Bunching at the kink: Implications for spending responses to health insurance contracts

Liran Einav, Amy Finkelstein, Paul Schrimpf

ARTICLE INFO

Article history:
Received 23 June 2016
Received in revised form 16 November 2016
Accepted 16 November 2016
Available online 26 November 2016

JEL classification:
D12
G22

Keywords:
Bunching
Medicare
Health insurance
Health care

1. Introduction

Over the last decade, there has been an increased reliance in public economics on evidence that is based on observed “bunching” around kink points in budget sets; Kleven (2016) provides an overview of this growing literature. The key underlying idea is simple and tractable: if rational individuals face a non-linear budget set with considerable kinks, they should bunch around the kinks, and the extent of bunching should be informative about relevant elasticities (or lack thereof). The existence of bunching or excess mass around kink points of a budget set can thus provide compelling, visual evidence against the null hypothesis of no behavioral response of individual to the incentives; likewise, the lack of such bunching suggests the opposite.

Many of the applications of this idea have been in the context of the behavioral response to non-linear income tax schedules (Chetty et al., 2011, Saez, 2010; Chetty et al., 2013a; Bastani and Selin, 2014, Kleven and Waseem, 2013, Kleven et al., 2014). But similar ideas have been widely applied in other settings that generate non-linear budget sets, including pensions (Manoli and Weber, forthcoming), electricity (Ito, 2014), fuel economy policy (Sallee and Slemrod, 2012), mortgages (Best et al., 2015), cell phones (Grubb, 2015; Grubb and Osborne, 2015), broadband (Nevo et al., 2016), taxes on home sales (Kopczuk and Munroe, 2015; Best and Kleven, forthcoming), healthcare procurement (Bajari et al., 2016), and – the subject of this current paper – health insurance contracts (Abaluck et al., 2015; Dalton et al., 2015; Einav et al., 2015).

A likely key factor behind this recent popularity of bunching estimates is the seminal contribution of Saez (2010), which illustrates how one may convert an observed bunching pattern to an economic object of interest: a “structural” behavioral elasticity parameter. Using data on individuals’ annual earning, which bunch around convex kinks in the income tax schedule, Saez used a stylized, static, frictionless model of labor supply to provide a simple, transparent, and easy-to-implement mapping from the observed bunching pattern to an estimate of the elasticity of labor supply (or earning) with respect to...
the marginal tax rate. This allows one to take the compelling visual evidence of bunching and move beyond merely rejecting the null of non-behavioral response to estimating a quantitative economic object of interest that can be used to predict behavioral responses to counterfactual scenarios. Not surprisingly, this compelling and tractable idea has been quite influential, and has been frequently used to translate various bunching estimates into “structural” elasticities (Bastani and Selin, 2014; Chetty et al., 2011; Kleven et al., 2011; Kleven and Waseem, 2013).

The Saez (2010) approach is very appealing. It is transparent and easy to implement. Of course, the simplicity comes at the cost of potentially abstracting from a host of real-world features that may be important in a particular context. An alternative to this approach would be to develop a more complete model of a given context, which includes dynamics, uncertainty, and other relevant frictions. Manoli and Weber (forthcoming) provide such a model in the context of labor supply, and our earlier work (Einav et al., 2015) provides another example in the context of demand for prescription drugs. As the Saez approach is so much simpler and easier to implement, it seems useful to ask how well of an approximation to the main object of interest, a simpler, Saez-style approach can provide. Naturally, the reasonableness of the approximation will depend on the specific context.

This is precisely the goal of the current paper, where we explore this question in the context of demand for prescription drugs under Medicare Part D, the public prescription drug insurance program for elderly and disabled individuals in the United States. Our substantive question concerns the spending (or “moral hazard”) effects of alternative insurance contracts. This is a topic that has attracted considerable attention both for health insurance contracts in general, and more recently in our specific Part D context.

We begin in Section 2 by describing the setting and the data. An important feature of Medicare Part D coverage is the donut hole in the basic benefit design, which generates a large, discontinuous increase in the marginal price. Consequently, individuals’ annual drug expenditures bunch around this kink, making it a natural context to explore the implication of different bunching estimates.

In Section 3 we present and estimate two different models of prescription drug purchasing behavior. The first is our adaptation of the static, frictionless Saez (2010) model to the Medicare Part D context; we refer to it hereafter as a Saez-style model and the resultant elasticities as Saez-based elasticities. The second is the dynamic model we developed in our earlier work (Einav et al., 2015); we refer to it hereafter as the dynamic model and the resultant elasticities as the dynamics-based elasticities. Both models match the basic bunching pattern; however, the implied elasticity from the dynamic model is an order of magnitude greater than the Saez-based elasticity estimate. This is the key result of the paper.

There are multiple differences between the two, non-nested settings. The Saez-style framework assumes continuous spending decisions (i.e. no lumpiness in drug purchases), perfect foresight of future health shocks (i.e. no uncertainty), and no discounting of the future. None of these assumptions are made in the dynamic model. It is interesting to explore which features of the model are most important for the differences in implied elasticities, which is the focus of Section 4. There we develop two modifications of the dynamic model, which bring it closer to the Saez-style framework. We then re-estimate each of these versions of the model using the same data. Our main finding is that a static, perfect foresight version of the full dynamic model – which comes quite close to the Saez-style model except that it allows for lumpiness in spending decisions – results in implied elasticities that are about half way between the Saez-based estimates and the dynamics-based ones. Interestingly, once we allow for lumpiness, allowing for uncertainty essentially allows us to recover the magnitude of the elasticity implied by the full dynamic model; as it turns out, allowing for discounting is not quantitatively important.

We emphasize that the results we present in this paper should be viewed as illustrative. They are specific to our particular (Medicare Part D) context, as well as to the modeling choices we have made. Nonetheless, they highlight what we believe to be an important and broader point: in-sample bunching patterns may be rationalized by a host of modeling assumptions, and these assumptions can, at least in some contexts, have very different quantitative implications for the out-of-sample objects of interest.1 This is a general issue that the “bunching” literature has grappled with: the immense sensitivity of results to the particular modeling assumptions used to translate the observed excess mass at a convex kink into an economic object that can be used in other contexts. Indeed, in an early, working paper version of his original contribution, Saez (1999) shows that the translation of the excess mass at kinks in the US income tax schedule to an underlying labor supply elasticity can be greatly affected by adding to the baseline model either earnings uncertainty or making labor supply choices discrete and lumpy rather than continuous.

This is essentially an under-identification problem, which is more formally characterized by Blomquist et al. (2015), who emphasize the need for additional moments in the data that would allow us to select among models. The subsequent bunching literature has broadly pursued two main strategies for handling this identification issue. One is to use the frequency of observations in dominated regions of the budget set (“notches”) to identify the extent of frictions; Kleven and Waseem (2013) develop this approach. Another is to parametrize the frictions – or other relevant features of the setting – and use additional moments in the data to identify them. In the working paper version of Chetty et al. (2011), the authors explore such an approach, developing a labor supply model and parameter-search costs that have quantitatively important implications for translating observed excess mass into an underlying, “structural” elasticity (Chetty et al., 2009). The exercise in our paper is similar in spirit; our focus is less on “frictions” but more on the dynamic nature of our problem, and we therefore use non-bunching moments, associated with the way spending patterns vary over the calendar year, to identify additional parameters of the dynamic model.

More generally, our paper speaks to the growing interest in our profession in developing approaches to translate compelling, transparent, “reduced form” evidence of a behavioral response into an economic object of interest. The bunching literature following Saez (2010) is one specific application of the influential “sufficient statistics” literature popularized by Chetty (2009) – which attempts to use simple models to directly and transparently map reduced form parameters into welfare analyses. But the phenomenon is more general. For example, randomized controlled trials have the ability to deliver compelling “causal effect” estimates, but translating the experimental treatment effects into economic objects that can be applied out-of-sample to make counterfactual predictions or analyses often requires additional economic modeling assumptions (Aron-Dine et al., 2013). Our (modest) goal here is to illustrate in a particular context that these modeling assumptions can be quite consequential. As we have demonstrated, two “reasonable” (in our subjective view) alternative models can match the basic reduced form bunching facts, while giving very different out-of-sample predictions. Sufficient statistics, in other words, are sufficient conditional on the model (or a set of models). This is an obvious point, made clearly by Chetty (2009), but it is sometimes forgotten in applications and interpretations.

1 Kleven (2016) emphasizes a similar point in his review article. In the labor supply context, he describes several empirical papers which show that, with potential optimization frictions, a given excess mass in the earnings distribution can be rationalized with virtually any underlying, “structural” labor supply elasticity.
دریافت فوری متن کامل مقاله

امکان دانلود نسخه تمام متن مقالات انگلیسی
امکان دانلود نسخه ترجمه شده مقالات
پذیرش سفارش ترجمه تخصصی
امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
امکان دانلود رایگان ۲ صفحه اول هر مقاله
امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
دانلود فوری مقاله پس از پرداخت آنلاین
پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات