firms.

Agostini et al (2022) begins with a two-period model where firms choose capital in period 2, K_2, to maximize shareholder value. Firms have productivity A_i and fixed costs F_i. Profits that are net of depreciation costs in the second period are given by,

$$Y_i(K_2) = \frac{1+e}{e} A_i^{1/(1+e)} K_2^{\frac{e}{1+e}} - F_i.$$

Firms are subject to the tax rate t_0 if their taxable income is below the kink κ and t_1 if their taxable income is above the kink, where $t_0 < t_1$.

Firms maximize shareholder value that is given by

$$\begin{aligned} max_{K_{2,i}} \quad V =& K_{1,i} - \frac{r}{1+r} K_{2,i} \\ &+ \mathbb{I}(Y_i(K_{2,i}) \leq \kappa) \frac{(1-t_0)Y_i(K_{2,i})}{1+r} \\ &+ \mathbb{I}(Y_i(K_{2,i}) > \kappa) \frac{(1-t_0)\kappa + (1-t_1)(Y_i(K_{2,i}) - \kappa)}{1+r}, \end{aligned} \tag{14}$$

where $\mathbb{I}(Y_i(K_{2,i}))$ and $\mathbb{I}(Y_i(K_{2,i}))$ are indicator functions for taxable income being below or above the kink and r is the discount rate.

The resulting distribution of taxable income for firms can be written in three pieces,

$$Y = \begin{cases} \frac{1+e}{e} r^{-e}(1-t_0)^e A - F, & A \leq \underline{A} \\ \kappa, & \underline{A} < A < \overline{A} \\ \frac{1+e}{e} r^{-e}(1-t_1)^e A - F, & A \geq \overline{A}. \end{cases} \tag{15}$$

The thresholds are found by setting the optimal taxable income equal to the kink κ with both tax rates: $\underline{A} = (\kappa + F)/(1-t_0)$, and $\overline{A} = (\kappa + F)/(1-t_1)$.

The distribution of firms features bunching at the kink due to the discontinuous incentives created by the tax rate increase at the kink. If the kink is at a positive income level, then the methods described in Section 2.3 can be used to recover the elasticity of taxable income for firms. However, Agostini et al (2022) consider the kink at \$0, where firms' statutory tax rates typically change from zero to positive. The models in Section 2.1 do not cover this case, therefore, to recover the elasticity, that paper develops two methods building on the methods in Section 2.3.

2.5.6 Models With a Notch and Without Knowledge of the Budget Constraint

In some settings, the budget constraint parameters are unknown and are often the parameters of interest. A typical question in this area is the size of a given notch (Bertanha et al, 2022a). Another example is Ewens et al (2021b), which estimates the cost of disclosure and governance regulations using a threshold based on a firm's level of equity. This type of question is common in finance and accounting settings, where due to a threshold, there is bunching in the observed distribution, and the object of interest is the change in incentives causing that behavior.

The approach taken by Ewens et al (2021b) is to use a model that has been previously calibrated from the literature. Then, by combining this model and bunching methods, the costs of regulation can be recovered. Specifically, their model consists of a set of firms choosing equity Y subject to a regulation that imposes a cost of t if their equity is greater than some threshold K. In the absence of the regulations, firms would choose Y^*, and firms bear a cost for the deviation of equity from its level without the regulation according to the penalty function $\Phi(Y; Y^*)$. The objective of the firm is to maximize,

$$\max_Y - \Phi(Y; Y^*) - t\mathbb{I}(Y > K). \tag{16}$$

There is a marginal firm that is indifferent between bunching $Y = K$ and setting their equity to be what it would have been in the absence of the notch $Y = Y^*$. The change in equity for this marginal firm can be recovered using bunching methods; then it can be used to calculate the regulatory costs given the indifference condition for the marginal firm,

$$t = \Phi(K; Y^*). \tag{17}$$

Ewens et al (2021b) approach works well if the other parameters, for example, the penalty function, are known from the literature. Bertanha et al (2022a) take a different route and do not assume that these parameters are known. Instead, Bertanha et al (2022a) build on the methods in Section 2.3 to provide estimates of the size of the notch.

2.5.7 Models With Multiple Kinks

In some settings there exist multiple kinks, perhaps across years, that can be used to isolate the parameters of interest. Denning et al (2023) provide an example in the context of student loans. Specifically, students are subject to a discontinuity in their interest rates based on how much debt they take out because they exhaust subsidized loans with lower interest rates. In their case, they observe changes in the interest rates across years and use this additional variation to estimate an elasticity and a misperception parameter that affects the perceived interest rate change.

Following Denning et al (2023), consider a two-period model where individuals choose how large of a loan L to take out. Individuals are heterogeneous in their utility preferences over debt, parameterized by N. They have exogenous income in both periods, Y_1 and Y_2, which they use to buy consumption in periods 1 and 2, C_1 and C_2 respectively, with a discount factor β. The interest rate that individuals face depends in part on how large of a loan they take out. Specifically, for $L < K$, the interest rate is r_1, and for $L > K$, the interest rate is r_1 for the first K amount and r_2 afterward. Individuals, however, may misperceive the interest rate for large loans to differ from r_2 by a factor θ.

Individuals choose L to maximize their utility over two periods by solving the problem

$$\begin{aligned}
\max_{L} \quad & C_1 + \beta C_2 - \frac{N}{1+1/\varepsilon}\left(\frac{L}{N}\right)^{1+\frac{1}{\varepsilon}} \\
s.t. \quad & \\
& C_1 = Y_1 + L \\
& C_2 = Y_2 - \mathbb{I}\{L \leq K\}(1+r_1)L - \mathbb{I}\{L > K\}\left[(1+\theta r_2)L - (r_2 - r_1)K\right].
\end{aligned}$$

The first-order condition determines each individual's loan amount as a function of their heterogeneous parameter N, the elasticity ε that determines the utility cost of the loan, and other parameters:

$$L = \begin{cases} N\left(1-\beta(1+r_1)\right)^{\varepsilon}, & \text{if} \quad N < K\left(1-\beta(1+r_1)\right)^{-\varepsilon} \\ K, & \text{if} \quad N \in \left[K\left(1-\beta(1+r_1)\right)^{-\varepsilon}, K\left(1-\beta(1+r_2)\right)^{-\varepsilon}\right] \\ N\left(1-\beta(1+\theta r_2)\right)^{\varepsilon}, & \text{if} \quad N > K\left(1-\beta(1+\theta r_2)\right)^{-\varepsilon}. \end{cases}$$

The methods discussed in Section 2.3 recover εs_0 and εs_1. In the typical case, s_0 and s_1 are observed and the elasticity can be recovered. In this case, the misperception parameter θ can be recovered by taking the ratio of the estimates. The elasticity in the numerator and denominator cancels out, thus identifying the misperception parameter as the ratio of the observed interest rates.

2.5.8 Future Work

This subsection discusses additional directions for extensions, noting that our list is far from exhaustive. Many extensions will require bespoke models to capture key features of the novel setting. For example, Einav et al (2017b) develop a model in the context of prescription drug insurance for the elderly in Medicare Part D and Agostini et al (2022) develop methods to focus on a kink in the corporate tax schedule at zero. 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.

Three areas that have received attention are the extensive margin, dynamic effects, and decomposing elasticities into different components. Gelber et al (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. Pollinger (2021) allows agents to choose the amount of wattage of solar panels (the intensive margin) and participation in the subsidy program (the extensive margin). Identification in this case relies on a local analytic function building on results in Goff (2022); for example, see Proposition 6.

Le Maire and Schjerning (2013) and Marx (2022) extend bunching methods 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 bunching methods 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 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) use 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). Velayudhan (2018) uses a VAT notch where small firms are not required to file. This paper then considers how the distribution of firms would look like if the bunching was caused by real production changes or misreporting.

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 et al (2021b) and Bertanha et al (2022a) 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 2.2.3.

Hong (2023) extends the basic model with a single elasticity to a model with a distribution of elasticities. The paper estimates this distribution in the setting of medical expenditures in South Korea. It exploits variation from a control group to recover the conditional cumulative distribution below a certain elasticity. This extension allows the paper to provide more realistic counterfactual policy simulations. In this setting, Hong (2023) finds that patient welfare can be improved by replacing a notch with a linear rate structure.

Goff (2022) provides a generalization of current methods that captures bunching with multiple-choice variables. The key insight in this work is to recast the parameter of interest as a choice of counterfactuals rather than a preference parameter. Through this extension, Goff (2022) shows that the bunching design is more general than the typical isoelastic model typically employed. This work extends some of the earliest work, including Saez (2010) and Kleven (2016), which discusses a generalization with heterogeneous elasticities. See, for example, Lemma SMALL in Goff (2022), which explores the usefulness of a "small-kink" approximation. Finally, Blomquist et al (2015) explore quasi-concave models without a parametric form.

In some contexts estimating an elasticity may not be necessary to provide policy-relevant insights. Moore (2022) shows that the bunching mass can be used as a sufficient statistic for the revenue effect of behavioral responses to small changes of the threshold without making assumptions necessary to identify an elasticity. Goff (2022) shows the effect of a marginal change in the threshold on bunching as well as on mean counterfactual choices are also point identified without any extrapolation assumptions. Future work can apply these methods to provide policy-relevant estimates and extend these methods to different types of policy changes.

Many parameters of interest can be identified using observed distributions and nonlinear budget constraints, incentives, or tax schedules. The bunching field is still full of examples where bunching in a variable has not yet been explored in applications. The menu of assumptions and data structures discussed in the previous sections offers some guidance about how one might approach a problem where bunching in the outcome variable is found and what can be identified in such a setting. The several applications in this section give a map of areas where the frontier is being pushed forward, where much is yet unknown.

3 Bunching in the Treatment Variable and the Smoking Example

A new branch of the literature focuses on how bunching in the treatment variable 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. Figures 11 and Figure 12 show 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 11: Education (years)

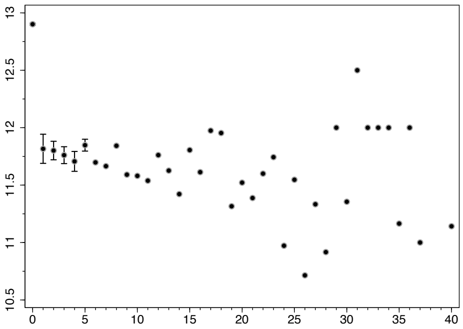
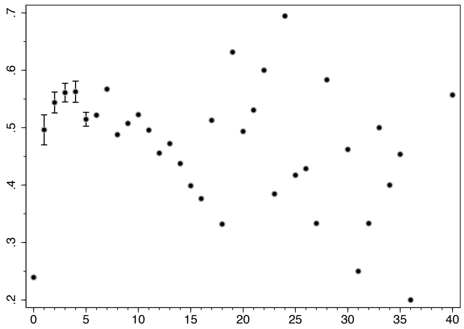


Figure 12: Likelihood Unmarried



Figures 11 and 12: 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.

3.1 Testing Identification Assumptions

Define $Y_i(t)$ as the potential outcome of observation i under treatment level t, and let T_i denote observation i's actual treatment. Suppose that $Y_i(t)$ is differentiable with respect to t, and denote its derivative $Y_i'(t)$. For a given vector of controls X_i, define

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

If T_i is exogenous, then the average conditional marginal treatment effect function $E[Y_i'(t)|T_i = t, X_i] = dE[Y_i|T_i = t, X_i]/dt$ can be identified. Estimation of the treatment effects depends on assumptions on $E[Y_i|T_i = t, X_i]$, as there are a wide range of options available in the selection-on-observables literature. For example, if $E[Y_i|T_i = t, X_i]$ is assumed to be linear on T_i and X_i, then the marginal treatment effect estimator is simply the coefficient of T_i on a regression of Y_i on T_i and X_i. A researcher interested in using such methods would therefore like to test if T_i is exogenous.

Caetano (2015) showed that, if the distribution of T_i has bunching at $T_i = \bar{t}$, it is possible to test the exogeneity of T_i. When one compares 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, since the treatment hardly varies. 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 T_i is endogenous). If one can assume that $Y_i(t)$ is continuous in t at $\bar{t}$ with probability one conditional on X_i, then the first possibility is ruled out. One can then test the exogeneity of T_i by checking if $E[Y_i|T_i = t, X_i] =$ is continuous in t at $\bar{t}$. This is because, if T_i is exogenous, then $E[Y_i|T_i = t, X_i] = E[Y_i(t)|T_i = t, X_i] = E[Y_i(t)|X_i]$ must be continuous in t at $\bar{t}$.

We translate the idea above to the smoking example. If the expected birth weight given the amount of smoking and controls is discontinuous at zero cigarettes, then either cigarettes have a discontinuous causal effect on birth weight, or unobservables are discontinuously different among those who do not smoke and those who smoke a marginally small amount. If we believe that marginal amounts of smoking have a marginal effect on birth weight, then the birth weight of those who do not smoke should be at most marginally different from the birth weight of those who smoke a very small amount. If the difference in birth weight is large, then the only possible explanation is differences in unobservables. This explanation implies that smoking is endogenous since, in that case, the unobservables of those who do not smoke are different from the unobservables of those who smoke a marginally small amount. This implies that, if the effect of smoking on birth weight is continuous at zero cigarettes, then the expected birth weight given the number of cigarettes and controls must be continuous at zero cigarettes.

This test makes sense when T_i has bunching because, as discussed above, confounders tend to be discontinuously different at the bunching point. Specifically, if T_i is endogenous and has bunching at $\bar{t}$, then $Y_i(t)|T_i = t, X_i$ will usually be discontinuous in t at $\bar{t}$. Therefore, if T_i is endogenous, $E[Y_i|T_i = t, X_i]$ will usually be discontinuous in t at $\bar{t}$. Summarizing, if T_i is exogenous, then $E[Y_i|T_i = t, X_i]$ is continuous in t at $t = \bar{t}$ and, if T_i is endogenous and has bunching at $T_i = \bar{t}$, $E[Y_i|T_i = t, X_i]$ is usually discontinuous in t at $t = \bar{t}$. We can thus test if T_i is endogenous by checking if $E[Y_i|T_i = t, X_i]$ is discontinuous in t at $t = \bar{t}$.

To implement this non-parametric test, the dimension of X_i is a concern. Caetano (2015) proposes aggregating the discontinuities across values of X_i, and testing instead if an average of the discontinuities is different from zero. A convenient aggregation explored in that paper yields the following testing quantity:

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

This strategy is particularly convenient when there is a large amount of bunching at $\bar{t}$, as is the case with smoking. Estimation can be done in a two step process: (1) estimate $E[Y_i|T_i = \bar{t}, X_i]$ non-parametrically (or, in practice, as non-parametrically as possible, using machine learning strategies or a kitchen sink regression) and (2) do a local linear regression of $\hat{E}[Y_i|T_i = \bar{t}, X_i] - Y_i$ onto T_i at $\bar{t}$, using only observations such that $T_i > \bar{t}$. The approach just described is known as the Discontinuity Test. See empirical implementations in Caetano (2015), Rozenas et al (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 (2021a), which is known as the Dummy Test. This test is suitable to cases where some semi-parametric assumptions on $Y_i(t)$ are made, and estimation takes these into account. For example, suppose that $Y_i(t) = \beta t + U_i$, where U_i is not observed, and one intends to estimate β as the coefficient of T_i in a regression of Y_i onto T_i and X_i. In this case, the main identification assumption is that $E[U_i|T_i, X_i] = X_i'\alpha$, which implies that $Y_i = \beta T_i + X_i'\alpha + U_i$, a standard linear model. 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 (this is because X_i can include fixed effects). The Dummy Test consists of adding the dummy $1(T_i = \bar{t})$ to the regression (i.e. regress Y_i onto T_i, X_i and $1(T_i = \bar{t})$) and implementing a simple t-test that the coefficient of $1(T_i = \bar{t})$ is significant.

The Dummy Test operates under the same principles as the Discontinuity Test, in leveraging the idea that, if T_i has bunching and is endogenous, the distribution of $U_i|T_i = t, X_i$ is likely to be discontinuous in t at $\bar{t}$. This would then generate a discontinuity in $E[Y_i|T_i = t, X_i]$ at the bunching point, which can be detected by including the dummy $1(T_i = \bar{t})$ in the regression. While the Discontinuity Test tests exclusively the exogeneity of T_i, the Dummy Test is a joint test of the exogeneity of T_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 et al (2018), Lavetti and Schmutte (2018), De Vito et al (2019), Caetano et al (2019), Kaneko and Noguchi (2020), Caetano et al (2021b), 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, and discrete-choice models. In all these cases, the main identification assumptions can be tested by including $1(T_i = \bar{t})$ in the set of controls and performing a simple t-test of the significance of its coefficient.

Similar ideas are leveraged in Khalil and Yıldız (2019) to build a test of the exogeneity of T_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 et al (2016) showed that bunching on the treatment variable can be used to test the validity of the instrumental variable in a triangular model.

3.2 Identifying Treatment Effects

In the general model discussed in the previous section, all observations at the bunching point have the same treatment. After controlling for observables, any variation in the outcomes of those observations must be due to the unobservables. Therefore, bunching provides a glimpse into the effects of confounder variation without contamination from the treatment variation. With some additional structure, it may be possible to use the variation at the bunching point to correct estimators so that they are consistent under 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. This is because there is often bunching at zero for variables that cannot be negative, as is the case with smoking, drinking, consuming coffee, almost all types of time use (e.g. exercising, studying), financial variables like debt, savings, specific types of investments, etc. Often law and other artificial restrictions generate this type of bunching, for example minimum wage, schooling, age to work laws, maximum 401K and Roth IRA contributions etc. For simplicity of notation, suppose that $\bar{t} = 0$ and is the lower extreme, so that $T_i \geq 0$. 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 T_i^*, as

$$T_i = T_i^* \cdot 1(T_i^* \geq 0), \quad \text{where } P(T_i^* < 0) > 0.$$

Although the approaches discussed below are agnostic about the structural interpretation of T_i^*, it is useful to think of this variable as the optimal choice in an unconstrained optimization problem. For example, in the maternal smoking example, we can suppose that the number of cigarettes T_i is decided as the result of a constrained optimization problem, where mothers maximize a utility function which takes into account her preferences, as well as her observable and unobservable characteristics, subject to any budget and other standard constraints, and additionally a constraint that T_i must not be negative. Then, T_i^* is the result of the optimization when the non-negativity constraint is lifted. In other words, T_i^* is the optimal choice the mother would have made if she took into account all factors of concern in her decision except the fact that one cannot smoke negative amounts. If the maximization yields $T_i^* \geq 0$, then she smokes that number, and thus $T_i = T_i^*$. If the maximization yields a negative number, then she smokes $T_i = 0$.

If conceiving of a negative amount of smoking is difficult, one can think of T_i^* as an ordinal index of the amount of indifference between smoking versus not smoking among the mothers when all factors are taken into account, including budget constraints. Therefore, a mother with $T_i^* = -2$ is closer to indifference between smoking versus not smoking than a mother with $T_i^* = -3$. Both mothers prefer not to smoke, but the latter mother would have to be paid more to smoke the first cigarette than the former mother. The condition $P(T_i^* < 0) > 0$ means that there are mothers who are not indifferent, and strictly prefer not smoking to smoking any amount.

Caetano et al (2020) consider the model

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

where

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

In this model, T_i and U_i are correlated, and therefore X_i is endogenous. However, the same relationship between T_i and U_i when $T_i > 0$ is also maintained between T_i^* and U_i when $T_i = 0$ (i.e. δ is the same for all values of T_i^*). This model states that, in the same way that T_i must be the index of all confounded variation for $T_i > 0$, T_i^* indexes all confounded variation in the bunching point. This structure implies

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

Although it is not possible to separate β and δ from the variation of T_i, the discontinuity of $E[Y_i|T_i, X_i]$ at $T_i = 0$ generated by the bunching reveals only the magnitude of δ, which can then be used to disentangle β. In other words, if it were feasible to estimate $E[T_i^*|T_i^* \leq 0, X_i]$, then $T_i + \hat{E}[T_i^*|T_i^* \leq 0, X_i]1(T_i = 0)$ could be added to the regression to correct for the endogeneity of T_i, in the sense that the coefficient of T_i in a regression of Y_i onto T_i, X_i and the "correction term" $T_i + \hat{E}[T_i^*|T_i^* \leq 0, X_i]1(T_i = 0)$ is a consistent estimator of β.

The same type of strategy can be used in other, more general, models also considered in Caetano et al (2020). See, for example, Caetano et al (2023), which implements a model with parametric and non-parametric correlated random effects to study the effect of the hours the mother works on the child's skills. This paper showcases one of the main advantages of leveraging bunching for identification instead of instrumental variables, in that the correction method is not prone to weak identification, and thus allows the division of the sample into subgroups to study heterogeneity. Caetano et al (2023) use this property to study heterogeneity of maternal labor supply effects by the mother's skills and pre-birth income. Another example can be seen in Caetano et al (2021b).

Caetano et al (2020) propose identifying $E[T_i^*|T_i^* \leq 0, X_i]$ using models on the shape of the conditional distribution of T_i^*. For example, if $T_i^*|X_i \sim \mathcal{N}(\mu(X_i), \sigma^2(X_i))$, for arbitrary functions μ and σ, then $E[T_i^*|T_i^* \leq 0, X_i] = \mu(X_i) - \sigma(X_i)^2 \cdot \lambda(-\mu(X_i)/\sigma(X_i))$, where $\lambda(\cdot)$ is the inverse Mills ratio (the PDF divided by the CDF of the standard normal distribution). A weaker assumption than normality is tail symmetry. If the distribution of $T_i^*|X_i$ is symmetric in the tails and $P(T_i = 0|X_i) \leq 0.5$, then $E[T_i^*|T_i^* \leq 0, X_i] = F_i^{-1}(1 - F_i(0))$ $-E[T_i|T_i \geq F_i^{-1}(1 - F_i(0)), X_i]$, where $F_i(t) = P(T_i \leq t|X_i)$.

All the quantities in the two cases above can be identified, and may be estimated with standard non-parametric methods, but Caetano et al (2020) propose a simpler empirical strategy in two steps: (1) discretize the X_i using hierarchical clustering (Hastie et al (2009)), and (2) do the estimation within each cluster under the assumption that, for all i

in cluster $C_i = c$, $E[T_i^*|T_i^* \leq 0, X_i] = E[T_i^*|T_i^* \leq 0, C_i = c]$ (i.e. assume that the expectation is the same for all observations within the cluster). In the normality case, the second step is equivalent to running a Tobit regression on a constant within each cluster c. The estimator of the constant is $\hat{\mu}(X_i)$ for all i such that $C_i = c$, and the estimator of the standard deviation is $\hat{\sigma}(X_i)$. 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 among observations in cluster c, $F_i^{-1}(q)$ as the quantile q of T_i among all observations in cluster c, and $E[T_i|T_i \geq a, X_i]$ as the mean of the $T_i \geq a$ among all observations in cluster c.

Caetano et al (2022b) considers partial identification strategies when the assumptions on the distribution of $T_i^*|X_i$ are relaxed. They show that a sharp bound on β can be obtained under no distributional assumption. An opposite sharp bound can be obtained under mild assumptions such as that, for $t \leq 0$, the density of $T_i^*|X_i$, $f_{T^*|X}(t)$, has no peaks larger than the right limit of the density of $T_i|X_i$ at the bunching point. Bounds can be narrowed if assumptions on $f_{T^*|X}(t)$ for $t \leq 0$ are strengthened, such as assuming concavity or convexity, both of which are testable conditions, or that $f_{T^*|X}$ belongs to families such as bi-log concave or log concave. These bounds are easy to calculate, and may be visually displayed in a compelling manner. For example, the plots in Caetano et al (2023) show that all the points in that paper stand even if no assumption is made on the distribution of $T_i^*|X_i$.

3.3 Future Work

Recently, Caetano et al (2022a) showed that it is possible to obtain non-parametric identification of treatment effects using bunching. Specifically, they show that $E[Y_i'(0)|T_i^* = 0]$, the average marginal treatment effect at the bunching point for the population with $T_i^* = 0$, can be identified if (1) the treatment effects are continuously differentiable at the bunching point; (2) the endogenous selection as a function of T_i^* $(Y_i(0)|T_i^* = t)$ is continuously differentiable at the bunching point; (3) the endogeneity bias is monotonic on T_i^* for $T_i^* \leq 0$; and (4) the distribution of the idiosyncratic variation conditional on T_i^* $(Y_i - E[Y_i|T_i^*])$ is right-continuous at the bunching point and independent of T_i^* at the bunching point. Identification is obtained by an innovative use of the change of variables theorem. They show that the bias of endogeneity for $T_i^* = 0$ can be written as the ratio of $\lim_{t \downarrow 0} f_{T|X}(t)$ and the density of the selection bias term evaluated at $T_i^* = 0$. The latter term can be identified through the distribution of the outcome at the bunching point: at the bunching point, any variation in outcome is due to the variation in $Y_i(0)$. The density of the selection can be deconvoluted from the density of the idiosyncratic noise term using the observations near the bunching point.

The advancements in Caetano et al (2022a) show that the potential of bunching as a source of identification is very promising. There is ample opportunity of contribution in the search for bunching identification strategies with weaker assumptions. Moreover, much remains to be done with regards to estimation of these models. For example, the use of clustering as a technique for bringing non-parametric flexibility to the standard models needs to be further studied. Additionally, the estimation in the non-parametric identification strategy in Caetano et al (2022a) uses limit deconvolution estimators which need to be studied, and may perhaps be improved or altogether avoided.

4 Summary

In this chapter, we review the literature on bunching methods. We discuss the limits of non-parametric identification of taxable earnings elasticity under continuity assumptions in the settings of kinks and notches. We also examine 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 growing branch of this literature that leverages bunching in the treatment variable in standard reduced-form causal models. We discuss how bunching in the treatment variable makes it possible to test for endogeneity and to correct for endogeneity without instrumental variables or panel data.

5 Acknowledgements

Responsible Section Editor: Alfonso Flores-Lagunes. The article has benefited from the valuable comments from the editors, anonymous referees, Andrew McCallum, Daniel Hungerman, and Mu-Jeung Yang. There is no conflict of interest.

References

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

Allen EJ, Dechow PM, Pope DG, Wu G (2017) Reference-dependent preferences: Evidence from marathon runners. Management Science 63(6):1657–1672, DOI {10.1287/mnsc.2015.2417}

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

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

Bertanha M, McCallum AH, Seegert N (2018) Better bunching, nicer notching, DOI http://dx.doi.org/10.2139/ssrn.3144539, working paper

Bertanha M, Gaulin M, Seegert N, Yang MJ (2022a) Estimating incentives using observed distributions of earnings, work in progress

Bertanha M, McCallum AH, Payne A, Seegert N (2022b) Bunching estimation of elasticities using Stata. The Stata Journal 22(3):597–624

Bertanha M, McCallum AH, Seegert N (2022c) Better bunching, nicer notching, DOI https://doi.org/10.48550/arxiv.2101.01170, working paper

Bleemer Z (2018) The effect of selective public research university enrollment: Evidence from California. Research & Occasional Paper Series: CSHE. 11.18. Center for Studies in Higher Education

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

Blomquist S, Newey W (2017) The bunching estimator cannot identify the taxable income elasticity. Working Paper 40/17, Cemmap

Blomquist S, Kumar A, Liang CY, Newey WK (2015) Individual heterogeneity, nonlinear budget sets, and taxable income. Tech. rep.

Burgstahler D, Dichev I (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, Rothe C, Yıldız N (2016) A discontinuity test for identification in triangular nonseparable models. Journal of Econometrics 193(1):113–122

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

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

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

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

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

Caetano C, Caetano G, Nielsen E, Sanfelice V (2023) The effect of maternal labor supply on children's skills, working paper

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

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

Cattaneo M, Jansson M, Ma X, Slemrod J (2018) Bunching designs: Estimation and inference. Working paper, UCSD Bunching Workshop

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

Chernozhukov V, Hong H (2002) Three-step censored quantile regression and extramarital affairs. Journal of the American Statistical Association 97(459):872–882

Chetty R, Friedman JN, Olsen T, Pistaferri L (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 JL, Patel E, Seegert N, Smith M (2022) How do firms respond to corporate taxes? Journal of Accounting Research 60(3):965–1006

Collier BL, Ellis C, Keys BJ (2021) The cost of consumer collateral: Evidence from bunching, working paper

De Vito A, Jacob M, Müller MA (2019) Avoiding taxes to fix the tax code, working paper

Dee TS, Dobbie W, Jacob BA, Rockoff J (2019) The causes and consequences of test score manipulation: Evidence from the new york regents examinations. American Economic Journal: Applied Economics 11(3):382–423

Denning J, Turner L, Seegert N (2023) Work in progress

Dowd C (2021) Donuts and distant cates: Derivative bounds for rd extrapolation. Available at SSRN 3641913

Dube A, Manning A, Naidu S (2018) Monopsony and employer mis-optimization explain why wages bunch at round numbers. Tech. rep., National Bureau of Economic Research

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

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

Erhardt EC (2017) Microfinance beyond self-employment: Evidence for firms in Bulgaria. Labour economics 47:75–95

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

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

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

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

Garicano L, Lelarge C, Van Reenan J (2016) Firm size distortions and the productivity distribution: Evidence from France. American Economic Review 106(11):3439–3479

Gelber AM, Jones D, Sacks DW (2013) Earnings adjustment frictions: Evidence from the social security earnings test, working Paper

Gelber AM, Jones D, Sacks DW (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 AM, Jones D, Sacks DW, Song J (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, Shen S, Zhang J (2019) A censored maximum likelihood approach to quantifying manipulation in china's air pollution data. Working paper, University of California - Davis

Ghanem D, Shen S, Zhang J (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, working Paper, arXiv 2205.10310

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

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

Hong Hy (2023) Estimating the distribution of elasticities of medical expenditures using a notch in out-of-pocket cost. Work in progress

Hungerman DM, Ottoni-Wilhelm M (2021) Impure impact giving: Theory and evidence. Journal of Political Economy 129(5):1553–1614

Hussein SM (2021) Educational time use and the cognitive development of children: Evidence from the longitudinal study of australian children. PhD thesis, The Australian National University (Australia)

Ito K (2014) Do consumers respond to marginal or average price? Evidence from nonlinear electricity pricing. American Economic Review 104(2):537–563

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

Jales H, Yu Z (2017) Identification and estimation using a density discontinuity approach. Regression Discontinuity Designs

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

Kaneko S, Noguchi H (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, Yıldız N (2019) A test of selection on observables assumption using a discontinuously distributed covariate, working paper

Kleven HJ (2016) Bunching. Annual Review of Economics 8:435–464

Kleven HJ, 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

Kopczuk W, Munroe D (2015) Mansion tax: The effect of transfer taxes on the residential real estate market. American Economic Journal: Economic Policy 7(2):214–57

Kostøl AR, Myhre AS (2021) Labor supply responses to learning the tax and benefit schedule. American Economic Review 111(11):3733–66

Kuhn PJ, Yu L (2021) Kinks as goals: Accelerating commissions and the performance of sales teams. Tech. rep., National Bureau of Economic Research

Lavetti K, Schmutte IM (2018) Estimating compensating wage differentials with endogenous job mobility, working paper

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

Marx BM (2022) Dynamic bunching estimation with panel data, working Paper

Meyer RH, Wise DA (1983) Discontinuous distributions and missing persons: The minimum wage and unemployed youth. Econometrica 51(6):1677–1698

Moore DT (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. PhD thesis, Syracuse University

Pollinger S (2021) Kinks know more: Policy evaluation beyond bunching with an application to solar subsidies. Tech. rep., TSE Working Paper

Rozenas A, Schutte S, Zhukov Y (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

Sallee JM, Slemrod J (2012) Car notches: Strategic automaker responses to fuel economy policy. Journal of Public Economics 96(11):981–999

Velayudhan T (2018) Misallocation or misreporting? evidence from a value added tax notch in india, working Paper