보험금융연구 제24권 제2호 (2013. 5) pp. 71-98

# 실손의료보험 손해액 극단값 혼합분포의 베이지언 추정*

## Bayesian Inference of a Mixture Model with Extreme Value Distributions in Korean Medical Insurance Applications

조 재 훈** · 이 근 창***
Jae Hoon Jho · Keunchang Lee

본 연구는 국내 실손의료보험 지급보험금을 극단값 혼합분포로 모형화하여 위험조정(risk-adjusted) 보험료를 산출하는 방법을 모색하고 있다. 의료비 지급보험금의 두터운 꼬리분포 특성을 반영하기 위해 손해분포함수를 혼합모형으로 설정하고 극단값 분포함수를 혼합분포의 구성함수로 사용하였으며, 두터운 꼬리의 특징이 나타나는 임계점(threshold)을 경험적 선택이 아닌 베이지언 방법을 통해 추정하였다. 연구결과의 실증분석으로써 민영건강보험회사 최근 2개년 장기손해보험의 실손의료보험 지급보험금 자료를 활용하여 혼합분포모형의 모수를 메트로폴리스-헤이스팅즈 알고리듬을 이용하여 추정하였으며, 추정결과를 이용한 실손의료보험 위험조정 순보험료 산출방법을 예시하였다. 분석 결과 지급보험금 심도 분포의 상이함에 따른 위험을 고려한 보험료는 기존의 대수의 법칙에 토대로 둔 기대값으로 산출한 보험료 보다 10.5% 정도의 위험할증이 필요한 것으로 나타나고 있다.

**국문 색인어:** 극단값 분포, 극단값 이론, 메트로폴리스-헤이스팅즈, 베이지언 추정, 실손의료보험 일반화파레토분포, 임계점, 혼합분포

**한국연구재단 분류 연구분야 코드:** B051603

* 이 논문은 대산 신용호 기념사업회에서 지원하는 연구비에 의하여 수행되었음.
** 영남대학교 국제통상학부 조교수(jaehoonjho@yu.ac.kr), 주저자, 교신저자
*** 영남대학교 국제통상학부 교수(kclee@ynu.ac.kr)
논문 투고일: 2013. 01. 04, 논문 최종 수정일: 2013. 02. 03, 논문 게재 확정일: 2013. 02. 22

# I. 서론

보험업감독규정의 개정에 따라 2009년 10월 1일 실손의료보험 표준화 이후 손해액[1]의 빈도와 심도를 분석하여 수리적으로 타당하며 실무적으로 활용 가능한 보험료 산출방법에 대한 관심이 증대되고 있다. 실손의료보험은 2012년 말 기준 가입자가 2,500만 명을 넘는 국민적 보험으로 자리 잡았으며 3년 갱신계약으로 사망보장과 암 진단비, 입원 일당 등 각종 보장을 포함하는 상품으로 판매되고 있으며, 2013년부터는 갱신주기가 1년으로 줄어들고 실손의료보험만으로도 가입이 가능하게 되었다[2]. 따라서 선택적으로 가입이 가능해지고 보험계약자가 부담하는 총보험료가 낮아져 실손의료보험 시장규모의 확대로 이어질 것으로 예상된다.

실손의료보험의 주요 연구 관심은 전통적인 빈도와 심도의 추정뿐 아니라, 피보험자의 보험금청구 패턴의 변화 여부 및 도덕적 해이와 역선택의 가능성 여부에 초점이 모아지고 있다. 민영건강보험의 실손의료비 보험료는 연령, 성별, 가입금액 등으로 매우 세분화 되어 있으며 방대한 데이터를 활용하여 산출되고 있다. 이렇게 산출된 보험료는 통계자료가 희박한 몇몇 연령구간을 제외하면 통계적으로 의미 있는 보험료로 간주될 수 있으나, 손해분포(loss distribution)를 고려하지 않은 대수의 법칙에 의존한 기대값의 추정치로 산출된 보험료라는 한계를 지니고 있다.

이 같은 방식으로는 손해분포가 서로 다름에도 불구하고 기대값이 같을 경우 동일한 보험료가 산출될 수 있어 보험료가 위험노출단위에 비례하지 않는 바람직하지 못한 결과를 초래할 수 있다. 다시 말해서 손해분포의 상이함을 고려하여 보험료를 차등 반영하는 수리적 방법이 위험노출에 비례하는 보험료 산출 원칙을 보다 충족시키기 때문에, 심도의 확률분포 자체를 고려하는 방법을 모색할 필요가 있다.

특히 실손의료보험에서 상대적으로 고액 의료비를 발생시키는 위험들은 발생

1) 손해액은 일반적으로 지급보험금, 발생손해액, 최종손해액 등으로 단계적으로 구분되며 본 연구에서는 지급보험금 실증자료에 손해진전계수를 적용하여 사용하였다.
2) 표준형 단독 실손의료보험 판매(2013년 1월1일), 금융위원회 보도자료(2012.12.21).

가능한 고액의 의료비와 관련하여 위험에 대한 조정(risk-adjustment)을 하지 않을 경우 보험회사의 위험관리에 부정적인 요소로 작용할 뿐만 아니라, 실손의료보험 가입자에게도 보험으로 누릴 수 있는 사회적비용 절감의 효과를 감소시키는 요인이 될 수 있다. 최근 관측되는 국민 개개인의 건강에 대한 관심증가, 복지제도 개선에 따른 건강검진 횟수와 적용대상 범위의 확대, 실손의료보험 제도 개선의 영향 및 손해액 심도의 변동과 심화 등 실손의료보험이 제도적으로 확장되고 세분화되고 있는 시점에서 보험료 역시 이에 상응하는 보다 정밀한 수리적 모형을 바탕으로 산출할 필요가 있다.

이러한 관점에서 본 연구에서는 실손의료보험 지급보험금을 가벼운 꼬리분포(light-tailed)와 두터운 꼬리분포(heavy-tailed)의 혼합분포모형으로 설명하고 그 모형의 모수를 추정하는 기법을 제시하였다. 즉, 실손의료보험 지급보험금 심도의 변동성을 설명하기 위해 손해분포가 두터운 꼬리일 때와 그렇지 않은 경우로 구분하여 일반화파레토분포함수(generalized Pareto distribution)와 일반적 손해분포의 혼합구성함수가 특정한 임계점(threshold)을 기준으로 혼합된 모형을 제시하고자 한다. 이러한 손해분포의 두터운 꼬리분포 특성은 극단값 이론(extreme value theory)에서 그 이론적 배경을 찾을 수 있다.

혼합 과정에는 두 구성함수가 혼합되는 임계점이 반드시 필요하다. 본 연구에서는 혼합분포를 기존의 연구 중 Behrens·Lopes·Gamerman(2004)과 Cebrián·Denuit·Lambert(2003)의 모형을 활용하였으며, 임계점의 추정은 메트로폴리스-헤이스팅즈(Metropolis-Hastings) 알고리듬을 이용한 베이지언 추정을 시도하였다. 이들의 연구를 포함하여, 극단값 이론을 이용한 극단값 분포의 모수추정은 기존 연구들에서도 찾아 볼 수 있지만, 구성함수들이 혼합되는 기준인 임계점을 추정하는 엄밀한 방법을 제시하지 못하고 있거나 임계점 이하의 구성함수의 모수추정을 수행하지 않고 있다. 따라서 본 연구에서는 임계점과 두 구성함수의 모수를 동시에 추정하여 기존 연구들의 단점을 보완하여 실무적 활용이 가능한 모형으로 발전시키는 시도를 하였다.

## II. 선행연구

Cebrián · Denuit · Lambert(2003)는 손해보험에서 관측되는 거대손해에 대한 확률분포함수를 극단값 이론을 이용하여 설명하는 시도를 하였으며, 미국의 단체의료보험(group medical insurance)의 실증자료를 이용하여 분포함수 도출과정과 모수추정 결과를 자세히 제시하고 있다. 이들의 연구가 Behrens · Lopes · Gamerman(2004)의 연구와 다른 점은 혼합분포를 이용하지 않고 특정한 임계점을 초과하는 손해액(excess over threshold)을 극단값 분포로 설정하여 의료비 손해액의 오른쪽 꼬리분포를 설명하고 있으며 꼬리분포 두터움의 정도를 나타내는 극단값지수(extreme value index)를 정밀하게 추정하는 시도를 하였다. 또한 임계점 초과손해액의 극단값 분포모형은 재보험의 초과손해액특약(excess of loss treaties), 발생가능손해액(probable maximum loss), value-at-risk 점추정 등을 계산하는데 응용될 수 있음을 언급하였다.

동일한 실증자료를 이용하여 의료비 손해액을 극단값 분포로 설명하는 시도는 Balasooriya · Low(2008)의 연구에서도 찾아볼 수 있으며, 그들은 일반화파레토분포와 일반화람다분포(generalized lambda distribution)를 비교하여 분석한 바 있다. 손해보험의 영역에서도 이러한 극단값 이론을 적용한 연구가 진행되었는데, McNeil(1997)은 화재보험 고액 손해액(Danish large fire insurance losses)을 이용하여 일반화파레토분포의 형태모수를 추정하고 고분위점을 계산하는 방법을 예시하고 있다.

윤석훈(2011)은 포아송-GPD 모형으로 코스피지수 자료의 베이지안 극단값 분석을 통하여 일별 로그수익률 분포의 오른쪽 꼬리는 비교적 짧은 반면 로그손실률 꼬리분포는 다소 두텁다는 결론을 도출하였고 그 현상을 가격제한폭 15% 정률제의 운용결과로 유추한 바 있다. 김수영 · 송종우(2012)는 Peaks-Over-Threshold 방법론을 이용하여 국내 자동차보험의 손해율을 서로 다른 3가지 기법으로 추정하여 그 결과를 비교분석하였다.

위의 선행연구, Cebrián · Denuit · Lambert(2003), Balasooriya · Low(2008), 윤석훈

(2011), 김수영·송종우(2012)는 극단값 이론에 근거하여 우선적으로 임계점의 수준을 결정한 뒤, 임계점을 초과하는 손해액을 극단값 분포로 모형화하여 손해분포함수의 꼬리부분에 대한 설득력 있는 수리적 설명을 하였고 다양한 분포함수와 추정기법을 활용하고 있으나 극단값 이론을 적용하기 위한 임계점을 선택하는 엄밀한 통계적 방법을 제시하지는 못하고 있다는 한계점이 있다.

임계점 추정에 관한 시도는 Behrens·Lopes·Gamerman(2004)에서 찾아볼 수 있는데, 그들은 일반화파레토분포를 이용하여 어떠한 임계점을 넘어서는 비정상적 크기의 확률변수의 분포를 극단값 이론으로 모형화하여 이 임계점의 불확실성에 대하여 연구하였다. 이들은 손해액 확률변수가 임계점을 넘지 않는 경우에는 일반적으로 사용되는 손해분포함수를 가정하고, 손해액의 확률변수가 임계점을 넘어서는 경우에는 일반화파레토분포를 이용하였으며, 특정한 임계점을 접점으로 두 손해분포구성함수의 혼합분포(mixture distribution)를 고안하였다.

기대값이 유한값으로 항상 존재하는 감마분포, 로그정규분포 등 일반적인 손해분포함수만을 이용하면 비정상적으로 큰 손해가 발생하는 경우를 분포함수모델로 설명하기 어려운 한계가 존재하게 된다. 이에 Behrens·Lopes·Gamerman(2004)은 이러한 문제점을 개선하기 위해 분포의 중심과 분포의 오른쪽 꼬리를 구분하여 혼합분포구성함수를 설정하였다. 이들은 혼합분포함수의 모수 추정에 대한 방법으로는 베이지언 추정을 이용하였는데 임계점을 경험적으로 또는 임의로 결정하였던 기존의 연구와는 다르게 임계점을 확률변수로 간주하여 베이지언 모수 추정 과정에 포함시켜 혼합분포함수의 모든 모수와 유기적으로 연결시켜 추정하는 효과적인 방법을 시도하였다.

Behrens·Lopes·Gamerman(2004)은 임계점을 모형의 확률변수로 포함시켜 혼합분포함수모형이 구조적으로 결함이 없다는 장점이 있으나, 임계점 이하의 일반적 크기의 손해액 자료들이 임계점의 추정을 왜곡시킬 수 있는 단점이 존재하기 때문에 임계점의 변화에 따른 극단값지수의 변동과 다른 모수들의 변동을 면밀히 관찰하여 결론을 내려야할 필요성을 언급하고 있다. 한편, 이와 관련된 임계점을 추정하는 방법에 대하여 Bermudez·Turkman·Turkman(2001)은 심도 있게 연구를

진행하였는데 임계점과 상응하는 순서확률변수(order statistics)를 추정모수에 포함시킨 베이지언 모형으로 고분위점(high quantile)을 예측하는 방법을 제시하고 있다.

# III. 극단값 혼합분포 모형

## 1. 꼬리분포 모델(Tail Distribution Modeling)

극단값이론은 20세기 초반부터 이론적인 토대가 마련되었으며 1970년대 POT(Peaks Over Threshold)방법이 소개되면서부터 공학분야 및 금융분야에서 자주 응용되기 시작하였고, 21세기에 들어서는 보험수리 분야에서도 활발하게 적용되고 있다. 특히, 금융리스크의 심도가 과거에 비해 변동성이 심해지고 그 크기가 과거 경험을 이용한 고전적인 예측을 상당히 초과하여 나타나는 경우가 많아지고 있는 최근의 시점에서 극단값 분포가 다양한 금융리스크 추정을 위해 많은 역할을 할 수 있을 것으로 전망된다. 극단값이론의 확률변수는 다양하게 정의가 가능하지만, 본 연구에서는 앞으로 모든 확률변수와 분포함수를 손해확률변수(loss random variable)와 손해분포함수(loss distribution)의 범위로 제한하였다.

손해분포함수의 임계점($u$)을 초과하는 초과손해액 확률변수의 조건부확률분포함수를 구하기 위한 극단값 이론과 손해분포함수의 꼬리분포 두터움(heavy-tailedness)을 나타내는 지표인 극단값지수(extreme value index)를 추정하는 기존의 방법을 우선 소개한다.

손해확률변수를 $X$, 손해확률분포함수를 $F_X$, 추후에 추정되어지는 임계점을 $u(u>0)$ 라고 하면, 임계점을 초과하는 손해액의 확률분포함수는 다음과 같이 표시된다.

$$F_u(x) = P(X-u \le x \mid X > u) = \frac{F(x+u)-F(u)}{1-F(u)},$$

$$x \geqq 0, \quad u < x_F$$

이때 $x_F$는 분포함수의 오른쪽 끝점으로서, $x_F = \sup\{x \in R | F(x) < 1\}$ .

극단값 이론에서 자주 사용되는 Pickand-Balkema-de Hann 정리[3]에 의하면 임계점 $u$가 증가함에 따라 초과손해액의 분포함수는 일반화파레토분포에 아래와 같이 수렴하게 된다.

$$\lim_{u \to x_F} sup_{\ 0 < x < x_F - u} |F_u(x) - G_{\xi,\sigma(u)}(x)| = 0$$

이 때 $G$는 일반화파레토함수로서 $G(x|\xi,\sigma,u) = 1 - (1 + \xi x/\sigma(u))^{-1/\xi}$ 이며, $\sigma$는 $u$의 양의 함수이다.

위의 극한방정식에서 임계점이 충분히 클 때 조건부 초과손해액$\{X - u | X > u\}$의 분포함수를 일반화파레토함수로 설정하여 꼬리분포를 추정하는 것이 가능하게 되는 이론적인 근거를 찾을 수 있다. 극단값 이론에 관한 자세한 내용은 Balkema · de Haan(1974), Pickands(1975), Embrechts · Klüppelberg · Mikosch(1997) 등에서 참조할 수 있다.

## 2. 혼합분포함수와 우도함수

Behrens · Lopes · Gamerman(2004)의 모형에 따르면, 손해확률변수 $X$가 임계점$(u)$보다 작은 경우, $\{X | X \leq u\}$의 분포함수를 양의 정수 $m$개의 모수 $\eta = (\eta_1, ..., \eta_m)$ 로 설명되는 $H_X(x|\eta)$로 설정하고[4], $X$가 임계점$(u)$ 보다 큰 경우, $\{X | X > u\}$의 분포함수가 일반화파레토분포, $G(x|\xi,\sigma(u))$를 따른다고 가정하면, 손해확률변수 $X$의 전 영역에서의 손해분포함수는 아래와 같은 혼합분포함수로 나타낼 수 있다.

$$\overline{F_X}(x) = I_{\{x \leqq u\}} \overline{H}(x|\eta) + I_{\{x > u\}} \overline{H}(u|\eta)\, \overline{G}(x - u|\xi,\sigma)\ , \quad x \geqq 0$$

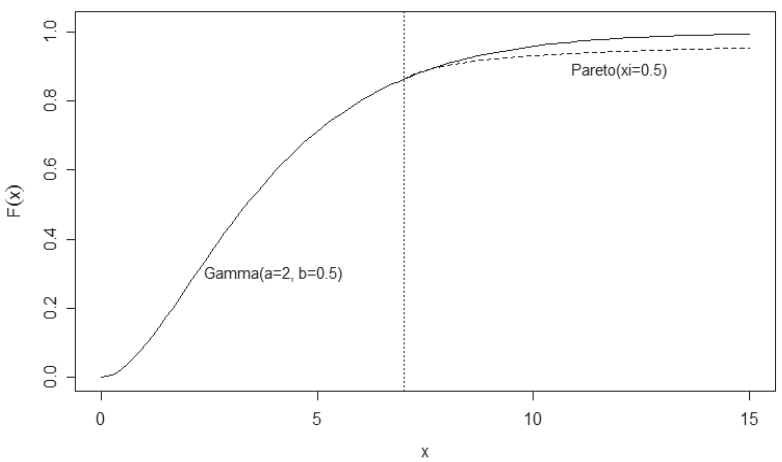
3) Balkema · de Haan(1974), Pickands(1975).

4) Behrens · Lopes · Gamerman(2004)은 임계점 이하를 설명하는 구성함수를 감마분포를 이용하였다. 즉, 감마분포의 형상모수 $\alpha$와 척도모수 $\beta$로서, $H(x|\eta) = G(x|\alpha, \beta)$.

이때 $G(x-u|\xi,\sigma)=1-\left(1+\frac{\xi(x-u)}{\sigma}\right)^{-1/\xi}, \quad x \geqq u$

극단값 이론에 의하면 임계점($u$)의 값에 관계없이 위의 혼합함수는 동일한 극단값지수를 갖는 것으로 알려져 있으며, 이러한 경우를 동일한 MDA(Maximum Domain of Attraction)에 속한다고 표현한다. 감마분포($\alpha=2, \beta=0.5$)와 일반화파레토분포($\xi=0.5, \sigma=1$)의 혼합분포를 아래의 그림에서 살펴볼 수 있다.

〈그림 1〉 감마분포와 일반화파레토분포의 혼합분포 예시




혼합분포함수의 구성함수 $H(x|\eta)$, $G(x|\xi,\sigma)$의 확률밀도함수를 각각 $h(x|\eta)$, $g(x|\xi,\sigma)$로 표시하면, 모수 $\eta$, $\xi$, $\sigma$를 베이지언 모형에 포함시킨 우도함수(likelihood function)를 구할 수 있다. 본 연구에서는 우도함수에 포함된 임계점을 임의로 선택된 상수가 아닌 확률변수로 포함시키기 위해서, $X_{i,n}$을 $i$-번째 순서확률변수라 하고, 두터운 꼬리분포를 따르는 순서확률변수의 개수를 확률변수 $k$ 로 정의하면, 임계점 확률변수 $u_k$ 를 다음과 같이 도출할 수 있다.

$$u_k = X_{n-k+1,n},$$
$$X_{1,n} \leq \cdots \leq X_{n-k,n} < u_k \leq X_{n-k+1,n} \leq \cdots \leq X_{n,n} \quad 5)$$

따라서, 확률변수 $k$ 와 임계점 확률변수 $u_k$ 를 이용한 우도함수와 추정모수를 모두 포함한 베이지언 구조를 아래와 같이 도출하였다.

$$l(x_{1,n},x_{2,n},...,x_{n,n};\eta,\xi,\sigma,k)=\prod_{i=1}^{n-k}h(x_{i,n}|\eta)\prod_{i=n-k+1}^{n}\overline{H}(u_k|\eta)g(x_{i,n}-u_k|\xi,\sigma),$$

$$X\sim I_{\{x\leqq u_k\}}\overline{H}(x|\eta)+I_{\{x>u_k\}}\overline{H}(u_k|\eta)\,\overline{G}(x-u_k|\xi,\sigma)\,,\quad x\geqq 0,$$

$$\eta_j\sim\pi_j(\eta_j),\quad j=1,2,...\,m,$$

$$\xi\sim\pi_\xi(\xi),\quad \sigma\sim\pi_\sigma(\sigma),\quad k\sim\pi_k(k),$$

이때 $u_k=x_{n-k+1,n}$, $\quad g(x|\xi,\sigma)=\dfrac{1}{\sigma}\left(1+\dfrac{\xi}{\sigma}x\right)^{-(1+\xi)/\xi}$.

위에서 정의된 베이지언 구조는 임계점을 임의의 상수로 선택하는 경우를 배제하며, 베이지언 실험을 통해 $k$의 선험적분포를 다양하게 변화시켜 임계점($u$)의 변동에 따른 극단값지수의 변화를 관찰할 수 있어서 기존의 연구방법에 비하여 수리적으로 정밀한 임계점 계산방법이라 할 수 있다. 본 연구의 실증분석은 실손의료보험 지급보험금의 심도를 혼합분포모형으로 추정하는 것에 초점을 맞추었으며 특히, 위의 이론적 배경을 바탕으로 지급보험금의 손해분포함수의 두터운 꼬리분포 특성이 나타나는 임계점을 베이지언 기법을 통해 추정하였다.

## 3. 모수의 결합확률분포와 사후분포

앞에서 정의한 우도함수에서 베이지언 추정이 필요한 모수들을 요약하면 다음과 같으며

$$\theta=(\eta_1,...,\eta_m,\xi,\sigma,k),$$

5) $x_{i,n}$은 서로 동일한 값이 없음을 가정하였다. 즉, 모든 $i\neq j$ 에 대하여 $x_{i,n}\neq x_{j,n}$

모수들의 선험적 분포($\pi$)와 혼합분포의 우도함수를 곱하여 아래와 같이 모수($\theta$)의 결합확률밀도함수를 구할 수 있다.

$$\pi(\eta,\xi,\sigma,k|\, x_{1,n}, x_{2,n}, \dots, x_{n,n}) \;\propto\; \left(\prod_{i=1}^{n-k} h(x_{i,n}|\eta) \prod_{i=n-k+1}^{n} \overline{H}(u_k|\eta) g(x_{i,n}-u_k|\xi,\sigma)\right) \times \left(\prod_{j=1}^{m} \pi_j(\eta_j)\right) \times \pi_\xi(\xi)\pi_\sigma(\sigma)\pi_k(k)$$

결국 각각의 모수들의 사후분포는 위의 결합확률밀도함수를 이용하여 각각의 모수 관점에서 상수로 간주되는 부분을 제외하여 다음과 같이 비례적 관계로 나타낼 수 있다.

$$\pi(\eta_i|\,\cdot\,) \;\propto\; \overline{H}(u_k|\eta)^k \prod_{i=1}^{n-k} h(x_{i,n}|\eta) \times \prod_{j=1}^{m} \pi_j(\eta_j),$$

$$\pi(\xi|\,\cdot\,) \propto \left(\prod_{i=n-k+1}^{n} g(x_{i,n}-u_k|\xi,\sigma)\right) \times \pi_\xi(\xi),$$

$$\pi(\sigma|\,\cdot\,) \propto \left(\prod_{i=n-k+1}^{n} g(x_{i,n}-u_k|\xi,\sigma)\right) \times \pi_\sigma(\sigma),$$

$$\pi(k|\,\cdot\,) \;\propto\; \overline{H}(u_k|\eta)^k \prod_{i=1}^{n-k} h(x_{i,n}|\eta) \prod_{i=n-k+1}^{n} g(x_{i,n}-u_k|\xi,\sigma) \times \pi_k(k)$$

위의 사후분포함수들은 일반적으로 함수적 형태가 알려져 있지 않기 때문에 근사값을 계산하는 즉, 사후분포의 적분을 계산하는 수리적 알고리듬이 필요하다.

마지막으로 위에서 정의된 베이지언 구조 하에서 시뮬레이션을 통한 모수들의 사후분포와 사후평균을 구하기 위해서는 적절히 선택된 모수의 초기값이 필요하다. 두 구성함수, 감마분포와 일반화파레토분포의 모수들은 그 성질이 잘 알려져 있어 자료의 요약통계 만으로도 그 수준을 가늠할 수 있어 별다른 문제를 발생시키지 않으나, 임계점을 초과하는 자료의 개수로 정의된 $k$ 확률변수는 사후평균을 결정짓는 매우 중요한 요소이나 초기값 선택에 대한 정보는 자료의 관찰만으로

얻기 어렵다. 극단값 이론을 완벽히 적용하기 위해 너무 큰 임계점을 선택하면 극단값 분포의 추정모수의 분산과 표준오차가 증가하게 되어 정확성이 떨어지며, 반대로 분산을 조절하기 위해 임계점을 너무 낮게 선택하면 극단값 분포로 설명되지 않는 자료가 모수추정에 포함되어 극단값 이론을 적용할 수 없게 된다. 본 연구에서는 이러한 분산-편의 상충관계(variance bias trade-off)를 최소화하기 위해 초기값($k_0$)을 10 부터 실손의료보험 지급보험금 실증자료의 총갯수 $n$ 까지 반복 대입하여 모수들의 사후분포를 계산하고 초기값-극단값지수 도표 $\{(k_0, \hat{\xi})\}$를 얻은 뒤, $k_0$가 증가하더라도 $\hat{\xi}$의 추정량이 변하지 않는 구간을 관찰하여 극단값지수와 상응하는 임계점을 설정하여 적용하였다.

다음 장에서는 실손의료보험 지급보험금 실증자료를 극단값 혼합분포로 추정하기 위해 혼합분포의 구성함수들을 정의하고, 우도함수, 추정모수의 선험적분포, 그리고 프로포잘분포를 이용하여 메트로폴리스-헤이스팅즈 알고리듬을 수행하는 구체적인 방법을 서술하였다.

# IV. 실증분석 및 결과

## 1. 실증분석 자료

실증자료의 객관성과 안정성을 위해 전체 민영건강보험회사 최근 5개년도 장기손해보험의 기초통계자료를 조사하였다. 우리나라의 실손형 의료보험상품의 역사를 간략히 살펴보면, 1984년 장기손해보험의 '상해의료비'가 최초로 도입되었으며 1992년에 상해와 질병을 모두 보장하는 '입원의료비', 1999년에 보상한도가 대폭 증가한 '입통원의료비'를 거쳐 2009년 10월 1일부터 표준화된 '실손의료보험'이 판매되고 있다. 특히, 실손의료보험은 상해 · 질병 · 종합의 3가지 담보형태와 입원 · 통원 두 가지 의료형태로 구분하여 총 6개의 담보종목으로 구분되어진다.

본 연구에서는 5개년 기초통계 조사자료 중 '입원의료비' 질병담보, '입통원의료비' 질병입원 담보, '실손의료보험' 질병입원 담보에 대하여 사고날짜 기준 2009년

4월 1일 이후의 지급보험금 자료를 활용하였다. 아래의 〈표 1〉은 자료 추출 기준을 요약한 것이다.

〈표 1〉 실증분석 자료 출처, 추출조건, 추출기준 및 추출기간

| 구분 | 내용 |
|---|---|
| 출처 | 보험개발원 장기손해보험 기초통계자료 |
| 추출조건 | 입원의료비의 질병담보,<br>입통원의료비의 질병담보,<br>실손의료보험의 질병입원담보 |
| 추출기준 | 계약: 보험개시일 기준 가입금액별 신계약건수 합계<br>사고: 가입금액 및 지급보험금 금액구간별 사고건수와 지급보험금 합계 |
| 추출기간 | 사고날짜 기준 2009.4.1 ~ 2011.3.31 |

최근 실손의료보험의 기초통계 요약자료는 아래의 〈표 2〉, 〈표 3〉 및 〈표 4〉와 같이 계약의 가입금액과 지급보험금의 크기를 기준으로 신계약건수합계, 사고건수합계, 지급보험금합계로 요약할 수 있다. 2009년 10월에 실손의료보험이 표준화 되어 계약시점을 기준으로 보상하는 손해의 보험금 산정기준이 상이할 수 있어(예를 들어 실손의료비 입원의료비 10% 본인부담) 자료의 연속성이 보장되지 않으나, 이 연구의 목적이 의료비 손해액에 대한 혼합분포모형 개발 및 추정이기 때문에 실손의료보험 관련 제도개정으로 인한 영향은 적을 것으로 판단되어 자료의 보정 없이 사용하였다. 아래의 〈표 2〉, 〈표 3〉 및 〈표 4〉는 2009년도 추출자료 중 계약기준 신계약건수, 사고날짜기준 사고건수와 지급보험금의 합계를 가입구간 및 금액구간별로 요약한 자료이다.

〈표 2〉 FY2009[6] 신계약건수 합계

(단위: 건)

| 가입금액 | 3천만 이하 | 3천만~5천만 | 5천만~1억 | 1억 초과 |
|---|---|---|---|---|
| 신계약건수 | 247,508 | 4,214,688 | 6,884,038 | 150,663 |

6) 보험계약일 기준 2009.4.1.~2010.3.31

〈표 3〉 AY2009[7] 사고건수 합계

(단위: 건)

| 지급보험금 구간 | 가입금액 | | | |
|---|---|---|---|---|
| | 3천만 이하 | 3천만~5천만 | 5천만~1억 | 1억 초과 |
| 1백만 이하 | 629,893 | 48,359 | 178,875 | 12,833 |
| 1백만~5백만 | 111,847 | 11,229 | 37,002 | 1,597 |
| 5백만~1천만 | 7,648 | 817 | 2,666 | 117 |
| 1천만~3천만 | 1,954 | 208 | 664 | 28 |
| 3천만~1억 | 14 | 14 | 26 | - |
| 합계 | 751,356 | 60,627 | 219,233 | 14,575 |

〈표 4〉 AY2009 지급보험금 합계

(단위: 백만 원)

| 지급보험금 구간 | 가입금액 | | | |
|---|---|---|---|---|
| | 3천만 이하 | 3천만~5천만 | 5천만~1억 | 1억 초과 |
| 1백만 이하 | 195,611 | 17,017 | 59,860 | 3,814 |
| 1백만~5백만 | 217,916 | 21,840 | 71,674 | 3,007 |
| 5백만~1천만 | 51,306 | 5,535 | 17,917 | 821 |
| 1천만~3천만 | 29,034 | 2,972 | 9,515 | 407 |
| 3천만 초과 | 508 | 509 | 1,072 | - |
| 합계 | 494,375 | 47,873 | 160,039 | 8,049 |

기초통계 자료는 금액구간별 지급보험금합계와 사고건수의 합계의 요약통계로서 본 연구의 목적에 맞도록 사고건별 자료로 재생성하였다. 이 때 원 자료의 지급보험금 금액 구간별 평균과 재생성된 사고건별자료의 지급보험금 평균이 일치하도록 하였다. 또한 사고건별 자료의 생성 중 자기회귀 가능성 등에 의해 무작위 특성이 저해될 가능성을 최소화하기 위해 전체 사고건별 자료에서 각 연도마다 5,000개의 자료를 무작위로 다시 선별하여 시뮬레이션에 사용할 최종 사고건별 지급보험금을 결정하였다.

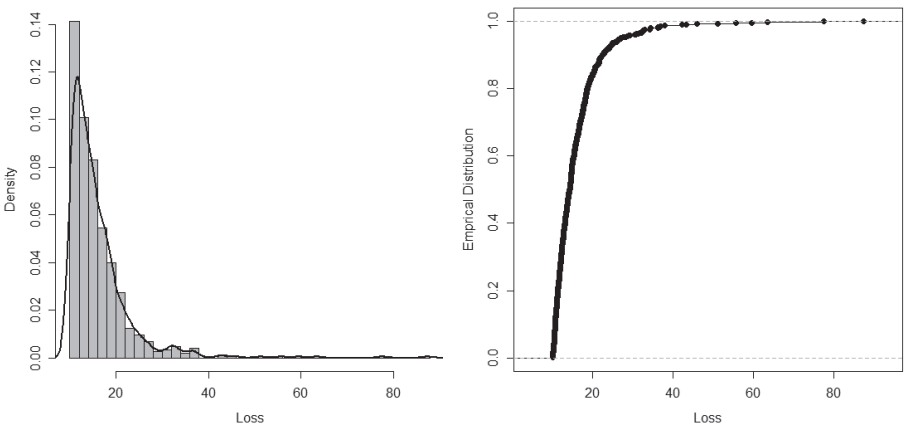
7) 사고날짜 기준 2009.4.1.~2010.3.31

## 2. 자료의 특성

〈그림 2〉는 2009.4.1.~2010.3.31. 사이에 발생한 지급보험금 중 1천만원 이상 자료의 히스토그램과 누적경험분포로서 지급보험금 손해분포의 두터운 꼬리 특성을 시각적으로 가늠해 볼 수 있다. 또한 〈표 5〉의 실증자료 요약통계를 살펴보면 중간값(0.4203)과 평균(0.7713)의 차이가 매우 크며 평균이 3분위(0.6549)보다 큰 것으로 보아 손해분포의 오른쪽 꼬리가 길게 늘어져있는 것으로 분석된다.

〈그림 2〉 AY2009 질병입원 손해액 히스토그램 및 누적경험분포

(단위:백만 원)




〈표 5〉 AY2009 질병입원의료비 요약통계

(단위: 백만 원)

| 최소값 | 1분위 | 중간값 | 평균 | 3분위 | 최대값 |
|---|---|---|---|---|---|
| 0.0042 | 0.2126 | 0.4203 | 0.7713 | 0.6549 | 87.3900 |

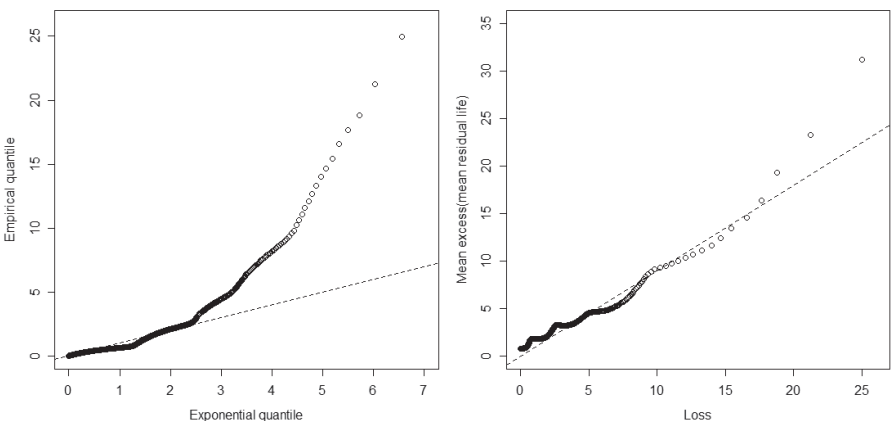
요약통계를 통한 질병담보 지급보험금의 두터운 꼬리분포의 단순한 추측을 보완하기 위해 자료의 분석을 통하여 얻을 수 있는 지수분위도표(exponential quantile plot) 및 초과평균도표(mean excess plot)를 이용하였다. 이 방법들은 엄밀하게 극단값지수의 통계적 추정치를 제공하지는 않으나, 실증자료의 초기 분석에

서 간단하게 사용할 수 있다는 장점이 있다.

먼저 지수분포함수가 두터운 꼬리분포함수와 그렇지 않은 함수들을 구분하는 기준으로 고려되므로 동일한 분위확률에 대하여 자료의 경험적 분위가 지수함수의 분위보다 크게 나타나는 경향은 두터운 꼬리 특성이 존재하기 때문인 것으로 판단된다. AY2009 질병입원의료비 손해액 실증자료의 지수분위도표는 아래 〈그림 3〉의 왼쪽 그림에 예시하였는데 두터운 꼬리의 특징을 확인할 수 있다.

〈그림 3〉 AY2009 질병입원의료비의 지수분위도표와 초과평균도표

(단위: 백만 원)




꼬리분포 특징을 파악하는 두 번째 방법은 특정한 임계점을 초과하는 손해액의 평균을 살펴봄으로써 가능한데 이는 초과평균도표(mean excess plot) 또는 평균여명도표(mean residual life) 방법으로 불린다. 일반화파레토분포의 특정한 임계점($u$)을 초과하는 손해액의 평균 함수는 다음과 같이 임계점에 대한 1차함수로 표시할 수 있기 때문에,

$$e(u) = E(X-u|X>u) = \frac{\sigma+\xi u}{1-\xi}, \quad \xi < 1,$$

이러한 선형 특징을 이용해 아래의 경험적 산개도 $(u, \tilde{e}(u))$ ,

$$\tilde{e}(u) = \frac{\sum_{i=1}^{n}(X_i - u)^+}{\sum_{i=1}^{n} 1_{X_i > u}}$$

의 모양이 일정한 기울기를 갖는 직선의 모양을 보인다면 두터운 꼬리 특성의 가능성을 배제할 수 없다.

AY2009 지급보험금 자료의 초과평균도표는 〈그림 3〉의 오른쪽 그림에서 찾아볼 수 있는데 실증자료의 가장 큰 몇 개의 자료를 제외한다면 선형함수의 모양을 띠고 있어 임계점을 초과하는 손해액을 일반화파레토함수로 모형화 할 수 있는 이론적 배경을 제공해준다.

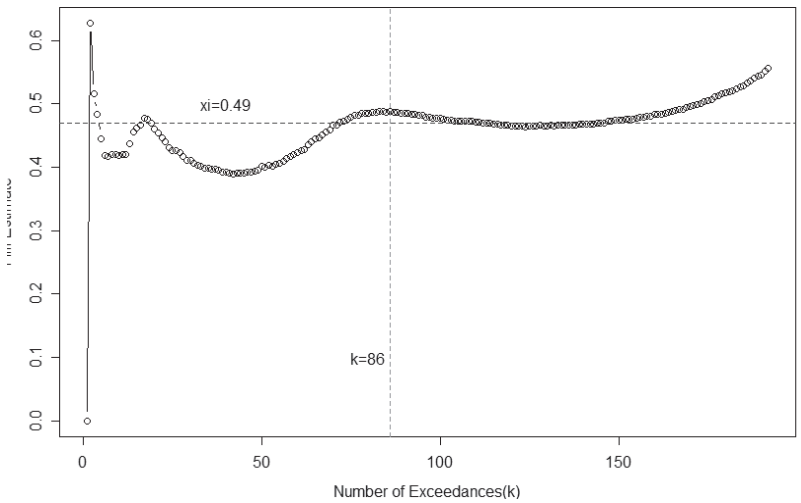
마지막으로 Hill 추정치를 이용하여 꼬리분포를 설명하는 극단값 분포함수의 극단값지수($\xi$)의 수준을 가늠할 수 있다. 주어진 $n$개의 실증자료 $X_1, \cdots, X_n$ 를 이용하여 임계점($u$)을 초과하는 자료의 개수($k$)를 이용하여, Hill 추정량을 다음과 같이 계산할 수 있다(Hill, 1975).

$$\hat{\xi}_{k,n} = \frac{1}{k} \sum_{j=n-k+1}^{n} (log\, X_{j,n} - log\, X_{n-k+1,n}),$$

$$X_{1,n} \leqq X_{2,n} \leqq \cdots \leqq X_{n,n} \quad (X_{i,n} : i-th \text{ order statistic})$$

$k$ : 임계점($u$)을 초과하는 자료의 개수.

〈그림 4〉 AY2009 질병입원의료비 손해액 Hill 도표




임계점을 초과하는 자료의 갯수를 증가시키며 그에 따른 Hill 추정량의 관계를 그림으로 나타낸 Hill 도표의 관찰을 통하여 안정적으로 유지되는 극단값 지수를 추정값으로 선택할 수 있다. 〈그림 4〉에서 AY2009 지급보험금 꼬리분포의 극단값 지수인 Hill 추정량은 임계점을 초과하는 자료의 갯수가 86개 이상일 경우 큰 변동이 없이 $\hat{\xi}= 0.49$ 수준을 유지하는 것을 확인할 수 있다. 극단값 이론에 의해 극단값 지수가 양수인 경우 MDA 분류기준에 의해 프레체함수군(Fréchet distribution family)에 속하여, AY2009 지급보험금은 두터운 꼬리분포의 특성을 지니는 것으로 판단된다.

## 3. 극단값 혼합분포 모수의 추정

Ⅲ장에서 도출한 혼합분포함수 모수들의 사후분포는 각 모수 관점의 상수를 제외한 비례적 관계로 표현이 가능하나, 우도함수가 이질적인 두 구성함수의 혼합분포이기 때문에 구체적인 함수의 형태(analytic form)는 일반적으로 알려져 있지 않으며 더욱이 사후분포평균을 구하기 위한 적분의 가능여부도 불분명하다. 앞에

서 언급한 것과 같이 본 연구에서는 마르코프사슬 몬테카를로(MCMC, Markov Chain Monte Carlo) 방법 중 널리 사용되는 메트로폴리스-헤이스팅즈(Metropolis-Hastings) 알고리듬을 이용하여 각 모수의 사후분포를 계산하고 그 결과를 바탕으로 사후분포평균을 계산하였다.

메트로폴리스-헤이스팅즈 알고리듬을 수행하기 위해서는 우도함수가 우선 정의 되어야 하며, 우도함수에 포함된 모수들의 선험적 분포와 프로포잘분포의 적절한 선택이 필요하다. 또한 알고리듬의 첫 단계에서 사용되는 모수의 초기값을 설정하여야 한다. 알고리듬을 단계별로 구분하여 간략히 살펴보면 다음과 같다.

- 0단계: 추정모수의 초기값 $\theta_0$ 설정, $j=0$.
- 1단계: 추정모수의 프로포잘분포 $\alpha$ 를 이용하여 후보모수값[8] $\theta_c$ 생성
  $\theta_c \sim \alpha(\theta_c|\theta_j)$
- 2단계: 균일분포 확률변수를 무작위 생성하고, $w \sim U(0,1)$, 생성된 후보모수값 $\theta_c$의 채택 또는 기각을 아래와 같이 결정.
  $$\theta_{j+1} = \begin{cases} \theta_c, & R > w \\ \theta_j, & R \le w \end{cases}, \qquad R = \frac{\pi(\theta_c)\alpha(\theta_j|\theta_c)}{\pi(\theta_j)\alpha(\theta_c|\theta_j)}$$
- $j$ 를 1씩 증가시키며($j \leftarrow j+1$) 1단계로 돌아가 충분한 횟수로 반복

우선, 우도함수를 구성하는 혼합구성함수를 살펴보면, 임계점 이하의 손해액의 확률분포는 자료의 특성에 따라 다양하게 설정이 가능하며 소손해(small losses)와 중간크기손해액(medium losses)을 관찰하여 적절한 손해함수를 선택할 수 있다. 본 연구의 목적은 손해확률변수가 임계점 이상에 해당하는 경우의 손해확률분포함수를 극단값 분포로 추정하는 방법에 관한 것이므로, 손해액이 임계점 이하인 경우 해당되는 손해확률분포함수 $H(x|\eta)$의 선택에 관한 내용은 다루지 않았다. 본 연구에서는 임계점 이하의 손해액을 두 개의 모수를 갖는 감마분포로 설정하였으며, 극단값 혼합분포 모형에 포함되는 두 개의 구성함수 $H(x|\eta)$, $G(x|\xi,\sigma)$

8) candidate parameter.

를 요약하면 아래와 같다.

$$H(x|\eta_1,\eta_2)=\int_0^x \frac{1}{\eta_2^{\eta_1}\Gamma(\eta_1)} y^{\eta_1-1}e^{-y/\eta_2}dy,\quad \eta_1>1,\ \eta_2>0\,,$$

$$G(x-u|\xi,\sigma)=1-\left(1+\frac{\xi(x-u)}{\sigma}\right)^{-1/\xi},\quad x\geqq u.$$

모수의 선험적 분포는 일반적으로 주관적(subjective), 주관적 무정보접근(subjective & non-informative), 무정보접근(non-informative) 방법으로 분류되는데, 본 연구에서는 Behrens·Lopes·Gamerman(2004)의 주관적 방법을 차용하여, $k$를 제외한 모수의 선험적분포를 특정한 하이퍼파라미터(hyper parameter)에 의한 감마분포로 선택하였다. 그러나 임계점보다 큰 자료의 개수를 나타내는 확률변수 $k$의 선험적분포는 Behrens·Lopes·Gamerman(2004)이 제시한 모수에 대한 해당분야 전문가의견[9]에 관한 정보가 없어 무정보접근 방법을 적용하였다. 마지막으로 프로포잘분포(proposal distribution)는 일반적으로 많이 사용되는 확률보행(random walk) 프로포잘을 설정하였다[10].

〈표 6〉 혼합분포 우도함수 모수의 선험적 분포

| 모수 | 선험적분포 $\pi(\cdot)$ |
|---|---|
| $\eta_1,\ \eta_2$ | 감마분포: $\eta_1\sim\Gamma(a_1,b_1),\quad \eta_2\sim\Gamma(a_2,b_2)$ |
| $\xi,\ \sigma$ | 감마분포: $\xi\sim\Gamma(c_1,d_1),\quad \sigma\sim\Gamma(c_2,d_2)$ |
| $k$ | 부적절균일분포 $k\sim U(0,\infty)$ |

주: $a_i,b_i,c_i,d_i$ 는 하이퍼파라미터로서 자료의 요약통계 관찰을 통하여 선택함.

마지막으로 각 모수들의 초기값을 설정하여야 하는데, 임계점 이하 손해액의 구성함수인 감마분포의 모수들은 실증자료의 표본평균과 분산을 통하여 계산된

9) "eliciting prior information" (Coles and Tawn, 1996: 467-469)

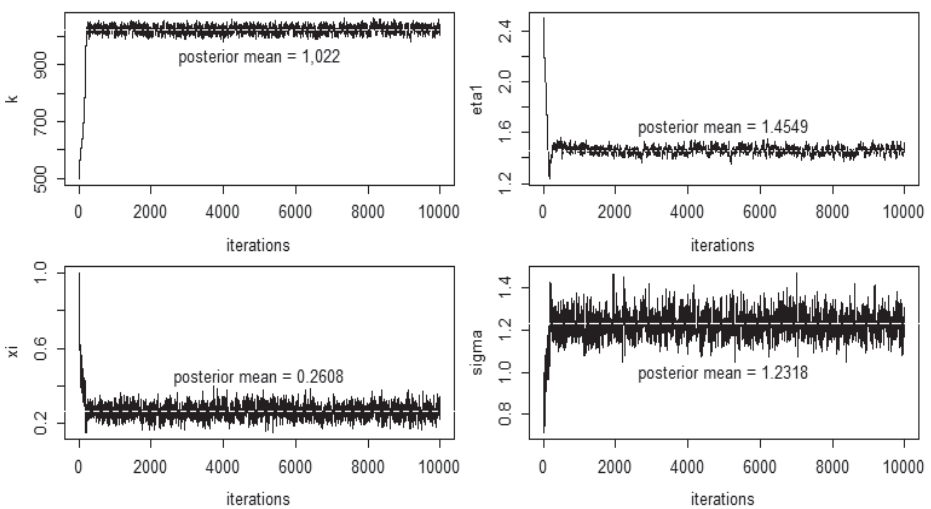
10) 프로포잘분포 $\alpha$를 평균이 $\theta_j$인 정규분포를 이용하는 방법으로 $\theta_c\sim\theta_j+N(0,s^2)$의 관계로 후보모수값이 생성된다. 이 때 표준편차 $s$는 후보자모수 채택 빈도가 30%~80% 사이에 형성되도록 조정하여 선택한다.

추정값을 사용하였다. 하이퍼파라미터와 임계점 이상의 구성함수, 즉 일반화파레토분포의 형상모수와 척도모수는 시뮬레이션 결과를 관측하여 시행착오를 통해 빠르게 수렴하도록 선택하였다. 다만 임계점을 초과하는 자료의 개수($k$)의 초기값은 시뮬레이션에 의한 사후평균의 변동을 관찰하기 위해 변수로 설정하여 가각의 초기값에 따른 결과를 모두 기록하였다. 〈표 7〉과 〈그림 5〉는 AY2009 질병입원 지급보험금 자료(단위: 백만 원)에 초기값 $k_0 = 500$인 경우의 1만 번 시뮬레이션 결과로서 혼합분포 모수 $k, \xi, \sigma, \gamma_1, \gamma_2$ 의 사후평균과 표준오차를 확인할 수 있다.

〈표 7〉 2009 혼합분포 모수의 사후평균과 표준오차[11] (초기값 $k_0 = 500$)

| 모수 | $\hat{k}$ | $\hat{\xi}$ | $\hat{\sigma}$ | $\hat{\eta_1}$ | $\hat{\eta_2}$ |
|---|---|---|---|---|---|
| 사후평균 | 1,022 | 0.2608 | 1.2318 | 1.4549 | 0.3446 |
| 표준오차 | 0.1366 | 0.00040 | 0.00067 | 0.00035 | 0.00011 |

〈그림 5〉 2009 혼합분포 모수의 1만 번 시뮬레이션 생성값(초기값 $k_0 = 500$)




11) 시뮬레이션 초기의 불안정한 생성 값을 제외하기 위해 10,000 번의 시뮬레이션 중 첫 2,500 생성값을 제외하여 사후평균과 표준오차를 계산하였음(burn-in period 2,500).

## 4. 임계점 추정 및 결과 분석

위의 시뮬레이션 실험에서 얻은 사후평균은 다른 모수들의 초기값에 상관없이 동일한 수렴결과를 보여주는 것으로 확인되었다. 그러나 $k$의 초기값 $k_0$을 서로 다르게 설정하여 반복 실험한 경우에는 각 모수의 사후평균이 아래 〈표 8〉의 예시와 같이 서로 다른 값으로 수렴하였다.

〈표 8〉 초기값 $k_0$에 대한 혼합분포 모수의 사후평균

| $k_0$ | $\hat{k}$ | $\hat{u}$ | $\hat{\xi}$ | $\hat{\sigma}$ | $\hat{\eta_1}$ | $\hat{\eta_2}$ |
|---|---|---|---|---|---|---|
| 100 | 247 | 4.47(백만 원) | 0.3356 | 1.9608 | 1.0330 | 0.6353 |
| 1,000 | 1,022 | 0.75(백만 원) | 0.2613 | 1.2326 | 1.4561 | 0.3441 |
| 3,000 | 2,356 | 0.34(백만 원) | 0.8351 | 0.3121 | 1.3100 | 0.4234 |

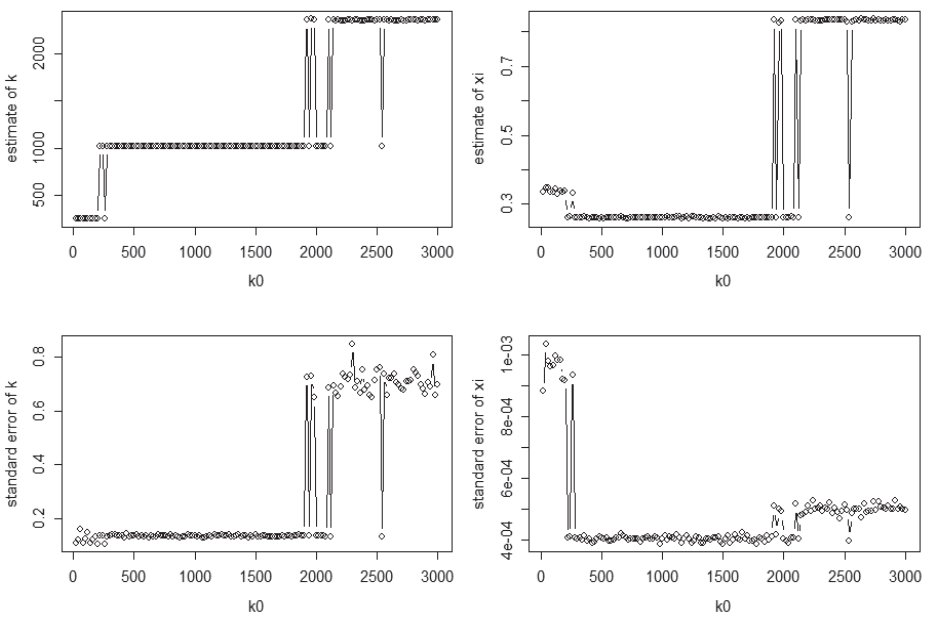
이러한 현상은 우도함수가 임계점 이하는 감마함수의 속성을 지니고 임계점 초과는 이질적인 일반화파레토함수로서 두터운 꼬리특성을 지니기 때문이며, 초기값 $k_0$는 두 구성함수의 가중치를 결정하는 계수처럼 작용하여 혼합분포함수의 우도함수를 특정 값 주변에서 급격히 변화시키는 것으로 분석된다. 이러한 이유로, 관측이 불가능한 임계점보다 크거나 작은 초기값의 설정에 따라서 $k$와 $\xi$의 시뮬레이션 생성값이 서로 다른 값으로 수렴하는 현상이 발생하였다. 흥미로운 사실은 초기값 $k_0$의 값을 1부터 자료의 총 개수 $n$까지 변화시키더라도 $k$와 $\xi$ 생성 값은 특정한 몇 개의 값으로만 수렴하는 것을 〈그림 6〉에서 확인할 수 있다.

초기값 $k_0$가 (0, 200) 구간 사이의 경우에는 $\hat{k}=246$, $\hat{\xi}=0.3374$로 일정한 수준에서 수렴하지만 극단값 분포에 포함되는 자료의 개수가 너무 적어 표준오차가 크게 관측되었으며, 두 구성함수가 혼합되는 것으로 보이는 (200, 280) 구간에서는 불규칙한 모양을 보이다가 (280, 1900) 구간 사이에서는 이전 수렴값과는 다른 수준, $\hat{k}=1022$, $\hat{\xi}=0.2619$ 에서 안정적으로 수렴하고 있다. 마지막으로 초기값 $k_0$가 1900 이상인 경우에는 또 다른 값들 $\hat{k}=2356$, $\hat{\xi}=0.8351$로 수렴하고 있다. 요약하면, 초기값 $k_0$가 (0, 3000) 구간에서 변화할 때 모수 $\hat{k}$, $\hat{\xi}$의 수렴값은 오직 3가지로

만 나타나고 있다.

위의 결과를 분석하기 위해서 혼합분포함수의 결합 구조를 다시 한 번 살펴볼 필요가 있다. $H_X$와 $G_{X-u}$가 혼합되는 임계점에서 확률밀도함수를 계산해보면 일반적으로 불연속인 것을 알 수 있는데[12] 이것은 손해분포함수를 두 개의 이질적인 구성함수의 혼합으로 정의하였기 때문에 귀결되는 당연한 사실이다. 초기값 $k_0$가 극단값 분포의 실제 모수 근처에서 시작되는 경우 정상적으로 수렴하는 것으로 보이며, 반대로 $k_0$가 너무 큰 값에서 시작되는 경우에는 포함되지 않아야 할 자료가 우도함수의 극단값 분포를 따르는 것으로 간주되어 심한 편의(bias)를 보여줘 점점 가벼운 꼬리 분포로 옮아가는 것으로 판단된다.

〈그림 6〉 초기값 $k_0$와 $\hat{k}, \hat{\xi}$의 사후평균도표와 표준오차




12) 누적확률분포는 연속임.

결국 AY2009 질병입원 지급보험금 자료의 혼합분포 극단값지수는 $\hat{\xi}=0.2619$로 선택되는 것이 가장 적합한 것으로 판단되며 상응하는 임계점은 안정적인 수렴구간 중 가장 작은 $\hat{k}=1021$을 선택할 경우 $\hat{u}=X_{n,n-\hat{k}}=0.73$(백만 원)로 추정할 수 있다. 결국 혼합분포함수의 추정결과는 아래와 같이 표현할 수 있다.

$$\overline{F_X}(x) \approx I_{\{x \leq \hat{u}\}}\overline{\Gamma}(x|\hat{\eta}_1,\hat{\eta}_2)+I_{\{x>\hat{u}\}}\overline{\Gamma}(u|\hat{\eta}_1,\hat{\eta}_2)\overline{G}(x-\hat{u}|\hat{\xi},\hat{\sigma})$$

이 때 $\hat{u}=0.73$(백만 원), $\hat{\eta}_1=1.4529$, $\hat{\eta}_2=0.3451$, $\hat{\xi}=0.2619$, $\hat{\sigma}=1.2315$,

위의 과정을 연도별로 반복 시행하여 얻은 질병입원 지급보험금의 극단값 혼합분포 추정결과를 아래의 〈표 9〉에 요약하였다.

〈표 9〉 연도별 질병입원의료비 극단값 혼합분포 추정결과

| 연도 | $\hat{k}$ $(\hat{k}/n)$ | $\hat{u}$ | $\hat{\xi}$ | $\hat{\sigma}$ | $\hat{\eta_1}$ | $\hat{\eta_2}$ |
|---|---|---|---|---|---|---|
| 2009 | 1021 (20.4%) | 73만 원 | 0.2619 | 1.2315 | 1.4529 | 0.3451 |
| 2010 | 1102 (22.0%) | 95만 원 | 0.2357 | 1.4863 | 1.3948 | 0.4708 |

주: $n$: 각 연도별 사고건수 총합계. 2009년 $\hat{k}/n$ 비율은 질병입원 사고자료 중 추정 임계점 보다 큰 자료의 비중이 20.4% 임을 의미함.

## 5. 추정 결과의 활용

이상의 극단값 혼합분포 추정결과를 이용하여 실손의료보험 질병입원 담보 지급보험금의 두터운 꼬리 특성을 반영한 위험조정(risk-adjustment) 보험료를 계산할 수 있다. 이는 보험회사에서 널리 활용하는 표준편차 방식의 위험보험료 부가방식에 비하여 계산과정이 매우 복잡하지만 서론에서 언급하였던 것처럼 정밀한 수리적 기법을 통한 보험료 산출방법의 한가지로 활용될 수 있을 것이다.

사고빈도에 관한 추가적인 정보로서 2010년 실손의료보험 질병입원 담보를 포함한 유효계약건수($m$)가 주어진 경우, 위험조정 보험료는 다음과 같은 공식을 이

용하여 계산할 수 있다[13].

$$\hat{f} = \sum_{i=1}^{n} X_i / m,$$

$$\begin{aligned}\hat{s} &= \int_0^{\infty} 1 - F_X(x)\, dx \\ &= \int_0^{\hat{u}} \overline{H}(x|\hat{\eta}_1, \hat{\eta}_2)\, dx + \hat{H}(\hat{u}|\hat{\eta}_1, \hat{\eta}_2) \times \int_{\hat{u}}^{\infty} \overline{G}(x - \hat{u}|\hat{\xi}, \hat{\sigma})\, dx \\ &= \int_0^{\hat{u}} \overline{H}(x|\hat{\eta}_1, \hat{\eta}_2)\, dx + \overline{H}(\hat{u}|\hat{\eta}_1, \hat{\eta}_2) \times \left[\hat{u} + \frac{\hat{\sigma}}{1-\hat{\xi}}\right]\end{aligned}$$

$$\hat{P} = f \times s$$

2010년도 질병입원 담보의 심도($\hat{s}$)는 위의 공식을 이용하여 아래와 같이 계산되며,

$$\hat{s} = 0.4125 + 0.2304 \times 2.8919 = 1.0789 \text{ (단위: 백만 원)}$$

따라서 극단값 혼합분포 모형의 위험조정에 의한 보험료는

$$\frac{\hat{s} - \frac{1}{n}\sum_{i=1}^{n} X_i}{\frac{1}{n}\sum_{i=1}^{n} X_i} = \frac{1.0789\text{-}0.9763}{0.9763} = 10.5\%$$

의 증가분이 필요한 것으로 계산되었다.

이와 같이 손해액 심도 분포의 상이함에 따른 위험을 고려한 보험료는, 기존의 대수의 법칙에 토대로 둔 손해액의 기대값으로 산출한 보험료 보다 10.5% 정도의 위험할증이 필요한 것으로 분석되었다.

13) 일반화파레토분포 $G_{\xi,\sigma,u}$의 기대값은 $0 < \xi < 1$ 일 때, $E[X] = u + \sigma/(1-\xi)$.

# V. 결론 및 시사점

최대우도함수 추정법에 의해 계산되는 Hill 추정량의 단점은 극단값 분포를 적용하기 위한 충분히 큰 임계점의 선택이 필요하다는 점이다. 지수분위도표 또는 초과평균도표를 이용하여 임계점을 선택하는 기법이 이러한 단점을 일부분 보완해주지만 도표를 시각적으로 판단할 경우 분석하는 사람에 따라 다른 결론을 내릴 수 있다. 앞 장에서 베이지언 방법으로 추정한 임계점($\hat{u}= u_{\hat{k}}$)에 의한 극단값 지수를 Hill 추정량과 아래의 〈표 10〉에서 비교해 보았다.

〈표 10〉 손해액 꼬리분포 극단값지수 추정결과

| 자료 | | 극단값지수 및 임계점초과 자료개수 $\hat{\xi}$ $(\hat{k})$ | |
|---|---|---|---|
| | | 혼합분포 베이지언 추정 | Hill 추정 |
| 질병 입원 | 2009년 | 0.2618 (1021) | 약 0.49 (약 86) |
| | 2010년 | 0.2357 (1102) | 약 0.47 (약 90) |

위의 두 추정치가 차이를 보는 이유 중의 하나는 혼합분포 추정치는 임계점보다 작은 실증자료가 혼합분포함수의 구조 때문에 우도함수에 포함되어 있지만 Hill 추정량은 주어진 임계점 이하의 자료는 추정에 사용되지 않기 때문이다. 따라서 선택된 임계점이 상대적으로 작은 경우에는 Hill 추정량이 극단값 분포를 따르지 않는 표본의 자료가 포함되어 극단값 이론에 의한 일반화파레토분포로의 점근성을 저해하며, 반대의 경우 임계점이 너무 크게 선택된 경우에는 추정에 사용되는 자료의 개수가 줄어들어 Hill 추정량의 표준오차(standard error)가 증가하게 되어 추정의 신뢰도가 떨어지게 된다.

〈표 10〉의 결과에 의하면 장기손해보험 실손의료비 손해액은 2009년도 대비 2010년도에 혼합모형에 의한 극단값 지수가 증가하였음을 알 수 있다. 이것은 연도별 손해액 분포의 꼬리가 점점 두터워지고 있는 하나의 예로서 여러 연도에 대하여 추가적인 연구를 통해 그 근거를 강화시킬 수 있을 것이다. 혼합분포 추정 임계점을 두터운 꼬리분포의 기준으로 설정한 경우, 연도별 질병입원 담보의 지

급보험금 평균값은 아래와 같이 각 연도에서 10% 이상의 위험조정이 필요한 것으로 판단된다.

〈표 11〉 질병입원 의료비 손해액 평균값 조정

| 자료 | | 지급보험금 경험평균 | 위험조정 평균 | 임계점이상 평균 $E[X\|X>\hat{u}]$ |
|---|---|---|---|---|
| 질병 입원 | 2009년 | 77만 원 | 85만 원(+10.5%) | 246만 원 |
| | 2010년 | 98만 원 | 108만 원(+10.2%) | 292만 원 |

보험회사는 과거의 자료를 관찰하여 실제 비용(loss cost)에 가까운 보험료를 산출하기 위해 손해액의 증가추세를 반영하고 의료수가의 인상률을 함께 반영하는데 선형모형을 이용하는 경우가 일반적이나, 매년 보험회사가 경험하는 손해액이 증가하는 속도가 빨라지고 있는 현 시점에서 본 연구의 모형인 혼합분포 추정방법이 실손의료비 보험료 산출을 보완하는 방법으로 활용될 수 있을 것이다.

# 참고문헌

윤석훈,「코스피 지수 자료의 베이지안 극단값 분석」,『응용통계연구』, 2011, 24, pp. 833-845.

김수영·송종우,「POT방법론을 이용한 자동차보험 손해율 추정」,『응용통계연구』, 2012, 25, pp. 101-114.

Balasooriya, U. and Low, C., "Modeling Insurance Claims with Extreme Observations: Transformed Kernel Density and Generalized Lambda Distribution", *North American Actuarial Journal*, 12:2, 2008, pp. 129-142.

Balkema, A. and de Haan, L., "Residual Life Time at Great Age", *Annals of Probability*, 2:5, 1974, pp. 792-804.

Behrens, C.N., Lopes, H. and Gamerman, D., "Bayesian Analysis of Extreme Events with Threshold Estimation", *Statistical Modelling*, 4, 2004, pp. 227-244.

Bermudez, P.Z., Turkman, M.A.A. and Turkman, K.F., "A Predictive Approach to Tail Probability Estimation", *Extremes*, 4:4, 2001, pp. 295-314

Cebrián, A., Denuit, M. and Lambert, P., "Generalized Pareto Fit to the Society of Actuaries' Large Database", *North American Actuarial Journal*, 7:3, 2003, pp. 18-36.

Coles S.G. and Tawn J.A., "A Bayesian Analysis of Extreme Rrainfall Data", *Applied Statistics* 45, 1996, pp. 463-78.

Embrechts, P., Klüppelberg, C. and Mikosch, T., *Modelling Extremal Events for Insurance and Finance*, Berlin: Springer Verlag, 2000.

Hill, B., "A Simple General Approach to Inference about the Tail of a Distribution", *Annals of Statistics,* 3, 1975, pp. 1163-1174.

McNeil, A., "Estimation the Tails of Loss Severity Distributions Using Extreme Value Theory", *Astin Bulletin*, 27, 1997, pp. 117-137.

Pickands, J., "Statistical Inference Using Extreme Order Statistics", *Annals of Statistics*, 3, 1975, pp. 119-131.

## Abstract

In this paper, we introduce a practical method of estimating the threshold over which a heavy-tailed distribution is approximated asymptotically for the underlying distribution of extreme events. We introduce a mixture model of a loss distribution of a certain parametric form below a threshold and a heavy-tailed distribution above the threshold. The number of exceedances over a threshold are considered a random variable for a prior distribution in the Bayesian framework in order to estimate the threshold and corresponding extreme value index. A numerical example is given to illustrate the Bayesian estimation of the parameters by applying the mixture model to losses in medical insurance policies in Korea. About a 10.5% extra charge over traditionally calculated premiums seems necessary to hedge the risk embedded in the heavy-tailed loss distribution.


※ **Key words:** Bayesian estimation, distribution, extreme value distributions, extreme value theory, generalized Pareto mixture distribution, Korean medical insurance, Metropolis-Hastings, threshold