Estimating the Distribution of Elasticities of Medical Expenditures Using a Notch in Out-of-Pocket Cost

Hae-young Hong*

November 5, 2024

Abstract

This paper develops a novel method for estimating the distribution of elasticities of medical expenditures with respect to out-of-pocket costs, leveraging patient responses at a notch in South Korea. Using a natural experiment that exploits differences in the out-of-pocket costs for 64- and 65-year-olds within the same calendar year, a strategy is developed to characterize the conditional cumulative distribution of elasticities given medical expenditures as a function of observable variables. A copula approach is employed to account for the dependence between elasticities and medical expenditures. Using Korean health insurance administrative data from 2013-2017, the study finds the upper bound of the elasticities is 0.17 and the mean is 0.1. Counterfactual policy simulations show that introducing a linear coinsurance rate of 23.1% instead of a notch can improve the welfare of patients and clinics without increasing the insurer's spending.

*Postdoctoral Fellow, Seoul National University, econhonx@gmail.com

I would like to express my gratitude to my primary advisor, Jeffrey Smith, for his support and guidance throughout the dissertation process. I would also like to thank my committee members, Naoki Aizawa, Corina Mommaerts, and John Mullahy, for their valuable feedback and insights on my research. I am also grateful to the participants of the seminars where I presented my research, which has significantly contributed to the development of this paper.

This material is reviewed and accepted by Korea National Institute for Bioethics Policy (P01-202204-01-011). The use of Korean National Health Information Database is reviewed and accepted (NHIS-2022-1-796). Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the Korean National Health Insurance Service.

JEL: I13, I11, C51

Keywords: Out-of-Pocket Costs; Elasticities of Medical Expenditures; Bunching

1 Introduction

The elasticity of medical expenditure with respect to out-of-pocket costs is one of the most important parameters for health systems. In particular, for universal healthcare systems, it is essential to predict changes in national budget spending resulting from policy changes. However, policymakers often overlook the elasticity of medical expenditure when making decisions. For instance, when the out-of-pocket system was changed in South Korea (Korea, hereafter), government agencies predicted the impact on medical expenditure based on the assumption that the distribution of medical expenses would remain the same as the existing system (e.g., Ministry of Health and Welfare, 2017; National Assembly Budget Office, 2016). However, this prediction is only valid if patients do not respond to price changes at all. If patients do respond to price changes, this could lead to highly inaccurate predictions.

This paper develops a novel method for estimating the distribution of the elasticity of medical expenditure with respect to out-of-pocket costs, using bunching responses to a discontinuity in out-of-pocket costs in Korea. In particular, the out-of-pocket costs for a clinic visit, which were applied until 2017, are different for those aged under 65 and those aged 65 or over. Patients aged 65 or over pay a copayment of 1.5K KRW when the total expenditure is 15K KRW or less, and pay a coinsurance rate of 30% when the total expenditure is more than 15K KRW.[1] This creates a *notch* in which the out-of-pocket payment jumps from 1.5K KRW to 4.5K KRW.[2] On the other hand, patients under the age of 65 pay a coinsurance of 30% regardless of their total expenditure (see Figure 1).[3]

[1]The average exchange rate for 1 United States dollar was 1,145 South Korean won (KRW) in 2021. To make it easier to compare monetary values to U.S. dollars, I will use the letter K to represent 1,000 for KRW.

[2]To clarify terminology, total expenditure refers to the sum of fees of all medical services provided to a patient at one visit, and out-of-pocket cost refers to the amount paid directly by the patient out of the total expenditure. In Korea, patients pay their out-of-pocket costs at the hospital reception desk. I focus on the visit-level costs because the notch applies to the total expenditure of each visit.

[3]There is no deductible, so these out-of-pocket systems are directly applied to any visits. There is an

This paper builds on the bunching estimation literature (see for review Kleven (2016)). In particular, Saez (2010), Kleven and Waseem (2013), and Einav et al. (2017) are closely related to this paper. Saez (2010) develops a theoretical model using a utility function with quasi-linear and isoelastic preferences to measure the bunching response to a kink. Kleven and Waseem (2013) extend the bunching method to use a notch where a price schedule discontinuously changes. Einav et al. (2017) modify Saez (2010)'s utility function so that the bunching estimation can be applied to the context of health insurance. In this paper, I adopt the utility function of Einav et al. (2017) and apply a strategy of Kleven and Waseem (2013) to characterize the optimization problem in a notch design.

The institutional setting in Korea has advantages in estimating elasticities compared to the existing bunching literature for the following reasons. First, since the prices of medical services in Korea are fixed within a calendar year in most cases, individuals turning 65 face different out-of-pocket costs in the same calendar year even when they receive the same services. Therefore, the distribution of medical expenditures for 64-year-olds can be used as a counterfactual distribution to that of 65-year-olds. Many of the bunching papers only observe the distribution in which bunching occurs, so the counterfactual distribution in the absence of a kink or a notch is estimated through a polynomial approximation. Second, the notch at 15K KRW has been maintained at a nominal level since 2001, while prices of medical services have increased every year. A price of a medical service is determined by multiplying the relative value unit of the service by a conversion factor of the year. Since the conversion factor increases every year, while the relative value units are unchanged for most cases, the distribution of medical expenditure shifts upward over time while maintaining a similar shape. This creates variation in the distribution of medical expenditures across years.

The key idea in this paper is that the ratio of two densities for the treatment and control groups reveals the conditional cumulative distribution of expenditure below a certain elasticity. This is due to the characteristic that the notch is fixed at 15K KRW. When the notch

annual maximum out-of-pocket cost. However, the out-of-pocket costs of outpatient treatment are too low for the upper limit to be applied, so the effects of the upper limit are not considered in this paper.

is introduced, a rational patient would compare the utility of initial health services with reducing health services to 15K KRW. If the patient bunches at the notch, it indicates that their marginal benefit of paying a lower cost exceeds that of getting more medical services at a given visit. Since the notch is 15K KRW, the amount a patient must give up for bunching varies by the initial medical expenditure they would have chosen without the notch. For example, a patient who would have chosen 20K KRW must give up medical services by 5K KRW in order to bunch. Likewise, a patient who would have chosen 25K KRW must give up medical services by 10K KRW for bunching. This suggests that the minimum elasticity for bunching from 25K KRW will be larger than that of 20K KRW. Thus, the density impacted by the notch comprises those who are less price-sensitive than the minimum elasticity for bunching and those who are price-sensitive but remain due to some friction. The density of the control group is constructed by the distribution of health needs with no restriction on elasticities. Therefore, the ratio of treated density and control density is closely related to the conditional cumulative distribution below the minimum elasticity at each amount of medical expenditure, given the probability of spending the amount in the absence of the notch.

The main contribution of this paper is to extend the bunching estimation method to estimate the elasticity distribution in the presence of a control group facing a linear price schedule. To the best of my knowledge, the existing literature has not actively used the ratio of treated density and counterfactual density, except for Hamilton (2018). Most papers in the bunching literature rely on a polynomial approximation to construct the counterfactual distribution (see, for example: 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) This is because they observe only the treated density. However, Blomquist et al. (2021) point out that one cannot identify if a bunching density is due to price sensitivity or simply if the underlying counterfactual density is high, when using a single treated density. In addition, the polynomial approximation mostly comes with the

assumption that the counterfactual distribution is locally constant to normalize the size of bunching by the density at the kink or notch. However, the local constancy assumption does not fit my setting in which the control distributions are not smooth (see Figure 2). I use the ratios of treated and control densities at each medical expenditure instead of a density at a single point. Moreover, I suggest a framework in which the density ratio is the conditional cumulative distribution of elasticity given medical expenditure. Hamilton (2018) uses density ratios to estimate the bunching window but does not interpret them as the conditional cumulative distribution.

To estimate the conditional cumulative distribution, I adopt the copula approach pioneered by Sklar (1973). Previous studies assume that the probability density function is locally constant at a kink or a notch to estimate the mean response, and then substitute it into the theoretical model to derive an estimate of the elasticity at the mean response. However, since I can observe control and treated densities thanks to my institutional setting, I can exploit information from two densities at each point without relying on the local constancy assumption. Moreover, I can observe multiple years of data in which the notch is fixed at a nominal level while prices of medical services increase every year. The variation in the cumulative distribution of medical expenditures is a key to estimating the dependence between elasticities and medical expenditures in the copula approach.

The other closely-related literature is to estimate the price elasticity of health care demand. Specifically, I find that medical expenditures per outpatient visit respond to a discontinuity in out-of-pocket costs in Korea. My findings are contrary to the finding of the RAND Health Insurance Experiment that cost-sharing affects the number of episodes but not the cost per episode (see, for example: Manning et al., 1987; Lohr et al., 1986, Keeler and Rolph, 1988; and Aron-Dine et al., 2013). In addition, another experimental study, the Oregon Health Insurance Experiment (e.g., Finkelstein et al., 2012), and empirical studies mostly focus on the extensive margin, looking into the probability of any medical use, the number of medical uses, or the total spending for a certain period (see, for example: Brot-

Goldberg et al., 2017; Ellis et al. (2017); Choi et al., 2010; and Choi, 2018). On the other hand, Duarte (2012) quantifies that two-thirds of the total elasticity is explained by the intensity of each visit in Chile. This paper focuses on the responses on the intensive margin, finding evidence of bunching in response to the discontinuity in out-of-pocket costs. Second, in my setting, the change in out-of-pocket costs that patients respond to is from 1.5K KRW to 4.5K KRW. Although the change is three-fold in proportion, it is still a small amount in absolute terms. Thus, this paper supplements previous studies that have examined price elasticity at relatively large amounts, such as total annual expenditures (e.g., Einav et al., 2017) and monthly expenditures (e.g., Ellis et al. (2017)). Third, there are studies that analyze the same system as my setting in Korea (see, for example, Kim and Kwon, 2010; Na, 2020; and Kim, 2021). In particular, Kim (2021) has the most similarity with my approach in that it uses the bunching estimation method. However, Kim (2021) uses a different data set, the Korea Health Panel Survey, and does not use the age-64 density as a counterfactual.

I use Korean health insurance administrative data in 2013-2017. There is evidence that patients with higher expenditure tend to be less elastic, with the estimated Kendall's rank correlation of -0.52. The upper bound of the elasticities is 0.17, the unconditional mean is 0.1, and the standard deviation is 0.07, assuming that elasticities are distributed as a beta distribution. Based on these estimates, counterfactual policy simulations show that a linear coinsurance rate of 21.3% instead of a notch can improve the welfare of patients and clinics without increasing the insurer's spending.

The paper proceeds as follows. In Section 2, I provide institutional background about cost-sharing systems in Korea and the analysis sample. Section 3 describes a theoretical model of bunching responses to a notch in the context of health insurance. Section 4 proposes a method for identifying the distribution of elasticities that utilizes the copula approach. In Section 5, I propose an estimation strategy. Section 6 presents the Monte Carlo simulation results and estimation results using the data of 2013-2017. In Section 7, I simulate policy counterfactuals using the estimates of elasticity distribution. Section 8 concludes.

2 Background and Data

2.1 Cost-sharing Systems in the Korean Health Insurance System

South Korea has a universal health insurance system operated by a single insurer, the National Health Insurance Service (NHIS). Patients directly pay a portion of the total medical expenditure for each visit, and health providers receive reimbursement for the remaining amount from the NHIS. And, there is no deductible in the cost-sharing system.

This study exploits a setting in the out-of-pocket system applied to outpatient treatments in clinic-level providers. The out-of-pocket costs for outpatient care at a clinic between 2007 and 2017 are different for ages 6–64 and ages 65 or older. For each visit, patients aged 6–64 pay 30% of the total expenditure. Those aged 65 or older pay a copayment of 1.5K KRW if the total expenditure is 15K KRW or below; they pay 30% of the total expenditure otherwise. Figure 1(a) shows the out-of-pocket cost function according to the total expenditure. The out-of-pocket cost function is linear for those aged 6–64, while it is notched at 15K KRW for those aged 65 or older.[4]

This results in different out-of-pocket costs for those turning 65 within one calendar year, even if they receive the same services. On the other hand, medical institutions earn the same revenue for the same services because the NHIS reimburses the rest regardless of how much the patient pays.

In this system, the notch point has been fixed at 15K KRW since 2001, while the prices of medical services have risen every year. Physicians have insisted on a change in the system. The Korean Medical Clinic Association (2014) issued a press release that when out-of-pocket payment increases from 1.5K to 4.5K KRW, "some patients treat doctors as thieves, throw money into the reception desk, and even go out yelling and paying only 1.5K KRW. Some doctors, tired of these complaints, provide free injections or physical therapy to keep the total expenditure below 15K KRW, reduce essential prescriptions, or reduce the amount of

[4]Although there exists a maximum out-of-pocket cap, I will ignore the effects of the cap in this study because the amount of coinsurance for outpatient treatments is too small to exceed the cap.

out-of-pocket payment even though they know it is illegal." This is an anecdote that shows that patients react sensitively to prices in reality.

Total expenditure is determined on a fee-for-service basis. Total expenditure is the sum of the prices of all services provided to a patient in a single visit: a basic fee for each visit and fees for additional services. The fee of a service is the product of the relative value score of the service and the conversion factor of the year. The relative value score is almost fixed over time, and the conversion factor is increased every year. The fee of each service is set by the government and is applied equally to all medical institutions across the nation.[5] Because patients and medical institutions cannot adjust the fees of items, an adjustment in the total expenditure can be made only by adding or subtracting items.

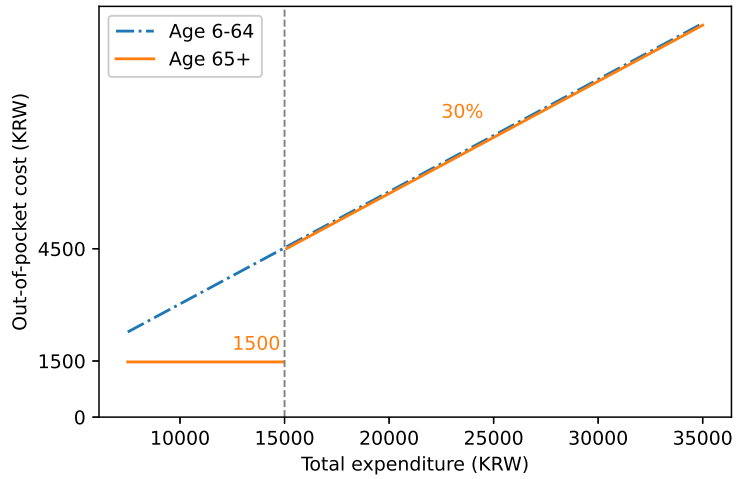
Table A.2 shows an example of physical therapy, which accounts for a large portion of outpatient visits costing around 15K KRW. The most common package of physical therapy consists of four items—visit fee, deep heat therapy, superficial heat therapy, and electrical nerve stimulation. The package cost less than 15K KRW until 2014. However, in 2015, the package cost becomes 15.21K KRW. Therefore, to adjust the total expenditure to the highest amount under 15K KRW, superficial heat therapy of 0.437K KRW needs to be excluded. In 2016, instead of superficial heat therapy, more expensive deep heat therapy needs to be excluded. In 2017, both deep heat therapy and superficial heat treatment therapy must be excluded. This example shows that the items excluded to make the total expenditure below 15K KRW can be changed across years. Also, it shows cases where the highest total expenditure under 15K KRW is not exactly 15K KRW because the fees of services cannot be adjusted by patients or medical institutions.

2.2 Data

I use the National Health Information Database of the NHIS. The data are collected by the NHIS for all residents in Korea. The variables I use are medical institutions, total medical

[5]There are no medical institutions outside the NHIS system; all institutions must be registered to operate.

Figure 1: Out-of-Pocket Costs per Visit



Note: The figure shows the out-of-pocket costs for those aged 6-64 (blue) and those aged 65 or older (orange) from 2007 to 2017.

expenses, out-of-pocket payments, main diagnosis, and age in the medical claims data.

The main analysis sample is individuals turning 65 years old in each calendar year between 2013 and 2017. I only use the medical claims of the last visit at age 64 and the first visit at age 65 for the same disease category. This has two implications. First, the age-64 distribution of total expenditures is used as a counterfactual distribution for 65-years-olds. It is based on the assumption that, within a calendar year, individuals who go to the clinic with the same disease have the same characteristics except for the change in out-of-pocket rules. Second, I exclude the responses at the extensive margin. 65-year-olds have an incentive to visit the clinic more frequently because they face lower out-of-pocket payments when they adjust the total expenditure to less than 15K KRW. If this is the case, the age-65 density below 15K KRW will be affected by both the intensive-margin and extensive-margin responses, making it difficult to compare with the age-64 distribution. Therefore, I focus on the intensive-margin responses by selecting the data just before and after the new out-of-pocket cost system is applied.

In the process of cleaning the medical claims data, claims that violate the out-of-pocket

cost formulas are excluded. These are simply due to a reporting error, or cases where other exceptions are applied. Medical Aid beneficiaries are excluded as they pay copayments.[6]

Table 1 shows the characteristics of the analysis sample. In Panel A, the numbers of bills are equal for age 64 and age 65 in the same year since I extract the data for the last visit at age 64 and the first visit at age 65. However, about 23% of bills are only for visiting a doctor without using additional services, and it is inappropriate to estimate price elasticities including those bills. So I use bills with at least one additional service for analysis. Panel B summarizes the bills with any additional services. This shows that small amounts of medical expenses account for a large proportion of clinic visits. For example, the proportions of medical expenses under 15K KRW are 49-71% in 2013–2017. This suggests that how individuals respond to the notch of 15K KRW is important in terms of health insurance spending.

Figure 2 displays the histogram of the total medical expenses for 64-year-olds and 65-year-olds in 2013–2018. The figure provides graphical evidence of bunching. In 2013–2017 when the notch exists at 15K KRW, the age-65 densities are more concentrated below the notch than the age-64 densities. In contrast, in 2018 when the notch is replaced with the kink at 15K KRW, the two densities become more alike around 15K KRW.[7] Also, the age-64 histograms show that the counterfactual densities are not locally constant around the notch and there are many peaks and pits.

The graphical evidence of bunching is more vivid in the plots of density ratios. Figure 3(a) shows the empirical cumulative distribution functions (CDFs) of total expenditures for age-64 by year. The empirical CDFs shift to the right each year because the conversion factors are raised every year while most of relative value units remain the same. Table A.1 shows the conversion factors in 2013-2017. Figures 3(b) and 3(c) plot the density ratios at each point of medical expenditures. In the figures, the closer to the notch, the more

[6]Medical Aid is a welfare system that assists medical expenses for low-income people, people with disabilities, and veterans. Medical Aid beneficiaries are about 3% of the total population.

[7]The out-of-pocket cost functions since 2018 are presented in Figure A.3. Years beyond 2017 are not directly considered in this study, but indirectly considered in the context of policy counterfactual analysis.

the proportion of patients bunching, and the density ratios gradually increase in medical expenditures and eventually converge to one.

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

Notes: The table shows statistics for the main analysis sample between 2013 and 2017. Panel A shows the number of medial bills. Bills without additional services are bills only with visit fee. Bills with additional services are bills with visit fee and fees for additional services. Panel B shows descriptive statistics of total expenditure for bills with additional services.

3 Model of Bunching Responses

In this section, I characterize the relationships between elasticities and medical expenditures using a structural model. In particular, through the structural model, I find the elasticity of the marginal buncher at each medical expenditure. This helps to link the observed densities of treated and control groups and the conditional cumulative distribution of elasticity given medical expenditure.

I assume a quasi-linear and constant elasticity utility function, which is widely used in the bunching literature (See, for example: Saez, 2010; and Kleven and Waseem, 2013). In particular, I use the utility function suggested by Einav et al. (2017), which adapts Saez (2010)'s utility function to a health insurance context. An advantage of using this type of utility function is that the relationship between the choice variable and the elasticity parameter is represented in a simple and intuitive form.

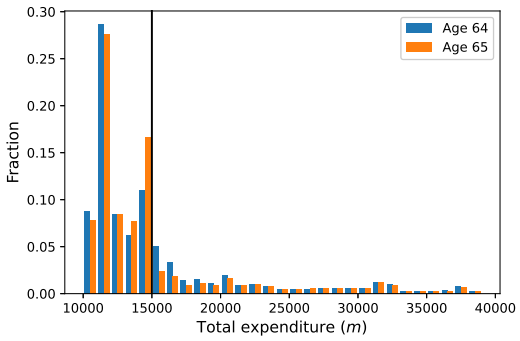
$$u\left(m;\zeta,\eta\right)=g\left(m\right)+c=\left[2m-\frac{\zeta}{1+\frac{1}{\eta}}\left(\frac{m}{\zeta}\right)^{1+\frac{1}{\eta}}\right]+\left[y-s\left(m\right)\right] \tag{1}$$

where m is the total expenditure for a visit, ζ is a parameter for health needs, η is a parameter for elasticity $\eta>0$, y is income, $s\left(m\right)$ is the patient's out-of-pocket cost given m, and c is the residual income after paying $s\left(m\right)$. By assuming a quasi-linear utility function, I ignore income effects on medical expenses. I allow for heterogeneity in ζ and η. The distribution of ζ is not necessarily smooth. When ζ and η are given, individuals make a choice for m.

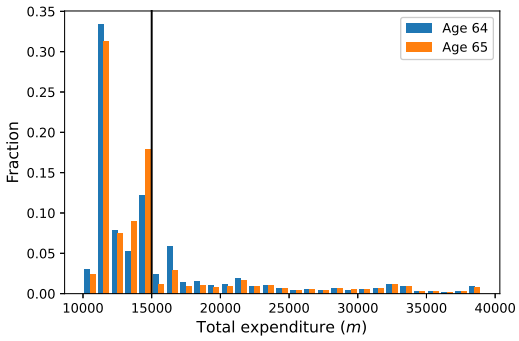
The choice variable m in this model has the following features. First, m is the total expenditure for a visit. This is because the institutional setting I use for estimation applies to the total expenditure per visit rather than total annual spending. Therefore, this model is static and focuses on responses on the intensive margin. Second, m includes only medical expenses observed on medical claims. Non-medical costs, such as opportunity costs and travel costs, are not considered in this paper. Third, m includes only services covered by the NHIS. Uncovered services are not of interest in this paper because they are not observed in

Figure 2: Histograms of Total Expenditures

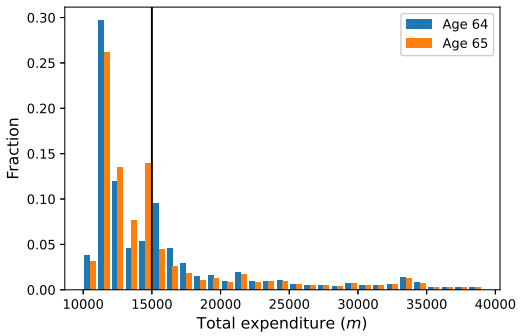
(a) 2013



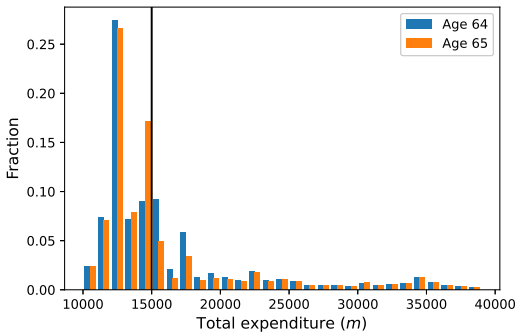
(b) 2014



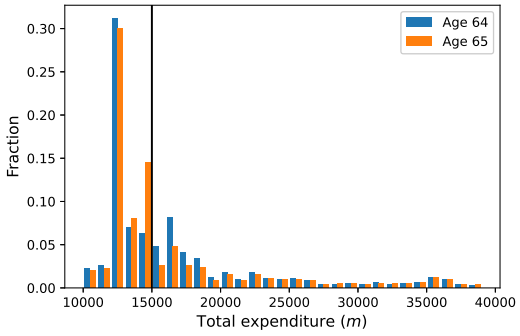
(c) 2015



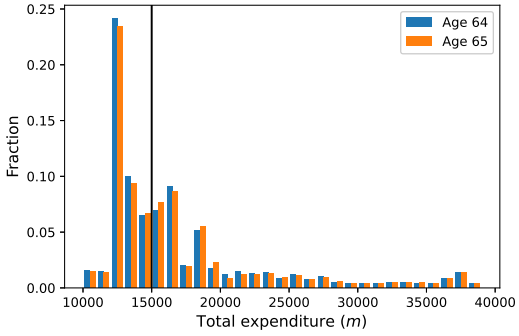
(d) 2016



(e) 2017



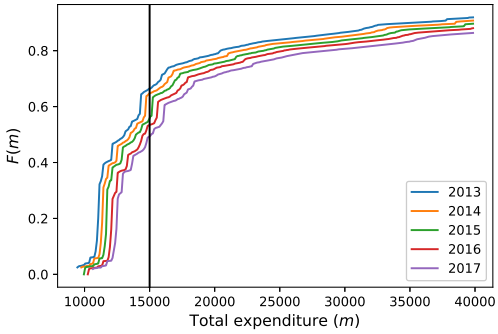
(f) 2018



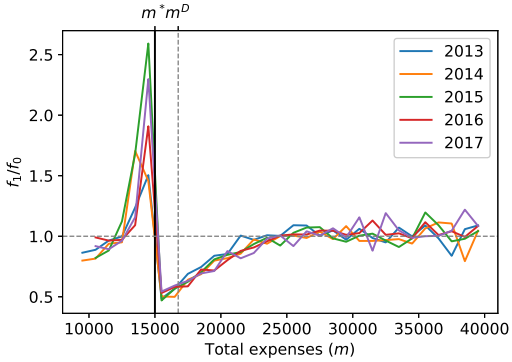
Notes: The figure shows histograms of total expenses for age 64 (blue) and age 65 (orange) between 2013 and 2018. The bin size is 1,000 KRW.

Figure 3: Estimates of Density Ratios between 2013 and 2017

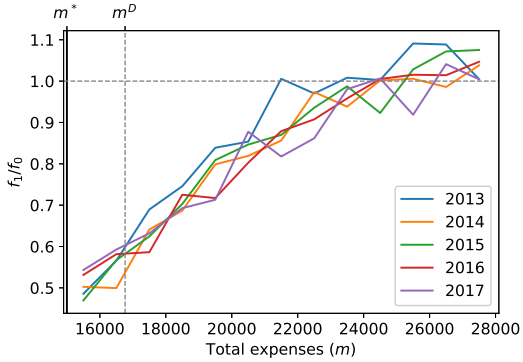
(a) $F_0^t(m)$



(b) f_1^t/f_0^t



(c) f_1^t/f_0^t above the Notch



Notes: The figure plots estimates of density ratios via histogram estimation between 2013 and 2017 with bin size of 1,000 KRW. m^* is the notch at 15K KRW. m^D is the upper bound of the dominated region. (a) shows the CDF of total expenditure for 64-year-olds. (b) shows density ratios for total expenditures between 9,000 and 40,000. (c) shows density ratios for total expenditures between 15,000 and 28,000.

the data and are paid 100% by patients.

To understand the characteristics of parameters ζ and η, I start with a linear out-of-pocket cost function $s(m) = sm$ where s is a coinsurance rate between 0 and 1. From the first-order condition for maximizing the patient's utility in equation (1), the optimal choice under a linear coinsurance is given by:

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

The individual type parameter ζ can be interpreted as total expenses the patient would

choose if there were no health insurance, i.e., $s = 1$. The elasticity parameter η is the elasticity of total expenses with respect to $(2 - s)$, $\eta = \partial \log m / \partial \log (2 - s)$. Since the range of possible coinsurance rate s is between 0 and 1, 2 is used as a normalization constant so that $2 - s$ ranges between 1 and 2. This elasticity parameter can be mapped into the elasticity of total expenses with respect to the coinsurance rate s, $\varepsilon \equiv |\partial \log m / \partial \log s| = \eta \times s / (2 - s)$.

Now, I consider the optimization problem with a notch in out-of-pocket costs. In particular, the function of out-of-pocket costs for age 65 has the following form:

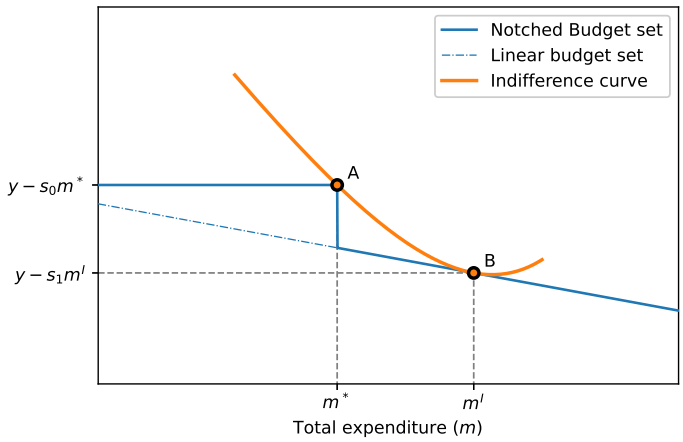
$$s(m) = 1\{m \leq m^*\} S_0 + 1\{m > m^*\} s_1 m \text{ where } S_0 < s_1 m^* \tag{3}$$

Patients pay a copayment S_0 when total expenses are lower than or equal to m^* and a coinsurance $s_1 m$ when total expenses exceed m^*. m^* is a notch point at which the patient's out-of-pocket costs discretely jump from S_0 to $s_1 m$. I define $s_0 \equiv S_0 / m^*$ as the "coinsurance rate" at the notch for convenience. This setting is a mix of copayment and coinsurance, but the identification is the same as the case of two coinsurance rates changing from s_0 to s_1 at the notch.

Suppose that starting with a linear out-of-pocket $s(m) = s_1 m$, a notch is introduced at m^* and the out-of-pocket costs are lowered below m^*. Then, price-sensitive patients who would have chosen total expenses above m^* have an incentive to choose m^* because the out-of-pocket costs are discretely reduced. The distribution of elasticities would determine how many patients initially above the notch would bunch at m^*.

To understand the distribution of elasticities, I characterize the marginal elasticity for patients initially choosing m^I to have the same utility level as m^*. The following conditions characterize the marginal elasticity. In the previous literature, this characterization is interpreted as the conditions for the "marginal buncher" (for example, Saez, 2010; Kleven and Waseem, 2013). By rearranging equation (2), I obtain an equation for the individual type choosing m^I under a linear coinsurance rate s_1: $\zeta^I = \frac{m^I}{(2 - s_1)^\eta}$. The utility level when the

Figure 4: Optimization Problem with a Notched Budget Set



Notes: The figure shows an optimization problem with a notched budget set. The vertical axis shows the residual income after paying out-of-pocket payments, $y - s(m)$, where y is individual income, and the out-of-pocket cost function is given by $s(m) = s_0 m^* \times 1\{m \leq m^*\} + s_1 m \times 1\{m > m^*\}$. The horizontal axis shows the total expenditure per visit. The solid orange line presents an indifference curve between residual income and total expenditure per visit. s_0 is the proportion of the out-of-pocket payment at m^*, and s_1 is the coinsurance rate when total expenditure exceeds m^*. The solid blue line is the budget set with a notch, and the dotted blue line is the counterfactual budget set in the absence of a notch.

patient of type ζ^I bunches at m^* and pay $s_0 m^*$ is written as follows:

$$u^* = (2 - s_0)\, m^* - \frac{m^I (2 - s_1)}{1 + 1/\eta} \left(\frac{m^*}{m^I}\right)^{1+\frac{1}{\eta}} + y. \tag{4}$$

The utility level when the patient of type ζ^I chooses an interior point m^I and pay $s_1 m^I$ is given by:

$$u^I = \frac{m^I (2 - s_1)}{1 + \eta} + y \tag{5}$$

Now, the condition that $u^* = u^I$ leads to the following equation:

$$(2 - s_0) \left(\frac{m^*}{m^I}\right) - \frac{2 - s_1}{1 + 1/\eta} \left(\frac{m^*}{m^I}\right)^{1+\frac{1}{\eta}} - \frac{2 - s_1}{1 + \eta} = 0 \tag{6}$$

Equation (6) is the condition in which a patient who would choose m^I under a linear coinsurance rate of s_1 is indifferent between bunching at m^* and choosing m^I. It is worth noting that equation (6) does not depend on unobservable individual type ζ^I, and s_0, s_1 and m^* are known policy-related values.

For given s_0, s_1 and m^*, let $\eta(m)$ denote η solving equation (6) when $m^I = m$. In other words, $\eta(m)$ is the minimum elasticity for those who would choose m under a linear coinsurance $s(m) = s_1 m$ to bunch at m^*. In addition, define $\epsilon(m)$, as $\epsilon(m) \equiv \eta(m) \times \frac{s}{2-s}$, so that $\epsilon(m)$ is the elasticity of medical expenditure with respect to s for the marginal buncher at m.

Proposition 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^*. (A proof is in Section A.1.)*

Proposition 1 is to show that each m can be mapped to a unique ϵ.

Proposition 2. *$\epsilon(m)$ is strictly increasing in m. (A proof is in Section A.2.)*

Propositions 1 and 2 together show that for any ϵ there exists a unique m that satisfies $\epsilon(m) = \epsilon$ for $m > \left(\frac{2-s_0}{2-s_1}\right) m^*$.

If a patient initially chooses m but has an elasticity ϵ higher than $\epsilon(m)$, they will be willing to bunch at m^* when the notch is introduced, because the marginal costs from higher out-of-pocket costs outweigh the marginal benefits from receiving more medical services. This implies that the larger m is, the higher $\epsilon(m)$ is. When m gets larger, a higher elasticity is required for patients initially at m to bunch as they must give up more medical services, which is $m - m^*$.

In a notch design, there exists "a dominated region" just above the notch that is not rationalized by any elasticity value (Kleven and Waseem, 2013). Let $\left(m^*, m^D\right]$ indicate the dominated region. The upper bound of the dominated region m^D is the value of m^I solving

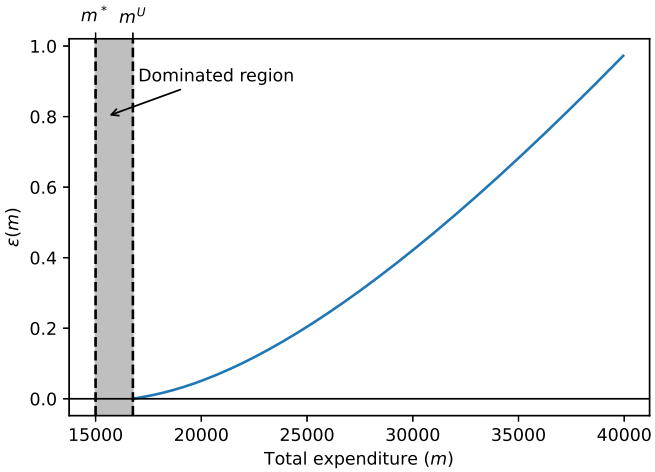
equation (6) as η approaches to zero:

$$m^D = \left(\frac{2 - s_0}{2 - s_1}\right) m^* \tag{7}$$

When a notch exists, the costs increase discretely whereas the utility level changes smoothly at the notch. Therefore, patients located slightly above the notch can always increase their utility by lowering the total medical expenses to m^* regardless of their elasticities. In other words, even a perfectly inelastic patient in the dominated region can be better off if they bunch.

Figure (5) illustrates the marginal buncher's elasticity function $\epsilon(m)$ along the m-axis for the case where $s_0 = 0.1$, $s_1 = 0.3$, and $m^* = 15,000$. Those values of the policy-related variables imply an upper bound of dominated region at $m^D \approx 16,765$. The figure also shows that $\epsilon(m)$ is monotonically increasing in m starting from 0 at $m = m^D$.

Figure 5: Function of Marginal Elasticity for Bunching, $\epsilon(m)$



Notes: The figure shows the function of marginal elasticity for bunching for the case where $s_0 = 0.1$, $s_1 = 0.3$, and $m^* = 15,000$. $\epsilon(m) = \eta(m) \times s_1/(2 - s_1)$ where $\eta(m)$ solves equation (6) for each $m^I = m$ given s_0, s_1 and m^*. The dominated region is where no solution for equation (6) exists. m^D is the upper bound of the dominated region, $m^D \approx 16,765$. m is in KRW.

4 Identification of the Elasticity Distribution

In this section, I suggest a new strategy to identify the distribution of elasticities ϵ, as a function of observable variables.

I begin with a frictionless model for simplicity. Let $f_0(\epsilon, m)$ denote the counterfactual distribution of (ϵ, m) given a linear out-of-pocket cost function $s(m) = s_1 m$, and $f_1(\epsilon, m)$ denote the distribution given an out-of-pocket costs function with a notch defined in equation (3). Let $f(\epsilon|m)$ and $f(\epsilon)$ respectively denote the conditional and unconditional probability density function of ϵ. And, $F(\epsilon|m)$ denotes the conditional cumulative probability of ϵ given m. And as defined earlier, $\epsilon(m)$ is the minimum elasticity at m for bunching given s_0, s_1, and m^*. $f_0(m)$ is the integral of $f_0(\epsilon, m)$ over all elasticities at m. $f_1(m)$ is the density of those who remain at m even when the notch is introduced. In other words, those who are included in $f_1(m)$ have elasticity lower than $\epsilon(m)$. The difference between $f_0(m)$ and $f_1(m)$ is the density of those moving to the notch from m. This leads to the following equation:

$$\begin{aligned} f_0(m) - f_1(m) &= \int_{\epsilon \geq \epsilon(m)} f_0(\epsilon, m)\, d\epsilon \\ &= \int_{\epsilon \geq \epsilon(m)} f_0(m)\, f_{\epsilon|m}(\epsilon|m)\, d\epsilon \\ &= f_0(m) \left[1 - F_{\epsilon|m}(\epsilon(m)|m)\right] \\ \Rightarrow \frac{f_1(m)}{f_0(m)} &= F_{\epsilon|m}(\epsilon(m)|m) \end{aligned} \tag{8}$$

Equation (8) indicates that the ratio of f_1 and f_0 at m is the probability of ϵ being below $\epsilon(m)$ conditional on m in a frictionless world. If m and ϵ are independently distributed, (8) becomes $F_\epsilon(\epsilon(m)) = \frac{f_1(m)}{f_0(m)}$. Thus, F_ϵ is simply identified by f_1/f_0, 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 $f_1(m)$ and $f_0(m)$ are observed.

Now I consider the case where there are optimization frictions. Since a linear coinsurance

system is applied to individuals until the age of 64 and a notch out-of-pocket system is introduced from the age of 65, I consider frictions as cases where individuals do not or cannot reoptimize after age 65 and stick to their previous decisions. Therefore, $f_0(m) - f_1(m)$, the realized mass of bunching from m, reflects the mass of individuals who have an elasticity greater than $\epsilon(m)$ and do not have friction. Let $\phi(\epsilon, m)$ denote the fraction of individuals with optimization friction at (ϵ, m).

$$f_0(m) - f_1(m) = \int_{\epsilon \geq \epsilon(m)} f_0(\epsilon, m)\,(1 - \phi(\epsilon, m))\,d\epsilon \tag{9}$$

Assumption 3. *$\phi(\epsilon, m)$ is constant on $\left(m^*, m^U\right)$, i.e. $\bar{\phi} \equiv \phi(\epsilon, m)$.*

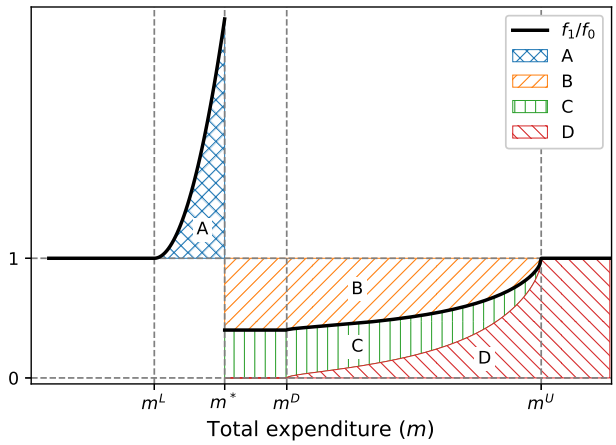
Assumption 3 is the same assumption as in Kleven and Waseem (2013). With the constancy of ϕ, ϕ is separated out of the integral in equation (9). If ϕ is increasing in m, then assuming the constancy of ϕ will result in the underestimation of elasticities.

$$\begin{aligned} f_0(m) - f_1(m) &= \int_{\epsilon \geq \epsilon(m)} f_0(m)\, f_{\epsilon|m}(\epsilon|m)\left(1 - \bar{\phi}\right) d\epsilon \\ &= f_0(m)\left(1 - \bar{\phi}\right)\left[1 - F_{\epsilon|m}(\epsilon(m)|m)\right] \\ \Rightarrow \frac{f_1(m)}{f_0(m)} &= \underbrace{F_{\epsilon|m}(\epsilon(m)|m)}_{\text{Inelastic}} + \underbrace{\bar{\phi}\left[1 - F_{\epsilon|m}(\epsilon(m)|m)\right]}_{\text{Elastic, but with friction}} \end{aligned} \tag{10}$$

Equation (10) decomposes the density ratio at m. The first term is the conditional probability of those who are less elastic than $\epsilon(m)$ given m. The second term is the probability of those who have higher elasticity than $\epsilon(m)$ but have a friction. Figure 6 illustrates what the density ratio is composed of.

Following Kleven and Waseem (2013), I identify the fraction of friction $\bar{\phi}$ using the dominated region in equation (7). If everyone re-optimized in response to the notch when turning 65 years old, no one would be in the dominated region $\left(m^*, m^D\right]$, because even a perfectly in-

Figure 6: Decomposition of Ratios of $f_1(m)$ and $f_0(m)$



Notes: The figure illustrates the ratio of f_1 and f_0 following equation (10). Area A represents the result of bunching. Area B represents the conditional probability of bunching given m. Area C represents the conditional probability of having friction despite of willingness to bunch given m. Area D represents the conditional probability of having smaller elasticities than $\epsilon(m)$ given m.

elastic patient can be better off when bunching from the dominated region. However, if there are patients still remaining in the dominated region at age 65, it is considered due to a friction. Thus, for each $m \in \left(m^*, m^D\right]$, the fraction of friction at m is defined as $\phi(m) = \frac{f_1}{f_0}(m)$. If it is further assumed that $\phi(m)$ is constant over $\left(m^*, m^U\right)$, $\bar{\phi}$ is approximated as the ratio of densities in the dominated region as in Kleven and Waseem (2013).

$$\bar{\phi} = \frac{\int_{m^*}^{m^D} f_1(m)\, dm}{\int_{m^*}^{m^D} f_0(m)\, dm} \tag{11}$$

If ϵ and m are independent, $F(\epsilon)$ is a function of observed variables.

$$F_\epsilon(\epsilon(m)) = 1 - \frac{1}{1-\bar{\phi}}\left[1 - \frac{f_1(m)}{f_0(m)}\right] \tag{12}$$

Since $\bar{\phi}$ is identified using equation (11), and f_1 and f_0 are observed, $F_\epsilon(\epsilon(m))$ is identified.

But the independence of ϵ and m is a strong assumption.[8] To relax the independence

[8]Nonetheless, the independence of ϵ and m is a weaker assumption than the local constancy of f_0, as it

assumption, I adopt the copula approach pioneered by Sklar (1973). The copula approach parameterizes the joint distribution as a function of marginal distributions with dependence parameters. In this setting, the copula approach is appropriate in that the conditional CDF is observed through the ratio of f_1 and f_0, and one of marginal CDFs $F_0(m)$ is observed. So, when the copula function is well defined, the other marginal CDF, $F(\epsilon)$ can be identified. The superscript t is attached to variables in year t hereinafter.

Assumption 4. *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.*

Note. The conditional CDF of ϵ given m is the partial derivative of the copula with respect to the second argument, $F_0(m)$.[9] Let $h(\cdot,\cdot)$ denote the partial derivative of $C(\cdot,\cdot)$ with respect to the second argument, $h \equiv C_2$.

In this paper, I will use four copula families: Clayton (1978), Gumbel (1960), Frank (1978), and the Gaussian. The four copulas are widely used, and their properties are well established. Table A.3 summarizes selected properties relevant to this paper.

Assumption 5. *The marginal CDF of ϵ is stationary. $F_\epsilon^t(\epsilon) = F_\epsilon(\epsilon)$ for all year t.*

Assumption 5 indicates that the distribution of elasticities is stable regardless of time because it is an individual type. Using assumptions 4 and 5, equation (8) can be written as the following equation.

$$\frac{f_1^t(m)}{f_0^t(m)} = h\left(F_\epsilon(\epsilon(m)), F_0^t(m); \theta\right) \tag{13}$$

If the dependence between ϵ and m is allowed, the ratio of f_1^t and f_0^t is represented as a function of the marginal CDFs $F_\epsilon(\epsilon(m))$ and $F_0^t(m)$ with a copula with dependence parameter θ.

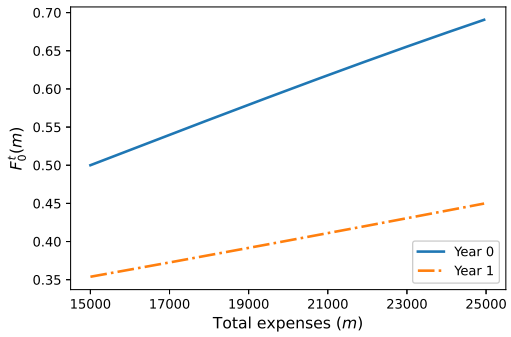
---

assumes $f_0(\epsilon, m) = f_0(m^*)$ for $m \in \left(m^*, m^U\right)$.

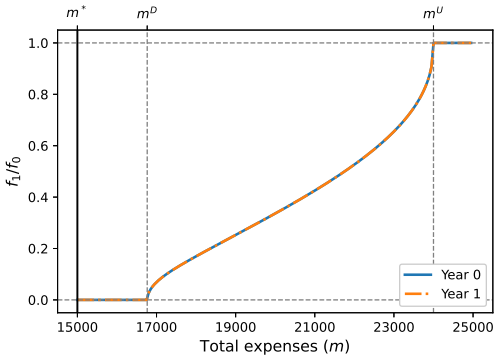
[9] $\frac{\partial}{\partial F_0(m)} C\left(F(\epsilon), F_0(m); \theta\right) = \frac{\partial F(\epsilon,m)/\partial m}{f_0(m)} = F(\epsilon|m)$

Figure 7: Hypothetical Density Ratios by Existence of Frictions and Dependence between ϵ and m

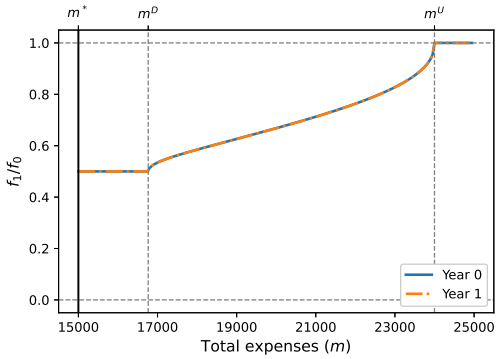
(a) $F^t(m)$



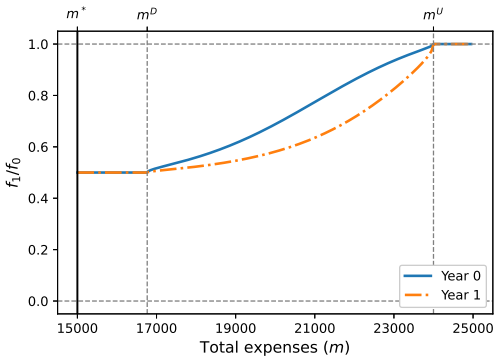
(b) $\bar{\phi} = 0, \tau = 0$



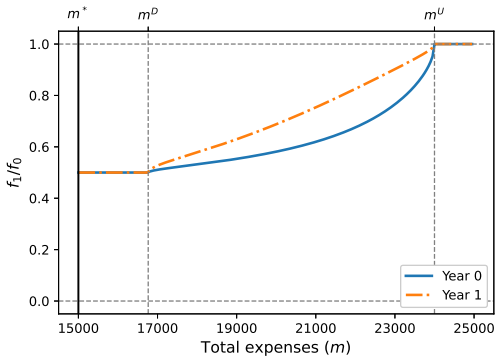
(c) $\bar{\phi} = 0.5, \tau = 0$



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



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



Notes: These figures show hypothetical density ratios by year, existence of frictions, and dependence between η and m. $m^* = 15,000$, m^D is the upper bound of the dominated region, $m^D \approx 16,765$, and $m^U = 24,000$. (a) plots two hypothetical CDFs of m for year 0 and year 1 in which the CDF of m for year 1 is below the CDF of m for year 0 for each m. (b) is when there is no friction and η and m are independent. (c) is when there is friction and η and m are independent. (d) is when there is friction and η and m are negatively dependent. (d) is when there is friction and η and m are positively dependent.

Figure 7 illustrates hypothetical density ratios f_1/f_0 by existence of frictions and dependence between ϵ and m. For year 0 and 1, I assume that the CDF of m for year 1 is below that of year 0. This is when the conversion factor of medical services increases over time. Figure 7(b) is when there is no friction and ϵ and m are independent. In this case, the density ratios in the dominated region must be zero, and the density ratios must be identical across years at each m. Figure 7(c) shows when there is friction, and ϵ and m are still independent. Due to frictions, the density ratios in the dominated region are strictly greater than zero. Figure 7(d) and Figure 7(e) are when there is friction, and ϵ and m are dependent. If ϵ and m are negatively dependent, for $F^0(m) > F^1(m)$, $h\left(F_\epsilon(\epsilon), F_0^0(m)\right) > h\left(F_\epsilon(\epsilon), F_0^1(m)\right)$; if ϵ and m are positively dependent, vice versa.

Assumption 6. *ϵ is distributed as Beta with parameters (α, β) on a support $\left(0, \epsilon^U\right)$ where $\epsilon^U = \epsilon\left(m^U\right)$.*

Assumption 6 parameterizes the distribution of ϵ. The Beta distribution is chosen because it has a bounded support and has two parameters, which allows a flexible estimation of the mean and the variance. Let Ω be the set of parameters $\left(\bar{\phi}, \epsilon^U, \alpha, \beta, \theta\right)$. Then, the ratio of f_1 and f_0 at m is represented as a function of m and $F_0^t(m)$ given parameters Ω.

$$\frac{f_1^t(m)}{f_0^t(m)} = \underbrace{1 - \left(1 - \bar{\phi}\right)\left\{1 - h\left[F_\epsilon\left(\epsilon(m); \alpha, \beta, \epsilon^U\right), F_0^t(m); \theta\right]\right\}}_{\equiv R\left(m; F_0^t, \Omega\right)} \tag{14}$$

Previous research has estimated the elasticity at the mean response of the choice variable (for example, Saez, 2010; Chetty et al., 2011; Kleven and Waseem, 2013). The approximation of the mean response defined by Kleven and Waseem (2013) can be written in the following

way using my notation:

$$\mathbb{E}\left(\Delta m\right) \approx \underbrace{\int_{(m^*, m^U)} \left[f_0\left(m\right) - f_1\left(m\right)\right] dm}_{B} \frac{1}{f_0\left(m^*\right)} \left(\frac{1}{1-\bar{\phi}}\right) \tag{15}$$

where Δm denotes the change in m in response to a notch. Unlike the strategy in this paper, they use the *bunching mass* B which is the integral of $f_0 - f_1$ over the whole bunching segment $\left(m^*, m^U\right)$, while I start with $f_0\left(m\right) - f_1\left(m\right)$ at each m in equation (14). And they approximate $f_0\left(m, \eta\right)$ by a single value of $f_0\left(m^*\right)$ whereas I use $f_0\left(m\right)$ for each m. There are two key identifying assumptions in the previous works: 1) the fraction of frictions ϕ is locally constant on $\left(m^*, m^U\right)$ (Kleven and Waseem, 2013), and 2) the counterfactual density f_0 is locally constant on $\left(m^*, m^U\right)$ (Saez, 2010; Chetty et al., 2011; Kleven and Waseem, 2013). Their final estimate for the elasticity parameter is the value of η that solves equation (6) with plugging $\mathbb{E}\left(\Delta m\right)$ into m^I.

In this study, I assume 1) $\phi\left(\epsilon, m\right)$ is locally constant on $\left(m^*, m^U\right)$ as in Kleven and Waseem (2013), and 2) $F\left(\epsilon, m\right) = C\left(F_\epsilon\left(\epsilon\right), F\left(m\right); \theta\right)$ where C is a copula function with dependence θ. The assumption that $f_0\left(\epsilon, m\right)$ is locally constant and well approximated as $f_0\left(\epsilon, m\right) = f_0\left(m^*\right)$ for $m \in \left(m^*, m^U\right)$, is much stronger than the assumption that ϵ and m are independent.

5 Estimation Strategy

To estimate the distribution of elasticities, I use equation (14). I first obtain estimates of f_0, f_1, and F_0 for a discretized set of m, and derive $\bar{\phi}$ and m^U using estimates of f_1 and f_0. Then, I replace f_0, f_1, $\bar{\phi}$, and m^U with the estimates in equation (14). Lastly, using least squares (α, β, θ) are estimated.

Densities $f_0^t(m)$, $f_1^t(m)$, and $F_0^t(m)$ To estimate f_0 and f_1, I use data on total medical expenses per visit of 64-years-olds and 65-years-olds respectively. The age-65 density is the treated density under the out-of-pocket cost function with a notch. The age-64 density is regarded as a counterfactual density in the absence of the notch.

To estimate the densities, I use histogram estimates. I use the same estimation method to f_0^t and f_1^t for every t, so I will omit the subscripts and superscripts below for simplicity. The histogram bins are defined to separate the variable below and above the notch. I use $b = 100$ to make a finely spaced histogram. The smallest unit of total expenditure is 10 KRW.

Given the notch m^*, define bins as $\mathcal{B}_j = (m^* + (j-1)\,b, m^* + jb]$ for $j = \ldots, -1, 0, 1, \ldots$ so that $\mathcal{B}_0 = (m^* - b, m^*]$ contains points just below the notch and $\mathcal{B}_1 = (m^*, m^* + b]$ contains points just above the notch. In this way, no bins contain points both below and above the notch. Let M_j denote the midpoint of bin $\mathcal{B}_j$, which is $M_j = m^* + \left(j - \frac{1}{2}\right) b$. For each bin $\mathcal{B}_j$, calculate the weighted histogram:

$$\hat{f}(M_j) = \frac{1}{Nb} \sum_{i=1}^{N} w_i 1\{m_i \in \mathcal{B}_j\} \tag{16}$$

where w_i is a weight of the observation i with $\sum_{i=1}^{N} w_i = 1$. I reweigh the analysis sample so that the disease composition is the same across time to compare results. The histogram estimator is a consistent estimator of $f(M_j)$ with asymptotic variance of $\mathbb{V}\left[\hat{f}(M_j)\right] = \frac{f(M_j)\{1 - bf(M_j)\}}{Nb}$.

The CDF of $m \leq M_j$ is defined using $\hat{f}(M_j)$. Since $\sum_{k \leq j} b\hat{f}_g^t(M_k)$ is the CDF of $m \leq M_j + \frac{b}{2}$, where $M_j + \frac{b}{2}$ is the upper bound of the bin $\mathcal{B}_j$, $\hat{F}_0^t(M_j)$ is estimated as the midpoint as follows.

$$\hat{F}_0^t(M_j) = \sum_{k \leq j} b\hat{f}_g^t(M_k) - \frac{b}{2}\hat{f}_g^t(M_j) \tag{17}$$

Fraction of friction $\bar{\phi}$ The fraction of optimization friction $\bar{\phi}$ is estimated under assumptions 1) observations in the dominated region $\left(m^*, m^D\right]$ are all due to optimization friction and 2) $\phi(\cdot)$ is locally constant. I estimate $\hat{\bar{\phi}}$ as the mean of the density ratio over the dominated region as follows:

$$\hat{\bar{\phi}} = \frac{\sum_t \sum_j \hat{f}_1^t(M_j) \, 1\left\{M_j \in \left(m^*, m^D\right]\right\}}{\sum_t \sum_j \hat{f}_0^t(M_j) \, 1\left\{M_j \in (m^*, m^D]\right\}} \tag{18}$$

where the upper bound m^D is calculated by equation (7).

The upper bound of bunching m^U If m^U is the upper bound of the bunching window, $F_\eta(\eta(m)) = 1$ for all $m \geq m^U$, and $f_1(m) = f_0(m)$ for all $m \geq m^U$ by equation (14). However, since there is noise in the estimated density function, I use a strategy to compare the smoothed densities with a bandwidth of h. Theoretically, since f_1 and f_0 are the same above m^U, the integrals of each density on $\left[m^U, m^U + h\right]$ should also be the same. Using this property, the minimum value of M_j at which the sum of $\hat{f}_1$ on $[M_j, M_j + h]$ becomes greater than or equal to that of $\hat{f}_0$ is regarded as the upper bound of the bunching window.

$$\hat{m}^U = \min\left\{ M_j - \frac{b}{2} : \sum_t \sum_{M_k \in [M_j, M_j + h]} \hat{f}_1^t(M_k) \geq \sum_t \sum_{M_k \in [M_j, M_j + h]} \hat{f}_0^t(M_k) \right\} \tag{19}$$

Distribution of elasticities $F(\epsilon)$ **and dependence parameter** θ The last step is to estimate the cumulative distribution function of elasticity ϵ and dependence parameter θ. I impose parametric assumptions on the distribution of η because η is not directly observed. I assume a beta distribution, since the distribution of η has a bounded support on the interval $\left(0, \epsilon^U\right)$ where $\epsilon^U = \epsilon\left(\hat{m^U}\right)$. And the beta distribution has two parameters (α, β), which allows flexible estimation of the mean and variance.

I also use four popular families of copula: Clayton, Gumbel, Frank and Gaussian. Table A.3 shows properties of four selected copulas. Those four copulas have different tail dependence and the range of dependence. The Clayton copula has lower tail dependence but zero

upper tail dependence while the Gumbel copula has upper tail dependence but zero lower tail dependence. Frank and Gaussian have zero tail dependence. The Clayton and Frank copulas do not have a negative dependence, so I use rotated Clayton and Gumbel by 90 and 270 degrees to estimate a negative dependence.

This is to estimate the parameters that best fit equation (14). By using equation (14), I construct a set of moment conditions.

$$\begin{aligned} \hat{g}^t\left(M_j;\Omega\right) &= \hat{f}_1^t\left(M_j\right) - \hat{f}_0^t\left(M_j\right)\hat{R}\left(M_j;\hat{F}_0^t,\Omega\right) \\ \text{where } \hat{R}^t\left(M_j;\Omega\right) &= 1-\left(1-\hat{\bar{\phi}}\right)\left\{1-h\left[F_\epsilon\left(\epsilon\left(M_j\right);\alpha,\beta,\hat{\epsilon^U}\right),\hat{F}_0^t\left(M_j\right);\theta\right]\right\} \end{aligned} \tag{20}$$

Assumption 7. $\mathbb{E}\left[\hat{g}^t\left(M_j;\Omega\right)\right]=0$, $\mathbb{V}\left[\hat{g}^t\left(M_j;\Omega\right)\right]=\mathbb{V}\left[\hat{f}_1^t\left(M_j\right)\right]$

α, β and θ are estimated via least squares.

$$\left(\hat{\alpha},\hat{\beta},\hat{\theta}\right)=\underset{\alpha,\beta,\theta}{\arg\min}\sum_{j,t}W_{j,t}\left[\hat{g}^t\left(M_j;\Omega\right)\right]^2 \tag{21}$$

where $W_{j,t}$ is the inverse of the variance of $\hat{f}_1^t\left(M_j\right)$.

6 Results

This section reports the results of Monte Carlo simulations to validate the estimation method and the results of applying the estimation method to the observed data. When showing the results, I will convert the parameters to more intuitive and comparable values. First, the structural elasticity parameter η will be converted to ϵ, where $\varepsilon = \eta \times \frac{s}{2-s}$ and s is a coinsurance rate. η is less intuitive in that it is the elasticity of medical expenditures with respect to $(2-s)$, while ϵ is the elasticity of medical expenditures with respect to coinsurance rates. Also, instead of parameters of the beta distribution, α and β, I will report the mean and standard deviation of ϵ. Lastly, since the dependence parameter of a copula, θ, is not

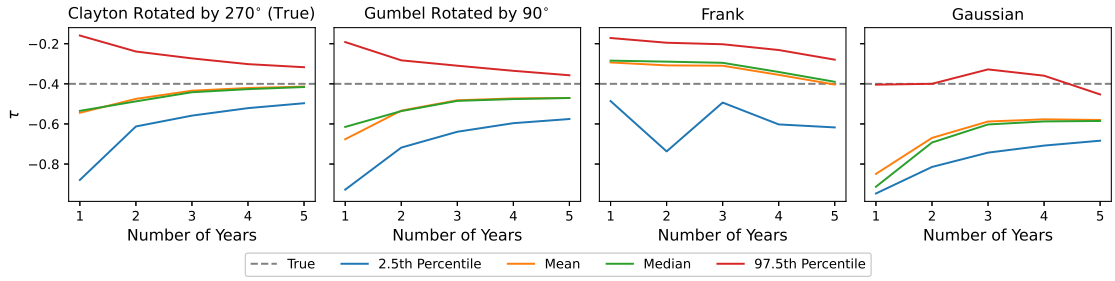
comparable across different copulas, I will report the Kendall's τ, a type of rank correlation ranging between -1 and 1.

6.1 Monte Carlo Simulation

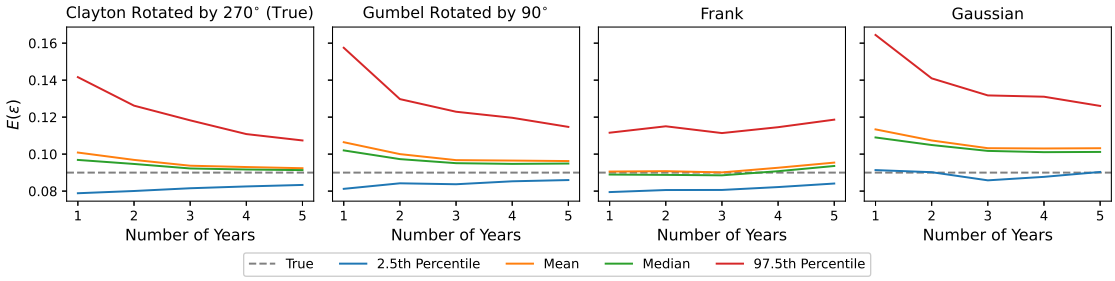
To learn about the behavior of my estimator, I perform a Monte Carlo simulation. The purpose of the simulation is to see how this methodology performs according to the number of years, and how the performance changes when the copula model is misspecified. The parameters to be used for validation are $\mathbb{E}(\epsilon)$, $\sigma(\epsilon)$, and τ. I will run two versions of the simulation that differ only in the copula family. In both versions, ϵ follows a beta distribution with $\mathbb{E}(\epsilon) = 0.09$ and $\sigma(\epsilon) = 0.055$, with the upper bound $\epsilon^U \approx 0.167$, which corresponds to the marginal buncher at $m^U = 24,000$. The distribution of m is the observed age-64 distribution of total expenditure. I set the Kendall's τ between ϵ and m to -0.5, and the fraction of patients with frictions $\bar{\phi}$ to 0.5 to make them similar to the estimates from my data. Six copula functions are used for estimation: Clayton 90, Clayton 270, Gumbel 90, Gumbel 270, Frank, and Gaussian. The bandwidth used for estimating the upper bound of bunching is 1,000. To find the distribution of the estimates, I repeat the same estimation for 1,000 replicates of the simulation sample.

The first simulation sample is generated from the Clayton copula rotated by 270 degrees. Figure 8 shows the distribution of estimates using 1,000 replicates. The simulation shows that the 95% confidence intervals get tighter as the number of years increases, regardless of the choice of the copula model. Clayton 270, the true copula model, has the shortest confidence intervals for all parameters, and the estimates converge to the true values as the number of years increases. Gumbel 90, which has a similar tail dependence to Clayton 270, exhibits similar performance to Clayton 270. Clayton 270 has the smallest mean squared errors for 950 out of 1,000 replicates. Therefore, in this case, even if the true copula family is unknown, the best model can often be found by selecting the model with the smallest mean squared errors among various alternative copulas.

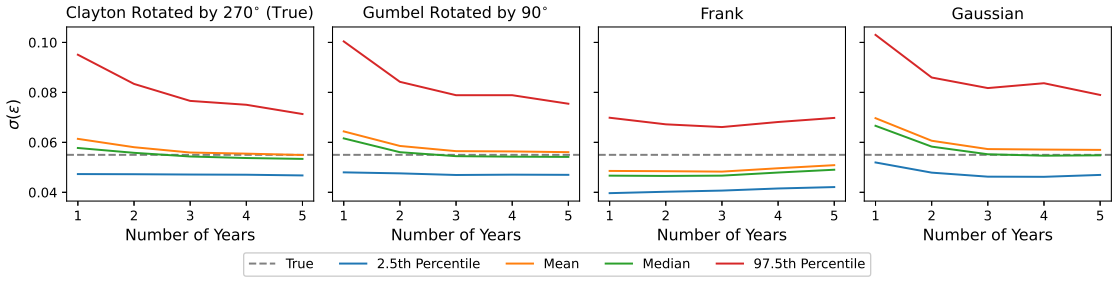
Figure 8: Monte Carlo Simulation of Parameter Estimates by Number of Years and Copula



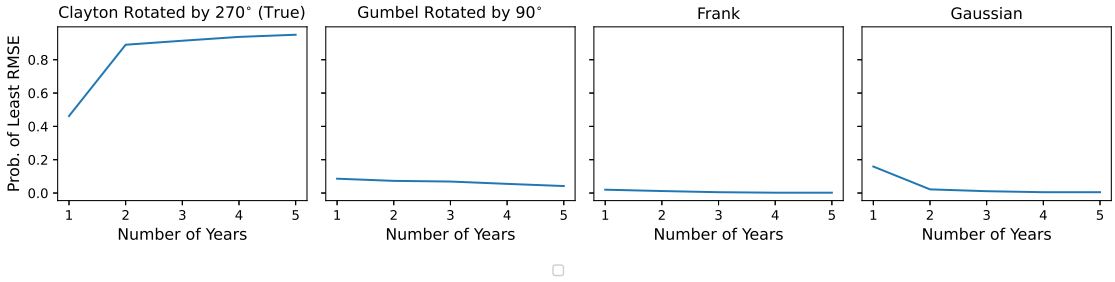
(a) τ



(b) $\mathbb{E}(\epsilon)$



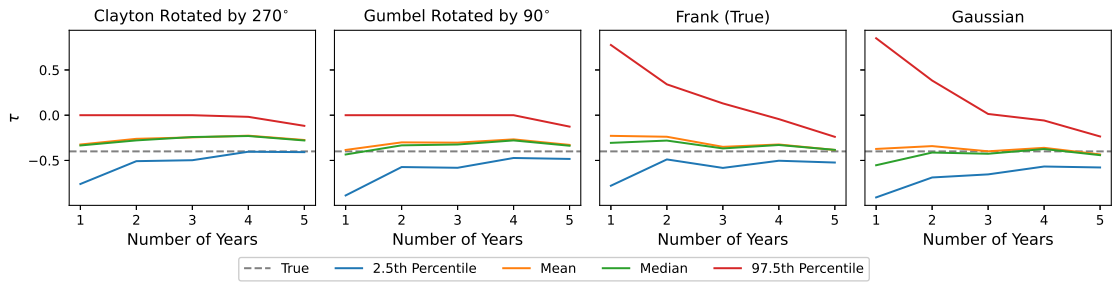
(c) $\sigma(\epsilon)$



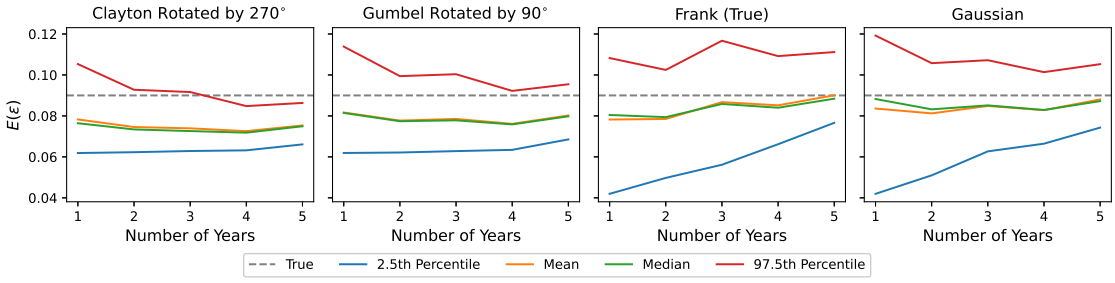
(d) Probability of Having Least Mean Squared Errors

Notes: The figures illustrate the Monte Carlo simulation results of estimation of τ, $\mathbb{E}(\epsilon)$, and $\sigma(\epsilon)$ by choices of number of years and copula family. True sample is generated by Clayton copula rotated by 270 degrees with $\tau = -0.5$. 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.

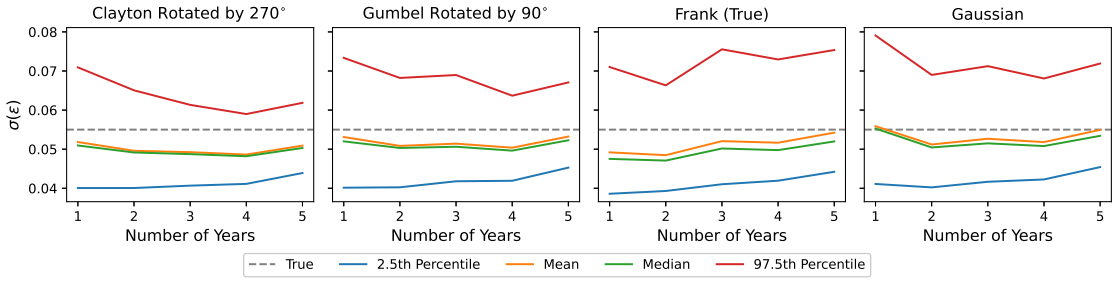
Figure 9: Monte Carlo Simulation of Parameter Estimates by Number of Years and Copula



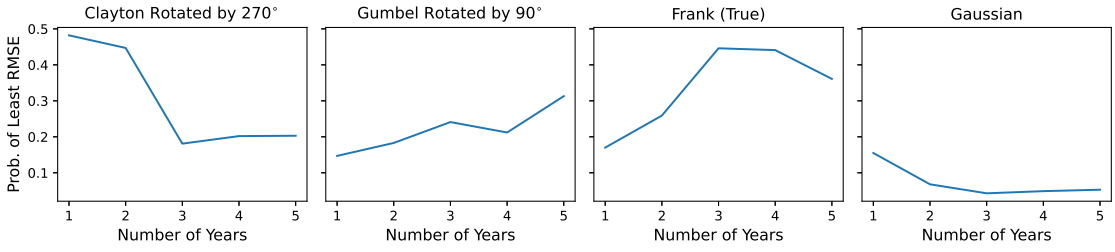
(a) τ



(b) $\mathbb{E}(\epsilon)$



(c) $\sigma(\epsilon)$



(d) Probability of Having Least Mean Squared Errors

Notes: The figures illustrate the Monte Carlo simulation results of estimation of τ, $\mathbb{E}(\epsilon)$, and $\sigma(\epsilon)$ by choices of number of years and copula family. True sample is generated by Frank copula with $\tau = -0.5$, $\mathbb{E}(\epsilon) = 0.09$, and $\sigma(\epsilon) = 0.055$, and the sample size is 200,000 for each year and age.

The second simulation sample is generated from the Frank copula of ϵ and m. Unlike the first sample, the probability of having the smallest MSE is highest for the Clayton 270 when one or two years of data are used. When three or more years are used, the Frank is most likely to have the smallest MSE, but the probability is only 30-40%, which is not as dominant as the first sample. This results from the fact that Frank is unstable in the estimation of τ when fewer years are used because the Frank allows τ to range from -1 to 1, whereas the Clayton 270 restricts τ to negative numbers. When three or more years are used, the Frank gets more accurate in estimating τ and produces lower MSEs. Fortunately, the models with the smallest MSE for each number of years have the 95% confidence intervals containing the true values of the parameters. Therefore, even if the Frank is not selected as the best model due to the small number of years used, the model with the smallest MSE can reliably estimate the moments of ϵ distribution and τ.

6.2 Main Results

Table 2 shows the estimates using data for 2013-2017. In this estimation, the bin size used for the histogram estimation is 100, and the bandwidth used for the upper bound of bunching is 1,000. Since the true copula is ex-ante unknown, I use six candidate copula functions. The Clayton and Gumbel cannot estimate a negative dependence, and the tail dependence is asymmetric in the lower tail and upper tail, I use the Clayton and Gumbel rotated by 90 and 270 degrees. The reported confidence intervals are 2.5th and 97.5th percentiles of estimates using 1,000 bootstrap replicates.

In Table 2, $\bar{\phi}$ and m^U have the same value regardless of the selection of a copula function, because they are estimated before a copula function is used. The proportion of patients with friction is estimated as 53% and the marginal buncher is at 24.1K KRW. For bootstrap replicates, the probability that each copula model has the smallest MSE is 75% for Clayton270 and 23% for Gumbel90, and the RMSE of the two models is the smallest among the six models. These two copulas have a higher dependence on the lower tail than the upper tail.

Figure A.2 shows the scatter plots of the marginal CDFs of η and m when τ is -0.5. The fact that Clayton270 and Gumbel90 are the two best models suggests that those who are less ill are more dispersed in term of elasticity and those who are more ill tend to be more homogeneous. The estimate of τ using Clayton270 is -0.52, indicating that ϵ and m are negatively dependent. It means that the sicker they are, the less elastic they tend to be. The unconditional mean elasticity is estimated to be 0.1.

Figure 10 illustrates how θ is estimated using Clayton 270. The dots in the figure represent the estimated values of conditional CDF of ϵ below $\epsilon(m)$ as a function of $\hat{\bar{\phi}}$, $\hat{f}_0(m)$, and $\hat{f}_1(m)$.[10] Dots of the same color represent values observed at the same m for five years. For each m, $F(\epsilon(m))$ has the same value according to the stationary assumption, but $F_0^t(m)$ varies from year to year because of fees for health services increase every year. In addition, $f_1^t(m)/f_0^t(m)$ varies every year as well. If the dependence between ϵ and m is negative, $h(F(\epsilon), F_0^t(m); \theta)$ is increasing in $F_0^t(m)$ for a fixed $F(\epsilon)$. The curved lines show the shape of $h(u_1, u_2; \theta)$ for a fixed $u_1 = F(\epsilon(m))$ with u_2 varying. θ determines the curvature of the h function, and the least squares estimation finds the best θ that fits the data.

Figure 11 shows how much the estimates of the upper bound of the bunching window vary across bandwidth choices as a sensitivity check. If there is no noise in the data, the estimates of the upper bound will not depend on the bandwidth. Even if there is noise, the estimates of the upper bound will not change much above a certain bandwidth where the noise is sufficiently smoothed out. The figure shows that the estimates of the upper bound are stabilized at around 24K KRW for bandwidths above 1K KRW. Thus, I choose the bandwidth of 1K KRW in the estimation of the upper bound.

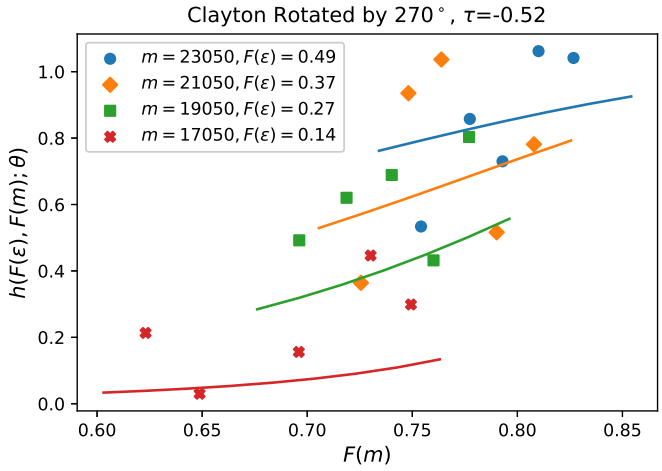
[10] It is obtained by rearranging equation (14) and replacing $\bar{\phi}$, $f_0(m)$, and $f_1(m)$ with their estimates, $1 - \left[1 - \hat{f}_1^t(m)/\hat{f}_0^t(m)\right] / \left(1 - \hat{\bar{\phi}}\right)$.

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

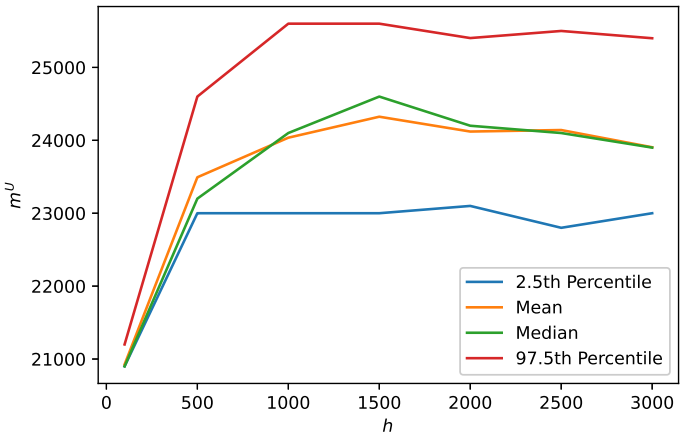
Notes: The table shows estimates of parameters using data between 2013 and 2017. $\bar{\phi}$ is the fraction of individuals with optimization frictions. m^U is the upper bound of the bunching window. $\mathbb{E}(\epsilon)$ is the mean of ϵ and $\sigma(\epsilon)$ is the standard deviation of ϵ estimated using beta distribution. τ is the Kendall's tau. RMSE is the squared root of mean squared error. 95% confidence intervals are estimated with 1,000 bootstrap replicates. The 2.5th and 97.5th percentiles of estimates are reported in square brackets. Probability of least RMSE is the probability of each copula having the least RMSE among six copulas for 1,000 bootstrap replicates.
Clayton90 and Gumbel90 are Clayton and Gumbel copulas rotated by 90 degrees, $C^{90}(u_1, u_2) = u_2 - C(1 - u_1, u_2)$. Clayton270 and Gumbel270 are Clayton and Gumbel copulas rotated by 270 degrees, $C^{270}(u_1, u_2) = u_1 - C(u_1, 1 - u_2)$. $C(u_1, u_2)$ is defined in Table A.3, and $u_1 = F_\eta(\eta)$ and $u_2 = F_0^t(m)$.

Figure 10: Estimation of Kendall's τ



Notes: This figure illustrates the estimation of Kendall's τ using Clayton copula $C\left(F\left(\epsilon\right), F\left(m\right)\right)$ rotated by 270 degrees. The x-axis is $F\left(m\right)$ and the y-axis is the conditional cumulative distribution of $F\left(\epsilon\right)$ given m. The dots are the estimates of the conditional cumulative distribution, which are obtained by rearranging equation (14), $1 - \left[1 - \hat{f}_1^t\left(m\right) / \hat{f}_0^t\left(m\right)\right] / \left(1 - \hat{\hat{\phi}}\right)$.

Figure 11: Estimates of the Upper Bound of Bunching Window by Bandwidth



Notes: This figure plots bootstrapped estimates of the upper bound of the bunching window by bandwidth h using data between 2013 and 2017.

7 Simulations of Counterfactual Policies

In this section, I conduct counterfactual simulations to compare the welfare associated with alternative out-of-pocket cost systems. I use the parameters of Clayton270 in Table 2. I draw a simulation sample $\{(m_k, \epsilon_k, d_k)\}_{k=1}^{K}$ with $K = 200,000$. First, $\{(F(\epsilon_k), F(m_k))\}$ 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 equation (2) as 64-years-olds are paying a linear coinsurance. For each (ζ_k, ϵ_k), I solve the patient optimization problem for alternative out-of-pocket systems. When $d_k = 1$ is drawn, I assume that patient k cannot re-optimize in response to the counterfactual system, but continue to choose the same expenditure as if they were 64 years old, or equivalently as if $s(m) = 0.3m$. Let m_k^* denote the total expenditure chosen by patient k under the counterfactual system. Then, I calculate the patients' utilities, the physicians' revenues, and the insurer's budget spending. The physician's revenue for patient k is the total expenditure m_k^* and the insurer's spending is $m_k^* - s(m_k^*)$. Patient k's utility is calculated using the utility function in equation (1). I impose $y = 0$ for simplicity because there is no income effect in this type of utility function.

Table 3 shows the simulation results for five alternative policies. Column (1) shows the utilities under the 2017 system in which there is a single notch at 15K KRW (see Figure 1(a)). I set the 2017 system as the baseline, and the rest of columns report the difference in welfare compared to the 2017 system.

Column (2) shows the change in welfare under the 2018 system, in which there are four intervals with different coinsurance rates creating one kink and two notches. Figure A.3 depicts this system.[11] In the new system, patients and physicians gain utility by 3% and

[11] The out-of-pocket costs are 1) 1.5K KRW when total expenditure is less than or equal to 15K KRW, 2) 10% coinsurance when total expenditure is over 15K and less than or equal to 20K KRW, 3) 20% coinsurance when total expenditure is over 20K and less than or equal to 25K KRW, and 4) 30% thereafter.

2.3% respectively, but the insurer has to spend 6.8% more. The new system is the least favorable for the insurer among the five alternatives. The utility gains for patients result from lower out-of-pocket costs between 15K KRW and 25K KRW than the 2017 system. And, the physician's revenue is greater than the baseline model, because the density of total expenditure between 15K KRW and 25K KRW increases as coinsurance rates are lower than the baseline model.

In columns (3)-(5), I derive linear coinsurance rates that guarantee the same welfare level as the baseline model for each of three agents. The insurer-equivalent coinsurance rate is 23.1%. The insurer-equivalent coinsurance model yields the highest sum of welfare among the five models, and it could increase the welfare of all three agents. The patient-equivalent coinsurance model is almost the same as the insurer-equivalent model, because the coinsurance rate is similar.

Lastly, I consider a nonlinear but smoothly changing out-of-pocket cost function, which varies from 1.5K KRW to 30% of coinsurance rate. Such a function must meet two conditions so that there are neither kinks nor notches: the out-of-pocket cost function itself is continuous, and the slope of out-of-pocket cost function is also continuous. To get an analytical solution, I assume a third-order polynomial function satisfying the following four restrictions: $s'(15000) = 0$, $s'(25000) = 0.3$, $s(15000) = 1500$, and $s(25000) = 7500$. Figure A.4 shows what the function looks like. The smoothly changing function is one of the functions suggested by a physician organization as an alternative to the 2018 system. However, it is not beneficial to physicians compared to the 2018 system, while the insurer could spend less.

Table 3: Policy Counterfactuals

	Baseline Welfare	Difference in Welfare				
	(1)	(2)	(3)	(4)	(5)	(6)
	2017 System	2018 System	Patient Equivalent	Clinic Equivalent	Insurer Equivalent	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)

Notes: The table shows the simulation results of policy counterfactuals. The sample size for simulations is 200,000. Column (1) shows the average welfare levels under the system that existed in 2007–2017 (see Figure 1(a)). Columns (2)–(6) show the changes in the average welfare levels compared to column (1) in each model. The out-of-pocket cost function in column (2) is the system introduced in 2018 (see Figure 1(b)). The out-of-pocket cost functions in columns (3)–(5) are linear coinsurance systems that make patients, clinics, and the NHIS indifferent to the baseline welfare, respectively. The out-of-pocket cost function in column (6) is the smooth cubic function in Figure A.4. Standard deviations are reported in parentheses.

8 Conclusion

This paper presents a novel method for estimating the elasticity distribution using a notch in the out-of-pocket costs in Korea. I take advantage of the institutional setting where the out-of-pocket costs applied to 64-year-olds and 65-year-olds in one calendar year are different. Since most existing studies observe only a single treated density, the counterfactual density must be interpolated. In addition, although the existing studies suggest a method to link the bunching responses to the elasticities through a quasi-linear utility function, they only use it to plug in the average bunching response in the last step. However, I suggest a way to link the observed ratios of the treated density and the counterfactual density at each value of medical expenditure with the distribution of elasticities. In addition, I present a method to estimate the dependence between elasticities and medical expenditures using the copula approach, when there is variation in the distributions of medical expenditures across multiple years.

In counterfactual policy simulations, I find that the new system introduced in 2018 increases the welfare of patients and physicians by 3% and 2.3%, respectively. However, the insurer needs to spend more by 6.8% than the old system before 2017. And, a linear coinsurance rate of 21.3% can achieve a Pareto improvement in that it increases welfare of patients and physicians while holding the insurer's spending constant.

The method in this paper has two limitations in the context of the bunching estimation. First, this method is a result of partial optimization because it does not take into account extensive-margin responses. Patients will have an incentive to adjust the frequency of visits if the out-of-pocket costs on one visit vary. Therefore, in future studies, it will be necessary to develop an estimation method under the full optimization that considers extensive-margin responses. Second, this method assumes that the fraction of patients with frictions is constant regardless of the elasticities or the medical expenditures. However, the evidence for this is lacking. If the friction increases in medical expenditures, the estimates in this paper will give the lower bound of the actual elasticities.

References

Aron-Dine, A., Einav, L., and Finkelstein, A. (2013). The RAND Health Insurance Experiment, Three Decades Later. *Journal of Economic Perspectives*, 27(1):197–222.

Bastani, S. and Selin, H. (2014). Bunching and Non-Bunching at Kink Points of the Swedish Tax Schedule. *Journal of Public Economics*, 109:36–49.

Blomquist, S., Newey, W. K., Kumar, A., and Liang, C.-Y. (2021). On Bunching and Identification of the Taxable Income Elasticity. *Journal of Political Economy*, 129(8):2320–2343.

Brot-Goldberg, Z. C., Chandra, A., Handel, B. R., and Kolstad, J. T. (2017). What Does a Deductible Do? The Impact of Cost-Sharing on Health Care Prices, Quantities, and Spending Dynamics. *The Quarterly Journal of Economics*, 132(3):1261–1318.

Chetty, R., Friedman, J. N., Olsen, T., and Pistaferri, L. (2011). Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records. *The Quarterly Journal of Economics*, 126(2):749–804.

Choi, J., Jo, C., Kim, S., and Choi, Y.-j. (2010). An Estimation of the Price Elasticity of Demand for Health Care after Implementation of Office Visit Copayment for Medical Aid Beneficiaries. *The Korean Journal of Health Economics and Policy*, 16(3):91–114.

Choi, S. (2018). A Study on the Health Service Utilization and Cost-Sharing. *The Journal of Women and Economics*, 15(1):25–47.

Clayton, D. G. (1978). A Model for Association in Bivariate Life Tables and Its Application in Epidemiological Studies of Familial Tendency in Chronic Disease Incidence. *Biometrika*, 65(1):141–151.

Duarte, F. (2012). Price Elasticity of Expenditure across Health Care Services. *Journal of Health Economics*, 31(6):824–841.

Einav, L., Finkelstein, A., and Schrimpf, P. (2017). Bunching at the Kink: Implications for Spending Responses to Health Insurance Contracts. *Journal of Public Economics*, 146:27–40.

Ellis, R. P., Martins, B., and Zhu, W. (2017). Health Care Demand Elasticities by Type of Service. *Journal of Health Economics*, 55:232–243.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., and Oregon Health Study Group (2012). The Oregon Health Insurance Experiment: Evidence from the First Year. *The Quarterly Journal of Economics*, 127(3):1057–1106.

Frank, M. J. (1978). On the Simultaneous Associativity of f(x,y) and X+y-F(x,y). *Aequationes Mathematicae*, 18(1-2):266–267.

Gumbel, E. J. (1960). Bivariate Exponential Distributions. *Journal of the American Statistical Association*, 55(292):698–707.

Hamilton, S. (2018). Optimal Deductibility: Theory, and Evidence from a Bunching Decomposition. *Working Paper*, page 38.

Keeler, E. B. and Rolph, J. E. (1988). The Demand for Episodes of Treatment in the Health Insurance Experiment. *Journal of Health Economics*, 7(4):337–367.

Kim, M. and Kwon, S. (2010). The Effect of Outpatient Cost Sharing on Health Care Utilization of the Elderly. *Journal of Preventive Medicine and Public Health*, 43(6):496.

Kim, W. (2021). The Effects of Out-of-Pocket Cost Change on Healthcare Utilization for the Elderly Outpatients in Korea. *The Korean Journal of Health Economics and Policy*, 27(2):95–120.

Kleven, H. J. (2016). Bunching. *Annual Review of Economics*, 8(1):435–464.

Kleven, H. J. and Waseem, M. (2013). Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan. *The Quarterly Journal of Economics*, 128(2):669–723.

Lohr, K. N., Brook, R. H., Kamberg, C. J., Goldberg, G. A., Leibowitz, A., Keesey, J., Reboussin, D., and Newhouse, J. P. (1986). Use of Medical Care in the Rand Health Insurance Experiment: Diagnosis- and Service-Specific Analyses in a Randomized Controlled Trial. *Medical Care*, 24(9):S1–S87.

Lu, Y., Shi, J., and Yang, W. (2019). Expenditure Response to Health Insurance Policies: Evidence from Kinks in Rural China. *Journal of Public Economics*, 178:104049.

Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E. B., and Leibowitz, A. (1987). Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment. *The American Economic Review*, 77(3):251–277.

Ministry of Health and Welfare (2017). Applying health insurance to infertility treatment and dementia neurocognitive tests [press release].

Mortenson, J. A. and Whitten, A. (2020). Bunching to Maximize Tax Credits: Evidence from Kinks in the US Tax Schedule. *American Economic Journal: Economic Policy*, 12(3):402–432.

Na, Y.-K. (2020). The Effect of Changes in Medical Use by Changing Copayment of Elderly. *Health Policy and Management*, 30(2):185–191.

National Assembly Budget Office (2016). Financial Forecasts on the Health Insurance Coverage Expansion Policy.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer Series in Statistics. Springer, New York Berlin Heidelberg, 2. ed edition.

Saez, E. (2010). Do Taxpayers Bunch at Kink Points? *American Economic Journal: Economic Policy*, 2(3):180–212.

Seim, D. (2017). Behavioral Responses to Wealth Taxes: Evidence from Sweden. *American Economic Journal: Economic Policy*, 9(4):395–421.

Sklar, A. (1973). Random Variables, Joint Distribution Functions, and Copulas. *Kybernetika*, 09(6):(449)–460.

The Korean Medical Clinic Association (2014). Immediately improve the copayment system for those over 65 or older [press release].

Trivedi, P. K. and Zimmer, D. M. (2007). Copula Modeling: An Introduction for Practitioners. *Foundations and Trends in Econometrics*, 1(1):1–111.

A Appendix

A.1 Existence and Uniqueness of $\epsilon(m)$

Proof. Since $\epsilon(m)$ is a linear transformation of $\eta(m)$, as $\epsilon(m) \equiv \eta(m) \times \frac{s_1}{2-s_1}$ for $s_1 > 0$, I will show the existence and the uniqueness of $\eta(m)$ instead. Let $A\left(\eta, m^I, s_0, s_1, m^*\right)$ denote the left-hand side of equation (6). Given s_0, s_1, m^*, and m^I, the partial derivative of $A\left(\eta, m^I, \cdot\right)$ with respect to η is strictly greater than zero.

$$\begin{aligned}
\frac{\partial}{\partial \eta} A\left(\eta, m^I, \cdot\right) &= -\frac{(2-s_1)}{(1+\eta)^2}\left(\frac{m^*}{m^I}\right)^{1+\frac{1}{\eta}} + \frac{2-s_1}{\eta(1+\eta)}\left(\frac{m^*}{m^I}\right)^{1+\frac{1}{\eta}} \log\left(\frac{m^*}{m^I}\right) + \frac{2-s_1}{(1+\eta)^2} \\
&= \frac{2-s_1}{(1+\eta)^2}\left[1 - \left(\frac{m^*}{m^I}\right)^{1+\frac{1}{\eta}} + \left(\frac{m^*}{m^I}\right)^{1+\frac{1}{\eta}} \log\left(\frac{m^*}{m^I}\right)^{1+\frac{1}{\eta}}\right] \\
&> 0 \qquad (22)
\end{aligned}$$

The last inequality results from the lower bound of natural logarithm, $\log x > 1 - \frac{1}{x}$, where $x = \left(\frac{m^*}{m^I}\right)^{1+\frac{1}{\eta}} < 1$.

As $\eta \to 0$, $A\left(\eta, m^I, \cdot\right)$ is strictly negative.

$$\lim_{\eta \to 0} A\left(\eta, m^I, \cdot\right) = (2-s_0)\left(\frac{m^*}{m^I}\right) - (2-s_1) < 0$$

for any $m^I > \frac{2-s_0}{2-s_1} m^*$. $\frac{2-s_0}{2-s_1} m^*$ is the upper bound of the dominated region.

Lastly, as $\eta \to \infty$, $A\left(\eta, m^I, \cdot\right)$ is strictly positive.

$$\lim_{\eta \to \infty} A\left(\eta, m^I, \cdot\right) = (2-s_0)\left(\frac{m^*}{m^I}\right) - (2-s_1)\left(\frac{m^*}{m^I}\right) = (s_1 - s_0)\left(\frac{m^*}{m^I}\right) > 0$$

since $s_1 > s_0$.

In sum, $A\left(\eta, m^I, \cdot\right)$ is strictly increasing in η, ranging from negative to positive numbers. Thus, there exists a unique η that solves $A\left(\eta, m^I, s_0, s_1, m^*\right) = 0$ given s_0, s_1, m^*, and m^I.

Thus, there exists a unique $\epsilon(m)$ for each m, given s_0, s_1, m^*, and m^I. □

A.2 $\epsilon(m)$ is strictly increasing in m

Proof. Since $\epsilon(m)$ is a linear transformation of $\eta(m)$, as $\epsilon(m) \equiv \eta(m) \times \frac{s_1}{2-s_1}$, and $\frac{s_1}{2-s_1} > 0$ for $s_1 > 0$, I will show that $\eta(m)$ is strictly increasing in m instead. $A\left(\eta, m^I, \cdot\right)$ is the same function as defined in Section A.1. The partial derivative of $A\left(\eta, m^I, \cdot\right)$ with respect to m^I is given by:

$$\begin{aligned}
\frac{\partial}{\partial m^I} A\left(\eta, m^I, \cdot\right) &= -(2-s_0)\left(\frac{m^*}{m^I}\right)\frac{1}{m^I} + (2-s_1)\left(\frac{m^*}{m^I}\right)^{1+\frac{1}{\eta}}\frac{1}{m^I} \\
&= \frac{1}{m^I}\left[(2-s_1)\left(\frac{m^*}{m^I}\right)^{1+\frac{1}{\eta}} - (2-s_0)\left(\frac{m^*}{m^I}\right)\right] \\
&< \frac{1}{m^I}\left[(2-s_1)\left(\frac{m^*}{m^I}\right) - (2-s_0)\left(\frac{m^*}{m^I}\right)\right] \\
&< 0
\end{aligned} \tag{23}$$

The last inequality results from $s_1 > s_0$. By the implicit function theorem, (22) and (23) imply $\partial\eta/\partial m > 0$. Thus, $\eta(m)$ is strictly increasing in m, and so is $\epsilon(m)$. □

A.3 Polynomial Approximations of the Counterfactual Density

In this section, I employ polynomial approximations to estimate the counterfactual density similar to previous studies on bunching estimation (e.g., Kleven and Waseem (2013)). This provides an example where the use of polynomial approximation fails to estimate the bunching mass due to the non-smoothness of the treated density. Following the existing literature, I estimate the counterfactual density only using the treated density, as if no control group is available. I run ordinary least squares regression to estimate the following:

$$\hat{f}_1(M_j) = \sum_{k=0}^{q} \beta_k M_j^k + \sum_{i \in \mathcal{I}_{\text{excl}}} \rho_i 1\{j = i\} + \sum_{i: M_i \in [m^L, m^U]} \tau_i 1\{j = i\} + v_j \tag{24}$$

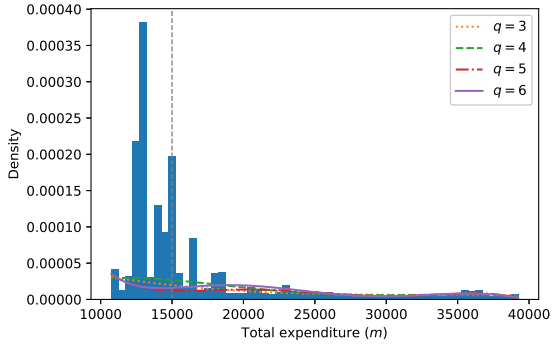
where $\hat{f}_1(M_j)$ is the density estimate for the jth bin, and M_j is the midpoint of the jth bin, q is the order of a polynomial, $\mathcal{I}_{\text{excl}}$ is a set of bin indicators that are excluded from the polynomial approximation such as outlier bins, m^L is the lower bound of the bunching window, and m^U is the upper bound of the bunching window. The parameters of interest are τ_is, which represent the differences from the estimated counterfactual density at each midpoint within the bunching window.

I use a bin size of 500. To simplify the analysis, I set $m^L = 13000$ and $m^U = 24100$ based on the main estimation results. [12] Additionally, I exclude two bins with M_j values of 12250 and 12750, showing peaks of the density. Figure A.1(a) shows the histogram of total expenses for age 65, along with the polynomial approximations using $q = 3, \ldots, 6$. Figure A.1(b) displays the values of τ_is for $M_i \in \left[m^L, m^U\right]$ for each q, as well as the observed difference between treated and control densities, $\hat{f}_1 - \hat{f}_0$. In the presence of any bunching response, τ_i should be positive for $M_i \leq m^*$ and negative for $M_i > m^*$. However, for all values of q, none of the polynomial approximations exhibit such patterns.

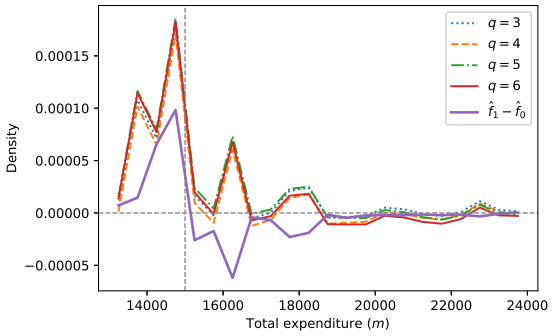
[12] In the literature, the lower and upper bounds of bunching window are parameters to be estimated.

Figure A.1: Polynomial Approximations of the Counterfactual Density

(a) Polynomial Approximations



(b) Coefficients for the Bunching Window



Notes: The figures shows the histogram of total expenses for age 65 and its polynomial approximations. q represents the degree of a polynomial. $\hat{f}_1 - \hat{f}_0$ represents the difference in the estimated densities for age 65 and age 64.

A.4 Additional Tables

Table A.1: Clinic Visit Fees and Conversion Factors in 2013-2017

Year	2013	2014	2015	2016	2017
Clinic Visit	9,430	9,710	10,000	10,300	10,620
Conversion Factor (CF)	70.1	72.2	74.4	76.6	79.0
CF Relative to 2015	0.94	0.97	1.00	1.03	1.06

Notes: The table shows clinic visit fees for follow-up visits, conversion factors, and conversion factors relative to 2015, for years between 2013 and 2017.

Table A.2: An Example of Fee-For-Service System: Physical Therapy

		(1) 2013	(2) 2014	(3) 2015	(4) 2016	(5) 2017	(6) 2018
A.	Outpatient Care - Follow-up Visits	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 $\leq$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)

Notes: The table shows the four selected items of physical therapy for follow-up visits and the fees of the items between 2013 and 2018. The fees are in KRW. The fees of superficial heat therapy are halved if deep heat therapy is accompanied. The total expenditures are the sum of fees of outpatient care, transcutaneous electrical nerve stimulation, deep heat therapy, and superficial heat therapy. The highest total expenditures less than or equal to 15K KRW are the sum of fees of items in parentheses. The total expenditures are rounded down to the nearest 10.

Table A.3: 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 $F(X, Y)$.

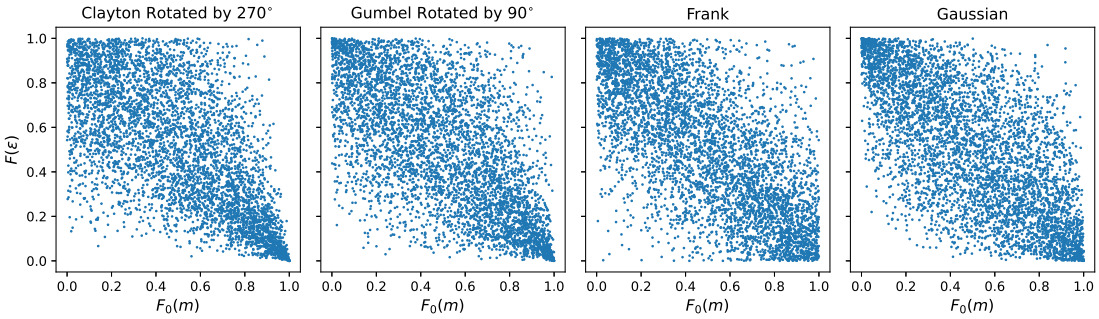
† Lower tail dependence λ_L is defined by $\lim_{v \to 0+} \Pr(u_1 < v | u_2 < v)$.

‡ 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.

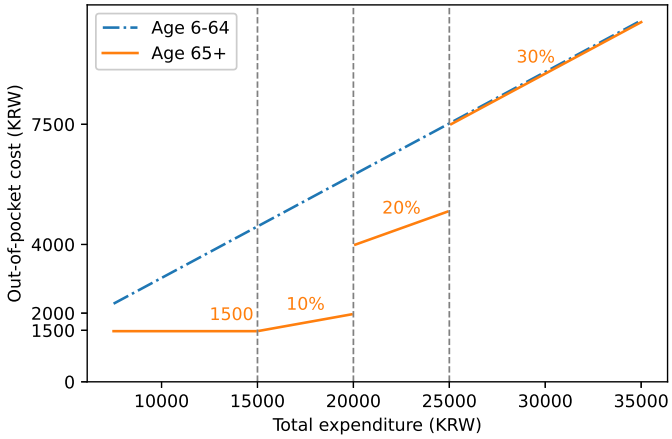
A.5 Additional Figures

Figure A.2: Scatter Plots of $F(\epsilon)$ and $F_0(m)$ by Copula Family when $\tau = -0.5$



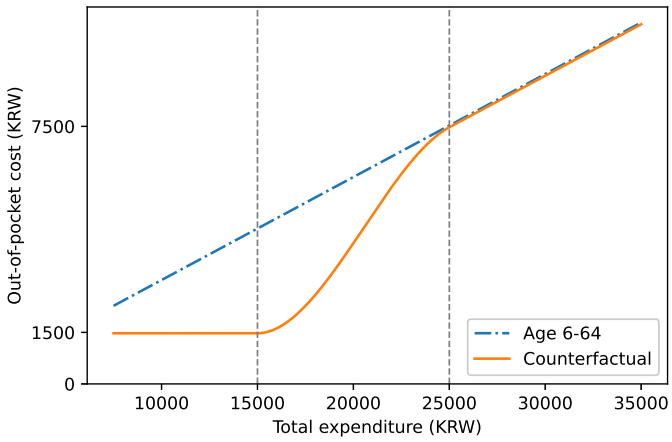
Notes: This figure illustrates the scatter plots of $F(\epsilon)$ and $F_0(m)$ generated by copula functions $C(F(\epsilon), F(m))$ when $\tau = -0.5$. The x-axis is $F(m)$ and the y-axis is $F(\epsilon)$.

Figure A.3: Out-of-Pocket Cost Function, 2018-Current



Notes: The function shows the out-of-pocket cost system since 2018.

Figure A.4: Smooth Cubic Out-of-Pocket Cost Function



Notes: The counterfactual function is $s(m) = 65625 - 10575\,(m/1000) + 555\,(m/1000)^2 - 9\,(m/1000)^3$. This is the cubic function satisfying $s'(15000) = 0$, $s'(25000) = 0.3$, $s(15000) = 1500$, and $s(25000) = 7500$. The function does not have a notch or a kink, but smoothly connects 1500 at $m = 15000$ and 7500 at $m = 25000$.