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

Liran Einav[1] Amy Finkelstein[2] **Paul Schrimpf**[3]

[1]Stanford

[2]MIT

[3]UBC

September 16, 2016

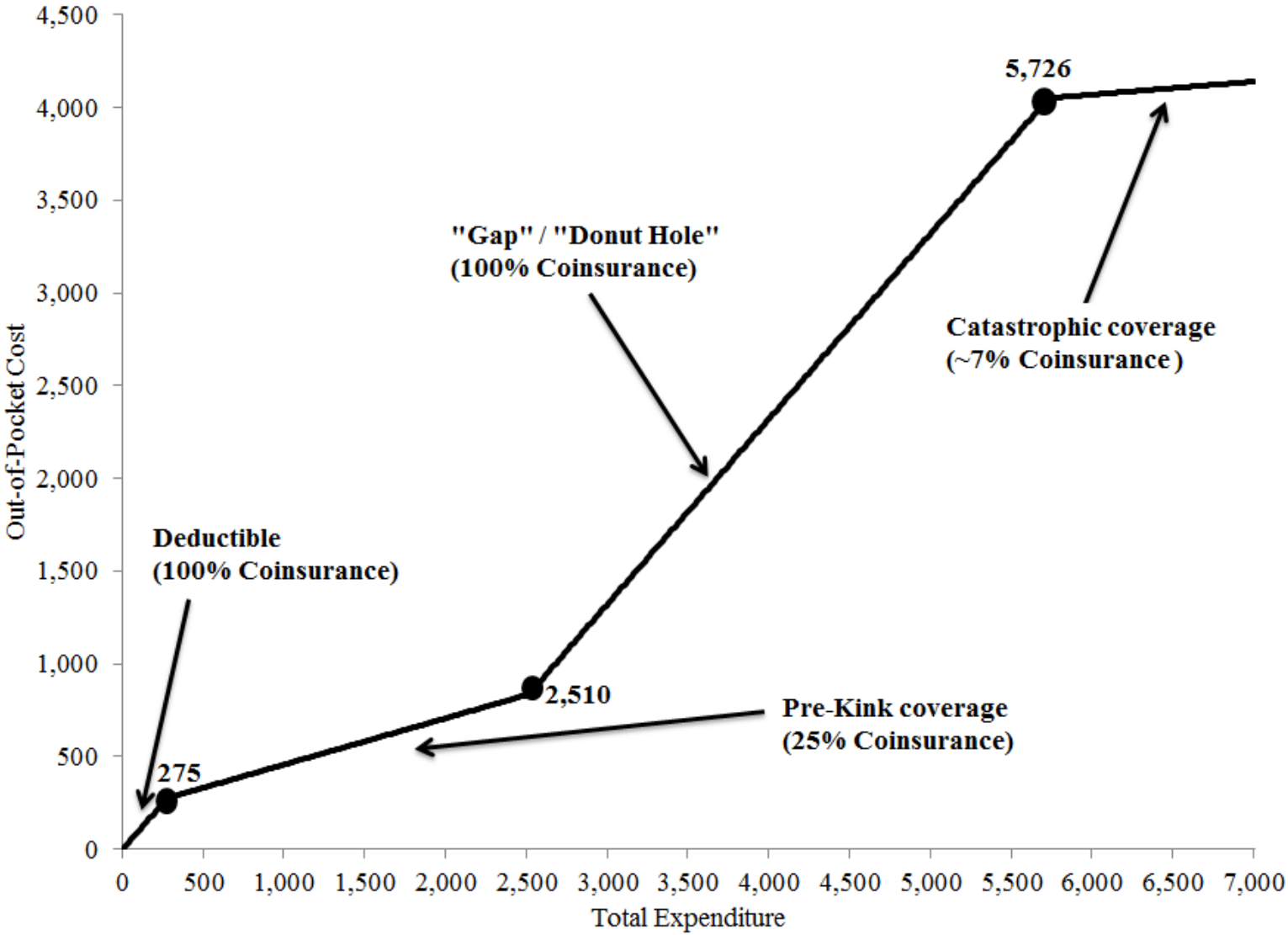
Motivation

- ”Credibility revolution” in economics
  - (Rightly) emphasize value of research design that produces compelling (often visual) evidence of a behavioral response
- ”Structural” models
  - (Rightly) emphasize defining and estimating economic objects that can be used to predict behavior in counterfactual environments
- ”Sufficient statistics” (Chetty, 2009)
  - Use simple models to directly and transparently map reduced form parameters into economic objects of interest
- Our paper: simple (not novel) point: choice of model can be consequential
  - Two ”reasonable” models can match the reduced form facts but produce very different counterfactual predictions

Specific application: Bunching

- Increased analysis in public economics of "bunching" at kink points in convex budget sets (Kleven 2016 Annual Reviews)
  - Existence of bunching (or "excess mass") can provide compelling, visual evidence against null of no behavioral response to incentives
  - Magnitude of excess mass often used to infer relevant elasticities
- Many applications with non-linear schedules: income taxes, home sale taxes, pensions, electricity, fuel economy, mortgages, cell phones, ...
- Two factors behind recent popularity:
  - Detecting bunching: Increased availability of rich, large administrative data
  - Interpreting bunching: Saez (2010) seminal paper
    - ★ Illustrates how to translate observed bunching into a "structural" behavioral elasticity parameter

Our context: Medicare Part D



Spending response to health

- Moral hazard in health economics: how does spending respond to health insurance contract design (e.g. consumer cost sharing)
  - Implications for health care spending and public sector budgets
- ”Donut hole” creates a large, discontinuous increase in marginal price of drugs
  - Allows us to link behavioral response (bunching) to the change in price
- Our specific question here: elasticity of spending with respect to a uniform change in cost sharing across the budget set
- Our approach: Estimate two different models of prescription drug purchasing
  - Both match the basic ”bunching at the kink”
  - But produce very different elasticity estimates

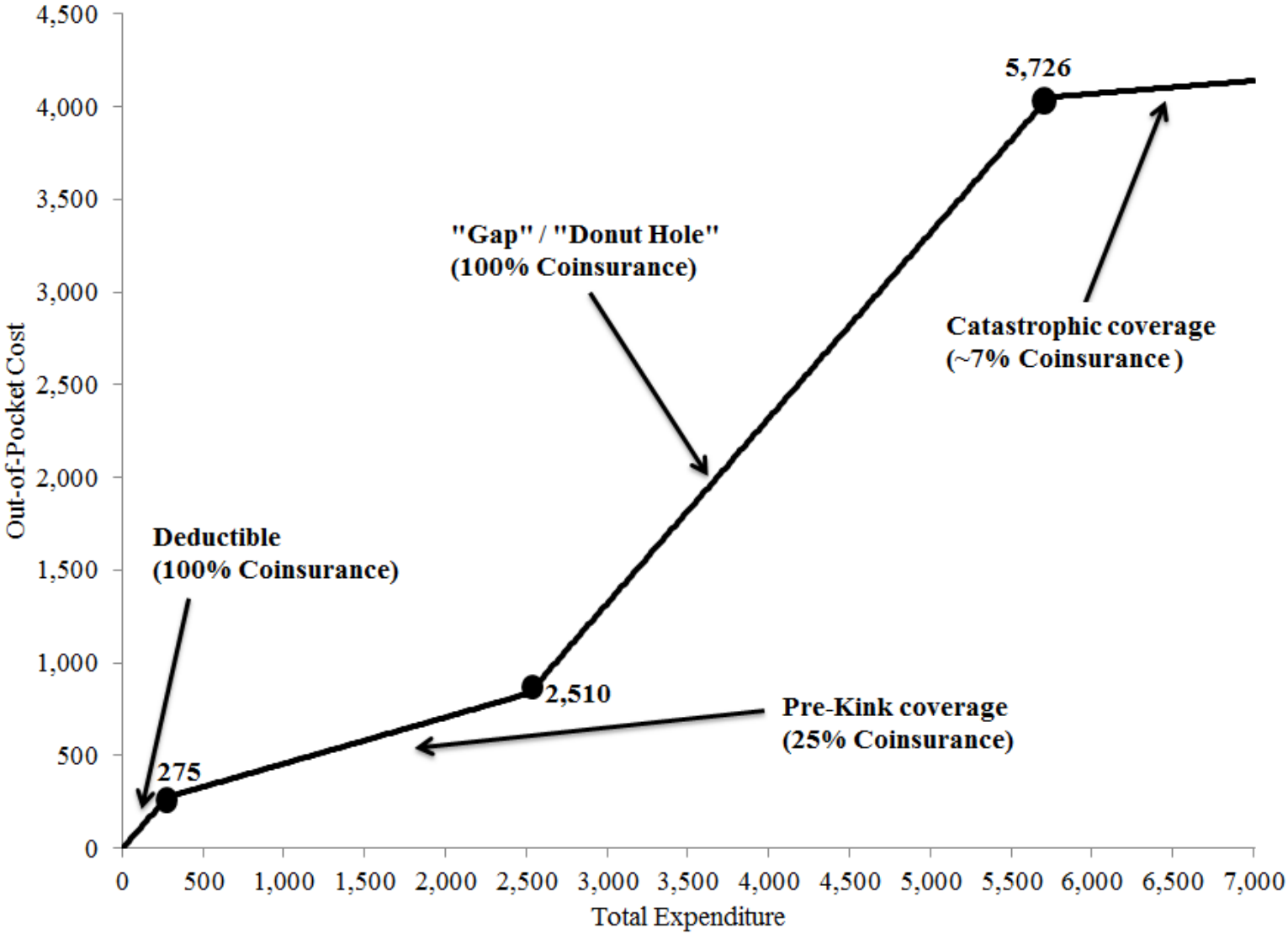
Outline

1. Data and Setting
2. Bunching estimates
3. Estimate two different models
   - Static, frictionless model (adaptation of Saez 2010)
   - Dynamic model (EFS 2015)
4. Conclusions

Medicare Part D

- Prescription drug coverage introduced in 2006
- Largest expansion of Medicare since inception
  - 32 million beneficiaries, 11% of Medicare spending
- Typical coverage highly non-linear
  - Government sets standard plan; actual (private) plans vary coverage details around this basic design

Standard plan



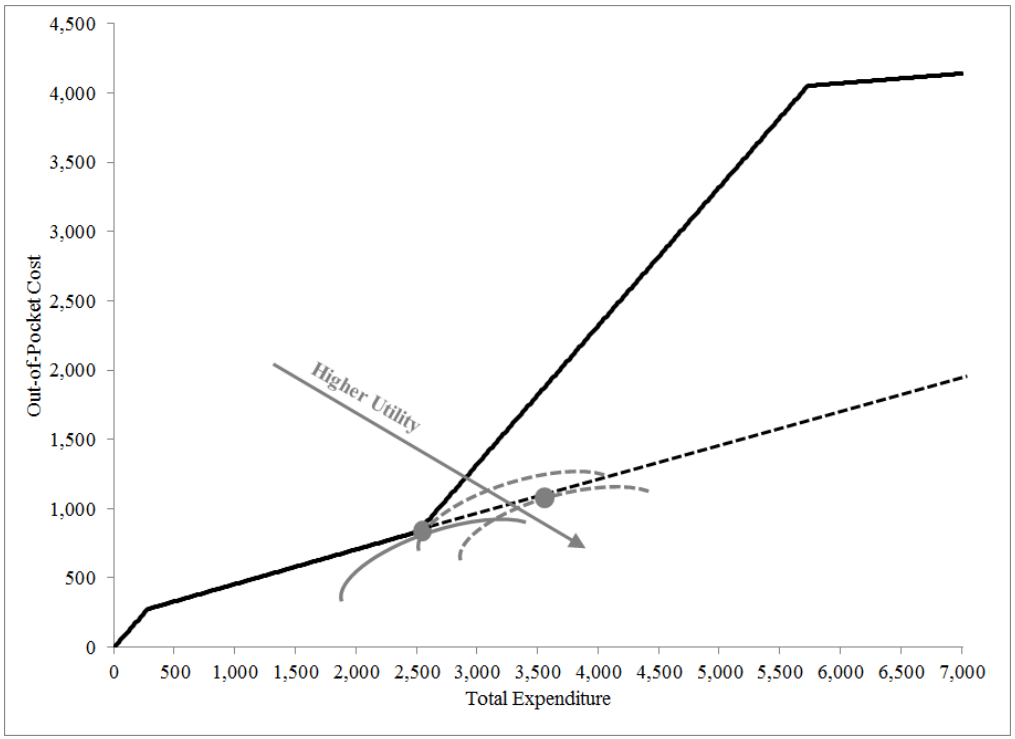
Data and sample

- 20% random sample of all Part D-covered individuals (2007 - 2009)
- Cost sharing features of each plan, basic demographics, and detailed, claim-level information on drugs purchased

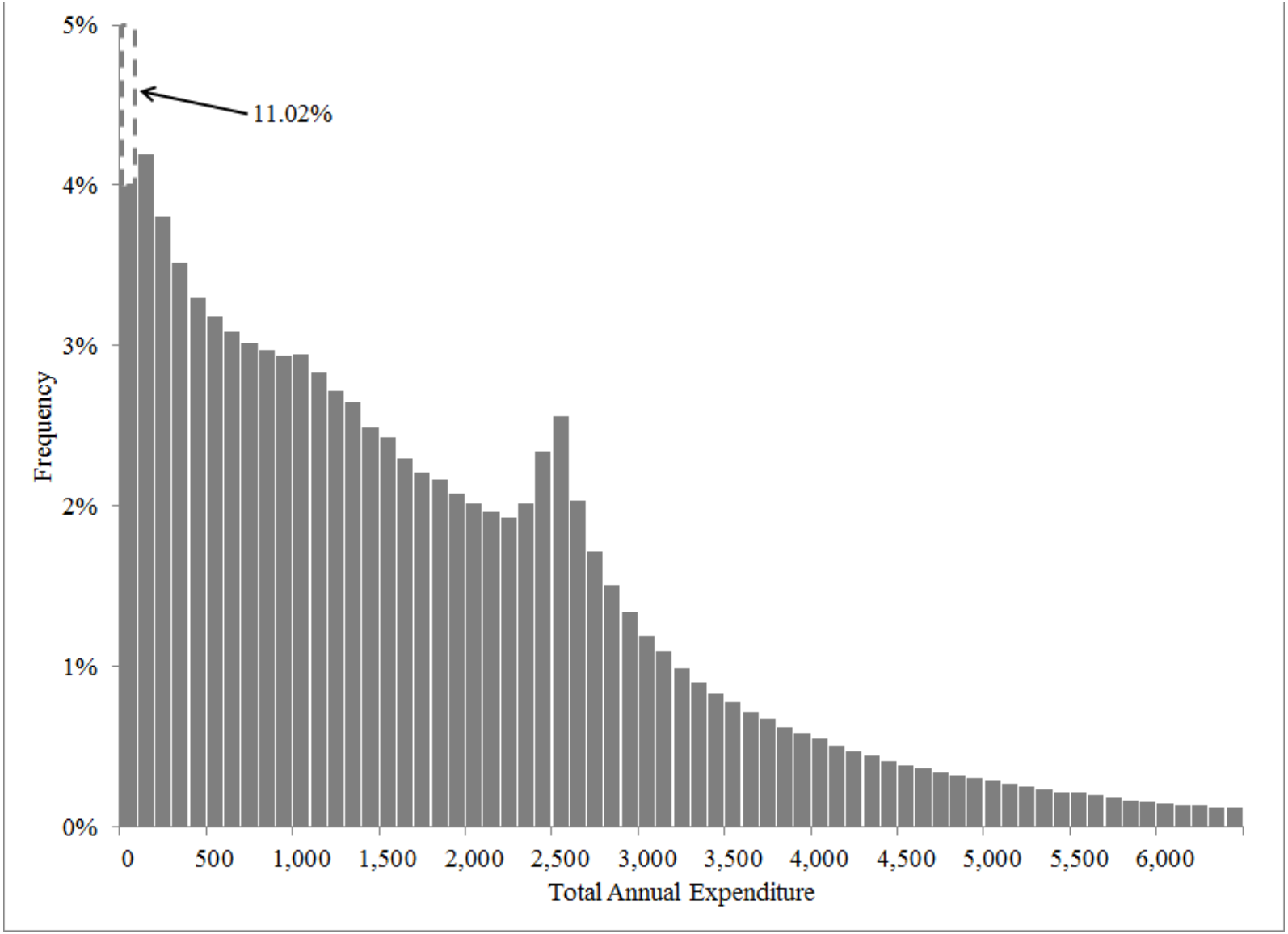
Response to price: bunching

- Sharp increase in price when go into donut hole
  - On average price goes from 34 to 93 cents for every dollar
- Standard economic theory: with convex preferences smoothly distributed in population, should see bunching at the convex kink

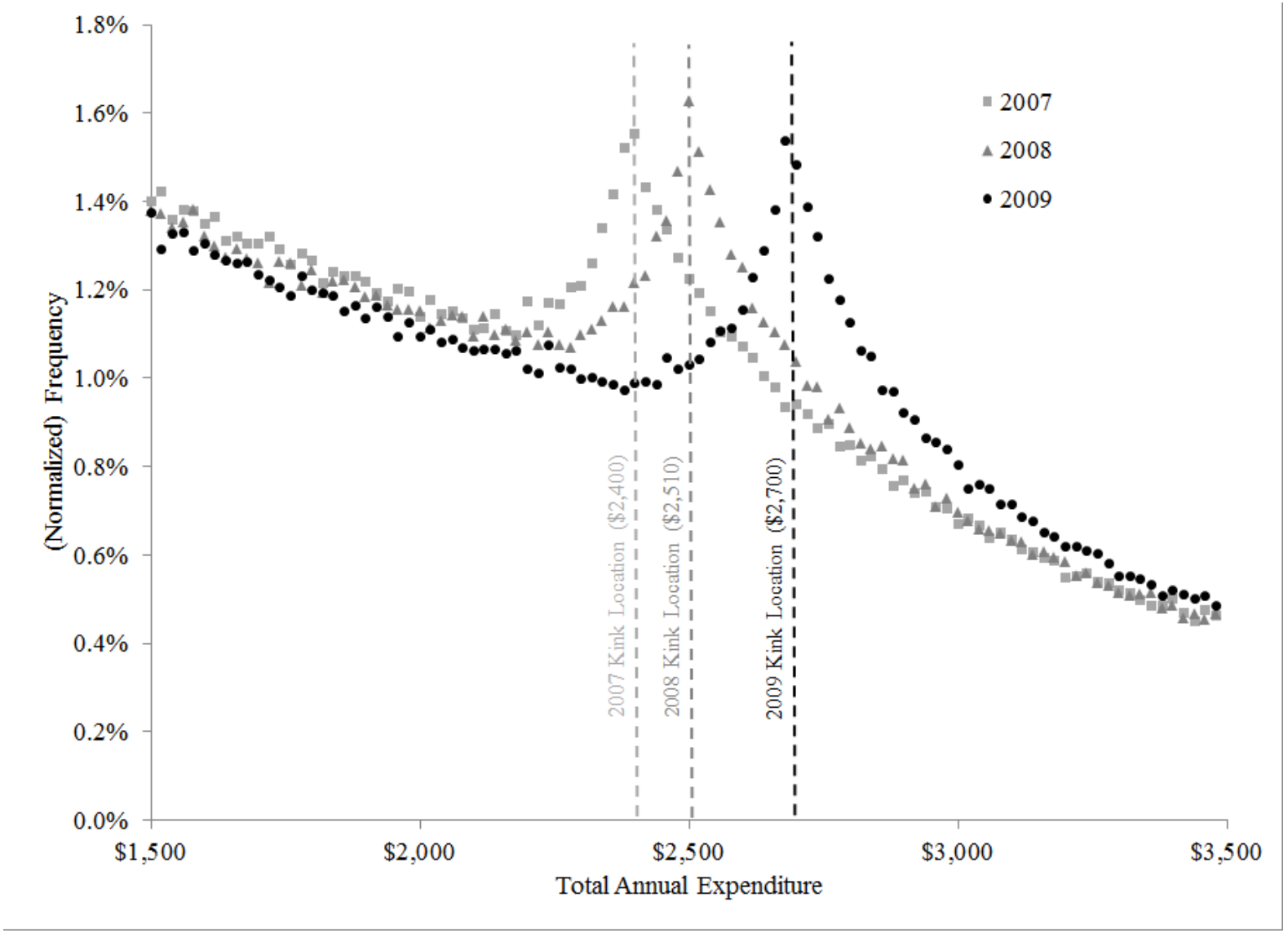
Figure A1: A Graphical Illustration for The Rationale to Observe Bunching at The Kink



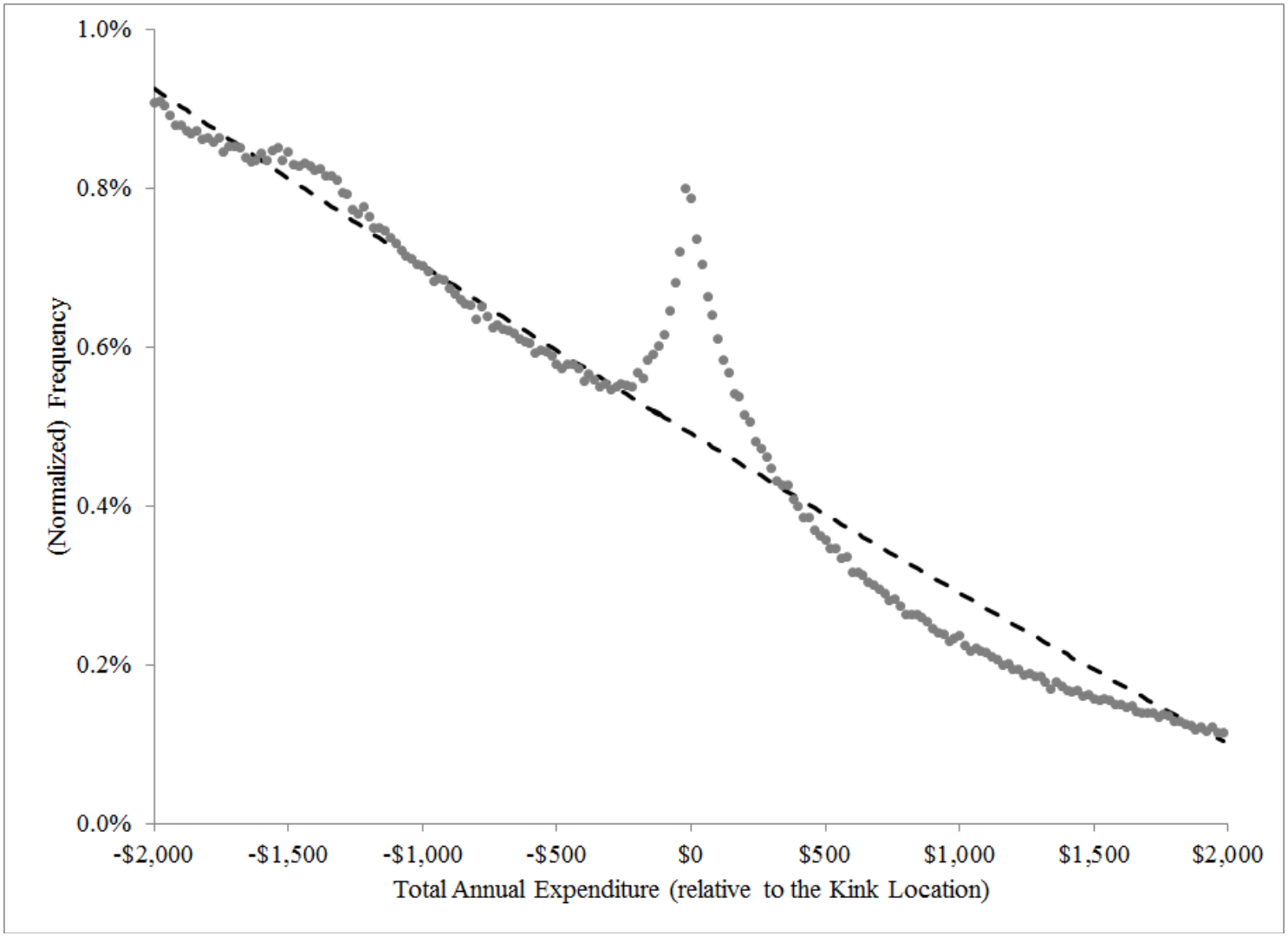
Bunching: 2008 Spending



Bunching: Year-to-year

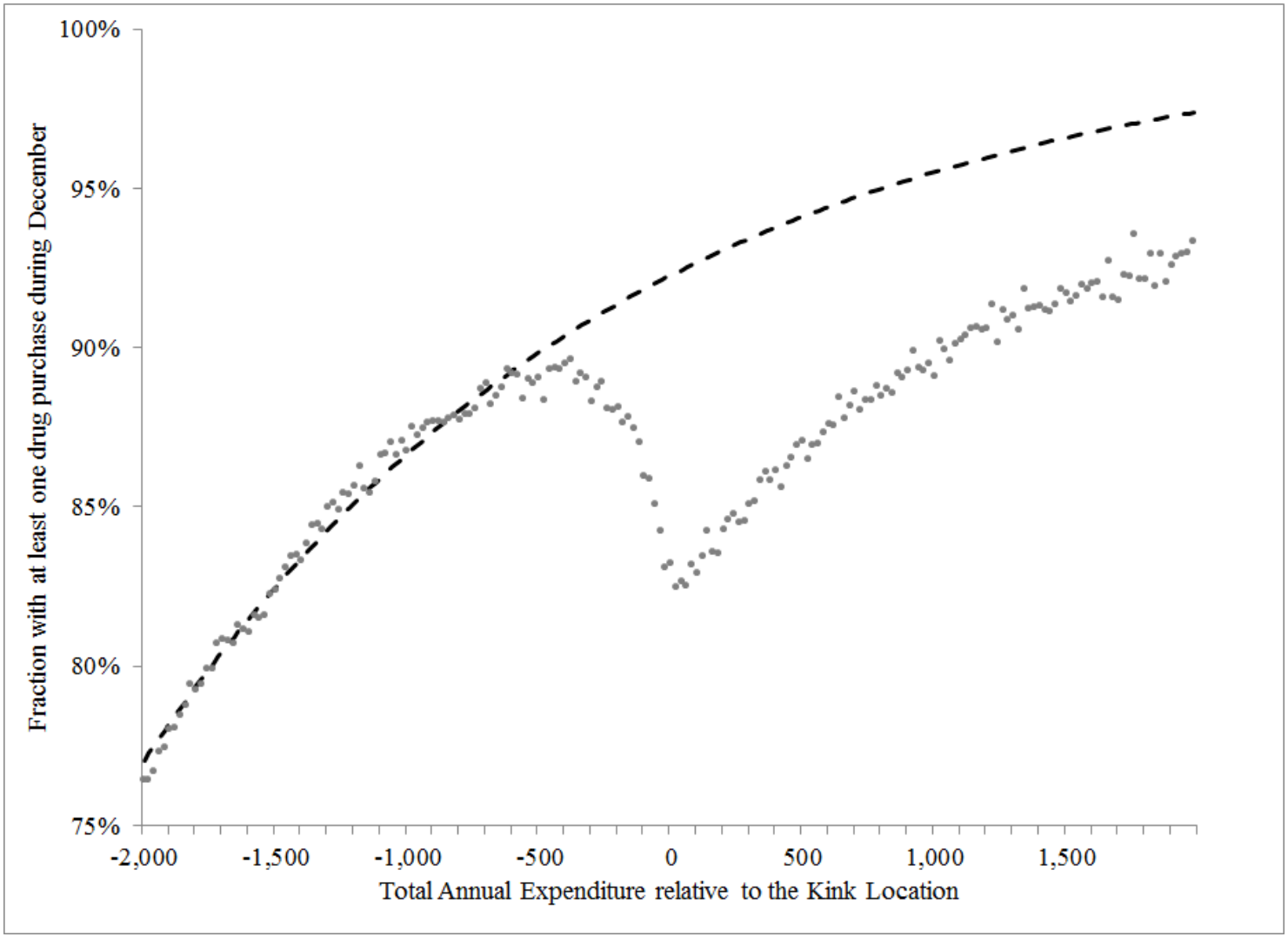


Bunching: All years, normalized

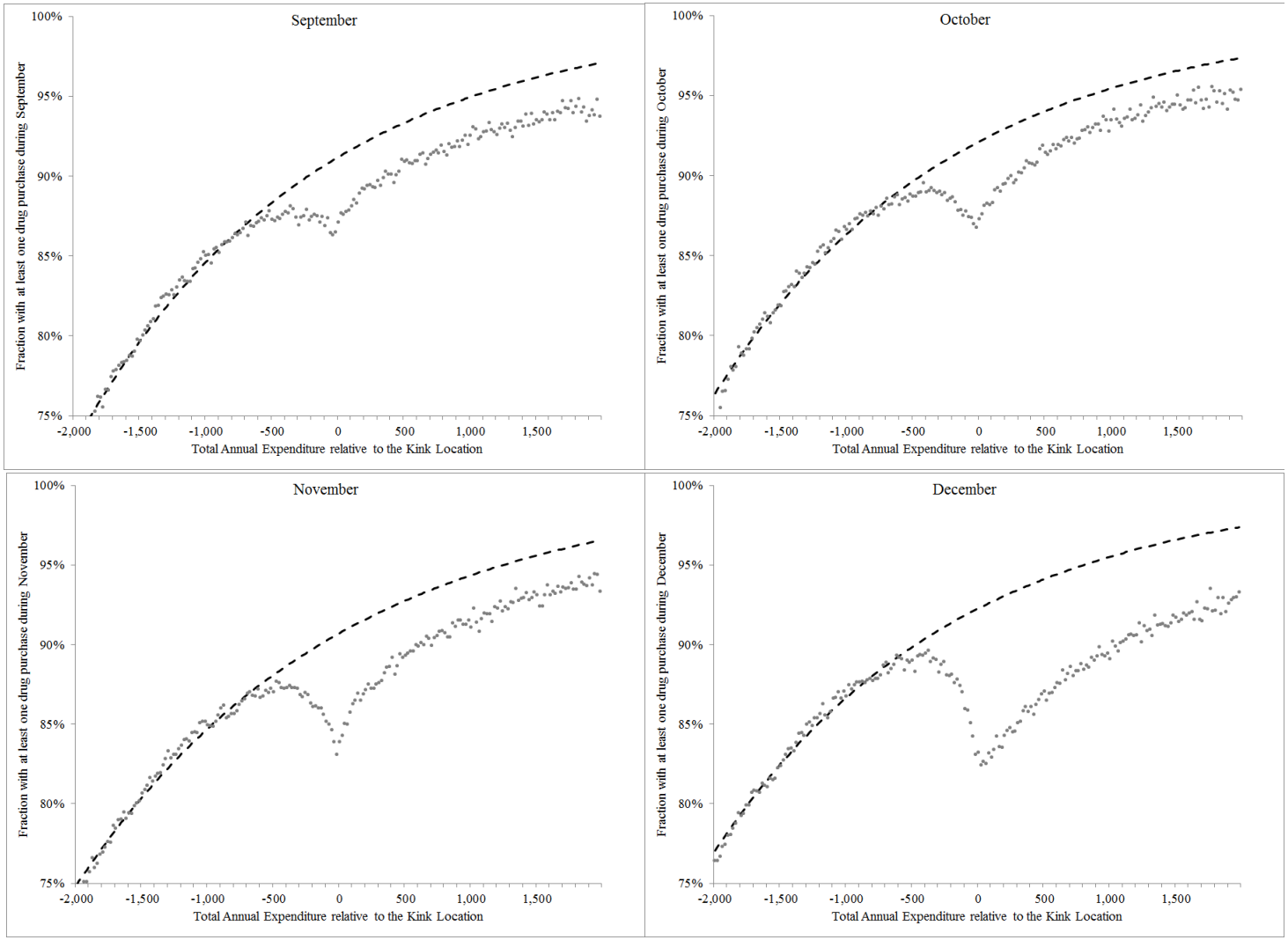


- Estimate excess mass of 29.1% (standard error = 0.003)
  - Rejects null of no behavioral response to price

Timing of purchases (December)



Timing of purchases (Sept - Dec)



Outline

1. Data and Setting
2. Bunching estimates
3. Estimate two different models
   - Static, frictionless model (adaptation of Saez 2010)
   - Dynamic model (EFS 2015)
4. Conclusions

Static model of drug spending

- Saez (2010)
  - Static, frictionless model of labor supply
  - Key insight: in this model, can translate observed bunching in annual earnings around convex kinks in income tax schedule into an estimate of labor supply elasticities
- We adapt it to our context, sticking as closely as possible to Saez's original model

Static model of drug use

- Assume individual i has quasi-linear utility over total drug spending m and residual income y
- Parametric assumptions:

$$u_i(m,y) = \underbrace{\left[2m - \frac{\zeta_i}{1+\frac{1}{\alpha}}\left(\frac{m}{\zeta_i}\right)^{1+\frac{1}{\alpha}}\right]}_{g_i(m)} + \underbrace{[I_i - C(m)]}_{y}$$

  where $C(m)$ maps total spending m into out of pocket spending
- $g_i(m)$ is chosen to obtain a tractable, constant elasticity form of the spending function similar to Saez

Static model (continued)

- Parametric assumptions:

$$u_i(m,y) = \underbrace{\left[2m - \frac{\zeta_i}{1+\frac{1}{\alpha}}\left(\frac{m}{\zeta_i}\right)^{1+\frac{1}{\alpha}}\right]}_{g_i(m)} + \underbrace{[I_i - C(m)]}_{y}$$

- With linear coverage ($C(m) = c \cdot m$, $c \in [0,1]$) optimal drug expenditure is

$$m = \zeta_i(2-c)^{\alpha}$$

- Specification implies a constant elasticity α of spending with respect to $(2-c)$
  - Very similar to Saez: constant elasticity with respect to $(1-t)$ where t is marginal tax rate on income
  - Rest of our derivation follows his closely; derives mapping between empirical "bunching" and elasticity α

Elasticity estimates from static

Counterfactual distribution	Exclusion window[a]	Bin size[b]	Excess mass[c]	Elasticity[d]
Linear	200	40	0.401	-0.047
Cubic	200	40	0.314	-0.037
Linear	200	60	0.418	-0.049
Linear	100	40	0.586	-0.034

[a] Exclusion window refers to the distance from the kink location within which we calculate the response to the kink.

[b] Bin size refers to the spending size of bins, which is used to fit the pre-kink spending distribution.

[c] Excess mass: $\frac{B}{N_{counter}} = \frac{N_{actual} - N_{counter}}{N_{counter}}$.

[d] Elasticity of spending calculated wrt end-of-year cost-sharing rate c of each

Dynamic model of drug use

- EFS (2015, QJE)
- Risk-neutral fwd-looking individual faces uncertain health shocks
- Prescriptions are defined by (θ, ω), where $\theta > 0$ is the prescription's (total) cost and $\omega > 0$ is the monetized cost of not taking the drug
  - Arrive at weekly rate λ', drawn from $H(\lambda'|\lambda)$ which follows a Markov process to allow for serial correlation in health
  - When arrive, drawn from $G(\theta, \omega) = G_2(\omega|\theta)G_1(\theta)$
- Insurance specifies (discrete) covg length T and defines $c(\theta, x)$ – the out-of-pocket cost associated with a prescription that costs θ when total spending so far is x.
- When a shock arrives, individuals make binary choice (fill prescription or not)
- Flow utility given by

$$u(\theta, \omega; x) = \begin{cases} -c(\theta, x) & \textit{if filled} \\ -\omega & \textit{if not filled} \end{cases}$$

Dynamic model (continued)

- Individual choice: fill prescription or not
- Optimal behavior characterized by simple finite horizon dynamic problem
- Value function given by solution to following Bellman equation:

$$v(x,t,\lambda) =$$

$$\int \left[\begin{array}{c} (1-\lambda')\delta v(x,t-1,\lambda') + \\ +\lambda' \int \max \left\{ \begin{array}{l} -C(\theta,x) + \delta v(x+\theta,t-1,\lambda'), \\ -\omega + \delta v(x,t-1,\lambda') \end{array} \right\} dG(\theta,\omega) \end{array} \right] dH(\lambda'|\lambda)$$

with terminal condition $v(x,0,\lambda) = 0$ for all x

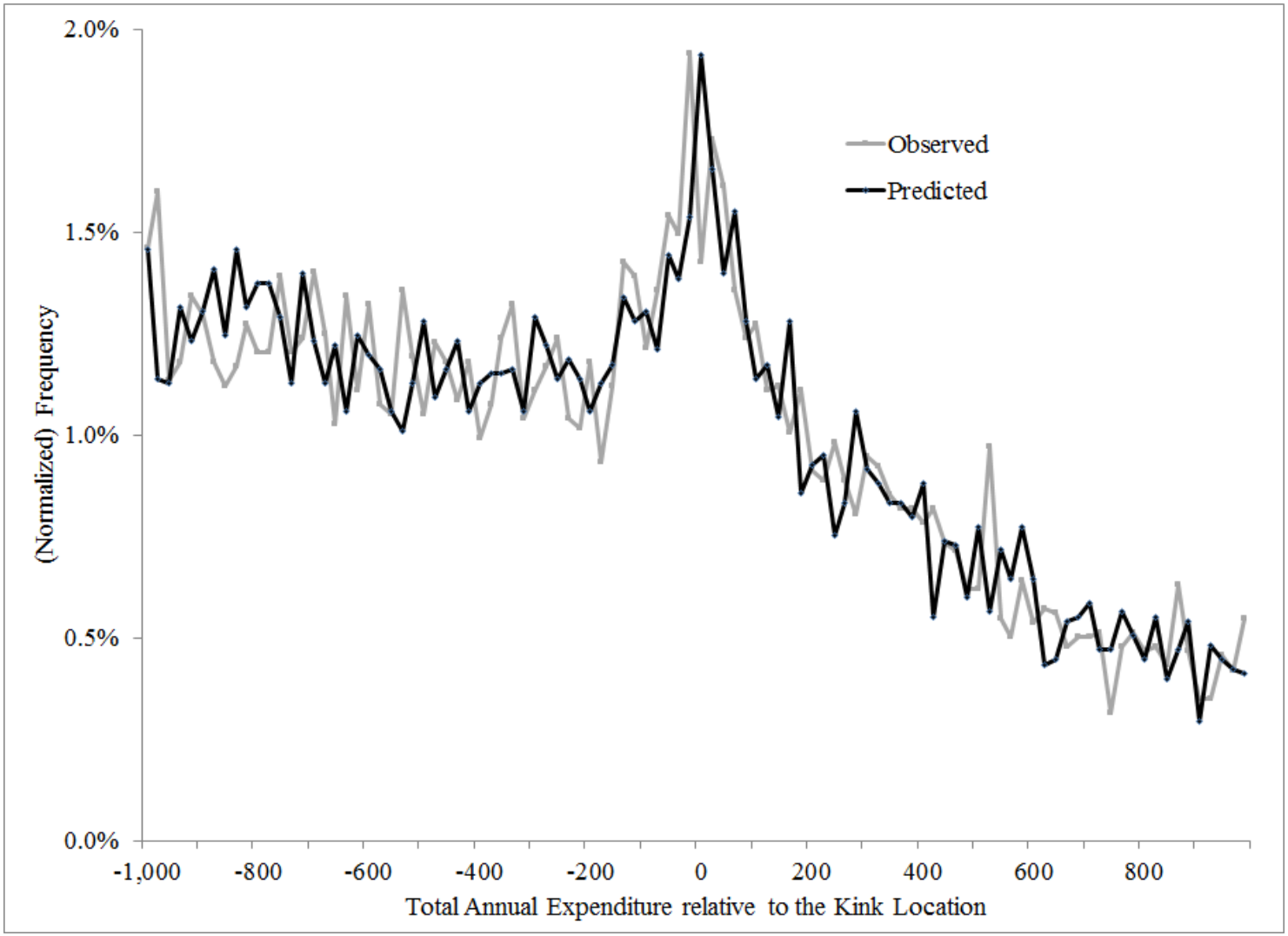
Three key economic objects

- Statistical description of distribution of health shocks: λ and $G_1(\theta)$
- "Primitive" price elasticity capturing substitution between health and income: $G_2(\omega|\theta)$
- Extent to which individuals understand and respond to dynamic incentives in non-linear contract: $\delta \in [0,1]$

Estimation

- We parameterize the model with distributional and functional form assupmtions, and model heterogeniety using a discrete type space
- Estimate using simulated minimum distance
- Moments:
  - Distribution of annual spending: average, standard deviation, pct zero, etc..
  - **Bunching**: Histogram of total spending around the kink ($\pm$ \$500)
  - Claim timing pattern around kink
  - Covariance in spending between first half and second half of year

Model fit: bunching



Elasticities from dynamic model

- Consider uniform % price reduction on all arms of standard plan
- Simulate spending for each indivdiual under original coverage plan and modified plan and use these to compute elasticities

(Uniform) Price Reduction[a]	Average Annual Spending	"Elasticity"[b]
0% (Baseline)	1,838	
1.0%	1,842	-0.22
5.0%	1,860	-0.24
10.0%	1,883	-0.24
15.0%	1,906	-0.25
25.0%	1,958	-0.26

[a] "Uniform price reduction" achieved by reducing price in every arm of each plan by the percent shown in the table.

[b] Implied "elasticity" calculated as ratio of percent change in spending (relative to the baseline) to percent change in price (relative to the baseline).

Comparing static and dynamic

- Key point:
  - Both models match bunching estimates
  - Deliver different elasticity estimates: dynamic model elasticity about five times larger than static model (-0.25 vs -0.05)
- We don't think models are vertically rankable
  - Saez model
    - ★ Simple and transparent mapping from descriptive fact to economic object of interest
    - ★ Relatedly, can be implemented quickly and easily
  - Dynamic model:
    - ★ More computationally challenging and time consuming to implement
    - ★ More "black box" relationship between underlying data objects and economic objects of interest
    - ★ Allows us to account for potentially important economic forces that Saez-style model abstracts from

Models are conceptually different

- Saez model is frictionless
  - Implementation allows for some frictions since bunching is measured with some bandwidth (vs kink)
    - ★ Will miss any behavioral response outside this bandwidth
  - Dynamic model allows lumpiness by modeling a discrete series of (weekly) health shocks and purchase decisions
  - Static model will miss any behavioral response outside the bandwidth used to measure bunching
- Saez model is static
  - All uncertainty realized prior to any spending decision
  - Dynamic model: individuals make sequential purchase decisions throughout the year as information is revealed
    - ★ Potential anticipatory behavior - set of people "at risk" of bunching may be endogeneously affected by presence of kink
    - ★ Previous work suggests existence and importance of anticipatory behavior (Aron-Dine et al. 2016, EFS 2015)

Reconciling the models

- Considered two "restricted" versions of the "full" dynamic model
  - "No dynamics model": assume no discounting or uncertainty; continue to allow for frictions in the form of lumpy spending
  - "No discounting model": allows for lumpiness in spending and also uncertainty in timing and nature of shocks throughout year but imposes $\delta = 1$
    - ★ All dynamic behavior due to uncertainty about future, rather than to time preferences
- Estimate each model, again fitting bunching patterns
- Elasticity results suggest allowing for lumpiness and uncertainty important; discounting less so
  - Full dynamics: -0.25 (vs Saez -0.05)
  - "No dynamics": -0.13
  - "No discounting": -0.22

Challenges and opportunities

- Current "frontier" of research
  - Focus on compelling evidence of behavioral reponse
  - Map the reduced form / compelling evidence to an economic object of interest
- Key point: mapping choice can be consequential
  - Illustrated here in context of bunching estimators
  - Previous work illustrating in context of randomized controlled trials: uses RAND HIE (Aron-Dine et al. 2013 JEP)
- Path forward?
  - Find the right model?
  - Find the right question?
    - ★ Are there underlying primitives to recover?

Excess Mass and "Bunching"

- Derive (a la Saez) expression relating elasticity (α) to a bunching estimate B :

$$B = m^* \left[\left(\frac{2 - c_0}{2 - c_1} \right)^{\alpha} - 1 \right] \frac{h(m^*)_- + h(m^*)_+ / \left(\frac{2-c_0}{2-c_1} \right)^{\alpha}}{2}$$

  - $B = N_{actual} - N_{counter}$; number of people empirically around kink over and above number we (counterfactually) estimate would be in this area if kink did not exist
  - $c_1 >> c_0$ are marginal price of drugs after and before gap, respectively
  - m^* location of kink

Implementation

- Approximate counterfactual distribution of spending near kink by fitting a polynomial approximation to spending below the kink, subject to integration constraint
  - Use counterfactual to project into \$200 winow around kink to estimate B
  - Explore sensitivity to polynomial choice, spending size bin, exclusion window
- Use model to map estimates of B to α

Parameterization

- Assume $G_1(\theta)$ is lognormal: $\log \theta \sim N(\mu, \sigma^2)$.
- Assume $\omega|\theta$ is stochastic, and is drawn from a mixture distribution:
  - $\omega \geq \theta$ with probability $1 - p$ (prescription is filled for sure)
  - $\omega \sim U[0, \theta]$ with probability p (decision responds to price)
- Allow for serially correlated shocks:
  - two possible values of for weekly event probability λ (H and L) with relatively probabilities depending on last period's draw (Markov process)
- Heterogeneity modeled using a finite (five types) mixture:
  - Individual is of type m with probability $\pi_m = \exp(z_i'\beta_m) / \sum_{k=1}^{M} \exp(z_i'\beta_k)$
  - All parameters but δ vary with type: λ_m, μ_m, σ_m^2, p_m
  - Baseline z_i: constant, risk score and 65-year-old indicator