Why Don't Taxpayers Bunch at Kink Points?*

Andrew H. McCallum †

January 15, 2025

Preliminary and Incomplete: Comments Welcome.

Abstract

I study agents' responses to changes in the slope or the intercept of a piecewise linear schedule of incentives. For example, past theory predicts that for increasing marginal income tax rates, many taxpayers will report income exactly at the threshold where the tax rate increases. Many empirical settings that would in theory predict a mass-point, instead have a diffused mass near the threshold, or no excess mass at all. I attribute these diffuse mass points to optimizing frictions and introduce new theory and new estimation methods to allow these frictions to depend on observables. Our methods are not limited to public finance and apply to a general class of mixture models and any of the four possible piecewise linear constraints, 1) slope increase, 2) slope decrease, 3) intercept increase, or 4) intercept decrease. I demonstrate these methods in three out of the four settings. I document which covariates account for a substantial share of optimizing frictions and provide elasticity estimates that explicitly control for optimizing frictions in the context of a slope decrease due to the Earn Income Tax Credit.

JEL codes: C14, H24, J20
Keywords: bunching, notching, optimizing frictions, EM algorithm

*Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the Board of Governors of the Federal Reserve System, or any other person associated with the Federal Reserve System. Michael A. Navarrete contributed to early stages of related research with the same paper title. Hudson Hinshaw, Alexis Payne, and Benjamin Stagoff-Belfort, provided excellent research assistance.

†andrew.h.mccallum@frb.gov, Board of Governors of the Federal Reserve System, Washington, DC 20551

1 Introduction

In several distributions I see a mass point where individuals make decisions that place them at or near this point in the distribution. This type of behavior is referred to as bunching in the literature. Typically when bunching is present, there is an incentive for individuals to do so. The stronger the incentive to bunch, the more likely I are to see a larger mass point in the distribution. Bunching happens because there is a change in the incentive schedule. If the slope of the incentive schedule changes, then I refer to that point as a kink. If the intercept of the incentive schedule changes, then I refer to that point as a notch. In general, notches lead to more bunching because notches have strong disincentives to be just to the right of that point by creating dominated regions. Individuals that are optimizing with perfect information should never be just to the right of a notch because their payoff will lower than if they just bunched at the notch point. Convex kinks do not create this dominated region where it would never be optimal for agents to be located just to the right of a kink point. However, concave kinks which occur less commonly than convex kinks empirically do exhibit a dominated region.

One gap in the literature is that I observe bunching only in a few cases. Several kink points exhibit no bunching despite individuals having an incentive to do so. Much of the research on bunching has focused only when bunching is especially pronounced: sharp bunching. When researchers apply their bunching estimators to non-sharp bunching cases, "fuzzy" bunching, they imprecisely estimate their parameters of interest. I construct a micro founded model that allows us to explain the observed "fuzzy" bunching, while simultaneously retrieving parameters of interest from these kink points. Although I are not the first paper to develop a bunching estimator that allows for "fuzzy" bunching (Alvero and Xiao, 2020), I are the first to develop a bunching estimator that incorporates optimizing frictions that are micro-founded with covariates both on the individuals' ability and individuals' optimizing frictions. I are also the first to use an estimator, maximum likelihood estimation (MLE), in conjunction with the Expectation Maximization (EM) algorithm

(Dempster, Laird, and Rubin, 1977) extension to missing data (Ruud, 1991) to all 4 cases: (1) convex kink, (2) concave kink, (3) convex notch, and (4) concave notch.

Although the tax salience literature (Chetty, Looney, and Kroft (2009); Farhi and Gabaix (2020); Kroft, Laliberté, Leal-Vizcaíno, and Notowidigdo (2020)) directly influenced the optimizing frictions literature (Kleven and Waseem (2013); Chetty (2012); Chetty, Friedman, and Saez (2013)), I view our work as directly contributing to the optimizing friction literature; I develop a bunching estimator that incorporates these optimizing frictions. The tax salience literature typically focuses on individuals misperceiving prices by not taking sales tax into account. Our bunching estimator is applicable to settings where non-linearities in the incentive schedule are present such as taxpayers facing marginal tax rates. In our model agents will face optimizing frictions about the location of the kink (or notch) point, which will result in "fuzzy" bunching (or notching).

Workers respond to tax rates, which affect how many hours they choose to supply and whether they choose to participate in the labor force. Estimating the elasticity of earnings with respect to the net of tax rate is a central problem in public economics. Much of the literature considers a continuous univariate distribution of agents who face a piece-wise linear schedule of incentives such as the one imposed by the Earned Income Tax Credit (EITC). Other work has shown that agents face "optimizing frictions" that prevent them from responding to the incentive schedule as implied by utility maximization (Chetty, 2012; Kleven and Waseem, 2013). Our contribution will be to build a micro founded model that incorporates these optimization frictions and to retrieve the elasticity of earnings even when no apparent bunching occurs.

The method of using bunching to estimate an elasticity is referred to as a bunching estimator. In the tax setting bunching estimators estimate the elasticity of reported earnings with respect to the net of tax rate. Bunching estimators began with Saez (2010), which was followed by several influential bunching estimator papers such as Kleven and Waseem (2013) and Chetty, Friedman, Olsen, and Pistaferri (2011). However, a limitation in this literature is that these estimators are based on sharp bunching which is not the case when optimizing frictions affect how agents bunch.

Our paper has two goals. First, I seek to explain why agents do not appear to optimize utility. I do this by introducing a micro foundation for a second distribution of heterogeneity of agents' optimizing friction that affects how close agents can bunch to the locations of kinks or notches in the incentive schedule. I estimate the effects that agents' observable characteristics have on the probability of not reporting at the kink by making these optimizing frictions a function of observable as well unobservable heterogeneity. Second, our method improves on the methods used to estimate the tax elasticity by controlling for optimizing frictions explicitly. This improvement addresses results from work that documents that estimates of the tax elasticity will be attenuated in the presence of optimizing frictions (Kostøl and Myhre, 2021).

Optimizing frictions are present in cases of bunching; often agents experience optimizing frictions in both kink and notch points. For example, Saez (2010, figure 3A) finds that individuals bunch at the first EITC kink but do not bunch at higher EITC kinks (reproduced in figure 1). The third kink point in figure 1 corresponds to a concave kink. At this point there should be a hole in the data (dominated region), but empirically I do not observe a perceptible change in the distribution. The existence of optimizing frictions are even more striking in most empirical studies of notches. Utility maximization subject to the true incentive schedule predicts that no individuals should report in a dominated region just to the right of the notch. Empirically, agents facing a notch are often observed in this dominated region Kopczuk and Munroe (2015, figure A.1) (reproduced in figure 2).

I introduce optimizing frictions as a friction that prevents an agent from bunching at a kink or notch point. These optimizing frictions could take a myriad of forms ranging from agents misperception about the location of the kink or notch point to labor market frictions that prevent a worker from controlling the exact hours that they supply (inflexible work hours). Agents in our model choose the optimal level of consumption and labor supply (consequently income) to maximize utility subject to their budget constraint. This constraint is a piece-wise function of tax rates and agents' optimizing frictions in relation to the location of where tax rates change. These optimizing frictions can be set to zero, which would result in agents having perfect knowledge of the tax schedule along with perfect flexibility on their labor supplied. In this way, I nest the prior literature that assumes agents face no optimizing frictions (Bertanha, McCallum, and Seegert, 2020). Agents will optimize utility with the incorrect constraint when they face optimizing frictions and potentially report non-optimal taxable income (labor supply).

The optimal reported income that results in our model is a function of the heterogeneous earnings ability and heterogeneous optimizing frictions. I can write the probability of observing reported earnings as a function of these two sources of heterogeneity and then develop the likelihood function. By making heterogeneous earnings ability and optimizing frictions functions of observable covariates and unobserved errors, I can simultaneously estimate the non-tax determinants of earnings and optimizing frictions while also recovering the elasticity. I then employ the expectation-maximization (EM) algorithm to maximize the likelihood and estimate all the parameters of the model. Previous work incorporating optimization frictions has often used a two step process that first cleans the data of the optimizing frictions before estimating the elasticity, but our method can do both in one step. I do not want to clean the data given that I are particularly interested in those optimizing frictions.

Recent work such as Bertanha et al. (2020) have shown that a kink does not identify an elasticity when the distribution of agents is non-parametric and continuous. The elasticity can be partially identified in the kink case with semi-parametric conditions that use covariates. By introducing uncertainty around the kink and having covariates determine agents' degree of uncertainty, I can better estimate the elasticity parameter.

The paper proceeds with an utility maximization model subject to a piecewise-linear constraint in section 2. I propose multiple estimation strategies in Section 3. I apply our model to the EITC case in and estimate elasticity of taxable income in Section 4. Section 5 concludes.

2 Optimization and the data generating process

I use the well-known public finance theoretical frameworks of Saez (2010) and Kleven and Waseem (2013) to motivate the data generating processes (DGPs) but I emphasize that the estimation method is applicable to these DGPs regardless of the underlying microeconomic model.

Each agent i maximizes utility, $U(C_i, Y_i)$, by jointly choosing the level of consumption C_i and labor supply that results in labor income Y_i. Formally, the microeconomic model is given by the following utility maximization problem:

$$\max_{C_i, Y_i} U(C_i, Y_i) = C_i - \frac{N_i^*}{1+1/\varepsilon}\left(\frac{Y_i}{N_i^*}\right)^{1+1/\varepsilon} \tag{1a}$$

subject to:

$$C_i = \mathbb{1}_{Y_i \leq K_i^*}[I_0 + (1-t_0)\,Y_i] + \mathbb{1}_{Y_i > K_i^*}\left[I_0 + (1-t_0)\,K_i^* - \Delta + (1-t_1)\,(Y_i - K_i^*)\right], \tag{1b}$$

in which $\mathbb{1}$ is the indicator function. The agent faces a budget constraint and may consume all of their labor income net of income taxes t_j for $j = \{0, 1\}$ plus an exogenous endowment or lump sum tax, I_0. The budget constraint is a piecewise-linear function with intercept I_0 and slope $1 - t_0$ if $Y_i \leq K_i^*$, but intercept $I_0 + K_i^*(1 - t_0) - \Delta$ with slope $1 - t_1$ if $Y_i > K_i^*$.

Ability, N_i^*, and slope change (kink) location K_i^*, are heterogeneous to each agent and known to the agent when they choose optimal consumption and income but are unknown to the econometrician. Importantly, I do not take a stand on the theoretical source of heterogeneity in N_i^* and K_i^*. Instead of imposing those interpretations on the theatrical model, I will let estimates from the data inform our interpretation of heterogeneity. Heterogeneity in earning ability could, for example, be related to industry of employment or a measure of human capital possessed by the agent. Heterogeneity in the kink location, could arise, for example, from agents not knowing the intricacies of the true tax schedule.

As in much of the public finance literature and for simplicity, Equation (1) assumes that consumption prices and wages are both equal to one. That assumption implies labor income

Y_i is equal to L_i and the value of consumption C_i is equal to the quantity consumed Q_i. Appendix A.1 solves Equation (1) for any prices and wages and shows this assumption does not affect the DGP.

2.0.1 Data generating process

The solutions to the optimization problem in Equation (1) provides the DGP. The specific form of that solution, however, depends on which of the four possible piecewise linear tax schedules the agent faces given by

$$s_0 \geq s_1 \text{ and } \Delta \geq 0, \qquad (2a) \qquad\qquad s_0 \geq s_1 \text{ and } \Delta \leq 0, \qquad (2b)$$

$$s_0 \leq s_1 \text{ and } \Delta \leq 0, \qquad (2c) \qquad\qquad s_0 \leq s_1 \text{ and } \Delta \geq 0. \qquad (2d)$$

in which I define $s_0 = \ln(1 - t_0)$ and $s_1 = \ln(1 - t_1)$ and at least one of the inequalities in each expression is strict. The general solution to Equation (1) for tax schedules given by Equations (2a), (2b), and (2c) is

$$y_i = \begin{cases} \varepsilon s_0 + n_i^* & \text{, if } \quad n_i^* < \underline{n}_i \left(k_i^*, \varepsilon, s_0, \Delta\right) \\ k_i^* & \text{, if } \quad \underline{n}_i \left(k_i^*, \varepsilon, s_0, \Delta\right) \leq n_i^* \leq \bar{n}_i \left(k_i^*, \varepsilon, s_1, \Delta\right) \\ \varepsilon s_1 + n_i^* & \text{, if } \quad n_i^* > \bar{n}_i \left(k_i^*, \varepsilon, s_1, \Delta\right), \end{cases} \qquad (3)$$

in which I define $y_i = \ln(Y_i)$, $n_i^* = \ln(N_i^*)$, $\underline{n}_i = \ln(\underline{N}_i)$, $\bar{n}_i = \ln(\bar{N}_i)$, and $k_i = \ln(K_i)$. The elasticity of reported income with respect to one minus the tax rate when the solution is interior is ε. The thresholds that determine the reporting cases for the agents are functions of the tax rates and the agent's heterogeneous kink locations. If an agent's log ability, n_i^*, is lower than $\underline{n}_i \left(k_i^*, \varepsilon, s_0, \Delta\right)$, then the agent will supply labor that will place them below their individual kink location, k_i^*. Similarly, an agent of ability greater than $\bar{n}_i \left(k_i^*, \varepsilon, s_1, \Delta\right)$ will supply labor that will place them above their individual kink location. For values of $n_i^* \in \left[\underline{n}_i \left(k_i^*, \varepsilon, s_0, \Delta\right), \bar{n}_i \left(k_i^*, \varepsilon, s_1, \Delta\right)\right]$ inside the individual's bunching interval, the agent's indifference curve is never tangent to the budget constraint. Instead, the agent supplies the non-interior solution $y_i = k_i^*$.

Thresholds are defined by the abilities that make agents indifferent. An agent with ability $\underline{n}_i(k_i^*, \varepsilon, s_0, \Delta)$ is indifferent between earning income $y_i = \varepsilon s_0 + \underline{n}_i(k_i^*, \varepsilon, s_0, \Delta)$ and $y_i = k_i^*$. Likewise, an agent with ability $\bar{n}_i(k_i^*, \varepsilon, s_1, \Delta)$ is indifferent between earning income $y_i = \varepsilon s_1 + \bar{n}_i(k_i^*, \varepsilon, s_1, \Delta)$ and $y_i = k_i^*$. Generally, these thresholds are implicit functions as shown in Appendix **??**.

Special cases of Equation (3) are well known in the literature. For example, Kleven and Waseem (2013) have Equation (3) with a tax schedule given by Equation (2a) and Saez (2010) has Equation (3) with a schedule given by Equation (2a) and restricted to $\Delta = 0$, which also allows the thresholds to be obtained as explicit functions, namely, $\underline{n}_i(k_i^*, \varepsilon, s_0) = k_i^* - \varepsilon s_0$ and $\bar{n}_i(k_i^*, \varepsilon, s_1) = k_i^* - \varepsilon s_1$.

When the tax schedule is given by Equation (2d), the solution to Equation (1) can be either Equation (3) or

$$y_i = \begin{cases} \varepsilon s_0 + n_i^* & \text{, if } \; n_i^* \leq \check{n}_i(k_i^*, \varepsilon, s_0, s_1, \Delta) \\ \varepsilon s_1 + n_i^* & \text{, if } \; n_i^* > \check{n}_i(k_i^*, \varepsilon, s_0, s_1, \Delta). \end{cases} \tag{4}$$

in which I define $\check{n}_i = \ln(\check{N}_i)$, $\check{\bar{y}}_i = \ln\check{\bar{Y}}_i$, and $\underline{\check{y}}_i = \ln\underline{\check{Y}}_i$. There is only one threshold, $\check{n}_i(k_i^*, \varepsilon, s_0, s_1, \Delta)$, and it is defined as the ability that make the agent indifferent between incomes $\underline{\check{y}}_i = \varepsilon s_0 + \check{n}_i(k_i^*, \varepsilon, s_0, s_1, \Delta)$ and $\check{\bar{y}}_i = \varepsilon s_1 + \check{n}_i(k_i^*, \varepsilon, s_0, s_1, \Delta)$ as derived in Appendix **??**. Importantly, data generated by Equation (4) does not have a mass point in the distribution of y_i even when $k_i^* = k$.

For the special case of tax schedule Equation (2d) with $\Delta = 0$, the threshold can be obtained as an explicit function given by $\check{n}_i(k_i^*, \varepsilon, s_0, s_1) = k_i^* + \ln(\varepsilon + 1) + b(s_0, s_1, \varepsilon)$ in which $b(s_0, s_1, \varepsilon) = \ln\left[(\exp(s_0) - \exp(s_1)) / \left(\exp(s_0)^{\varepsilon+1} - \exp(s_1)^{\varepsilon+1}\right)\right]$.

In the special case of tax schedule of Equation (2d) with $\Delta = 0$ and $k_i^* = k$, Bertanha et al. (2020) show the simple formula, $\hat{\varepsilon} = \left(\check{\bar{y}}_i - \underline{\check{y}}_i\right) / (s_1 - s_0)$, non-parametrically point-identifies ε because $\check{\bar{y}}_i$ and $\underline{\check{y}}_i$ are observed as the start and end, respectively, of a zero mass region in the distribution of y_i. When $k_i^* \neq k$, this non-parametric identification scheme is unavailable because $\check{\bar{y}}_i$ and $\check{\underline{y}}_i$ are not observed in the data.

When the tax schedule is given by Equation (2d), Equation (3) will be optimal for some combinations of Δ and ε while Equation (4) will be optimal for other combinations. As such, both Equations (3) and (4) must be estimated and then measures of model fit should be used to select the correct DGP and attendant parameter values.

In addition to distorting reported income, I emphasize that heterogeneous $k_i^* \neq k$ imply that the budget constraint Equation (1b) will not provide the actual level of consumption —which is equivalent to after tax income —attained by the agent. Instead, agents optimize considering their own kink location, k_i^*, but then pay taxes according to the tax code with the true location, k. As such, actual consumption is given by Equation (1b) with k replacing k_i^* but with the choice of income by agents given by the relevant optimal choice given by Equation (3) or Equation (4).

In the public finance literature, differences between data from Equation (3) or Equation (4) and that same equations with $k_i^* = k$ are often attributed to "optimizing frictions." I will sometimes refer to heterogeneous location k_i^* using that terminology.

2.0.2 Graphical depiction of solutions

Figure 3 provides graphical intuition for how agents' optimizing frictions affect the optimal income to report when they face a the tax schedule defined by $\Delta = 0$ and Equation (2a) , which is the setting considered by Saez (2010). I call this case a "convex kink." I provide the graphical intuition for this case in levels, denoted in capital letters, which contrasts with the solution from Equation (3) that is clearest in logs, denoted in lower case letters.

Figure 3a shows the convex kink without optimizing frictions and Figure 3b shows the convex kink with optimizing frictions, K_i^*, along with the relevant indifference curves. These indifference curves correspond to two individuals, one with the lowest ability, $\underline{N}(K, \varepsilon, \Delta)$, and the other with the highest ability, $\bar{N}(K, \varepsilon, \Delta)$, that report the same level of income, $Y_i = K$, in each case.

Figure 3c and Figure 3d provide histograms of income generated by optimizing behavior defined by Equation (3) and a distribution of N_i^* from a log-normal distribution with mean zero and standard deviation one. As is well known, the kink generates a mass point at K. However, the the shape of that mass point depends on whether or not one allows for optimizing frictions. In the case without optimizing frictions, Figure 3c, I observe sharp bunching where all the mass point is exactly at the kink point. In the case with optimizing frictions, Figure 3d, I observe fuzzy bunching where the mass point is diffused around the kink point. In a scenario with more optimization frictions, the diffusion could be large enough to eliminate the mass point at the kink. For example, the second kink point in Figure 1 does not exhibit a mass point.

Figure 3b shows the indifference curve of $\overline{N}$ (light blue line) and the indifference curve of $\underline{N}$ (light green line). Both of these agents will choose to bunch at the kink point, $K = 3$. Let $\tilde{N}$ denote an individual with ability between $\overline{N}$ and $\underline{N}$. All agents with ability between $\overline{N}$ and $\underline{N}$ including $\tilde{N}$ will optimally choose to bunch at the kink point, K, when there are no optimizing frictions are present. However, if $\tilde{N}$ faces optimizing frictions such that the location of the kink point with optimizing frictions moves from 3 to 7, $K^* = 7$, then their budget constraint would shift from the solid black line to the dashed black line. When $\tilde{N}$ solves Equation (3), they choose to supply labor such that they will receive an income around 5. The infeasible indifference curve for $\tilde{N}$ is shown by pink line. However, that consumption-income pair is not feasible, so $\tilde{N}$ will shift down to the original budget constrain (solid black line). This will result in $\tilde{N}$ receiving lower utility than if they had bunched at the kink point, $K = 3$.

Even though optimizing frictions such as the ones experienced by $\tilde{N}$ potentially lead agents to a lower indifference curve, the model in Figure 3b is more akin to behaviors exhibited by agents such as taxpayers bunching at EITC kink points than the model in Figure 3a. Even in the example of sharp bunching used by Saez (2010) of the first EITC kink point, Figure 1, there is diffused mass around the first kink point. In a model without optimizing frictions such as the one in Figure 3a, agents always attain the highest utility level possible, which results in all individuals with ability levels between $\overline{N}$ and $\underline{N}$ to bunch exactly at the kink point. This is unrealistic because agents could face a multitude of optimizing frictions. For example, taxpayers in Figure 1 may use a heuristic to recall the location of the first kink point. Instead of recalling the location of the first kink point as \$8,580, taxpayers may recall the location of the first kink point as \$8,600.

With only unobserved ability, agents are still optimizing correctly and should be bunching at the kink point, K, when they have ability between $\overline{N}$ and $\underline{N}$. Once taking optimizing frictions into account, agents become more dispersed around the kink points as reflected by Figure 3d. If individuals face optimizing frictions where K^* is to the right of K, then this would lead individuals to potentially work more and end up at a lower indifference curve as shown in Figure 3b. Similarly, if individuals face optimizing frictions where K^* is to the left of K, then the individual could work less than the optimal amount. It is important to note that optimizing frictions do not unambiguously change agents' allocation of consumption and income. An agent could fall within the bunching interval both with and without optimizing frictions. It could also be that optimizing frictions do not affect the agent's optimal allocation. For example, if the agent is to the left of the kink point and has optimizing frictions where the perceived kink is further to the right, then the agent will remain at the same allocation.

2.1 Model solution for a concave kink

2.1.1 Graphical depiction of solution for a concave kink

Figure 4 provides graphical intuition for how agents' optimizing frictions affect the optimal income to report when they face a concave kink. This intuition is clearest in levels, denoted in capital letters, but the solution from Equation (4) is clearest in logs, denoted in lower case letters.

Figure 4a shows the concave kink without optimizing frictions and Figure 4b shows the concave kink with optimizing frictions ($K_i^* = K$) along with indifference curves. These indifference curves correspond to the individual, $\check{N}$, that is indifferent at reporting at $\underline{Y}$ and

$\bar{Y}$. In Figure 4a, an $\check{N}$ will be tangent at $\underline{Y}$ (1.68) and $\bar{Y}$ (4). An individual with ability less than $\check{N}$ will report income weakly less than $\underline{Y}$. Conversely, an individual with ability more than $\check{N}$ will report income weakly more than $\bar{Y}$. This results in a hole in the distribution between $\underline{Y}$ and $\bar{Y}$.

Figure 4c and Figure 4d provide histograms of reporting income that would be observed by this optimizing behavior and a distribution of N_i^* from a normal distribution with mean zero and standard deviation one. When there is no optimizing frictions, the concave kink generates a hole in the distribution, which is a region where no individual will choose to supply labor that will result in an income between $\underline{Y}$ and $\bar{Y}$. However, the the size and potential existence of said hole in the distribution depends on whether or not one allows for optimizing frictions. In the case without optimizing frictions, Figure 4c, I observe a hole in the distribution that corresponds to the difference between $\underline{Y}$ and $\bar{Y}$. In the case with optimizing frictions, Figure 4d, I see the hole disappear. There is still less mass between $\underline{Y}$ and $\bar{Y}$, but that depends on the elasticity, tax rates, and optimization frictions.

Figure 4b shows the indifference curve of $\check{N}$ (dotted blue line) when they face no optimization frictions or face optimization frictions such that $K^* = K = 3$ and the indifference curve of $\check{N}$ (dashed pink line) when they face optimization frictions such that $K^* = 7$. Both of these agents will choose to bunch at the K^*. However, $\check{N}$ with $K^* = 7$ will be tangent at $Y = 3.92$ and $Y = 9.3$. For this agent this corresponds to their $\underline{Y}$ and $\bar{Y}$, respectively. If this agent chooses to report income at their $\underline{Y}$ of 3.92, then they will end up in the region where there was a hole in Figure 4a. By definition, $\check{N}$ with $K^* = 7$ is indifferent between reporting income at 3.92 and 9.34. With these optimizing frictions, $\check{N}$ will end up at a lower indifference curve than if they were optimizing under no optimization frictions.

With only unobserved ability, agents are still optimizing in a manner that results in them reaching the highest utility possible. This leads all agents to choose to supply labor such that their income is weakly less than $\underline{Y}$ or weakly more than $\bar{Y}$. Once taking optimizing frictions into account, agents have individual levels of $\underline{Y}$ and $\bar{Y}$. This leads to more mass to the right of $\underline{Y}$ and to the left of $\bar{Y}$ as reflected by Figure 4d. In the scenario with no optimization friction and all agents have a homogeneous $\underline{Y}$ and $\bar{Y}$, which results in there being no mass between $\underline{Y}$ and $\bar{Y}$. It could also be that optimizing frictions do not affect the agent's optimal allocation of labor to supply. For example, if the agent has ability marginally less than $\check{N}$ and face optimization frictions such that $K^* > K$, then the agent will remain at the same allocation.

3 Estimation

3.1 General estimation problem

The optimal behavior by the agent given optimizing frictions k_i^* for convex and concave budget constraints provided by Equations (3) and (4) are the two DGPs I will use for estimation. The equations are closely related and our estimation methodology encompass both of them. Our estimation methodology applies to any DGP that takes the form of Equations (3) or (4) and the specific micro-economic theory outlined above is not a necessary condition for estimation.

The main challenge for estimation created by the introduction of optimizing frictions, k_i^*, is that observations of y_i are not partitioned into being above, below, and at the kink for Equation (3), or being below or above the kink in Equation (4). In particular, this feature of the data precludes using the methods developed in Bertanha et al. (2020).

Our approach is to use the EM algorithm as the method for recovering the maximum likelihood estimates from a missing-data problem. One useful way to think of our problem is as a missing-data problem. For example, each reported income observation is the agent choosing to be below, at, or above the kink location, but I do not know with certainty to which group that observation belongs when there is heterogeneity in location, k_i^*. As such, I are missing data about which group the agent belongs —and groups are defined by the different budget constraint slopes that the agent faces given optimizing frictions as described in Sections 2.0.2 and 2.1.1. Estimation will need to estimate membership in each group and estimate the parameters that determine the income level reported within each group. Ruud

(1991) shows applications of the EM algorithm to missing-data problems such as ours by writing a latent model.

In our setting, it I will define three latent variables given by,

$$y_0^* = \varepsilon s_0 + n^*, \qquad k^*, \qquad y_1^* = \varepsilon s_1 + n^*. \tag{5}$$

These latent variables are the latent location, k_i^*, income below that location, y_0^*, and income above that location, y_0^*.

The microeconomic theory and estimation methodology presented above is general to many problems outside of public finance and outside of the response to taxes. Before applying these methods to a specific dataset, I make a few more assumptions. I assume that ability, n_i^*, and location, k^*, are given by

$$n_i^* = X_i'\beta + \nu_i \quad \text{(6a)} \qquad \text{and} \qquad k_i^* = k + Z_i'\gamma + \psi_i, \quad \text{(6b)}$$

in which X_i and Z_i are observable characteristics. The true location of the kink is k. The error terms ν_i and ψ_i are i.i.d. unobservable errors with a bivariate normal distribution given by

$$\begin{pmatrix} \nu_i \\ \psi_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\nu^2 & \rho\sigma_\nu\sigma_\psi \\ \rho\sigma_\nu\sigma_\psi & \sigma_\psi^2 \end{pmatrix}\right]. \tag{7}$$

The Expectation-Maximization (EM) algorithm is an iterative method that recovers the Maximum Likelihood Estimates (MLE) of the parameters of a latent variable model. The EM algorithm was developed by Dempster et al. (1977), which has approximately 64,000 citations on Google Scholar as of writing. The EM algorithm has wide application across many areas of science and has many other interpretations, not least of those is the cross-entropy loss in machine-learning.

3.2 The likelihood

Define $\Omega = (\varepsilon, \theta)$ to include the tax elasticity, ε, and parameters of the distribution of ability and locations $g\left(n_i^*, k_i^* \mid \theta\right)$. The likelihood of observing any given level of income for an individual is $L\left[\Omega \mid y_i, s_0, s_1, k\right] = \sum_{h=1}^{H} f_{ih} P_{ih}$. For the convex kink case, this sum is over the $H = 3$. In the concave kink case, this sum is over the $H = 2$. For readability, I abuse notation by reusing f_{ih} and P_{ih} for both convex and concave settings but it is important to note that, for example, f_{i2} from the convex case is certainly different from f_{i2} in the concave case.

The likelihood of observing an $M \times 1$ vector of M agents' incomes y is then the product of the agents' individual likelihoods given by

$$L\left[\Omega \mid y, s_0, s_1, k\right] = \prod_{i=1}^{M} \sum_{h=1}^{H} f_{ih} P_{ih}. \tag{8}$$

One can use the traditional direct log-likelihood estimation to solve for the parameters of interest by taking the log of Equation (8) and then using optimization technique such as gradient descent methods.

3.2.1 EM Algorithm Following Ruud Equations

An alternative to the direct log-like estimation procedure is the direct maximization with the EM algorithm following Ruud (1991). This procedure follows four steps broad steps. First, choose initial values for the parameters of interest: (ε, θ). Second, calculate the expectation steps based on the latent variables. Third, calculate the maximization step where the parameters get updated. The fourth and final step is to iterate between steps two and three until the difference in parameter values between iterations is below a pre-specified tolerance level. One of the nice features of the EM algorithm is that these iterative steps are guaranteed to converge to a local maximum.

The second step with the conditional expectation, conditional variance, and covariance terms are based on the latent variables defined in Equation (5). The conditional expectation terms of these three latent variables when I condition on the observed income level are as follows:

$$\begin{aligned} E\left[y_0^* \mid y\right] &= E\left[y_0^* \mid y, y < k^*\right] P\left[y < k^* \mid y\right] \\ &+ E\left[y_0^* \mid y, y = k^*\right] P\left[y = k^* \mid y\right] \\ &+ E\left[y_0^* \mid y, y > k^*\right] P\left[y > k^* \mid y\right] \end{aligned} \tag{9}$$

$$\begin{aligned} E\left[k^* \mid y\right] &= E\left[k^* \mid y, y < k^*\right] P\left[y < k^* \mid y\right] \\ &+ E\left[k^* \mid y, y = k^*\right] P\left[y = k^* \mid y\right] \\ &+ E\left[k^* \mid y, y > k^*\right] P\left[y > k^* \mid y\right] \end{aligned} \tag{10}$$

$$\begin{aligned} E\left[y_1^* \mid y\right] &= E\left[y_1^* \mid y, y < k^*\right] P\left[y < k^* \mid y\right] \\ &+ E\left[y_1^* \mid y, y = k^*\right] P\left[y = k^* \mid y\right] \\ &+ E\left[y_1^* \mid y, y > k^*\right] P\left[y > k^* \mid y\right] \end{aligned} \tag{11}$$

The conditional variance and covariance terms are as follows:

$$\begin{aligned} V\left[y_0^* \mid y\right] = V\left[y_1^* \mid y\right] = V\left[n^* \mid y\right], V\left[k^* \mid y\right], \\ C\left[y_0^*, k^* \mid y\right] = C\left[y_1^*, k^* \mid y\right] = C\left[n^*, k^* \mid y\right], \\ C\left[y_0^*, y_1^* \mid y\right] = V\left[n^* \mid y\right] \end{aligned}$$

in which $C[\cdot]$, denotes the covariance and some of the second moments are equal because y_0^* and y_1^* only differ by a constant. The conditional first moments are further described in Section C.2 of the Appendix. The conditional variance and conditional covariance terms are further described in Section C.3 of the Appendix.

In the the third step, I maximize the likelihood function for the latent dependent variable as described in Ruud (1991). Specifically, I use non-linear seemingly unrelated regressions

(NLSUR) of $E[y_0^* \mid y]$ and $E[y_1^* \mid y]$ on X. These regressions update our parameter estimates of β and ε. I regress $E[k^* \mid y]$ on Z to update our parameter estimate of γ.

Intuition for the EM algorithm for a convex kink is as follows. A small observed income level is likely, but not certainly, from the agents having low ability and having a kink location such that they choose to face s_0. Thus the probability that the observation should be classified as coming from $y_i = s_0 + n_i^*$ is relatively high. Computing similar probabilities of classifications for all the observations and then holding those fixed, I can re-estimate the parameters. Those new and more precise parameter estimates can then improve the classification of observations, which can again be used to estimate even more precise parameters. This approach nests the setting without optimizing frictions ($k^* = k$) because observations are either below, at, or above k so classification is directly observed instead of estimated.

3.2.2 Numerical optimization

Solving the maximization step of the EM algorithm requires a numerical optimization algorithm. Many of these algorithms can be classified broadly into three types based on the number of derivatives of the objective function that are used. These types are: 1) first derivative methods (for example, Gauss-Newton, Levenberg (1944)-Marquardt (1963), and (Berndt, Hall, Hall, and Hausman, 1974, BHHH), 2) second derivative methods (for example, Newton-Raphson (NR), Broyden (1970)-Fletcher (1970)-Goldfarb (1970)-Shanno (1970) (BFGS), and Goldfeld, Quandt, and Trotter (1966), and 3) derivative free methods (for example, (Nelder and Mead, 1965, Simplex) and (Goffe, Ferrier, and Rogers, 1994, Simulated Annealing).

I undertake the maximization step of the EM algorithm using the BHHH algorithm. The NR algorithm has a quadratic convergent rate but requires computing the inverse of the second derivative (Hessian) of the likelihood. The BHHH is similar to the NR but uses the outer product of the gradient (OPG) of the likelihood in place of the Hessian. This substitution is only valid when the function being optimized is a likelihood function because the expectation of the OPG is the information matrix, which is equal to the negative of the Hessian under general regularity conditions. The BHHH algorithm has the advantages of being computationally simple because only the first derivative of the likelihood is needed and it is guaranteed to converge to a local minimum because the OPG is always positive semi-definite. One BHHH disadvantage is that it can be slow to converge to the optimum because it has an approximately linear rate of convergence near the solution. In our application in Section 4, I calculate the gradient analytically to further improve the precision and speed of each BHHH maximization step.

3.2.3 Inference

The EM alogirthm recovers the MLE parameters which are unbiased, efficient (they achieve the Cramér–Rao lower bound), and asymptotically normal. I report robust standard errors for the parameter estimates using the typical "sandwich" form of the likelihood (extensions to correct for clusters are straightforward). These standard errors can be used for inference individually on the parameters of the model.

I can use the likelihood ratio (LR) to test hypothesis on many parameters jointly. I are particularly interested in testing the null hypothesis that the determinants of optimizing frictions do not belong in the model or if I can reject that null hypothesis. The LR test is performed by estimating two models and checking for a statistically different fit between the models. Estimating two models is computationally expensive, however, and the Wald test approximates the LR test and can be calculated after estimating the model only once. I report both Wald and LR test results. All inference likelihoods or truncated likelihoods with parameters estimated using augmented likelihood.

4 Application to EITC

4.1 Data

I use annual cross-section data from the Individual Public Use Tax Files, constructed by the Internal Revenue Service (IRS) initially used by Saez (2010). The data are individual tax returns for each year 1995 to 2004 that have been inflation-adjusted to 2008 dollars using the IRS inflation parameters. The data include basic demographic information and sampling weights which allow interpretation of any estimates as being based on the population of U.S. individual income tax returns.

During 1995 to 2004, the EITC schedule for taxpayers with one child provided taxpayers that earn between 0 and \$8,580 a marginal subsidy of -34 percent. The maximum subsidy is $\$2,917 = \$8,580 \times 0.34$. Between \$8,580 and \$15,740 of reported income, the marginal EITC rate is zero, between \$15,740 and \$33,995, the marginal rate is 0.11. For reported incomes higher than \$33,995, the marginal rate is zero again. The change in marginal rates generates a convex kink at each of \$8,580 and \$15,740 and a concave kink at \$33,995.

I use two sub-samples for estimation. What I will call the "self-employed" sub-sample includes taxpayers that report self-employment income but no wage income and have one child. The "wage earners" sub-sample includes taxpayers that report wage income but no self-employment income and have one child. I do not restrict the sample based on other sources of income such as capital gains, dividends, or pension benefits.

Figure 5a displays a histogram of earnings for the self-employed sample. The EIC amount (based on the EITC schedule) for the taxpayer is depicted by the black dashed line. The solid red vertical lines depict the locations of the convex or concave kink in the tax schedule. The literature commonly refers to the first kink point at \$8,580 as an example of sharp bunching, but the bunching mass is diffused around the convex kink point. This histogram is more similar to figure 3d than figure 3c. Within the framework of our model these self-employed tax filers are facing optimization frictions that prevent them from bunching exactly at \$8,580.

However, the optimization frictions for the self-employed are not uniform within this group of tax filers; one covariate where I see a large difference in optimization frictions is marriage. In figure 5b, the solid black kernel density is self-employed not married taxpayers and the dashed black line is the self-employed married taxpayers. Figure 5b shows that not being married amongst tax filers who only report self-employment income dramatically increases the probability of reporting at the $8,580 kink relative to being married, while leaving much of the rest of the distribution unchanged. The similarity of the kernel densities away from the kink suggests that optimization frictions are playing a large role. If the married distribution had a distribution shifted to the right, then I would expect less bunching given that tax filers would be less willing to bunch at the kink point because it would represent a larger loss in income.

Figure 5c shows the percentage of taxpayers that are married with one dependent child and only self-employment income peaks very near the first kink and rapidly drops just after it. From prior work I know that many government programs such as the EITC have a marriage penalty, Alm, Dickert-Conlin, and Whittington (1999). These marriage penalties affect whether a spouse decides to join the labor market Eissa and Hoynes (2004) to increase their EIC. However, most of these effects of the EITC on the marriage market is on the extensive margin whether a spouse decides to work. Given that, it's surprising that the share of not married tax filers almost peaks near the first kink point. One might expect that the married households have a stronger incentive to bunch at the kink point.

Similar to figure 5a, figure 5d displays a histogram of earnings, but for the wage earners subsample. Unlike figure 5a, figure 5d does not display sharp nor diffused bunching at any of the three kink points. Given that the first and second kink points are convex kinks, I might expect to see bunching in one or both of these locations. The third kink point at $33,995 corresponds to a concave kink where I would expect to see less mass in the earnings distribution. The third kink point is a concave kink because instead of the marginal tax rate increasing at $33,995, it decreases. Theoretically, I should see a hole in the distribution at the concave kink.

However, the optimization frictions for the wage earners are also not uniform within this group of tax filers; one covariate where I see a large difference in optimization frictions is whether a tax filer receives social security benefits. Figure 5e shows that receiving social security benefits amongst tax filers who only report wage income dramatically increases the probability of reporting at the $33,995 kink compared to not receiving social security benefits, while leaving much of the rest of the distribution unchanged. The solid black kernel density corresponds to wage earners without social security benefits and the dashed black line corresponds to wage earners with social security benefits. Wage earners with social security benefits appear to face larger optimization frictions given that they have less control over their social security benefits than their wage income. Receiving social security benefits is an important covariate for wage earning taxpayers and has the potential to be an important covariate in explaining why tax filers fail to move away from the third kink point.

Because the tax rate decreases at the $33,995 theory on concave kinks without frictions predicts that individual would try to report away from the that kink. At first glance if one were to only look at Figure 5d, then one would conclude that taxpayers do not report away from the concave kink as theory would predict. Figure 5d does not depict the ‘‘hole’’ in the distribution but conditioning on receipt of social security benefits, paints a different picture. Figure 5e shows that receiving social security benefits dramatically increases the probability of reporting at the $33,995 kink compared to not getting benefits. The solid black kernel density is wage-earners who also receive social security benefits and the dashed black line is wage earners that do not receive social security benefits. Individuals that do not get benefits could have more control of their income and find it easier to move away from the concave kink.

Figure 5f shows the percentage of taxpayers without benefits declines as it gets closer to the $33,995 kink and then rapidly recovers afterwards. This decline in the fraction without benefits echoes the region of zero mass that I would predict at a concave kink in a model without frictions. This example with social security benefits highlights the importance of our model in taking optimization frictions into account given that after conditioning on relevant covariates that affect optimization frictions, drastically different patterns are observed. Covariates on the ability distribution matter as Bertanha et al. (2020) showed, but so do covariates on the optimization frictions distribution.

4.2 Estimation results

4.3 Convex taxes estimation results

In figure 6, I apply the full version of our model to one of the subsamples that Saez (2010) uses where sharp bunching is present. I look at self-employed tax filers with 1 child. I add several covariates to the the ability help our model determine where taxfilers fall in the ability distribution. More importantly I add three covariates to help our model determine where taxfilers fall in the optimzing frictions distribution: whether the taxfiler is married, whether the taxfiler received unemployment insurance benefits, and whether the taxfiler received social security. Our model performs well in matching the data, given the close fit of both PDFs. Intuitively, the counterfactual distribution is shifted to the right because if negative marginal tax rate continued beyond the first kink point, that would induce taxfilers to work more hours.

In table 1, I can compare the estimates of the taxable income with respect to the net-of-tax rate from the previous subsample of self-employed tax filers with 1 child. I find a higher elasticity estimate than Saez (2010). This difference can be explained in part by the introduction of heterogeneous optimization frictions. Taxfilers who may not appear to be bunching are bunching once I take the optimizing frictions into account. [1] Our elasticity estimates are about 0.73 and Saez's estimate are about 0.48.

In figure 7, I apply our model to another subsample that Saez (2010) uses where where "fuzzy" bunching is present. I look at wage-earners tax filers with 1 child using 22 covariates in the model. I use all 22 covariates in both the ability distribution and the optimization

[1] Even when I do not include any covariates, I find that the coefficient on γ is statistically significant, which implies that optimizing frictions matter even when sharp bunching is present.

frictions distribution. I use the full subsample and focus on the second kink point of the EITC ($15,740). Again our model closely fits the empirical distribution in the PDF. At first glance there is no bunching present, but with our model that accounts for optimizing friction I are able to retrieve an elasticity estimate and other covariates of interest. There only appears to be no bunching because optimizing frictions are larger for the second kink point. One optimizing friction is knowledge about the location of the kink point and taxfilers may be less aware of other kink points besides the first one in the EITC schedule given that the first kink point is when the negative marginal tax rates end.

In table 2, I can compare the estimates of the taxable income with respect to the net-of-tax rate from the previous subsample of wage-earners tax filers with 1 child using our model. Unlike Saez (2010) I find a large elasticity estimate of about 1.2 when he finds an estimate close to 0. I also provide the coefficient on the constant term on *gamma* as well as coefficient values on covariates on optimizing frictions. I illustrate 4 binary variables with the largest coefficients such as receiving a tax credit. I have standardized the variables, so the magnitudes of each of the gammas are comparable. The coefficient should be compare to the constant to see if tax filers with these characteristics are more likely to be to the left or right of the kink point when they bunch due to optimizing frictions. One should interpret the coefficient of 0.3 on have capital gains income as a 1 standard deviation increase in capital gains is associated with being 30% to the right of the kink.

Note that in our examples of applying our model to the EITC I looked at two cases: (1) self-employed workers with one child at the first kink point and (2) wage earners with one child at the second kink point. The bunching literature focuses on the first kink point where one sees the most excess mass (Saez (2010) and Chetty et al. (2013)). However, our model can estimate an elasticity at other kink points where an excess mass point may be less obvious. Not only are elasticity estimates away from the first kink more interesting, but they may also be preferred.

4.4 Determinants of concave optimizing frictions

Mortenson and Whitten (2020) show that approximately two-thirds of bunchers always bunch at the refund maximizing kink. This is problematic because our model as well as other models in the literature have the wrong utility function for agents (equation 1a). This type of behavior of refund maximizing occurs at the tax refund maximum, which is the first kink point of the EITC and occurs much more for self-employed individuals than wage earners; self-employed individuals can more easily misreport their true income than wage earners. I are staying within the framework of the literature by assuming the same utility function that is quasi-linear in consumption and isoelastic in labor. However, our second case with wage earners at the second case should draw less concern given that these bunchers are not located at the refund maximizing kink and are less likely to misreport their income.

5 Conclusion

When I run our model on the same sample as Saez (2010) where he observes sharp bunching, I are able to retrieve similar elasticity estimates to Saez (2010). I are also able to retrieve an elasticity estimate that are non-zero for other kink points with "fuzzy" bunching. Our model explicitly takes these optimizing frictions into account and estimates an elasticity for wage earners with one child in the second kink point. These taxpayers are bunching at second kink point, but due to optimizing frictions, they are less precise in bunching exactly at the kink point.

In addition to being able to derive elasticity estimates that incorporate optimizing frictions and increasing the set of kinks and notches where I can estimate an elasticity, I can estimate the impact of optimizing frictions directly and see whether they matter and if so by how much. I find that these additional covariates help determine where an agent falls in the distribution of optimizing friction. I find that about 13% of the variance in optimizing frictions is explained by our covariates in optimizing frictions distribution. In addition I can determine whether these covariates shift agents to the left or right of the kink or the notch.

I see our contribution as developing a model that allows researchers to estimate an elasticity even in scenarios without sharp bunching. Our model can be generalized and applied to non-tax situations where bunching occurs. I also provide the Stata package to allow other researchers to easily apply our model to their settings.

References

Alm, James, Stacy Dickert-Conlin, and Leslie A. Whittington (1999). "Policy watch: The marriage penalty." *Journal of Economic Perspectives*, 13(3), pp. 193--204. doi:10.1257/jep.13.3.193. URL https://www.aeaweb.org/articles?id=10.1257/jep.13.3.193.

Alvero, Adrien and Kairong Xiao (2020). "Fuzzy bunching." *Available at SSRN 3611447*.

Berndt, Ernst R., Bronwyn H. Hall, Robert E. Hall, and Jerry A. Hausman (1974). *Estimation and Inference in Nonlinear Structural Models*, pp. 653--665. NBER. URL http://www.nber.org/chapters/c10206.

Bertanha, Marinho, Andrew H. McCallum, and Nathan Seegert (2020). "Better Bunching, Nicer Notching." Working Paper 3144539, SSRN. doi:http://dx.doi.org/10.2139/ssrn.3144539.

Broyden, C. G. (1970). "The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations." *IMA Journal of Applied Mathematics*, 6(1), pp. 76--90. doi:10.1093/imamat/6.1.76. URL https://doi.org/10.1093/imamat/6.1.76.

Chetty, Raj (2012). "Bounds on Elasticities With Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply." *ECONOMETRICA*, 80(3), pp. 969--1018. doi:10.3982/ECTA9043.

Chetty, Raj, John N. Friedman, Tore Olsen, and Luigi Pistaferri (2011). "Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records." *Quarterly Journal of Economics*, 126(2), pp. 749--804.

Chetty, Raj, John N. Friedman, and Emmanuel Saez (2013). "Using Differences in Knowledge across Neighborhoods to Uncover the Impacts of the EITC on Earnings." *American Economic Review*, 103(7), pp. 2683--2721. doi:10.1257/aer.103.7.2683. URL http://www.aeaweb.org/articles?id=10.1257/aer.103.7.2683.

Chetty, Raj, Adam Looney, and Kory Kroft (2009). "Salience and taxation: Theory and evidence." *American Economic Review*, 99(4), pp. 1145--77. doi:10.1257/aer.99.4.1145. URL https://www.aeaweb.org/articles?id=10.1257/aer.99.4.1145.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum likelihood from incomplete data via the em algorithm." *Journal of the Royal Statistical Society Series B (Methodological)*, 39(1), pp. 1--38. URL http://www.jstor.org/stable/2984875.

Eissa, Nada and Hilary Williamson Hoynes (2004). "Taxes and the labor market participation of married couples: the earned income tax credit." *Journal of Public Economics*, 88(9), pp. 1931--1958. doi:https://doi.org/10.1016/j.jpubeco.2003.09.005. URL https://www.sciencedirect.com/science/article/pii/S0047272703001440.

Farhi, Emmanuel and Xavier Gabaix (2020). "Optimal taxation with behavioral agents."

American Economic Review, 110(1), pp. 298--336. doi:10.1257/aer.20151079. URL https://www.aeaweb.org/articles?id=10.1257/aer.20151079.

Fletcher, R. (1970). ``A new approach to variable metric algorithms.'' *The Computer Journal*, 13(3), pp. 317--322. doi:10.1093/comjnl/13.3.317. URL https://doi.org/10.1093/comjnl/13.3.317.

Goffe, William L., Gary D. Ferrier, and John Rogers (1994). ``Global optimization of statistical functions with simulated annealing.'' *Journal of Econometrics*, 60(1), pp. 65--99. doi:https://doi.org/10.1016/0304-4076(94)90038-8. URL https://www.sciencedirect.com/science/article/pii/0304407694900388.

Goldfarb, Donald (1970). ``A family of variable-metric methods derived by variational means.'' *Mathematics of Computation*, 24(109), pp. 23--26. URL http://www.jstor.org/stable/2004873.

Goldfeld, Stephen M., Richard E. Quandt, and Hale F. Trotter (1966). ``Maximization by quadratic hill-climbing.'' *Econometrica*, 34(3), pp. 541--551. URL http://www.jstor.org/stable/1909768.

Kleven, Henrik J. and Mazhar Waseem (2013). ``USING NOTCHES TO UNCOVER OPTIMIZATION FRICTIONS AND STRUCTURAL ELASTICITIES: THEORY AND EVIDENCE FROM PAKISTAN.'' *QUARTERLY JOURNAL OF ECONOMICS*, 128(2), pp. 669--723. doi:10.1093/qje/qjt004.

Kopczuk, Wojciech and David Munroe (2015). ``Mansion Tax: The Effect of Transfer Taxes on the Residential Real Estate Market.'' *AMERICAN ECONOMIC JOURNAL-ECONOMIC POLICY*, 7(2), pp. 214--257. doi:10.1257/pol.20130361.

Kostøl, Andreas R. and Andreas S. Myhre (2021). ``Labor supply responses to learning the tax and benefit schedule.'' *American Economic Review*, 111(11), pp. 3733--66. doi:10.1257/aer.20201877. URL https://www.aeaweb.org/articles?id=10.1257/aer.20201877.

Kroft, Kory, Jean-William P Laliberté, René Leal-Vizcaíno, and Matthew J Notowidigdo (2020). ``Salience and taxation with imperfect competition.'' Working Paper 27409, National Bureau of Economic Research. doi:10.3386/w27409. URL http://www.nber.org/papers/w27409.

Levenberg, Kenneth (1944). ``A method for the solution of certain non-linear problems in least squares.'' *Quarterly of Applied Mathematics*, 2(2), pp. 164--168. doi:https://doi.org/10.1090/qam/10666. URL http://www.jstor.org/stable/43633451.

Marquardt, Donald W. (1963). ``An algorithm for least-squares estimation of nonlinear parameters.'' *Journal of the Society for Industrial and Applied Mathematics*, 11(2), pp. 431--441. doi:10.1137/0111030. URL https://doi.org/10.1137/0111030.

Mortenson, Jacob A. and Andrew Whitten (2020). ''Bunching to maximize tax credits: Evidence from kinks in the us tax schedule.'' *American Economic Journal: Economic Policy*, 12(3), pp. 402--32. doi:10.1257/pol.20180054. URL https://www.aeaweb.org/articles?id=10.1257/pol.20180054.

Nelder, John A and Roger Mead (1965). ''A simplex method for function minimization.'' *The computer journal*, 7(4), pp. 308--313.

Papakonstantinou, Joanna M. and Richard A. Tapia (2013). ''Origin and evolution of the secant method in one dimension.'' *The American Mathematical Monthly*, 120(6), pp. pp. 500--518. URL https://www.jstor.org/stable/10.4169/amer.math.monthly.120.06.500.

Ruud, Paul A. (1991). ''Extensions of estimation methods using the em algorithm.'' *Journal of Econometrics*, 49(3), pp. 305--341. doi:https://doi.org/10.1016/0304-4076(91)90001-T. URL https://www.sciencedirect.com/science/article/pii/030440769190001T.

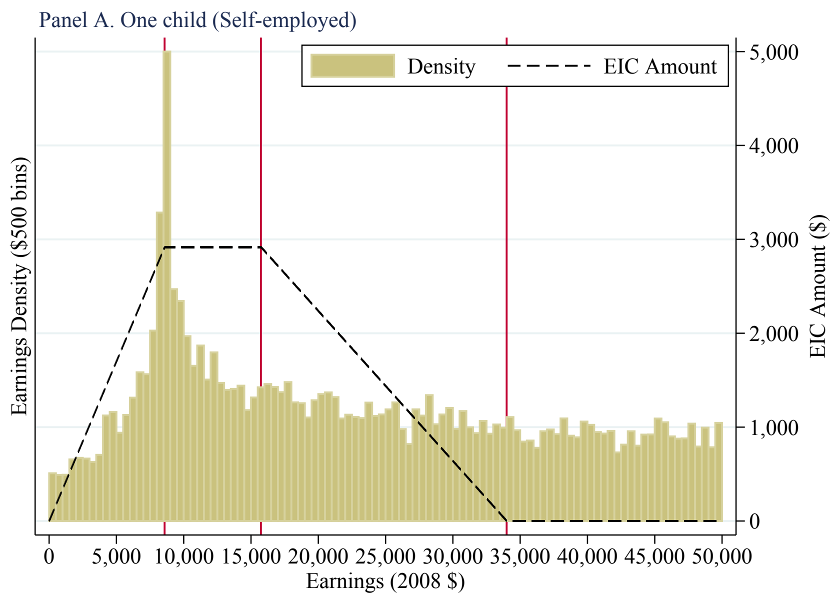
Saez, Emmanuel (2010). ''Do Taxpayers Bunch at Kink Points?'' *AMERICAN ECONOMIC JOURNAL-ECONOMIC POLICY*, 2(3), pp. 180--212. doi:10.1257/pol.2.3.180.

Scavo, T. R. and J. B. Thoo (1995). ''On the geometry of halley's method.'' *The American Mathematical Monthly*, 102(5), pp. 417--426. doi:10.1080/00029890.1995.12004594. URL https://doi.org/10.1080/00029890.1995.12004594.

Shanno, D. F. (1970). ''Conditioning of quasi-newton methods for function minimization.'' *Mathematics of Computation*, 24(111), pp. 647--656. URL http://www.jstor.org/stable/2004840.

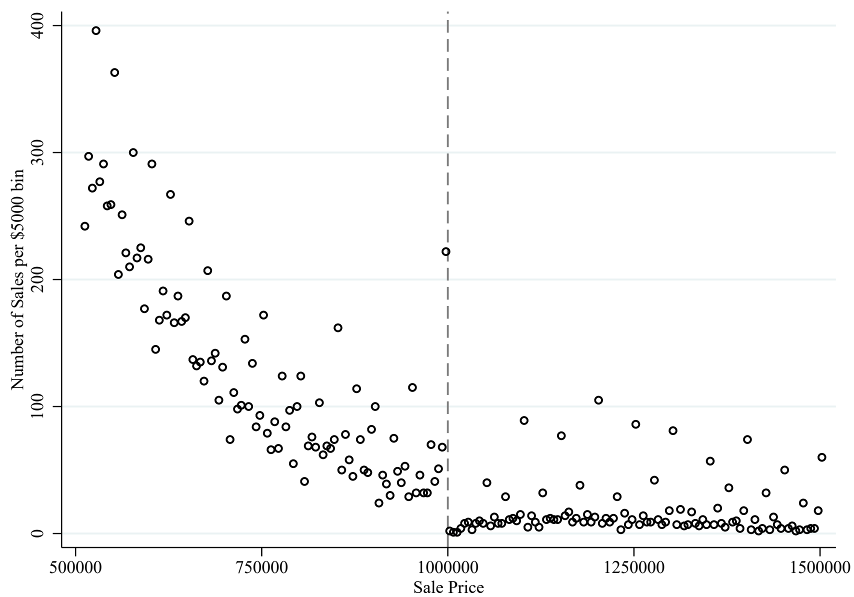
Figure 1: Lack of Bunching at Kinks

(a) Earnings Density Distributions and the EITC



Notes: This figure displays a histogram of earnings density (by $500 bins) for tax filers with one dependent child and who are self-employed. The EITC schedule is depicted by the black dashed line. When the black dashed line, changes slope, a kink point is formed. There are three kink points in this figure, which are depicted by red vertical lines. The literature commonly refers to the first kink point at $8,580 as an example of sharp bunching with no bunching apparent in the second or third kink points. However, even the first kink point is not a traditional example of sharp bunching given that the bunching mass is diffused around the kink point with the bins adjacent to the kink point having more mass than the counterfactual distribution would imply. Our model can account for this diffused bunching through optimizing frictions. This figure and data is similar to Saez (2010) where we also look at all years from 1995 to 2004 and inflate earnings to 2008 dollars using the IRS inflation parameters.

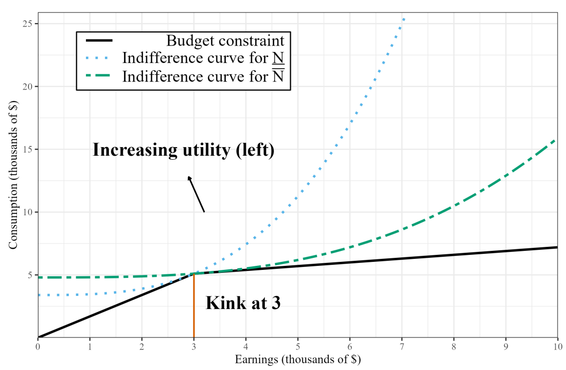
Figure 2: Mass in the Dominated Region after the Tax Notch



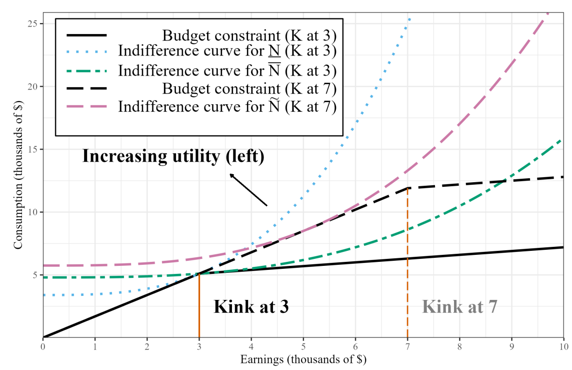
Notes: This figure is drawn from Figure 1 in Kopczuk and Munroe (2015) . In order to avoid a tax notch at $5000,000 where the tax rate changes from 1% to 1.425%, we exclude New York City from our analysis. We look at the remaining sales from New York state for 2004. The only tax notch these sales face are at $1,0000,000 where the tax rate changes from 0% to 1%. We also restrict to sales that arm's length transactions and one family year-round residence.

Figure 3: Convex kink with and without optimizing frictions

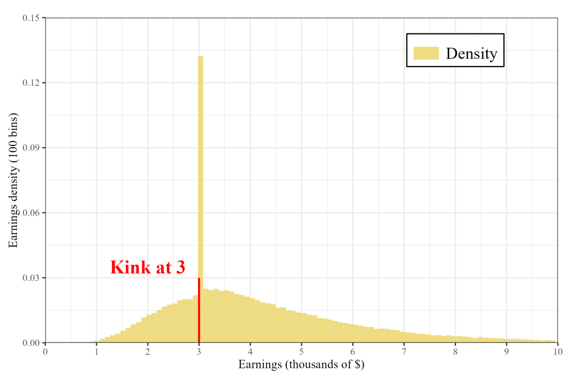
(a) Kink without Optimizing Frictions



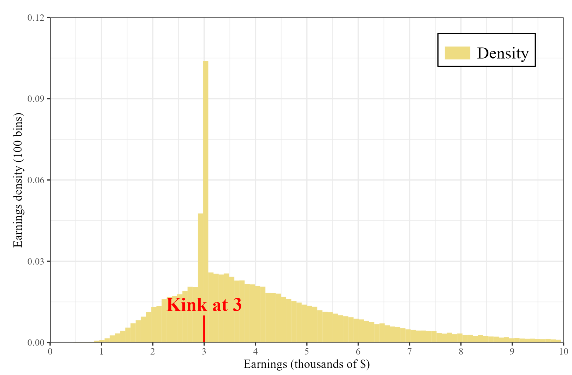
(b) Kink with Optimizing Frictions



(c) PDF of Kink without Optimizing Frictions



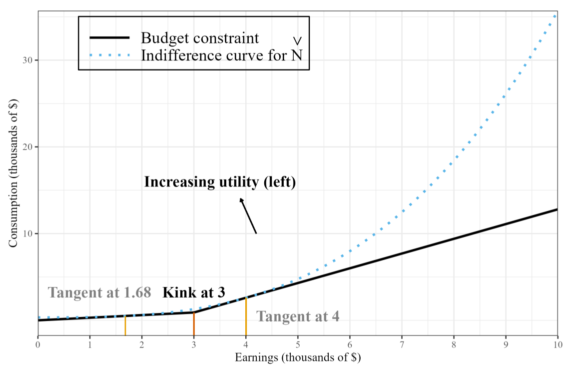
(d) PDF of Kink with Optimizing Frictions



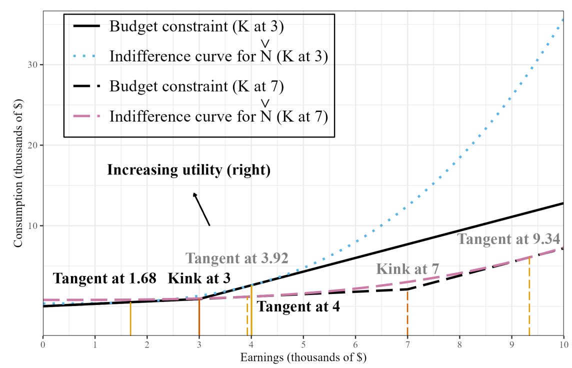
Notes: Panels A and B are related to convex kink points where panel A depicts two individuals without optimization frictions, while panel B depicts individual $\tilde{N}$ and how their decision varies depending on what optimization frictions they face. Panels C and D correspond to histograms of reporting that correspond to the optimizing behavior depicted in panels A and B, respectively.

Figure 4: Concave kink with and without optimizing frictions

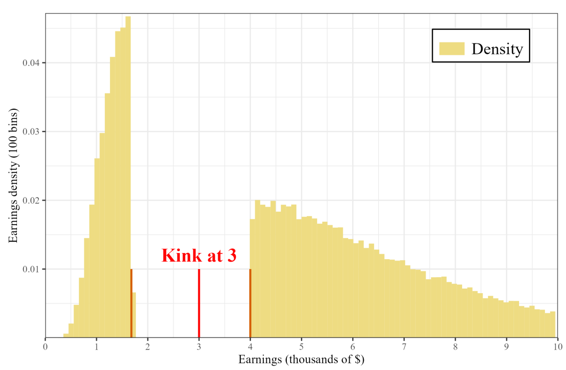
(a) Concave kink w/o frictions



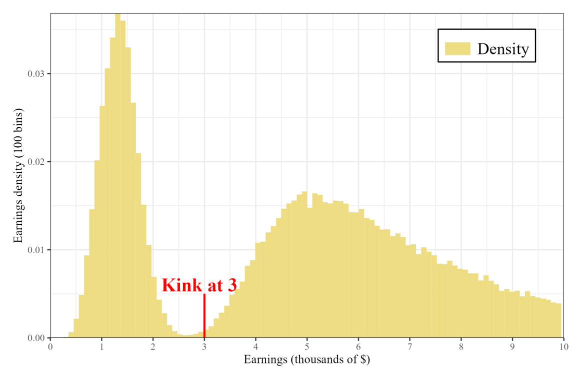
(b) Concave kink with frictions



(c) PDF of Concave kink w/o frictions



(d) PDF of Concave kink with frictions



Notes: Panels A and B are related to concave kink points where panel A depicts one individual without optimization frictions who is indifferent between reporting at the start and end of the zero reporting region. Panel B depicts an additional individual and how their decision varies depending on what optimization frictions they face. Panels C and D correspond to histograms of reporting that correspond to the optimizing behavior depicted in panels A and B, respectively.

Figure 5: Determinants that affect reporting at kinks

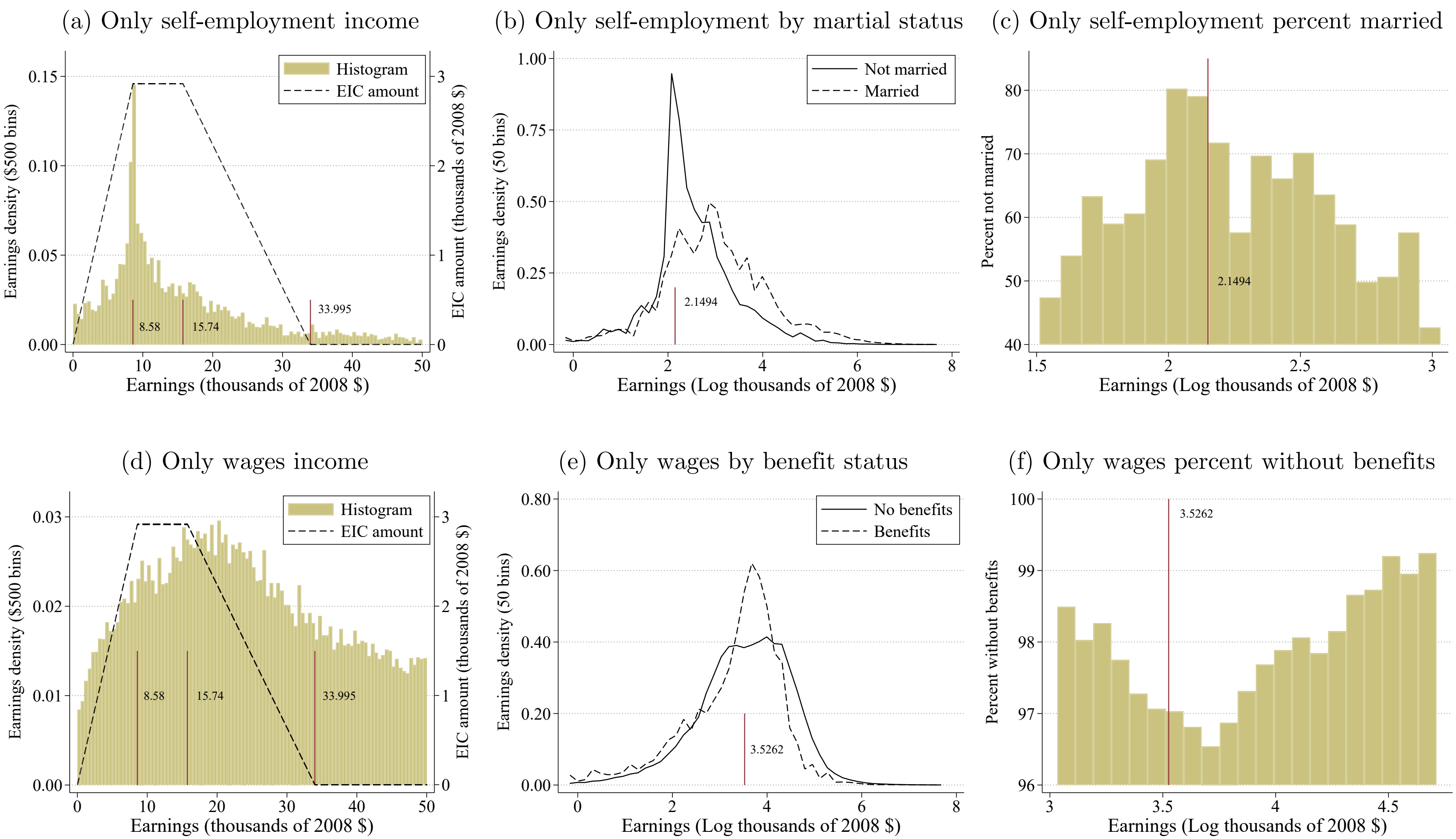
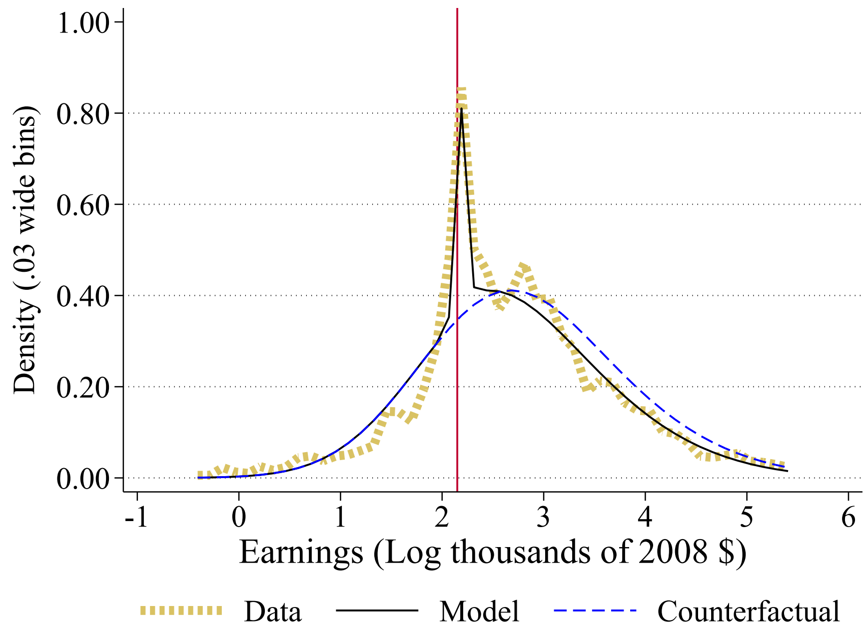


Figure 5 includes tax filers with one dependent child and who report only self-employment income and no wages (Figures 5a, 5b, 5c) or only wages and no self-employment income (Figures 5d, 5e, and 5f). There is diffuse bunching in the unconditional histogram of earnings at the $8,580 convex kink in Figure 5a but no missing mass at the $33,995 concave kink in Figure 5d. The black dashed line is the dollar amount of the EIC for reporting the income on the x-axis and the red vertical lines are two concave kinks created by the EITC at $8,580 and $15,740 and the one concave kink at $33,995. This figure and data is similar to Saez (2010) where we also look at all years from 1995 to 2004 and inflate earnings to 2008 $ using the IRS inflation parameters.

Figure 6: Self-employment income only, one child, \$8,580 kink, many covariates

(a) PDF

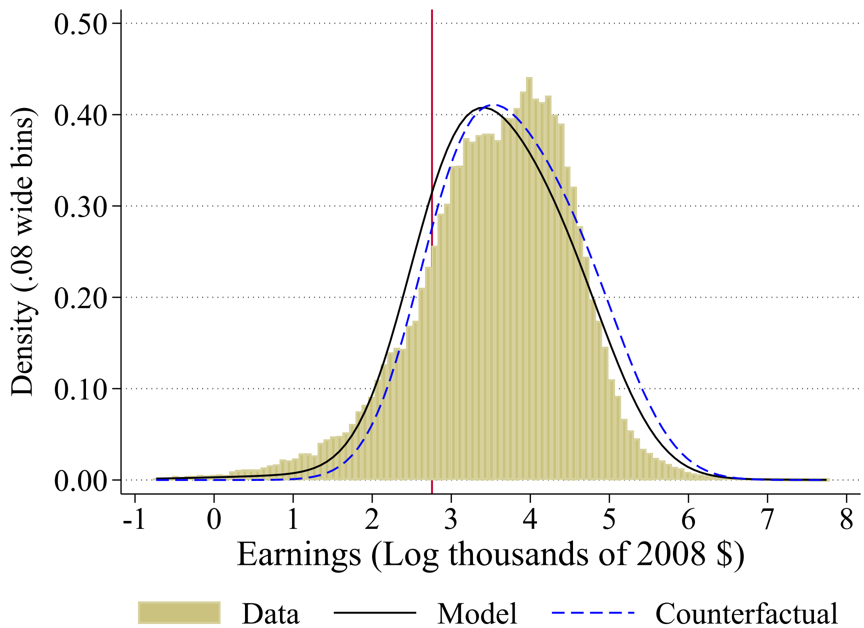


We

Notes: Figure 6 includes tax filers with one dependent child and who report only self-employment income and no wages. The dashed gold line depicts the pdf of the data, the solid black line is the pdf of our model, and the dashed blue line is the pdf of the counterfactual distribution if there was no change in the marginal tax rate at the first kink point of the EITC at \$8,580. The red vertical line corresponds to the first kink point of the EITC for self-employed individuals with one child. This figure and data is similar to Saez (2010) where we also look at all years from 1995 to 2004 and inflate earnings to 2008 \$ using the IRS inflation parameters.

Figure 7: Wage earners, one child, $15,740 kink, many covariates

(a) PDF



Note: Unobserved ability and knowledge are assumed to be uncorrelated so that ρ is constrained to zero. The data is the SOI PUFs (IRS) 1995-2004, wage earners, with one child and has 181.8 million weighted observations. These figures depict the second kink point of the EITC at $15,740. The figure depicts the pdf of the data, our model, and the counterfactual distribution for this sample.

Table 1: Self-employed earners, one child, $8,580 kink

Elasticity	0.73
	(0.00)
Married	0.08
	(0.00)
Unemployment	-0.05
	(0.00)
Social Security	0.04
	(0.00)
$\mathrm{COR}[\psi_i, \nu_i] = \rho$	0.22
	(0.01)
$\sqrt{\mathrm{V}[\psi_i]} = \sigma_\psi$	0.03
	(0.00)
Saez Elasticity	0.48
Observation	6.98m

p-values in parentheses

Note: The covariates are standardized so that the coefficients are comparable. The covariates are married, unemployment insurance, and social security are covariates that used. These covariates affect taxfilers location in the optimizing friction distribution distribution. The data is the SOI PUFs (IRS) 1995-2004, wage earners, with one child and has 181.8 million observations. p-values are in parentheses.

Table 2: Wage earners, one child, $15,740 kink, twenty-two covariates

Saez Elasticity	-0.001
	[0.025]
Elasticity	1.198^a
	[0.007]
Constant	-0.267^a
	[0.0001]
Have self-employment income	0.375^a
	[0.0007]
Have capital gains income	0.305^a
	[0.0005]
Have pension income	-0.113^a
	[0.0006]
Received a tax credit	0.318^a
	[0.0009]

Note: Unobserved ability and knowledge are assumed to be uncorrelated so that ρ is constrained to zero. The covariates are standardized so that the coefficients are comparable. The data is the SOI PUFs (IRS) 1995-2004, wage earners, with one child and has 181.8 million observations. Standard errors are in parentheses, with significance levels of 1%, 5%, and 10% denoted by a, b, and c, respectively.

A Theory appendix

A.1 Utility maximization appendix

The goal of this section is to demonstrate that models of supply or demand in many settings will generate data consistent with the processes presented in Section 2.0.1. Among the only restrictions on the underlying model is that it result in piecewise log-linear outcomes with at most two dimensions of unobserved heterogeneity. To provide an example with wages and prices, I solve a utility maximization problem similar to Equation (1) given by:

$$\max_{Q_i,L_i} U\left(Q_i,L_i\right) = Q_i - \frac{N_i^*}{1+1/\varepsilon}\left(\frac{L_i}{N_i^*}\right)^{1+1/\varepsilon} \tag{12a}$$

subject to:

$$C_i = \mathbb{1}_{Y_i \le K_i^*}\left[I_0 + \left(1-t_0\right)Y_i\right] + \mathbb{1}_{Y_i > K_i^*}\left[I_0 + \left(1-t_0\right)K_i^* - \Delta + \left(1-t_1\right)\left(Y_i - K_i^*\right)\right], \tag{12b}$$

$$Y_i = W_i L_i, \tag{12c}$$

$$C_i = P_i Q_i, \tag{12d}$$

Utility is increasing in the quantity of goods, Q_i, consumed and decreasing in labor, L_i. Labor earns wage, W_i, per unit and goods prices are, P_i, per unit so that Y_i and C_i are in currency units. The agent takes wages and prices as given but they can potentially differ for each agent and are therefore indexed by i. The Lagrangian function for the problem defined in Equation (12) is given by:

$$\begin{aligned}\mathcal{L}\left(Q_i,L_i\right) &= Q_i - \frac{N_i^*}{1+1/\varepsilon}\left(\frac{L_i}{N_i^*}\right)^{1+1/\varepsilon} \\ &+ \lambda \mathbb{1}_{W_iL_i \le K_i^*}\left[I_0 + \left(1-t_0\right)W_iL_i\right] \\ &+ \lambda \mathbb{1}_{W_iL_i > K_i^*}\left[I_0 + \left(1-t_0\right)K_i^* - \Delta + \left(1-t_1\right)\left(W_iL_i - K_i^*\right)\right] \\ &- \lambda P_i Q_i,\end{aligned} \tag{13}$$

in which Constraints (12c) and (12d) are substituted into Constraint (12b) and λ is the Lagrange multiplier. The derivative of the Lagrangian with respect to L_i is

$$\mathcal{L}_L = \begin{cases} -N_i^{*-1/\varepsilon}L_i^{1/\varepsilon} + \lambda\left(1-t_0\right)W_i & \text{if} \quad W_iL_i \le K_i^* \\ -N_i^{*-1/\varepsilon}L_i^{1/\varepsilon} + \lambda\left(1-t_1\right)W_i & \text{if} \quad W_iL_i > K_i^*. \end{cases} \tag{14}$$

The first order condition with respect to Q_i is $\mathcal{L}_Q = 1 - \lambda P_i = 0$ so that $1/P_i = \lambda$. Using Equation (14) and that $1/P_i = \lambda$, the first order conditions with respect to L_i provide labor supply

$$L_i = \begin{cases} N_i^*\left(\left(1-t_0\right)\dfrac{W_i}{P_i}\right)^{\varepsilon} & \text{if} \quad W_iL_i \le K_i^* \\ N_i^*\left(\left(1-t_1\right)\dfrac{W_i}{P_i}\right)^{\varepsilon} & \text{if} \quad W_iL_i > K_i^*. \end{cases} \tag{15}$$

The quantity of labor supplied by agents is a piecewise function that is increasing in the after-tax real wage and differs for each agent according to their unobserved heterogeneity,

N_i^* and K_i^*. Multiplying Equation (15) by W_i converts labor supply to labor income

$$Y_i = \begin{cases} N_i^* W_i \left((1-t_0) \dfrac{W_i}{P_i} \right)^{\varepsilon} & \text{if} \quad Y_i \le K_i^* \\ N_i^* W_i \left((1-t_1) \dfrac{W_i}{P_i} \right)^{\varepsilon} & \text{if} \quad Y_i > K_i^*. \end{cases} \tag{16}$$

Equation (16) does not describe the optimal policy for agents but instead presents circular logic in that it describes the optimal income choice for different ranges of optimal income. Instead, the optimal policy for agents will depend on thresholds that are derived in Section **??**. Throughout the remainder of this appendix, I omit the i subscript from N_i^*, W_i, P_i, Y_i, and K_i^* to simplify notation.

A.2 Indifference points

There are three indifference points for N^* defined by N_0, N_1, and $\check{N}$. Heterogeneity N^* indexes agents and so I can refer to these points as specific agents. N_0 is the agent that is indifferent between reporting income K^* and reporting income according to the $Y \le K^*$ case of Equation (16) with constraint slope $1 - t_0$. N_1 is the agent that is indifferent between reporting income K^* and reporting income according to the $Y > K^*$ case of Equation (16) with constraint slope $1 - t_1$. $\check{N}$ is the agent that is indifferent between reporting income according to the two cases of Equation (16).

A.2.1 Solving for N_0

An agent with heterogeneity value N_0 gets the same utility from reporting income K^* or from reporting income according to the $Y \le K^*$ case of Equation (16) with constraint slope $1 - t_0$ from Equation (12b). As such, I solve for N_0 by setting the difference in utility from Equation (12a) evaluated at these those relevant values equal to zero. Formally, I calculate:

$$U(Q_K, L_K) - U(Q_0, L_0) = Q_K - \frac{N_0}{1+1/\varepsilon} \left(\frac{L_K}{N_0} \right)^{1+1/\varepsilon} - Q_0 + \frac{N_0}{1+1/\varepsilon} \left(\frac{L_0}{N_0} \right)^{1+1/\varepsilon}, \tag{17}$$

in which, from Equation (12b) and the $WL \le K^*$ case of Equation (15),

$$Q_K = \frac{I_0}{P} + (1-t_0) \frac{K^*}{P}, \tag{18a}$$

$$L_K = \frac{K^*}{W}, \tag{18b}$$

$$Q_0 = \frac{I_0}{P} + (1-t_0) \frac{WL_0}{P}, \tag{18c}$$

$$L_0 = N_0 \left((1-t_0) \frac{W}{P} \right)^{\varepsilon}. \tag{18d}$$

After substituting Equation (18) into Equation (17) simplifying gives:

$$f(N_0, t_0) = -\varepsilon \left(\frac{K^*}{W} \right)^{1+1/\varepsilon} N_0^{-1/\varepsilon} - \left((1-t_0) \frac{W}{P} \right)^{1+\varepsilon} N_0 + (\varepsilon + 1)(1-t_0) \frac{K^*}{P}, \tag{19}$$

in which I define $f(N_0, t_0)$ only in terms of the two arguments N_0 and t_0 for notational simplicity. I find N_0 by solving

$$f(N_0, t_0) = 0. \tag{20}$$

A solution to Equation (20) is:

$$N_0 = \frac{K^*}{W}\left((1-t_0)\frac{W}{P}\right)^{-\varepsilon}, \tag{21}$$

which can be verified by substituting Equation (21) into Equation (20). I now show that Equation (21) is the unique solution to Equation (20). The first derivative of Equation (19) with respect to N_0 is

$$\frac{\partial f(N_0, t_0)}{\partial N_0} = \left(\frac{K^*}{W}\right)^{1+1/\varepsilon} N_0^{-(1+1/\varepsilon)} - \left((1-t_0)\frac{W}{P}\right)^{1+\varepsilon}. \tag{22}$$

Setting Equation (22) equal to zero and solving for N_0 gives Equation (21), so that it is a critical point of Equation (19). The second derivative of Equation (19) with respect to N_0 is:

$$\frac{\partial^2 f(N_0, t_0)}{\partial N_0^2} = -\left(1+\frac{1}{\varepsilon}\right)\left(\frac{K^*}{W}\right)^{1+1/\varepsilon} N_0^{-(2+1/\varepsilon)}, \tag{23}$$

which is always strictly negative because $K^*, W, N_0 > 0$ and $\varepsilon \geq 0$. Therefore, Equation (19) is globally concave down in N_0 and it attains a global maximum at N_0 defined by Equation (21). In summary, the solution to Equation (20) defined by Equation (21) is unique.

A.2.2 Solving for N_1

An agent with heterogeneity value N_1 gets the same utility from reporting income K^* or from reporting income according to the $Y > K^*$ case of Equation (16) with constraint slope $1 - t_1$ from Equation (12b). As such, I solve for N_1 by setting the difference in utility from Equation (12a) evaluated at these those relevant values equal to zero. Formally, I calculate:

$$U(Q_K, L_K) - U(Q_1, L_1) = Q_K - \frac{N_1}{1+1/\varepsilon}\left(\frac{L_K}{N_1}\right)^{1+1/\varepsilon} - Q_1 + \frac{N_1}{1+1/\varepsilon}\left(\frac{L_1}{N_1}\right)^{1+1/\varepsilon}, \tag{24}$$

in which, from Equation (12b) and the $WL > K^*$ case of Equation (15),

$$Q_K = \frac{I_0}{P} + (1 - t_0)\frac{K^*}{P}, \tag{25a}$$

$$L_K = \frac{K^*}{W}, \tag{25b}$$

$$Q_1 = \frac{I_0}{P} + (1 - t_0)\frac{K^*}{P} - \frac{\Delta}{P} + (1 - t_1)\left(\frac{WL_1}{P} - \frac{K^*}{P}\right), \text{ and} \tag{25c}$$

$$L_1 = N_1\left((1 - t_1)\frac{W}{P}\right)^{\varepsilon}. \tag{25d}$$

Substituting Equation (25) into Equation (24) and simplifying gives:

$$g(N_1) = f(N_1, t_1) + (\varepsilon + 1)\frac{\Delta}{P}, \tag{26}$$

in which,

$$f(N_1, t_1) = -\varepsilon\left(\frac{K^*}{W}\right)^{1+1/\varepsilon} N_1^{-1/\varepsilon} - \left((1 - t_1)\frac{W}{P}\right)^{1+\varepsilon} N_1 + (\varepsilon + 1)(1 - t_1)\frac{K^*}{P}. \tag{27}$$

I find N_1 by solving

$$g(N_1) = f(N_1, t_1) + (\varepsilon + 1)\frac{\Delta}{P} = 0. \tag{28}$$

Notice that Equation (27) is Equation (19) just evaluated at N_1 and t_1 instead of N_0 and t_0. As such, the first and second derivatives of Equation (26) with respect to N_1 are just the derivatives of $f(N_0, t_0)$—Equations (22) and (23) —evaluated at N_1 and t_1;

$$g'(N_1) = \frac{\partial g(N_1)}{\partial N_1} = \frac{\partial f(N_1, t_1)}{\partial N_1} = \left(\frac{K^*}{W}\right)^{1+1/\varepsilon} N_1^{-(1+1/\varepsilon)} - \left((1 - t_1)\frac{W}{P}\right)^{1+\varepsilon}, \tag{29}$$

and

$$g''(N_1) = \frac{\partial^2 g(N_1)}{\partial N_1^2} = \frac{\partial^2 f(N_1, t_1)}{\partial N_1^2} = -\left(1 + \frac{1}{\varepsilon}\right)\left(\frac{K^*}{W}\right)^{1+1/\varepsilon} N_1^{-(2+1/\varepsilon)}. \tag{30}$$

Setting Equation (29) equal to zero and solving for N_1 gives:

$$N_1 = \frac{K^*}{W}\left((1 - t_1)\frac{W}{P}\right)^{-\varepsilon}, \tag{31}$$

which is therefore a critical point of Equation (26). Equation (30) is always strictly negative because $K^*, W, N_1 > 0$ and $\varepsilon \geq 0$. Therefore, Equation (26) is globally concave down in N_1 and it attains a global maximum at N_1 defined by Equation (31).

The solution to Equation (28) depends on if $\Delta = 0$, $\Delta > 0$, or $\Delta < 0$. The first case is when $\Delta = 0$. When $\Delta = 0$, Equation (26) equals Equation (27) and, by the logic of Section A.2.1 and Equations (29) and (30), Equation (31) is the unique solution to Equation (28).

The second case is when $\Delta > 0$. When $\Delta > 0$, there is no closed form solution for N_1 so I solve Equation (28) using Halley's method. Halley's method is a third-order iterative root finding algorithm similar to the second-order Newton-Raphson method (Scavo and Thoo, 1995). The formula used to iterate is:

$$N_1^{(r+1)} = N_1^{(r)} - \frac{2g\left(N_1^{(r)}\right)g'\left(N_1^{(r)}\right)}{2\left[g'\left(N_1^{(r)}\right)\right]^2 - g\left(N_1^{(r)}\right)g''\left(N_1^{(r)}\right)}, \tag{32}$$

which uses Equations (26), (29), and (30). Next, rewrite Equation (28) as

$$f\left(N_1, t_1\right) = -\left(\varepsilon + 1\right)\frac{\Delta}{P}. \tag{33}$$

Because Equation (27) is globally concave down, attains its global maximum of zero at Equation (31), and $P, \Delta > 0$, $\varepsilon \geq 0$, I know that Equation (33) is satisfied exactly twice, once with a value smaller and once with a value larger than Equation (31). Using the $Y > K^*$ case of Equation (16), it must be that:

$$N_1 \geq \frac{K^*}{W}\left(\left(1 - t_1\right)\frac{W}{P}\right)^{-\varepsilon}, \tag{34}$$

so I eliminate the solution that is smaller than Equation (31).

To ensure that Halley's method converges and converges to the solution that is larger than Equation (31), it is important to start Halley's iterations at a carefully chosen value. To that end, for some small $h > 0$, I define

$$N_1^{(0)}\left(h\right) = \left(1 + h\right)\frac{K^*}{W}\left(\left(1 - t_1\right)\frac{W}{P}\right)^{-\varepsilon}. \tag{35}$$

I use two starting values, $N_1^{(0)}\left(0\right)$ and $N_1^{(0)}\left(h\right)$ in Equation (32) so that the first update is

$$N_1^{(1)} = N_1^{(0)}\left(0\right) - \frac{2g\left(N_1^{(0)}\left(0\right)\right)g'\left(N_1^{(0)}\left(h\right)\right)}{2\left[g'\left(N_1^{(0)}\left(0\right)\right)\right]^2 - g\left(N_1^{(0)}\left(0\right)\right)g''\left(N_1^{(0)}\left(0\right)\right)} = N_1^{(0)}\left(0\right) + \frac{2g'\left(N_1^{(0)}\left(h\right)\right)}{g''\left(N_1^{(0)}\left(0\right)\right)}. \tag{36}$$

Equation (32) simplifies to Equation (36) because $g'\left(N_1^{(0)}\left(0\right)\right) = 0$. The update is positive because $g'\left(N_1^{(0)}\left(h\right)\right) < 0$ and $g''\left(N_1^{(0)}\left(0\right)\right) < 0$. I cannot start at $N_1^{(0)}\left(0\right)$ for all the terms in Equation (32) because $g'\left(N_1^{(0)}\left(0\right)\right) = 0$ (making the numerator zero and not providing an update). In summary, I start at $N_1^{(0)}\left(0\right)$ and $N_1^{(0)}\left(h\right)$ then $N_1^{(1)} > N_1^{(0)}\left(0\right)$. For all following iterations, I use only Equation (32) until $|N_1^{(r+1)} - N_1^{(r)}| < h$, in which h is the same or a different very small number.

The third case is when $\Delta < 0$. In this case, because $P > 0$ and $\varepsilon \geq 0$, I know $-\left(\varepsilon + 1\right)\frac{\Delta}{P} > 0$ from Equation (33) but Equation (27) has a global maximum of zero and so there is no solution. In other words, there is no N_1 agent that is indifferent between reporting K^* and reporting according to the $Y > K^*$ case of Equation (16).

Instead, because Equation (27) is at most zero and $(\varepsilon+1)\dfrac{\Delta}{P} < 0$, $g(N_1) < 0$ from Equation (26). Because $g(N_1) < 0$, Equation (24) implies the utility from reporting income according to the $Y > K^*$ case of Equation (16) is always higher than utility from reporting income K^*. N_1 simply does not exist when $\Delta < 0$.

A.2.3 Solving for $\check{N}$

An agent with heterogeneity value $\check{N}$ gets the same utility from reporting income according to the $Y \leq K^*$ or $Y > K^*$ case of Equation (16). As such, I solve for $\check{N}$ by setting the difference in utility from Equation (12a) evaluated at these those relevant values equal to zero. Formally, I calculate:

$$U\left(\check{Q}_0, \check{L}_0\right) - U\left(\check{Q}_1, \check{L}_1\right) = \check{Q}_0 - \frac{\check{N}}{1+1/\varepsilon}\left(\frac{\check{L}_0}{\check{N}}\right)^{1+1/\varepsilon} - Q_1 + \frac{\check{N}}{1+1/\varepsilon}\left(\frac{\check{L}_1}{\check{N}}\right)^{1+1/\varepsilon}, \quad (37)$$

in which, from Equation (12b) and the two cases of Equation (15),

$$\check{Q}_0 = \frac{I_0}{P} + (1-t_0)\frac{W\check{L}_0}{P}, \quad (38a)$$

$$\check{L}_0 = \check{N}\left((1-t_0)\frac{W}{P}\right)^{\varepsilon}, \quad (38b)$$

$$\check{Q}_1 = \frac{I_0}{P} + (1-t_0)\frac{K^*}{P} - \frac{\Delta}{P} + (1-t_1)\left(\frac{W\check{L}_1}{P} - \frac{K^*}{P}\right), \text{ and} \quad (38c)$$

$$\check{L}_1 = \check{N}\left((1-t_1)\frac{W}{P}\right)^{\varepsilon}. \quad (38d)$$

Substituting Equation (38) into Equation (37), setting equal to zero, and then solving for $\check{N}$ gives,

$$\check{N} = \frac{(\varepsilon+1)\,A(t_0,t_1)\,K^*/P}{B(\varepsilon,t_0,t_1)\,(W/P)^{1+\varepsilon}} - \frac{(\varepsilon+1)\,\Delta/P}{B(\varepsilon,t_0,t_1)\,(W/P)^{1+\varepsilon}}, \quad (39)$$

in which,

$$A(t_0,t_1) = (1-t_0) - (1-t_1), \quad (40a) \qquad B(\varepsilon,t_0,t_1) = (1-t_0)^{\varepsilon+1} - (1-t_1)^{\varepsilon+1}. \quad (40b)$$

A.3 Binding thresholds

Depending on the intercept and slopes of the constraint defined in Equation (12b), the indifference points N_0, N_1, and $\check{N}$ defined in Section A.2 may or may not be the binding threshold that defines optimal policy for each agent. For example, when $\Delta < 0$ and $s_1 \geq s_0$ XXXX double check XXXX, N_1 does not exist and the binding threshold is instead $\check{N}$. This section defines the functions that map Equation (2) and the indifference points N_0, N_1, and $\check{N}$ points N_0, N_1, and $\check{N}$ into the binding thresholds $\underline{N}$, $\overline{N}$, and $\check{N}$. These binding thresholds do define the optimal policy and the data generating process explained in Section 2.0.1.

A.3.1 XXXX

A.4 Changing Among General Solutions

There are two general solutions to Equation (1) depending the incentive schedule that the agent faces. The two general solutions correspond to Equation (3) and Equation (4). Equation (3) corresponds to cases (2a), (2b), and (2c). Equation (4) corresponds to cases (2d).

Specifically, the concern is that there exists $\tilde{\varepsilon}$ such that the agent is indifferent between reporting $\underline{Y} = \check{\underline{Y}}$. If that were the case, then the agent would switch from reporting according to Equation (3) to reporting according to equation (4). To show that the agent cannot be indifferent between reporting according to these two equations I would need to show that utility at these two points cannot be equal, which implies$\underline{Y} \neq \check{\underline{Y}}$.

If the agent is indifferent between reporting $\underline{Y} = \check{\underline{Y}}$, then

The left hand side of Equation (**??**) could equal 0 given that $(\varepsilon + 1)\,\Delta\,(1 - t_0)^{\tilde{\varepsilon}}$ is always positive and the sign of $(\tilde{\varepsilon} + 1)\,(1 - t_0)^{\tilde{\varepsilon}}\,[(1 - t_0) - (1 - t_1)] - \left[(1 - t_0)^{\varepsilon+1} - (1 - t_1)^{\varepsilon+1}\right]$ is ambiguous.

A.5 Inverting

XXXX I do not yet need the inverse XXXX

Equation (39) is a linear function of K^*. To find the inverse, $\check{N}\,(K^*)^{-1}$, I substitute $\check{N}^{-1}$ for K^* and then solve for $\check{N}^{-1}$, which is

$$\check{N}^{-1} = \frac{PK^*\,(W/P)^{1+\varepsilon}\,B\,(\varepsilon, s_0, s_1)}{(\varepsilon + 1)\,A\,(s_0, s_1)} + \frac{\Delta}{A\,(s_0, s_1)}. \tag{41}$$

It is easy to verify that this is the inverse by substituting Equation (39) into (41) to get $\check{N}^{-1}\left(\check{N}\,(K^*)\right) = K^*$.

A.6 Scrap text

by Equations (21), (31), and (39), respectively

(32)

defined in Section A.2

Appendix Section A.3 explains these cases in detail and defines the functions that map Equation (2) and the points N_0, N_1, and $\check{N}$ into the binding thresholds $\underline{N}$, $\overline{N}$, and $\check{N}$ that define the optimal policy and the data generating process explained in Section 2.0.1.

and binding thresholds

XXXXX need to edit this section once the sections for each N0, N1, Ncheck are done XXXXX

Importantly, depending on the intercept and slopes of the constraint defined in Equation (12b), the points N_0, N_1, or $\check{N}$ may or may not be the binding threshold that defines optimal policy for each agent.

Appendix Section A.3 explains these cases in detail and defines the functions that map Equation (2) and the points N_0, N_1, and $\check{N}$ into the binding thresholds $\underline{N}$, $\overline{N}$, and $\check{N}$ that define the optimal policy and the data generating process explained in Section 2.0.1.

K^* and reporting income according to the $Y \leq K^*$ case of Equation (16). agent that is indifferent between reporting income according to $N^*(1-t_0)^\varepsilon$ and reporting income according to $N^*(1-t_1)^\varepsilon$.

For this threshold, the agent is deciding whether to bunch at K^*, which corresponds to the cases of Equations (2a), (2b), and (2c). In Section **??** I solve for the threshold $\overline{N}$, which is the agent that is indifferent between reporting income K^* and reporting income according to $N^*(1-t_1)^\varepsilon$. For this threshold, the agent is deciding whether to bunch at K^*, which corresponds to the cases of Equations (2a), (2b), and (2c).

In Section **??** I solve for for the threshold

that are relevant for each agent will bunch or the agent will create a missing mass in the distribution depending on the relevant case(s) from Equations (2a), (2b), (2c), and (2d):

The agent solves the utility maximization problem described in Equation (1). Depending on which of the four piecewise linear tax schedule schedules (as described in Equation (2)) that the agent faces, the agent will choose to report labor supply given by Equation (3) or by Equation (4).

In Section **??** I solve for for the threshold $\underline{N}$, which is the agent that is indifferent between reporting income K^* and reporting income according to $N^*(1-t_0)^\varepsilon$. For this threshold, the agent is deciding whether to bunch at K^*, which corresponds to the cases of Equations (2a), (2b), and (2c). In Section **??** I solve for the threshold $\overline{N}$, which is the agent that is indifferent between reporting income K^* and reporting income according to $N^*(1-t_1)^\varepsilon$. For this threshold, the agent is deciding whether to bunch at K^*, which corresponds to the cases of Equations (2a), (2b), and (2c). In Section **??** I solve for the threshold $\check{N}$, which is the agent that is indifferent between reporting income according to $N^*(1-t_0)^\varepsilon$ and reporting income according to $N^*(1-t_1)^\varepsilon$. For this threshold. the agent is deciding whether to create a missing mass at K^*, which corresponds to case of Equation (2d). At the end of each section I invert each threshold.

The utility function used to solve for these three thresholds is

$$U(C,Y) = C(Y) - \frac{N^*}{1+\frac{1}{\varepsilon}}\left(\frac{Y}{N^*}\right)^{1+\frac{1}{\varepsilon}}. \tag{42}$$

The general utility maximization problem shows that any model that takes a form similar to Equation (15) will generate data consistent with the processes presented in Section 4. These models could include models of supply or demand in many settings and as long as they piecewise log-linear.

This appendix solves a utility maximization problem similar to Equation (1) but with general wages and prices. I emphasize that any model that takes a form similar to Equation (15) will generate data consistent with the processes presented in Section 4.

XXXXX this will go into the final DGP section when I get there XXXX

$$y_i = \begin{cases} n_i^* + \varepsilon s_0 + (\varepsilon+1) w_i - \varepsilon p_i & y_i \leq k_i^* \\ n_i^* + \varepsilon s_1 + (\varepsilon+1) w_i - \varepsilon p_i & y_i > k_i^* \end{cases} \tag{43}$$

in which lower case variables are logarithms of the same upper case variable and I define $s_0 = \ln(1-t_0)$ and $s_1 = \ln(1-t_1)$. Throughout the remainder of this appendix, I omit the i subscript from n_i^*, w_i, p_i, y_i, and k_i^* to simplify notation and transition between uppercase level or lower case log variables.

Equation (22) equals zero when N_0 is equal to Equation (21) so that Equation (21) is a critical point of Equation (19). and

$$\frac{\partial^2 f(N_0, t_0)}{\partial N_0^2} = -\left(1+\frac{1}{\varepsilon}\right)\left(\frac{K^*}{W}\right)^{1+1/\varepsilon} N_0^{-(2+1/\varepsilon)} \tag{44}$$

Finally, I show that the critical value $\underline{N} = K^*(1-t_0)^{-\varepsilon}$ is the global maximum and consequently the unique solution to $f(\underline{N}, t_0) = 0$. The first and second derivatives of $f(\underline{N}, t_0)$ with respect to $\underline{N}$ are

and

$$f_{\underline{N}\underline{N}}(\underline{N}, t_0) = -K^{*\left(1+\frac{1}{\varepsilon}\right)}\left(1+\frac{1}{\varepsilon}\right)\underline{N}^{-\left(2+\frac{1}{\varepsilon}\right)} \tag{45}$$

respectively. If I set Equation (**??**) equal to zero, then I get $K^*(1-t_0)^{-\varepsilon} = \underline{N}$. Equation (45) is negative for any $K^* > 0, \underline{N} > 0$. This means that our function $f(\underline{N}, t_0)$ is globally concave, which means that $\underline{N} = K^*(1-t_0)^{-\varepsilon}$ is a global maximum.

XXXXXXXXXXXXXXXXXXXXXXX

The $Y \leq K^*$ case of Equation (16) ensures that

$$N_0 \geq \frac{K^*}{W}\left((1-t_0)\frac{W}{P}\right)^{-\varepsilon}, \tag{46}$$

and substituting Equation (21) into Equation (19) shows that it is a solution.

substituting Equation (21) into Equation (19) shows that it is a solution.

Then substituting Equation (21) into Equation (19) shows that it is a solution.

$\underline{N} = K^*(1-t_0)^{-\varepsilon}$ into $f(\underline{N}, t_0) = 0$ yields the solution. I can transform the solution of $\underline{N}$ into logs, which yields

$$\underline{n}(k^*, \varepsilon, s_0) = k^* - \varepsilon s_0, \tag{47}$$

By taking the first order condition of Equation (42) with respect to Y where I substitute the first part of the piecewise budget constraint in Equation (1b) for $C(Y)$, yields $Y = (1-t_0)^{\varepsilon} N$. The first part of the budget constraint is applicable when $Y \leq K^*$. To solve for the threshold $\underline{N}$, I set the utility of reporting K^* equal to the utility of reporting income according to $Y = (1-t_0)^{\varepsilon} N$. This equality along with simplifying notation corresponds to

$$\underline{Y} = \underline{N}(1-t_0)^{\varepsilon}, \qquad C(\underline{Y}) = I_0 + (1-t_0)\underline{Y}, \qquad C(K^*) = I_0 + (1-t_0)K^*$$

$$C\left(\underline{Y}\right)-\frac{\underline{N}}{1+\frac{1}{\varepsilon}}\left(\frac{\underline{Y}}{\underline{N}}\right)^{1+\frac{1}{\varepsilon}}=C\left(K^{*}\right)-\frac{\underline{N}}{1+\frac{1}{\varepsilon}}\left(\frac{K^{*}}{\underline{N}}\right)^{1+\frac{1}{\varepsilon}}. \tag{48}$$

With further algebraic steps I can simplify Equation (48) to

$$-\varepsilon K^{*\left(1+\frac{1}{\varepsilon}\right)}\underline{N}^{-\frac{1}{\varepsilon}}-\left(1-t_{0}\right)^{1+\varepsilon}\underline{N}+\left(\varepsilon+1\right)\left(1-t_{0}\right)K^{*}=f\left(\underline{N},t_{0}\right). \tag{49}$$

I define this difference in utility functions in Equation (49) as $f\left(\underline{N},t_{0}\right)$.

Then substituting in $\underline{N}=K^{*}(1-t_{0})^{-\varepsilon}$ into $f\left(\underline{N},t_{0}\right)=0$ yields the solution. I can transform the solution of $\underline{N}$ into logs, which yields

$$\underline{n}\left(k^{*},\varepsilon,s_{0}\right)=k^{*}-\varepsilon s_{0}, \tag{50}$$

which is the expression in the main text.

Additionally, I want to solve for the inverse function of $\underline{N}$. To find the inverse of the linear function, I substitute $\underline{N}^{-1}$ for K^{*} and then solve for $\underline{N}^{-1}$. Solving for $\underline{N}^{-1}$ yields $\underline{N}^{-1}=K^{*}(1-t_{0})^{\varepsilon}$.

By taking the first order condition of Equation (42) with respect to Y where I substitute the second part of the piecewise budget constraint in Equation (1b) for $C\left(Y\right)$, yields $Y=(1-t_{1})^{\varepsilon}N$. The second part of the budget constraint is applicable when $Y>K^{*}$. To solve for the threshold $\overline{N}$, I set the utility of reporting K^{*} equal to the utility of reporting income according to $Y=(1-t_{1})^{\varepsilon}N$. This equality along with simplifying notation corresponds to

$$\overline{Y}=\overline{N}\left(1-t_{1}\right)^{\varepsilon}, \qquad C\left(\overline{Y}\right)=I_{0}+\left(1-t_{0}\right)K^{*}-\Delta+\left(1-t_{1}\right)\left(\overline{Y}-K^{*}\right)$$

$$C\left(K^{*}\right)=I_{0}+\left(1-t_{0}\right)K^{*}$$

$$C\left(K^{*}\right)-\frac{\overline{N}}{1+\frac{1}{\varepsilon}}\left(\frac{K^{*}}{\overline{N}}\right)^{1+\frac{1}{\varepsilon}}=C\left(\overline{Y}\right)-\frac{\overline{N}}{1+\frac{1}{\varepsilon}}\left(\frac{\overline{Y}}{\overline{N}}\right)^{1+\frac{1}{\varepsilon}}. \tag{53}$$

where

$$C\left(\overline{Y}\right)=I_{0}+\left(1-t_{0}\right)K^{*}-\Delta+\left(1-t_{1}\right)\left(\bar{N}\left(1-t_{1}\right)^{\varepsilon}-K^{*}\right). \tag{54}$$

After simplifying Equation (53), I arrive at a nonlinear expression in terms of $\overline{N}$:

$$-\varepsilon K^{*\left(1+\frac{1}{\varepsilon}\right)}\overline{N}^{-\frac{1}{\varepsilon}} - (1-t_1)^{1+\varepsilon}\,\overline{N} + (\varepsilon+1)\,K^*\,(1-t_1) + (\varepsilon+1)\,\Delta = g\left(\overline{N},t_1\right) \qquad (55)$$

I define this difference in utility functions in Equation (55) as $g\left(\overline{N},t_1\right)$. Furthermore $g\left(\overline{N},t_1\right)$ can be written as $f\left(\overline{N},t_1\right)+(\varepsilon+1)\,\Delta$.

Additionally, in order to invert $\overline{N}$, I will rely on the secant method (Papakonstantinou and Tapia, 2013).

Finally when $\Delta=0$, $f\left(\overline{N},t_1\right)=g\left(\overline{N},t_1\right)$. This equality implies that the critical value $\overline{N}=K^*\,(1-t_1)^{-\varepsilon}$ is a unique global maximum.

When $\Delta>0$, I know from the expression $g\left(\overline{N},t_1\right)=0$ that $f\left(\overline{N},t_1\right)=-(\varepsilon+1)\,\Delta$. When $\Delta>0$, the function $f\left(\overline{N},t_1\right)$ crosses $-(\varepsilon+1)\,\Delta<0$ in two places giving one root that is less than, and one root that is greater than, $K^*\,(1-t_1)^{-\varepsilon}$. I use Halley's method to find these two roots numerically. The larger root is the solution when $\Delta<0$ and $g\left(\overline{N}\right)>0$ because higher utility is obtained. When $-(\varepsilon+1)\,\Delta<0, g\left(\overline{N}\right)>0$ between the two roots implies the utility obtained at K^* is higher than the utility obtained at $\overline{N}$ because $g\left(\overline{N}\right)$ is utility obtained at K^* minus utility obtained at $\overline{N}$. There is no $\overline{N}$ that yields $g\left(\overline{N}_N,t_1\right)=0$ when $\Delta<0$. This is due to $f\left(\overline{N},t_1\right)$ never equaling $-(\varepsilon+1)\,\Delta>0$. In order for $f\left(\overline{N},t_1\right)=-(\varepsilon+1)\,\Delta>0$, $g\left(\overline{N}\right)<0$ for all $\overline{N}$. However there is no $\overline{N}>K^*$ that makes the agent indifferent between reporting K^* and $\overline{N}$.

Additionally, in order to invert $\overline{N}$, I will rely on the secant method (Papakonstantinou and Tapia, 2013).

Finally when $\Delta=0$, $f\left(\overline{N},t_1\right)=g\left(\overline{N},t_1\right)$. This equality implies that the critical value $\overline{N}=K^*\,(1-t_1)^{-\varepsilon}$ is a unique global maximum. When $\Delta>0$, I know from the expression $g\left(\overline{N},t_1\right)=0$ that $f\left(\overline{N},t_1\right)=-(\varepsilon+1)\,\Delta$. When $\Delta>0$, the function $f\left(\overline{N},t_1\right)$ crosses $-(\varepsilon+1)\,\Delta<0$ in two places giving one root that is less than, and one root that is greater than, $K^*\,(1-t_1)^{-\varepsilon}$. I use Halley's method to find these two roots numerically. The larger root is the solution when $\Delta<0$ and $g\left(\overline{N}\right)>0$ because higher utility is obtained. When $-(\varepsilon+1)\,\Delta<0, g\left(\overline{N}\right)>0$ between the two roots implies the utility obtained at K^* is higher than the utility obtained at $\overline{N}$ because $g\left(\overline{N}\right)$ is utility obtained at K^* minus utility obtained at $\overline{N}$. There is no $\overline{N}$ that yields $g\left(\overline{N}_N,t_1\right)=0$ when $\Delta<0$. This is due to $f\left(\overline{N},t_1\right)$ never equaling $-(\varepsilon+1)\,\Delta>0$. In order for $f\left(\overline{N},t_1\right)=-(\varepsilon+1)\,\Delta>0$, $g\left(\overline{N}\right)<0$ for all $\overline{N}$. However there is no $\overline{N}>K^*$ that makes the agent indifferent between reporting K^* and $\overline{N}$.

B Estimation appendix

B.1 Convex taxes and bivariate normal errors

B.1.1 Convex taxes augmented log-likelihood

This section provides the ingredients for the augmented log-likelihood given in Equation (??). First, consider the probability of observing a data set generated by Equation (??). The probability that agent i has ability level n_i^* and location k_i^* that places them to the left of the kink with bivariate normal errors is denoted by Q_{i1}. The probability of observing earnings for agent i when they believe they will face the lower tax rate, s_0, is $f_{i1}Q_{i1}$:

$$f_{i1} = \frac{1}{\sigma_\nu}\phi\left(\frac{y_i - \varepsilon s_0 - X_i'\beta}{\sigma_\nu}\right) \text{ and } Q_{i1} = \Phi\left(\frac{k - \varepsilon s_0 - X_i'\beta + Z_i'\gamma}{\sigma_{\nu-\psi}}\right).$$

The probabilities of f_{i2} and Q_{i2} are defined similarly to f_{i1} and Q_{i1}. The probability that the agent finds it optimal to bunch at k_i^* and therefore report $y_i = k_i^*$ is $f_{i2}Q_{i2}$. Each of these components are defined by,

$$f_{i2} = \frac{1}{\sigma_\psi}\phi\left(\frac{y_i - k - Z_i'\gamma}{\sigma_\psi}\right) \text{ and } Q_{i2} = \Phi\left(\frac{k - \varepsilon s_1 - X_i'\beta + Z_i'\gamma}{\sigma_{\nu-\psi}}\right) - \Phi\left(\frac{k - \varepsilon s_0 - X_i'\beta + Z_i'\gamma}{\sigma_{\nu-\psi}}\right).$$

The probabilities of f_{i3} and Q_{i3} are defined similarly to f_{i1} and Q_{i1}. The probability of observing earnings for agent i when they optimize trying to face the higher tax rate, s_1, is $f_{i3}Q_{i3}$:

$$f_{i3} = \frac{1}{\sigma_\nu}\phi\left(\frac{y_i - \varepsilon s_1 - X_i'\beta}{\sigma_\nu}\right), \; Q_{i3} = 1 - \Phi\left(\frac{k - \varepsilon s_1 - X_i'\beta + Z_i'\gamma}{\sigma_{\nu-\psi}}\right).$$

B.2 Properties of Bivariate Normal Distribution

The random variables ν and ψ follow a joint bivariate normal distribution

$$\begin{pmatrix} \nu \\ \psi \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_\nu \\ \mu_\psi \end{pmatrix}, \begin{pmatrix} \sigma_\nu^2 & \rho_{\nu\psi}\sigma_\nu\sigma_\psi \\ \rho_{\nu\psi}\sigma_\nu\sigma_\psi & \sigma_\psi^2 \end{pmatrix}\right]$$

This implies that the errors for the selection equation $\nu - \psi$ have the following moments and properties

$$E[\nu] = \mu_\nu,\; E[\psi] = \mu_\psi, \qquad COV[\nu, \nu-\psi] = \sigma_\nu^2 - \rho_{\nu\psi}\sigma_\nu\sigma_\psi,$$

$$E[\nu-\psi] = \mu_\nu - \mu_\psi, \qquad COR(\nu, \nu-\psi) = \rho_{\nu,\nu-\psi},$$

$$V[\nu-\psi] = \sigma_\nu^2 + \sigma_\psi^2 - 2\rho_{\nu\psi}\sigma_\nu\sigma_\psi = \sigma_{\nu-\psi}^2, \qquad \rho_{\nu,\nu-\psi} = \frac{\sigma_\nu - \rho_{\nu\psi}\sigma_\psi}{\sqrt{\sigma_\nu^2 + \sigma_\psi^2 - 2\rho_{\nu\psi}\sigma_\nu\sigma_\psi}}$$

Therefore, the random variables ν and $\nu - \psi$ follow the joint bivariate normal distribution

$$\begin{pmatrix} \nu \\ \nu-\psi \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_\nu \\ \mu_\nu - \mu_\psi \end{pmatrix}, \begin{pmatrix} \sigma_\nu^2 & \rho_{\nu,\nu-\psi}\sigma_\nu\sigma_{\nu-\psi} \\ \rho_{\nu,\nu-\psi}\sigma_\nu\sigma_{\nu-\psi} & \sigma_{\nu-\psi}^2 \end{pmatrix}\right]$$

It is useful to note the following conditional distributions

$$(\nu \mid \nu-\psi) \sim N\left[\mu_\nu + \frac{\sigma_\nu}{\sigma_{\nu-\psi}}\rho_{\nu,\nu-\psi}(\nu - \psi - \mu_\nu + \mu_\psi), \left(1 - \rho_{\nu,\nu-\psi}^2\right)\sigma_\nu^2\right]$$

and

$$(\nu-\psi \mid \nu) \sim N\left[\mu_\nu - \mu_\psi + \frac{\sigma_{\nu-\psi}}{\sigma_\nu}\rho_{\nu,\nu-\psi}(\nu - \mu_\nu), \left(1 - \rho_{\nu,\nu-\psi}^2\right)\sigma_{\nu-\psi}^2\right].$$

The marginal distributions of ν and ψ are given by

$$(\nu) \sim N\left[\mu_\nu, \sigma_\nu^2\right], \ (\nu - \psi) \sim N\left[\mu_\nu - \mu_\psi, \sigma_{\nu-\psi}^2\right].$$

C Ruud equations appendix

Implementing the Expectation-Maximization (EM) algorithm following Ruud (1991) requires conditional moments of the latent variables n^* and k^* defined in Equation (51). These derivations are simplified because the latent variables are conditionally joint normally distributed.

Normality ensures that the conditional expectations leverage simple formulas for truncated normal distributions.

C.1 Conditional probabilities

XXXX These probabilities are all messed up. To fix go back to convex functions and don't worry about the LHS definitions of these terms. Write out what you coded and then try to figure out who the Pcondy columns are the correct conditionals for the EM. XXXX I use Bayes rule to write the probability of each case in Equation (60). The probability that an observed value y is below the kink is,

$$P\left(y < k^* \mid y\right) = \frac{P\left(y \mid y < k^*\right) P\left(y < k^*\right)}{P\left(y\right)}, \tag{56}$$

in which the three terms are

$$P\left(y \mid y < k^*\right) = \frac{1}{\sigma_\nu}\phi\left(\frac{y - \varepsilon s_0 - X'\beta}{\sigma_\nu}\right),$$

$$P\left(y < k^*\right) = 1 - \Phi\left(a_0\right),$$

and

$$P\left(y\right) = P\left(y \mid y < k^*\right) P\left(y < k^*\right) + P\left(y \mid y = k^*\right) P\left(y = k^*\right) + P\left(y \mid y > k^*\right) P\left(y > k^*\right). \tag{57}$$

The probability that observed value y is at the kink is,

$$P\left(y = k^* \mid y\right) = \frac{P\left(y \mid y = k^*\right) P\left(y = k^*\right)}{P\left(y\right)}, \tag{58}$$

in which the denominator is defined in Equation (57) and the remaining two terms are,

$$P\left(y \mid y = k^*\right) = \frac{1}{\sigma_\psi}\phi\left(\frac{y - k - Z'\gamma}{\sigma_\psi}\right)$$

and

$$P\left(y = k^*\right) = \Phi\left(b_1\right) - \Phi\left(b_0\right).$$

The probability than observed value y is above the kink is,

$$P\left(y > k^* \mid y\right) = \frac{P\left(y \mid y > k^*\right) P\left(y > k^*\right)}{P\left(y\right)}, \tag{59}$$

in which the denominator is defined in Equation (57) and the remaining two terms are,

$$P\left(y \mid y > k^*\right) = \frac{1}{\sigma_\nu}\phi\left(\frac{y - \varepsilon s_1 - X'\beta}{\sigma_\nu}\right)$$

and

$$P\left(y > k^*\right) = \Phi\left(a_1\right).$$

C.2 Conditional first moments

C.2.1 Conditional first moments of location

The conditional first moment of the location latent variable defined in Equation (5) can be written in three additive terms as,

$$\begin{aligned} E\left[k^* \mid y\right] &= E\left[k^* \mid y, y < k^*\right] P\left[y < k^* \mid y\right] \\ &+ E\left[k^* \mid y, y = k^*\right] P\left[y = k^* \mid y\right] \\ &+ E\left[k^* \mid y, y > k^*\right] P\left[y > k^* \mid y\right]. \end{aligned} \tag{60}$$

Using our assumptions about n^* and k^* from Equation (51), the first expectation part of Equation (60) is,

$$E\left[k^* \mid y, y < k^*\right] = k + Z'\gamma + E\left[\psi \mid y, y < k^*\right], \tag{61}$$

and because conditional on y and $y = k^*$, k^* just is y, the second part is,

$$E\left[k^* \mid y, y = k^*\right] = y. \tag{62}$$

The third expectation part of Equation (60) is

$$E\left[k^* \mid y, y > k^*\right] = k + Z'\gamma + E\left[\psi \mid y, y > k^*\right]. \tag{63}$$

The conditional expectations of ψ in Equations (61) and (63) rely on the fact that ν and ψ are distributed according to the bivariate normal given in Equation (7), the definition of k^* given in Equation (6b), and the implicit definition of ψ from combining Equations (3), (5), and (6a). Combining these ingredients provides,

$$E\left[\psi \mid y, y < k^*\right] = \rho\frac{\sigma_\psi}{\sigma_\nu}\left(y - \varepsilon s_0 - X'\beta\right) + \sigma_\psi\sqrt{1-\rho^2}\left[\frac{\phi\left(a_0\right)}{1 - \Phi\left(a_0\right)}\right], \tag{64}$$

and also,

$$E\left[\psi \mid y, y > k^*\right] = \rho\frac{\sigma_\psi}{\sigma_\nu}\left(y - \varepsilon s_1 - X'\beta\right) - \sigma_\psi\sqrt{1-\rho^2}\left[\frac{\phi\left(a_1\right)}{\Phi\left(a_1\right)}\right], \tag{65}$$

in which,

$$a_j = \frac{y - k - Z'\gamma - \rho\frac{\sigma_\psi}{\sigma_\nu}\left(y - \varepsilon s_j - X'\beta\right)}{\sigma_\psi\sqrt{1-\rho^2}} \text{ for } j = \{0,1\}. \tag{66}$$

For tax schedules that follow Equations (2a), (2b), and (2c), Equation (61) combined with Equation (64) provides the final expression for the first expectation, which is,

$$E\left[k^* \mid y, y < k^*\right] = k + Z'\gamma + \rho\frac{\sigma_\psi}{\sigma_\nu}\left(y - \varepsilon s_0 - X'\beta\right) + \sigma_\psi\sqrt{1-\rho^2}\left[\frac{\phi\left(a_0\right)}{1-\Phi\left(a_0\right)}\right]. \tag{67}$$

Equation (62) provides the second expectation. Combining Equations (63) and (65) provides the final expression for the third expectation, which is,

$$E\left[k^* \mid y, y > k^*\right] = k + Z'\gamma + \rho\frac{\sigma_\psi}{\sigma_\nu}\left(y - \varepsilon s_1 - X'\beta\right) - \sigma_\psi\sqrt{1-\rho^2}\left[\frac{\phi\left(a_1\right)}{\Phi\left(a_1\right)}\right]. \tag{68}$$

For tax schedules that follow Equation (2d), have $\Delta \neq 0$, and have parameter values such that Equation (4) generates the data also have $P\left[y = k^* \mid y\right] = 0$ so that Equation (60) omits Equation (62) and only has two expectation terms given by Equations (67) and (68).

C.2.2 Conditional first moments of income pre- and post-kink

The conditional first moments of the variables y_0 and y_1 defined in Equation (5) can be written in three additive terms analogous to Equation (60) as,

$$\begin{aligned} E\left[y_j^* \mid y\right] &= E\left[y_j^* \mid y, y < k^*\right] P\left[y < k^* \mid y\right] \\ &+ E\left[y_j^* \mid y, y = k^*\right] P\left[y = k^* \mid y\right] \\ &+ E\left[y_j^* \mid y, y > k^*\right] P\left[y > k^* \mid y\right] \text{ for } j = \{0,1\}. \end{aligned} \tag{69}$$

The first additive expectation in Equation (69) for $j = 0$ is,

$$E\left[y_0^* \mid y, y < k^*\right] = y. \tag{70}$$

because conditional on $y < k^*$ and the level of y, I know that $y_0^* = y$. The second expectation term is,

$$E\left[y_0^* \mid y, y = k^*\right] = \varepsilon s_0 + X'\beta + E\left[\nu \mid y, y = k^*\right], \tag{71}$$

and the third term is,

$$E\left[y_0^* \mid y, y > k^*\right] = \varepsilon s_0 + X'\beta + E\left[\nu \mid y, y > k^*\right]. \tag{72}$$

The three terms of Equation (69) for $j = 1$ are similarly defined as,

$$E\left[y_1^* \mid y, y < k^*\right] = \varepsilon s_1 + X'\beta + E\left[\nu \mid y, y < k^*\right], \tag{73}$$

$$E\left[y_1^* \mid y, y = k^*\right] = \varepsilon s_1 + X'\beta + E\left[\nu \mid y, y = k^*\right], \tag{74}$$

and,

$$E\left[y_1^* \mid y, y > k^*\right] = y, \tag{75}$$

because conditional on $y > k^*$ and the level of y, I know that $y_1^* = y$.

The conditional expectation of ν given in Equations (71), or equivalently Equation (74), rely on the properties of the truncated bivariate normal and can be computed in closed form as,

$$E\left[\nu \mid y, y = k^*\right] = \rho\frac{\sigma_\nu}{\sigma_\psi}\left(y - k - Z'\gamma\right) - \sigma_\nu\sqrt{1-\rho^2}\left[\frac{\phi\left(b_1\right) - \phi\left(b_0\right)}{\Phi\left(b_1\right) - \Phi\left(b_0\right)}\right], \tag{76}$$

in which,

$$b_j = \frac{y - \varepsilon s_j - X'\beta - \rho\frac{\sigma_\nu}{\sigma_\psi}\left(y - k - Z'\gamma\right)}{\sigma_\nu\sqrt{1-\rho^2}} \text{ for } j = \{0,1\}. \tag{77}$$

The conditional expectations of ν given in Equations (72) and (73) require computing two integrals. The integral over ν conditional on a specific ψ can be computed in closed form using the properties of the truncated bivariate normal but the integral over ψ cannot. As such, I employ Gauss-Hermite quadrature to compute the integral that cannot be computed in closed form. The resulting computation of the conditional expectations of ν in Equations (72) and (73) are, respectively,

$$E\left[\nu \mid y, y > k^*\right] = \frac{\sum_{l=1}^{L}\omega_l \mathbb{1}\left(y > k + Z'\gamma + \psi_{1l}\right)\left(\frac{\rho\sigma_\nu}{\sigma_\psi}\psi_{1l} + \sigma_\nu\sqrt{1-\rho^2}\left[\frac{\phi\left(c_{1l}\right)}{1-\Phi\left(c_{1l}\right)}\right]\right)}{\sqrt{\pi}\Phi\left(a_1\right)}, \tag{78}$$

and

$$E\left[\nu \mid y, y < k^*\right] = \frac{\sum_{l=1}^{L}\omega_l \mathbb{1}\left(y < k + Z'\gamma + \psi_{0l}\right)\left(\frac{\rho\sigma_\nu}{\sigma_\psi}\psi_{0l} - \sigma_\nu\sqrt{1-\rho^2}\left[\frac{\phi\left(c_{0l}\right)}{\Phi\left(c_{0l}\right)}\right]\right)}{\sqrt{\pi}\left(1-\Phi\left(a_0\right)\right)}, \tag{79}$$

in which,

$$c_{jl} = \frac{k + Z'\gamma - \varepsilon s_j - X'\beta + \psi_{jl}\left(1 - \frac{\rho\sigma_\nu}{\sigma_\psi}\right)}{\sigma_\nu\sqrt{1-\rho^2}} \text{ for } j = \{0,1\}, l = 1\ldots L, \tag{80}$$

$$\psi_{jl} = \frac{\rho\sigma_\psi}{\sigma_\nu}\left(y - \varepsilon s_j - X'\beta\right) + \alpha_l\sigma_\psi\sqrt{2\left(1-\rho^2\right)} \text{ for } j = \{0,1\}, l = 1\ldots L, \tag{81}$$

a_j for $j = \{0, 1\}$ is from Equation (66), α_l in Equation (81) are the standard normal Gaussian abscissae and ω_l in Equations (78) and (79) are the corresponding weights for the L quadrature points for each $j = \{0, 1\}$.

I use Bayes rule to write the probability of each case in Equation (69). These three probability terms are defined in Equation (56), Equation (58), and Equation (59).

C.2.3 Conditional first moments of ability

The conditional first moment of the ability latent variable defined in Equation (5) can be written in three additive terms as,

$$\begin{aligned} E\left[n^* \mid y\right] &= E\left[n^* \mid y, y < k^*\right] P\left[y < k^* \mid y\right] \\ &+ E\left[n^* \mid y, y = k^*\right] P\left[y = k^* \mid y\right] \\ &+ E\left[n^* \mid y, y > k^*\right] P\left[y > k^* \mid y\right]. \end{aligned} \tag{82}$$

Using the definition in Equation (6a), I can write the three expectations in Equation (82) as,

$$E\left[n^* \mid y, y < k^*\right] = y - \varepsilon s_0, \tag{83}$$

$$E\left[n^* \mid y, y = k^*\right] = X'\beta + E\left[\nu \mid y, y = k^*\right], \tag{84}$$

and,

$$E\left[n^* \mid y, y > k^*\right] = y - \varepsilon s_1, \tag{85}$$

because conditional on the values of y and ε, I can observe n^* when $y > k^*$ and $y < k^*$.

I use Bayes rule to write the probability of each case in Equation (82). These three probability terms are defined in Equation (56), Equation (58), and Equation (59).

C.3 Conditional second moments

C.3.1 Conditional second moments of location

The conditional second moment of the location latent variable defined in Equation (5) can be written in three additive terms as,

$$\begin{aligned} E\left[k^{*2} \mid y\right] &= E\left[k^{*2} \mid y, y < k^*\right] P\left[y < k^* \mid y\right] \\ &+ E\left[k^{*2} \mid y, y = k^*\right] P\left[y = k^* \mid y\right] \\ &+ E\left[k^{*2} \mid y, y > k^*\right] P\left[y > k^* \mid y\right] \end{aligned} \tag{86}$$

Using our assumptions about n^* and k^* from Equation (51) and the derivation of the first conditional of ψ in Equation (64), the first expectation part of Equation (86) is,

$$E\left[k^{*2} \mid y, y < k^*\right] = (k + Z'\gamma)^2 + 2(k + Z'\gamma) E[\psi \mid y, y - k - Z'\gamma < \psi] + E\left[\psi^2 \mid y, y - k - Z'\gamma < \psi\right]. \tag{87}$$

The last term in Equation (87) is,

$$E\left[\psi^2 \mid y, y - k - Z'\gamma < \psi\right] = V[\psi \mid y, y - k - Z'\gamma < \psi] + E[\psi \mid y, y - k - Z'\gamma < \psi]^2 \tag{88}$$

in which,

$$V[\psi \mid y, y - k - Z'\gamma < \psi] = \sigma_\psi^2 \left(1 - \rho^2\right) \left[1 + a_0 \left(\frac{\phi(a_0)}{1 - \Phi(a_0)}\right) - \left(\frac{\phi(a_0)}{1 - \Phi(a_0)}\right)^2\right] \tag{89}$$

and

$$\begin{aligned} E[\psi \mid y, y - k - Z'\gamma < \psi]^2 &= \rho^2 \left(\frac{\sigma_\psi}{\sigma_\nu}\right)^2 (y_i - \varepsilon s_0 - X_i'\beta)^2 \\ &+ 2\rho \left(\frac{\sigma_\psi}{\sigma_\nu}\right) (y_i - \varepsilon s_0 - X_i'\beta)\, \sigma_\psi \sqrt{1 - \rho^2} \left(\frac{\phi(a_0)}{1 - \Phi(a_0)}\right) \\ &+ \sigma_\psi^2 \left(1 - \rho^2\right) \left(\frac{\phi(a_0)}{1 - \Phi(a_0)}\right)^2 \end{aligned} \tag{90}$$

therefore,

$$\begin{aligned} E\left[\psi^2 \mid y, y - k - Z'\gamma < \psi\right] &= \sigma_\psi^2 \left(1 - \rho^2\right) \left[1 + a_0 \left(\frac{\phi(a_0)}{1 - \Phi(a_0)}\right)\right] \\ &+ \rho^2 \left(\frac{\sigma_\psi}{\sigma_\nu}\right)^2 (y_i - \varepsilon s_0 - X_i'\beta)^2 \\ &+ 2\rho \left(\frac{\sigma_\psi}{\sigma_\nu}\right) (y_i - \varepsilon s_0 - X_i'\beta)\, \sigma_\psi \sqrt{1 - \rho^2} \left(\frac{\phi(a_0)}{1 - \Phi(a_0)}\right) \end{aligned} \tag{91}$$

in which a_0 comes from Equation (66) for j = 0.
Because conditional on y and $y = k^*$, k^* just is y, the second part is,

$$E\left[k^{*2} \mid y, y = k^*\right] = y^2 \tag{92}$$

or the square of the first conditional moment of location where $y = k$.

Using the derivation of the first conditional moment of ψ in Equation (65), the third expectation part of Equation (86) is

$$\begin{aligned} E\left[k^{*2} \mid y, y > k^*\right] &= (k + Z'\gamma)^2 + 2(k + Z'\gamma) E[\psi \mid y, y > k^*] \\ &+ E\left[\psi^2 \mid y, y - k - Z'\gamma > \psi\right]. \end{aligned} \tag{93}$$

The last term in Equation (93) is,

$$E\left[\psi^2 \mid y, y-k-Z'\gamma > \psi\right] = \sigma_\psi^2\left(1-\rho^2\right)\left[1+a_1\left(\frac{\phi(a_1)}{1-\Phi(a_1)}\right)\right] + \rho^2\left(\frac{\sigma_\psi}{\sigma_\nu}\right)^2\left(y_i-\varepsilon s_1-X_i'\beta\right)^2 + 2\rho\left(\frac{\sigma_\psi}{\sigma_\nu}\right)\left(y_i-\varepsilon s_1-X_i'\beta\right)\sigma_\psi\sqrt{1-\rho^2}\left(\frac{\phi(a_1)}{1-\Phi(a_1)}\right) \quad (94)$$

in which a_1 comes from Equation (66) for j = 1.

I use Bayes rule to write the probability of each case in Equation (86). These three probability terms are defined in Equation (56), Equation (58), and Equation (59).

C.3.2 Conditional second moments of income pre- and post-kink

The conditional second moments of the variables y_0 and y_1 defined in Equation (5) can be written in three additive terms analogous to Equation (86) as,

$$E\left[y_j^{*2} \mid y\right] = E\left[y_j^{*2} \mid y, y<k^*\right]P\left[y<k^* \mid y\right] + E\left[y_j^{*2} \mid y, y=k^*\right]P\left[y=k^* \mid y\right] + E\left[y_j^{*2} \mid y, y>k^*\right]P\left[y>k^* \mid y\right] \text{ for } j=\{0,1\}. \quad (95)$$

The first additive expectation in Equation (95) for $j = 0$ is,

$$E\left[y_0^{*2} \mid y, y<k^*\right] = y^2 \quad (96)$$

because conditional on $y < k^*$ and the level of y, I know that $y_0^* = y$, analogous to (70). The second expectation term is,

$$E\left[y_0^{*2} \mid y, y=k^*\right] = (\varepsilon s_0 + X'\beta)^2 + 2(\varepsilon s_0 + X'\beta)E\left[\nu \mid y, y=k^*\right] + E\left[\nu^2 \mid y, y=k^*\right]. \quad (97)$$

The last term in Equation (97) is

$$E\left[\nu^2 \mid y, y=k^*\right] = V\left[\nu \mid y, y=k^*\right] + E\left[\psi \mid y, y=k^*\right]^2 \quad (98)$$

in which,

$$V\left[\nu \mid y, y=k^*\right] = \sigma_\nu^2\left(1-\rho^2\right)\left[1-\left(\frac{b_1\phi(b_1)-b_0\phi(b_0)}{\Phi(b_1)-\Phi(b_0)}\right)-\left(\frac{\phi(b_1)-\phi(b_0)}{\Phi(b_1)-\Phi(b_0)}\right)^2\right] \quad (99)$$

and

$$E\left[\nu \mid y, y = k^*\right]^2 = \rho^2 \frac{\sigma_\nu^2}{\sigma_\psi^2} \left(y - k - Z'\gamma\right)^2 - 2\rho \frac{\sigma_\nu}{\sigma_\psi} \left(y - k - Z'\gamma\right) \sigma_\nu \sqrt{1-\rho^2} \left(\frac{\phi\left(b_1\right) - \phi\left(b_0\right)}{\Phi\left(b_1\right) - \Phi\left(b_0\right)}\right) + \sigma_\nu^2 \left(1-\rho^2\right) \left(\frac{\phi\left(b_1\right) - \phi\left(b_0\right)}{\Phi\left(b_1\right) - \Phi\left(b_0\right)}\right)^2 \quad (100)$$

therefore,

$$E\left[\nu^2 \mid y, y = k^*\right] = \sigma_\nu^2 \left(1-\rho^2\right) \left[1 - \frac{b_1 \phi\left(b_1\right) - b_0 \phi\left(b_0\right)}{\Phi\left(b_1\right) - \Phi\left(b_0\right)}\right] + \rho^2 \frac{\sigma_\nu^2}{\sigma_\psi^2} \left(y - k - Z'\gamma\right)^2 - 2\rho \frac{\sigma_\nu}{\sigma_\psi} \left(y - k - Z'\gamma\right) \sigma_\nu \sqrt{1-\rho^2} \left(\frac{\phi\left(b_1\right) - \phi\left(b_0\right)}{\Phi\left(b_1\right) - \Phi\left(b_0\right)}\right). \quad (101)$$

The third term is,

$$E\left[y_0^{*2} \mid y, y > k^*\right] = (\varepsilon s_0 + X'\beta)^2 + 2(\varepsilon s_0 + X'\beta) E\left[\nu \mid y, y > k^*\right] + E\left[\nu^2 \mid y, y > k^*\right]. \quad (102)$$

The three terms of Equation (95) for $j = 1$ are similarly defined as,

$$E\left[y_1^{*2} \mid y, y < k^*\right] = (\varepsilon s_1 + X'\beta)^2 + 2(\varepsilon s_1 + X'\beta) E\left[\nu \mid y, y < k^*\right] + E\left[\nu^2 \mid y, y < k^*\right], \quad (103)$$

$$E\left[y_1^{*2} \mid y, y = k^*\right] = (\varepsilon s_1 + X'\beta)^2 + 2(\varepsilon s_1 + X'\beta) E\left[\nu \mid y, y = k^*\right] + E\left[\nu^2 \mid y, y = k^*\right], \quad (104)$$

and

$$E\left[y_1^{*2} \mid y, y > k^*\right] = y^2. \quad (105)$$

The last term in Equation (104) is equal to the last term in Equation (97), which is defined in Equation (97). In regards to Equation (105), conditional on $y > k^*$ and the level of y, I know that $y_1^* = y$. This is analogous to Equation (75).

As in equations (78) and (79), the conditional expectations of ν given in Equations (102) and (103) require computing two integrals. Using Gauss-Hermite quadrature, the resulting computation of the conditional expectations of ν in Equations (102) and (103) are, respectively,

$$\begin{aligned} E\left[\nu^2 \mid y, y > k^*\right] &= \frac{1}{\sqrt{\pi}\Phi\left(a_1\right)}\left\{\sum_{l=1}^{L}\omega_l \mathbb{1}\left(y > k + Z'\gamma + \psi_{1l}\right)\right. \\ &\times \left[\sigma_\nu^2(1-\rho^2)\left(1 + \frac{c_{1l}\phi\left(c_{1l}\right)}{1-\Phi\left(c_{1l}\right)} - \left(\frac{\phi\left(c_{1l}\right)}{1-\Phi\left(c_{1l}\right)}\right)^2\right)\right. \\ &+ \left.\left.\left(\frac{\rho\sigma_\nu}{\sigma_\psi}\psi_{1l} + \sigma_\nu\sqrt{1-\rho^2}\frac{\phi\left(c_{1l}\right)}{1-\Phi\left(c_{1l}\right)}\right)^2\right]\right\} \end{aligned} \tag{106}$$

and

$$\begin{aligned} E\left[\nu^2 \mid y, y < k^*\right] &= \frac{1}{\sqrt{\pi}\left(1-\Phi\left(a_0\right)\right)}\left\{\sum_{l=1}^{L}\omega_l \mathbb{1}\left(y < k + Z'\gamma + \psi_{0l}\right)\right. \\ &\times \left[\sigma_\nu^2(1-\rho^2)\left(1 + \frac{c_{0l}\phi\left(c_{0l}\right)}{\Phi\left(c_{0l}\right)} - \left(\frac{\phi\left(c_{0l}\right)}{\Phi\left(c_{0l}\right)}\right)^2\right)\right. \\ &+ \left.\left.\left(\frac{\rho\sigma_\nu}{\sigma_\psi}\psi_{0l} - \sigma_\nu\sqrt{1-\rho^2}\frac{\phi\left(c_{0l}\right)}{\Phi\left(c_{0l}\right)}\right)^2\right]\right\} \end{aligned} \tag{107}$$

in which c_{0l} and c_{1l} come from Equation (80) for $j = \{0, 1\}$.

I use Bayes rule to write the probability of each case in Equation (95). These three probability terms are defined in Equation (56), Equation (58), and Equation (59).

C.3.3 Conditional second moments of ability

The conditional second moment of the ability latent variable defined in Equation (5) can be written in three additive terms as,

$$\begin{aligned} E\left[n^{*2} \mid y\right] &= E\left[n^{*2} \mid y, y < k^*\right] P\left[y < k^* \mid y\right] \\ &+ E\left[n^{*2} \mid y, y = k^*\right] P\left[y = k^* \mid y\right] \\ &+ E\left[n^{*2} \mid y, y > k^*\right] P\left[y > k^* \mid y\right] \end{aligned} \tag{108}$$

Using the definition in Equation (6a), I can write the three expectations in Equation (82) as,

$$E\left[n^{*2} \mid y, y < k^*\right] = (y - \varepsilon s_0)^2, \tag{109}$$

$$E\left[n^{*2} \mid y, y = k^*\right] = (X'\beta)^2 + 2X'\beta E\left[\nu \mid y, y = k^*\right] + E\left[\nu^2 \mid y, y = k^*\right], \tag{110}$$

and

$$E\left[n^{*2} \mid y, y > k^*\right] = (y - \varepsilon s_1)^2. \tag{111}$$

The second term and third term in Equation (110) are defined in Equations (76) and (97), respectively.

I use Bayes rule to write the probability of each case in Equation (108). These three probability terms are defined in Equation (56), Equation (58), and Equation (59).

C.3.4 Conditional mixed second moments of location and income pre- and post-kink

The conditional second mixed moments of the variables y_0 and y_1 with the location latent variable defined in (5) defined in Equation (5) can be written in three additive terms analogous to Equations (69) and (60) as,

$$\begin{aligned} E\left[y_j^* k^* \mid y\right] &= E\left[y_j^* k^* \mid y, y < k^*\right] P\left[y < k^* \mid y\right] \\ &+ E\left[y_j^* k^* \mid y, y = k^*\right] P\left[y = k^* \mid y\right] \\ &+ E\left[y_j^* k^* \mid y, y > k^*\right] P\left[y > k^* \mid y\right] \text{ for } j = \{0, 1\}. \end{aligned} \tag{112}$$

The first additive expectation in Equation (112) for $j = 0$ and $j = 1$ is, respectively,

$$E\left[y_0^* k^* \mid y, y < k^*\right] = y(k + Z'\gamma + E\left[\psi \mid y, y < k^*\right]) \tag{113}$$

and

$$\begin{aligned} E\left[y_1^* k^* \mid y, y < k^*\right] &= (k + Z'\gamma)(\varepsilon s_1 + X'\beta) \\ &+ (k + Z'\gamma) E\left[\nu \mid y, y < k^*\right] \\ &+ (\varepsilon s_1 + X'\beta) E\left[\psi \mid y, y < k^*\right]) \\ &+ E\left[\nu\psi \mid y, y < k^*\right]. \end{aligned} \tag{114}$$

in which $E\left[\psi \mid y, y < k^*\right]$ is defined in Equation (64), $E\left[\nu \mid y, y < k^*\right]$ is defined in Equation (79), and $E\left[\nu\psi \mid y, y < k^*\right]$ is defined in Equation (125).

The second part for $j = 0$ and $j = 1$ is, respectively,

$$E\left[y_0^* k^* \mid y, y = k^*\right] = y(\varepsilon s_0 + X'\beta + E\left[\nu \mid y, y = k^*\right]) \tag{115}$$

$$E\left[y_1^* k^* \mid y, y = k^*\right] = y(\varepsilon s_1 + X'\beta + E\left[\nu \mid y, y = k^*\right]) \tag{116}$$

in which $E\left[\nu \mid y, y = k^*\right]$ is defined in Equation (76).

The third part is,

$$\begin{aligned} E\left[y_0^* k^* \mid y, y > k^*\right] &= (k + Z'\gamma)(\varepsilon s_0 + X'\beta) \\ &+ (k + Z'\gamma) E\left[\nu \mid y, y > k^*\right] \\ &+ (\varepsilon s_0 + X'\beta) E\left[\psi \mid y, y > k^*\right] \\ &+ E\left[\nu\psi \mid y, y > k^*\right] \end{aligned} \tag{117}$$

and

$$E\left[y_1^* k^* \mid y, y > k^*\right] = y(k + Z'\gamma + E\left[\psi \mid y, y > k^*\right]), \tag{118}$$

in which $E\left[\nu \mid y, y > k^*\right]$ is defined in Equation (78), $E\left[\psi \mid y, y > k^*\right]$ is defined in Equation (65), and $E\left[\psi \mid y, y > k^*\right]$ is defined in Equation (127).

I use Bayes rule to write the probability of each case in Equation (112). These three probability terms are defined in Equation (56), Equation (58), and Equation (59).

The second mixed moment of pre-tax income with post-tax income is,

$$E\left[y_0^* y_1^* \mid y\right] = \varepsilon(s_1 - s_0) E\left[y_0^* \mid y\right] + E\left[y_0^{*2} \mid y\right] \tag{119}$$

in which $E\left[y_0^{*2} \mid y\right]$ is defined in Equation (95) for $j = 0$ and $E\left[y_0^* \mid y\right]$ is defined in Equation (69) for $j = 0$.

C.4 Conditional mixed second moments of ability and location

The conditional second mixed moment of the ability latent variable with the location latent variable, both defined in Equation (5), can be written in three additive terms as,

$$\begin{aligned} E\left[n^* k^* \mid y\right] &= E\left[n^* k^* \mid y, y < k^*\right] P\left[y < k^* \mid y\right] \\ &+ E\left[n^* k^* \mid y, y = k^*\right] P\left[y = k^* \mid y\right] \\ &+ E\left[n^* k^* \mid y, y > k^*\right] P\left[y > k^* \mid y\right] \end{aligned} \tag{120}$$

Given our assumptions that $n^* = X'\beta + \nu$ and $k^* = k + Z'\gamma + \psi$, I can apply these to the parts of Equation (120) where,

$$E\left[n^* k^* \mid y, y < k^*\right] = (y - \varepsilon s_0)(k + Z'\gamma + E\left[\psi \mid y, y < k^*\right]), \tag{121}$$

$$E[n^*k^* \mid y, y = k^*] = y(X'\beta + E[\nu \mid y, y = k^*]), \tag{122}$$

and

$$E[n^*k^* \mid y, y > k^*] = (y - \varepsilon s_1)(k + Z'\gamma + E[\psi \mid y, y > k^*]) \tag{123}$$

in which $E[\psi \mid y, y < k^*])$ is defined in Equation (64), $E[\nu \mid y, y = k^*]$ is defined in Equation (76), and $E[\psi \mid y, y > k^*])$ is defined in Equation (65).

I use Bayes rule to write the probability of each case in Equation (120). These three probability terms are defined in Equation (56), Equation (58), and Equation (59).

C.5 Conditional mixed second moments of $\nu\psi$

The conditional second mixed moment of ν with ψ uses properties of the joint bivariate normal distribution of these variables defined in (7) such that they can be written in three additive terms as,

$$\begin{aligned} E[\nu\psi \mid y] &= E[\nu\psi \mid y, y < k^*] P[y < k^* \mid y] \\ &+ E[\nu\psi \mid y, y = k^*] P[y = k^* \mid y] \\ &+ E[\nu\psi \mid y, y > k^*] P[y > k^* \mid y]. \end{aligned} \tag{124}$$

The first part is,

$$E[\nu\psi \mid y, y < k^*] = E[\nu \mid y, y < k^*] \psi_{0l}, \tag{125}$$

in which $E[\nu \mid y, y < k^*]$ comes from Equation (79).

The second part is,

$$E[\nu\psi \mid y, y = k^*] = y(\varepsilon s_0 + X'\beta + E[\nu \mid y, y = k^*]) \tag{126}$$

in which $E[\nu \mid y, y = k^*]$ comes from Equation (76).

The third part is,

$$E[\nu\psi \mid y, y > k^*] = E[\nu \mid y, y > k^*] \psi_{1l}, \tag{127}$$

in which $E[\nu \mid y, y > k^*]$ is defined in Equation (78).

I use Bayes rule to write the probability of each case in Equation (124). These three probability terms are defined in Equation (56), Equation (58), and Equation (59).

C.6 Conditional second centered moments of location, pre- and post-tax income, and ability

The conditional second centered moments of k^* can be written as

$$V\left[k^{*}\mid y\right]=E\left[k^{*2}\mid y\right]-E\left[k^{*}\mid y\right]^{2} \tag{128}$$

and for y_0 and y_1,

$$V\left[y_j^{*}\mid y\right]=E\left[y_j^{*2}\mid y\right]-E\left[y_j^{*}\mid y\right]^{2} \text{ for } j=\{0,1\} \tag{129}$$

in which $E\left[y_j^{*2}\mid y\right]$ and $E\left[y_j^{*}\mid y\right]$ come from Equations (95) and (69), respectively.

Finally, I can use definitions of the first and second moments of ability to define the second centered moment of ability,

$$V\left[n^{*}\mid y\right]=E\left[n^{*2}\mid y\right]-E\left[n^{*}\mid y\right]^{2} \tag{130}$$

in which $E\left[n^{*2}\mid y\right]$ and $E\left[n^{*}\mid y\right]$ come from Equations (108) and (82), respectively.

Notch Thresholds

Case 1: $N_i^{*}=\bar{N}$

$$C_k-\frac{\bar{N}}{1+\frac{1}{\varepsilon}}\left(\frac{Y_k}{\bar{N}}\right)^{1+\frac{1}{\varepsilon}}-\bar{C}+\frac{\bar{N}}{1+\frac{1}{\varepsilon}}\left(\frac{\bar{Y}}{\bar{N}}\right)^{1+\frac{1}{\varepsilon}}=0$$

$$\bar{Y}-\bar{N}\left(1-t_1\right)^{\varepsilon}=0$$

$$\bar{C}-I_0-\left(1-t_0\right)K+\Delta-\left(1-t_1\right)(\bar{Y}-K)=0$$

$$C_k-I_0-\left(1-t_0\right)K=0$$

$Y_k - K = 0$

$$\bar{C} = I_0 + (1-t_0)K - \Delta + (1-t_1)\left(\bar{N}(1-t_1)^{\varepsilon} - K\right)$$

$$I_0 + (1-t_0)K - \frac{\bar{N}}{1+\frac{1}{\varepsilon}}\left(\frac{K}{\bar{N}}\right)^{1+\frac{1}{\varepsilon}} - \bar{C} + \frac{\bar{N}}{1+\frac{1}{\varepsilon}}\left(\frac{\bar{N}(1-t_1)^{\varepsilon}}{\bar{N}}\right)^{1+\frac{1}{\varepsilon}} = 0$$

$$-\frac{\bar{N}}{1+\frac{1}{\varepsilon}}\left(\frac{K}{\bar{N}}\right)^{1+\frac{1}{\varepsilon}} + \Delta - (1-t_1)\left(\bar{N}(1-t_1)^{\varepsilon} - K\right) + \frac{\bar{N}}{1+\frac{1}{\varepsilon}}\left(\frac{\bar{N}(1-t_1)^{\varepsilon}}{\bar{N}}\right)^{1+\frac{1}{\varepsilon}} = 0$$

$$-\frac{\bar{N}^{-\frac{1}{\varepsilon}}K^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}} + \Delta - (1-t_1)(1-t_1)^{\varepsilon}\bar{N} + (1-t_1)K + \frac{\bar{N}^{-\frac{1}{\varepsilon}}\bar{N}^{1+\frac{1}{\varepsilon}}(1-t_1)^{\varepsilon+1}}{1+\frac{1}{\varepsilon}} = 0$$

$$-\frac{\bar{N}^{-\frac{1}{\varepsilon}}K^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}} + \Delta - (1-t_1)(1-t_1)^{\varepsilon}\bar{N} + (1-t_1)K + \frac{\bar{N}(1-t_1)^{\varepsilon+1}}{1+\frac{1}{\varepsilon}} = 0$$

$$-\frac{\bar{N}^{-\frac{1}{\varepsilon}}K^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}} - (1-t_1)(1-t_1)^{\varepsilon}\bar{N} + \frac{\bar{N}(1-t_1)^{\varepsilon+1}}{1+\frac{1}{\varepsilon}} + (1-t_1)K + \Delta = 0$$

$$-\left(\frac{K^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}}\right)\bar{N}^{-\frac{1}{\varepsilon}} + \left(-(1-t_1)^{\varepsilon+1} + \frac{(1-t_1)^{\varepsilon+1}}{1+\frac{1}{\varepsilon}}\right)\bar{N} + (1-t_1)K + \Delta = 0$$

Case 2: $N_i^* = \underline{N}$

$C_k - \frac{\underline{N}}{1+\frac{1}{\varepsilon}}\left(\frac{Y_k}{\underline{N}}\right)^{1+\frac{1}{\varepsilon}} - \underline{C} + \frac{\underline{N}}{1+\frac{1}{\varepsilon}}\left(\frac{\underline{Y}}{\underline{N}}\right)^{1+\frac{1}{\varepsilon}} = 0$

$\underline{Y} - \underline{N}(1-t_0)^{\varepsilon} = 0$

$\underline{C} - I_0 - (1-t_0)\underline{Y} = 0$

$C_k - I_0 - (1-t_0)K + \Delta = 0$

$Y_k - K = 0$

$$\begin{aligned} \underline{C} &= I_0 + (1-t_0)(1-t_0)^{\varepsilon}\underline{N} \\ &= I_0 + (1-t_0)^{\varepsilon+1}\underline{N} \end{aligned}$$

$$
\begin{aligned}
I_0 + (1-t_0)\,K - \Delta - \frac{\underline{N}}{1+\frac{1}{\varepsilon}}\left(\frac{K}{\underline{N}}\right)^{1+\frac{1}{\varepsilon}} - \underline{C} + \frac{\underline{N}}{1+\frac{1}{\varepsilon}}\left(\frac{\underline{N}\,(1-t_0)^{\varepsilon}}{\underline{N}}\right)^{1+\frac{1}{\varepsilon}} &= 0 \\
I_0 + (1-t_0)\,K - \Delta - \frac{\underline{N}}{1+\frac{1}{\varepsilon}}\left(\frac{K}{\underline{N}}\right)^{1+\frac{1}{\varepsilon}} - I_0 - (1-t_0)^{\varepsilon+1}\,\underline{N} + \frac{\underline{N}}{1+\frac{1}{\varepsilon}}\left(\frac{\underline{N}\,(1-t_0)^{\varepsilon}}{\underline{N}}\right)^{1+\frac{1}{\varepsilon}} &= 0 \\
(1-t_0)\,K - \Delta - \frac{\underline{N}}{1+\frac{1}{\varepsilon}}\left(\frac{K}{\underline{N}}\right)^{1+\frac{1}{\varepsilon}} - (1-t_0)^{\varepsilon+1}\,\underline{N} + \frac{\underline{N}}{1+\frac{1}{\varepsilon}}\left(\frac{\underline{N}\,(1-t_0)^{\varepsilon}}{\underline{N}}\right)^{1+\frac{1}{\varepsilon}} &= 0 \\
(1-t_0)\,K - \Delta - \frac{\underline{N}\underline{N}^{-1-1/\varepsilon}}{1+\frac{1}{\varepsilon}}K^{1+\frac{1}{\varepsilon}} - (1-t_0)^{\varepsilon+1}\,\underline{N} + \frac{\underline{N}}{1+\frac{1}{\varepsilon}}\,(1-t_0)^{\varepsilon+1} &= 0 \\
\left[(1-t_0)\,K - \Delta\right] - \left(\frac{K^{1+\frac{1}{\varepsilon}}}{1+\frac{1}{\varepsilon}}\right)\underline{N}^{-1/\varepsilon} + \underline{N}\left(\frac{(1-t_0)^{\varepsilon+1}}{1+\frac{1}{\varepsilon}} - (1-t_0)^{\varepsilon+1}\right) &= 0
\end{aligned}
$$