14.472 Public Finance II

Topic II_c: More on Moral Hazard

Amy Finkelstein

Fall 2022

Taking Baily Chetty to data

- Formula offers a potential road map for empirical work: to tell you if locally should raise or lower benefits:

$$\underbrace{\frac{u'(c_l) - v'(c_h)}{v'(c_h)}}_{\text{welfare gain from increase in insurance}} = \underbrace{\frac{\varepsilon_{1-e,b}}{e}}_{\text{fiscal cost of increasing insurance}}$$

- LHS: Gap in MU across states (challenging - see last lecture)
- RHS: Impact of benefit on government budget (in principle, straightforward).

Focus of Lecture

- Emphasize some methodological issues with estimation
- Context: Empirical literature on moral hazard, mainly focusing on health insurance (sorry :-)).

Overview

Testing: Does moral hazard exist?

Implications of moral hazard for behavior under alternative contracts

Implications of moral hazard for welfare

Testing: Does moral hazard exist?

Methodological themes

- Reduced form evidence as complement (not substitutes) for "structural modeling" - and vice versa
- Value (and limitations of) reduced form work
- Uses (and abuses) of structural estimation
- Importance of modeling choices: a given reduced form object can have very different out-of-sample implications depending on the model
- Underappreciated importance of attention to *measurement:* what is the process by which the data were generated?

Moral hazard in health insurance

- In health insurance "moral hazard" typically refers to impact of health insurance on medical spending (i.e. "price elasticity of demand for medical care")
  - Hidden type: how much medical care I would consume if I faced the full marginal cost of the medical care consumption
  - What is the hidden action?
- Intellectual history:
  - Arrow (1963 AER) first use of moral hazard to mean "medical insurance increases demand for medical care"
  - Pauly (1968 AER): first explicit use of term to refer to impact of insurance on medical care via its effect on reducing the price

Moral hazard in health insurance (con't)

- Two distinct phenomena referred to as "moral hazard" in health insurance context
- "Ex ante moral hazard": I invest less effort in my health (smoking, drinking, exercise) and so my health worsens (Ehrlich and Becker 1972)
  - Seems a compelling example of "hidden action"
- "Ex post moral hazard": Conditional on my health, I consume more medical care (Pauly 1968)
  - Focus of empirical literature
  - What's the hidden action?

Ex Ante Moral Hazard

- "Ex ante moral hazard": I invest less effort in my health (smoking, drinking, exercise) and so my health worsens (Ehrlich and Becker 1972)
- Little empirical work / evidence
- Spenkuch (JHE 2012)
  - RCT in Mexico (geographic level) - Seguro Popular (King et al. 2009)
  - Re-analyzes, grouping for power and finds some evidence of *declines* in preventive care (flu shots, mammograms etc)
  - Question of interpretation: ex ante mh or congestion? [Cleaner tests?]
- Note: even "full" health insurance only insures medical expenses (not health)

"Ex post" moral hazard

- "Ex post moral hazard": Conditional on my health, I consume more medical care (Pauly 1968)
- What is the hidden action?
  - You or your physician make less effort to shop around for a good price (Arrow 1963)
  - Lower price $\rightarrow$ consume higher quantity of care (Pauly 1968).
    - How is this hidden action? Cutler and Zeckhauser (2000): Action not hidden but motivation is (re-writing hidden type problems as hidden action problems... Milgrom 1979)
- From now on will focus on "ex post" moral hazard: response of health care spending (P*Q) or health care use to consumer price
  - Only recently have we started to try to disentangle role of "price shopping" (i.e. P a la Arrow) - Brot-Goldberg et al. (QJE 2017) find no evidence that high deductibles encourage search for lower prices

Testing / Existence: Is there moral hazard in health insurance

- Does health insurance increase healthcare spending
- Claim: credible reduced form techniques are ideally for testing such nulls

Conceptually what do we expect?

- Health insurance, by design, lowers the price indivdiuals pay for their healthcare use
- Conjectures:
  - Health insurance will increase healthcare spending (Demand curves slope down)
  - Health insurance will not affect healthcare spending which is determined by "needs"
    - "No body wants to go to the doctor"
  - Health insurance will reduce healthcare spending
    - Reduce inappropriate and inefficient use of (expensive) emergency rooms
    - Improve health and therefore reduce health care use

Ultimately this is an empirical question

- Empirical challenge: people w/ and w/o insurance differ for reasons that may be related to the outcome of interest (i.e. health care use)
  - In particular, recall adverse selection: sick select into market
- Inference issue: separating selection from treatment (moral hazard)
- Arguably no better way to convincingly test the null of no moral hazard than with a randomized experiment
  - Randomly assign insurance across individuals
  - *By construction*, the insured and uninsured are on average identical except for whether or not they have insurance.
- Three RCTs in US health insurance
  - RAND Health Insurance Experiment (1970s)
  - Oregon Health Insurance Experiment (2008)
  - AB Demonstration project (Michalapoulous et al. 2011)

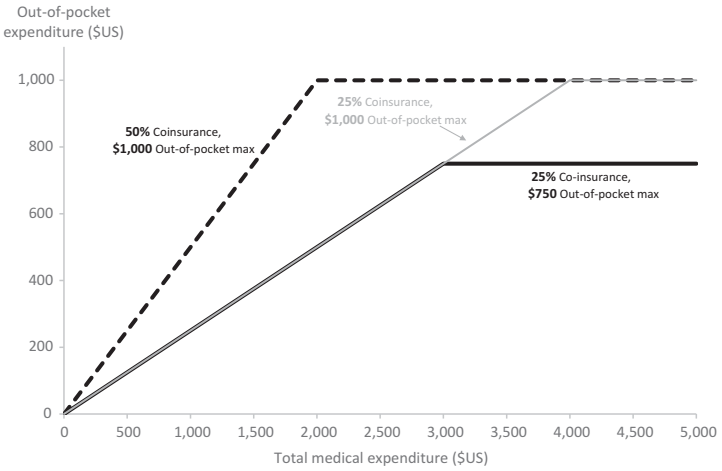
Rand HIE

- 1974-1981: Randomize approx 2,000 families (5,800 people) into plans with different consumer cost-sharing
  - Conducted at 6 different locations across the US
  - Designed to be representative of families w/ adults under age 62
  - Assigned to experiment for 3 to 5 years
- Designed to study effects of consumer cost sharing on health care spending and health
- Pioneering – One of the earliest RCTs in the US
  - PI: Joe Newhouse
  - To date, still one of the largest (~$300 million in 2011 $)

Random assignment to contracts

- Main feature: randomly assign familes to plans with different consumer cost-sharing
  - ranging from full coverage (free care plan) to a plan with almost no coverage for first ~$4,000 (in 2011 $) incurred during year
- 6 main plans. Coinsurance:
  - 0% (free plan); 25%; 50%; 95%
  - Mixed plans:
    - 25% except mental and dental (50%)
    - "Individual deductible plan": 95% outpatient, 0% inpatient.
- To limit participants' financial exposure, randomly assigned (w/in coinsurance rates) to plans with different Maximum Dollar Expenditure (MDE) for years
  - Typically 5, 10 or 15% of income up to max of ~$3,000 - $4,000 in 2011 dollars
  - One average, ~1/3 of families hit stop loss within year
  - Interpretation implications: Variation not over catastrophic coverage

Examples of contracts offered



Randomization and Measurement

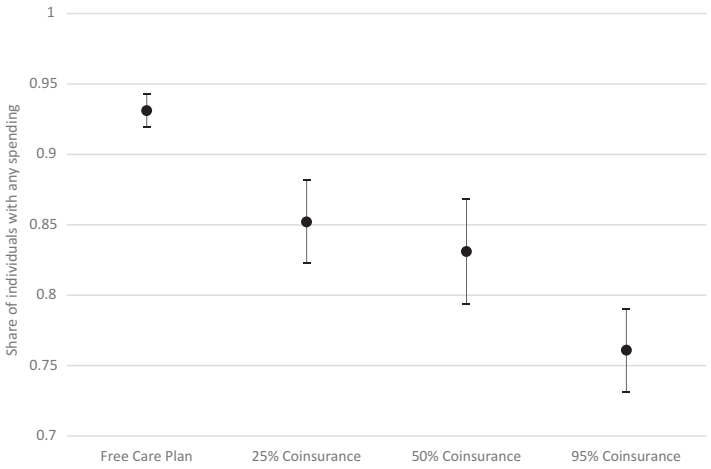
- Not simple random assignment
  - Within a site and enrollment month assigned by stratified random sampling designed to achieve better balance across a set of baseline covariates than would likely be achieved by chance alone
- Data on medical use and medical spending come from claims filed by participants with the experiment
  - During the duration of the experiment the experiment acts as your insurer.
  - To be reimbursed, one needs to file claims
  - Claims provide detailed information on health care use and spending

Experimental treatment effects: empirical framework

$$y_{i,t} = \lambda_p + \tau_t + \alpha_{l,m} + \varepsilon_{i,t}$$

- Outcome $y_{i,t}$ (e.g. medical expenditures) regressed on plan, year, and location x start month fixed effects
- Key coefficients of interest are the six plan fixed effects (λ_p)
- Because plan assignment was random conditional on location and start month, include full set of interactions
  - Need to condition on anything correlated with assignment (site, start and interaction)
  - Year fe's (to account for multiple years of study)
  - Analyze individual level data but cluster on family (level of assignment)

Experimental treatment effects: example results



Three main threats to validity

- Was assignment random?
  - "checking" that randomization occurred through balance tests
- Differential attrition (participation and/or refusal) by plan assignment
- Differential measurement of outcomes by plan assignment
- These are all discussed in more detail in Aron-Dine et al. (2013)

Experimental validity II: Differential Attrition?

- Now well-known as a key potential issue in RCTs
  - Similar issue in negative income tax experiments from 1970s (Ashenfelter and Plant 1990)
- Key point: attrition undermines the essence of random assignment
  - Particularly concerning when attrition rates vary across treatment arms
  - But even if attrition *rates* same, have to worry that composition of participants differs

Differential attrition: in RAND

- Reason to expect differential attrition:
  - Individuals assigned to more comprehensive plan have greater incentive to participate
- Only 76% of individuals assigned to plans participated
- Completion rates substantially and systemetically higher in more comprehenisve plans
  - 88% in (most comprehensive) free care
  - 63% in (least comprehenisve) 95% coinsurance plan

Strategies for "addressing" differential attrition

- Approach 1: Administrative data on outcomes (e.g. health care utilization) for all people in study, including non-participants - e.g. administrative data
  - Would allow comparison of outcomes based on assignment, regardless of participation (intent-to-treatment)
  - Could instrument for plan enrollment with plan assignment
- Approach 2: Make (economic) assumptions about likely economic model of selection and use this to adjust point estimates
  - Note: this is moving beyond the pure statistical nature of the RCT to impose economic modeling
- Approach 3: Statistical exercise designed to find lower bound of treatment effects (under some statistical assumptions)
  - e.g. Lee (2009) bounding procedure

Lee bounding procedure

- Statistical question: how bad could bias be from differential participation?
- Drop highest spenders in the lower cost sharing (more comprehensive) plan until participation rates equalized with higher cost sharing (less comprehensive) plan
  - e.g. since have 88% participation in free care but only 63% participation in 95% coinsurance plan, drop highest 28% (=(88-63)/88) of spenders from free care sample to equalize participation rates
- Lee (2009) shows that this gives worst case lower bound for treatment effects under assumption that any participant who refused participation in a given plan would have also refused participation in a less comphreneisve plan (monotonicity assumption)
- Note: This approach does not get you an alternative point estimate. It gets you a *lower bound*

Experimental validity III: Differential measurement

- Data on spending comes from claims filed.
- Participants in more comprehensive plans have greater incentive to file
- Newhouse and Rogers (1985) audit study of roughly 1/3 of enrollees found under-reporting of outpatient spending ranges from 4% in free care plan to 12% in 95% coinsurance plan
- Source of upward bias in experimental treatment effects of cost sharing

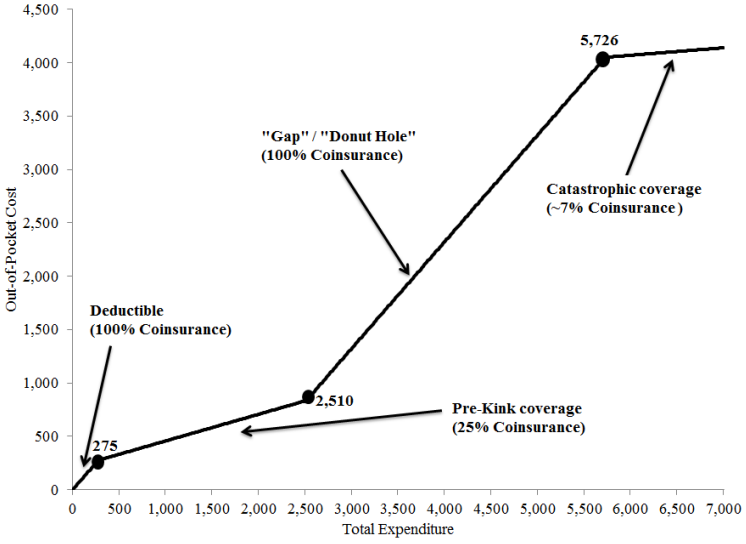
Importance of understanding where the data come from

- We think a lot about inference problems that arise because "correlation is not causation"
- But another pitfall in inference is bias that arises because of what we are measuring
  - Increasingly important as we start to work with novel, exciting new data sets becoming available
- Examples
  - Spending comes from claims filed in RAND
  - Measuring health using insurance claims data (Song et al. 2010 NEJM)
  - Abraham Wald and WWII: Where to armor the plane?

Quasi-experimental evidence of moral hazard

- Medicare Part D: Prescription Drug Coverage for elderly and disabled
- Provided by private insurers who are required to offer coverage that is actuarially equivalent or more generous than a government designed standard benefit

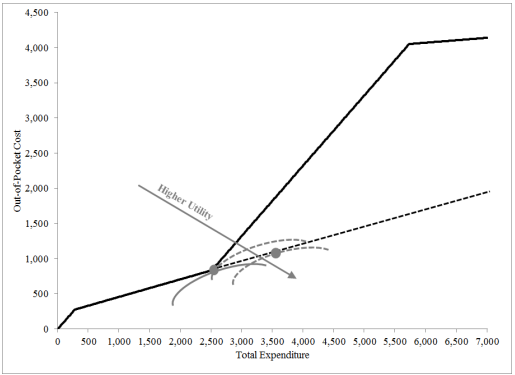
Standard Benefit Design (2008)



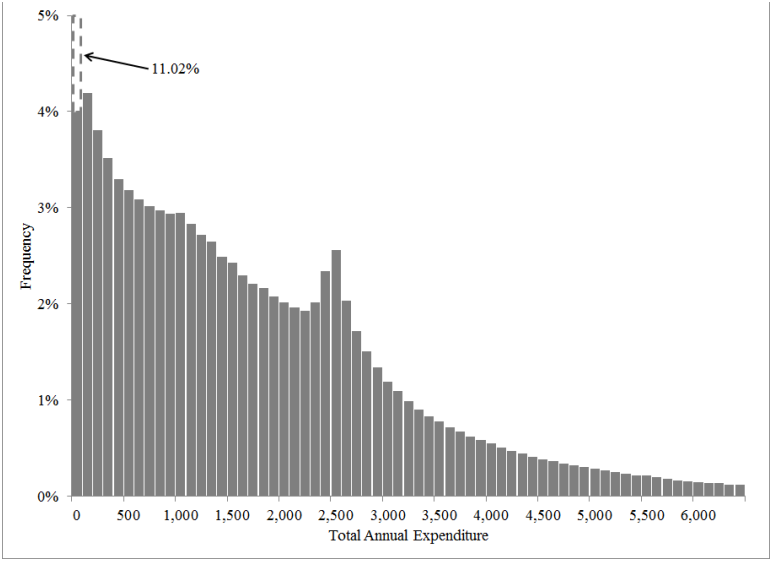
Response to price: bunching at the kink

- 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

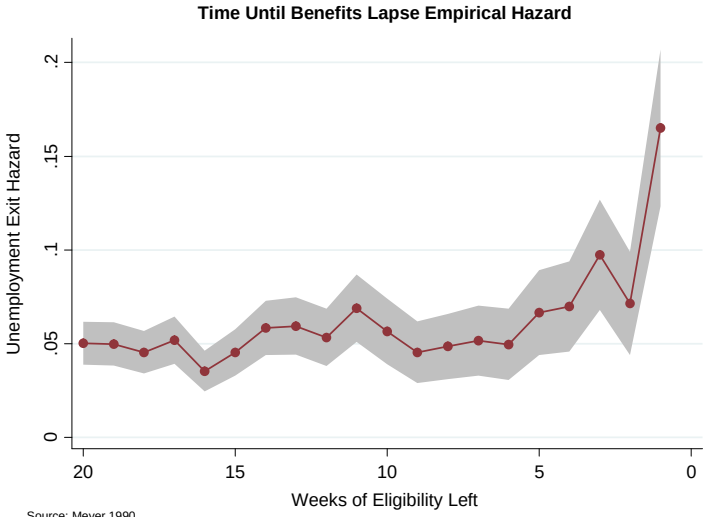
Figure A1: A Graphical Illustration for The Rationale to Observe Bunching at The Kink



Evidence of Bunching



UI: Spike in exit when benefits are exhausted



Source: Meyer 1990

Spike in exit when benefits are exhausted

- One of the most striking pieces of evidence for distortionary effects of UI is the spike in the hazard rate of exiting unemployment when benefits end
- Spike in exit rates is one of most robust findings in UI research: replicated in many countries and data sets.
- Caveat (Card, Chetty Weber AEA P&P 2007)
  - Austrian data suggest spike is smaller when "hazard" is re-employment as opposed to "exit from (registered) unemployment"
    - Classification change ("unemployed" vs "out of labor force") larger than "real" change (are you employed?).
    - Distortionary cost of program depends on real behavior
  - Don't infer too much from this study, this divergence between UI exit and re-employment is more pronounced in the Austrian data than it is other countries.

Recall: Importance of Measurement

- Card et al. - distinction between "exit from registered unemployment" and "re-employment"
- Recall RAND HIE: only measure spending if file a claim (and insurance coverage may affect incentives to file)

Some comments on reduced form evidence

- Compelling evidence of existence of moral hazard
  - Relative strengths of RCT vs compelling quasi-experimental design?
- Random (or quasi-random) assignment shuts down selection - is it clear we want to?

Question: How much will high deductible health insurance plans reduce health-care spending?

- In most settings, individuals are offered a choice of health insurance options and the policy question is how to design the choice set
  - E.g. Introduce a high deductible health savings account option
  - Choice even within social insurance programs (e.g. Medicare Part D)
- Random assignment of health insurance solves causal inference problem: gives impact of consumer cost sharing on medical spending
- If choices are based in part on one's anticipated behavioral response to the contract (i.e. one's "moral hazard type"), then magnitude of spending effect of *offering* high deductible plan in option set may be very different than effect of randomly *assigning* high deductible plan
- Einav, Finkelstein, Ryan, Schrimpf and Cullen (AER 2013) "Selection on moral hazard in health insurance"
  - Increasing emphasis /interest in applied micro on existence and substantive importance of heterogeneous treatment effects

Selection on moral hazard

- Selection on risk has two separate components:
  - Risk that's invariant to covg choice ("traditional" selection)
  - Risk that arises b/c of coverage ("selection on moral hazard")
- In addition to "traditional" selection based on one's health risk (sicker individuals choose more comprehensive coverage) individual's may also select health insurance on basis of their moral hazard type
- Implications for reducing health care spending / combatting mh through higher consumer cost sharing:
  - Selection on mh implies non-random selection into plans - e.g. high deductible plans selected by low "moral hazard" types
  - Abstracting from selection on moral hazard could lead to substantial over-estimation of spending reduction associated with offering a high deductible plan as one option

Selection on "level" vs. "slope": all you can eat restaurants

- Who goes to all you can eat restaurants?
- People with big appetites (level)
- People who will eat a lot more than usual when the food is free on the margin (slope)

Selection on "level" vs. "slope": implications

- IO vs. Labor: How useful is a "good" IV LATE vs. a "bad" OLS ATT?
- Imagine 80% of people are enrolled in a program and you're interested in the ATE
- Imagine an RCT that randomizes a subsidy to encourage take up of a program
  - Suppose it induces 1 pctg pt more people to enroll (and statistical power is not an issue)
  - The RCT solves the "selection on levels" problem but how useful is this estimate if there is substantial heterogeneity in treatment effects?
- Imagine an OLS regression comparing all people on vs off the program, attempting to control for "stuff"
  - Selection on levels is a concern, but selection on slope is less of a concern because most are treated

Ideal experiment?

- [If you can do it]
- Proposal
  - (Endogenous) choice of two linear coverage (constant coinsurance) plans (high and low)
  - Within each (endogenously chosen plan): randomly assign a new (constant coinsurance) plan $\rightarrow$ estimate behavioral response of those with each old plan
  - Do those who chose higher coverage endogenously have different estimated moral hazard effect?
- Estimate treatment effects "cleanly" (via random assignment) and purged of selection

Ideal experiment?

- Medicare created a 5 year (2016-2020) (mandatory participation) bundled payment intervention randomized across MSAs
  - Then-representative Price objected to "experimenting with the health of the american public"
  - In October 2017, HHS Secretary Price converted intevention to voluntary in half the MSAs (starting in 2018).
- Who selected into intervention and how is this correlated with level and slope?
  - Einav, Finkelstein, Ji and Mahoney (QJE 2022): "Voluntary Regulation".

"Selection on Treatment"

- Selection on moral hazard is a specific (economic) application of a more general (econometric) point - "selection on gains"
- Heterogeneity in treatment effects + selection into treatment based on anticipated treatment effects
  - Heckman, Urzua and Vytlacil (2006) discuss properties of IV in this setting ("essential heterogeneity")
  - Growing interest in estimating marginal treatment effects (MTEs). See Mogstad and Torgovitsky (2018 Annual Review).

Implications of moral hazard for behavior under alternative contracts

How to interpret / use RAND treatment effects

- Testing: can reject null of no impact of cost sharing on health care spending
- Quantifying: How do we learn from these treatment effects beyond rejection of the null?
  - How to transform plan fixed effects into an economically interpretable construct that can be applied out of sample (to impact of plan designs not observed in the experiment)
  - RAND investigators made one such attempt and concluded: price elasticity of demand for medical care = -0.2 (FAMOUS. Treated with reverence by profession)
  - Key point: This famous result derives from the experimental data plus a large number of (economic and statistical) assumptions.
    - True of any out-of-sample extrapolation of experimental treatment effects more generally.
    - To learn the most from an experiment (or any reduced form estimate) we often have to layer on additional assumptions.

How to translate to a price elasticity of demand?

- Challenge (in real world and in RAND experiment): health insurance contracts are highly-non-linear
- Price faced by consumer falls as total medical spending cumulates during the year
- In general:
  - 100% during deductible
  - 10-20% in coinsurance rate
  - 0% once out of pocket spending limit has been reached
- With a non-linear budget set, what is "the price" of medical care (or the elasticity w.r.t "the price")

What price?

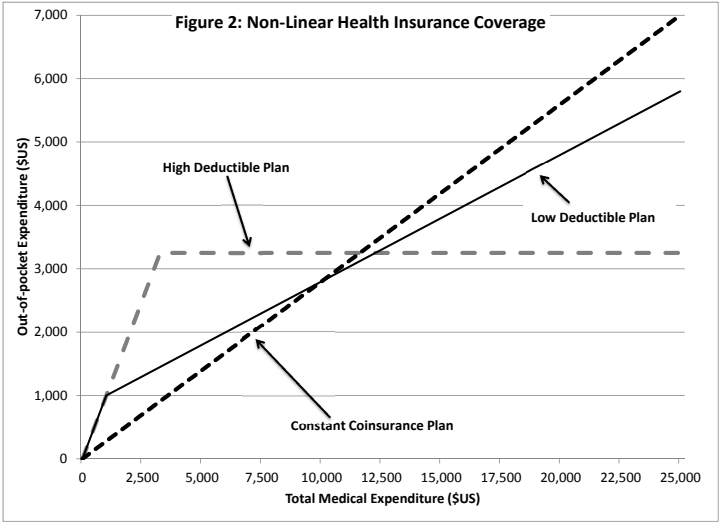


Figure 2: Non-Linear Health Insurance Coverage
Out-of-pocket Expenditure ($US)
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
High Deductible Plan
Low Deductible Plan
Constant Coinsurance Plan
0
2,500
5,000
7,500
10,000
12,500
15,000
17,500
20,000
22,500
25,000
Total Medical Expenditure ($US)

Analytical decisions

- How to analyze medical expenditures that occur at different times, under potentially different cost sharing rules but stem from same underlying health event?
- What price does individual respond to in making medical spending decision?
  - Current "spot" price of care (fully myopic)
  - Expected end-of-year price (fully forward looking + model of expectation formation)
  - Realized end-of-year price (health care consumption happens on the margin)
  - Weighted average of the prices paid over year (ad hoc)
- NB: These modeling challenges are inherent in problem of extrapolating from any study of impact of a particular health insurance policy on spending.
  - Not unique to RAND.
  - Will come up again in course (and perhaps in your research...)

What price is used can be important

- Example: What would be spending effect of replacing 28% constant coinsurance plan with RAND's 25% coinsurance plan up to the MDE
- Some ways to summarize RAND 25% coinsurance plan with a single price
  - Dollar-weighted average price (10%)
  - Person-weighted average price (17%)
  - person-weighted average end of year price (13%)
- Applying -0.2 estimate to changing from each of these prices to 28 cents yields predicted reduction in health spending of 18%, 9%, and 14% respectively
- In this example, decision of how to apply the price leads to differences in the predicted reduction of spending that vary by a factor of 2

Dangerous to use a single price for a non-linear contract

- In general no "right" way to summarize a non-linear budget set with a single price
- And we just saw how reasonable yet ad hoc "fixes" can have very different implications
- Modeling the response to the non-linear budget set induced by health insurance (or other!) contracts is an open / active area of research

Non-linear budget set: which price?

- Consider impact on spending of introducing high deductible plan (previously no deductible)
  - Completely myopic individual: reacts to "price" increase to 100%
  - Fully forward looking individuals with annual expenditures typically above the new deductible might not change his behavior much
- Which plan provides more incentives to economize on medical spending:
  - Plan A (10% coinsurance and $5,000 out of pocket max) vs. Plan B (50% coinsurance and $5,000 out of pocket max)
  - Naïve answer: B is less generous and will lead to lower medical utilization
  - But depends on distribution of medical spending w/o insurance and how ff looking people are: if have $10,000 surgery early in year (or expect to have it during year), for rest of year spot price lower under plan B

Before we get too carried away by the theory...

- Basic empirical question: do individuals take dynamic incentives into account in their medical consumption decisions?
  - i.e. do they respond to "future price" of medical care?
- Why you might be affected by the current ("spot") price:
  - Myopic (completely discount future)
  - Liquidity constrained
  - Misunderstanding of price schedule
- Empirically challenging to test null that individuals respond only to spot price
  - Challenge: how to separately identify effect of spot and future price?
- Spot price and future price often vary jointly
  - Low spending individual faces both a higher spot price and a higher future price
  - Variation in insurance contracts (e.g. changes in deductibles, coinsurance etc) will change both spot price and future price

Ideal experiment

- Randomize people across plans w same spot, different future price
  - Can test whether respond to future price
- RAND actually did this!!
  - Randomized MDE within coinsurance amount
  - Sadly, though issues of low power (and also low MDEs $\rightarrow$ can affect spot price almost immediately)

Approximating the ideal experiment

- Aron-Dine et al. (ReStat 2015)
- Idea: typical health insurance contracts offer coverage for a fixed duration, and reset on pre-specified dates
  - Generate identifying variation in future price when they get applied to indiviudals whose initial coverage period is shorter
- Example: Annual health insurance contract (January 1- December 31) with annual deductible
  - When people join mid-year, deductible remains at annual level but applies only till end of calendar year
  - Individuals who join the plan later in the year face:
    - higher future price (have fewer months to spend past the deductible)
    - same spot price
  - Examine initial care utilization across individuals who join in different months

Two applications

- Health insurance contracts are annual, with open enrollment periods (typically Oct and Nov) to change coverage for following calendar year
  - How do we find variation in contract length / when join plan?
- Application #1: Employer provided health insurance
  - New hire: Individuals who join the firm mid-year
- Application #2: Medicare Part D prescription drug coverage
  - Newly eligibles: Can join in the birth-month you turn 65

Graphical evidence: Medicare Part D

Figure 2: Probability of initial claim and expected end-of-year-price by enrollment month

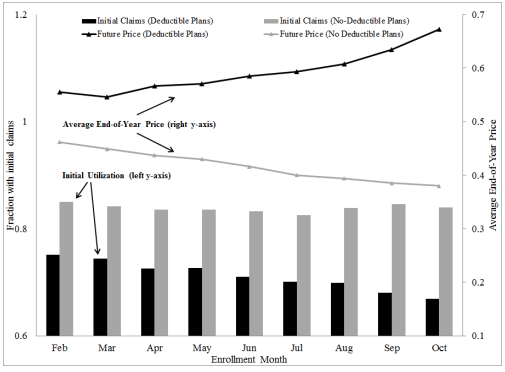


Figure graphs the pattern of expected end-of-year price and of any initial drug claim by enrollment month for individuals in Medicare Part D during their first year of eligibility (once they turn 65). We graph results separately for individuals in deductible plans and no deductible plans. We calculate the expected end-of-year price separately for each individual based on his plan and birth month, using all other individuals who enrolled in the same plan that month. The fraction with initial claim is measured as the share of individuals (by plan type and enrollment month) who had at least one claim over the first three months. See Appendix B for more details on the construction of variables used in this figure. N =137,536 (N=108,577 for no deductible plans, and N=28,959 for deductible plans).

Beyond testing: quantifying response to dynamics

- How to quantify spending response to non-linear budget set?
  - Reduced form results suggest don't want to summarize budget set with a single "price"
- Quantifying requires additional economic and statistical modeling assumptions
  - Two related papers: Einav, Finkelstein and Schrimpf (2015 QJE; 2017 JPubEc)

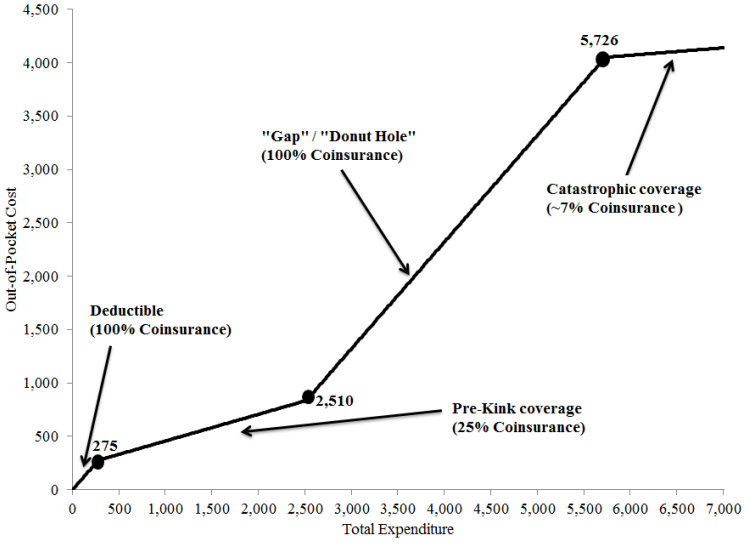
Broad Motivation

- "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
- Simple (not novel) point: choice of model can be consequential
  - Will show how two "reasonable" models can match the reduced form facts but produce very different counterfactual predictions

Specific application: Bunching estimators

- 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

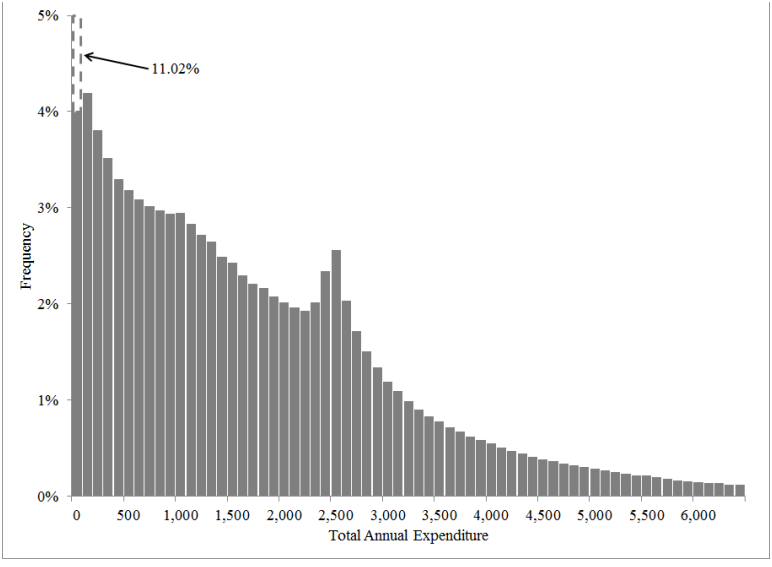
Specific context: highly non-linear nature of Part D contracts



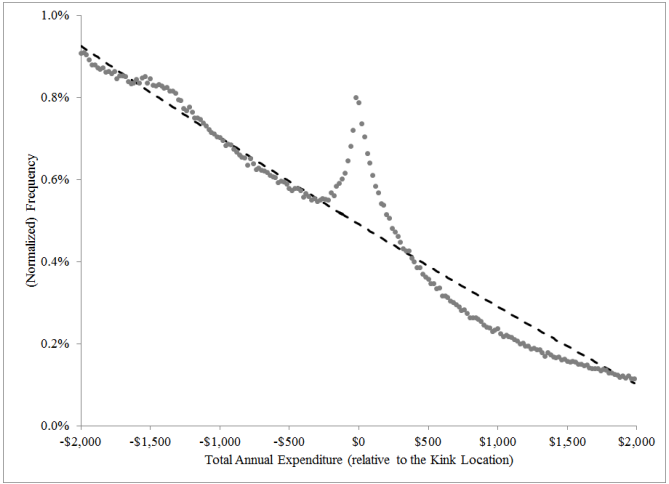
Medicare Part D

- 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, but actual plans often provide different coverage
- Many planned and potential changes to contract design
  - E.g., under ACA, "donut hole" "filled" by 2020
- Objective: Develop a model that allows us to assess responses of drug spending to non-linear contracts ("moral hazard")
  - Conceptual: e.g. anticipatory behavior
  - Quantitative: Impact of changes to contracts on drug spending

Recall: evidence of Bunching



Bunching at kink



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

Where do we go from here?

- Goal: want to make quantitative inferences about behavior under counterfactual contracts
- EFS (2017) consider two models
  - Static, frictionless model (adaptation of Saez 2010)
  - Dynamic model (EFS 2015)
- Punchline: Both models match (by construction) the bunching / excess mass, but produce very different out-of-sample predictions

Static model of drug spending, Saez-style

- 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 Part D 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:

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

  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 mode (con't)I

- Parametric assumptions:

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

- 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

Relation between 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 window around kink to estimate B
  - Explore sensitivity to polynomial choice, spending size bin, exclusion window
- Use model to map estimates of B to α

Elasticity estimates from the static model

Counterfacutal 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 individual and estimate of α. We report the average estimated elasticity across individuals.

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 $G(\theta, \omega) = G_2(\omega|\theta)G_1(\theta)$
  - $G_1(\theta)$ is distribution of financial cost of filling
  - $G_2(\omega|\theta)$ is distribution of monetized utility loss from not filling
- 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

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

Stylized model (cont.)

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

$$\begin{aligned} v(x,t) &= (1-\lambda)\delta v(x,t-1) + \\ &\quad \lambda \int max \left\{ \begin{array}{l} -c(\theta,x) + \delta v(x+\theta,t-1), \\ -\omega + \delta v(x,t-1) \end{array} \right\} dG(\theta,\omega) \end{aligned}$$

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

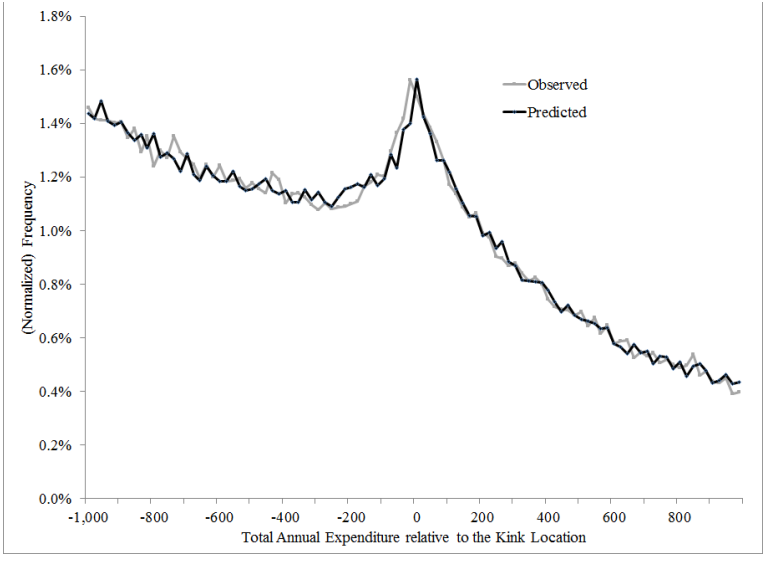
Three key economic objects

- Statistical description of distribution of health shocks: λ and $G_1(\theta)$
- "Primitive" price elasticity capturing substitution between health and income: $G_2(\omega|\theta)$
  - If $\omega \geq \theta$, fill even if have to pay full cost
  - If $\omega < \theta$, fill only if some portion of cost (effectively) paid by insurer
  - Convenient to think about the ratio ω/θ
- Extent to which individuals understand and respond to dynamic incentives in non-linear contract: $\delta \in [0, 1]$
  - "Full" myopia ($\delta = 0$): don't fill if $\omega < c(\theta, x)$
  - Dynamic response ($\delta > 0$): utilization depends on both spot and future price
  - δ is context specific! ... Captures salience, discounting, and perhaps liquidity constraints

Estimation

- We parameterize the model with distributional and functional form assumptions, and model heterogeneity 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 (+/- $500)
  - Claim timing pattern around kink
  - Covariance in spending between first half and second half of year

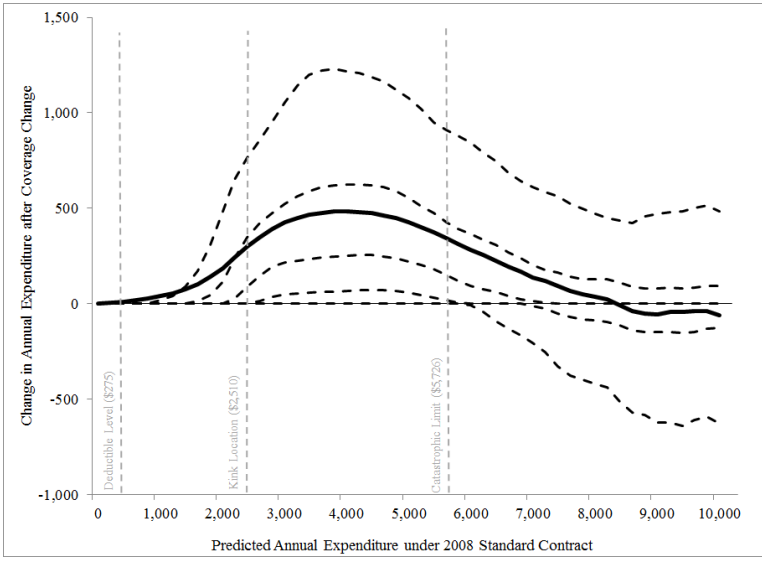
Model Fit: Spending around the kink



Counterfactual: "filling the gap"

- Main counterfactual exercise considers "filling the gap" as specified by ACA by 2020:
  - Coinsurance rate in standard contract will remain at its pre-gap level (of 25%) until out of pocket spending puts individual at catastrophic coverage limit
- Consider spending effect of "filling the gap" in the 2008 standard benefit design
  - On average, increases total spending by $204 (12%)
  - Insurer spending increases by $358; out of pocket declines by $154

Heterogeneity in who is affected



Some subtle implications of non-linear contracts

- Change in spending by people far from gap / endogeneity of people at risk of bunching
  - Arises due to dynamic considerations / forward looking behavior
  - Estimate that about 25% of average $204 / person increase in annual spending comes from "anticipatory" response by people more than $200 below kink location
  - Intuition: anticipate higher price in future => do not fill today => (sometimes) do not hit kink
- "Filling" donut hole causes some people to *decrease* spending
  - Catastrophic coverage limit held constant with respect to out of pocket (vs. total) spending, so for some people who used to hit the limit no longer do and their marginal price actually rises
  - General point: with non-linear contracts, a given contract change can provide more coverage on margin to some individuals but less coverage to others

Elasticity estimates from the dynamic model

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

(Uniform) Price Reduction[a]	Average Annual Spending	Implied "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 models

- 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)
- Models are not 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 (e.g. anticipatory responses)

Models are conceptually different (and non-nested)

- Saez model is frictionless
  - Implementation allows for some frictions since bunching is measured with some bandwidth (vs kink)
    - Will miss any behavioral response outside the bandwidth used to measure bunching
  - Dynamic model allows lumpiness by modeling a discrete series of (weekly) health shocks and purchase decisions
- 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 endogenously affected by presence of kink
    - Previous work suggests existence and importance of anticipatory behavior (Aron-Dine et al. 2016, EFS 2015)

Explored what features may be quantitatively important

- 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
  - Note: distinct from "comparative statics" (without re-estimation)
- 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

Dynamic vs static models in UI

- Unemployment is a dynamic problem
  - Any UI policy can be described as a vector b_t. Index policies b_1 and b_0.
  - Treatment effect of any UI policy is a vector $h(1) - h(0)$.
- Dimension reduction needed for static models
  - Schmieder and von Wachter (AR 2017) summarizes studies using elasticities suitable for static models $E(h(1) - h(0))/E(b_1 - b_0)$
- Is this the right way to do dimension reduction?
  - Even in a model with $u(c) - \psi(e)$ where ψ is a constant elasticity, this means elasticity is constant wrt to the value of finding a job, not a constant elasticity of duration wrt benefits. A single model of utility will deliver different elasticities in different contexts.

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, but point is more general
  - e.g. RAND HIE and recovering "the price elasticity" (Aron-Dine et al. 2013 JEP)
- Path forward?
  - Find the right model?
  - Find the right question?
    - Are there underlying primitives to recover?

Aside: two comments on paper writing

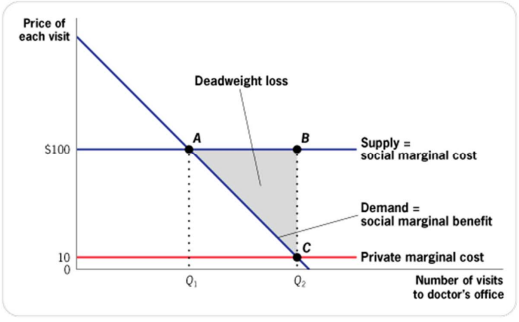
- A fun (to write and read) paper structure
  - "Reduced form" results (something to hang your hat on)
    - Evidence of forward looking behavior (Aron-Dine et al. 2015)
    - Bunching behavior at kinks (Einav et al. 2015)
  - Quantification / Interpretation / Counterfacutal Analysis - Usually requires additional modeling assumptions ("structural").
    - Arguably more speculative but more interesting
- Re-using designs for different questions - more bang for the buck (greater ratio of thinking to doing!). Examples:
  - Einav-Finkelstein-et-al: 4 donut hole papers
  - Handel et al. (AER 2013) and Handel-Hendal-Whinston (EMA 2015)
  - Doyle et al. (JPE 2015) and Hull (JMP 2017)

Some areas ripe for work

- Other aspects of nature of moral hazard
  - Inter-temporal dimension / multi-year contracts
    - What is health / utility cost of postponing medical care
  - "Source" of moral hazard - provider vs patient
  - Price shopping (Brot-Goldberg et al. QJE 2017)
  - Ex ante moral hazard (Spenkuch JHE 2012)
- Provider vs. consumer incentives
  - For reducing health care spending, have we been looking under the wrong lamppost?
  - top 5 percent of patients account for 50 percent of medical spending; top 1 percent account for almost 25%
- Welfare implications of MH - tricky but important (more on next few slides)

Implications of moral hazard for welfare

Recall classic textbook welfare analysis of moral hazard



Patient-side Moral Hazard • With no insurance, at a cost of $100 per visit, individuals would consume Q_1 doctor's office visits, where marginal costs and benefits are equal. With only a $10 copayment, however, individuals consume Q_2 worth of visits, where private marginal costs equal social marginal benefit; this overconsumption of health care leads to a deadweight loss of *ABC*.

Source: Gruber textbook

Departures from this framework

1. Dynamics
2. Price vs Social Marginal Cost
3. Behavioral

Departures from this framework I: Dynamics

- Framework is static
- Health insurance may also induce development and adoption of new technologies
  - e.g. Finkelstein (QJE 2007) on Medicare Introduction
  - But missing welfare analysis of induced innovation...

Departures from framework II: Price vs SMC

- "Overconsumption" of medical care: WTP for marginal unit of care is less than its social cost
  - But what insurance does is subsidize the price of medical care. So that only maps immediately to welfare if health care is priced at its social marginal cost.
- In many settings, $p_{hc} >> smc_{hc}$
- Example I: Prescription drugs: SMC of drug production ~0
  - Drug price distorted above MC due to patents
  - Classic patent analysis: tradeoff of dynamic vs static efficiency
  - Insurance as two part pricing undoes that inefficiency and promotes socially beneficial increase in consumption? (Lakadwalla and Sood JPubEc 2009)

Price vs SMC (con't)

- In many settings, $p_{hc} >> smc_{hc}$
- Example II: Use of ER
  - Popular view that lack of insurance $\rightarrow$ "inefficient" use of expensive ER vs lower priced doctor's offices / clinics
  - (Empirical evidence? see Oregon HIE...)
- But SMC of doc time for non emergency treatment may be close to 0 (despite high price)
  - Have to staff ER in case of emergency; what else are they doing at 3am?
  - Of course also consider opportunity cost of uninsured's time...

Dynamics or Price vs SMC: (Some) Implications

- We're not focused on the "right" behavioral response for welfare.
- The (socially) "costly" impact of insurance on behavior may be:
  - via impact on premiums and hence insurance demand (so price elasticity of demand for insurance is relevant)
  - via effect on expected market size / innovation (is this welfare increasing or welfare decreasing?)
- Not yet formalized or analyzed...

Departures from this framework III: "Behavioral"

- Individuals may not consume the privately optimal level of care absent insurance
  - May under-consume care (particularly preventive care) because of myopia, lack of understanding of long run benefits etc.
- In this second best world, subsidizing price of care through insurance and inducing increased utilization may be welfare increasing
- Baicker, Mullainathan, Schwartzstein (2015 QJE) "Behavioral hazard in health insurance"
  - What is welfare change from eliminating copays?
  - Standard demand analysis implies welfare loss of $80/person
  - Value of statistical life calculation implies welfare gain of $3,000/person
- Choudhry et al. (NEJM 2011)
  - RCT elimination of (small) Rx co-pays for post-heart attack patients
  - 4-6 percentage point $\uparrow$ in medication adherence
  - Rates of total major vascular events $\downarrow$ by 1.8 ppt, heart attacks by 1.1 ppt
- Chandra, Flack, and Obermeyer (forthcoming QJE): mortality consequences of donut hole

Some methodological points: Summary

- Experiments great for testing nulls
  - Issues in experimental design
- Limitations of (some) experiments
  - Endogeneity eliminated by an experiment may be important economically (selection on mh)
  - Recovering an economic object of interest -economic models an important complement (e.g. non linear contracts - RAND HIE)
- (Some) comments on modeling
  - Choice of model is consequential ("sufficient" statistic is "sufficient" conditional on a model)
  - Sensible counterfactuals (don't go too far out of sample)

Complementarities between RF and "structural" work

- Credibility / transparency in (good) RF estimates and presentation
  - Always good to see the existence of phenomenon of interest "as close to the data" as possible
- RF work can help guide (consequential) modeling choices - e.g. static vs. dynamic behavioral model
- Values of modeling
  - Sometimes can't run an experiment (e.g. merger analysis; GE effects of health insurance (or can you?)...)
  - Counterfactual analysis (can't run an experiment for every permutation of a question)
    - But important to be sensible / dont' go "too far" out of sample
  - Welfare analysis - need utility model