Better Bunching, Nicer Notching*

Marinho Bertanha†
University of Notre Dame

Andrew H. McCallum‡
Board of Governors of the Federal Reserve System

Nathan Seegert §
University of Utah

First draft: September 13, 2016
This draft: December 22, 2017

Preliminary and Incomplete: Comments Welcome.
Please do not cite or circulate.

Abstract

Bunching and notching estimation methods pioneered by Saez (2010) and Kleven and Waseem (2013a) use mass points in the distributions of income or profits that are caused by policies to estimate key parameters in public finance. The goal of this paper is to build on and enhance these estimators by combining advances in public finance, labor economics, and econometrics. The result is an updated method that retains the insights and advantages of the original bunching and notching estimators while incorporating covariates. In particular, we place our estimator within the econometrics literature by showing that bunching and notching are censoring models. Once our estimator is recast in this light, it is possible to leverage the significant econometric developments since Tobin (1958) and to place bunching and notching on a rigorous statistical foundation. We compare the updated and standard methods using Monte Carlo simulations to illustrate the relative performance of recovering parameter values. Finally, we apply our method in the context of the earned income tax credit to show it leads to quantitatively different estimates of the compensated elasticity of reported income with respect to (one minus) the marginal tax rate.

JEL: H23, H24, H26
Keywords: bunching, notching, tax kink, earned income tax credit

*The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Federal Reserve Board or the Federal Reserve System. Jessica C. Liu and Michael A. Navarrete provided excellent research assistance. All remaining errors are our own.

†mbertanha@nd.edu
‡andrew.h.mccallum@frb.gov
§nathan.seegert@gmail.com

1 Introduction

The goal of this paper is to enhance and improve ‘‘bunching’’ and ‘‘notching’’ estimators developed by Saez (2010) and Kleven and Waseem (2013a). By combining insights from public finance, labor economics, and econometrics we provide an updated method that retains the insights and advantages of the original method while also providing a statistically rigorous general model that allows for both ‘‘kinks’’ and ‘‘notches.’’ The public finance literature refers to a change in the intercept of a convex piecewise-linear constraint as a ‘‘notch’’ while calling a change in the slope a ‘‘kink.’’ Similarly, the ‘‘notch point’’ is the income level at which the intercept changes and the ‘‘kink point’’ is the income level at which the slope changes.

We begin by presenting a general model of utility maximization with a piecewise-linear constraint showing that it nests any public finance model with notches and kinks. The piecewise nature of the constraint can be a result of differential tax rates, or piecewise insurance reimbursement rates, for example. These constraints can be concave or convex and, because they are piecewise-linear, they are fully characterized by their intercepts and slopes. Well-known models fit into this category and include Burtless and Hausman (1978), Saez (2010), Kleven and Waseem (2013a), Best and Kleven (2017), Einav, Finkelstein, and Schrimpf (2017), among other.[1]

Using the general model, we derive the Saez (2010) elasticity bunching estimator. Saez (2010)'s insight is that the mass of agents reporting at the kink point is increasing the elasticity of taxable income for a given distribution of potential unobserved earning ability, also called the latent variable. Stated another way, the more agents that shift income to the kink point, the easier it must be to shift income. All current bunching and notching estimators use this insight to estimate the elasticity.

Turning to identification conditions, Lemma 3.1 proves there there is no way to identify

[1] Similar models have also been applied to electricity, water, and cellular phone markets (Ito, 2014; Huang, 2008; Reiss and White, 2005; Olmstead, Hanemann, and Stavins, 2007).

the elasticity if the latent distribution is completely unrestricted. In doing so, we clarify that many existing non-parametric estimators for ε are either implicitly restricting $\mathcal{F}_n$ or inconsistent for the true elasticity. We then discuss specific example restrictions on the latent variable distribution that give point and partial identification.

Next we present our key insight which is that we can rewrite the observable income that individuals report as a middle censoring model. Once we have linked bunching estimators to the literature on censored models, we can formalize the assumptions necessary to identify the structural elasticity and leverage the significant econometric developments made since Tobin (1958). We also add richness to the model by assuming that the latent variable is a function of some observable non-random and non-tax covariates. Extending the estimator in this way is novel to public finance and tractable in a censoring setting but much less so in a bunching setting. Our censoring perspective allows researchers to incorporate numerous kinks and notches at once, which is of practical importance in many settings including piecewise pricing.

Moving to estimation, we highlight there are many fully- and semi-parametric ways to estimate the parameters of a middle censoring model. We discuss those and derive a simple extension of a Heckit-style multi-step estimator that follows from Heckman (1976). All of the estimators we discuss are transparent, straightforward, statistically rigorous, and well-established. They are also widely implemented in statistical programs such as Stata and Matlab.

We demonstrate the practical advantages of our estimator in the context of the earned income tax credit (EITC) and through a Monte Carlo simulation. We use the Individual Public Use Tax Files constructed by the U.S. Internal Revenue Service (IRS) as a repeated annual cross-section previously used by Saez (2010). In this data, there is visual evidence of bunching at the lowest income level. To demonstrate the importance of controlling for covariates, we show graphically that the amount of bunching changes with observable characteristics. Our model allows us to control for these differences.

Estimating the elasticity of taxable income is a primary focus of public finance because it provides evidence on how agents respond to tax and transfer programs and cannot be directly computed from other estimates.[2]

Table 1 provides a detailed summary of the public finance literature that estimates the elasticity of taxable income. This table highlights two interesting patterns. First, there is a wide range of estimates for the elasticity, frequently anywhere between 0 and 1 with some estimates as low as -0.83 and as high as 3 Feldstein (1995); Goolsbee (1999). Second, the most common estimation method is an instrumental variables (IV) approach outlined by Auten and Carroll (1999) and Gruber and Saez (2002). Subsequent papers focused on income controls and base-year income splines to correct for challenges created by mean reversion and changes in income inequality (Gelber, 2014). Kopczuk (2005) and Giertz (2005) show that elasticity estimates are sensitive to these income controls. Weber (2014) shows that the income controls are ineffective at correcting the endogeneity concerns of the underlying model.[3] These later papers demonstrate the challenges associated with IV methods and spurred the development of new methods.

The bunching and notching methods pioneered by Saez (2010) and Kleven and Waseem (2013b) sought to overcome the limitations of the IV and difference-in-differences methods. Cross sectional data on income levels and piecewise linear structures in taxes and transfers are common across not only many tax jurisdictions but in other settings as well. As such, bunching and notching expanded the settings in which it was possible to estimate how agents respond to change in the slope or intercept of their budget sets. Studies using these methods have investigated poverty reducing polices like the EITC, taxes on real estate, and corporate income. They have also been used to study welfare programs, education funding, medical and social insurance, minimum wages, automobile fuel economy, and tiered pricing

[2]In one of the earlier papers estimating the elasticity of taxable income Lindsey (1987) states, "The response of taxable income to tax rates is not the same as the response of labor supply or other real economic factors. The response of taxable income includes, but is by no means limited to these factors. Existing parameters on labor supply response, for example, are not applicable to the problem at hand."

[3]Weber (2014) and Blomquist and Selin (2010) provide instrumental variables that are exogenous under testable assumptions of the degree of serial correlation.

in cellular service. Kleven (2016) provides a review of the many ways that these methods have been employed.

Despite the important insights, increased reliance, and numerous advantages of the bunching estimator, recent work has discussed limitations (see, e.g., Chetty, Friedman, Olsen, and Pistaferri, 2011; Kleven, 2016; Einav et al., 2017).[4] Chetty et al. (2011) find that micro estimates of elasticities that do not account for optimizing frictions produce biased estimates. Kleven (2016) highlights the complication of mapping bunching to structural parameters due to the existence of optimizing frictions in setting with both kinks and notches. Einav et al. (2017) find large differences between estimates from traditional bunching methods and estimates from a richer model with optimizing frictions, dynamics, and uncertainty. The stark contrast in estimates highlights the tradeoff between, ‘‘transparency, simplicity, and speed of communication’’ with ‘‘richness of the model’’ that more easily translates estimates into underlying economic objects. Since the seminal work by Chetty et al. (2011) there has been an increase in awareness of the differences in elasticity estimates by method. Chetty et al. (2011) demonstrate current bunching methods---but also most micro methods---produce biased estimates. As a test, Chetty et al. (2011) demonstrates that bunching elasticity estimates increase with the size of the difference in tax rates---something that should not occur if it consistently estimates the structural elasticity.

2 Optimization subject to piecewise-linear constraints

Firms' and individuals' optimization problems often face piecewise-linear constraints. The piecewise nature of the constraint can be a result of differential tax rates, piecewise insurance reimbursement rates, or contract bonuses, for example. These constraints can be concave or convex and, because they are piecewise-linear, they are fully characterized by their intercepts and slopes.

[4]There have also been a group of recent working papers on the practical complications of using bunching to estimate elasticities for example, Blomquist, Kumar, Liang, and Newey (2015); Patel, Seegert, and Smith (2016); Dekker, Strohmaier, and Bosch (2016).

The public finance literature refers to a change in the intercept of a convex piecewise-linear constraint as a "notch" while calling a change in the slope a "kink." Similarly, the "notch point" is the income level at which the intercept changes and the "kink point" is the income level at which the slope changes.

The following section presents a general model of utility maximization with a piecewise-linear constraint showing that it nests any public finance model with notches and kinks. Well-known models fit into this category and include Burtless and Hausman (1978), Saez (2010), Kleven and Waseem (2013a), Best and Kleven (2017), Einav et al. (2017), among others.[5] This general setup also allows us to easily consider parameterizations of the utility function that differ from the often used quasilinear and isoelastic utility. To illustrate this point, we present an example with constant elasticity of substitution (CES) utility subject to a piecewise-linear constraint with changes to the intercept in Appendix A.

2.1 General Model

Consider a population of agents that are heterogeneous with respect to N^* which is unobserved by the econometrician but known to the agent. This heterogeneity can be interpreted as ability, health, or differential preferences, for example. Each agent maximizes utility subject to a piecewise-linear budget constraint by jointly choosing a composite consumption good, C_i, with a price, P_i, selling, L_i at price W_i to increase resources, Y_i, that can be used to fund consumption. L_i is commonly interpreted as labor supply but could also be medical spending as in Einav et al. (2017), for example. We assume the agent takes the price of labor, W_i, as given.

[5] Similar models have also been applied to electricity, water, and cellular phone markets (Ito, 2014; Huang, 2008; Reiss and White, 2005; Olmstead et al., 2007).

Formally, the agent solves

$$\begin{aligned} \max_{C_i, L_i} & \quad U\left(C_i, L_i; N_i^*\right) \\ s.t. \quad Y_i &= W_i L_i \\ C_i &= S_0 Y_i + \sum_{j}^{J} \left[\Delta I_j + \Delta S_j \left(Y_j - D_j\right)\right] \mathbb{1}\left(Y_i > D_j\right) + R_i \end{aligned} \tag{1}$$

The intercept of the budget constraint can change by ΔI_j at D_j and the slope can also change by $\Delta S_j \equiv S_j - S_{j-1}$ at D_j. The changes enumerated by $j = 1, \ldots, J$ encompass all possible line segments. For example, at income level D_j, if the intercept does not change but the slope does then, $\Delta I_j = 0$ and $\Delta S_j \neq 0$.

The piecewise-linear budget constraint results in an optimal choice for L_i which is a piecewise function. It is typically more difficult to observe the optimal choice for L_i than to observe $Y_i = W_i L_i$ and so we use the first constraint of equation (1) to write optimal L_i as

$$Y_i = \begin{cases} D_j & N_i^* \in \left[\underline{N}_j, \overline{N}_j\right] \quad j = 1, \ldots, J \\ W_i L_i \left(S_{j-1}, N_i^*, \Delta I_j, R_i\right) & N_i^* \in \left(\overline{N}_{j-1}, \underline{N}_j\right) \quad j = 1, \ldots, J+1, \end{cases} \tag{2}$$

in which $\overline{N}_0 \equiv N_{min}$ and $\underline{N}_{J+1} \equiv N_{max}$. The optimal choice takes the form of a decision rule given by equation (2), which depends on the unobserved variable N_i^* and the thresholds $\underline{N}_j$ and $\overline{N}_j$. Agents with unobserved variable N_i^* outside of these thresholds report $W_i L_i \left(S_{j-1}, N_i^*, \Delta I_j, R_i\right)$ while agents with N_i^* inside these bounds set L_i such that optimal Y_i is equal to the discontinuity point D_j.

2.2 Examples from applied microeconomics

Equations (1) and (2) may appear unfamiliar so in the following section we show that these expressions nest well-known models from the labor and public finance literatures.

2.2.1 Quasi-linear and isoelastic Utility with only Slope Changes

The public finance literature often uses utility that is quasilinear in C_i and isoelastic with elasticity ε in L_i combined with a budget constraint that only has slope, and not intercept, changes. The general model in (1) and (2) becomes particularly tractable in this setting. In fact, we show in Appendix A that in this case, both the choice of Y_i and the thresholds $\underline{N}_j$ and $\overline{N}_j$ in equation (2) are log-linear functions of the elasticity ε, the slope changes in logs s_j, and the latent variable N_i for any number of J kink points.

To make the model in equations (1) and (2) abundantly clear, we relate them to the seminal work of Saez (2010) by considering the agent's problem written as

$$\begin{aligned} \max_{C_i, L_i} \quad & C_i - (N_i^*)^{-1/\varepsilon} \frac{L_i^{1+\frac{1}{\varepsilon}}}{1+1/\varepsilon} \\ s.t. \quad Y_i &= L_i \\ C_i &= (1-t_0)\, Y_i + (t_0 - t_1)\,(Y_j - D_j)\, \mathbb{1}\,(Y_i > D_j)\,. \end{aligned} \tag{3}$$

Our model nests Saez (2010) when $P_i = W_i = 1$, $Y_i = z$ is taxable income, the slopes are the net-of-tax rates, $S_j = 1 - t_j$, and heterogeneity $N_i^* = n$ is interpreted as a distribution of potential earnings ability.

As shown in the general model, the optimal labor supply choice resulting in income, Y_i, reported by agent i is a piecewise function and depends on the value of the unobserved heterogeneity variable N_i^* and is given by

$$y_i = \begin{cases} n_i^* + \varepsilon s_0 & n_i^* < \underline{n}_1 \\ d_1 & n_i^* \in [\underline{n}_1, \overline{n}_1] \\ n_i^* + \varepsilon s_1 & n_i^* > \overline{n}_1 \end{cases} \tag{4}$$

in which we use lower case letters to denote logs, for example, $s_j \equiv \ln S_j \equiv \ln(1 - t_j)$, $y_i \equiv \ln(Y_i)$, and $d_1 \equiv \ln(D_1)$. Because there is only one change in the slope $S_1 = S_0 + \Delta S_1$.

As discussed previously and shown in Appendix A, all models with isoelasticity utility and only slope changes in the budget constraint result in optimal reporting of income that is piecewise log-linear. In that setting, we know that $\underline{n}_1 = d_1 - \varepsilon s_0$ and $\overline{n}_1 = d_1 - \varepsilon s_1$.

The thresholds, $\underline{n}$ and $\overline{n}$ define the lowest and highest abilities that bunch at the discontinuity. Consider a geometrical interpretation of this equation. When the budget constraint has only slope changes, the thresholds are defined by indifference curves that are tangent to the budget constraint at each discontinuity D_j. For example, the indifference curve of ability $\underline{n}_j$ is tangent to the budget constraint at D_j using the slope below D_j and the indifference curve for ability $\overline{n}_j$ is tangent using the slope above the kink D_j.

There are several other examples of quasilinear and isoelastic utility with one change in the budget constraint's slope similar to the model of (3) with slightly different notation. Burtless and Hausman (1978) consider labor supply of hours worked, $Y_i = h_i$, with the slopes being the net-of-tax wages $S_j = (1 - t_j)\, w$, and heterogeneity $N_i^* = k_i$. Recently, Einav et al. (2017) uses this model with demand over total drug spending, $Y_i = m$, with slope equal to the net-of-insurance costs given by $S = (2 - c)$ and heterogeneity in an individual's health needs, $N_i^* = \xi$.

3 Inferences of parameters

Next we derive the bunching estimator to recover parameters of the utility function using observations of individuals' optimizing behavior characterized by equation (2). Assume that the latent heterogeneity variable N_i^* is distributed according to some probability density function (PDF) given by $g_N(N_i^*)$ with cumulative distribution function (CDF) $G_N(N_i^*)$. Based on equation (2), all agents with the latent variable $N_i^* \in [\underline{N}_j, \overline{N}_j]$ report $Y = D_j$ and the mass of agents at D_j is

$$B_j = P\left(Y_i = D_j\right) = P\left(\underline{N}_j \leq N_i^* \leq \overline{N}_j\right) = \int_{\underline{N}_j}^{\overline{N}_j} g_N\left(N^*\right) dN^* = G_N\left(\overline{N}_j\right) - G_N\left(\underline{N}_j\right). \quad (5)$$

Using the specific example from equation (4), the mass at D_1 given by (5), becomes

$$B^{Saez} = F_n\left(\overline{n}_1\right) - F_n\left(\underline{n}_1\right) = F_n\left(d_1 - \varepsilon \ln\left(1 - t_1\right)\right) - F_n\left(d_1 - \varepsilon \ln\left(1 - t_0\right)\right), \qquad (6)$$

in which we have used a change of variables to move from the distribution of the level of the latent variable $G_N\left(N^*\right)$ to the distribution of its logarithm $n^* = \ln\left(N^*\right)$ with CDF $F_n\left(n^*\right)$ and PDF $f_n\left(n^*\right)$. In appendix B, we prove that equation (6) is exactly (Saez, 2010, equation (4), p. 186) which he used to identify the elasticity.

3.1 Non-parametric identification of the elasticity

Saez (2010)'s insight is that the mass of agents bunching in equation (6), B^{Saez}, is increasing in the elasticity ε for a given distribution of the latent variable F_n. Stated another way, the more agents that shift income to the kink-point d_1, the easier it must be to shift income. All current bunching and notching estimators use this insight to estimate the elasticity.

An alternative interpretation of equation (6) is also possible, however. The mass of agents bunching, B^{Saez}, is increasing as F_n becomes more concentrated between $\underline{n}_1$ and $\overline{n}_1$ for a given elasticity ε. In other words, if more individuals happen to be distributed in $[\underline{n}_1, \overline{n}_1]$ because of the shape of F_n, more of them will move to the kink point for any given elasticity. Patel et al. (2016) demonstrate this empirically by showing that bunching estimates of ε are indeed sensitive to assumptions about the latent distribution.

The following section demonstrates there is no way to simultaneously estimate both the elasticity and distribution of the latent variable together using only equation (5) if the latent distribution is completely unrestricted. Intuitively, identification using only one equation to solve for two unknowns, ε and F_n, is impossible.

3.1.1 The elasticity is unidentified without latent variable restrictions

The data and model comprise five objects, (1) the distribution of earnings, F_y, (2) the kink point, d_1, (3) the slopes of the piecewise linear constraint, $s = (s_0, s_1)$, (4) the CDF of the latent variable, F_n, and (5) the elasticity, ε.

The agents observe the latter four objects and report optimal income according to equation (4) taking them as given. The resulting CDF of optimal incomes across agents, F_y, can be constructed from equation (4) and is a mapping, $F_y = T(F_n, \varepsilon, d_1, s_0, s_1)$, from the objects the agents take as given to the observable income distribution. Inspecting equation (4), the distribution of y has a mass point at d_1 but it is otherwise continuous with PDF f_y.

The researcher observes the first three of these five objects but does not observe the last two, F_n and ε. The problem of identification consists of inverting the mapping T such that the unobserved ε is a function that only depends on the three observed objects (F_y, d_1, s) regardless of what F_n may be. We denote the class of admissible distributions of n as $\mathcal{F}_n$. If the class $\mathcal{F}_n$ contains all possible continuous distributions of n, then identification of ε is impossible.

Lemma 3.1. *Let $\mathcal{F}_n$ be the class of all CDFs F_n that have continuous PDFs f_n with support $(-\infty, \infty)$. Let $\mathcal{F}_y$ be the class of all CDFs F_y that are mixed continuous-discrete with one mass point at d_1 and continuous PDFs f_y otherwise. Assume that $\varepsilon \in (0, \infty)$, $-\infty < s_1 < s_0 \leq 0$, and fix arbitrary values of F_y, d_1, s_0, and s_1. Then, there does **not** exist a unique ε such that*

$$F_y = T(F_n, \varepsilon, d_1, s_0, s_1) \quad \forall F_n \in \mathcal{F}_n.$$

Therefore, it is impossible to retrieve a unique ε just using (F_y, d_1, s_0, s_1).

Proof. See Appendix C.1 as illustrated in Figure 1. □

Figure 1 provides intuition for the proof of Lemma 3.1. It illustrates that the observable

PDF f_y in Figure 1a can be generated by applying (4) to two different combinations of latent variables distributions and elasticities, $f_{n,\varepsilon}$ and $f_{n,\varepsilon'}$ in Figures 1b and 1c, respectively. We developed Lemma 3.1 independently of and a month before we were aware of Blomquist and Newey (2017) which also shows that, as the title of their paper suggests, "The Bunching Estimator Cannot Identify the Taxable Income Elasticity."

3.2 Examples of identifying restrictions

Lemma 3.1 clarifies that many existing non-parametric estimators for ε are either implicitly restricting $\mathcal{F}_n$ or inconsistent for the true elasticity. A direct consequence of Lemma 3.1 is that it is impossible to evaluate the quality of different restrictions on $\mathcal{F}_n$ and how close they are to the truth. Below we consider a few examples of identifying restrictions.

Example 1. *Restrict F_n such that f_n is equal to a specific function inside the interval $[\underline{n}_1, \overline{n}_1]$. This is the approach used by Saez (2010). He assumes f_N is a linear function within $[D_1/(1-t_0)^{\varepsilon}, D_1/(1-t_1)^{\varepsilon}]$. That assumption is equivalent to an exponential shape on f_n, the PDF of $n = \ln(N)$, within $[\underline{n}_1, \overline{n}_1]$. Under this linearity restriction, the trapezoidal approximation utilized by Saez (2010) holds exactly and the elasticity ε is identified as the solution of his Equation (5).*

One may argue that the linear assumption is a good approximation to any potentially non-linear true density f_N in the neighborhood of the kink point if the interval $[D_1/(1-t_0)^{\varepsilon}, D_1/(1-t_1)^{\varepsilon}]$ is small. The problem with this argument is that the size of the interval is itself a function of the elasticity. It is impossible to state that the interval is small, and the linear approximation is a good one, without a priori knowledge of the elasticity.

There are more general restrictions one could impose on f_n. For example, one could say n follows a distribution inside a parametric family of distributions.

Example 2. *Restrict F_n to be a parametric class of distributions. This is the approach used by Chetty et al. (2011). In general, let $\mathcal{F}_n = \{F_n = F_\theta\ ,\ \theta \in \Theta\}$ where F_θ are CDFs* *indexed by a parameter θ, $\Theta \subseteq \mathbb{R}^p$ is the parameter space, and p is a positive integer. Jointly solving (7), (8), (9) for ε as a function of d_1, s_0, and s_1 identifies the elasticity.*

$$\mathbb{P}[y = d_1] = F_n\left(d_1 - \varepsilon s_1\right) - F_n\left(d_1 - \varepsilon s_0\right) \tag{7}$$

$$F_y(u) = F_n(u - \varepsilon s_0) \quad \text{for } u < d_1 \tag{8}$$

$$F_y(u) = F_n(u - \varepsilon s_1) \quad \text{for } u > d_1. \tag{9}$$

For example, if $\mathcal{F}_n$ is assumed to be the family of normal distributions with unknown mean and variance, the elasticity is identified. Section C.3 in the Appendix illustrates how to verify Equations (7), (8), (9) for the normal case.

Chetty et al. (2011) assume a flexible polynomial functional form for f_y, and that this same parametric functional form also holds for f_n up to a shift in location as seen in Equations (8) and (9). It is important to emphasize this is ***not*** *a non-parametric identification strategy for ε. There is an implicit parametric assumption made on f_n within $[\underline{n}_1, \overline{n}_1]$. Although f_y is non-parametrically identified, Lemma 3.1 makes clear that f_n is not non-parametrically identified within $[\underline{n}_1, \overline{n}_1]$.*

Thus far we have considered parametric restrictions on the class of distributions $\mathcal{F}_n$ that yield point identification of ε. It is possible to impose even weaker restrictions on $\mathcal{F}_n$ and obtain partial identification of ε. Restricting the class $\mathcal{F}_n$ to satisfy some shape restrictions yields partial identification. One such restriction is to assume the PDF f_n is continuous and has slope magnitude bounded by an a priori selected value $M \in (0, \infty)$. Equations (8) and (9) reveal that f_y is equal to f_n up to a shift in location. The researcher may observe f_y to get an idea of the value of M.

The following theorem gives the partially identified set for ε as a function of identified quantities and the maximum slope magnitude M.

Theorem 3. *Assume $\mathcal{F}_n$ contains all distributions with continuous PDF f_n such that the* *maximum slope magnitude of* f_n *is* $M \in (0, \infty)$*. Then, the elasticity* $\varepsilon \in \Upsilon$ *where*

$$\Upsilon = \begin{cases} \emptyset & , \text{ if } \mathbb{P}[y = d_1] < \frac{\left|f(d_1^+) - f(d_1^-)\right| \left[f(d_1^+) + f(d_1^-)\right]}{2M} \\ [\underline{\varepsilon}, \overline{\varepsilon}] & , \text{ if } \frac{\left|f(d_1^+) - f(d_1^-)\right| \left[f(d_1^+) + f(d_1^-)\right]}{2M} \leq \mathbb{P}[y = d_1] < \frac{f(d_1^+)^2 + f(d_1^-)^2}{2M} \\ [\underline{\varepsilon}, \infty) & , \text{ if } \frac{f(d_1^+)^2 + f(d_1^-)^2}{2M} \leq \mathbb{P}[y = d_1] \end{cases}$$

where $\emptyset$ *is the empty set, and*

$$\underline{\varepsilon} = \frac{2\left[f(d_1^+)^2/2 + f(d_1^-)^2/2 + M\ \mathbb{P}[y = d_1]\right]^{1/2} - \left(f(d_1^+) + f(d_1^-)\right)}{M(s_0 - s_1)}$$

$$\overline{\varepsilon} = \frac{-2\left[f(d_1^+)^2/2 + f(d_1^-)^2/2 - M\ \mathbb{P}[y = d_1]\right]^{1/2} + \left(f(d_1^+) + f(d_1^-)\right)}{M(s_0 - s_1)}$$

For the proof see Section C.4 in the Appendix.

In practice, a budget set may display several points where bunching occurs. On the one hand, the income distribution may be very different across different bunching points. On the other hand, the elasticity ε is assumed to be the same for all individuals. Variation in probability of bunching, tax rates and PDF values narrow down the partially indentified set. Variation could arise from one or multiple time periods.

Corollary 3.1. *Suppose the researcher observes a budget set with* K *kinks* $\mathbb{P}[y = d_j]$*,* $f_y(d_j^+)$*,* $f_y(d_j^-)$ *for* $j = 1, \ldots, K$ *along with tax prices* $s_0, s_1, \ldots, s_K$*. Assume the conditions of Theorem 3. Then, the elasticity* $\varepsilon \in \bigcap_{j=1}^{K} \Upsilon_j$ *where* Υ_j *is the partially identified set of Theorem 3 for kink* j*.*

Corollary 3.1 shows that variation arising from multiple kinks in one budget set or from budget sets at different time periods helps identify the elasticity ε without a parametric assumption on $\mathcal{F}_n$. Blomquist and Newey (2002) assume a population variation of budget sets to identify elasticities non-parametrically. Their result rely on an identification condition that could be hard to verify in practice. Corollary 3.1 constitutes a straightforward approach for practitioners because it produces the narrowest partially identified set given whatever variation the researcher has at hand. This variation could be little or none, or could be close to satisfying the identification condition of Blomquist and Newey (2002).

3.3 Identification with Covariates

Up to this point, we have only considered the general distribution of n^*. Here we show that if that distribution depends on other covariates, the effect of those covariates must be estimated simultaneously along with the estimate of ε. Consideration of covariates is related to another crucial assumption of Saez (2010) which is that the unobserved variable N_i^* is identically distributed across all agents. To see how this is related to identifying the elasticity, consider the possibility that the heterogeneous factor is a function of some observable covariates x_i and z_i in addition to a completely random characteristics with CDF $F_\nu(\nu_i)$ given by $N_i^* = N_i^*(x_i, z_i, \nu_i; \theta)$. A simple example of this function would be $N_i^* = \exp(x_i'\beta + z_i'\gamma + \nu_i)$. Ensure that x_i includes a constant term and denote the standardized distribution of ν_i as $\Phi_\nu(\nu_i)$. In this setting, the mass at the kink varies by *individual* and the analog to (6) is given by

$$B_i^{Saez} = \Phi_\nu\left(\frac{z^* - \varepsilon\ln(1-t_1) - x_i'\beta - z_i'\gamma}{\sigma_\nu}\right) - \Phi_n\left(\frac{z^* - \varepsilon\ln(1-t_0) - x_i'\beta - z_i'\gamma}{\sigma_\nu}\right). \quad (10)$$

We can state this point in another way. The first fundamental insight of Saez (2010) is that a larger elasticity results in a larger mass at the kink. Here we have an example which shows that a larger β can also result in a larger mass at the kink when $\beta > 0$ and $x_i > 0$. Conversely, a more negative conditional mean results in a smaller mass at the kink.

The probability that an individual reports income at the kink depends systematically on determinants of their potential earning ability and differs for each individual. The number of bunching masses at the kink conditional on covariates is now equal to the size of the sample of individuals and there is no such thing as ‘‘the’’ mass at the kink. Much of labor economics has taught us that there are many plausible observable covariates that effect potential earnings. These must be controlled for in any specification that seeks to estimate the response of all individuals to a change in marginal tax rates. Examples of plausible determinants of income are years of work history, years of education, U.S. state of residence, gender, and number of children, among others. In the following data section, we provide evidence that covariates do in fact lead to systematic differences in earnings potential and therefore probabilities of reporting income at the kink that differ systematically across individuals.

4 Parametric and semi-parametric estimation

The econometrics literature promises a way to estimate the agents' preferences parameters that overcomes the limitations of the bunching methods discussed previously. In particular, the data generating process described in equation (2) is a type of censoring model. Censoring models are typically discussed as having a mass point at the upper or lower end of the distribution, leading to an upper or lower censored model, respectively. The model proposed in equation (2) has instead a mass point that lies in the middle of the distribution. Recasting the model as middle censoring allows us to leverage the significant econometric developments since Tobin (1958) and to place both bunching and notching estimators on a rigorous statistical foundation.

As we did before, assume that the latent heterogeneity variable N_i^* is distributed according to some probability density given by $g_N(N_i^*)$ with cumulative distribution function $G_N(N_i^*)$. We can derive the likelihood for the model in equation (2) which is

$$\begin{aligned} L(Y,W,S,R \mid \theta) &= \prod_{i=1}^{N}\prod_{j=1}^{J} g_N\left(L_i^{-1}\left(\frac{Y_i}{W_i}, S_{j-1}, R_i\right)\right)^{1\left(W_i L_i\left(S_{j-1},\underline{N}_j,R_i\right)<Y_i<W_i L_i\left(S_{j-1},\overline{N}_{j-1},R_i\right)\right)} \\ &\times \left[G_N\left(\overline{N}_j\right) - G_N\left(\underline{N}_j\right)\right]^{1(Y_i=D_j)} \end{aligned} \tag{11}$$

in which we inverted $L_i^{-1}\left(\frac{Y_i}{W_i}, S_{j-1}, R_i\right) = N_i^*$.

Equation (11) is difficult to understand intuitively so we also present the likelihood function corresponding to the simpler model of equation (4). We also add the additional assumption that the heterogeneous factor is a function of some observable covariates x_i and z_i in addition to a completely random characteristics with CDF $F_\nu(\nu_i)$ given by $N_i^* = \exp\left(x_i'\beta + z_i'\gamma + \nu_i\right)$. Ensure that x_i includes a constant term and denote the standardized CDF of ν_i as $\Phi_\nu(\nu_i)$ with PDF ϕ_n

$$\begin{aligned} L\left(y,x,z \mid \varepsilon,\beta,\gamma,\sigma_\nu\right) &= \prod_{i=1}^{N}\phi_\nu\left(\frac{y_i-\varepsilon\ln\left(1-t_0\right)-x_i'\beta-z_i'\gamma}{\sigma_\nu}\right)^{1(y_i<d_1)} \\ &\times \left[\Phi_\nu\left(\frac{d_1-\varepsilon\ln\left(1-t_1\right)-x_i'\beta-z_i'\gamma}{\sigma_\nu}\right)-\Phi_\nu\left(\frac{d_1-\varepsilon\ln\left(1-t_0\right)-x_i'\beta-z_i'\gamma}{\sigma_\nu}\right)\right]^{1(y_i=d_1)} \\ &\times \phi_\nu\left(\frac{y_i-\varepsilon\ln\left(1-t_1\right)-x_i'\beta-z_i'\gamma}{\sigma_\nu}\right)^{1(y_i>d_1)} \end{aligned} \quad (12)$$

4.1 Parametric estimation

If we assume that ϕ_ν in (12) is a normal distribution, this becomes a middle censored Tobit type 1 model after Tobin (1958). Extending to multiple kink points with a normally distributed latent variable would lead to a mixed ordered probit model with level equations in each ordered region of the domain. Adding a second normal error term to the level of income being reported makes this a Tobit type 2 developed by Gronau (1973). The model could also be extended to the Tobit type 3 Heckman (1974), Tobit type 4 Nelson and Olson (1978), or Tobit type 5 Heckman (1978) models. Surveys of censoring models and their applications are provided by Maddala (1983), Amemiya (1984), Dhrymes (1986) Long (1997), DeMaris (2005), and Greene (2005)

Because parametric censoring models have been studied for many decades, estimation of the parameters of the model could proceed using any number of different techniques. Among these methods are Maximum Likelihood Estimation (MLE) via gradient ascent or the expectation maximization algorithm (Ruud, 1991). The parameters could also be estimated using Bayesian methods or the Method of Simulated Moments (MSM). In more restrictive cases, by assuming ϕ_ν is normal, for example, it is possible to develop a Heckit-style multi-step estimators similar to Heckman (1976).

A Heckit-style estimator based on the model in (12) is particularly appealing and so we derive those steps explicitly in Appendix D. Intuitively, the first step relies on estimating a probit model for observations above and below the kink. Constructing the inverse mills ratio from that first step and including it as covariate in the second step allows consistent estimation of the parameters determining the level of reported income. Finally, the elasticity ε can be recovered from a linear combination of the parameters estimated in the second step.

Other, context specific, parametric structural models also provide potential ways to consistently estimate the elasticity of taxable income with examples from Einav, Finkelstein, and Schrimpf (2015) and Einav et al. (2017). The advantages of structural models are explored in recent papers Chetty (2009); Angrist and Pischke (2010); Deaton (2010); Heckman (2010); Imbens (2010) and Keane (2010). For a longer discussion of structural models in public finance see the excellent summary by Thoresen and Vattø (2015).

Structural models rely on strong assumptions, particularly about the distributions of unobservables. We have shown that any method that point identifies the elasticity will necessarily make some parametric restrictions. This should attenuate the precived cost of using structural models. It should also emphasizes the importance of tests of external validity. To this end, several papers have made progress combining structural labor supply models and experimental evidence (Brewer, Duncan, Shephard, and Suarez, 2006; Pronzato, 2012; Geyer, Haan, and Wrohlich, 2015). Since being developed at least in the early 1980s there also tests of the parametric assumptions made for the error term in censoring models (Nelson, 1981).

4.2 Semi-parametric estimation

If the parametric assumptions are true, the most efficient way of estimating the elasticity is with MLE.[6] Trading fewer restrictions for less potential efficiency in a semi-parametric setting, could still be desireable.

[6]There is some evidence that prior concerns about the assumptions inherent in MLE in a censoring context are less severe than previously thought (Evers, De Mooij, and Van Vuuren, 2008).

Censoring models that relax strict assumptions on the distributions of the error terms necessary for MLE and related methods have been developed. Most of those also focus on upper and lower, instead of middle, censoring but adapting to middle censoring is straightforward. Another reason to commend their use is that standard statistical programs include code that implements them. (Pagan and Ullah, 1999, ch. 9) provide a nice outline of methods that were developed up to the late 1990s and there has been significant progress on semi-parametric extensions since that time.

Extending the Heckit-style estimator to a semi-parametric setting by following the control function estimation approaches is also possible. Instead of estimating a parametric binary outcome model in the first step, that step would estimate a semi-parametric binary model. That could be accomplished using a kernel or series estimator following insights by Powell (1987), Newey (1988), and Ahn and Powell (1993). The second step would include the semi-parametric equivalent of the inverse mills ratio as a control. These methods are more flexible than but are limited to estimating the first moment of the conditional income distribution. While that limitation does not affect the estimation of an elasticity, it does limit potential counterfactual analysis that rely on higher moments.

5 Application to EITC

5.1 Data

In order to emphasize the practical implications of our perspective, we employ the data originally used by Saez (2010) and replicate his results using code from the AER website. Those data are from the Individual Public Use Tax Files constructed by the IRS. The annual cross section for each year 1995 to 2004 includes sampling weights which allow interpretation of any estimates as being based on the population of U.S. income tax returns.

5.2 Graphical Evidence on the Importance of Covariates

Figure 2 graphs the income distribution for individuals with one child (Panel A) and two or more children (Panel B) and replicates Figure 3 from Saez (2010). These panels demonstrate clear bunching at the first kink point denoted by a red vertical line. The elasticity of taxable income with respect to the tax schedule is likely not zero given the mass in the distribution at the kink point.

Figures 3a through 6b graph the income distributions for individuals with one child in several subsamples to demonstrate that non-random and non-tax determinants change the likelihood an individual is observed at the lowest kink point which is $8580 in year 2008 dollars.

Figures 3a and 3b graph the income distribution for U.S. states with the lowest and highest average incomes which are Montana and Colorado, respectively. A mass point at the lowest kink point is observable for Montana but not for Colorado. This difference suggests that the state of residence is an important determinant for the probability that an individual is in the region $[\underline{n}_1, \overline{n}_1]$.

Figures 4a and 4b graph the income distributions of individuals that do and do not itemize their deductions. The income distributions of these two groups differs greatly. Compared to those that itemize deductions, a much higher proportion of individuals that do not itemize deductions have income near the first kink point. These figures stand in contrast to what theory might suggest because itemizers should have more opportunities to shift income, have a higher elasticity, and be more likely to be observed at the kink point.

Figures 5a and 5b graph the income distributions for individuals that are and are not self-employed. Saez (2010) made the intuitively appealing point that the mass at the kink point, and therefore the elasticity, for the self-employed is much larger than the mass for those who are not self-employed.

Figures 6a and 6b, however, allow an alternative explanation. Those figures show the income distributions for individuals that are self-employed and itemize compared to individuals who are self-employed and do not itemize. Theory would suggest that self-employed--itemizers have ample opportunities to shift income, yet there is little mass at the lowest kink point in the data. On the other hand, self-employed--non-itemizers pile up at the lowest kink point.

Together these figures demonstrate that covariates like state dummies, itemizing deductions, and self-employment status affect the likelihood that an individual has unobserved ability in the region that induces them to report income at the kink point. Saez (2010), and many others, try to control for this by estimating the elasticity separately for different subsamples. Using subsamples is a tried and true way to control for observable differences in econometrics. There are practical complications of using subsamples, however. These complications include estimating many elasticities (one for each state, for example), choosing how to define subsamples--including interaction terms, and the reduction in statistical precision caused by using smaller datasets. Finally, relying on subsamples makes an estimated elasticity less valuable for informing policy because the ideal estimator might seek to capture the average response of a hypothetical taxpayer to a change in the tax rate.

6 Conclusion

This paper has formalized bunching within the econometrics literature. Doing so has shown that elasticities cannot be identified using data locally around notch or kink points in budget constraints and without parametric restrictions. The insights from Saez (2010) and Kleven and Waseem (2013a), however, can be extended to a censoring setting. Extending bunching to these contexts provides several advanced methods to consistently estimate elasticities. In particular, we demonstrate the importance of including covariates.

Additional research is needed to estimate the elasticity of taxable income using maximum likelihood and control function methods. These estimates will hopefully inform the literature on the appropriate parametric restrictions.

Missing in this discussion is the role of frictions, which observationally are an important feature. Chetty et al. (2011) advocate estimating a structural model of labor supply with frictions to identify the structural elasticity relevant for policy. Advances in the labor literature on discrete choice models offer potential solutions (Van Soest, 1995; Hoynes, 1996; Creedy and Kalb, 2005; Thoresen and Vattø, 2015).

References

Aarbu, K. O. and T. O. Thoresen (2001). Income responses to tax changes—evidence from the norwegian tax reform. *National Tax Journal*, 319--335.

Ahn, H. and J. L. Powell (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics 58*(1-2), 3--29.

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

Angrist, J. D. and J.-S. Pischke (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *The Journal of Economic Perspectives 24*(2), 3--30.

Auten, G. and R. Carroll (1999). The effect of income taxes on household income. *Review of economics and statistics 81*(4), 681--693.

Auten, G., R. Carroll, and G. Gee (2008). The 2001 and 2003 tax rate reductions: An overview and estimate of the taxable income response. *National Tax Journal*, 345--364.

Best, M. C. and H. J. Kleven (2017). Housing market responses to transaction taxes: Evidence from notches and stimulus in the uk. *The Review of Economic Studies*, rdx032.

Blomquist, S., A. Kumar, C.-Y. Liang, and W. K. Newey (2015, 4). Individual Heterogeneity, Nonlinear Budget Sets, and Taxable Income. CESifo Working Paper Series 5320, CESifo Group Munich.

Blomquist, S. and W. Newey (2002). Nonparametric estimation with nonlinear budget sets. *Econometrica 70*(6), 2455--2480.

Blomquist, S. and W. Newey (2017, September). The bunching estimator cannot identify the taxable income elasticity. Technical report, MIT.

Blomquist, S. and H. Selin (2010). Hourly wage rate and taxable labor income responsiveness to changes in marginal tax rates. *Journal of Public Economics 94*(11), 878--889.

Brewer, M., A. Duncan, A. Shephard, and M. J. Suarez (2006). Did working families' tax credit work? the impact of in-work support on labour supply in great britain. *Labour economics 13*(6), 699--720.

Burtless, G. and J. A. Hausman (1978). The effect of taxation on labor supply: Evaluating the gary negative income tax experiment. *Journal of political Economy 86*(6), 1103--1130.

Chetty, R. (2009). Is the taxable income elasticity sufficient to calculate deadweight loss? the implications of evasion and avoidance. *American Economic Journal: Economic Policy 1*(2), 31--52.

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

Creedy, J. and G. Kalb (2005). Discrete hours labour supply modelling: specification, estimation and simulation. *Journal of Economic Surveys 19*(5), 697--734.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature 48*(2), 424--455.

Dekker, V., K. Strohmaier, and N. Bosch (2016, 5). A data-driven procedure to determine the bunching window: An application to the Netherlands. Hohenheim Discussion Papers in Business, Economics and Social Sciences 05-2016, University of Hohenheim, Faculty of Business, Economics and Social Sciences.

DeMaris, A. (2005). *Truncated and Censored Regression Models*, pp. 314--347. John Wiley & Sons, Inc.

Dhrymes, P. J. (1986). Limited dependent variables. In Z. Griliches† and M. D. Intriligator (Eds.), *Handbook of Econometrics* (1 ed.), Volume 3, Chapter 27, pp. 1567--1631. Elsevier.

Doerrenberg, P., A. Peichl, and S. Siegloch (2015). The elasticity of taxable income in the presence of deduction possibilities. *Journal of Public Economics*.

Einav, L., A. Finkelstein, and P. Schrimpf (2015). The response of drug expenditure to nonlinear contract design: evidence from medicare part d. *The Quarterly Journal of Economics 130*(2), 841--899.

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

Evers, M., R. De Mooij, and D. Van Vuuren (2008). The wage elasticity of labour supply: a synthesis of empirical estimates. *De Economist 156*(1), 25--43.

Feldstein, M. (1995). The effect of marginal tax rates on taxable income: a panel study of the 1986 tax reform act. *Journal of Political Economy 103*(3), 551--572.

Gelber, A. M. (2014). Taxation and the earnings of husbands and wives: evidence from sweden. *Review of Economics and Statistics 96*(2), 287--305.

Gelber, A. M., D. Jones, and D. W. Sacks (2013). Earnings adjustment frictions: Evidence from the social security earnings test. *National Bureau of Economic Research No. w19491*.

Geyer, J., P. Haan, and K. Wrohlich (2015). The effects of family policy on maternal labor supply: Combining evidence from a structural model and a quasi-experimental approach. *Labour Economics 36*, 84--98.

Giertz, S. (2005). Sensitivity analysis of the elasticity of taxable income. *CBO Working Paper 2005-01, Congressional Budget office, Washington D.C.*.

Giertz, S. H. (2007). The elasticity of taxable income over the 1980s and 1990s. *National Tax Journal*, 743--768.

Goolsbee, A. (1999). Evidence on the high-income laffer curve from six decades of tax reform. *Brookings Papers on Economic Activity 1999*(2), 1--64.

Greene, W. H. (2005). Censored data and truncated distributions. In T. Mills and K. Patterson (Eds.), *The Handbook of Econometrics*, Volume Vol. 1 Theoretical Econometrics of *5*, Chapter 20, pp. 695--736. London: Palgrave. An optional note.

Gronau, R. (1973, Part II,). The Effect of Children on the Housewife's Value of Time. *Journal of Political Economy 81*(2), 168--199.

Gruber, J. and E. Saez (2002). The elasticity of taxable income: evidence and implications. *Journal of Public Economics 84*(1), 1--32.

Hansson, Å. (2007). Taxpayers' responsiveness to tax rate changes and implications for the cost of taxation in sweden. *International Tax and Public Finance 14*(5), 563.

Heckman, J. (1974). Shadow prices, market wages, and labor supply. *Econometrica 42*(4), 679--694.

Heckman, J. J. (1976, May). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement, Volume 5, number 4*, NBER Chapters, pp. 475--492. National Bureau of Economic Research, Inc.

Heckman, J. J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica 46*(4), 931--959.

Heckman, J. J. (2010). Building bridges between structural and program evaluation approaches to evaluating policy. *Journal of Economic Literature 48*(2), 356--398.

Heim, B. T. (2009). The effect of recent tax changes on taxable income: Evidence from a new panel of tax returns. *Journal of Policy Analysis and Management 28*(1), 147--163.

Holmlund, B. and M. Söderström (2008). Estimating dynamic income responses to tax reforms: Swedish evidence. *IFAU-Institute for Labour Market Policy Evaluation Working paper 2008:28*.

Hoynes, H. (1996). Welfare transfers in two-parent families: Labor supply and welfare participation under afdc-up. *Econometrica 64*(2), 295--332.

Huang, C.-I. (2008). Estimating demand for cellular phone service under nonlinear pricing. *Quantitative Marketing and Economics 6*(4), 371.

Imbens, G. W. (2010). Better late than nothing: Some comments on deaton (2009) and heckman and urzua (2009). *Journal of Economic Literature 48*(2), 399--423.

Ito, K. (2014). Do consumers respond to marginal or average price? evidence from nonlinear electricity pricing. *The American Economic Review 104*(2), 537--563.

Keane, M. P. (2010). Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics 156*(1), 3--20.

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

Kleven, H. J. and E. A. Schultz (2014). Estimating taxable income responses using danish tax reforms. *American Economic Journal: Economic Policy 6*(4), 271--301.

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

Kleven, H. J. and M. Waseem (2013b). 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. (2005). Tax bases, tax rates and the elasticity of reported income. *Journal of Public Economics 89*(11), 2093--2119.

Kopczuk, W. (2012). The polish business “flat” tax and its effect on reported incomes: a pareto improving tax reform. *Working Paper Columbia*.

Lindsey, L. B. (1987). Individual taxpayer response to tax cuts: 1982--1984: with implications for the revenue maximizing tax rate. *Journal of Public Economics 33*(2), 173--206.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables* (2 ed.). SAGE Publications.

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

Moffitt, R. A. and M. Wilhelm (2000). Taxation and the labor supply: Decisions of the affluent. In J. Slemrod (Ed.), *Does Atlas Shrug: The Economic Consequences of Taxing the Rich*, pp. 193--234. New York: Russell Sage Foundation.

Nelson, F. and L. Olson (1978). Specification and estimation of a simultaneous-equation model with limited dependent variables. *International Economic Review 19*(3), 695--709.

Nelson, F. D. (1981). A test for misspecification in the censored normal model. *Econometrica 49*(5), 1317--1329.

Newey, W. K. (1988). Two step series estimation of sample selection models.

Olmstead, S. M., W. M. Hanemann, and R. N. Stavins (2007). Water demand under alternative price structures. *Journal of Environmental Economics and Management 54*(2), 181--198.

Pagan, A. and A. Ullah (1999). *Nonparametric Econometrics*. Cambridge University Press.

Patel, E., N. Seegert, and M. G. Smith (2016). At a loss: The real and reporting elasticity of corporate taxable income. *Available at SSRN: https://ssrn.com/abstract=2608166 or http://dx.doi.org/10.2139/ssrn.2608166*.

Powell, J. L. (1987). Semiparametric estimation of bivariate latent variable models.

Pronzato, C. (2012). Comparing quasi-experimental designs and structural models for policy evaluation: The case of a reform of lone parental welfare. *IZA Discussion Papers 6803*.

Reiss, P. C. and M. W. White (2005). Household electricity demand, revisited. *The Review of Economic Studies 72*(3), 853--883.

Ruud, P. A. (1991). Extensions of estimation methods using the em algorithm. *Journal of Econometrics 49*(3), 305 -- 341.

Saez, E. (2003). The effect of marginal tax rates on income: a panel study of 'bracket creep'. *Journal of Public Economics 87*(5), 1231--1258.

Saez, E. (2004). Reported incomes and marginal tax rates, 1960-2000: evidence and policy implications. *Tax Policy and the Economy 18*, 117--173.

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

Selén, J. (2002). Taxable income responses to tax changes-a panel analysis of the 1990/91 swedish reform. Technical report, Trade Union Institute for Economic Research.

Sillamaa, M.-A. and M. R. Veall (2001). The effect of marginal tax rates on taxable income: a panel study of the 1988 tax flattening in canada. *Journal of Public Economics 80*(3), 341--356.

Singleton, P. (2011). The effect of taxes on taxable earnings: evidence from the 2001 and related us federal tax acts. *National Tax Journal 64*(2), 323.

Thoresen, T. O. and T. E. Vattø (2015). Validation of the discrete choice labor supply model by methods of the new tax responsiveness literature. *Labour Economics 37*, 38--53.

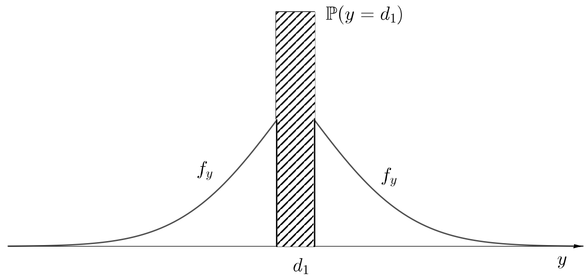
Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica 26*(1), 24--36.

Van Soest, A. (1995). Structural models of family labor supply: a discrete choice approach. *Journal of Human Resources*, 63--88.

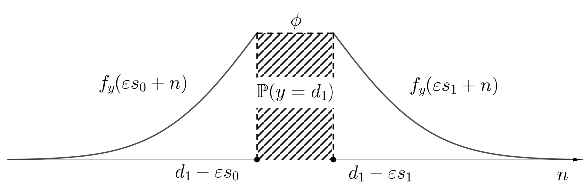
Weber, C. E. (2014). Toward obtaining a consistent estimate of the elasticity of taxable income using difference-in-differences. *Journal of Public Economics 117*, 90--103.

Figure 1: Non-identification of ε

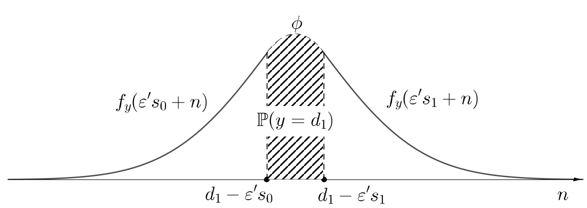
(a) Density and Probability Mass of an Observed Distribution F_y



(b) Density of a distribution $F_{n,\varepsilon}$ that is consistent with F_y along with ε

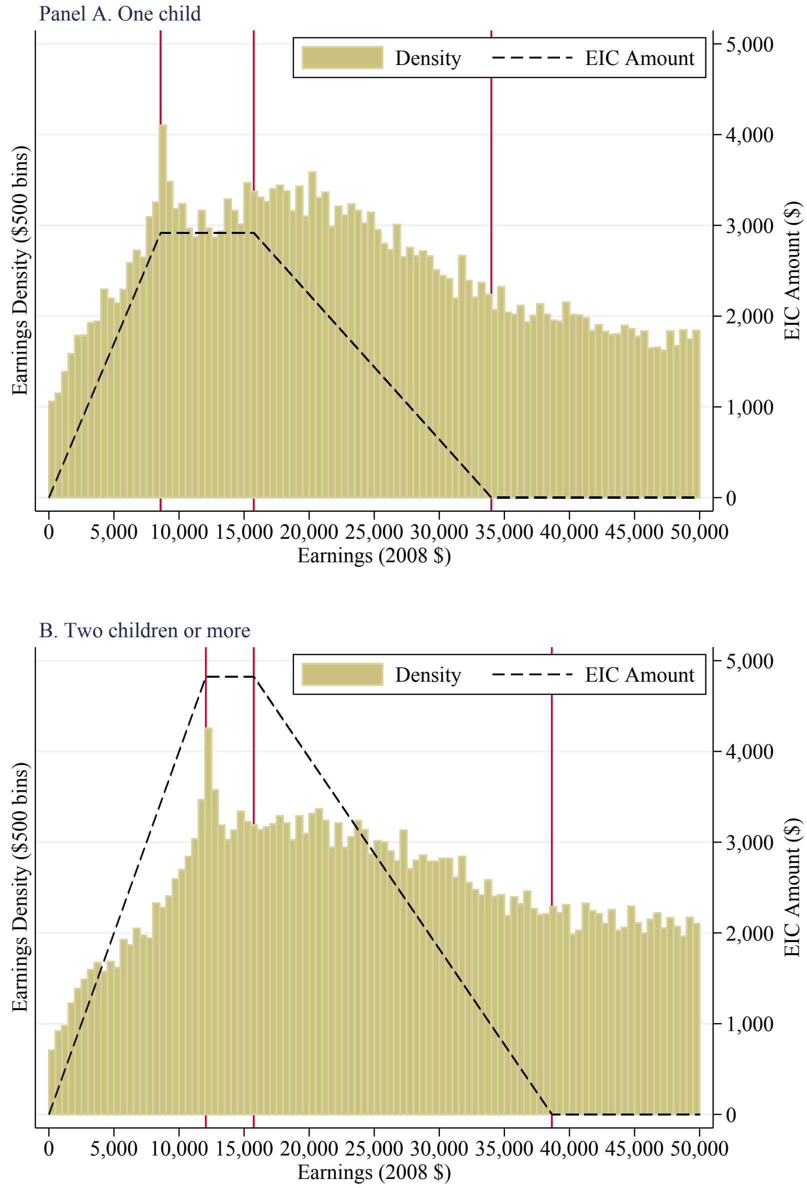


(c) Density of a distribution $F_{n,\varepsilon'}$ that is consistent with F_y along with $\varepsilon' < \varepsilon$



Notes: This figure illustrates two different distributions of n, $F_{n,\varepsilon}$ and $F_{n,\varepsilon'}$, that generate the same distribution F_y according to Equation (4) for two different values of the elasticity $\varepsilon > \varepsilon'$. The hatched area is the mass probability of bunching $\mathbb{P}(y = d_1)$. The shape of the left and right tails of 1b and 1c are identified up to location from the continuous part of 1a. The PDF of n in the hatched area is denoted ϕ. The function ϕ and its support are unknown except that ϕ integrates to $\mathbb{P}(y = d_1)$.

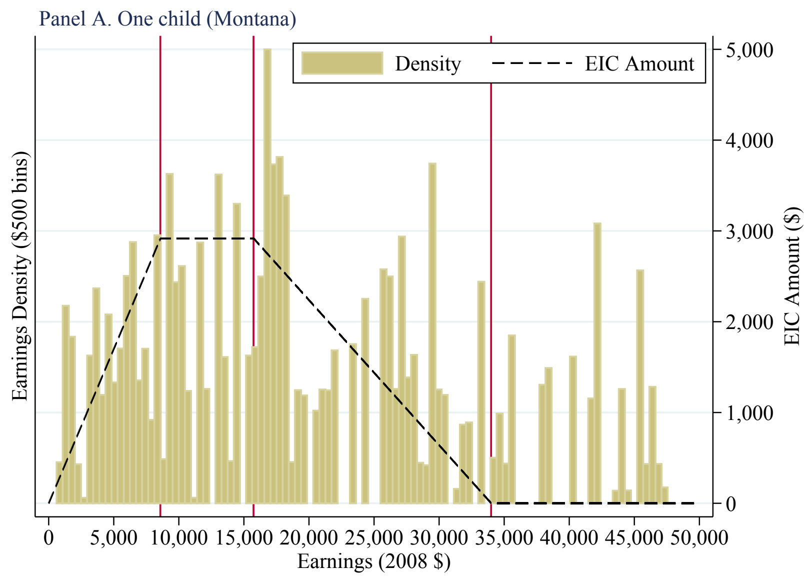
Figure 2



Note: Figure 2 replicates Figure 3 from Saez (2010). Each panel displays the histogram of earnings (by $500 bins) for tax filers by the number of dependent children for the years 1995 to 2004. Earnings are inflated to 2008 dollars using the IRS inflation parameters and are defined as wages and salaries plus self-employment income (net of one-half of the self-employed payroll tax). The EITC schedule is depicted with a dashed black line and the three kinks are depicted with vertical red lines. Panel A above includes 58,095 observations representing 116.3 million tax returns compared to the 57,692 observations from Saez (2010). Panel B above includes 67,426 observations representing 115.4 million tax returns compared to the 67,038 observations in the original.

Figure 3: The distribution of income across U.S. States

(a) One Child Montana



(b) One Child Colorado

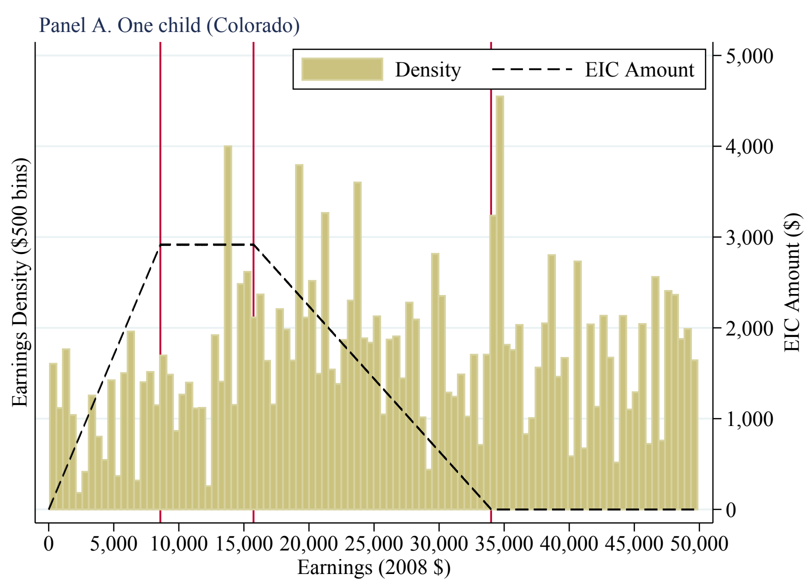
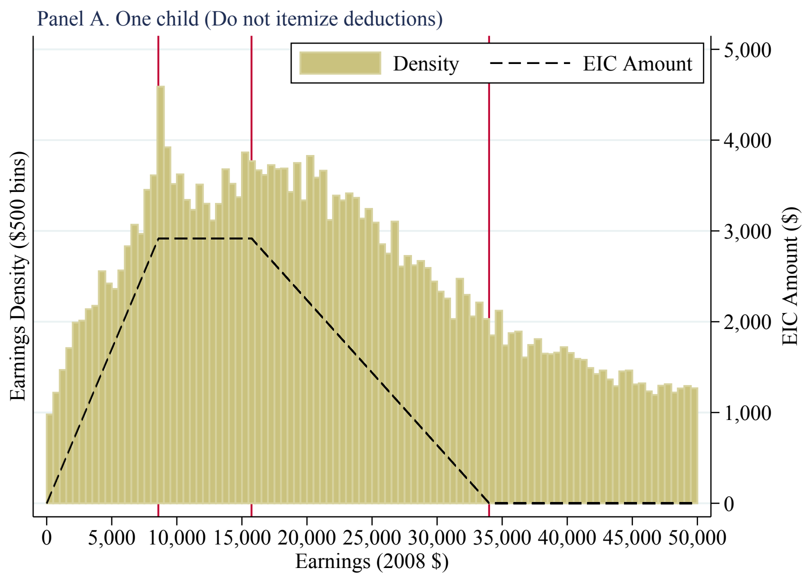


Figure 4: The distribution of income by itemization status

(a) One Child Do not Itemize



(b) One Child Itemize

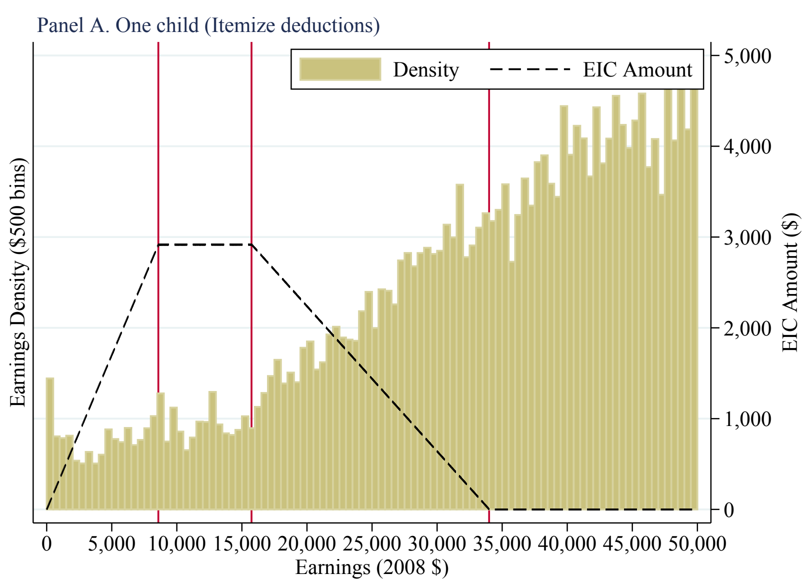
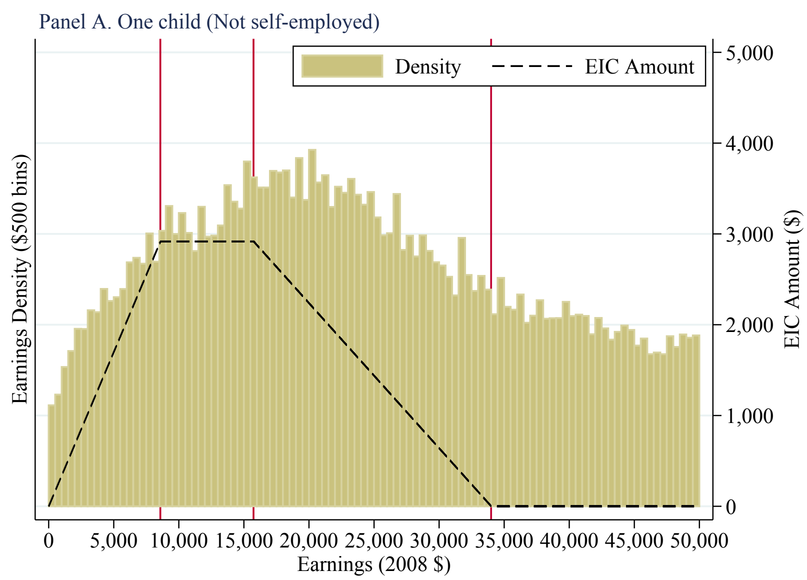


Figure 5: The distribution of income for employees or self-employed

(a) One Child Not Self-employed



(b) One Child Self-employed

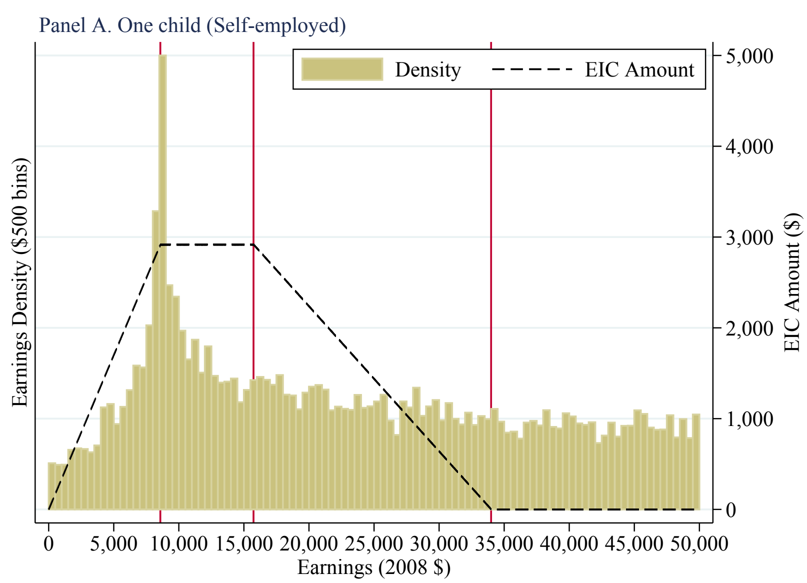
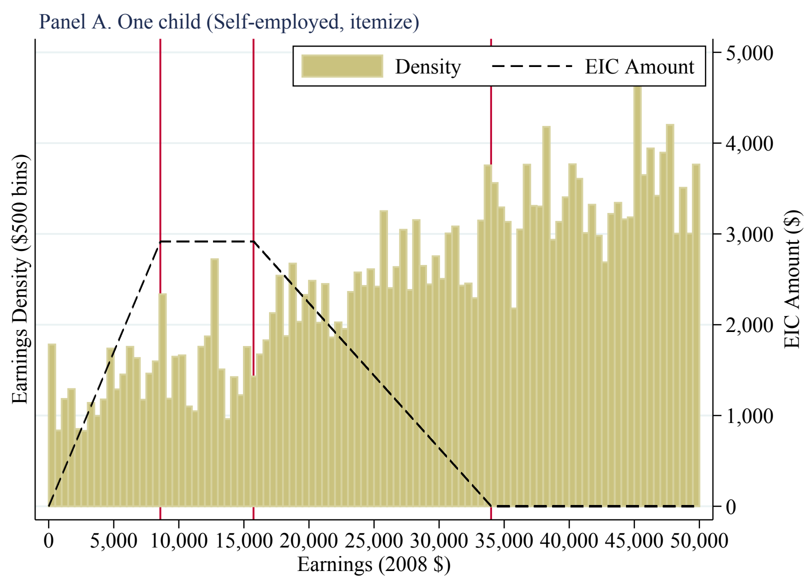


Figure 6: The distribution of income for interacted covariates

(a) One Child Self-employed and Itemize



(b) One Child Self-employed and Do Not Itemize

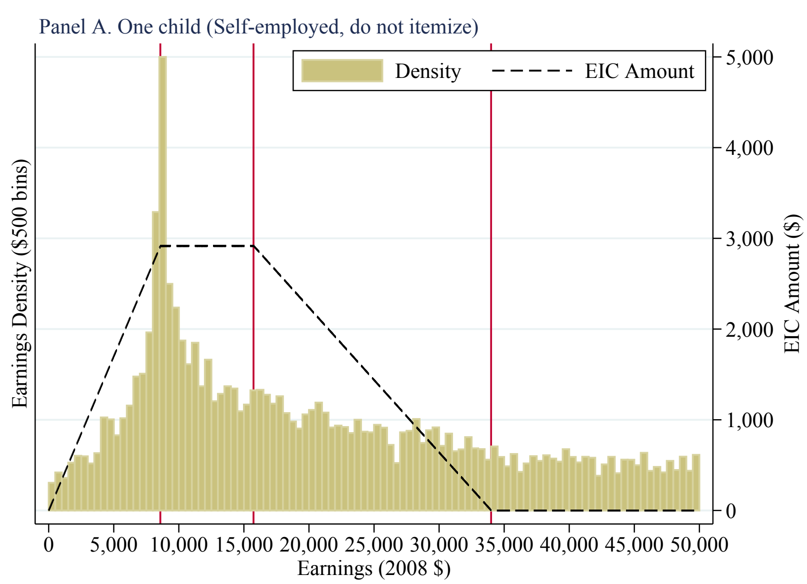
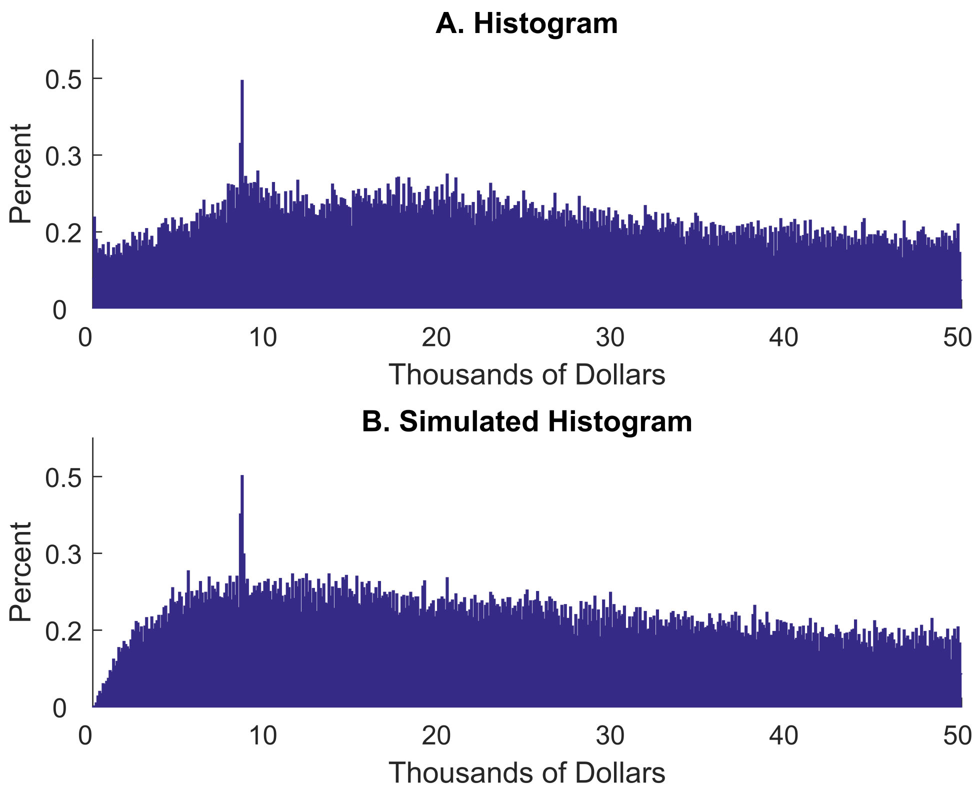
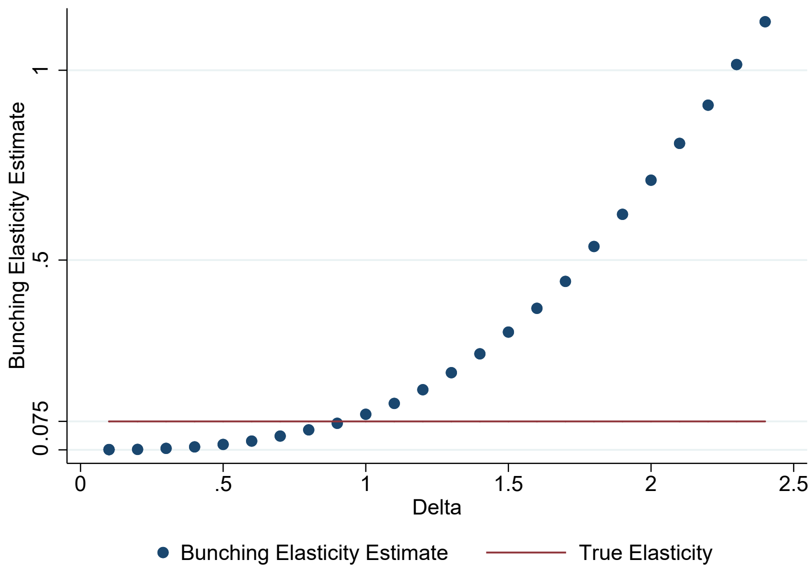


Figure 7: Actual and simulated income distributions



Note: Figure 7 replicates Figure 3A from Saez (2010). The top panel displays the histogram of earnings (by \$100 bins) for tax filers for the years 1995 to 2004 and includes 58,095 unweighted observations. Earnings are inflated to 2008 dollars using the IRS inflation parameters and are defined as wages and salaries plus self-employment income (net of one-half of the self-employed payroll tax). Panel B above includes 59,783 simulated observations based on a truncated Tobit 1 model with the tax elasticity, $\varepsilon = 0.075$, the optimizing friction standard deviation, $\sigma_\xi = 0.007$, the standard deviation of ability, $\sigma_\zeta = 1.4$ and the mean of ability conditional on parameters and covariates estimated for each individual.

Figure 8: Elasticity Estimates as the Window Increases



Note: This figure uses the Saez (2010) code to estimate the elasticity using data that is simulated to look like the actual EITC data. The simulated data includes 59,783 simulated observations based on a truncated Tobit 1 model with the tax elasticity, $\varepsilon = 0.075$, the optimizing friction standard deviation, $\sigma_\xi = 0.007$, the standard deviation of ability, $\sigma_\zeta = 1.4$ and the mean of ability conditional on parameters and covariates estimated for each individual. The horizontal axis gives different values of δ, which represents the window around the kink point used to estimate the amount of bunching.

Table 1: **Estimates of the Elasticity of Taxable Income 1/2**

Study	Years	Data	Method	Estimates
Lindsey (1987)	1981 ERTA	Tax returns (IRS), public	Dif-in-Dif	(1, 3)
Feldstein (1995)	1986 TRA	Tax returns (IRS), public	Dif-in-Dif	(1, 3)
Auten and Carroll (1999)	1985--1989	Tax returns (IRS), confidential	IV	0.57
Goolsbee (1999)	1920--1966	Statistics of Income (SOI)	Dif-in-Dif	(-0.83, 0.59)
Moffitt and Wilhelm (2000)	1983 and 1989	Survey of Consumer Finances (SCF)	Dif-in-Dif	(1.76, 1.99)
Moffitt and Wilhelm (2000)	1983 and 1989	Survey of Consumer Finances (SCF)	IV	(0.35, 0.97)
Aarbu and Thoresen (2001)	1991--1994	Statistics Norway's Income Distribution Survey	IV	(-0.6, 0.2)
Sillamaa and Veall (2001)	1986--1989	Statistics Canada Longitudinal Administrative Database	IV	0.25
Selén (2002)	1990--1995	household income survey (HINK) Statistics of Sweden	Dif-in-Dif	(0.2, 0.4)
Gruber and Saez (2002)	1979--1990	Panel, public-use	IV	0.4
Saez (2003)	1979--1981	Continuous Work History File (IRS)	IV using bracket creep	(0.2, 0.7)
Saez (2004)	1960--2000	Tax returns (IRS)	Time-series regression	0.2
Kopczuk (2005)	1979--1990	Michigan Tax Panel	IV	(0.2, 0.57)

Table 2: **Estimates of the Elasticity of Taxable Income 2/2**

Study	Years	Data	Method	Estimates
Giertz (2007)	1979--2001	Continuous Work History File (IRS)	IV	(0.26, 0.40)
Hansson (2007)	1989--1992	Longitudinell individdatabas (Sweden)	IV	(0.4, 0.5)
Holmlund and Söderström (2008)	1991--2002	Longitudinal Individual Dataset for Sweden	IV	(0.1, 0.3)
Auten, Carroll, and Gee (2008)	1999--2005	Tax Returns, Statistics of Income	IV	0.4
Heim (2009)	1999--2005	Tax Returns, 1999--2005 Edited Panel	IV	(0.3, 0.4)
Blomquist and Selin (2010)	1981--1991	Swedish Level of Living Survey	IV	0.21
Singleton (2011)	2002--2003	Current Population Survey lined to W-2 records	IV	(0.217, 0.304)
Kopczuk (2012)	2002--2005	Tax Return, Polish Ministry of Finance	IV	1
Gelber, Jones, and Sacks (2013)	1961--2006	Social Security	Bunching---updated	0.23
Kleven and Waseem (2013b)	2006--2008	Federal Board of Revenue in Pakiston	Bunching, notches	(0.07, 0.24)
Gelber (2014)	1988--1991	Longitudinal Individual Dataset for Sweden	IV	(0.41, 0.47)
Weber (2014)	1982--1990	Michigan IRS Tax Panel data	IV updated	1.046
Kleven and Schultz (2014)	1984--2005	Statistics Denmark	Dif-in-Dif	(0.2, 0.3)
Doerrenberg, Peichl, and Siegloch (2015)	2001--2008	German Taxpayer Panel	IV updated	(0.34, 0.68)

A Utility maximization details

This section begins with an iso-elastic and quasi-linear utility and derives the piecewise demand function when the budget constraint includes kinks and notches. The demand with a kink is reported in the text in example 1. The demand with the notch is not reported in the text and, as we demonstrate, is the demand from Kleven and Waseem (2013a). This section then derives the piecewise demand for a constant elasticity of substitution utility when the budget constraint includes kinks and notches. The demand with a notch is reported in the text in example 2.

A.1 Iso-Elastic and quasilinear Utility

Agents maximize their iso-elastic and quasi-linear utility subject to their budget constraint with discontinuities,

$$\begin{aligned} \max_{C_i,L_i} \quad & C_i - (N_i^*)^{-1/\varepsilon} \frac{L_i^{1+\frac{1}{\varepsilon}}}{1+1/\varepsilon} \\ s.t. \quad Y_i &= L_i \\ C_i &= S_0 Y_i + \sum_j^J \left[\Delta I_j + \Delta S_j \left(Y_j - D_j\right)\right] \mathbb{1}\left(Y_i > D_j\right) \end{aligned} \tag{13}$$

The budget constraint allows for a piecewise function where the slope changes by $\Delta S_j \equiv S_j - S_{j-1}$ and the intercept can change ΔI_j, often called notches. A proportional notch, as defined by Kleven and Waseem (2013a), is where above the discontinuity there is an additional proportional tax on the *entire* value of Y_i. In our model, a proportional notch is given by $\Delta S_j < 0$ and $\Delta I_j = \Delta S_j D_j$. For example, if a property sale price, Y_i is under \$1,000,000 the seller does not face a surtax and prices above \$1,000,0000 are subject to a 10 percent surtax $0.1 * Y_i$. In our model this is given by $\Delta S_j > -.1$ and $\Delta I_j = -.1 * 1,000,000 = -100,000$.

To better understand the budget constraint, we can rewrite the budget constraint assuming that $Y_i > D_1$ and $Y_i < D_2$, and $\Delta I_j = 0\ \forall j$,

$$\begin{aligned} C_i &= I + S_0 Y_i + \sum_{j=1}^J [\Delta I_j + \Delta S_j (Y_i - D_j)] \mathbb{1}(Y_i > D_j) \\ &= I + S_0 Y_i + (\Delta I_1 + (S_1 - S_0)(Y_i - D_1)) \mathbb{1}(Y_i > D_1) \\ &= I + (S_0 - S_0 + S_1) Y_i - (S_1 - S_0) D_1 \\ &= I + S_1 Y_i - (S_1 - S_0) D_1 \end{aligned}$$

From this expression, it is clear that the budget constraint can be written as the price in the relevant piece of the budget constraint times the demand Y_i minus a constant term equal to the difference in prices multiplied by the discontinuity level, D_1. This provides the intuition for the derived demand as a function of only the price in the relevant piece of the budget constraint. Specifically, the first-order condition is given by,

$$\frac{\partial u(Y_i)}{\partial Y_i} = \frac{\partial C_i(Y_i)}{\partial Y_i} - \left(\frac{Y_i}{N_i^*}\right)^{1/\varepsilon} = S_j - \left(\frac{Y_i}{N_i^*}\right)^{1/\varepsilon}.$$

The log of demand, denoted $y_i \equiv log(Y_i)$, is a piecewise function of the log of the heterogeneity variable $n_i \equiv \ln N_i^*$, and the log of the price, $s_j \equiv \ln S_j$,

$$y_i = \begin{cases} d_j & \text{if } n_i \in [\underline{n}_j, \overline{n}_j] \quad \text{for } j = 1, ..., J \\ n_i + \varepsilon p_{j-1} & \text{if } n_i \in [\overline{n}_{j-1}, \underline{n}_j] \quad \text{for } j = 1, ..., J+1 \end{cases}$$

where $d_j \equiv \ln D_j$, $\overline{n}_0 \equiv -\infty$, and $\underline{n}_{J+1} \equiv \infty$.

A.1.1 Kink

With a concave kink discontinuity, some agents decrease their Y, such that some mass bunch at the discontinuity. The agents that bunch at discontinuity D_j are those with $N_i^* \in [\underline{N}_j, \overline{N}_j]$. When the discontinuity is a kink, the thresholds are determined by noting that the lowest N_i^* that bunches, $\underline{N}$ chooses $Y_i = D_j$, with the price S_{j-1} and the highest N_i^* that bunches, $\overline{N}$ chooses $Y_i = D_j$, with the price S_j. This implies $D_j = \underline{N} S_{j-1}^{\varepsilon}$ and $D_j = \overline{N} S_j^{\varepsilon}$, and thus $\underline{N} = D_j S_{j-1}^{-\varepsilon}$ and $\overline{N} = D_j S_j^{-\varepsilon}$.

A.1.2 Notches

The lowest type of agent that bunches at the notch, denoted by $\underline{N}_1$, has an indifference curve that is tangent to the budget constraint at the notch. The highest type of agent that bunches at the notch, denoted by $\overline{N}$, has an indifference curve that intersects the budget constraint at the notch and is tangent to the budget constraint right of the notch. These thresholds can be written as,

$$\begin{aligned} \underline{N}_j &= D_j S_{j-1}^{-\varepsilon} \\ \overline{N}_j &= \frac{\Delta I (1+\varepsilon) S_{j-1}^{-1-\varepsilon}}{\varepsilon \gamma^{1+1/\varepsilon} - \gamma(1+\varepsilon) + S_j^{1+\varepsilon} S_{j-1}^{-1-\varepsilon}} \end{aligned}$$

where $\gamma = \underline{N}/\overline{N}$. To derive $\overline{N}$, we set the utility at the notch, $Y = D_j = \underline{N}S_{j-1}^{\varepsilon}$, and to the right of the notch, $Y = S_j^{\varepsilon}\overline{N}$, equal for this agent.

$$\begin{aligned}
S_{j-1}D_j - (1+1/\varepsilon)^{-1}\overline{N}_j^{-1/\varepsilon}D_j^{1+1/\varepsilon} &= S_jS_j^{\varepsilon}\overline{N}_j - (1+1/\varepsilon)^{-1}\overline{N}_j^{-1/\varepsilon}(S_j^{\varepsilon}\overline{N}_j)^{1+1/\varepsilon} - \Delta I\\
S_{j-1}D_j - (1+1/\varepsilon)^{-1}\overline{N}_j^{-1/\varepsilon}D_j^{1+1/\varepsilon} &= S_j^{1+\varepsilon}\overline{N}_j - (1+1/\varepsilon)^{-1}\overline{N}_jS_j^{1+\varepsilon} - \Delta I\\
\underline{N}_jS_{j-1}^{1+\varepsilon} - (1+1/\varepsilon)^{-1}\overline{N}_j^{-1/\varepsilon}\underline{N}_j^{1+1/\varepsilon}S_{j-1}^{1+\varepsilon} &= (1+\varepsilon)^{-1}\overline{N}_jS_j^{1+\varepsilon} - \Delta I\\
\overline{N}_j[(1+1/\varepsilon)^{-1}S_{j-1}^{1+\varepsilon}\gamma^{1+1/\varepsilon} - S_{j-1}^{1+\varepsilon}\gamma + (1+\varepsilon)^{-1}S_j^{1+\varepsilon}] &= \Delta I\\
\overline{N}_j[\varepsilon\gamma^{1+1/\varepsilon} - \gamma(1+\varepsilon) + S_j^{1+\varepsilon}S_{j-1}^{-1-\varepsilon}] &= \Delta I(1+\varepsilon)S_{j-1}^{-1-\varepsilon}\\
\overline{N}_j &= \frac{\Delta I(1+\varepsilon)S_{j-1}^{-1-\varepsilon}}{\varepsilon\gamma^{1+1/\varepsilon} - \gamma(1+\varepsilon) + S_j^{1+\varepsilon}S_{j-1}^{-1-\varepsilon}}
\end{aligned}$$

These expressions demonstrate several important characteristics about notches. First, the lower threshold in the notch case is exactly the same as in the kink case. Second, the upper threshold, however, differs. In particular the penultimate line shows that when $\Delta I_j = 0$ then $\gamma = 1$ and $\overline{N} = \underline{N}$.

Example Kleven and Waseem (2013a)

In this example we derive the same equation as equation (5) of Kleven and Waseem (2013a). To do this, we note that $\overline{N} = D_j(1+\Delta Y/D_j)S_{j-1}^{-\varepsilon}$ and $\gamma \equiv \underline{N}/\overline{N}$, can write as $1/(1+\Delta Y/D_j)$ to match the notation in Kleven and Waseem (2013a).[7]

$$\begin{aligned}
\overline{N}[\varepsilon\gamma^{1+1/\varepsilon} - \gamma(1+\varepsilon) + S_j^{1+\varepsilon}S_{j-1}^{-1-\varepsilon}] &= \Delta I(1+\varepsilon)S_{j-1}^{-1-\varepsilon}\\
(1+\varepsilon)^{-1}\varepsilon\gamma^{1+1/\varepsilon} - \gamma + (1+\varepsilon)^{-1}S_j^{1+\varepsilon}S_{j-1}^{-1-\varepsilon} &= \Delta IS_{j-1}^{-1-\varepsilon}\overline{N}^{-1}\\
(1+\varepsilon)^{-1}\varepsilon\left(\frac{1}{1+\Delta Y/Y}\right)^{1+1/\varepsilon} - \frac{1}{1+\Delta Y/Y} + (1+\varepsilon)^{-1}S_j^{1+\varepsilon}S_{j-1}^{-1-\varepsilon} &= \Delta IS_{j-1}^{-1}\frac{Y^{-1}}{1+\Delta Y/Y}\\
\frac{1}{1+\Delta Y/Y}\left(1+\frac{\Delta I/Y}{S_{j-1}}\right) - \frac{1}{1+1\varepsilon}\left(\frac{1}{1+\Delta Y/Y}\right)^{1+1/\varepsilon} - \frac{1}{1+\varepsilon}S_j^{1+\varepsilon}S_{j-1}^{-1-\varepsilon} &= 0
\end{aligned}$$

This equation is exactly the same as equation (5) in Kleven and Waseem (2013a).

A.2 CES Utility with only intercept changes

Much of the literature studying changes in the intercept of the budget constraint, called notches in public finance, uses the same isoelastic and quasi-linear utility function as Saez

[7] Note that $\overline{N}_j = D_j(1+\Delta Y/D_j)S_{j-1}^{-\varepsilon}$ with S_{j-1} not S_j because the marginal agent that bunches at the notch moves in two steps, the first step from the old price to the new price and in the second step from the new price interior solution to the notch.

(2010), for example, Kleven and Waseem (2013a); Best and Kleven (2017). To demonstrate the generality of our model, in this example, we use a constant elasticity of substitution (CES) utility, in which ψ is the constant elasticity of substitution between consumption and labor.

Agents maximize CES utility subject to a budget constraint with one intercept change such that $\Delta I_1 \neq 0$ while also facing a linear tax $S_1 = S_0 = (1-t)$ and earning exogenous non-labor income, R_i. Formally this is

$$\begin{aligned} \max_{C_i, L_i} \quad & \left(C_i^{\frac{\psi-1}{\psi}} - (N_i^*)^{\frac{-1}{\psi}} L_i^{\frac{\psi-1}{\psi}} \right)^{\frac{\psi}{\psi-1}} \\ s.t. \quad Y_i &= L_i \\ C_i &= (1-t) Y_i + \Delta I_1 \mathbb{1}(Y_i > D_1) + R. \end{aligned} \tag{14}$$

And optimal reported labor supply and therefore income is again a simple piecewise function,

$$Y_i = \begin{cases} \frac{(1-t)^{-\psi}}{N_i^* - (1-t)^{1-\psi}} \left((R + \Delta I \right) & N_i^* < \underline{N}_1 \\ D_1 & N_i^* \in \left[\underline{N}_1, \overline{N}_1 \right] \\ \frac{(1-t)^{-\psi}}{N_i^* - (1-t)^{1-\psi}} R & N_i^* > \overline{N}_1 \end{cases} \tag{15}$$

When there is a change in intercepts, a notch in the budget constraint, at the discontinuity D_j the threshold $\underline{N}_j$ is defined by the ability that has an indifference curves that is tangent to the budget constraint at D_j. The threshold $\overline{N}_j$ is similarly defined by the ability that is indifferent between reporting $Y_i = D_j$ at the discontinuity point and reporting their optimal location above it, $Y_i > D_j$.

A.2.1 CES Utility with intercept and slope changes

Agents maximize their constant elasticity of substitution utility, where ψ is the constant elasticity, subject to their budget constraint with exogenous income, R, and for simplicity one discontinuity, D_1,

$$\max_{Y_i, C_i} u\left(Y_i, C_i; N_i^*\right) = \left[C_i^{\frac{\psi-1}{\psi}} - (N_i^*)^{-\frac{1}{\psi}} Y_i^{\frac{\psi-1}{\psi}} \right]^{\frac{\psi}{\psi-1}}$$

subject to

$$C_i = S_0 Y_i + (\Delta I_1 + \Delta S_1 (Y_j - D_1)) \mathbb{1}(Y_i > D_1) + R_i.$$

The Lagrangian and first-order conditions can be written as,

$$\mathcal{L} = \left[C_i^{\frac{\psi-1}{\psi}} - (N_i^*)^{-\frac{1}{\psi}} Y_i^{\frac{\psi-1}{\psi}} \right]^{\frac{\psi}{\psi-1}} + \lambda \left[S_0 Y_i + (\Delta I_1 + \Delta S_1 (Y_j - D_1)) \mathbb{1}(Y_i > D_1) + R_i - C_i \right]$$

$$\frac{\partial \mathcal{L}}{\partial C_i} = C_i^{-\frac{1}{\psi}} \left[C_i^{\frac{\psi-1}{\psi}} - (N_i^*)^{-\frac{1}{\psi}} Y_i^{\frac{\psi-1}{\psi}} \right]^{\frac{1}{\psi-1}} - \lambda = 0$$

$$\frac{\partial \mathcal{L}}{\partial Y_i} = -(N_i^*)^{-\frac{1}{\psi}} Y_i^{-\frac{1}{\psi}} \left[C_i^{\frac{\psi-1}{\psi}} - (N_i^*)^{-\frac{1}{\psi}} Y_i^{\frac{\psi-1}{\psi}} \right]^{\frac{1}{\psi-1}} + \lambda \left[S_0 + \Delta S_1 \mathbb{1}(Y_i > D_1) \right] = 0.$$

Combining the first-order conditions and the budget constraint gives,

$$(N_i^*)^{-\frac{1}{\psi}} Y_i^{-\frac{1}{\psi}} = C_i^{-\frac{1}{\psi}} \left[S_0 + \Delta S_1 \mathbb{1}(Y_i > D_1) \right]$$

$$N_i^* Y_i = (S_0 Y_i + (\Delta I_1 + \Delta S_1 (Y_j - D_1)) \mathbb{1}(Y_i > D_1) + R) \left[S_0 + \Delta S_1 \mathbb{1}(Y_i > D_1) \right]^{-\psi}$$

$$Y_i = \frac{((\Delta I_1 - \Delta S_1 D_1) \mathbb{1}(Y_i > D_1) + R) \left[S_0 + \Delta S_1 \mathbb{1}(Y_i > D_1) \right]^{-\psi}}{N_i^* - (S_0 - \Delta S_1 \mathbb{1}(Y_i > D_1)) \left[S_0 + \Delta S_1 \mathbb{1}(Y_i > D_1) \right]^{-\psi}}$$

The demand, Y_i is a piecewise function depending on whether or not $Y_i > D_1$, which can be written in terms of the heterogeneity variable N_i^*,

$$Y_i = \begin{cases} \frac{S_1^{-\psi}}{N_i^* - S_1^{\psi-1}} (R_i + \Delta I - \Delta S_1 D_1) & \text{if } N_i^* > \overline{N}_1 \\ D_1 & \text{if } N_i^* \in [\underline{N}_1, \overline{N}_1] \\ \frac{S_0^{-\psi}}{N_i^* - S_0^{\psi-1}} R_i & \text{if } N_i^* < \underline{N}_1. \end{cases}$$

This demand function encompasses both kinks and notches. The following two subsections give the simplified demand functions with only a kink or only a notch and explicitly solves for the thresholds, $\overline{N}_i$ and $\underline{N}_i$

A.2.2 Kinks

With a concave kink discontinuity, some agents decrease their Y_i, such that some mass bunch at the discontinuity. The agents that bunch at discontinuity D_j are those with $N_i^* \in [\underline{N}_j, \overline{N}_j]$. When the discontinuity is a kink, the thresholds are determined by noting that the lowest N_i^* that bunches, $\underline{N}$ chooses $Y_i = D_j$, with the slope S_{j-1} and the highest N_i^* that bunches, $\overline{N}$ chooses $Y_i = D_j$, with the slope S_j. These thresholds can be written as,

$$\begin{aligned}
D_1 &= \frac{S_0^{-\psi}}{\underline{N} - S_0^{\psi-1}} R \\
(\underline{N} - S_0^{\psi-1}) D_1 &= S_0^{-\psi} R \\
\underline{N} D_1 &= S_0^{-\psi} R + S_0^{\psi-1} D_1 \\
\underline{N} &= \frac{S_0^{-\psi} R + S_0^{\psi-1} D_1}{D_1}
\end{aligned}$$

and

$$\overline{N} = \frac{S_1^{-\psi}(R - \Delta S_1 D_1) + S_1^{\psi-1} D_1}{D_1}$$

The demand function can then be written as,

$$Y_i = \begin{cases} \frac{S_1^{-\psi}}{N_i^* - S_1^{\psi-1}}(R_i - \Delta S_1 D_1) & \text{if } N_i^* > \frac{S_1^{-\psi}(R - \Delta S_1 D_1) + S_1^{\psi-1} D_1}{D_1} \\ D_1 & \text{if } N_i^* \in [\frac{S_0^{-\psi} R + S_0^{\psi-1} D_1}{D_1}, \frac{S_1^{-\psi}(R - \Delta S_1 D_1) + S_1^{\psi-1} D_1}{D_1}] \\ \frac{S_0^{-\psi}}{N_i^* - S_0^{\psi-1}} R_i & \text{if } N_i^* < \frac{S_0^{-\psi} R + S_0^{\psi-1} D_1}{D_1}. \end{cases}$$

A.2.3 Notches

The lowest type of agent that bunches at the notch, denoted by $\underline{N}_1$, chooses to be at the notch point given the budget constraint to the left of the notch. Said differently, their indifference curve is tangent to the notch point and can be written as

$$\begin{aligned}
D_1 &= \frac{S_0^{-\psi}}{\underline{N} - S_0^{\psi-1}} R \\
(\underline{N} - S_0^{\psi-1}) D_1 &= S_0^{-\psi} R \\
\underline{N} D_1 &= S_0^{-\psi} R + S_0^{\psi-1} D_1 \\
\underline{N}_1 &= \frac{S_0^{-\psi} R + S_0^{\psi-1} D_1}{D_1}.
\end{aligned}$$

The highest type of agent that bunches at the notch, denoted by $\overline{N}$, receives the same utility from bunching at the notch or consuming to the right of the notch given the budget constraint to the right of the notch. This condition can be written as an implicit function

$$\left[(D(S+(R/D))^{\frac{\psi-1}{\psi}} - \overline{N}^{-\frac{1}{\psi}} D_1^{\frac{\psi-1}{\psi}}\right]^{\frac{s}{s-1}} = \left[C_i^{\frac{\psi-1}{\psi}} - \overline{N}^{-\frac{1}{\psi}} (Y_i)^{\frac{\psi-1}{\psi}}\right]^{\frac{\psi}{\psi-1}}$$

$$\overline{N}_1 = \frac{D^{\frac{\psi-1}{\psi}} - Y^{\frac{\psi-1}{\psi}}}{\alpha_1 D^{\frac{\psi-1}{\psi}} - \alpha_2 Y^{\frac{\psi-1}{\psi}}}$$

where $\alpha_1 = (S + R/D)^{\frac{\psi-1}{\psi}}$, $\alpha_2 = (S + \tilde{R}/Y)^{\frac{\psi-1}{\psi}}$, and $\tilde{R} = R + \Delta I$. With these thresholds, the demand can be written as

$$Y_i = \begin{cases} \frac{S^{-\psi}}{N_i^* - S^{\psi-1}}(R_i + \Delta I) & \text{if } N_i^* > \overline{N}_1 \\ D_1 & \text{if } N_i^* \in [\underline{N}_1, \overline{N}_1] \\ \frac{S^{-\psi}}{N_i^* - S^{\psi-1}} R_i & \text{if } N_i^* < \underline{N}_1. \end{cases}$$

B Saez (2010)

This section presents the derivations behind key expressions in Saez (2010) using his notation. We include all details to make the solution abundantly accessible to the reader.

B.1 Utility Maximization in Saez (2010)

We begin with section, ‘‘B. Empirical Estimation of the Elasticity using Bunching’’ from (Saez, 2010, p. 185). Each agents solves a utility maximization problem of the form

$$\begin{aligned} \max_{c,z} \quad & c - \frac{n}{1+1/e}\left(\frac{z}{n}\right)^{1+1/e} \\ s.t. & \\ c \; = \; & z\left(1-t_0\right)1\left(z \leq z^*\right) + z\left(1-t_1\right)1\left(z > z^*\right) + R \end{aligned}$$

in which c is consumption, z is earnings, R is non-earning resources that can be spent on consumption, and the compensated elasticity of reported income with respect to (one minus) the marginal tax rate is $e \geq 0$. Without loss of generality, we assume that $t_0 \leq t_1$. This specification for utility gives that the higher the level of ability n the lower is the disutility of earning income level z for any $e > 0$. At $e = 0$, all agents solve the same problem because n no longer enters utility. The unobservable ability variable, n, is distributed according to some PDF $f(n)$ and some CDF $F(n)$. The utility function is decreasing in income because we implicitly assume that earning income labor supply which reduces utility. We could make this explicit by replacing z with labor supply l in the utility function then introducing a second constraint that sets $z = l$ so that nominal wage is one. Notice we need to have $z \in (0, \infty)$ and $n \in (0, \infty)$ as well. The Lagrangian to be maximized is

$$\mathcal{L} = c - \frac{n}{1+1/e}\left(\frac{z}{n}\right)^{1+1/e} + \lambda\left[z\left(1-t_0\right)1\left(z \leq z^*\right) + z\left(1-t_1\right)1\left(z > z^*\right) + R - c\right]$$

Differentiate with respect to c to get the first order condition (FOC)

$$\mathcal{L}_c = 1 - \lambda = 0$$

So that $\lambda = 1$ which captures the implicit assumption that the price of consumption, which is the price of relaxing the budget constraint, is normalized to one. Because we already set the nominal wage of labor to one, this also implies the real wage is equal to one. In order to maximize with respect to earned income, we consider three regions, which are $z < z^*$, $z = z^*$, and $z > z^*$. We can solve for the optimal z within these regions by solving the first order condition because $\mathcal{L}$ is differentiable with respect to z within them. Starting with the $z < z^*$ region the Lagrangian becomes

$$\mathcal{L} \mid z < z^* = c - \frac{n}{1+1/e}\left(\frac{z}{n}\right)^{1+1/e} + \lambda\left[z\left(1-t_0\right) + R - c\right]$$

So that the FOC within this region is

$$\mathcal{L}_z \mid z < z^* = -n\left(\frac{1}{n}\right)^{1+1/e} z^{1/e} + \lambda\left(1 - t_0\right) = 0$$

Solving this FOC for optimal z gives

$$z = n\lambda^e \left(1 - t_0\right)^e$$

We can follow the same steps for the $\mathcal{L} \mid z > z^*$ to get that

$$z = n\lambda^e \left(1 - t_1\right)^e$$

The solution to the case when $z = z^*$ is trivial because that's just it. Combining the solution in each region with the fact that that $\lambda = 1$ provides the case wise optimal solution for all regions of z

$$z = \begin{cases} n\left(1 - t_0\right)^e & z < z^* \\ z^* & z = z^* \\ n\left(1 - t_1\right)^e & z > z^* \end{cases} \tag{16}$$

Equation (16) is helpful but is not expressed as the optimal behavior by the agent because the cases define tautologies instead of decision rules. We can derive a decision rule for the $z < z^*$ and $z > z^*$ regions by substitution

$$z = \begin{cases} n\left(1 - t_0\right)^e & n\left(1 - t_0\right)^e < z^* \\ z^* & z = z^* \\ n\left(1 - t_1\right)^e & n\left(1 - t_1\right)^e > z^* \end{cases}$$

All that remains in order to characterize the full solution is to replace the tautology when $z = z^*$ with a decision rule. In order to do this, we must find the highest ability $\overline{n}$ that is indifferent between reporting $z = z^*$ and facing t_0 and reporting $z = \overline{n}\left(1 - t_1\right)^e$ but being taxed at t_0. This means we need to solve the following equality for $\overline{n}$

$$\begin{aligned} \mathcal{L}\left(z = z^*, \overline{n}, c\right) &= \mathcal{L}\left(z = \overline{n}\left(1 - t_1\right)^e, \overline{n}, c\right) \\ c - \frac{\overline{n}}{1 + 1/e}\left(\frac{z^*}{\overline{n}}\right)^{1+1/e} + \lambda\left[z^*\left(1 - t_0\right) + R - c\right] &= c - \frac{\overline{n}}{1 + 1/e}\left(\frac{\overline{n}\left(1 - t_1\right)^e}{\overline{n}}\right)^{1+1/e} + \lambda\left[\overline{n}\left(1 - t_1\right)^e\left(1 - t_0\right) + R - c\right] \end{aligned}$$

It is obvious that the solution is $\overline{n} = z^*/\left(1 - t_1\right)^e$. The steps to solve for $\underline{n}$ are analogous and lead to $\underline{n} = z^*/\left(1 - t_0\right)^e$. Combining these with equation (16) provides the final decision rule as a function of the latent and unobserved variable n

$$z = \begin{cases} n\left(1 - t_0\right)^e & n < z^*/\left(1 - t_0\right)^e \\ z^* & n \in \left[z^*/\left(1 - t_0\right)^e, z^*/\left(1 - t_1\right)^e\right] \\ n\left(1 - t_1\right)^e & n > z^*/\left(1 - t_1\right)^e \end{cases} \tag{17}$$

While expressed in a less compact way, equation (17) is precisely what is derived on (Saez, 2010, p. 186).

B.2 Inference in Saez (2010)

After solving the agent's utility maximization problem (Saez, 2010, equation (3)) defines Δz^* by considering the percent difference between the highest and lowest ability agents that report z^*. Namely, Δz^* is defined using

$$\frac{\overline{n} - \underline{n}}{\underline{n}} = \frac{z^*/\left(1-t_1\right)^e}{z^*/\left(1-t_0\right)^e} - 1 = \left(\frac{1-t_1}{1-t_0}\right)^e - 1 = \frac{\Delta z^*}{z^*}$$

which is exactly Saez (2010) equation (3). On that same page he defines the counterfactual linear tax PDF as

$$h_0\left(z\right) = H_0'\left(z\right)$$

in which

$$h_0\left(z\right) = f\left(z/\left(1-t_0\right)^e\right)/\left(1-t_0\right)^e$$

and

$$H_0\left(z\right) = F\left(z/\left(1-t_0\right)^e\right)$$

Following his notation, to the left hand side of the approximation in equation (4) we have

$$\begin{aligned}
B^{Saez} &= \int_{z^*}^{z^*+\Delta z^*} h_0\left(z\right) dz \\
&= H_0\left(z^*+\Delta z^*\right) - H_0\left(z^*\right) \\
&= F\left(\left(z^*+\Delta z^*\right)/\left(1-t_0\right)^e\right) - F\left(z^*/\left(1-t_0\right)^e\right) \\
&= F\left(\left(z^* + z^*\left(\frac{1-t_0}{1-t_1}\right)^e - z^*\right)/\left(1-t_0\right)^e\right) - F\left(z^*/\left(1-t_0\right)^e\right) \\
&= F\left(z^*\left(\frac{1-t_0}{1-t_1}\right)^e/\left(1-t_0\right)^e\right) - F\left(z^*/\left(1-t_0\right)^e\right) \\
&= F\left(z^*/\left(1-t_1\right)^e\right) - F\left(z^*/\left(1-t_0\right)^e\right)
\end{aligned}$$

Hence, the mass at the kink point z^* in (Saez, 2010, equation (4), p. 186), without approximating the integral, is identical to B^{Saez} defined in equation (6) of the main text.

C Proofs

C.1 Proof of Lemma 3.1

Proof. It suffices to show that for every $\varepsilon > 0$, there exists $F_{n,\varepsilon} \in \mathcal{F}_n$ such that $F_y = T\left(F_{n,\varepsilon}, \varepsilon, d_1, p_0, p_1\right)$ for fixed F_y, d_1, p_0, and p_1. To show the existence of such an $F_{n,\varepsilon}$, fix arbitrary $\varepsilon > 0$ and then construct $F_{n,\varepsilon}$ as follows:

1. First, define a continuous function $\phi : [d_1 - \varepsilon p_0, d_1 - \varepsilon p_1] \to \mathbb{R}_+$ such that:
   (a) $\phi(d_1 - \varepsilon p_0) = \lim_{u \uparrow d_1} f_y(u)$;
   (b) $\phi(d_1 - \varepsilon p_1) = \lim_{u \downarrow d_1} f_y(u)$;
   (c) $\int \phi(u)\ du = F_y(d_1) - \lim_{u \uparrow d_1} F_y(u)$.
2. Second, compute the CDF $F_{n,\varepsilon}$ as the integral of the following PDF:

$$f_{n,\varepsilon}(v) = \begin{cases} f_y\left(\varepsilon p_0 + v\right), & v \in (-\infty, d_1 - \varepsilon p_0) \\ \phi(v), & v \in [d_1 - \varepsilon p_0, d_1 - \varepsilon p_1] \\ f_y\left(\varepsilon p_1 + v\right) & v \in (d_1 - \varepsilon p_1, +\infty) \end{cases}$$

□

C.2 Example 1 - Saez Identification Restriction

A simple identification restriction on $\mathcal{F}_n$ is to assume that the density f_n follows a particular function on the interval $[\underline{n}_1, \overline{n}_1]$ where this function is known to the researcher. This is the type of identification restriction implicit in the analysis of Saez (2010).

Saez (2010) relies on the trapezoidal approximation of an integral of the density of Y_0, the income Y in the counter-factual scenario where t_1 is equal to t_0. More specifically,

$$\int_{D_1}^{D_1'} f_{Y_0}(y)\ dy \approx (D_1' - D_1)(f_{Y_0}(D_1') - f_{Y_0}(D_1))/2$$

where $D_1' = D_1((1 - t_0)/(1 - t_1))^{\varepsilon}$.

This approximation holds exactly if the density $f_{Y_0}(y)$ is a linear function of y for $y \in [D_1, D_1']$. Note that f_{Y_0} is a simple transformation of the density of N with $f_{Y_0}(y) = f_N\left(Y/(1 - t_0)^{\varepsilon}\right)/(1 - t_0)^{\varepsilon}$. Therefore, the trapezoidal approximation holds exactly if $f_N(n)$ is a linear function of n for $n \in [D_1/(1 - t_0)^{\varepsilon}, D_1/(1 - t_1)^{\varepsilon}]$. A linear shape on f_N within $[D_1/(1 - t_0)^{\varepsilon}, D_1/(1 - t_1)^{\varepsilon}]$ is equivalent to an exponential shape on f_n within $[\underline{n}_1, \overline{n}_1]$. Under this assumption, the elasticity ε is identified as the solution of Equation (5) by Saez (2010).

The linear approximation would be good a non-linear density f_N if the interval $[D_1/(1 - t_0)^{\varepsilon}, D_1/(1 - t_1)^{\varepsilon}]$ is very small. For a fixed value of ε, a observed small change in tax rates makes this interval small and the linear approximation may be good. The problem with this argument is that the quality of the approximation deteriorates the bigger the elasticity is. Therefore, we can never learn if the linear approximation is a good one because it depends on the unidentified quantity ε.

C.3 Example 2 - Parametric Restrictions

Other examples of parametric restrictions on $\mathcal{F}_n$ include a polynomial functional form on the PDF f_n. A polynomial form for f_n translates into a polynomial form for f_y. For example, Chetty et al. (2011) and Kleven and Waseem (2013a) identify the coefficients of a polynomial on f_y and take those coefficients to generate f_n. It is important to emphasize that even if f_y is non-parametrically identified, estimating a polynomial form for f_y is **not** a non-parametric identification strategy even if f_y is non-parametrically identified. Unless f_n over $[\underline{n}_1, \overline{n}_1]$ is restricted to be a parametric functional of the distribution of F_y, Lemma 3.1 says it is impossible to identify ε.

The researcher may impose restrictions on f_n that allow for more flexible shapes than linearity and still identify the elasticity. One may assume a parametric class of distributions such that $\mathcal{F}_n = \{F_n = F_\theta \ , \ \theta \in \Theta\}$ where F_θ are CDFs indexed by a parameter θ in a parameter space Θ. Identification consists of solving for the elasticity ε as a function of d_1, s_0, and s_1 using the following conditions.

Equation (7) is simply Equation (6) restated. Equations (8) and (9) come from the fact that the continuous part of the distribution of y has a PDF function that equals the PDF of n shifted. That is, $f_y(u) = f_n(u - \varepsilon s_0)$ for $u < d_1$, and $f_y(u) = f_n(u - \varepsilon s_1)$ for $u > d_1$. Aside from these unknown shifts, the shape of f_y is the same of the shape of f_n. This is valuable identifying information because the distribution of y is fully observed.

We demonstrate how to verify these conditions in the parametric Gaussian case. Suppose the distribution of n follows a normal distribution with unknown mean μ and unknown variance σ^2, such that $F_n(n) = F_{\mu,\sigma^2}(n) = \Phi\left(\frac{n-\mu}{\sigma}\right)$ where Φ denotes the standard normal CDF.

First, Equation (8) plus the normality assumption gives $F_y(u) = \Phi\left(\frac{u-\varepsilon s_0-\mu}{\sigma}\right)$. Inverting the CDF yields $(u - \varepsilon s_0 - \mu)/\sigma = \Phi^{-1}(F_y(u))$. The function on LHS is linear in u. The function on the RHS is identified from the data. Two distinct points $u' < u < d_1$ suffice to identify the slope $1/\sigma$ and intercept $-(\varepsilon s_0 + \mu)/\sigma$ of the LHS function. Hence, σ is identified, and $k_0 \equiv \varepsilon s_0 + \mu$ is identified. Similarly, Equation (9) identifies $k_1 \equiv \varepsilon s_1 + \mu$. Combining k_0 and k_1 identifies the elasticity $\varepsilon = (k_0 - k_1)/(s_1 - s_0)$.

C.4 Proof of Theorem 3 - Partial Identification

Assume $\mathcal{F}_n$ contains all distributions with continuous PDF f_n such that the maximum slope magnitude of f_n is $M \in (0, \infty)$. Then, the elasticity $\varepsilon \in \Upsilon$ where

$$\Upsilon = \begin{cases} \emptyset & , \text{ if } \mathbb{P}[y = d_1] < \frac{\left|f(d_1^+) - f(d_1^-)\right| \left[f(d_1^+) + f(d_1^-)\right]}{2M} \\ [\underline{\varepsilon}, \overline{\varepsilon}] & , \text{ if } \frac{\left|f(d_1^+) - f(d_1^-)\right| \left[f(d_1^+) + f(d_1^-)\right]}{2M} \leq \mathbb{P}[y = d_1] < \frac{f(d_1^+)^2 + f(d_1^-)^2}{2M} \\ [\underline{\varepsilon}, \infty) & , \text{ if } \frac{f(d_1^+)^2 + f(d_1^-)^2}{2M} \leq \mathbb{P}[y = d_1] \end{cases}$$

where $\emptyset$ is the empty set, and

$$\underline{\varepsilon} = \frac{2\left[f(d_1^+)^2/2 + f(d_1^-)^2/2 + M\ \mathbb{P}[y=d_1]\right]^{1/2} - \left(f(d_1^+) + f(d_1^-)\right)}{M(s_0 - s_1)}$$

$$\overline{\varepsilon} = \frac{-2\left[f(d_1^+)^2/2 + f(d_1^-)^2/2 - M\ \mathbb{P}[y=d_1]\right]^{1/2} + \left(f(d_1^+) + f(d_1^-)\right)}{M(s_0 - s_1)}$$

First, let's fix $\varepsilon > 0$. We look at all possible PDFs in $\mathcal{F}_n$ and compute the maximum and minimum integrals over the interval $[\underline{n}_0, \overline{n}_1]$. The length of this interval is $\varepsilon(s_0 - s_1)$. Thus, without loss of generality, we restrict our attention to f_n over the interval $[0, \varepsilon(s_0 - s_1)]$ such that:

(i) f_n is continuous, and it connects the point $(0, f_y(d_1^-))$ to $(\varepsilon(s_0 - s_1), f_y(d_1^+))$ in the (x,y) plane;

(ii) the absolute value of the slope of f_n is bounded by M.

First, start with f_n being a line. The magnitude of the slope is $\frac{|f_y(d_1^+) - f_y(d_1^-)|}{\varepsilon(s_0 - s_1)}$. Suppose this magnitude is bigger than M. Then, any f_n satisfying (i) will have a slope magnitude higher than M at some point. Therefore, we need to look at $\varepsilon \geq \varepsilon_1$ where $\varepsilon_1 = \frac{|f_y(d_1^+) - f_y(d_1^-)|}{M(s_0 - s_1)}$.

For fixed $\varepsilon \geq \varepsilon_1$, the slope of the line will be less or equal to M. The maximum possible area is attained when the function has the shape of hat whose lines attain the maximum slope. That is, a combination of two line segments. One that starts $(0, f_y(d_1^-))$ and has slope $+M$, and another that ends at $(\varepsilon(s_0 - s_1), f_y(d_1^+))$ and has slope $-M$. Call this function $\overline{f}_n$. These lines intersect at x^* where

$$x^* = \frac{f_y(d_1^+) - f_y(d_1^-) + M\varepsilon(s_0 - s_1)}{2M}.$$

Note that x^* is always such $0 \leq x^* \leq \varepsilon(s_0 - s_1)$ because $\varepsilon \geq \varepsilon_1$. Note that it is impossible to find another f_n that satisfies (i), it is greater than $\overline{f}_n$, and that has slope magnitude less or equal than M. The maximum area is

$$\overline{A}(\varepsilon) = \int_0^{\varepsilon(s_0 - s_1)} \overline{f}_n(v)\ dv$$
$$= (1/4M)\left[M^2\varepsilon^2 s_0^2 - 2M^2\varepsilon^2 s_0 s_1 + M^2\varepsilon^2 s_1^2 + 2M\varepsilon f_y(d_1^-)\right.$$
$$\left. s_0 - 2M\varepsilon f_y(d_1^-)s_1 + 2M\varepsilon f_y(d_1^+)s_0 - 2M\varepsilon f_y(d_1^+)s_1 - f_y(d_1^-)^2 + 2f_y(d_1^-)f_y(d_1^+) - f_y(d_1^+)^2)\right]$$

The function $\overline{A}(\varepsilon)$ is strictly increasing with respect to ε over $\varepsilon \geq \varepsilon_1$. In fact, the derivative is $((s_0 - s_1)(f_y(d_1^-) + f_y(d_1^+) + M\varepsilon(s_0 - s_1))/2$ which is strictly positive.

The minimum possible area is attained when the function has the shape of an inverted hat whose lines attain the maximum slope. That is, a combination of two line segments. One that starts $(0, f_y(d_1^-))$ and has slope $-M$, and another that ends at $(\varepsilon(s_0 - s_1), f_y(d_1^+))$ and has slope $+M$. Differently the hat function, the intersection (x^{**}, y^{**}) of this inverted hat function may or may not happen above the x-axis. That is, y^{**} may be negative, but f_n is always positive. In that case, we simply set the function to zero in the region where it would be negative. Call this function $\underline{f}_n$.

The intersection occurs at

$$x^{**} = \frac{f_y(d_1^-) - f_y(d_1^+) + M\varepsilon(s_0 - s_1)}{2M}.$$

Note that x^{**} is always such $x^{**} \geq 0$ because $\varepsilon \geq \varepsilon_1$. The y-value of the intersection is

$$y^{**} = \frac{f_y(d_1^-) + f_y(d_1^+) - M\varepsilon(s_0 - s_1)}{2M}.$$

and this is positive as long as $\varepsilon \leq \varepsilon_2$ where $\varepsilon_2 = \frac{|f_y(d_1^+)+f_y(d_1^-)|}{M(s_0-s_1)}$. Note also that $\varepsilon_1 < \varepsilon_2$.

For $\varepsilon_1 \leq \varepsilon \leq \varepsilon_2$, the minimum area is

$$\begin{aligned} \underline{A}(\varepsilon) &= \int_0^{\varepsilon(s_0-s_1)} \underline{f}_n(v)\ dv \\ &= (-1/4M)\left[M^2\varepsilon^2 s_0^2 - 2M^2\varepsilon^2 s_0 s_1 + M^2\varepsilon^2 s_1^2 - 2M\varepsilon f_y(d_1^-)s_0 \right. \\ &\left. +2M\varepsilon f_y(d_1^-)s_1 - 2M\varepsilon f_y(d_1^+)s_0 + 2M\varepsilon f_y(d_1^+)s_1 - f_y(d_1^-)^2 + 2f_y(d_1^-)f_y(d_1^+) - f_y(d_1^+)^2\right] \end{aligned}$$

The function $\underline{A}(\varepsilon)$ is strictly increasing with respect to ε over $\varepsilon_1 \leq \varepsilon < \varepsilon_2$. In fact, the derivative is $((s_0 - s_1) * (f_y(d_1^-) + f_y(d_1^+) - M\varepsilon(s_0 - s_1)))/2$ which is strictly positive once we take into account $\varepsilon < \varepsilon_2$. The function $\underline{A}(\varepsilon)$ is constant with respect to ε over $\varepsilon \geq \varepsilon_2$.

Therefore, we have characterized the maximum and minimum areas $\underline{A}(\varepsilon)$ and $\overline{A}(\varepsilon)$ for any given ε. These areas are undefined if $\varepsilon < \varepsilon_1$, they are equal if $\varepsilon = \varepsilon_1$, they are strictly increasing wrt ε and $\underline{A}(\varepsilon) \leq \overline{A}(\varepsilon)$ for $\varepsilon \in (\varepsilon_1, \varepsilon_2)$. For $\varepsilon \geq \varepsilon_2$, $\overline{A}(\varepsilon)$ continues to grow wrt ε but $\underline{A}(\varepsilon)$ stays constant at $\underline{A}(\varepsilon_2)$. The expression for $\underline{A}(\varepsilon_2)$ is $(f_y(d_1^-)^2 + f_y(d_1^+)^2)/2M$.
Finally, we define the partially identified set. Call the probability of bunching $p = \mathbb{P}[y = d_1]$.

Case I: If $p < \underline{A}(\varepsilon_1) = \overline{A}(\varepsilon_1)$, there does not exist any function f_n consistent with any elasticity ε, so the set is empty. The expression for $\underline{A}(\varepsilon_1) = \overline{A}(\varepsilon_1)$ is
$(|f_y(d_1^-) - f_y(d_1^+)|(f_y(d_1^-) + f_y(d_1^+)))/(2M)$.

Case II: Suppose $p \geq \underline{A}(\varepsilon_1)$ and $p < \underline{A}(\varepsilon_2)$. There is an interval range for ε such that for any ε in this interval there exists a function f_n whose integral equals p. The minimum possible elasticity solves $\overline{A}(\underline{\varepsilon}) = p$. That gives

$$\underline{\varepsilon} = \frac{2\left[f(d_1^+)^2/2 + f(d_1^-)^2/2 + M\ \mathbb{P}[y = d_1]\right]^{1/2} - \left(f(d_1^+) + f(d_1^-)\right)}{M(s_0 - s_1)}.$$

The maximum possible elasticity solves $\underline{A}(\overline{\varepsilon}) = p$. That gives

$$\overline{\varepsilon} = \frac{-2\left[f(d_1^+)^2/2 + f(d_1^-)^2/2 - M\ \mathbb{P}[y = d_1]\right]^{1/2} + \left(f(d_1^+) + f(d_1^-)\right)}{M(s_0 - s_1)}$$

Case III: Suppose $p \geq \underline{A}(\varepsilon_2)$. It is still possible to find a minimum elasticity that solves $\overline{A}(\underline{\varepsilon}) = p$. However, for any elasticity $\varepsilon \geq \underline{\varepsilon}$ we have $\underline{A}(\varepsilon) \leq p$, so $\overline{\varepsilon}$ is infinity.

D Multi-step estimation

The model proposed in equation (2) can also be estimated using a multi-step procedure similar to Heckman (1976). Starting with the value of labor supplied in logs we can write the general model as

$$y_i = \begin{cases} \varepsilon \ln(1-t_0) + x_i'\beta + \eta_i & \nu_i < d_1 - \varepsilon \ln(1-t_0) - x_i'\beta - z_i'\gamma \\ d_1 & \nu_i = [d_1 - \varepsilon \ln(1-t_0) - x_i'\beta - z_i'\gamma, d_1 - \varepsilon \ln(1-t_1) - x_i'\beta - z_i'\gamma] \\ \varepsilon \ln(1-t_1) + x_i'\beta + \eta_i & \nu_i > d_1 - \varepsilon \ln(1-t_1) - x_i'\beta - z_i'\gamma \end{cases} \tag{18}$$

We could derive this model from equation (4) by taking two steps. First, we would assume that the latent variable process which determines log labor supply l_i is given by

$$n_i^* = x_i'\beta + z_i'\gamma + \nu_i \tag{19}$$

Second, we would assume that $z_i'\gamma$ affects the probability of supplying labor at the kink but does affect the level of earnings. We can motivate this assumption by treating the labor market in our model as being perfectly competitive so that $z_i'\gamma$ does not affect the worker's marginal productivity. Because of this, the workers wage will not include compensation or penalty for those factors in their wages. As such, the log of earnings can be written as

$$\begin{aligned} y_i &= w_i + l_i \\ &= w_i + \varepsilon \ln(1-t_0) + n_i^* \\ &= w_i + \varepsilon \ln(1-t_0) + x_i'\beta + z_i'\gamma + \nu_i \\ &= -z_i'\gamma + \zeta_i + \varepsilon \ln(1-t_0) + x_i'\beta + z_i'\gamma + \nu_i \\ &= \zeta_i + \varepsilon \ln(1-t_0) + x_i'\beta + \nu_i \\ &= \varepsilon \ln(1-t_0) + x_i'\beta + \nu_i + \zeta_i \\ &= \varepsilon \ln(1-t_0) + x_i'\beta + \eta_i \end{aligned}$$

in which $w_i = -z_i'\gamma + \zeta_i$ and $\eta_i \equiv \nu_i + \zeta_i$. Next we make the simplifying assumption that

$$\begin{bmatrix} \eta_i \\ \nu_i \end{bmatrix} \sim N\left(\begin{matrix} 0 \\ 0 \end{matrix}, \begin{matrix} 1 & \sigma_\nu\rho \\ \sigma_\nu\rho & \sigma_\nu \end{matrix} \right) \tag{20}$$

So that the model is a type 2 Tobit model with middle censoring. Using the properties of a truncated normal distribution allows us to write the conditional expectations for this model as

$$E[y_i \mid y_i < d_1, x_i', z_i', t_0, t_1] = \varepsilon \ln(1-t_0) + x_i'\beta - \rho\lambda_i(d_1, x_i', z_i', t_0)$$

$$E[y_i \mid y_i > d_1, x_i', z_i', t_0, t_1] = \varepsilon \ln(1-t_1) + x_i'\beta + \rho\lambda_i(d_1, x_i', z_i', t_1)$$

in which the inverse mills ratio using the lower marginal rate is

$$\lambda_i\left(d_1, x_i', z_i', t_0\right) = \frac{\phi_\nu\left(\dfrac{d_1 - \varepsilon \ln\left(1 - t_0\right) - x_i'\beta - z_i'\gamma}{\sigma_\nu}\right)}{\Phi_\nu\left(\dfrac{d_1 - \varepsilon \ln\left(1 - t_0\right) - x_i'\beta - z_i'\gamma}{\sigma_\nu}\right)} \tag{21}$$

and the inverse mills ratio using the upper marginal rate, $\lambda_i\left(d_1, x_i', z_i', t_0\right)$, is symmetrically defined by repalcing t_0 with t_1. After deriving these defintions, we can propose a multi-step estimator. In doing so, it will be useful to partition $\beta' = \left(\beta_0, \beta_1\right)'$ in which β_0 is the scalar coefficient on a constant term in x_i' and β_1 is a $K \times 1$ vector of other coefficient. Define $x_i' = \left(x_{i0}', x_{i1}'\right)$ similarly. For simplicity, we also use population formulas instead of estimates.

- Step 1: Recover a linear combination of the elasticity and the overall constant term: $\alpha_0 = \varepsilon \ln\left(1 - t_0\right) + \beta_0$.
  - Step 1. a.) Estimate a probit model in which you define an indicator for observing income below the kink d_1

$$\begin{aligned} E\left[1\left(y_i < d_1\right) \mid x_i', z_i', t_0\right] &= P\left[y_i < d_1 \mid x_i', z_i', t_0\right] \\ &= P\left[n_i^* < \bar{n}_i \mid x_i', z_i', t_0\right] \\ &= P\left[\nu_i < d_1 - \varepsilon \ln\left(1 - t_0\right) - x_i'\beta - z_i'\gamma \mid x_i', z_i', t_0\right] \\ &= \Phi\left[\frac{d_1 - \alpha_0 - x_{i1}'\beta_1 - z_i'\gamma}{\sigma_\nu}\right] \end{aligned}$$

    By including a constant term in x_i', it will be impossible to recover the elasticity directly from this regression. We will denote the constant term from this probit model as $\alpha_0 = \varepsilon \ln\left(1 - t_0\right) + \beta_0$. With the estimate of these parameters from the probit, form the inverse mills ratio

$$\lambda_i\left(d_1, x_i', z_i', \alpha_0\right) = \frac{\phi_\nu\left(\dfrac{d_1 - \alpha_0 - x_{i1}'\beta_1 - z_i'\gamma}{\sigma_\nu}\right)}{\Phi_\nu\left(\dfrac{d_1 - \alpha_0 - x_{i1}'\beta_1 - z_i'\gamma}{\sigma_\nu}\right)} \tag{22}$$

  - Step 1. b.) Estimate a linear regression using income levels below the kink d_1 and include the inverse mills ratio from 1. a.) as a covariate

$$E\left[y_i \mid y_i < d_1, x_i', z_i'\right] = \alpha_0 + x_{i1}'\beta_1 + \rho\lambda_i\left(d_1, x_i', z_i', \alpha_0\right)$$

  Store the constant term from this regression α_0.

- Step 2: Recover a linear combination of the elasticity and the overall constant term: $\alpha_1 = \varepsilon \ln\left(1 - t_1\right) + \beta_0$.

– Step 2. a.) Estimate a probit model in which you define an indicator for observing income above the kink d_1

$$E\left[1\left(y_i > d_1\right) \mid x_i', z_i'\right] \;=\; \Phi\left[\frac{d_1 - \alpha_1 - x_{i1}'\beta_1 - z_i'\gamma}{\sigma_\nu}\right]$$

We will denote the constant term from this probit model as $\alpha_1 = \varepsilon \ln\left(1 - t_1\right) + \beta_0$. With the estimate of these parameters in hand, form the inverse mills ratio

$$\lambda_i\left(d_1, x_i', z_i', \alpha_1\right) = \frac{\phi_\nu\left(\frac{d_1 - \alpha_1 - x_{i1}'\beta_1 - z_i'\gamma}{\sigma_\nu}\right)}{\Phi_\nu\left(\frac{d_1 - \alpha_1 - x_{i1}'\beta_1 - z_i'\gamma}{\sigma_\nu}\right)} \tag{23}$$

– Step 2. b.) Estimate a linear regression using income levels above the kink d_1 and include the inverse mills ratio from 2. a.) as a covariate

$$E\left[y_i \mid y_i > d_1, x_i', z_i'\right] \;=\; \alpha_1 + x_{i1}'\beta_1 + \rho\lambda_i\left(d_1, x_i', z_i', \alpha_1\right)$$

Store the constant term from this regression α_1.

- Step 3: Recover the elasticity using the constant terms estimated 1. b.) and 2. b.) as

$$\varepsilon \;=\; \frac{\alpha_1 - \alpha_0}{\ln\left(1 - t_1\right) - \ln\left(1 - t_0\right)}$$

Standard errors for ε can be easily recovered by using the fact that both α_1 and α_0 are asymptotically normally distributed.

E EM ordered approach

Why did you finally make progress? It is because you realized that the best way to write the problem was as treating the latent variable as the two index binary variable for if the ability is within a specific case. When you define the latent variable that way, you have a discrete mixture model.

E.1 The statistical model

The selection equation is given by

$$n_i^* = x_i'\beta + w_i'\gamma + \nu_i \tag{24}$$

in which n_i^* is a latent variable for each of N individuals, x_i is a $K \times 1$ vector of covariates that can affect the unobserved selection variable n_i^* and the outcome variable y_i while w_i is a $L \times 1$ vector of covariates that are excluded from the outcome equation. The outcome equation takes different values y_i based on on the value of the selection variable n_i^* acccording to

$$y_i = \begin{cases} e\ln(1-t_0) + x_i'\alpha + \zeta_i + \xi_i & n_i^* < k - e\ln(1-t_0) \\ \mu + \delta k + \xi_i & n_i^* \in [k - e\ln(1-t_1), k - e\ln(1-t_0)] \\ e\ln(1-t_1) + x_i'\alpha + \zeta_i + \xi_i & n_i^* > k - e\ln(1-t_1) \end{cases} \quad (25)$$

in which ξ_i is the optimizing friction/classical measurement error. Equations (1) and (2) define the model. It is straightfoward to see that this model nests Saez (2010), page 186 just above equation (3), when $\nu_i = \zeta_i$ and $\alpha = \beta = \gamma = \mu = \xi_i = 0$ and $\delta = 1$.

The likelihood for this problem is given by

$$\begin{aligned} L(\theta \mid x, w, y) &= \prod_{i=1}^{N} P[y_i \mid t_0, t_1, x_i, w_i, k, \theta] \\ &= \prod_{i=1}^{N} \int P[y_i, n_i^* \mid t_0, t_1, x_i, w_i, k, \theta] P[n_i^* \mid t_0, t_1, x_i, w_i, k, \theta] \, dn_i^* \end{aligned}$$

It will be helpful to transform the latent variable into cases as

$$q_{ij} = \begin{cases} 1[n_i^* < k - e\ln(1-t_0)] & j = 1 \\ 1[k - e\ln(1-t_0) \le n_i^* \le k - e\ln(1-t_1)] & j = 2 \\ 1[n_i^* > k - e\ln(1-t_1)] & j = 3 \end{cases} \quad (26)$$

which allows us to write (2) as

$$y_i = \begin{cases} e\ln(1-t_0) + x_i'\alpha + \zeta_i + \xi_i & q_{i1} = 1 \\ \mu + \delta k + \xi_i & q_{i2} = 1 \\ e\ln(1-t_1) + x_i'\alpha + \zeta_i + \xi_i & q_{i3} = 1 \end{cases} \quad (27)$$

The likelihood for this problem is given by

$$\begin{aligned} L(\theta \mid x, w, y) &= \prod_{i=1}^{N} \int P[y_i, n_i^* \mid t_0, t_1, x_i, w_i, k, \theta] P[n_i^* \mid t_0, t_1, x_i, w_i, k, \theta] \, dn_i^* \\ &= \prod_{i=1}^{N} \sum_{j=1}^{J} P[y_i \mid t_0, t_1, x_i, w_i, k, q_{ij} = 1] P[q_{ij} = 1 \mid t_0, t_1, x_i, w_i, k] \end{aligned}$$

Maximizing the product of a sum is very difficult so we will use the EM algorithm.

E.2 Parametric assumptions

For now assume that the errors are uncorrelated and normally distributued so that $\nu_i \sim N(0, \sigma_\nu^2)$, $\zeta_i \sim N\left(0, \sigma_\zeta^2\right)$, and $\xi_i \sim N\left(0, \sigma_\xi^2\right)$. Using this fact we know that

$$P[y_i \mid t_0, t_1, x_i, w_i, k, q_{i1} = 1] = \left(\sqrt{\sigma_\nu^2 + \sigma_\zeta^2}\right)^{-1} \phi\left(\frac{y_i - e\ln(1 - t_0) - x_i'\alpha}{\sqrt{\sigma_\nu^2 + \sigma_\zeta^2}}\right)$$

$$P[y_i \mid t_0, t_1, x_i, w_i, k, q_{i2} = 1] = \sigma_\xi^{-1}\phi\left(\frac{y_i - \mu - \delta k}{\sigma_\xi}\right)$$

$$P[y_i \mid t_0, t_1, x_i, w_i, k, q_{i3} = 1] = \left(\sqrt{\sigma_\nu^2 + \sigma_\zeta^2}\right)^{-1} \phi\left(\frac{y_i - e\ln(1 - t_1) - x_i'\alpha}{\sqrt{\sigma_\nu^2 + \sigma_\zeta^2}}\right)$$

Using the normality assumption and latent variables we defined above, we know that

$$\begin{aligned} P[q_{i1} = 1 \mid t_0, t_1, x_i, w_i, k] &= P[n_i^* < k - e\ln(1 - t_0) \mid t_0, t_1, x_i, w_i, k] \\ &= \Phi\left(\frac{e\ln(1 - t_0) - x_i'\beta - w_i'\gamma}{\sigma_\nu}\right) \end{aligned}$$

$$\begin{aligned} P[q_{i2} = 1 \mid t_0, t_1, x_i, w_i, k] &= P[k - e\ln(1 - t_0) \leq n_i^* \leq k - e\ln(1 - t_1) \mid t_0, t_1, x_i, w_i, k] \\ &= \Phi\left(\frac{e\ln(1 - t_1) - x_i'\beta - w_i'\gamma}{\sigma_\nu}\right) - \Phi\left(\frac{e\ln(1 - t_0) - x_i'\beta - w_i'\gamma}{\sigma_\nu}\right) \end{aligned}$$

$$\begin{aligned} P[q_{i3} = 1 \mid t_0, t_1, x_i, w_i, k] &= P[n_i^* > k - e\ln(1 - t_1) \mid t_0, t_1, x_i, w_i, k] \\ &= 1 - \Phi\left(\frac{e\ln(1 - t_1) - x_i'\beta - w_i'\gamma}{\sigma_\nu}\right) \end{aligned}$$

These assumptions are not so bad that we couldn't maximize the likelihood directly but it will be much easier using the EM algoritm.

E.3 EM approach

Define a latent variable unobserved $N \times J$ matrix p

$$p^* = \begin{bmatrix} p_{11}^* & p_{12}^* & \dots & p_{1J}^* \\ p_{21}^* & p_{22}^* & \dots & \\ \vdots & \dots & \ddots & \vdots \\ p_{N1}^* & p_{N2}^* & \dots & p_{NJ}^* \end{bmatrix}$$

in which

$$p_{ij}^* = \begin{cases} 1, & y_i \; from \; P\left[y_i \mid t_0, t_1, x_i, w_i, k, q_{ij} = 1\right] P\left[q_{ij} = 1 \mid t_0, t_1, x_i, w_i, k\right] \\ 0, & otherwise \end{cases}$$

Then the likelihood becomes

$$\begin{aligned} L\left(\theta \mid x, w, y\right) &= \prod_{i=1}^{N} \sum_{k=1}^{K} P\left[y_i \mid t_0, t_1, x_i, w_i, k, q_{ij} = 1\right] P\left[q_{ij} = 1 \mid t_0, t_1, x_i, w_i, k\right] \\ &= \prod_{i=1}^{N} \prod_{j=1}^{J} \left(P\left[y_i \mid t_0, t_1, x_i, w_i, k, q_{ij} = 1\right] P\left[q_{ij} = 1 \mid t_0, t_1, x_i, w_i, k\right]\right)^{p_{ij}^*} \\ &= \sum_{i=1}^{N} \sum_{j=1}^{J} p_{ij}^* \ln\left(P\left[y_i \mid t_0, t_1, x_i, w_i, k, q_{ij} = 1\right] P\left[q_{ij} = 1 \mid t_0, t_1, x_i, w_i, k\right]\right) \end{aligned}$$

which

$$\begin{aligned} l\left(\theta \mid x, w, y\right) &= \sum_{i=1}^{N} \sum_{j=1}^{J} p_{ij}^* \ln\left(P\left[y_i \mid t_0, t_1, x_i, w_i, k, q_{ij} = 1\right] P\left[q_{ij} = 1 \mid t_0, t_1, x_i, w_i, k\right]\right) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{J} p_{ij}^* \ln\left(P\left[y_i \mid t_0, t_1, x_i, w_i, k, q_{ij} = 1\right]\right) \\ &+ \sum_{i=1}^{N} \sum_{j=1}^{J} p_{ij}^* \ln\left(P\left[q_{ij} = 1 \mid t_0, t_1, x_i, w_i, k\right]\right) \end{aligned}$$

E.4 Writing out the terms explicitly

$$\sum_{i=1}^{N}\sum_{j=1}^{J} p_{ij}^{*}\ln\left(P\left[y_i \mid t_0,t_1,x_i,w_i,k,q_{ij}=1\right]\right) = \sum_{i=1}^{N} p_{i1}^{*}\ln\left[\left(\sqrt{\sigma_\xi^2+\sigma_\zeta^2}\right)^{-1}\phi\left(\frac{y_i-e\ln\left(1-t_0\right)-x_i'\alpha}{\sqrt{\sigma_\xi^2+\sigma_\zeta^2}}\right)\right] + \sum_{i=1}^{N} p_{i2}^{*}\ln\left[\sigma_\xi^{-1}\phi\left(\frac{y_i-\mu-\delta k}{\sigma_\xi}\right)\right] + \sum_{i=1}^{N} p_{i3}^{*}\ln\left[\left(\sqrt{\sigma_\xi^2+\sigma_\zeta^2}\right)^{-1}\phi\left(\frac{y_i-e\ln\left(1-t_1\right)-x_i'\alpha}{\sqrt{\sigma_\xi^2+\sigma_\zeta^2}}\right)\right]$$

$$\sum_{i=1}^{N}\sum_{j=1}^{J} p_{ij}^{*}\ln\left(P\left[q_{ij}=1 \mid t_0,t_1,x_i,w_i,k\right]\right) = \sum_{i=1}^{N} p_{i1}^{*}\ln\left[\Phi\left(\frac{e\ln\left(1-t_0\right)-x_i'\beta-w_i'\gamma}{\sigma_\nu}\right)\right] + \sum_{i=1}^{N} p_{i2}^{*}\ln\left[\Phi\left(\frac{e\ln\left(1-t_1\right)-x_i'\beta-w_i'\gamma}{\sigma_\nu}\right)-\Phi\left(\frac{e\ln\left(1-t_0\right)-}{\sigma_\nu}\right.\right. + \sum_{i=1}^{N} p_{i3}^{*}\ln\left[\Phi\left(\frac{e\ln\left(1-t_1\right)-x_i'\beta-w_i'\gamma}{\sigma_\nu}\right)\right]$$

I guess you don't know for sure if this is the correct E step but I think going with it is fine for now.

$$E\left[p_{ij}^{*(m)} \mid \cdot\right] = \frac{P\left[y_i \mid t_0,t_1,x_i,w_i,k,q_{ij}=1,\theta^{(m)}\right]P\left[q_{ij}=1 \mid t_0,t_1,x_i,w_i,k,\theta^{(m)}\right]}{\sum_{j=1}^{J}P\left[y_i \mid t_0,t_1,x_i,w_i,k,q_{ij}=1,\theta^{(m)}\right]P\left[q_{ij}=1 \mid t_0,t_1,x_i,w_i,k,\theta^{(m)}\right]}$$

F Alternative Methods

General steps (which I will then elaborate on).

PDF and CDF

Everything in this section will be discretized into bins. Therefore a PDF (e.g., $g(x)$) is just the number of observations (e.g., people or corporations) at x (e.g., income) which is in bin $i = x/w$. A CDF is the integral (in discrete talk the sum) of people over some range.

To find the difference in CDFs between two points (e.g., $G(z^*+\Delta)-G(z^*)$) it is common for some of these alternative methods to take the average value of the PDF in the range (e.g., $\bar{g(x)} = (1/((z^*+\Delta)-z^*))\sum_{z^*}^{(z^*+\Delta)} g(x)$) and multiply it by the range (e.g., $(z^*+\Delta)-z^*$). Geometrically, this is calculating the difference in the CDFs as the area of the rectangle with base equal to the range $(z^*+\Delta)-z^*$ and the height equal to the average PDF level $\bar{g(x)}$. Together this looks like $G(z^*+\Delta)-G(z^*) = \bar{g}(\delta)\Delta z$

DEFINITIONS

1) w is the size of each bin (for example \$50).
2) M_i : Number of observations in bin i

$$M_i = \sum_{j=i}^{i+w} 1((i-1)w \leq z_i < wi) \tag{28}$$

3) Define the ‘‘number’’ of people in the bunching region $(z^* - \delta, z^* + \delta)$, noting, I am not scaling here.

$$h(\delta) = \sum_{i=(z^*-\delta)/w}^{(z^*+\delta)/w} 1(z^* - \delta \leq z_i < z^* + \delta) \tag{29}$$

F.1 Chetty et al. (2011)

This section presents the model and estimation method of Chetty et al. (2011).
1) Estimate the counterfactual distribution

The counterfactual distribution (this is important to find B, and Δz). The key difference between methods is how the counterfactual distribution is estimated.

i) Estimate the distribution from the left and to the right of the bunching region, using T which is a matrix with columns $(1, t, t^2, t^3, t^4, t^5, t^6, t^7)$, where t is income (or any running variable). The following regression regresses the number of observations in bin i, defined as M_i, with income between $(i-1)w$ and iw, on a polynomial (order 7 typically and in this example) of income levels.

$$M_i = T\beta_L + \varepsilon \quad \text{if} \quad i < (z^* - \delta)/w \tag{30}$$

$$M_i = T\beta_R + \varepsilon \quad \text{if} \quad i > (z^* + \delta)/w \tag{31}$$

ii) Use the estimates $\hat{\beta_L}$ and $\hat{\beta_R}$ to ‘‘estimate’’ the counterfactual distribution in the bunching region $(z^* - \delta, z^* + \delta)$,

$$g(\delta) = (T|_{(z^*-\delta)/w \leq i < z^*/w}\hat{\beta_L} + T|_{z^*/w \leq i < (z^*+\delta)/w}\hat{\beta_R})/(2\delta) \tag{32}$$

$g(\delta)$ is a number (e.g., 10 observations). This estimation adds the expected value of the distribution from the left and the right from the regressions above and takes the average. So, this is just a number, which is the average predicted value from the out-of-sample prediction. This is particularly problematic because the regression in equation (31) is based on an observed distribution that is affected by the kink (which is being used to estimate the counterfactual distribution in the absence of the kink).

iii) Typically, the distribution to the right of the kink is assumed to be the same as under the bunching region.

$$g(M_i)|_{z > z^* + \delta} = g(\delta) \tag{33}$$

This is of course a bad assumption. The issue is it is hard to estimate this distribution because the observed distribution to the right of the kink is *not* the distribution in the absence of the kink. Basically this just says we expect the number of people in each bin to the right of the kink to be $g(\delta)$. Maybe a better notation would be $\hat{M}$.

2) Find B

Take the difference between the "number" of people in the bunching region and the counterfactual number of people in the bunching region.

$$B = h(\delta) - g(\delta)2\delta \tag{34}$$

3) Find ΔZ

$$\int_{z^*}^{z^*+\Delta Z} g(z)dz = B \tag{35}$$

which can be rewritten as

$$G(z^* + \Delta Z) - G(z^*) = B \tag{36}$$

With the assumption that the distribution to the right of the kink is just $g(\delta)$, this can be rewritten as, (for a longer discussion please see above under the section PDF and CDF),

$$g(\delta)\Delta Z = B \tag{37}$$

which gives,

$$\Delta Z = B/g(\delta) \tag{38}$$

4) Now calculate elasticity of income with respect to the net of tax rate $\theta = 1 - t$,

$$\varepsilon_{Z,\theta} = \frac{log(\Delta z/z^* + 1)}{log(\frac{1-t_0}{1-t_1})} = \frac{\Delta z/z^*}{log(\frac{1-t_0}{1-t_1})} = \frac{B/g(\delta)}{z^* log(\frac{1-t_0}{1-t_1})} \tag{39}$$

This last equation is exactly the equation in Chetty et al. (2011), Weber (2012), and Devereux et al. (2014).

F.2 Patel et al. (2016)

Here, the method from Patel et al. (2016) has the same basic steps with one crucial difference, the counterfactual distribution is estimated from a control group. Here we assume the distribution of the latent potential income, n, is the same for all "people," but because different groups face different kink points, their observed distribution of income is different.

For simplicity, assume two groups $k = t, c$ where $z_t^*(1-t_1)^{-e} < z_c^*(1-t_0)^{-e}$, which ensures that individuals it latent potential income $n \in [z_t^*/(1-t_0)^e, z_t^*/(1-t_1)^e]$ bunch if they are in the treatment group and do not bunch if they are in the control group.

$$z = \begin{cases} n(1-t_0)^e, & n < z_k^*/(1-t_0)^e \\ z_k^*, & n \in [z_k^*/(1-t_0)^e, z_k^*/(1-t_1)^e] \\ n(1-t_0)^e, & n > z_k^*/(1-t_1)^e \end{cases}$$

1) Estimate the counterfactual distribution.

i) Here we estimate the distribution, using a polynomial captured by T which is a matrix with columns $(1, t, t^2, t^3, t^4, t^5, t^6, t^7)$, using only the control group's observations

$$M_i = T\beta + \varepsilon \quad \text{if} \quad k = c \ \ \text{and} i < (z_c^* - \delta)/w \tag{40}$$

ii) Use the estimate of β to calculate the counterfactual distribution,

$$g(M_i) = T\hat{\beta}. \tag{41}$$

$$g(\delta) = \sum_{i=(z_t^*-\delta)/w}^{(z_t^*+\delta)/w} g(M_i) \tag{42}$$

This estiamted counterfactual distribution is used to calculate bunching, B, and Δz.

2) Find B

Take the difference between the "number" of people in the bunching region and the counterfactual number of people in the bunching region.

$$B = h(\delta) - g(\delta) \tag{43}$$

3) Find ΔZ

$$\int_{z^*}^{z^*+\Delta Z} g(z)dz = B \tag{44}$$

which can be rewritten as

$$G(z^* + \Delta Z) - G(z^*) = B \tag{45}$$

This is slightly more complicated in the PSS method because we use the estimated distribution to the right of the kink (not just a number).

$$min_{\Delta z} \quad \left(\sum_{i=z_t^*/w}^{(z_t^*+\Delta z)/w} g(M_i) - B \right)^2 \quad \text{s.t.} \quad \sum_{i=z_t^*/w}^{(z_t^*+\Delta z)/w} g(M_i) - B > 0 \tag{46}$$

4) Now calculate elasticity of income with respect to the net of tax rate $\theta = 1 - t$,

$$\varepsilon_{z,\theta} = \frac{log(\Delta z/z^* + 1)}{log(\frac{1-t_0}{1-t_1})} = \frac{\Delta z/z^*}{log(\frac{1-t_0}{1-t_1})} \tag{47}$$

F.3 Andrew math

F.4 Identification in the notching case

F.4.1 Relating to change in income in Saez

The first thing to do is to relate the change in the highest and lowest latent variables to the elasticity. The log percent difference in this model is

$$\begin{aligned}\overline{n}-\underline{n} &= d_1-\varepsilon s_1-(d_1-\varepsilon s_0)\\ &= d_1-\varepsilon s_1-d_1+\varepsilon s_0\\ &= \varepsilon\left(s_0-s_1\right)\\ &= \varepsilon\left(\ln\left(1-t_0\right)-\ln\left(1-t_1\right)\right)\end{aligned}$$

The percent difference in this model is

$$\begin{aligned}\frac{\overline{N}-\underline{N}}{\underline{N}} &= \frac{D_1/S_1^{\varepsilon}-D_1/S_0^{\varepsilon}}{D_1/S_0^{\varepsilon}}\\ &= \frac{D_1/S_1^{\varepsilon}}{D_1/S_0^{\varepsilon}}-1\\ &= \frac{S_0^{\varepsilon}}{S_1^{\varepsilon}}-1\\ &= \left(\frac{1-t_0}{1-t_1}\right)^{\varepsilon}-1\end{aligned}$$

In the Saez model the definition of ΔD_1 relies on using the upper and lower latent variables that give the same reported income. This means that we use $D_1^-=NS_0^{\varepsilon}$ so that $D_1^-/S_0^{\varepsilon}=N$ and similarly $D_1^+=NS_1^{\varepsilon}$ so that $D_1^+/S_1^{\varepsilon}=N$. The definition of $\Delta D_1=D_1^+-D_1^-=NS_1^{\varepsilon}-NS_0^{\varepsilon}$ and

$$\begin{aligned}\frac{\Delta D_1}{D_1^-} &= \frac{D_1^+-D_1^-}{D_1^-}\\ &= \frac{NS_1^{\varepsilon}-NS_0^{\varepsilon}}{NS_0^{\varepsilon}}\\ &= \frac{NS_1^{\varepsilon}-NS_0^{\varepsilon}}{NS_0^{\varepsilon}}\end{aligned}$$

$$\begin{aligned}
\frac{Y-Y}{D_1} &= \frac{NS_1^{\varepsilon}-NS_0^{\varepsilon}}{D_1} \\
&= \frac{D_1/S_1^{\varepsilon}-D_1/S_0^{\varepsilon}}{D_1/S_0^{\varepsilon}} \\
&= \frac{D_1/S_1^{\varepsilon}}{D_1/S_0^{\varepsilon}} \\
&= \frac{S_0^{\varepsilon}}{S_1^{\varepsilon}} \\
&= \frac{(1-t_0)^{\varepsilon}}{(1-t_1)^{\varepsilon}} \\
&= \left(\frac{1-t_0}{1-t_1}\right)^{\varepsilon}
\end{aligned}$$

The claim is that we can identify the elasticity in the notching case directly from the data by looking at $D_j+\Delta D_j$.

F.5 Bunching delta

We want to define a term similar to Δz^* in Saez. This is done by considering the highest ability person that reports D_1 in our model. That person has ability $\bar{N}=D_1/\left(1-t_1\right)^{\varepsilon}$. Remember that utility maximization with his utility function under a linear tax implied the person reports $\bar{Y}=\bar{N}\left(1-t\right)^{\varepsilon}$. Hence, we can construct a counterfactual reported income that would obtain if the tax rate, t_0, below the kink extended past the discrete income level D_1

$$\begin{aligned}
\bar{Y} &= \bar{N}\left(1-t_0\right)^{\varepsilon} \\
\bar{Y} &= D_1\left(\frac{1-t_0}{1-t_1}\right)^{\varepsilon}
\end{aligned}$$

Similarly, the lowest ability person that reports $z=z^*$ has ability $N=D_1/\left(1-t_0\right)^{\varepsilon}$. If they had faced the linear tax of t_0, they would report

$$\begin{aligned}
Y &= N\left(1-t_0\right)^{\varepsilon} \\
Y &= \frac{D_1\left(1-t_0\right)^{\varepsilon}}{\left(1-t_0\right)^{\varepsilon}} \\
&= D_1
\end{aligned}$$

Combining these two expression defines

$$\begin{aligned}
\Delta D_1 &\equiv \bar{Y}-Y \\
&= D_1\left(\frac{1-t_0}{1-t_1}\right)^{\varepsilon}-D_1
\end{aligned}$$

Notice that both $\bar{Y}$ and Y are counterfactual levels of reported income that will never be observed even if all the assumptions of the model are correct. Saez writes counterfactual

ΔD_1 as equation (3)

$$\frac{\Delta D_1}{D_1} = \left(\frac{1-t_0}{1-t_1}\right)^{\varepsilon} - 1$$

which we can show is equivalent to

$$\frac{\Delta D_1}{D_1} = \left(\frac{1-t_0}{1-t_1}\right)^{\varepsilon} - 1 = \frac{\bar{N}}{N} - 1$$

F.6 Notching delta

We want to define a term similar to Δz^* in Saez but for the notching case. This is done by considering the highest ability person that reports D_1 in the notching case. For simplicity use the model presented in equation (5). XXXX it must be the case that $\Delta I_1 < 0$, right? If $R = 0$ this means the person supplies negative labor when $t = 0$? Something is wrong.

That person has ability $\bar{N} = f(D_1, \varepsilon, t)$. Remember that utility maximization with his utility function without a notch the person would report $\bar{Y} = \frac{1}{\bar{N}}(R + \Delta I_1)$ with $t = 0$. Why do higher ability people report lower income?

F.7 Andrew's math

Introducing a kink in the tax rate changes the problem into

$$\begin{aligned} \max_{c,z} \quad & c - \frac{n}{1+1/e}\left(\frac{z}{n}\right)^{1+1/e} \\ s.t. \quad & \\ c \;=\; & z(1-t_0)\,1(z \leq z^*) + z(1-t_1)\,1(z > z^*) + R \end{aligned}$$

in which c is consumption, z is earnings, and R is non-earning resources that can be spent on consumption. An unobservable ability variable, n, is distributed according to some PDF $f(n)$ and some CDF $F(n)$ and affects the level of optimal consumption for each consumer. Notice we need to have $z \in (0, \infty)$ and $n \in (0, \infty)$. The Lagrangian is

$$\begin{aligned} \mathcal{L} &= c - \frac{n}{1+1/e}\left(\frac{z}{n}\right)^{1+1/e} + \lambda\left[z(1-t_0)\,1(z \leq z^*) + z(1-t_1)\,1(z > z^*) + R - c\right] \\ &= c - \frac{n}{1+1/e}\left(\frac{1}{n}\right)^{1+1/e} z^{1+1/e} + \lambda\left[z(1-t_0)\,1(z \leq z^*) + z(1-t_1)\,1(z > z^*) + R - c\right] \end{aligned}$$

We want to maximize this lagrangian we can differentiate with respect to c easily to get started

$$\begin{aligned} \mathcal{L}_c = 1 - \lambda &= 0 \\ 1 &= \lambda \end{aligned}$$

We have three regions, they are $z < z^*$ $z = z^*$ and $z > z^*$. We can differentiate with respect to z as long we as are in regions in which the Lagrangian is not discontinuous. These are the $z < z^*$ and $z > z^*$ regions. We can solve it by considering, cases, however. Start with the $z < z^*$ region then the $\mathcal{L}$ becomes

$$\mathcal{L} \mid z < z^* \quad = \quad c - \frac{n}{1+1/e}\left(\frac{z}{n}\right)^{1+1/e} + \lambda\left[z\left(1-t_0\right) + R - c\right]$$

$$\begin{aligned}
\mathcal{L}_z \mid z < z^* = -n\left(\frac{1}{n}\right)^{1+1/e} z^{1/e} + \lambda\left(1-t_0\right) &= 0 \\
n\left(n\right)^{-1-1/e} z^{1/e} &= \lambda\left(1-t_0\right) \\
n^{-1/e} z^{1/e} &= \lambda\left(1-t_0\right) \\
\left(\frac{z}{n}\right)^{1/e} &= \lambda\left(1-t_0\right) \\
\frac{z}{n} &= \lambda^e\left(1-t_0\right)^e \\
z &= n\lambda^e\left(1-t_0\right)^e
\end{aligned}$$

Since we know that $\lambda = 1$ the optimal value when $z < z^*$ is

$$z = n\left(1-t_0\right)^e$$

We can follow the same steps for the $\mathcal{L} \mid z > z^*$ to get that

$$z = n\left(1-t_1\right)^e$$

The solution to the case when $z = z^*$ is trivial because that's just it. This implies that the casewise optimal solution is therefore

$$z \quad = \quad \begin{cases} n\left(1-t_0\right)^e & z < z^* \\ z^* & z = z^* \\ n\left(1-t_1\right)^e & z > z^* \end{cases} \tag{48}$$

By substitution, we can derive

$$z \quad = \quad \begin{cases} n\left(1-t_0\right)^e & n\left(1-t_0\right)^e < z^* \\ z^* & z = z^* \\ n\left(1-t_1\right)^e & n\left(1-t_1\right)^e > z^* \end{cases} \tag{49}$$

Our goal is to find the person with highest ability $\bar{n}$ that optimizes constrained utility by reporting $z = z^*$. Because the agent reports $z = z^*$, we can be sure that constrained utility is maximized at that point. If the agent is indifferent between reporting $z = z^*$ and facing t_1 or reporting $z = \bar{n}(1-t_1)^e$ and facing that same tax rate t_1, given that they are optimizing at z^*, that means we only need to compare the alternate budget constraints which must provide the same resources in each case. The general budget constraint is

$$c = z(1-t_0)1(z \leq z^*) + z(1-t_1)1(z > z^*) + R$$

The agent must have the same resources from reporting either case so this means we compare

$$\begin{aligned} c(z = z^*, \bar{n} \mid t_1) &= c\left(z = \bar{n}(1-t_1)^e, \bar{n} \mid t_1\right) \\ z^*(1-t_1) + R &= \bar{n}(1-t_1)^e(1-t_1) + R \\ z^*(1-t_1) &= \bar{n}(1-t_1)^e(1-t_1) \\ z^*(1-t_1) &= \bar{n}(1-t_1)^{1+e} \\ z^* &= \bar{n}(1-t_1)^e \end{aligned}$$

This might be easier to see by consider the full Lagrangian when we use the FOC w.r.t. c to get that $\lambda = 1$

$$\begin{aligned} \mathcal{L}(z^* = z, \bar{n}, c) &= \mathcal{L}\left(z^* = \bar{n}(1-t_1)^e, \bar{n}, c\right) \\ -\frac{\bar{n}}{1+1/e}\left(\frac{z^*}{\bar{n}}\right)^{1+1/e} + z^*(1-t_1) + R &= -\frac{\bar{n}}{1+1/e}\left(\frac{z^*}{\bar{n}}\right)^{1+1/e} + \bar{n}(1-t_1)^{1+e} + R \\ z^*(1-t_1) &= \bar{n}(1-t_1)^{1+e} \\ z^* &= \bar{n}(1-t_1)^e \end{aligned}$$

we take similar steps to get the lower cutoff. This leads to the final decision rule as a function of the ability variable n.

$$z = \begin{cases} n(1-t_0)^e & n < z^*/(1-t_0)^e \\ z^* & n = \left[z^*/(1-t_0)^e, z^*/(1-t_1)^e\right] \\ n(1-t_1)^e & n > z^*/(1-t_1)^e \end{cases} \tag{50}$$

F.8 With a Notch

Introducing a notch in the tax rate changes the problem into

$$\begin{aligned} \max_{c,z} \quad & c - \frac{n}{1+1/e}\left(\frac{z}{n}\right)^{1+1/e} \\ s.t. & \\ c &= z(1-t) + \Delta T 1(z > z^*) + R \end{aligned} \tag{51}$$

in which c is consumption, z is earnings, and R is non-earning resources that can be spent on consumption. An unobservable ability variable, n, is distributed according to some

PDF $f(n)$ and some CDF $F(n)$ and affects the level of optimal consumption for each consumer. Notice we need to have $z \in (0,\infty)$ and $n \in (0,\infty)$. The Lagrangian is

$$\begin{aligned}\mathcal{L} &= c - \frac{n}{1+1/e}\left(\frac{z}{n}\right)^{1+1/e} + \lambda\left[z(1-t) + \Delta T 1(z > z^*) + R - c\right] \\ &= c - \frac{n}{1+1/e}\left(\frac{1}{n}\right)^{1+1/e} z^{1+1/e} + \lambda\left[z(1-t) + \Delta T 1(z > z^*) + R - c\right]\end{aligned}$$

We want to maximize this lagrangian we can differentiate with respect to c easily to get started

$$\begin{aligned}\mathcal{L}_c = 1 - \lambda &= 0 \\ 1 &= \lambda\end{aligned}$$

We have three regions, they are $z < z^*$ $z = z^*$ and $z > z^*$. We can differentiate with respect to z as long we as are in regions in which the Lagrangian is not discontinuous. These are the $z < z^*$ and $z > z^*$ regions. We can solve it by considering, cases, however. Start with the $z < z^*$ region then the $\mathcal{L}$ becomes

$$\mathcal{L} \mid z < z^* = c - \frac{n}{1+1/e}\left(\frac{z}{n}\right)^{1+1/e} + \lambda\left[z(1-t) + R - c\right]$$

$$\begin{aligned}\mathcal{L}_z \mid z < z^* = -n\left(\frac{1}{n}\right)^{1+1/e} z^{1/e} + \lambda(1-t) &= 0 \\ n(n)^{-1-1/e} z^{1/e} &= \lambda(1-t) \\ n^{-1/e} z^{1/e} &= \lambda(1-t) \\ \left(\frac{z}{n}\right)^{1/e} &= \lambda(1-t) \\ \frac{z}{n} &= \lambda^e (1-t)^e \\ z &= n\lambda^e (1-t)^e\end{aligned}$$

Since we know that $\lambda = 1$ the optimal value when $z < z^*$ is

$$z = n(1-t)^e$$

We can follow the same steps for the $\mathcal{L} \mid z > z^*$ and will do so carefully because this can get confusing

$$\mathcal{L} \mid z > z^* = c - \frac{n}{1+1/e}\left(\frac{z}{n}\right)^{1+1/e} + \lambda\left[z(1-t) + \Delta T + R - c\right]$$

$$\begin{aligned} \mathcal{L}_z \mid z > z^* = -n\left(\frac{1}{n}\right)^{1+1/e} z^{1/e} + \lambda(1-t) &= 0 \\ n(n)^{-1-1/e} z^{1/e} &= \lambda(1-t) \\ n^{-1/e} z^{1/e} &= \lambda(1-t) \\ \left(\frac{z}{n}\right)^{1/e} &= \lambda(1-t) \\ \frac{z}{n} &= \lambda^e (1-t)^e \\ z &= n\lambda^e (1-t)^e \end{aligned}$$

$$z = n(1-t)^e$$

The solution to the case when $z = z^*$ is trivial because that's just it. This implies that the casewise optimal solution is therefore

$$z = \begin{cases} n(1-t)^e & z < z^* \\ z^* & z = z^* \\ n(1-t)^e & z > z^* \end{cases} \tag{52}$$

Our goal is to find the person with highest ability $\bar{n}$ that optimizes constrained utility by reporting $z = z^*$. Because the agent reports $z = z^*$, we can be sure that constrained utility is maximized at that point. If the agent is indifferent between reporting $z = z^*$ and facing t plus ΔT or reporting $z = \bar{n}(1-t)^e$ and facing that same tax rate t plus ΔT, given that they are optimizing at z^*, that means we only need to compare the alternate budget constraints which must provide the same resources in each case. The general budget constraint is

$$c = z(1-t) + \Delta T 1(z > z^*) + R$$

The agent must have the same resources from reporting either case so this means we compare

$$\begin{aligned} c(z = z^*, \bar{n} \mid t, \Delta T) &= c\left(z = \bar{n}(1-t)^e, \bar{n} \mid t, \Delta T\right) \\ z^*(1-t) + \Delta T + R &= \bar{n}(1-t)^e (1-t) + \Delta T + R \\ z^*(1-t_1) &= \bar{n}(1-t_1)^e (1-t_1) \\ z^*(1-t_1) &= \bar{n}(1-t_1)^{1+e} \\ z^* &= \bar{n}(1-t_1)^e \end{aligned}$$

This might be easier to see by consider the full Lagrangian when we use the FOC w.r.t. c to get that $\lambda = 1$

$$
\begin{aligned}
\mathcal{L}\left(z^*=z,\bar{n},c\right) &= \mathcal{L}\left(z^*=\bar{n}\left(1-t_1\right)^e,\bar{n},c\right) \\
-\frac{\bar{n}}{1+1/e}\left(\frac{z^*}{\bar{n}}\right)^{1+1/e}+z^*\left(1-t_1\right)+R &= -\frac{\bar{n}}{1+1/e}\left(\frac{z^*}{\bar{n}}\right)^{1+1/e}+\bar{n}\left(1-t_1\right)^{1+e}+R \\
z^*\left(1-t_1\right) &= \bar{n}\left(1-t_1\right)^{1+e} \\
z^* &= \bar{n}\left(1-t_1\right)^e
\end{aligned}
$$

Our goal is to find the person with highest ability $\bar{n}$ that is indifferent between reporting $z=z^*$ or reporting $z=n\left(1-t\right)^e$ and facing the tax notch ΔT. $z>z^*$. We know that reporting $z=z^*$ is optimal so we know the Lagrangian is

$$
\begin{aligned}
\mathcal{L}\left(z^*=z,\bar{n},c\right) &= c-\frac{\bar{n}}{1+1/e}\left(\frac{z^*}{\bar{n}}\right)^{1+1/e}+\lambda\left[z^*\left(1-t\right)+R-c\right] \\
&= -\frac{\bar{n}}{1+1/e}\left(\frac{z^*}{\bar{n}}\right)^{1+1/e}+z^*\left(1-t\right)+R
\end{aligned}
$$

In which the last step comes from the fact that the FOC w.r.t. c ensures $\lambda=1$. This person should get the same utility from reporting $z=n\left(1-t\right)^e$

$$
\begin{aligned}
\mathcal{L}\left(z^*=\bar{n}\left(1-t\right)^e,\bar{n},c\right) &= c-\frac{\bar{n}}{1+1/e}\left(\frac{\bar{n}\left(1-t\right)^e}{\bar{n}}\right)^{1+1/e}+\lambda\left[\bar{n}\left(1-t\right)^e\left(1-t\right)+\Delta T+R-c\right] \\
&= c-\frac{\bar{n}}{1+1/e}\left(\left(1-t\right)^e\right)^{1+1/e}+\lambda\left[\bar{n}\left(1-t\right)^{1+e}+\Delta T+R-c\right] \\
&= c-\frac{\bar{n}}{1+1/e}\left(1-t\right)^{1+e}+\lambda\left[\bar{n}\left(1-t\right)^{1+e}+\Delta T+R-c\right] \\
&= -\frac{\bar{n}}{1+1/e}\left(1-t\right)^{1+e}+\bar{n}\left(1-t\right)^{1+e}+\Delta T+R
\end{aligned}
$$

Setting these equal and solving for $\bar{n}$

$$\begin{aligned}
\mathcal{L}\left(z^*=z,\bar{n},c\right) &= \mathcal{L}\left(z^*=\bar{n}\left(1-t\right)^e,\bar{n},c\right)\\
-\frac{\bar{n}}{1+1/e}\left(\frac{z^*}{\bar{n}}\right)^{1+1/e}+z^*\left(1-t\right)+R &= -\frac{\bar{n}}{1+1/e}\left(1-t\right)^{1+e}+\bar{n}\left(1-t\right)^{1+e}+\Delta T+R\\
-\frac{\bar{n}}{1+1/e}\left(\frac{z^*}{\bar{n}}\right)^{1+1/e}+z^*\left(1-t\right) &= -\frac{\bar{n}}{1+1/e}\left(1-t\right)^{1+e}+\bar{n}\left(1-t\right)^{1+e}+\Delta T\\
-\frac{\bar{n}}{1+1/e}\left(z^*\right)^{1+1/e}\bar{n}^{-1-1/e}+z^*\left(1-t\right) &= -\frac{\bar{n}}{1+1/e}\left(1-t\right)^{1+e}+\bar{n}\left(1-t\right)^{1+e}+\Delta T\\
-\frac{\left(z^*\right)^{1+1/e}}{1+1/e}\bar{n}^{-1/e}+z^*\left(1-t\right) &= -\bar{n}\left(\frac{1}{1+1/e}\left(1-t\right)^{1+e}+\left(1-t\right)^{1+e}\right)+\Delta T\\
-\frac{\left(z^*\right)^{1+1/e}}{1+1/e}\bar{n}^{-1/e}+z^*\left(1-t\right) &= -\bar{n}\left(\frac{1}{1+1/e}\left(1-t\right)^{1+e}+\left(1-t\right)^{1+e}\right)+\Delta T
\end{aligned}$$

We also know from the FOC w.r.t. c that $\lambda=1$ so this simplfies to

We also know from the FOC w.r.t. c that $\lambda=1$ so this simplfies to

$$\begin{aligned}
\mathcal{L}\left(\bar{n},c\right) &= c-\frac{\bar{n}}{1+1/e}\left(\frac{z^*}{\bar{n}}\right)^{1+1/e}+z^*\left(1-t\right)+R-c\\
&= -\frac{\bar{n}}{1+1/e}\left(\frac{z^*}{\bar{n}}\right)^{1+1/e}+z^*\left(1-t\right)+R
\end{aligned}$$

We also know that the lower bound person also reports the same

$$\mathcal{L}\left(\tilde{n}\right)=c-\frac{\tilde{n}}{1+1/e}\left(\frac{z^*}{\tilde{n}}\right)^{1+1/e}+\lambda\left[z^*\left(1-t\right)+R-c\right]$$

$$\mathcal{L}\left(\bar{n}\right)=c-\frac{\bar{n}}{1+1/e}\left(\frac{z^*}{\bar{n}}\right)^{1+1/e}+\lambda\left[z^*\left(1-t\right)+R-c\right]$$

Postulate that the region is this

$$n=\left[z^*/\left(1-t\right)^e+\Delta T,z^*/\left(1-t\right)^e\right]$$

Nate said that the budget constraint alone is not the right thing to look at and I believe that is a correct statement.

The steps one would take in the kinked case are simply

$$\begin{aligned}
\mathcal{L}\left(z^*=z,\bar{n},c\mid t_0\right) &= \mathcal{L}\left(z^*=\bar{n}\left(1-t_1\right)^e,\bar{n},c\mid t_0\right)\\
-\frac{\bar{n}}{1+1/e}\left(\frac{z^*}{\bar{n}}\right)^{1+1/e}+z^*\left(1-t_0\right)+R &= -\frac{\bar{n}}{1+1/e}\left(\frac{\bar{n}\left(1-t_1\right)^e}{\bar{n}}\right)^{1+1/e}+\bar{n}\left(1-t_1\right)^e\left(1-t_0\right)+R\\
-\frac{\bar{n}}{1+1/e}\left(\frac{z^*}{\bar{n}}\right)^{1+1/e}+z^*\left(1-t_0\right) &= -\frac{\bar{n}}{1+1/e}\left(\frac{\bar{n}\left(1-t_1\right)^e}{\bar{n}}\right)^{1+1/e}+\bar{n}\left(1-t_1\right)^e\left(1-t_0\right)
\end{aligned}$$

So clearly the answer is $z^*=\bar{n}\left(1-t_1\right)^e$. Do the same for the lower kink.

So then in the notches case

$$\begin{aligned}
\mathcal{L}\left(z^*=z,\bar{n},c\mid t,\Delta T=0\right) &= \mathcal{L}\left(z^*=\bar{n}\left(1-t_1\right)^e,\bar{n},c\mid t,\Delta T\neq 0\right)\\
-\frac{\bar{n}}{1+1/e}\left(\frac{z^*}{\bar{n}}\right)^{1+1/e}+z^*\left(1-t\right)+R &= -\frac{\bar{n}}{1+1/e}\left(\frac{\bar{n}\left(1-t\right)^e}{\bar{n}}\right)^{1+1/e}+\bar{n}\left(1-t\right)^e\left(1-t\right)+\Delta T+R\\
-\frac{\bar{n}}{1+1/e}\left(\frac{z^*}{\bar{n}}\right)^{1+1/e}+z^*\left(1-t\right) &= -\frac{\bar{n}}{1+1/e}\left(\frac{\bar{n}\left(1-t\right)^e}{\bar{n}}\right)^{1+1/e}+\bar{n}\left(1-t\right)^{e+1}+\Delta T\\
-\frac{\bar{n}}{1+1/e}\left(\frac{z^*}{\bar{n}}\right)^{1+1/e}+z^*\left(1-t\right) &= \bar{n}(1-t)^{1+e}\left(\frac{1+1/e}{1+1/e}-\frac{1}{1+1/e}\right)+\Delta T\\
-\frac{\bar{n}}{1+1/e}\left(\frac{z^*}{\bar{n}}\right)^{1+1/e}+z^*\left(1-t\right) &= \bar{n}(1-t)^{1+e}\frac{1}{1+e}+\Delta T
\end{aligned}$$

So the goal is to try to write this expression as a simple closed form algebraic expression.

Find $\underline{n}$ such that these two equations are equal

$$\begin{aligned}
\mathcal{L}\left(z^*=z,\underline{n},c\mid t,\Delta T=0\right) &= \mathcal{L}\left(z=\underline{n}\left(1-t\right)^e,\underline{n},c\mid t,\Delta T=0\right)\\
-\frac{\underline{n}}{1+1/e}\left(\frac{z^*}{\underline{n}}\right)^{1+1/e}+z^*\left(1-t\right)+R &= -\frac{\underline{n}}{1+1/e}\left(\frac{\underline{n}\left(1-t\right)^e}{\underline{n}}\right)^{1+1/e}+\underline{n}\left(1-t\right)^e\left(1-t\right)+R\\
-\frac{\underline{n}}{1+1/e}\left(\frac{z^*}{\underline{n}}\right)^{1+1/e}+z^*\left(1-t\right) &= -\frac{\underline{n}}{1+1/e}\left(\frac{\underline{n}\left(1-t\right)^e}{\underline{n}}\right)^{1+1/e}+\underline{n}\left(1-t\right)^e\left(1-t\right)
\end{aligned}$$

so the lower bound $\underline{n}$ has to be $z^*=\underline{n}\left(1-t\right)^e$.

Next, Saez defines Δz^*. This is done by considering the highest ability person that reports $z=z^*$. That person has ability $n^+=z^*/\left(1-t_1\right)^e$. Remember that utility maximization with his utility function under a linear tax implied they would report $z=n\left(1-t\right)^e$. Hence, if this person had faced a linear tax of t_0, they would report

$$\begin{aligned}
z^+ &= n^+\left(1-t_0\right)^e\\
z^+ &= z^*\left(\frac{1-t_0}{1-t_1}\right)^e
\end{aligned}$$

Similarly, the lowest ability person that reports $z=z^*$ has ability $n_-=z^*/\left(1-t_0\right)^e$. If they had faced the linear tax of t_0, they would report

$$\begin{aligned} z_- &= n_- (1-t_0)^e \\ z_- &= z^* \end{aligned}$$

Combining these two expression defines

$$\begin{aligned} \Delta z^* &\equiv z^+ - z_- \\ &= z^* \left(\frac{1-t_0}{1-t_1}\right)^e - z^* \end{aligned}$$

Notice that both z^+ and z_- are counterfactual levels of reported income that will NEVER be observed even if all the assumptions of the model are correct. Saez writes counterfactual Δz^* as equation (3)

$$\frac{\Delta z^*}{z^*} = \left(\frac{1-t_0}{1-t_1}\right)^e - 1$$

This can be rewritten as

$$e = \frac{ln\left(1+\frac{\Delta z^*}{z^*}\right)}{ln\left(\frac{1-t_0}{1-t_1}\right)}$$

Now, we let

$$\Delta z^* = B/f_-() = \frac{F()-F()}{f_-()} \equiv b$$

Which allows us to write with the approximation that $log(1+x) = x$,

$$e = \frac{b}{Kln\left(\frac{1-t_0}{1-t_1}\right)}$$

which is exactly the equation (6) in Chetty et al. (2011).

Remember that he defines the counterfactual linear tax PDF as

$$h_0(z) = H_0'(z)$$

where

$$h_0(z) = f\left(z/(1-t_0)^e\right)/(1-t_0)^e$$

and

$$H_0(z) = F\left(z/(1-t_0)^e\right)$$

Following his notation, to the left hand side of the approximation in equation (4) we have

$$
\begin{aligned}
B^{Saez} &= \int_{z^*}^{z^*+\Delta z^*} h_0(z)\,dz \\
&= H_0(z^*+\Delta z^*) - H_0(z^*) \\
&= F\left((z^*+\Delta z^*)/(1-t_0)^e\right) - F\left(z^*/(1-t_0)^e\right) \\
&= F\left(\left(z^* + z^*\left(\frac{1-t_0}{1-t_1}\right)^e - z^*\right)/(1-t_0)^e\right) - F\left(z^*/(1-t_0)^e\right) \\
&= F\left(z^*\left(\frac{1-t_0}{1-t_1}\right)^e/(1-t_0)^e\right) - F\left(z^*/(1-t_0)^e\right) \\
&= F\left(z^*/(1-t_1)^e\right) - F\left(z^*/(1-t_0)^e\right)
\end{aligned}
$$

which you can see is the exact expression for B given in equation (2) above and our model exactly recovers the same theoretical mass at the kink as long as $\beta = 0$.

My next move would be to apply a change of variables to get this expression in logs so that I can worry about $z^* - e\ln(1-t_1)$ we hope that will give us

$$
B^{Saez} = F\left(z^* - e\ln(1-t_1)\right) - F\left(z^* - e\ln(1-t_0)\right)
$$

Just to make the point easy assume the log linear ability distribution has a mean μ and standard deviation σ that doesn't depend on the tax rate then the standardized distribution would be

$$
B^{Saez} = F\left(\frac{z^* - e\ln(1-t_1) - \mu}{\sigma}\right) - F\left(\frac{z^* - e\ln(1-t_0) - \mu}{\sigma}\right)
$$

$$
\begin{aligned}
\frac{\partial B^{Saez}}{\partial \mu} &= \frac{\partial}{\partial \mu}F\left(\frac{z^* - e\ln(1-t_1) - \mu}{\sigma}\right) - \frac{\partial}{\mu}F\left(\frac{z^* - e\ln(1-t_0) - \mu}{\sigma}\right) \\
&= \frac{\partial}{\partial x_1}F(x_1)\frac{\partial x_1}{\partial \mu} - \frac{\partial}{\partial x_0}F(x_0)\frac{\partial x_0}{\partial \mu} \\
&= f(x_1)\left(-\frac{\mu}{\sigma}\right) - f(x_0)\left(-\frac{\mu}{\sigma}\right) \\
&= \frac{\mu}{\sigma}\left(f(x_0) - f(x_1)\right)
\end{aligned}
$$

which may have any sign but is certainly non-zero as long as $\mu = 0$.

My next move would be to apply a change of variables to get this expression in logs so that I can worry about $z^* - e\ln(1-t_1)$ we hope that will give us

$$
B^{Saez} = F\left(z^* - e\ln(1-t_1)\right) - F\left(z^* - e\ln(1-t_0)\right)
$$

Just to make the point easy assume the log linear ability distribution has a mean μ and standard deviation σ that doesn't depend on the tax rate then the standardized distribution

would be

$$B^{Saez} = F\left(\frac{z^* - e\ln(1-t_1) - \mu}{\sigma}\right) - F\left(\frac{z^* - e\ln(1-t_0) - \mu}{\sigma}\right)$$

$$\begin{aligned}\frac{\partial B^{Saez}}{\partial \mu} &= \frac{\partial}{\partial \mu}F\left(\frac{z^* - e\ln(1-t_1) - \mu}{\sigma}\right) - \frac{\partial}{\mu}F\left(\frac{z^* - e\ln(1-t_0) - \mu}{\sigma}\right)\\ &= \frac{\partial}{\partial x_1}F(x_1)\frac{\partial x_1}{\partial \mu} - \frac{\partial}{\partial x_0}F(x_0)\frac{\partial x_0}{\partial \mu}\\ &= f(x_1)\left(-\frac{\mu}{\sigma}\right) - f(x_0)\left(-\frac{\mu}{\sigma}\right)\\ &= \frac{\mu}{\sigma}\left(f(x_0) - f(x_1)\right)\end{aligned}$$

which may have any sign but is certainly non-zero as long as $\mu = 0$.

F.8.1 Utility maximization with notches

Introducing a notch in the tax rate changes the problem into

$$\begin{aligned}&\max_{c,z} \quad c - \frac{n}{1+1/e}\left(\frac{z}{n}\right)^{1+1/e}\\ &s.t.\\ &c = z(1-t) + \Delta T 1(z > z^*) + R\end{aligned} \tag{53}$$

in which c is consumption, z is earnings, and R is non-earning resources that can be spent on consumption. An unobservable ability variable, n, is distributed according to some PDF $f(n)$ and some CDF $F(n)$ and affects the level of optimal consumption for each consumer. Notice we need to have $z \in (0,\infty)$ and $n \in (0,\infty)$. The Lagrangian is

$$\begin{aligned}\mathcal{L} &= c - \frac{n}{1+1/e}\left(\frac{z}{n}\right)^{1+1/e} + \lambda\left[z(1-t) + \Delta T 1(z > z^*) + R - c\right]\\ &= c - \frac{n}{1+1/e}\left(\frac{1}{n}\right)^{1+1/e} z^{1+1/e} + \lambda\left[z(1-t) + \Delta T 1(z > z^*) + R - c\right]\end{aligned}$$

We want to maximize this lagrangian we can differentiate with respect to c easily to get started

$$\begin{aligned}\mathcal{L}_c = 1 - \lambda &= 0\\ 1 &= \lambda\end{aligned}$$

We have three regions, they are $z < z^*$ $z = z^*$ and $z > z^*$. We can differentiate with respect to z as long we as are in regions in which the Lagrangian is not discontinuous. These are the $z < z^*$ and $z > z^*$ regions. We can solve it by considering, cases, however. Start with the $z < z^*$ region then the $\mathcal{L}$ becomes

$$\mathcal{L} \mid z < z^* \quad = \quad c - \frac{n}{1+1/e}\left(\frac{z}{n}\right)^{1+1/e} + \lambda\left[z(1-t) + R - c\right]$$

$$\begin{aligned}
\mathcal{L}_z \mid z < z^* = -n\left(\frac{1}{n}\right)^{1+1/e} z^{1/e} + \lambda(1-t) &= 0 \\
n(n)^{-1-1/e} z^{1/e} &= \lambda(1-t) \\
n^{-1/e} z^{1/e} &= \lambda(1-t) \\
\left(\frac{z}{n}\right)^{1/e} &= \lambda(1-t) \\
\frac{z}{n} &= \lambda^e(1-t)^e \\
z &= n\lambda^e(1-t)^e
\end{aligned}$$

Since we know that $\lambda = 1$ the optimal value when $z < z^*$ is

$$z = n(1-t)^e$$

We can follow the same steps for the $\mathcal{L} \mid z > z^*$ and will do so carefully because this can get confusing

$$\mathcal{L} \mid z > z^* \quad = \quad c - \frac{n}{1+1/e}\left(\frac{z}{n}\right)^{1+1/e} + \lambda\left[z(1-t) + \Delta T + R - c\right]$$

$$\begin{aligned}
\mathcal{L}_z \mid z > z^* = -n\left(\frac{1}{n}\right)^{1+1/e} z^{1/e} + \lambda(1-t) &= 0 \\
n(n)^{-1-1/e} z^{1/e} &= \lambda(1-t) \\
n^{-1/e} z^{1/e} &= \lambda(1-t) \\
\left(\frac{z}{n}\right)^{1/e} &= \lambda(1-t) \\
\frac{z}{n} &= \lambda^e(1-t)^e \\
z &= n\lambda^e(1-t)^e
\end{aligned}$$

$$z = n(1-t)^e$$

The solution to the case when $z = z^*$ is trivial because that's just it. This implies that the casewise optimal solution is therefore

$$z = \begin{cases} n\left(1-t\right)^{e} & z<z^{*} \\ z^{*} & z=z^{*} \\ n\left(1-t\right)^{e} & z>z^{*} \end{cases} \tag{54}$$

Our goal is to find the person with highest ability $\overline{n}$ that optimizes constrained utility by reporting $z=z^{*}$. Because the agent reports $z=z^{*}$, we can be sure that constrained utility is maximized at that point. If the agent is indifferent between reporting $z=z^{*}$ and facing t plus ΔT or reporting $z=\overline{n}\left(1-t\right)^{e}$ and facing that same tax rate t plus ΔT, given that they are optimizing at z^{*}, that means we only need to compare the alternate budget constraints which must provide the same resources in each case. The general budget constraint is

$$c = z\left(1-t\right)+\Delta T1\left(z>z^{*}\right)+R$$

The agent must have the same resources from reporting either case so this means we compare

$$\begin{aligned} c\left(z=z^{*},\overline{n}\mid t,\Delta T\right) &= c\left(z=\overline{n}\left(1-t\right)^{e},\overline{n}\mid t,\Delta T\right) \\ z^{*}\left(1-t\right)+\Delta T+R &= \overline{n}\left(1-t\right)^{e}\left(1-t\right)+\Delta T+R \\ z^{*}\left(1-t_{1}\right) &= \overline{n}\left(1-t_{1}\right)^{e}\left(1-t_{1}\right) \\ z^{*}\left(1-t_{1}\right) &= \overline{n}\left(1-t_{1}\right)^{1+e} \\ z^{*} &= \overline{n}\left(1-t_{1}\right)^{e} \end{aligned}$$

This might be easier to see by consider the full Lagrangian when we use the FOC w.r.t. c to get that $\lambda=1$

$$\begin{aligned} \mathcal{L}\left(z^{*}=z,\overline{n},c\right) &= \mathcal{L}\left(z^{*}=\overline{n}\left(1-t_{1}\right)^{e},\overline{n},c\right) \\ -\frac{\overline{n}}{1+1/e}\left(\frac{z^{*}}{\overline{n}}\right)^{1+1/e}+z^{*}\left(1-t_{1}\right)+R &= -\frac{\overline{n}}{1+1/e}\left(\frac{z^{*}}{\overline{n}}\right)^{1+1/e}+\overline{n}\left(1-t_{1}\right)^{1+e}+R \\ z^{*}\left(1-t_{1}\right) &= \overline{n}\left(1-t_{1}\right)^{1+e} \\ z^{*} &= \overline{n}\left(1-t_{1}\right)^{e} \end{aligned}$$

Our goal is to find the person with highest ability $\overline{n}$ that is indifferent between reporting $z=z^{*}$ or reporting $z=n\left(1-t\right)^{e}$ and facing the tax notch ΔT. $z>z^{*}$. We know that reporting $z=z^{*}$ is optimal so we know the Lagrangian is

$$\begin{aligned} \mathcal{L}\left(z^{*}=z,\overline{n},c\right) &= c-\frac{\overline{n}}{1+1/e}\left(\frac{z^{*}}{\overline{n}}\right)^{1+1/e}+\lambda\left[z^{*}\left(1-t\right)+R-c\right] \\ &= -\frac{\overline{n}}{1+1/e}\left(\frac{z^{*}}{\overline{n}}\right)^{1+1/e}+z^{*}\left(1-t\right)+R \end{aligned}$$

In which the last step comes from the fact that the FOC w.r.t. c ensures $\lambda = 1$. This person should get the same utility from reporting $z = n\left(1-t\right)^{e}$

$$\begin{aligned}
\mathcal{L}\left(z^{*} = \overline{n}\left(1-t\right)^{e}, \overline{n}, c\right) &= c - \frac{\overline{n}}{1+1/e}\left(\frac{\overline{n}\left(1-t\right)^{e}}{\overline{n}}\right)^{1+1/e} + \lambda\left[\overline{n}\left(1-t\right)^{e}\left(1-t\right) + \Delta T + R - c\right] \\
&= c - \frac{\overline{n}}{1+1/e}\left(\left(1-t\right)^{e}\right)^{1+1/e} + \lambda\left[\overline{n}\left(1-t\right)^{1+e} + \Delta T + R - c\right] \\
&= c - \frac{\overline{n}}{1+1/e}\left(1-t\right)^{1+e} + \lambda\left[\overline{n}\left(1-t\right)^{1+e} + \Delta T + R - c\right] \\
&= -\frac{\overline{n}}{1+1/e}\left(1-t\right)^{1+e} + \overline{n}\left(1-t\right)^{1+e} + \Delta T + R
\end{aligned}$$

Setting these equal and solving for $\overline{n}$

$$\begin{aligned}
\mathcal{L}\left(z^{*} = z, \overline{n}, c\right) &= \mathcal{L}\left(z^{*} = \overline{n}\left(1-t\right)^{e}, \overline{n}, c\right) \\
-\frac{\overline{n}}{1+1/e}\left(\frac{z^{*}}{\overline{n}}\right)^{1+1/e} + z^{*}\left(1-t\right) + R &= -\frac{\overline{n}}{1+1/e}\left(1-t\right)^{1+e} + \overline{n}\left(1-t\right)^{1+e} + \Delta T + R \\
-\frac{\overline{n}}{1+1/e}\left(\frac{z^{*}}{\overline{n}}\right)^{1+1/e} + z^{*}\left(1-t\right) &= -\frac{\overline{n}}{1+1/e}\left(1-t\right)^{1+e} + \overline{n}\left(1-t\right)^{1+e} + \Delta T \\
-\frac{\overline{n}}{1+1/e}\left(z^{*}\right)^{1+1/e}\overline{n}^{-1-1/e} + z^{*}\left(1-t\right) &= -\frac{\overline{n}}{1+1/e}\left(1-t\right)^{1+e} + \overline{n}\left(1-t\right)^{1+e} + \Delta T \\
-\frac{\left(z^{*}\right)^{1+1/e}}{1+1/e}\overline{n}^{-1/e} + z^{*}\left(1-t\right) &= -\overline{n}\left(\frac{1}{1+1/e}\left(1-t\right)^{1+e} + \left(1-t\right)^{1+e}\right) + \Delta T \\
-\frac{\left(z^{*}\right)^{1+1/e}}{1+1/e}\overline{n}^{-1/e} + z^{*}\left(1-t\right) &= -\overline{n}\left(\frac{1}{1+1/e}\left(1-t\right)^{1+e} + \left(1-t\right)^{1+e}\right) + \Delta T
\end{aligned}$$

We also know from the FOC w.r.t. c that $\lambda = 1$ so this simplfies to

We also know from the FOC w.r.t. c that $\lambda = 1$ so this simplfies to

$$\begin{aligned}
\mathcal{L}\left(\overline{n}, c\right) &= c - \frac{\overline{n}}{1+1/e}\left(\frac{z^{*}}{\overline{n}}\right)^{1+1/e} + z^{*}\left(1-t\right) + R - c \\
&= -\frac{\overline{n}}{1+1/e}\left(\frac{z^{*}}{\overline{n}}\right)^{1+1/e} + z^{*}\left(1-t\right) + R
\end{aligned}$$

We also know that the lower bound person also reports the same

$$\mathcal{L}\left(\tilde{n}\right) = c - \frac{\tilde{n}}{1+1/e}\left(\frac{z^{*}}{\tilde{n}}\right)^{1+1/e} + \lambda\left[z^{*}\left(1-t\right) + R - c\right]$$

$$\mathcal{L}(\overline{n}) = c - \frac{\overline{n}}{1+1/e}\left(\frac{z^*}{\overline{n}}\right)^{1+1/e} + \lambda\left[z^*(1-t) + R - c\right]$$

Postulate that the region is this

$$n = \left[z^*/(1-t)^e + \Delta T, z^*/(1-t)^e\right]$$

Nate said that the budget constraint alone is not the right thing to look at and I believe that is a correct statement.

The steps one would take in the kinked case are simply

$$\begin{aligned}
\mathcal{L}(z^* = z, \overline{n}, c \mid t_0) &= \mathcal{L}\left(z^* = \overline{n}(1-t_1)^e, \overline{n}, c \mid t_0\right) \\
-\frac{\overline{n}}{1+1/e}\left(\frac{z^*}{\overline{n}}\right)^{1+1/e} + z^*(1-t_0) + R &= -\frac{\overline{n}}{1+1/e}\left(\frac{\overline{n}(1-t_1)^e}{\overline{n}}\right)^{1+1/e} + \overline{n}(1-t_1)^e(1-t_0) + R \\
-\frac{\overline{n}}{1+1/e}\left(\frac{z^*}{\overline{n}}\right)^{1+1/e} + z^*(1-t_0) &= -\frac{\overline{n}}{1+1/e}\left(\frac{\overline{n}(1-t_1)^e}{\overline{n}}\right)^{1+1/e} + \overline{n}(1-t_1)^e(1-t_0)
\end{aligned}$$

So clearly the answer is $z^* = \overline{n}(1-t_1)^e$. Do the same for the lower kink.

So then in the notches case

$$\begin{aligned}
\mathcal{L}(z^* = z, \overline{n}, c \mid t, \Delta T = 0) &= \mathcal{L}\left(z^* = \overline{n}(1-t_1)^e, \overline{n}, c \mid t, \Delta T \neq 0\right) \\
-\frac{\overline{n}}{1+1/e}\left(\frac{z^*}{\overline{n}}\right)^{1+1/e} + z^*(1-t) + R &= -\frac{\overline{n}}{1+1/e}\left(\frac{\overline{n}(1-t)^e}{\overline{n}}\right)^{1+1/e} + \overline{n}(1-t)^e(1-t) + \Delta T + R \\
-\frac{\overline{n}}{1+1/e}\left(\frac{z^*}{\overline{n}}\right)^{1+1/e} + z^*(1-t) &= -\frac{\overline{n}}{1+1/e}\left(\frac{\overline{n}(1-t)^e}{\overline{n}}\right)^{1+1/e} + \overline{n}(1-t)^{e+1} + \Delta T \\
-\frac{\overline{n}}{1+1/e}\left(\frac{z^*}{\overline{n}}\right)^{1+1/e} + z^*(1-t) &= \overline{n}(1-t)^{1+e}\left(\frac{1+1/e}{1+1/e} - \frac{1}{1+1/e}\right) + \Delta T \\
-\frac{\overline{n}}{1+1/e}\left(\frac{z^*}{\overline{n}}\right)^{1+1/e} + z^*(1-t) &= \overline{n}(1-t)^{1+e}\frac{1}{1+e} + \Delta T
\end{aligned}$$

So the goal is to try to write this expression as a simple closed form algebraic expression.

Find $\underline{n}$ such that these two equations are equal

$$\begin{aligned}
\mathcal{L}(z^* = z, \underline{n}, c \mid t, \Delta T = 0) &= \mathcal{L}\left(z = \underline{n}(1-t)^e, \underline{n}, c \mid t, \Delta T = 0\right) \\
-\frac{\underline{n}}{1+1/e}\left(\frac{z^*}{\underline{n}}\right)^{1+1/e} + z^*(1-t) + R &= -\frac{\underline{n}}{1+1/e}\left(\frac{\underline{n}(1-t)^e}{\underline{n}}\right)^{1+1/e} + \underline{n}(1-t)^e(1-t) + R \\
-\frac{\underline{n}}{1+1/e}\left(\frac{z^*}{\underline{n}}\right)^{1+1/e} + z^*(1-t) &= -\frac{\underline{n}}{1+1/e}\left(\frac{\underline{n}(1-t)^e}{\underline{n}}\right)^{1+1/e} + \underline{n}(1-t)^e(1-t)
\end{aligned}$$

so the lower bound $\underline{n}$ has to be $z^* = \underline{n}(1-t)^e$.

Notice that both z^+ and z_- are counterfactual levels of reported income that will

NEVER be observed even if all the assumptions of the model are correct. Saez writes counterfactual Δz^* as equation (3)

$$\frac{\Delta z^*}{z^*} = \left(\frac{1-t_0}{1-t_1}\right)^e - 1$$

This can be rewritten as

$$e = \frac{ln\left(1+\frac{\Delta z^*}{z^*}\right)}{ln\left(\frac{1-t_0}{1-t_1}\right)}$$

Now, we let

$$\Delta z^* = B/f_-() = \frac{F()-F()}{f_-()} \equiv b$$

Which allows us to write with the approximation that $log(1+x) = x$,

$$e = \frac{b}{Kln\left(\frac{1-t_0}{1-t_1}\right)}$$

which is exactly the equation (6) in Chetty et al. (2011).

That person has ability $\overline{n} = z^*/\left(1-t_1\right)^e$. Remember that utility maximization with his utility function under a linear tax implied they would report $z = n\left(1-t\right)^e$. Hence, if this person had faced a linear tax of t_0, they would report

$$\begin{aligned} z^+ &= n^+\left(1-t_0\right)^e \\ z^+ &= z^*\left(\frac{1-t_0}{1-t_1}\right)^e \end{aligned}$$

Similarly, the lowest ability person that reports $z = z^*$ has ability $n_- = z^*/\left(1-t_0\right)^e$. If they had faced the linear tax of t_0, they would report

$$\begin{aligned} z_- &= n_-\left(1-t_0\right)^e \\ z_- &= z^* \end{aligned}$$

Combining these two expression defines

$$\begin{aligned} \Delta z^* &\equiv z^+ - z_- \\ &= z^*\left(\frac{1-t_0}{1-t_1}\right)^e - z^* \end{aligned}$$

G Evasion problem

Saez problem

$$\begin{aligned}
\max_{C_i, L_i} \quad & C_i - (N_i^*)^{-1/\varepsilon} \frac{L_i^{1+\frac{1}{\varepsilon}}}{1+1/\varepsilon} \\
s.t. \quad Y_i &= L_i \\
L_i &= H_i^\theta R_i^{1-\theta} \\
C_i &= Y_i - t_0 R_i + (t_0 - t_1)(R_i - D_j) \mathbb{1}(R_i > D_j)
\end{aligned}$$

The agent sells into the labor market quantity L_i but is only taxed on reported quantity of labor R_i and not on hidden quantity of labor H_i. Notice we have already assumed that the numeraire is the real wage which is equal to one. Substituion will allow us to write

$$\begin{aligned}
\max_{C_i, H_i, R_i} \quad & C_i - (N_i^*)^{-1/\varepsilon} \frac{\left(H_i^\theta R_i^{1-\theta}\right)^{1+\frac{1}{\varepsilon}}}{1+1/\varepsilon} \\
s.t. \quad C_i &= H_i^\theta R_i^{1-\theta} - t_0 R_i + (t_0 - t_1)(R_i - D_j) \mathbb{1}(R_i > D_j)
\end{aligned}$$

The Lagrangian becomes

$$\mathcal{L} = C_i - (N_i^*)^{-1/\varepsilon} \frac{\left(H_i^\theta R_i^{1-\theta}\right)^{1+\frac{1}{\varepsilon}}}{1+1/\varepsilon} + \lambda \left(H_i^\theta R_i^{1-\theta} - t_0 R_i + (t_0 - t_1)(R_i - D_j) \mathbb{1}(R_i > D_j) - C_i\right)$$

Diff w.r.t. consumption to get

$$\mathcal{L}_c = 1 - \lambda = 0$$

Notice that regardless of the level of reported income R_i the Lagrangian as a function of H_i remains the same. As such, it is always the case that diff w.r.t. H_i gives

$$\begin{aligned}
\mathcal{L}_H = -(N_i^*)^{-1/\varepsilon} \left(H_i^\theta R_i^{1-\theta}\right)^{\frac{1}{\varepsilon}} H_i^{\theta-1} R_i^{1-\theta}\theta + \lambda H_i^{\theta-1} R_i^{1-\theta}\theta &= 0 \\
\lambda H_i^{\theta-1} R_i^{1-\theta}\theta &= (N_i^*)^{-1/\varepsilon} \left(H_i^\theta R_i^{1-\theta}\right)^{\frac{1}{\varepsilon}} H_i^{\theta-1} R_i^{1-\theta}\theta \\
\lambda\theta &= (N_i^*)^{-1/\varepsilon} \left(H_i^\theta R_i^{1-\theta}\right)^{\frac{1}{\varepsilon}} \theta
\end{aligned}$$

I guess everything is pinned down by this equation

$$\begin{aligned}
\lambda &= (N_i^*)^{-1/\varepsilon} \left(H_i^\theta R_i^{1-\theta}\right)^{\frac{1}{\varepsilon}} \\
1 &= (N_i^*)^{-1/\varepsilon} \left(H_i^\theta R_i^{1-\theta}\right)^{\frac{1}{\varepsilon}} \\
(N_i^*)^{1/\varepsilon} &= \left(H_i^\theta R_i^{1-\theta}\right)^{\frac{1}{\varepsilon}} \\
N_i^* &= H_i^\theta R_i^{1-\theta} \\
N_i^* &= L_i
\end{aligned}$$

And so we did not generate a kink in reported income. Maybe if we use CES of H_i and $R_i$we will be able to figure it out.

The Lagrangian differs for different values of the reported hours work, however. Consider the region in which $R_i < D_j$

$$\mathcal{L} \mid R_i \leq D_j = C_i - \left(N_i^*\right)^{-1/\varepsilon} \frac{\left(H_i^\theta R_i^{1-\theta}\right)^{1+\frac{1}{\varepsilon}}}{1+1/\varepsilon} + \lambda \left(H_i^\theta R_i^{1-\theta} - t_0 R_i - C_i\right)$$

Diff w.r.t. R_i for $R_i \leq D_j$ to get

$$\begin{aligned}
\mathcal{L}_R \mid R_i \leq D_j = -\left(N_i^*\right)^{-1/\varepsilon} \left(H_i^\theta R_i^{1-\theta}\right)^{\frac{1}{\varepsilon}} H_i^\theta R_i^{-\theta} (1-\theta) + \lambda H_i^\theta R_i^{-\theta} (1-\theta) - \lambda t_0 &= 0 \\
-\lambda H_i^\theta R_i^{-\theta} (1-\theta) + \lambda H_i^\theta R_i^{-\theta} (1-\theta) - \lambda t_0 &= 0 \\
-H_i^\theta R_i^{-\theta} (1-\theta) + H_i^\theta R_i^{-\theta} (1-\theta) - t_0 &= 0 \\
t_0 &= 0
\end{aligned}$$

which I think means we have a corner solution? Lagrangian for $R_i > D_j$

$$\mathcal{L} \mid R_i > D_j = C_i - \left(N_i^*\right)^{-1/\varepsilon} \frac{\left(H_i^\theta R_i^{1-\theta}\right)^{1+\frac{1}{\varepsilon}}}{1+1/\varepsilon} + \lambda \left(H_i^\theta R_i^{1-\theta} - t_1 R_i - C_i\right)$$

Diff w.r.t. R_i for $R_i > D_j$ to get

$$\begin{aligned}
\mathcal{L}_R \mid R_i > D_j = -\left(N_i^*\right)^{-1/\varepsilon} \left(H_i^\theta R_i^{1-\theta}\right)^{\frac{1}{\varepsilon}} H_i^\theta R_i^{-\theta} (1-\theta) + \lambda H_i^\theta R_i^{-\theta} (1-\theta) - \lambda t_1 &= 0 \\
-\lambda H_i^\theta R_i^{-\theta} (1-\theta) + \lambda H_i^\theta R_i^{-\theta} (1-\theta) - \lambda t_1 &= 0 \\
-H_i^\theta R_i^{-\theta} (1-\theta) + H_i^\theta R_i^{-\theta} (1-\theta) - t_1 &= 0 \\
t_1 &= 0
\end{aligned}$$

$$\mathcal{L}_R \mid R_i > D_j \quad = \quad -\left(N_i^*\right)^{-1/\varepsilon} \left(H_i^\theta R_i^{1-\theta}\right)^{\frac{1}{\varepsilon}} H_i^\theta R_i^{-\theta} (1-\theta) + \lambda H_i^\theta R_i^{-\theta} (1-\theta) = 0$$

call H_i work hours for which you are paid W_i and call R_i hours that you spend evading taxes. You get paid W_i for before taxes for each hour worked. You get no wage income for hours spent evading taxes but you do get to pay less taxes so I think the implicit wage for that is $1-t$

$$\begin{aligned}
\max_{C_i, L_i, R_i} \quad & C_i - \left(N_i^*\right)^{-1/\varepsilon} \frac{L_i^{1+\frac{1}{\varepsilon}}}{1+1/\varepsilon} \\
s.t. \quad Y_i &= W_i L_i \\
L_i &= H_i^\theta R_i^{1-\theta} \\
C_i &= (1-t_0) Y_i + (t_0 - t_1)(Y_j - D_j) \mathbb{1}(Y_i > D_j)
\end{aligned}$$

Substitution will give you

$$\begin{aligned} \max_{C_i,H_i,R_i} \quad & C_i - \left(N_i^*\right)^{-1/\varepsilon} \frac{\left(H_i^{\theta} R_i^{1-\theta}\right)^{1+\frac{1}{\varepsilon}}}{1+1/\varepsilon} \\ s.t. \quad Y_i &= W_i H_i \\ C_i &= (1-t_0)\, Y_i + (t_0 - t_1)\,(Y_j - D_j)\, \mathbb{1}\,(Y_i > D_j) \end{aligned}$$

H Optimizing frictions: a policy relevant elasticity

H.1 Previous incorporations of optimizing frictions

Models without optimizing frictions produces stark predictions on the observed distribution of Y_i around the discontinuity. For example, models with a notch predict that for some region to the right of the notch there will be zero density, a hole and models with a kink predict that agents will move exactly to the kink point and the excess mass will occur only at the kink. In practice neither of these predictions are true in the data: to the right of a notch we observe mass where there should not be any and we observe excess mass in a window around a kink, not only at the kink point.

Most previous work explains the difference between the model's predictions and the observed data by appealing to optimizing frictions that limit agents' ability to adjust or adjust finely enough to locate exactly at a discontinuity (see, e.g., Chetty, Friedman, Olsen, and Pistaferri, 2011; Gelber, Jones, and Sacks, 2013). Allowing for optimizing frictions is critical for identifying the preference parameters in the model. In appendix XX, we demonstrate that as a practical matter there are large deviations between the true and estimated preference parameters in the model when the data generating process includes optimizing frictions but the estimator does not, Therefore, accounting for optimizing frictions is a necessary condition for any estimator using discontinuities.

Current methods of incorporating optimizing frictions have been criticized for being ad hoc. For example, many studies using kink points arbitrarily choose a window around the kink from which to estimate the excess mass. The implicit assumption is that optimizing frictions cause agents to bunch around rather than exactly at the kink point. Unfortunately, the choice of window exactly identifies the preference parameters. Said differently, almost any positive value for a preference parameter can be estimated given a different window. We demonstrate this point through a series of examples of current methods in Appendix XX.

Chetty, Friedman, Olsen, and Pistaferri (2011) demonstrates that current bunching methods produce elasticity estimates that cannot be used for policy inference precisely because they do not adequately account for optimizing frictions. Chetty, Friedman, Olsen, and Pistaferri (2011) demonstrate this point by showing that current bunching methods have some peculiar implications. First, if the elasticity estimate is recovering the true elasticity, then the estimate should be invariant to the size of the tax difference. Chetty, Friedman, Olsen, and Pistaferri (2011) show, however, that the elasticity estimates from current bunching estimators increase with the size of the change in tax rates at the discontinuity. This prediction arises in their model because individuals only move to the discontinuity if the gains are larger than the fixed costs of moving. This implies that more individuals move and bunch when the tax rate difference is larger because that is when the gains from moving are larger.

To produce estimates of the elasticity that are useful for policy, we incorporate optimizing frictions into the model and estimator. As a test that our estimator produces policy relevant elasticities---those that recover the true elasticity---we show that our estimate is invariant to the size of the tax change.

H.2 Incorporating optimizing frictions into the model

To construct an estimator that recovers policy relevant elasticities, we incorporate optimizing frictions in the model in two ways. The first, is a heterogeneous fixed cost of adjustment that causes some agents to not adjust. Specifically, with probability π_d, $\zeta_i = 0$ and the agent's fixed costs are sufficiently low that the agent re-optimizes and with probability $(1 - \pi_d)$, $\zeta_i = 1$ and the agent's fixed costs are sufficiently high that the agent does not re-optimize. Most studies focused on kinks assume $\pi_d = 1$, while studies focused on notches consider $\pi_d \in (0, 1)$ to account for the mass of agents just right of the notch.

We also model optimizing frictions as an error term that causes desired demand to differ from actual demand according to the parameter ξ, with cumulative distribution function $F_\xi(\xi_i)$ and probability distribution $f_\xi(\xi_i)$. The optimizing frictions, ξ_i are unknown to individuals when they choose Y_i and capture, for example when Y_i is taxable income, the lumpiness of pay and deductions, bonuses, forced overtime, forced short-time, and uncertainty that may exist over items like capital gains. This modeling technique also captures the possibility that individuals, especially in the short-run, may only have discrete choices along their budget constraint.

With these optimizing frictions, the log of Y_i, defined as y_i, with J discontinuities can be written as,

$$y_i = \begin{cases} d_j + \xi_i & n_i^* \in \left[\underline{n}_j, \overline{n}_j\right] \text{ and } \zeta_i = 0 \quad j = 1, \ldots, J \\ w_i + l_i\left(s_{j-1}, n_i^*, r_i; \varepsilon\right) + \xi_i & n_i^* \in \left(\overline{n}_{j-1}, \underline{n}_j\right) \text{ and } \zeta_i = 1 \quad j = 1, \ldots, J \\ w_i + l_i\left(s_{j-1}, n_i^*, r_i; \varepsilon\right) + \xi_i & n_i^* \in \left(\overline{n}_{j-1}, \underline{n}_j\right) \quad j = 1, \ldots, J+1, \end{cases} \tag{55}$$

in which $\overline{n}_0 \equiv n_{min}$ and $\underline{n}_{J+1} \equiv n_{max}$. We index these cases by $q_{i,h}$, which equals one in case h and zero otherwise such that, $q_{i,1} = 1$ if $y_{0,i} \in (n_{min}, d_1]$, and zero otherwise; $q_{i,2} = 1$ if $y_{0,i} \in (d_1, d_1 + \varepsilon s_0 - \varepsilon s_1]$ and $\zeta = 1$, and zero otherwise; $q_{i,3} = 1$ if $y_{0,i} \in (d_1, d_1 + \varepsilon s_0 - \varepsilon s_1]$ and $\zeta = 0$, and zero otherwise; and $q_{i,H} = 1$ if $y_{0,i} \in (d_J + \varepsilon s_{J-1} - \varepsilon s_J, n_{max})$ and zero otherwise, where $H = 3J - 2$.

H.3 Quasi-linear and Iso-elastic Utility with only Slope Changes

When individuals maximize iso-elastic and quasilinear utility with a budget constraint that has kinks, demand can be written as,

$$y_i = \begin{cases} d_j + \xi_i & \text{if } y_{0,i} \in [d_j, d_j + \varepsilon s_{j-1} - \varepsilon s_j] \text{ and } \zeta_i = 0 & \text{for } j = 1, ..., J \\ n_i + \varepsilon s_{j-1} + \xi_i & \text{if } y_{0,i} \in [d_j, d_j + \varepsilon s_{j-1} - \varepsilon s_j] \text{ and } \zeta_i = 1 & \text{for } j = 1, ..., J \\ n_i + \varepsilon s_{j-1} + \xi_i & \text{if } y_{0,i} \in [d_{j-1} + \varepsilon s_{j-2} - \varepsilon s_{j-1}, d_j,] & \text{for } j = 1, ..., J+1 \end{cases}$$

where $d_0 \equiv n_{min}$, and $d_{J+1} \equiv n_{max}$.