Estimating the Distribution of Elasticity of Medical Expenditure Using a Notch in Out-of-Pocket Costs

Hae-young Hong

Seoul National University

Jan 4, 2025
AEA Annual Meeting

Overview

1. This paper develops **a novel method to estimate the joint distribution of price elasticities and medical expenditures,** using patient bunching behavior at a notch
   - Unlike traditional bunching estimation methods relying on polynomial approximations, this approach utilizes **a control group without a notch**
2. Applies the method to South Korea's policy, which features an **age-based shift in out-of-pocket (OOP) costs from linear to discontinuous**
   - The upper bound of elasticities is 0.17, the mean elasticity is 0.1, and the rank correlation between elasticity and medical expenditure is -0.52
3. Simulates **policy counterfactuals**
   - A linear coinsurance rate of 23.1% improves patient welfare and clinic revenue without increasing insurer spending

Overview

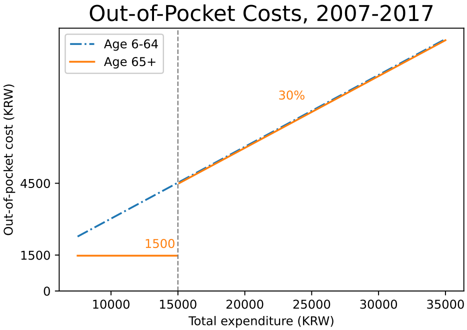
1. This paper develops **a novel method to estimate the joint distribution of price elasticities and medical expenditures,** using patient bunching behavior at a notch
   - Unlike traditional bunching estimation methods relying on polynomial approximations, this approach utilizes **a control group without a notch**
2. Applies the method to South Korea's policy, which features an **age-based shift in out-of-pocket (OOP) costs from linear to discontinuous**
   - The upper bound of elasticities is 0.17, the mean elasticity is 0.1, and the rank correlation between elasticity and medical expenditure is -0.52
3. Simulates **policy counterfactuals**
   - A linear coinsurance rate of 23.1% improves patient welfare and clinic revenue without increasing insurer spending

Overview

1. This paper develops **a novel method to estimate the joint distribution of price elasticities and medical expenditures,** using patient bunching behavior at a notch
   - Unlike traditional bunching estimation methods relying on polynomial approximations, this approach utilizes **a control group without a notch**
2. Applies the method to South Korea's policy, which features an **age-based shift in out-of-pocket (OOP) costs from linear to discontinuous**
   - The upper bound of elasticities is 0.17, the mean elasticity is 0.1, and the rank correlation between elasticity and medical expenditure is -0.52
3. Simulates **policy counterfactuals**
   - A linear coinsurance rate of 23.1% improves patient welfare and clinic revenue without increasing insurer spending

Institutional Setting: OOP Cost System in South Korea

- Age-based OOP cost system for outpatient visits in 2007-2017
  - Ages 6-64: Patients paid 30% of total expenditure (linear coinsurance)
  - Ages 65+:
    - For visits costing≤15,000 KRW: Fixed payment of 1,500 KRW
    - For visits costing>15,000 KRW: Patients paid 30% of total expenditure



- A "notch" at 15,000 KRW drives behavioral changes

Note: 1,067 KRW = 1 USD as of Dec 31, 2017.

Institutional Setting: Fee-for-Service System

How is total expenditure per visit determined?

- Total expenditure per visit is the sum of the fees for all services provided during a visit
- Fees are set by a national committee and are generally increased once a year
- Patients or physicians may exclude certain services to keep total expenditure below 15,000 KRW

Table 5: An Example of Fee-For-Service System: Physical Therapy

		(1) 2013	(2) 2014	(3) 2015	(4) 2016	(5) 2017	(6) 2018
A.	Outpatient Care - Established Patient	9,430	9,710 +	10,000	10,300	10,620	10,950
B.	Transcutaneous Electrical Nerve Stimulation	3,370	3,473 +	3,577	3,680	3,795	3,876
C.	Deep Heat Therapy	1,127	1,162	1,196	1,231	1,265	1,265
D.	Superficial Heat Therapy (with Deep Heat Therapy)	414	426	437	460	472	460
E.	Superficial Heat Therapy (without Deep Heat Therapy)	828	863 =	886	909	943	920
	Total Expenditure (A+B+C+D)	14,340	14,770 ≤15,000	15,210	15,670	16,150	16,550
	Highest Total Expenditure ≤15K	14,340	14,770	14,770	14,880	14,410	14,820
		(A+B+C+D)	(A+B+C+D)	(A+B+C)	(A+B+E)	(A+B)	(A+B)

Institutional Setting: Fee-for-Service System

How is total expenditure per visit determined?

- Total expenditure per visit is the sum of the fees for all services provided during a visit
- Fees are set by a national committee and are generally increased once a year
- Patients or physicians may exclude certain services to keep total expenditure below 15,000 KRW

Table 5: An Example of Fee-For-Service System: Physical Therapy

		(1) 2013	(2) 2014	(3) 2015	(4) 2016	(5) 2017	(6) 2018
A.	Outpatient Care - Established Patient	9,430	9,710	10,000 +	10,300	10,620	10,950
B.	Transcutaneous Electrical Nerve Stimulation	3,370	3,473	3,577 +	3,680	3,795	3,876
C.	Deep Heat Therapy	1,127	1,162	1,196	1,231	1,265	1,265
D.	Superficial Heat Therapy (with Deep Heat Therapy)	414	426	437	460	472	460
E.	Superficial Heat Therapy (without Deep Heat Therapy)	828	863	886 =	909	943	920
	Total Expenditure (A+B+C+D)	14,340	14,770	15,210 >15,000	15,670	16,150	16,550
	Highest Total Expenditure ≤15K	14,340	14,770	14,770	14,880	14,410	14,820
		(A+B+C+D)	(A+B+C+D)	(A+B+C)	(A+B+E)	(A+B)	(A+B)

Institutional Setting: Fee-for-Service System

How is total expenditure per visit determined?

- Total expenditure per visit is the sum of the fees for all services provided during a visit
- Fees are set by a national committee and are generally increased once a year
- Patients or physicians may exclude certain services to keep total expenditure below 15,000 KRW

Table 5: An Example of Fee-For-Service System: Physical Therapy

		(1) 2013	(2) 2014	(3) 2015	(4) 2016	(5) 2017	(6) 2018
A.	Outpatient Care - Established Patient	9,430	9,710	10,000 +	10,300	10,620	10,950
B.	Transcutaneous Electrical Nerve Stimulation	3,370	3,473	3,577 +	3,680	3,795	3,876
C.	Deep Heat Therapy	1,127	1,162	1,196	1,231	1,265	1,265
D.	Superficial Heat Therapy (with Deep Heat Therapy)	414	426	~~437~~	460	472	460
E.	Superficial Heat Therapy (without Deep Heat Therapy)	828	863	886	909	943	920
	Total Expenditure (A+B+C+D)	14,340	14,770	15,210	15,670	16,150	16,550
				=			
	Highest Total Expenditure ≤15K	14,340	14,770	14,770 ≤15,000	14,880	14,410	14,820
		(A+B+C+D)	(A+B+C+D)	(A+B+C)	(A+B+E)	(A+B)	(A+B)

Main Contribution to the Literature

I extend the **bunching estimation literature** by developing a novel method to estimate the elasticity distribution

- Most studies construct counterfactual distributions using **polynomial approximations**
  - e.g., Saez, 2010; Chetty et al., 2011; Kleven and Waseem, 2013; Seim, 2017; Bastani and Selin, 2014; Einav et al., 2017; Lu et al., 2019; Mortenson and Whitten, 2020; and Kim, 2021
- **Limitations** of existing bunching estimation methods
  1. Only the mean elasticity is estimated, **not the full distribution**
  2. It is **impossible to distinguish** the elasticity and the underlying distribution with a single budget set (Blomquist et al., 2021)
  3. For non-smooth underlying distributions, **estimation may fail entirely**
- I propose a framework linking the **ratio of treated and control densities** to the **joint distribution of elasticity and medical expenditure**

Main Contribution to the Literature

I extend the **bunching estimation literature** by developing a novel method to estimate the elasticity distribution

- Most studies construct counterfactual distributions using **polynomial approximations**
  - e.g., Saez, 2010; Chetty et al., 2011; Kleven and Waseem, 2013; Seim, 2017; Bastani and Selin, 2014; Einav et al., 2017; Lu et al., 2019; Mortenson and Whitten, 2020; and Kim, 2021
- **Limitations** of existing bunching estimation methods
  1. Only the mean elasticity is estimated, **not the full distribution**
  2. It is **impossible to distinguish** the elasticity and the underlying distribution with a single budget set (Blomquist et al., 2021)
  3. For non-smooth underlying distributions, **estimation may fail entirely**
- I propose a framework linking the **ratio of treated and control densities** to the **joint distribution of elasticity and medical expenditure**

Main Contribution to the Literature

I extend the **bunching estimation literature** by developing a novel method to estimate the elasticity distribution

- Most studies construct counterfactual distributions using **polynomial approximations**
  - e.g., Saez, 2010; Chetty et al., 2011; Kleven and Waseem, 2013; Seim, 2017; Bastani and Selin, 2014; Einav et al., 2017; Lu et al., 2019; Mortenson and Whitten, 2020; and Kim, 2021
- **Limitations** of existing bunching estimation methods
  1. Only the mean elasticity is estimated, **not the full distribution**
  2. It is **impossible to distinguish** the elasticity and the underlying distribution with a single budget set (Blomquist et al., 2021)
  3. For non-smooth underlying distributions, **estimation may fail entirely**
- I propose a framework linking the **ratio of treated and control densities** to the **joint distribution of elasticity and medical expenditure**

Notations

- m: **Total expenditure per visit**
- ϵ: **Elasticity** of total expenditure per visit with respect to OOP costs
- $\epsilon(m)$: Elasticity of the **marginal buncher** at m
  - Individuals choosing m under a linear coinsurance system are willing to bunch if their elasticity is greater than or equal to $\epsilon(m)$
- $\phi(\epsilon, m)$: Proportion of individuals with **friction** in optimal choices at (ϵ, m)
- Subscript 0: denotes distributions under **a linear coinsurance system** (e.g., $f_0(m)$, $F_0(m)$)
- Subscript 1: denotes distributions under **an OOP system with a notch** (e.g., $f_1(m)$, $F_1(m)$)
- If the subscript is omitted (e.g., $F(\epsilon, m)$, $F_{\epsilon|m}(\epsilon|m)$), it is assumed to refer to 0 (linear coinsurance system) for simplicity

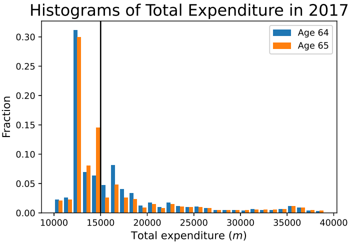
Notations

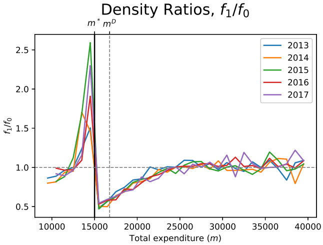
- m: **Total expenditure per visit**
- ϵ: **Elasticity** of total expenditure per visit with respect to OOP costs
- $\epsilon(m)$: Elasticity of the **marginal buncher** at m
  - Individuals choosing m under a linear coinsurance system are willing to bunch if their elasticity is greater than or equal to $\epsilon(m)$
- $\phi(\epsilon, m)$: Proportion of individuals with **friction** in optimal choices at (ϵ, m)
- Subscript 0: denotes distributions under **a linear coinsurance system** (e.g., $f_0(m)$, $F_0(m)$)
- Subscript 1: denotes distributions under **an OOP system with a notch** (e.g., $f_1(m)$, $F_1(m)$)
- If the subscript is omitted (e.g., $F(\epsilon, m)$, $F_{\epsilon|m}(\epsilon|m)$), it is assumed to refer to 0 (linear coinsurance system) for simplicity

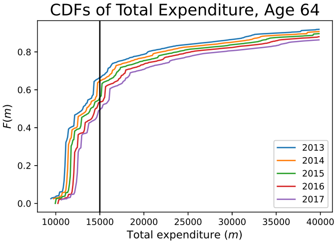
Data

- Data: **The Korean National Health Information Database**
  - Administrative data collected by the National Health Insurance Service (NHIS)
  - Covers the entire population of residents in South Korea ($\because$ NHIS is the single insurer)
  - Covers the entire medical providers ($\because$ there is no private sector)
- Analysis Sample
  - Individuals turning 65 years old in each calendar year between 2013 and 2017
  - **Medical claims of the last visit at age 64 and the first visit at age 65 for the same disease category**
  - Claims violating the OOP cost formulas are excluded
  - Medical Aid beneficiaries are excluded
- Variables: total expenditure, OOP cost, principle diagnosis, age
- Descriptive Statistics

Bunching Patterns in the Data



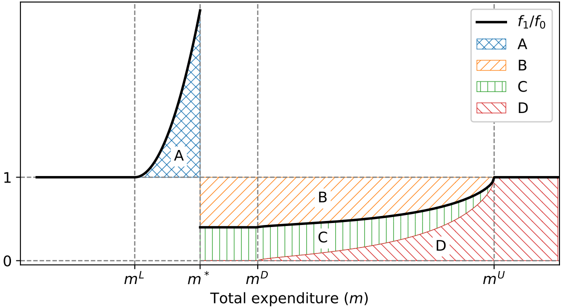




1. Age-64 density (under a linear coinsurance) is not smooth
2. There is the upper bound of bunching responses
3. There is a variation in the CDFs across years

Identification (1): Strategy

How is f_1/f_0 decomposed?



$$1 - \frac{f_1(m)}{f_0(m)} = \underbrace{\left(1-\bar{\phi}\right)\Pr\{\epsilon \geq \epsilon(m)\,|m\}}_{\text{B: more elastic than marginal buncher, without friction}}$$

$$\frac{f_1(m)}{f_0(m)} = \underbrace{\Pr\{\epsilon < \epsilon(m)\,|m\}}_{\text{D: less elastic than marginal buncher}} + \underbrace{\bar{\phi}\Pr\{\epsilon \geq \epsilon(m)\,|m\}}_{\text{C: more elastic than marginal buncher, with friction}}$$

Identification (1): Strategy

$$\frac{f_1(m)}{f_0(m)} = F_{\epsilon|m}(\epsilon(m)|m) + \bar{\phi}\left[1 - F_{\epsilon|m}(\epsilon(m)|m)\right]$$

$$\Rightarrow F_{\epsilon|m}\left(\epsilon(m)|m\right) = 1 - \left(1 - \bar{\phi}\right)^{-1}\left(1 - \frac{f_1(m)}{f_0(m)}\right)$$

$F_{\epsilon|m}$: Independence, or Copula

$\epsilon(m)$, $\bar{\phi}$: Structural model

$\frac{f_1(m)}{f_0(m)}$: Observed

Identification (2): Structural Model of Bunching

- Quasi-linear and constant elasticity preferences (e.g. Saez, 2010; Kleven and Waseem, 2013)
- I adopt Einav et al. (2017)'s utility function

$$u(m;\zeta,\eta) = g(m) + c = \left[2m - \frac{\zeta}{1+\frac{1}{\eta}}\left(\frac{m}{\zeta}\right)^{1+\frac{1}{\eta}}\right] + [y - s(m)] \quad (1)$$

m: total expenditure for a visit, $s(m)$: out-of-pocket cost,
ζ: health needs, η: elasticity, y: income

- Optimal choice under **a linear coinsurance system,** $s(m) = sm$

$$m(\zeta,\eta,s) = \zeta(2-s)^{\eta}. \quad (2)$$

  - If $s = 1$ (no insurance coverage), $m = \zeta$
  - $\eta = \partial \log m / \partial \log(2-s)$
  - $\varepsilon \equiv |\partial \log m / \partial \log s| = \eta \times \frac{s}{2-s}$

Identification (2): Structural Model of Bunching

- Quasi-linear and constant elasticity preferences (e.g. Saez, 2010; Kleven and Waseem, 2013)
- I adopt Einav et al. (2017)'s utility function

$$u(m;\zeta,\eta) = g(m) + c = \left[2m - \frac{\zeta}{1+\frac{1}{\eta}}\left(\frac{m}{\zeta}\right)^{1+\frac{1}{\eta}}\right] + [y - s(m)] \qquad (1)$$

m: total expenditure for a visit, $s(m)$: out-of-pocket cost,
ζ: health needs, η: elasticity, y: income

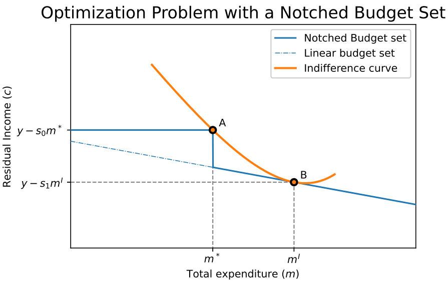
- Optimal choice under **a linear coinsurance system,** $s(m) = sm$

$$m(\zeta,\eta,s) = \zeta(2-s)^{\eta}. \qquad (2)$$

  - If $s = 1$ (no insurance coverage), $m = \zeta$
  - $\eta = \partial \log m / \partial \log(2-s)$
  - $\varepsilon \equiv |\partial \log m / \partial \log s| = \eta \times \frac{s}{2-s}$

Identification (2): Structural Model of Bunching

Optimal choice under **an OOP with a notch,** $s(m) = \begin{cases} s_0 m & \text{if } m \leq m^* \\ s_1 m & \text{if } m > m^* \end{cases}$



The conditions for the **marginal buncher** at $m^I > m^*$

1. $\zeta^I = \frac{m^I}{(2-s_1)^\eta}$ (Under a linear coinsurance system, m^I is chosen)
2. $u^* = u(m^*) = (2 - s_0)\, m^* - \frac{m^I (2-s_1)}{1+1/\eta} \left(\frac{m^*}{m^I}\right)^{1+\frac{1}{\eta}} + y$ (Utility at m^* under an OOP with a notch)
3. $u^I = u(m^I) = \frac{m^I (2-s_1)}{1+\eta} + y$ (Utility at m^I under an OOP with a notch)

Identification (2): Structural Model of Bunching

- $\epsilon(m)$: Elasticity of the marginal buncher
  - The condition that $u^* = u^I$ leads to the following equation:

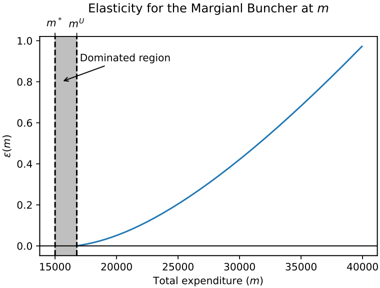
$$(2 - s_0)\left(\frac{m^*}{m^I}\right) - \frac{2 - s_1}{1 + 1/\eta}\left(\frac{m^*}{m^I}\right)^{1+\frac{1}{\eta}} - \frac{2 - s_1}{1 + \eta} = 0 \qquad (3)$$

  - $\eta(m)$ denotes η solving equation (3) when $m^I = m$ for given s_0, s_1 and m^*
    - s_0, s_1, m^* are given as policy variables
    - In the Korea's age-based OOP system, $s_0 = 0.1$, $s_1 = 0.3$, $m^* = 15,000$
  - By the relationship between η and ϵ,

$$\epsilon(m) = \eta(m) \times \frac{s_1}{1 - s_1}$$

Identification (2): Structural Model of Bunching

- Properties of $\epsilon(m)$
  1. There exists a unique $\epsilon(m)$ for any $m > \left(\frac{2-s_0}{2-s_1}\right) m^*$ and given s_0, s_1 and m^*.
  2. $\epsilon(m)$ is strictly increasing in m



- Dominated Region
  : The region (m^*, m^D) is not rationalized by any value of elasticity, where

$$m^D = \left(\frac{2-s_0}{2-s_1}\right) m^* \tag{4}$$

  - e.g., $s_0 = 0.1$, $s_1 = 0.3$, and $m^* = 15,000 \Rightarrow m^D \approx 16,765$

Identification (3): Probability Distribution

- **Case 1: If ϵ and m are independent, and $\phi = 0$ (no friction)**

$$F_\epsilon(\epsilon(m)) = \frac{f_1(m)}{f_0(m)} \quad (5)$$

  - F_ϵ is identified
    $\because$ for any $\epsilon \in [0, \epsilon^U]$ there exists a unique $m \in [m^D, m^U]$ such that $\epsilon = \epsilon(m)$ and $f_1(m)$ and $f_0(m)$ are observed.
  - Interpretation: The proportion of individuals who still choose m under a discontinuous OOP system represents the probability of having an elasticity less than that of the marginal buncher at m

Identification (3): Probability Distribution

- **Case 1: If ϵ and m are independent, and $\phi = 0$ (no friction)**

$$F_\epsilon \left(\epsilon \left(m\right)\right) = \frac{f_1 \left(m\right)}{f_0 \left(m\right)} \tag{5}$$

  - F_ϵ is identified
    $\because$ for any $\epsilon \in \left[0, \epsilon^U\right]$ there exists a unique $m \in \left[m^D, m^U\right]$ such that $\epsilon = \epsilon\left(m\right)$ and $f_1\left(m\right)$ and $f_0\left(m\right)$ are observed.
  - Interpretation: The proportion of individuals who still choose m under a discontinuous OOP system represents the probability of having an elasticity less than that of the marginal buncher at m

Identification (3): Probability Distribution

- **Assumption 1:** $\phi(\epsilon, m)$ is constant on $\left(m^*, m^U\right]$, i.e. $\bar{\phi} \equiv \phi(\epsilon, m)$.
- $\bar{\phi}$: Proportion of individuals with **friction**
  - Following Kleven and Waseem (2013), $\bar{\phi}$ is identified using the **dominated region**

$$\bar{\phi} = \frac{\int_{m^*}^{m^D} f_1(m)\, dm}{\int_{m^*}^{m^D} f_0(m)\, dm} \tag{6}$$

  - Interpretation: Observations in the dominated region $\left(m^*, m^D\right]$ are entirely attributed to optimization friction
- **Case2: If ϵ and m are independent and $\phi > 0$,**

$$\frac{f_1(m)}{f_0(m)} = \underbrace{F_\epsilon(\epsilon(m))}_{\text{Inelastic}} + \underbrace{\bar{\phi}\left[1 - F_\epsilon(\epsilon(m))\right]}_{\text{Elastic, but with friction}}$$

$$\Rightarrow F_\epsilon(\epsilon(m)) = 1 - \frac{1}{1-\bar{\phi}}\left[1 - \frac{f_1(m)}{f_0(m)}\right] \tag{7}$$

  - F_ϵ is identified $\because$ for any $\epsilon \in \left[0, \epsilon^U\right]$ there exists a unique $m \in \left[m^D, m^U\right]$ such that $\epsilon = \epsilon(m)$ and $f_1(m)$, $f_0(m)$, and $\bar{\phi}$ are observed.

Identification (3): Probability Distribution

- To allow dependence between ϵ and m, I adopt a copula approach (Sklar, 1973)
  - **Assumption 2**: There exists a twice differentiable bivariate copula C with dependence parameter θ such that $F^t(\epsilon, m) = C\left(F_\epsilon^t(\epsilon), F_0^t(m); \theta\right)$ where $F_\epsilon^t(\epsilon)$ and $F_0^t(m)$ are the marginal CDFs of ϵ and m in year t, respectively.
  - **Assumption 3**: The marginal CDF of ϵ is stationary. $F_\epsilon^t(\epsilon) = F_\epsilon(\epsilon)\ \forall t$.
  - **Assumption 4**: ϵ is distributed as Beta with parameters (α, β) on a support $\left(0, \epsilon^U\right)$ where $\epsilon^U = \epsilon\left(m^U\right)$.

- **Case 3: If ϵ and m are dependent and $\phi > 0$,** the conditional CDF of ϵ given m can be represented as a function of marginal CDFs of ϵ and m (By **Assumption 2**)

$$F_{\epsilon|m}^t(\epsilon|m) = h\left(\underbrace{F_\epsilon\left(\epsilon; \alpha, \beta, \epsilon^U\right)}_{\textbf{Assumptions 3\&4}}, F_0^t(m); \theta\right)$$

$$\Rightarrow \frac{f_1^t(m)}{f_0^t(m)} = \underbrace{1 - \left(1 - \bar{\phi}\right)\left\{1 - h\left[F_\epsilon\left(\epsilon(m); \alpha, \beta, \epsilon^U\right), F_0^t(m); \theta\right]\right\}}_{\equiv R\left(m; F_0^t, \Omega\right)} \qquad (8)$$

where $h(u_1, u_2) = \partial C(u_1, u_2)/\partial u_2 = F_{\epsilon|m}(\epsilon|m)$ and $\Omega = \left(\alpha, \beta, \epsilon^U, \theta\right)$

Identification (3): Probability Distribution

- To allow dependence between ϵ and m, I adopt a copula approach (Sklar, 1973)
  - **Assumption 2**: There exists a twice differentiable bivariate copula C with dependence parameter θ such that $F^t(\epsilon, m) = C\left(F^t_\epsilon(\epsilon), F^t_0(m); \theta\right)$ where $F^t_\epsilon(\epsilon)$ and $F^t_0(m)$ are the marginal CDFs of ϵ and m in year t, respectively.
  - **Assumption 3**: The marginal CDF of ϵ is stationary. $F^t_\epsilon(\epsilon) = F_\epsilon(\epsilon)\ \forall t$.
  - **Assumption 4**: ϵ is distributed as Beta with parameters (α, β) on a support $\left(0, \epsilon^U\right)$ where $\epsilon^U = \epsilon\left(m^U\right)$.

- **Case 3: If ϵ and m are dependent and $\phi > 0$,** the conditional CDF of ϵ given m can be represented as a function of marginal CDFs of ϵ and m (By **Assumption 2**)

$$F^t_{\epsilon|m}(\epsilon|m) = h\left(\underbrace{F_\epsilon\left(\epsilon; \alpha, \beta, \epsilon^U\right)}_{\textbf{Assumptions 3\&4}}, F^t_0(m); \theta\right)$$

$$\Rightarrow \frac{f^t_1(m)}{f^t_0(m)} = \underbrace{1 - \left(1 - \bar{\phi}\right)\left\{1 - h\left[F_\epsilon\left(\epsilon(m); \alpha, \beta, \epsilon^U\right), F^t_0(m); \theta\right]\right\}}_{\equiv R\left(m; F^t_0, \Omega\right)} \qquad (8)$$

where $h(u_1, u_2) = \partial C(u_1, u_2)/\partial u_2 = F_{\epsilon|m}(\epsilon|m)$ and $\Omega = \left(\alpha, \beta, \epsilon^U, \theta\right)$

Estimation

- Densities $f_0^t(m)$, $f_1^t(m)$, and $F_0^t(m)$
  - Using **weighted histogram estimates**

$$\hat{f}(M_j) = \frac{1}{Nb} \sum_{i=1}^{N} w_i 1\{m_i \in \mathcal{B}_j\}$$

  where $\mathcal{B}_j = (m^* + (j-1)b, m^* + jb]$ for $j = \ldots, -1, 0, 1, \ldots,$ and M_j is the midpoint of bin $\mathcal{B}_j$, $M_j = m^* + \left(j - \frac{1}{2}\right)b$
  - I reweight the sample to ensure consistent disease composition across years
  - The CDF of $m \leq M_j$ is defined using $\hat{f}(M_j)$.

$$\hat{F}_0^t(M_j) = \sum_{k \leq j} b\hat{f}_g^t(M_k) - \frac{b}{2}\hat{f}_g^t(M_j)$$

Estimation

- Fraction of friction $\bar{\phi}$
  - Proportion of individuals in the dominated region

$$\hat{\bar{\phi}} = \frac{\sum_t \sum_j \hat{f}_1^t(M_j) 1\left\{M_j \in \left(m^*, m^D\right]\right\}}{\sum_t \sum_j \hat{f}_0^t(M_j) 1\left\{M_j \in (m^*, m^D]\right\}}$$

- The upper bound of bunching m^U
  - The lowest point where the cumulative treated density converges to the cumulative control density indicates the end of bunching behavior
  - For a bandwidth h,

$$\hat{m}^U = \min\left\{M_j - \frac{b}{2} : \sum_t \sum_{M_k \in [M_j, M_j + h]} \hat{f}_1^t(M_k) \geq \sum_t \sum_{M_k \in [M_j, M_j+h]} \hat{f}_0^t(M_k)\right\}$$

Estimation

- Fraction of friction $\bar{\phi}$
  - Proportion of individuals in the dominated region

$$\hat{\bar{\phi}} = \frac{\sum_t \sum_j \hat{f}_1^t (M_j) 1 \left\{ M_j \in \left(m^*, m^D \right] \right\}}{\sum_t \sum_j \hat{f}_0^t (M_j) 1 \{ M_j \in (m^*, m^D] \}}$$

- The upper bound of bunching m^U
  - The lowest point where the cumulative treated density converges to the cumulative control density indicates the end of bunching behavior
  - For a bandwidth h,

$$\hat{m}^U = \min \left\{ M_j - \frac{b}{2} : \sum_t \sum_{M_k \in [M_j, M_j + h]} \hat{f}_1^t (M_k) \geq \sum_t \sum_{M_k \in [M_j, M_j + h]} \hat{f}_0^t (M_k) \right\}$$

Estimation

- Parametric models for F_ϵ and $h(\cdot, \cdot; \theta)$
  - $\epsilon \sim$ Beta(α, β) on the support $\left(0, \epsilon^U\right)$, where $\epsilon^U = \epsilon\left(m^U\right)$
  - $h(\cdot, \cdot; \theta)$ is defined by one of four popular copulas (Clayton, Gumbel, Frank, or Gaussian) and their rotations

    $\rightarrow$ The best-fit copula is selected based on the smallest RMSE
- α, β and θ are estimated via least squares

$$\left(\hat{\alpha}, \hat{\beta}, \hat{\theta}\right) = \underset{\alpha,\beta,\theta}{\arg\min} \sum_{j,t} W_{j,t} \left[\hat{g}^t\left(M_j; \Omega\right)\right]^2$$

  - $\hat{g}^t\left(M_j; \Omega\right)$:

$$\hat{g}^t\left(M_j; \Omega\right) = \overbrace{\hat{f}_1^t\left(M_j\right)}^{\text{Observed probability}} - \overbrace{\hat{f}_0^t\left(M_j\right)\hat{R}\left(M_j; \hat{F}_0^t, \Omega\right)}^{\text{Predicted probability from a parametric model}}$$

where $\hat{R}^t\left(M_j; \Omega\right) = 1 - \left(1 - \hat{\bar{\phi}}\right)\left\{1 - h\left[F_\epsilon\left(\epsilon\left(M_j\right); \alpha, \beta, \hat{\epsilon^U}\right), \hat{F}_0^t\left(M_j\right); \theta\right]\right\}$

  - $W_{j,t}$: the inverse of variance of $\hat{f}_1^t\left(M_j\right)$
- Bootstrap standard errors are calculated by repeating the estimation for 1,000 bootstrap replicates of the simulation sample

Estimation Results

Table 2: Estimates of Parameters of the Joint Distribution of ϵ and m

Copula	Gumbel90	Gumbel270	Clayton90	Clayton270	Frank0	Gaussian0
ϕ	0.53	0.53	0.53	0.53	0.53	0.53
	[0.52, 0.53]	[0.52, 0.53]	[0.52, 0.53]	[0.52, 0.53]	[0.52, 0.53]	[0.52, 0.53]
m^U	24,100	24,100	24,100	24,100	24,100	24,100
	[23,000, 25,600]	[23,000, 25,600]	[23,000, 25,600]	[23,000, 25,600]	[23,000, 25,600]	[23,000, 25,600]
$\mathbb{E}(\epsilon)$	0.11	0.11	0.12	0.10	0.11	0.11
	[0.09, 0.14]	[0.09, 0.14]	[0.10, 0.15]	[0.09, 0.13]	[0.09, 0.14]	[0.09, 0.14]
$\sigma(\epsilon)$	0.070	0.069	0.067	0.070	0.067	0.069
	[0.055, 0.092]	[0.054, 0.092]	[0.052, 0.089]	[0.055, 0.092]	[0.052, 0.089]	[0.054, 0.091]
τ	-0.60	-0.71	-0.75	-0.52	-0.62	-0.67
	[-0.66, -0.49]	[-0.77, -0.61]	[-0.81, -0.65]	[-0.58, -0.43]	[-0.70, -0.47]	[-0.73, -0.56]
RMSE	3.021	3.026	3.036	3.021	3.029	3.025
	[3.117, 3.472]	[3.123, 3.486]	[3.133, 3.503]	[3.115, 3.472]	[3.126, 3.493]	[3.122, 3.483]
Prob. of Least RMSE	0.232	0.002	0.000	0.754	0.004	0.008

- Best-fit copulas: Clayton rotated by 270°
  - Probability of achieving the least RMSE: 75.4%
- Upper bound of bunching response: 24,100 KRW Sensitivity
- $\mathbb{E}(\epsilon)$: 0.1
- Kendall's τ: -0.52

Counterfactual Simulation

- A simulation sample $\{(m_k, \epsilon_k, d_k)\}_{k=1}^{K}$ with $K = 200,000$
  - $\{(F(\epsilon_k), F(m_k))\}$ are drawn using the Clayton270
  - $\{m_k\}$ are drawn from the age-64 distribution of total expenditure in 2017
  - $\{\epsilon_k\}$ are drawn from the Beta distribution with estimated parameters
  - d_k is an indicator for the presence of optimization frictions, and $\{d_k\}$ are drawn from the binomial distribution of probability $\hat{\bar{\phi}}$
  - For each pair of (m_k, ϵ_k), individual type ζ_k is constructed using the equation (2)
  - For each (ζ_k, ϵ_k), I solve the patient optimization problem for alternative out-of-pocket systems $s^c(m) \Rightarrow m_k^c$
- Welfares
  - Patient welfare: $\sum_k u(m_k^c; \zeta_k, \eta_k)$
  - Clinic revenue: $\sum_k m_k^c$
  - Insurer spending: $\sum_k (m_k^c - s^c(m_k^c))$

Counterfactual Simulation

- Policy counterfactuals
  - Baseline welfare: 2017 System (a single notch at 15,000 KRW)
  - The OOP system reformed in 2018 (one kink and two notches)
  - Linear coinsurance that make patients, clinics, and the insurer indifferent to the baseline welfare, respectively
  - Smoothly changing coinsurance

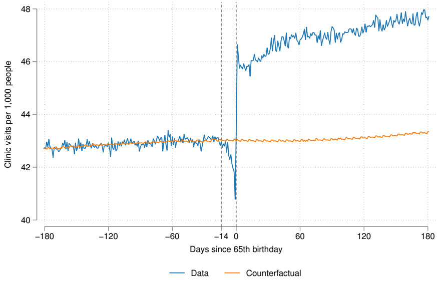
Table 3: Policy Counterfactuals

	Baseline Welfare	Difference in Welfare				
	(1) 2017 System	(2) 2018 System	(3) Patient Equivalent	(4) Clinic Equivalent	(5) Insurer Equivalent	(6) Smooth Cubic
Coinsurance Rate	-	-	0.232	0.313	0.231	-
Patient	26,619	799	0	-1,607	16	573
	(18,811)	(18,683)	(20,304)	(19,530)	(20,312)	(18,659)
Clinic	19,792	455	200	0	202	306
	(9,849)	(9,595)	(9,868)	(9,928)	(9,868)	(9,664)
Insurer	-15,373	-1,043	18	1,771	0	-699
	(6,010)	(5,741)	(7,580)	(6,823)	(7,587)	(5,733)

- Results
  - A linear coinsurance rate of 23.1% improves patient welfare and clinic revenue without increasing insurer spending
  - The 2018 reform worsened the financial burden on the NHIS

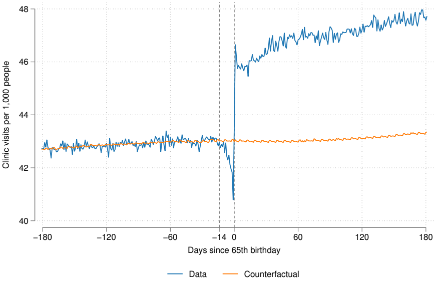
Conclusion

- Key takeaways
  - This study proposes a novel method to estimate the full distribution of elasticities and medical expenditure using a control group, avoiding the limitations of polynomial approximations
  - By eliminating the notch and transitioning to a linear coinsurance system (e.g., 23.1%), the system could reduce behavioral distortions, and enhance welfare without increasing public spending
- Ongoing Research: Responses on the number of visits (Hong, 2024)
  - People begin reducing clinic visits before their 65th birthday and increase visits immediately after turning 65



Conclusion

- Key takeaways
  - This study proposes a novel method to estimate the full distribution of elasticities and medical expenditure using a control group, avoiding the limitations of polynomial approximations
  - By eliminating the notch and transitioning to a linear coinsurance system (e.g., 23.1%), the system could reduce behavioral distortions, and enhance welfare without increasing public spending
- Ongoing Research: Responses on the number of visits (Hong, 2024)
  - People begin reducing clinic visits before their 65th birthday and increase visits immediately after turning 65



Thank you for your attention!

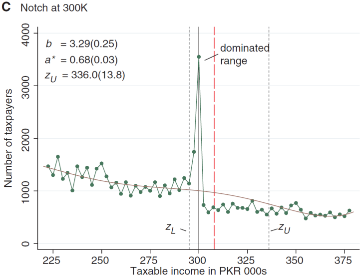
Appendix: Related Literature

Price elasticity of demand for medical care

- Responses on the intensive margin
  - Contrary to the RAND Health Insurance Experiment: cost-sharing affects the number of episodes but not the cost per episode (e.g., Manning et al., 1987; Lohr et al., 1986, Keeler and Rolph, 1988; and Aron-Dine et al., 2013
  - Oregon Health Insurance Experiment (e.g., Finkelstein et al., 2012), and empirical studies (e.g.Brot-Goldberg et al., 2017; Ellis et al., 2017; Choi et al., 2010; and Choi, 2018) mostly have focused on the extensive margin
- Reposes to a small amount of medical expenditure
  - Previous studies have focused on relatively large amounts, such as total annual expenditures (e.g., Einav et al., 2017) and monthly expenditures (e.g., Ellis et al., 2017)
  - However, this paper focuses on responses to small OOP changes, from 1,500 KRW to 4,500 KRW
- In Korea (e.g. Kim and Kwon, 2010; Na, 2020; and Kim, 2021)

Appendix: Limitations of Existing Bunching Estimation Methods

- Most studies in the bunching estimation literature construct counterfactual distributions using polynomial approximations
  - e.g. Saez, 2010; Chetty et al., 2011; Kleven and Waseem, 2013; Seim, 2017; Bastani and Selin, 2014; Einav et al., 2017; Lu et al., 2019; Mortenson and Whitten, 2020; and Kim, 2021



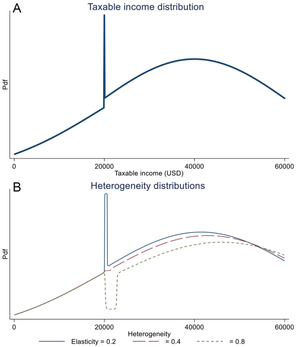
Source: Figure 6, Kleven and Waseem (2013)

- Assumptions:
  - Counterfactual density is **smooth around the threshold**
  - Counterfactual density is **locally constant**

Back

Appendix: Limitations of Existing Bunching Estimation Methods

- Criticism by Blomquist et al. (2021)
  - The size of the kink or notch probability depends on both elasticity and the distribution of individuals around the kink or notch
  - It is impossible to distinguish the elasticity from the underlying distribution with a single budget set



Source: Figure 3, Blomquist et al. (2021)

Back

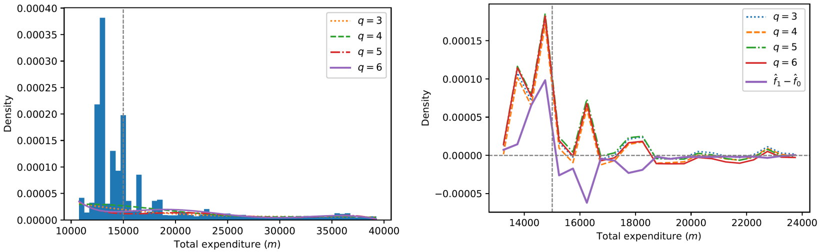
Appendix: Limitations of Existing Bunching Estimation Methods

If polynomial approximation is applied to Korea's age-based OOP system:

Figure A.1: Polynomial Approximations of the Counterfactual Density

(a) Polynomial Approximations

(b) Coefficients for the Bunching Window



- Non-smooth distributions cannot be accurately estimated, even with higher-order polynomials.
- Polynomial approximations fail to capture the bunching response

Back

Appendix: Descriptive Statistics

Table 1: Descriptive Statistics for Analysis Sample

Year	2013		2014		2015		2016		2017	
Age	64	65	64	65	64	65	64	65	64	65
Panel A: Number of Medical Bills										
Total Bills	278,412	278,412	272,635	272,635	263,834	263,834	260,536	260,536	354,906	354,906
Without Additional Services	65,245	57,850	63,385	54,145	61,894	53,987	57,215	54,054	81,428	74,892
With Additional Services	213,167	220,562	209,250	218,490	201,940	209,847	203,321	206,482	273,478	280,014
Panel B: Total Expenditure for Bills with Additional Services (KRW)										
Mean	19,207	19,721	20,364	20,625	21,686	21,999	22,886	23,650	24,597	25,614
Std. Dev.	20,351	21,884	21,979	23,271	24,305	25,723	25,577	28,354	28,568	32,375
Mean $\leq$ 40K	14,891	14,845	15,307	15,200	15,733	15,551	16,136	15,997	16,565	16,416
Std. Dev.$\leq$40K	6,168	6,077	6,234	6,073	6,327	6,153	6,136	6,207	6,184	6,276
25th Percentile	11,150	11,150	11,450	11,450	11,750	11,750	12,150	12,150	12,550	12,550
50th Percentile	13,150	13,250	13,550	13,650	14,050	13,950	14,550	14,450	15,150	14,750
75th Percentile	17,150	16,750	18,550	17,550	19,750	19,550	21,350	22,250	22,950	23,950
Fraction $\leq$15K	0.66	0.71	0.65	0.70	0.55	0.64	0.53	0.61	0.49	0.57
Fraction >40K	0.08	0.09	0.09	0.10	0.10	0.11	0.12	0.13	0.14	0.15

Back

Appendix: Copula Families

Table 6: Properties of Copula Families

Copula	$C(u_1, u_2; \theta)$	$\theta \in$	Kendall's τ^*	$\lambda_L{}^\dagger$	$\lambda_U{}^\ddagger$
Clayton	$\left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{-1/\theta}$	$(0, \infty)$	$\frac{\theta}{2+\theta}$	$2^{-1/\theta}$	0
Gumbel	$\exp\left(-\left[(-\log u_1)^\theta + (-\log u_2)^\theta\right]^{1/\theta}\right)$	$[1, \infty)$	$\frac{\theta-1}{\theta}$	0	$2 - 2^{1/\theta}$
Frank	$-\frac{1}{\theta}\log\left[1 + \left(e^{-\theta u_1} - 1\right)\left(e^{-\theta u_2} - 1\right)\left(e^{-\theta} - 1\right)^{-1}\right]$	$(-\infty, \infty)$	$1 + \frac{4}{\theta}\left(\frac{1}{\theta}\int_0^\theta \frac{t}{e^t - 1}dt - 1\right)$	0	0
Gaussian	$\Phi_G\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \theta\right)^\S$	$(-1, 1)$	$\frac{2}{\pi}\arcsin(\theta)$	0	0

Notes: The table shows properties of four copula families: Clayton (1978), Gumbel (1960), Frank (1978), and Gaussian. The properties are from Nelsen (2006) and Trivedi and Zimmer (2007).

* Kendall's τ is defined by $\Pr[(X_1 - X_2)(Y_1 - Y_2) > 0] - \Pr[(X_1 - X_2)(Y_1 - Y_2) < 0]$ where (X_1, Y_1) and (X_2, Y_2) are independent pairs from joint distribution $F(X, Y)$.

† Lower tail dependence λ_L is defined by $\lim_{v \to 0+} \Pr(u_1 < v | u_2 < v)$.

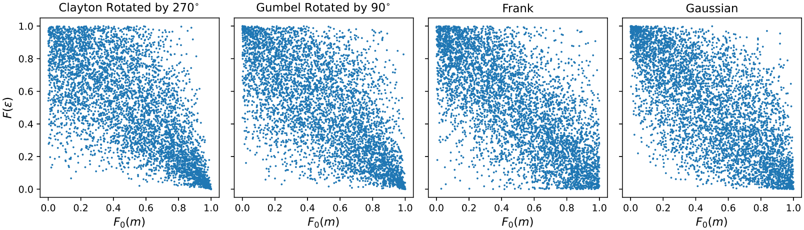
‡ Upper tail dependence λ_U is defined by $\lim_{v \to 1-} \Pr(u_1 > v | u_2 > v)$.

§ $\Phi_G(\cdot, \cdot; \theta)$ is the standard bivariate normal distribution with correlation θ. Φ^{-1} is the inverse standard normal distribution.

◂ Back

Appendix: Probability Distribution

Figure: Scatter Plots of $F(\epsilon)$ and $F_0(m)$ by Copula Family when $\tau = -0.5$

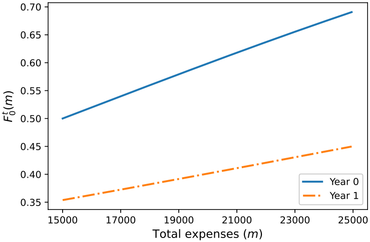


Notes: This figure illustrates the scatter plots of $F(\epsilon)$ and $F_0(m)$ generated by copula functions $C(F(\epsilon), F(m))$ when $\tau = -0.5$. The x-axis is $F(m)$ and the y-axis is $F(\epsilon)$.

Appendix: Probability Distribution

What does f_1/f_0 look like depending on the presence of frictions and the dependence between ϵ and m?

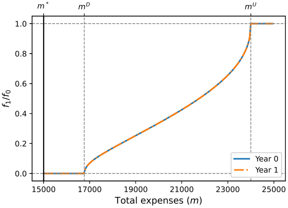
- When the cumulative distribution of total expenditure shifts to the right due to fee increases



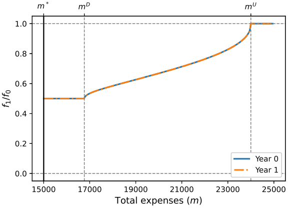
(a) $F^t\,(m)$

Appendix: Probability Distribution

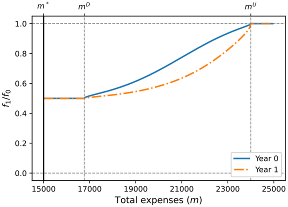
What does f_1/f_0 look like depending on the presence of frictions and the dependence between ϵ and m?



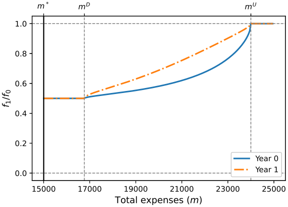
(b) $\bar{\phi} = 0$, $\tau = 0$



(c) $\bar{\phi} = 0.5$, $\tau = 0$

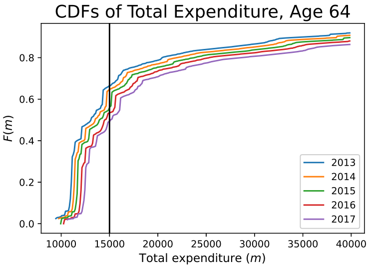


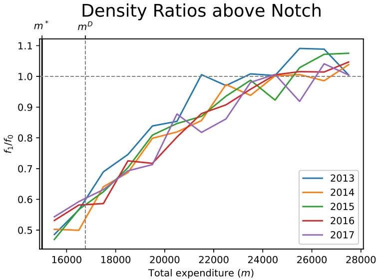
(d) $\bar{\phi} = 0.5$, $\tau = -0.5$



(e) $\bar{\phi} = 0.5$, $\tau = 0.5$

Appendix: Probability Distribution

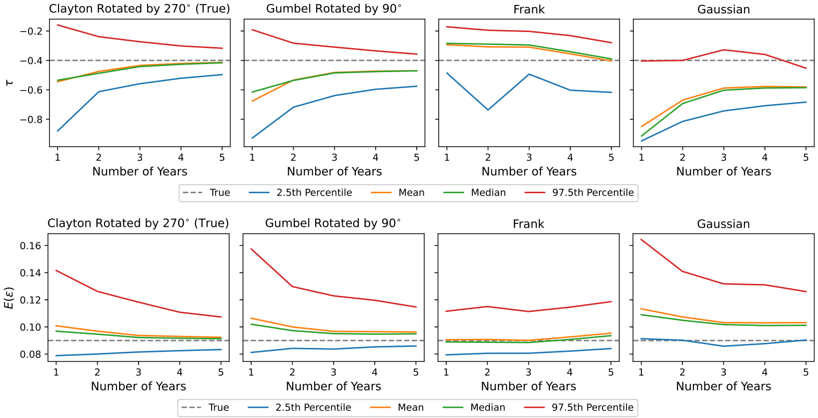




- The cumulative distribution of total expenditure shifts to the right due to fee increases
- Shows evidence of $\phi \neq 0$ and $\tau < 0$

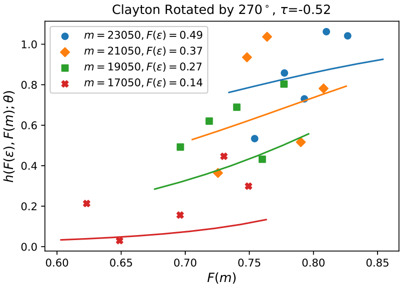
Appendix: Monte Carlo Simulation

- True sample is generated by Clayton copula rotated by 270 degrees with $\tau = -0.4$. The distribution of ϵ is beta distribution with $\mathbb{E}(\epsilon) = 0.09$, and $\sigma(\epsilon) = 0.055$.
- The sample size is 200,000 for each year and age.



Appendix: Estimation

Figure: Estimation of Kendall's τ

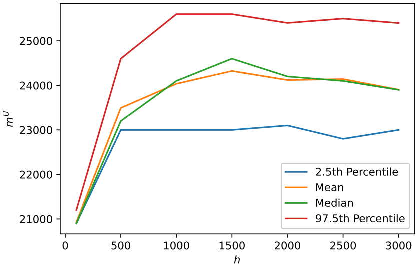


- Dots represent the conditional cumulative distribution function derived from observed data: $F_{\epsilon|m}\left(\epsilon\left(\hat{m}\right)|m\right) = 1 - \frac{1}{1-\hat{\hat{\phi}}}\left[1 - \frac{\hat{f}_1(m)}{\hat{f}_0(m)}\right]$
- Lines represent the conditional cumulative distribution function derived from parameter estimation: $h\left[F_\epsilon\left(\epsilon\left(m\right);\hat{\alpha},\hat{\beta},\epsilon^{\hat{U}}\right),\hat{F}_0^t\left(m\right);\hat{\theta}\right]$

Appendix: Estimation of the Upper Bound of Bunching Window

Sensitivity check by bandwidth choice

Figure: Estimates of the Upper Bound of Bunching Window by Bandwidth

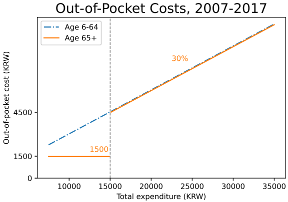


- The estimates of the upper bound are stabilized at around 24,000 KRW for bandwidths above 1,000 KRW

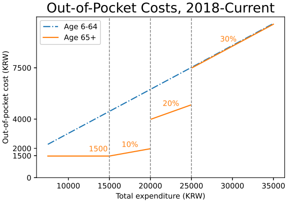
Back

Appendix: Counterfactual Policies

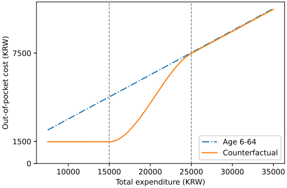
Figure: Counterfactual Policies



(a) Baseline OOP System



(b) The OOP system reformed in 2018



(c) Smoothly Changing OOP

Back