응용통계연구 (2012)
25(1), 101–114
DOI: http://dx.doi.org/10.5351/KJAS.2012.25.1.101

POT방법론을 이용한 자동차보험 손해율 추정

김수영[1] · 송종우[2]

[1]이화여자대학교 통계학과, [2]이화여자대학교 통계학과

(2011년 11월 접수, 2011년 12월 수정, 2011년 12월 재택)

요 약

자동차보험의 손해율이란 지급보험금의 수입보험료에 대한 비율을 의미한다. 손해율이 매우 큰 값을 갖는 대형손실이 일어나는 경우에는 보험회사의 재무적인 부분에 큰 악영향을 미치게 된다. 따라서 보험회사가 이에 대비할 수 있도록 하기 위하여 손해율의 극단 분위수(extreme quantile)를 추정하는 것은 매우 중요한 일이다. 다른 종류의 보험 관련 데이터와 같이 손해율의 분포는 오른쪽으로 긴 꼬리를 갖는 두꺼운 꼬리분포(heavy-tailed distribution)를 갖는다. 이런 자료에서 극단 분위수를 추정하기 위하여 가장 많이 사용되는 방법론은 POT(Peaks over threshold)와 Hill 추정(Hill estimation)이다. 본 논문에서는 일반화파레토분포(generalized Pareto distribution; GPD)의 다양한 모수추정방법론의 성능을 모의실험과 실제 손해율 데이터를 사용하여 비교, 분석하였다. 또한 Hill 추정치를 사용하여 극단 분위수를 추정하였다. 그 결과 대부분의 경우에 POT 방법론이 Hill 추정치를 이용한 방법보다 정확한 분위수를 추정하였고, 모수추정방법론 중에서는 MLE, Zhang, NLS-2 방법론이 가장 좋은 결과를 보여주었다.

주요용어: POT, 일반화파레토분포(GPD), 손해율, Hill 추정량, 극단분위수 추정.

1. 서론

손해율은 지급보험금의 수입보험료에 대한 비율을 말한다. 손해율은 보험회사의 영업수지를 나타내는 대표적인 변수로 정확한 손해율을 추정하는 하는 것이 가능하다면, 보험회사에서 적정한 지급준비금을 준비하여 경영안정을 유지할 수 있도록 도울 수 있다. 특히 보험회사에서 대형손실이 발생하는 경우는 빈번하지는 않지만 대형손실의 발생하게 되는 경우에는 보험회사에 미치는 재무적 영향이 매우 커지게 된다. 따라서 대형손실이 일어나는 경우의 손해율을 보다 안정적으로 추정하여 보험회사가 이를 미리 대처할 수 있도록 하는 것은 매우 중요한 과제중의 하나이다. 대형손실이 일어나는 경우를 보다 정확하게 예측하기 위해서는 우선 적절한 손해율의 분포를 추정해야 한다. 대체적으로 손해율의 분포는 두꺼운 꼬리를 갖는 분포(heavy-tailed distribution)이므로 극단치이론(Extreme value theory; EVT)을 이용하여 극단분위수(extreme quantile)를 추정하여 대형손실의 발생을 추정한다. 손해율 데이터는 넓은 구간에 걸쳐 분포되어 있으므로 전체적으로 모형에 적합시키는 방법보다는 임계점을 초과하는 꼬리부분의 데이터만을 일반화파레토분포(generalized Pareto distribution; GPD)에 근사화시키는 방법을 이용한다. 분위수 추정에 사용되는 방법은 여러 가지 방법이 있지만 그 중에서도 가장 많이 사용되는

이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업이며 (No. 2011-0026070) 제 1저자 이현의의 석사학위논문의 축약본임.
[2]교신저자: (120-750) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 통계학과, 교수.
E-mail: josong@ewha.ac.kr

POT(Peaks over threshold)방법론과 Hill 추정 방법을 이용하여 극단 분위수를 추정하는 것이 본 논문의 목표이다. 우선 모의실험을 통하여 GPD분포에서 생성된 자료를 이용하여 여러 극단적인 분위수를 추정하는 방법들의 정확도를 비교하여 본다. 그 후, 실제 손해율 데이터를 이용한 실증분석을 통해 모의실험에서 사용한 방법들이 얼마나 정확하게 극단 분위수를 추정하는지 알아보도록 한다.

2. 극단치 이론(Extreme value theory)

본 논문에서는 분포의 꼬리부분에 해당하는 대형손실을 추정하는 것에 대하여 관심을 가지고 있으므로 극단 값을 다루는 극단치이론을 이용한다. 극단치분포는 극단치를 정의하는 방법에 따라 두 가지로 나누어진다. 첫째는 전체 구간을 몇 개의 세부구간들(blocks)로 나눈 후 각 구간에서 발생하는 최대값을 극단치로 정의하는 Black Maxima 방법, 둘째는 임계점(threshold)을 초과하는 값들을 극단치로 정의하는 POT방법이다. 일반적으로 극단치 이론의 결과로 얻어지는 확률분포는 각각 일반화 극단치 분포(generalized extreme value distribution; GEV)모형과 일반화파레토분포모형의 두 가지로 구분된다.

2.1. POT(Peaks over threshold)방법

극단치 이론으로부터 두꺼운 꼬리를 갖는 임의의 분포에서 적절한 임계점을 초과하는 관측치들의 분포는 양의 형태모수(shape parameter) ξ를 갖는 일반화파레토분포에 수렴하는 것으로 알려져 있다. 그러므로 여기에서는 특정한 임계점을 초과하는 관측치들에 대해 극단치 분포를 모형화하는 방식인 POT방법을 이용할 수 있다.

확률변수의 수열 $X_1, \ldots, X_n$이 독립동등분포(i.i.d)이며 임의의 분포함수 F로부터 파생되었을 때, POT방법에서는 극단값의 분위수를 추정하기 위하여 임계점(threshold) u를 초과하는 값들의 분포에 대해서 관심을 갖는다. x_0를 임의의 분포 F의 무한 또는 유한의 오른쪽 끝점(right endpoint)이라고 하면, 이는 $x_0 = \sup\{x \in R : F(x) < 1\} \leq \infty$로 표현할 수 있다. 이때, 임계점 u에 대해서 임계치초과분포함수(excess distribution over the threshold)는 다음과 같이 정의할 수 있다.

$$F_u(x) = P[X - u \leq x | X > u] = \frac{F(x+u) - F(u)}{1 - F(u)}, \quad 0 \leq x \leq x_0 - u.$$

즉, $F_u(x)$는 확률변수 X가 임계점 u를 초과할 때 그 초과치가 x보다 작을 조건부 확률을 말한다. 이는 McNeil과 Saladin (1997)에서 확인 할 수 있다.

2.2. 일반화파레토분포(generalized Pareto distribution; GPD)

GPD는 주로 형태모수(shape parameter) ξ와 척도모수(scale parameter) σ를 갖는 다음과 같은 분포함수를 갖는 분포로 정의한다.

$$G_{\xi,\sigma}(x) = \begin{cases} 1 - \left(1 + \xi \frac{x}{\sigma}\right)^{-\frac{1}{\xi}}, & \text{if } \xi \neq 0, \\ 1 - \exp\left(-\frac{x}{\sigma}\right), & \text{if } \xi = 0. \end{cases}$$

이때 $\sigma > 0$이고, 받침(support)은 $\xi \geq 0$인 경우 $x \geq 0$이며, $\xi < 0$일 때 $0 \leq x \leq -\sigma/\xi$이다.

GPD는 ξ값에 따라 분포의 종류를 크게 세 가지로 분류할 수 있다. $\xi > 0$이면 파레토분포, $\xi < 0$이면 제2종 파레토분포(type II Pareto distribution) 그리고 $\xi = 0$인 경우에는 지수분포(exponential

distribution)로 분류할 수 있다. 그리고 위치모수(location parameter) μ를 추가하여 파레토분포를 확장할 수 있는데, 이 경우에는 GPD $G_{\xi,\mu,\sigma}(x)$는 $G_{\xi,\sigma}(x-\mu)$로 정의할 수 있다.

2.3. 일반화극단치분포(Generalized extreme value distribution: standard GEV)

일반화극단치분포는 다음과 같이 설명할 수 있다. 만약 확률변수들 $X_1,\ldots,X_n$이 독립동등분포(i.i.d)이며 임의의 분포함수 F를 가진다고 할 때, 최대확률변수를 $M_n = \max(X_1,\ldots,X_n)$이라고 하자. Fisher와 Tippett (1928)의 정리에 의하면 어떤 실수 $a_n > 0$, b_n으로 표준화된 최대확률변수 $(M_n - b_n)/a_n$는 H라는 비퇴화분포(non-degenerate distribution)로 수렴한다. 즉,

$$P\left\{\frac{(M_n - b_n)}{a_n} \le x\right\} = F^n(a_n x + b_n) \to H(x), \quad \text{as} \quad n \to \infty \tag{2.1}$$

로 나타낼 수 있다. 모든 분포가 전부 비퇴화분포로 수렴하는 것은 아니지만, 대부분의 통계학에서 많이 나오는 분포들은 비퇴화분포로 수렴한다.

이를 정리하자면, 만약 주어진 분포 F에서 표준화된 표본 최대확률변수가 위와 같이 비퇴화분포로 수렴한다면, F는 임의의 ξ를 갖는 극단치분포(extreme value distribution) H_ξ의 MDA(maximum domain of attraction)에 속한다고 말할 수 있고, $F \in \mathrm{MDA}(H)$와 같이 표현한다.

여기에서 일반화극단치분포는 다음과 같은 분포함수를 갖는다.

$$H_\xi(x) = \begin{cases} \exp(-(1+\xi x)^{-\frac{1}{\xi}}, & \text{if } \xi \neq 0, \\ \exp(-e^{-x}), & \text{if } \xi = 0. \end{cases}$$

이때, $1 + \xi x > 0$이다. 세 가지의 극단치분포를 GEV의 특별한 경우로 나타낼 수 있다. $H_\xi(x)$는 ξ에 따라 형태가 결정되는데, $\xi > 0$인 경우에는 프레셰 분포(Frechét distribution), $\xi < 0$인 경우에는 와이블 분포(Weibull distribution), 그리고 $\xi = 0$일 때에는 굼벨 분포(Gumble distribution)에 속하게 된다.

2.4. 분포함수의 수렴

Balkema와 de Haan (1974)과 Pickands (1975)에 의하면 X의 분포함수 F가 GEV의 분포 MDA에 속하는 경우(즉, $F \in \mathrm{MDA}(H_\xi), \xi \in R$)이면, 임계점 u가 오른쪽 끝점에 충분히 가까워지면 $F_u(x)$는 GPD함수 $G_{\xi,\sigma}(x)$에 의해 근사될 수 있다는 것을 다음의 식으로 표현하고 증명하였다.

$$\lim_{u \to x_0} \sup_{0 \le x < x_{0-u}} \left| F_u(x) - G_{\xi,\sigma(u)}(x) \right| = 0.$$

따라서 우리가 자료의 오른쪽 꼬리부분에 관심을 가지고 있을 때, 자료의 분포 F의 종류와 상관없이 임의의 분포는 GPD로 수렴한다는 것을 의미하며 분위수의 추정이 가능하다. 그런데 GPD 모형에 내포된 모수 (σ,ξ)에 대한 추정치는 임계점 u를 어떻게 정하는지에 따라 분산과 편의의 상충관계(variance-bias trade-off)가 나타날 수 있다. 즉, 임계점이 너무 작으면 극단치의 자료수가 많아지므로 모수 추정치의 분산(variance)은 작아지지만, 분포의 점근성이 떨어져서 편의(bias)가 커지게 된다. 반면 임계점을 너무 크게 정하면 편의는 작아지지만 분산은 커지는 현상이 나타날 수 있다. 따라서 적절한 임계점을 결정하는 것은 중요한 문제이고 이를 결정하기 위한 여러 가지 방법들이 있으나 주로 평균초과함수(mean excess function)의 그래프를 사용하여 적절한 임계점을 결정한다.

2.5. 임계점 u의 선택

적절한 임계점 u를 선택하는 방법으로 중 평균초과함수를 이용하는 것이다. 평균초과함수는 다음과 같다.

$$\{(u, e_n(u)), X_{n:n} < u < X_{1:n}\},$$

여기에서 $X_{1:n}$과 $X_{n:n}$은 각각 첫 번째, n번째의 순서통계량이며, $e_n(u)$는 표본평균초과함수(sample mean excess function)로 다음과 같이 정의된다.

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

즉 표본평균초과함수는 임계점을 넘는 값들의 초과분의 합을 임계점을 초과하는 데이터의 개수로 나눈 값이 된다. 표본평균초과함수 $e_n(u)$은 경험적으로 $e(u) = E[X - u|X > n]$로 정의되는 평균초과함수로 추정되며, 평균초과함수는 임계점을 초과하는 값들의 임계점 초과치의 기댓값을 나타낸다. 이 함수와 관련된 설명은 Beirlant 등 (1996), Emvrechts 등 (1997) 그리고 Hogg와 Klugman (1984)에서 자세히 살펴볼 수 있다. 임계점의 값을 다양하게 변화시키면서 경험적 확률초과함수 $e_n(u)$의 값을 구해 이를 분석하여 적절한 임계점을 찾는다. 만약 경험적 평균초과함수가 어떤 점을 초과하는 부분에서 대체적으로 양의 기울기를 갖는 직선으로 나타난다면, 이 점을 임계점으로 지정해 준다. 그리고 이때 임계점을 초과하는 값들은 양의 형태모수를 갖는 GPD를 따른다고 할 수 있다. 또한, GPD를 따르게 되면 평균초과함수는 $\sigma + \xi u > 0$인 경우 $e(u) = (\sigma + \xi u)/(1 - \xi)$로 표현할 수 있다. 임계점을 초과하는 값들의 분포는 $x \geq u$인 경우, $F_u(x - u)$로 정의한다. 그리고 $G_{\xi,\sigma}(x - u) = G_{\xi,u,\sigma}(x)$로 GPD를 근사할 수 있다.

2.6. 분위수의 추정

앞에서 언급한 바와 같이 POT방법은 일정한 임계점을 초과하는 극단치에 대한 확률분포를 모형화하는 방식이다. 따라서 우리는 주어진 자료에서 임계점 u를 추정한 후에 이 임계점을 넘는 관측치들을 이용해서 GPD의 모수 (σ, ξ)를 추정한다. 모수를 추정하는 방법에는 여러 가지 방법이 있지만 MLE, Pickands, Moments, Zhang, 그리고 NLS-2 방법 등을 이용하도록 한다. 일단 모수를 추정한 후에는 추정된 모수를 이용하여 분위수의 추정이 가능하다. Zhang 방법은 Zhang (2007, 2010)에서 NLS-2방법은 Song과 Song (2011)에서 찾아볼 수 있으며 모수를 추정하는 각 방법은 다음과 같이 간단히 설명할 수 있다.

Zhang 방법은 프로파일 가능도함수(profile likelihood function)를 기반으로 경험적 베이즈 방법(empirical Bayesian method)을 이용하는 방법이다. GPD의 모수 (σ, ξ)를 $\theta = \xi/\sigma$로 두고 (θ, ξ)로 재모수화 한다. 그 후 θ와 ξ를 추정한 후, σ를 $\hat{\sigma} = \hat{\xi}/\hat{\theta}$로 추정한다. 우선 θ의 추정치는 다음과 같이 정의한다.

$$\hat{\theta}^*_{NEW} = \sum_{j=1}^{m} w^*_j \theta^*_j \hat{\theta}_j = \frac{n-1}{n+1} X^{-1}_{(n)} + \frac{\sigma^*}{\xi^*}\left[1 - \left(\frac{j - 0.5}{m}\right)^{-\xi^*}\right], \quad j = 1, \ldots, m$$

$$w_j = \frac{L(\theta_j)}{\sum_{t=1}^{m} L(\theta_t)} = \frac{1}{\sum_{t=1}^{m} e^{l(\theta_t) - l(\theta_j)}},$$

여기에서 θ_j는 사전분포(prior distribution)의 $(j-0.5)/m$ 분위수이다 $(j=1,\ldots,m)$.

따라서 $(j-0.5)/m = 1 - F_{\sigma^*,\xi^*}((n-1)/(n+1)X_{(n)}^{-1} - \theta_j^*)$로 표현할 수 있고, $w_j^* = 1/\sum_{t=1}^{m} e^{l(\theta_t^*)-l(\theta_j^*)}$와 $m = 20 + [\sqrt{n}]$으로 표현된다.

위의 식으로부터 추정한 모수는 다음과 같다.

$$\xi_{NEW}^* = n^{-1}\sum_{i=1}^{n}\log\left(1-\hat{\theta}_{NEW}^* X_i\right),$$

$$\hat{\sigma}_{NEW}^* = \frac{\hat{k}_{NEW}^*}{\hat{\theta}_{NEW}^*}.$$

NLS-2방법은 경험적 누적분포함수(empirical cumulative distribution function)를 반응변수로 이론적 누적분포함수(theoretical cumulative distribution function)를 독립변수로 놓고, 잔차제곱합(residual sum of squares; RSS)를 최소화하는 모수 (σ,ξ)를 추정하는 방법이다. 즉 임계점을 초과하는 관측치 $z_1,\ldots,z_{N_u}$에 대하여

$$\arg\min_{(\sigma,\xi)}\sum_{i=1}^{N_u}\left(F_n(z_i)-F_{\xi,u,\sigma}(z_i)\right)^2$$

를 만족하는 모수 (σ,ξ)를 추정하는 방법이다. $\xi \geq 0$인 두꺼운 꼬리분포의 모수를 추정하는 경우에 적합하다.

위의 방법들을 사용하여 추정한 GPD의 모수와 다음의 성질을 이용해서 분위수를 추정할 수 있다. 우선 분포의 꼬리에 위치한 자료들$(x \geq u)$은 다음과 같은 분포를 갖는다.

$$F(x) = p\{X \leq x\} = (1-P\{X \leq u\})F_u(x-u) + P\{X \leq u\}.$$

임계점이 충분히 큰 경우에는 $F_u(x-u)$를 $G_{\xi,u,\sigma}(x)$로부터 추정할 수 있고, $P\{X \leq u\}$는 $F_n(u)$로 추정할 수 있다. 따라서 $x \geq u$인 경우 분포함수 $F(x)$를 추정하기 위해 다음의 꼬리 추정값(tail estimate)를 사용할 수 있다.

$$\widehat{F(x)} = (1-F_n(u))G_{\xi,u,\sigma}(x) + F_n(u)$$

$\widehat{F(x)}$ 역시 GPD이고 형태모수 $\tilde{\xi}$는 $F(x)$의 형태모수 ξ와 동일한 값을 갖고, 척도모수는 $\tilde{\sigma} = \sigma(1-F_n(u))^{\xi}$ 그리고 위치모수 $\tilde{\mu} = u - \tilde{\sigma}((1-F_n(u))^{-\xi}-1)/\xi$를 갖는다.

분포 F와 일반적인 p에 있어 $F^{\leftarrow}$가 일반화역함수(generalized inverse of F)인 경우, $x_p = F^{\leftarrow}(p)$를 나타낸다. 즉, $F^{\leftarrow}(p)$는 다음과 같다.

$$F^{\leftarrow}(p) = \inf\{x \in R : F(x) \geq p\}.$$

이때, 분포를 알고 있다면 $F^{\leftarrow}(p)$를 계산할 수 있지만, 분포를 모르는 경우에는 추정을 해야 한다.

POT 추정량 x_p는 위에 언급한 꼬리 추정 함수의 역함수와 미지의 GPD분포의 모수 대신 추정된 $\hat{\xi},\hat{\sigma}$를 사용하여 얻을 수 있다.

$$x_p = F^{\leftarrow}(p) = G_{\hat{\xi},u,\hat{\sigma}}^{-1}\left(\frac{p-F_n(u)}{1-F_n(u)}\right) = u + \frac{\hat{\sigma}}{\hat{\xi}}\left(\left(\frac{1-p}{1-F_n(u)}\right)^{-\hat{\xi}} - 1\right).$$

여기에서 N_u는 임계점 u를 초과하는 자료의 개수이고 n을 함수 F로부터 파생된 전체 자료 수라고 한다면, 분위수 추정량(quantile estimator)는

$$\hat{x}_p = u + \frac{\hat{\sigma}}{\hat{\xi}}\left(\frac{n}{N_u}(1-p)^{-\hat{\xi}} - 1\right)$$

으로 나타낼 수 있다.

3. Hill 추정량(Hill estimator)

Hill 추정량은 극단치 분포에서 꼬리지수(tail index)를 추정하는 대표적인 비모수 추정방법(nonparametric method)이다. Hill 추정량은 다음의 과정을 이용하여 분위수를 추정할 수 있다. 우선 두꺼운 꼬리를 갖는 X의 분포와 $\log(X)$의 분포는 각각 다음과 같이 정의된다.

$$P(X > x) = \bar{F}(x) = x^{-\alpha} I(x)$$
$$P(\log X > u) = e^{-\alpha u} I(e^u)$$

이때 분위수 함수(quantile function)는

$$Q(1-p) = p^{-\gamma} I^*\left(\frac{1}{p}\right)$$

이 되므로 분위수 함수의 로그값은 다음과 같다.

$$\log Q(1-p) = -\gamma \log(p) + \log I^*\left(\frac{1}{p}\right).$$

그리고 $X_1, X_2, \ldots, X_n$의 순서통계량(order statistics)을 $X_1^* \le X_2^* \le \cdots \le X_n^*$이라고 할 때, X_{n-j+1}^*는 $Q(1-j/(n+1))$의 일치추정량(consistent estimator)이 된다 $(j = 1, \ldots, n)$. 이때, 데이터가 파레토 분포에 가까운지 알아보기 위하여 Hill Plot을 그려 진단한다. Hill Plot은 $\log(X_{n-j+1}^*)$를 y축으로 $-\log(j/(n+1))$를 x축으로 하는 그래프로, Hill 추정치를 구하기 위하여 꼬리 부분의 자료가 직선에 잘 적합 될 때의 직선의 기울기 γ을 구한다.

하지만 그래프로는 정확한 기울기를 추정하는 것이 어렵기 때문에 다음의 기울기 추정치(slope estimator)를 사용하여 결정한다.

$$\hat{\gamma} = \frac{\frac{1}{k}\sum_{j=1}^{k}\left(\log X_{n-j+1}^* - \log X_{n-k}^*\right)}{-\frac{1}{k}\sum_{j=1}^{k}\left(\log\left(\frac{j}{n+1}\right) - \log\left(\frac{k}{n+1}\right)\right)},$$

여기에서 k는 분위수 추정에 사용되는 순서통계량의 개수이다. k가 충분이 크면 분모는 1에 가까워진다. 보다 자세한 설명은 Hill (1975)에서 살펴볼 수 있고 이를 이용한 Hill 추정치는 다음과 같다.

$$H_{k,n} = \frac{1}{k}\sum_{j=1}^{k}(\log X_{n-j+1}^* - \log X_{n-k}^*)$$

이 때 Hill 추정치의 경우에도 k의 선택에 따라 분산과 편의의 상충관계가 발생하게 되므로 적절한 k를 결정하도록 주의를 기울여야 한다.

표 **4.1.** $\xi = 0$, $\sigma = 1$인 GPD의 경우의 RMSE와 ARB

Quantile	0.95	0.99	0.999	0.9999
MLE	0.0385 (0.0101)	0.0810 (0.0139)	0.2383 (0.0269)	0.5555 (0.0465)
Pickands	0.0444 (0.0119)	0.1754 (0.0304)	0.9789 (0.1133)	2.6676 (0.2208)
Moments	0.0383 (0.0100)	0.0811 (0.0139)	0.2385 (0.0268)	0.5547 (0.0462)
Zhang	0.0384 (0.0101)	0.0810 (0.0139)	0.2369 (0.0266)	0.5626 (0.0470)
NLS-2	0.0369 (0.0097)	0.1046 (0.0176)	0.3000 (0.0349)	0.6035 (0.0533)

표 **4.2.** $\xi = 0.5$, $\sigma = 1$인 GPD의 경우의 RMSE와 ARB

Quantile	0.95	0.99	0.999	0.9999
MLE	0.1727 (0.0196)	0.8282 (0.0386)	7.6514 (0.0965)	48.9417 (0.1771)
Pickands	0.1755 (0.0202)	1.7059 (0.0755)	22.3856 (0.2712)	170.8143 (0.5552)
Moments	0.4395 (0.0553)	1.3001 (0.0569)	9.6943 (0.1335)	60.1915 (0.2711)
Zhang	0.1730 (0.0197)	0.8289 (0.0386)	7.6903 (0.0967)	49.4936 (0.1785)
NLS-2	0.1631 (0.0189)	0.9548 (0.0428)	6.9365 (0.0907)	34.9767 (0.1406)

앞에서 구한 Hill 추정치 $H_{k,n}$를 사용해서 추정된 분위수 $Q(1-p)$는 Weissman (1978)에서

$$\hat{Q}^*_{k,n} = X^*_{n-k}\left(\frac{k+1}{(n+1)p}\right)^{H_{k,n}}$$

과 같이 구하였다.

4. 모의실험 방법 및 결과

GPD를 따르는 데이터에서의 분위수 추정방법의 정확성을 살펴보기 위하여 모의실험을 수행하였다. 모의실험의 과정을 단계별로 살펴보면 다음과 같다.

(단계 1) 모수 (ξ, σ)인 GPD를 따르는 독립적인 100,000개의 데이터를 생성한다.

(단계 2) $0 \ll q < p < 1$ 확률을 결정하고, 추정하고자 하는 분위수의 실제값 x_p를 계산한다. q는 단계 3에서 임계점을 결정할 때 사용한다.

(단계 3) 생성된 데이터 중에서 10,000개를 임의로 선택하고 임계점 $u = x_q$을 계산한다.

(단계 4) 임계점을 초과하는 N_u를 지정하고, 각 방법을 이용하여 일반화파레토 분포의 모수 ξ와 σ를 추정한다.

(단계 5) POT방법으로 분위수 추정량 $\hat{x}_p$를 추정한다.

(단계 6) 단계3–단계5를 100번 반복하여 추정된 분위수의 평균제곱근오차(root mean square error; RMSE)와 절대상대편향(absolute relative bias; ARB)를 구하였다.

이때, GPD를 따르는 확률변수는 척도모수 σ는 1로 형상모수 ξ는 각각 0, 0.5, 1일 때의 3가지의 경우에 대하여 분석하였다. 임계점은 추출된 10,000개의 표본의 90% 분위수로 정하였다. 그리고 ARB는 $|\hat{\theta} - \theta|/\theta$로 정의되며 $\hat{\theta}$는 분위수의 추정치이고 θ는 실제 분위수의 값이다. 결과는 다음 표 4.1, 표 4.2, 표 4.3과 같다.

95%, 99.99% 분위수를 추정하는 경우에는 NLS-2방법이 실제 분위수를 가장 잘 추정하고 MLE와 Zhang방법도 실제 분위수와 비슷한 값을 추정하는 것을 알 수 있다. 99% 분위수와 99.9% 분위수

표 **4.3.** $\xi = 1$, $\sigma = 1$인 GPD의 경우의 RMSE와 ARB

Quantile	0.95	0.99	0.999	0.9999
MLE	0.7996 (0.0322)	8.624 (0.0672)	255.343 (0.1890)	4719.08 (0.3104)
Pickands	0.8533 (0.0356)	18.037 (0.1354)	743.551 (0.5116)	19169.23 (1.0686)
Moments	34.0406 (1.4608)	147.335 (1.0587)	456.189 (0.3930)	7219.33 (0.7246)
Zhang	0.7978 (0.0321)	8.635 (0.0672)	256.777 (0.1901)	4750.99 (0.3125)
NLS-2	0.7685 (0.0313)	9.028 (0.0697)	214.080 (0.1638)	3282.38 (0.2425)

표 **4.4.** Hill 추정치를 이용한 경우의 RMSE와 ARB (단, $\sigma = 1$)

	0.95	0.99	0.999	0.9999
$\xi = 0$	0.1171 (0.0368)	0.2575 (0.0513)	3.3111 (0.4722)	12.433 (1.3343)
$\xi = 0.5$	0.2686 (0.0330)	1.2073 (0.0539)	23.0390 (0.3580)	172.220 (0.8239)
$\xi = 1$	0.8755 (0.0383)	10.0591 (0.0797)	277.5039 (0.2205)	4418.528 (0.3061)

추정에서도 MLE, Zhang, NLS-2 방법이 실제 분위수와 비슷한 값을 추정한다는 것을 알 수 있다. Pickands 방법은 99.9% 또는 99.99%의 분위수를 추정하는 경우에는 적합하지 않은 것을 알 수 있다. Moments방법은 실제 $\xi = 0$인 경우 이외에는 실제 분위수와 추정된 분위수가 큰 차이가 있는 것을 알 수 있다.

Hill 추정지를 이용하여 분위수를 추정한 경우에는 POT방법을 이용하였을 때보다는 정확도가 떨어지는 것을 확인할 수 있다. 그리고 위치모수 ξ의 값이 1에 가까워질수록 더 정확도가 떨어진다.

5. 실증분석

본 장에서는 우리나라 자동차 보험회사의 손해율을 이용하여 실증분석을 실시하였다. 먼저 손해율은 지급보험금의 수입보험료에 대한 비율을 말한다. 지급보험금만으로는 정확한 보험사의 손해를 측정할 수 없기 때문에 손해율을 사용한다. 즉, 지급보험금이 매우 크더라도 보험 가입자가 납부한 보험료가 지급한 보험료보다 큰 경우에는 보험사는 이익을 얻을 수 있고, 지급보험금이 적더라도 보험료가 이보다 적게 되면 보험사는 손해를 볼 수 있다. 따라서 이런 경우에는 손해액만으로는 비교가 어렵기 때문에 손해율을 사용하여 보험사의 손익을 비교 하게 된다. 사고가 일어나지 않는 경우에는 지급하는 보험금이 없으므로 손해율은 0%로 나타나나고, 보험금과 보험료가 동일한 경우에는 손해율이 100%로 나타난다. 따라서 손해율이 100%이하인 경우에는 보험사에서는 이익을 보게 되고, 손해율이 100%를 초과하면 보험사는 손해를 보게 된다.

손해율의 분위수를 POT방법을 이용하여 추정하기 위해서 한국 자동차 보험회사의 실제 자료를 이용하여 분석하였다. 원 자료는 50만 개 이상의 데이터를 포함하고 있었으나 데이터 정제(data cleaning)를 통해 입력오차와 같은 자료처리오차를 포함하는 자료를 삭제하였다. 예를 들어, 보험가입자의 연령이 만 15세로 기록된 경우(운전나이 제한인 만 18세 보다 작은 값을 갖는 경우), 보험료가 음의 값을 갖는 경우, 또는 성별이 남자나 여자가 아닌 경우의 자료는 삭제하였다. 데이터 정제 과정을 통해 얻은 138,125개의 자료를 이용하여 분석을 하였다 (표 5.1).

그림 5.1의 손해율의 히스토그램를 살펴보면 극단적인 값들이 존재하기 때문에 뚜렷한 분포형태를 찾을 수가 없다. 따라서 손해율에 log를 취하여 그래프를 그려 분포를 알아보았다. 손해율은 최소값이 0%, 중간값이 10.28%, 평균값이 143.4%, 최대값이 173,900%이다. 0%인 자료는 전체 138,125개 중에서 3,021개로 전체의 21.87%를 차지한다. 그리고 최대값이 매우 크며 평균값이 중간 값보다 매우 큰 값을 가지므로 손해율의 분포는 오른쪽으로 꼬리가 매우 긴 분포라고 볼 수 있다.

표 **5.1.** 손해율 데이터의 기초통계량

최소값	제 1사분위수	중간값	평균	제 3사분위수	최대값
0	4.58	10.28	143.40	99.04	173900

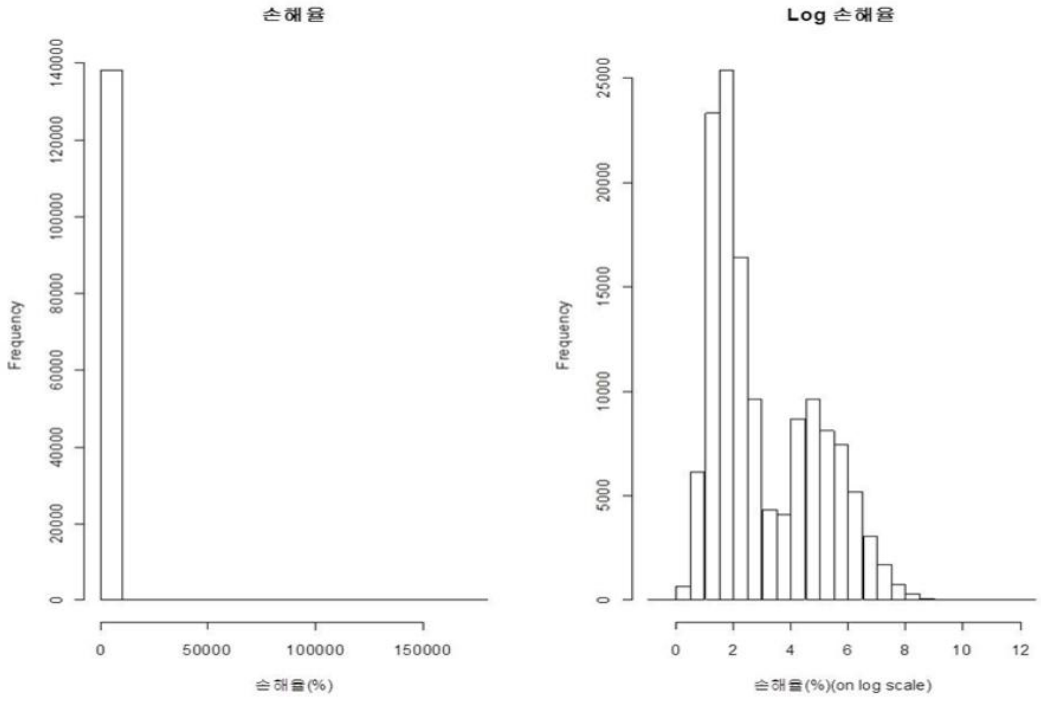

그림 **5.1.** 손해율의 히스토그램

5.1. 임계점 설정

손해율 자료를 POT방법을 이용하여 분위수를 추정하려면 우선 적절한 임계점의 값을 정해야 한다. 따라서 앞에서 언급한 평균초과함수 그래프를 그려보았다 (그림 5.2).

손해율 자료의 평균초과함수 그래프는 임계점이 매우 작은 경우를 제외하면 대체로 양의 기울기를 갖는 직선으로 볼 수 있다. 그래프의 실선은 손해율의 90% 분위수 값을 임계점 u로 지정한 경우이다. 그래프를 통해서는 정확한 임계점을 값을 알아내기 어렵지만 90% 분위수 이상에서 충분히 GPD분포를 따른다고 여기지기 때문에 90% 분위수 이상의 값을 임계점으로 설정하기로 하였다. 그리하여 임계점의 값을 손해율 자료의 90% 분위수, 92% 분위수, 94% 분위수, 96% 분위수, 96% 분위수로 지정하였다. 각 임계점 값과 임계점을 초과하는 데이터의 수(N_u)는 다음의 표 5.2에서 살펴볼 수 있다.

5.2. GPD 모형 하에서의 분위수 추정

앞에서 지정한 바와 같이 손해율 데이터의 90% 분위수, 92% 분위수, 94% 분위수, 96% 분위수, 96% 분위수의 값을 임계점으로 정하여 95%, 99%, 99.9% 99.99%의 분위수의 값을 추정하였다. 표 5.3에서는 각 임계점에 따른 분위수의 추정치가 실제 분위수 값을 비교하기 위한 절대오차(absolute error)와 절대

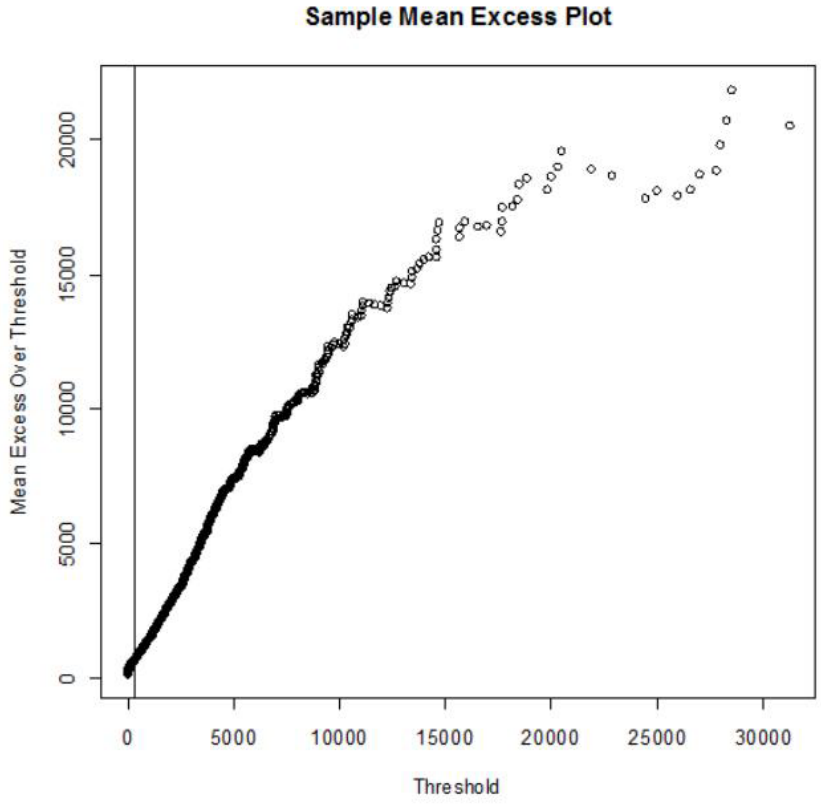

그림 **5.2.** 손해율의 평균초과함수

표 **5.2.** 임계점 값과 N_u

임계점		N_u
90% 분위수	335.1717	13813
92% 분위수	410.4140	11050
94% 분위수	519.2675	8288
96% 분위수	711.9452	5525
98% 분위수	1160.5950	2763

상대편향을 보여준다. 실제 data의 분위수를 계산해서 그 값을 참값으로 두고, 여러 가지 분위수 추정방법을 사용해서 나온 추정치와 비교하여 절대오차와 절대상대편향을 계산하였다.

여기에서 99.99% 분위수의 경우에 모형이 좋은 경우라면 손실 중 매 10,000개마다 1개의 손실이 해당 분위수를 초과한다고 예상할 수 있다. 이런 손실은 그 수는 매우 드물지만, 재무적으로 매우 큰 손실을 가져오므로 보험회사의 입장에서는 충분히 위협적일 수 있다. 표 5.3에서 분위수의 추정치는 임계점의 선택에 의존한다는 사실을 보여주고 있다. 임계점이 90%~95% 분위수인 경우에는 Pickands방법이 95% 분위수와 99% 분위수를 실제값과 가장 가깝게 추정하고, 99.9% 분위수와 99.99%분위수를 추정하는 경우에는 MLE와 NLS-2방법이 좋은 결과를 가져옴을 알 수 있다. 그리고 96%와 98% 분위수를 임계점으로 정한 경우에는 MLE방법과 NLS-2방법이 가장 좋은 결과를 가져오는 것을 알 수 있다.

표 **5.3.** 임계점에 따른 손해율 자료의 분위수 추정: 절대오차(Absolute error)와 절대상대편향(Absolute relative bias; ARB)

u		분위수			
		0.95	0.99	0.999	0.9999
90% 분위수	MLE	0.7977 (0.0013)	18.273 (0.01040)	120.684 (0.0170)	6999.52 (0.224)
	Pickands	0.0366 (0.0001)	55.997 (0.03170)	318.102 (0.0460)	4180.20 (0.134)
	Moments	40.6519 (0.0675)	112.753 (0.06390)	614.255 (0.0890)	12175.33 (0.389)
	Zhang	0.8140 (0.0014)	18.179 (0.01030)	121.148 (0.0170)	7001.41 (0.224)
	NLS-2	6.6854 (0.0111)	177.094 (0.10030)	2584.497 (0.3730)	14045.28 (0.449)
92% 분위수	MLE	0.5987 (0.0010)	17.586 (0.01000)	133.148 (0.0190)	7083.74 (0.226)
	Pickands	0.6316 (0.0010)	63.517 (0.03600)	482.780 (0.0700)	2941.05 (0.094)
	Moments	32.1871 (0.0535)	129.363 (0.07330)	522.449 (0.0750)	11915.21 (0.381)
	Zhang	0.4706 (0.0008)	18.105 (0.01030)	134.055 (0.0190)	7101.47 (0.227)
	NLS-2	5.0367 (0.0084)	122.630 (0.06950)	1873.277 (0.2700)	8547.14 (0.273)
94% 분위수	MLE	1.1200 (0.0019)	25.266 (0.01430)	248.729 (0.0360)	8216.10 (0.263)
	Pickands	0.0880 (0.0001)	7.718 (0.00440)	615.647 (0.0890)	10709.24 (0.342)
	Moments	12.8410 (0.0213)	148.935 (0.08440)	387.300 (0.0560)	11525.03 (0.368)
	Zhang	1.2650 (0.0021)	24.240 (0.01370)	242.894 (0.0350)	8148.17 (0.260)
	NLS-2	0.0590 (0.0001)	75.347 (0.04270)	1207.408 (0.1740)	3527.81 (0.113)
96% 분위수	MLE	-	25.231 (0.01430)	218.359 (0.0320)	7971.87 (0.255)
	Pickands	-	0.233 (0.00013)	1203.203 (0.1740)	14888.07 (0.476)
	Moments	-	162.026 (0.09181)	227.688 (0.0330)	11047.49 (0.354)
	Zhang	-	21.753 (0.01233)	211.372 (0.0300)	7826.18 (0.250)
	NLS-2	-	24.842 (0.01408)	380.746 (0.0550)	2522.77 (0.081)
98% 분위수	MLE	-	2.118 (0.00120)	53.820 (0.0078)	5064.36 (0.162)
	Pickands	-	0.233 (0.00013)	234.104 (0.0338)	6712.27 (0.215)
	Moments	-	145.602 (0.08250)	31.308 (0.0045)	10206.24 (0.326)
	Zhang	-	3.884 (0.00220)	57.052 (0.0082)	4784.05 (0.153)
	NLS-2	-	0.061 (0.00003)	390.089 (0.0563)	8088.42 (0.258)

5.3. Hill 추정치를 이용한 분위수 추정

그림 5.3은 손해율 데이터의 Hill plot이다. 그림에서 점선은 90% 분위수로 실선은 94% 분위수를 임계점으로 정한 경우이다. 94% 분위수를 임계점으로 정한 경우에 임계점을 초과하는 부분들이 직선에 더 적합한 것을 알 수 있다. POT방법으로 추정한 것과 같이 90%, 92%, 94%, 96%, 98% 분위수의 값을 임계점으로 하는 경우 Hill 추정치를 이용하여 극단 분위수를 추정하였고, 각 임계점에 따른 절대오차와 절대상대편향 값은 표 5.4와 같다.

표 5.4에서는 그림 5.3에서 확인 했던 것처럼 임계점의 분위수가 커질수록 일반적으로 오차가 줄어드는 것을 알 수 있다. 그리고 대체로 POT방법보다 추정의 정확도는 떨어지지만, 임계점이 98%인 경우의 99.9% 분위수 추정치는 POT방법보다 정확하게 추정된 것을 알 수 있다.

6. 결론

지금까지 POT방법과 Hill 추정치를 이용하여 꼬리가 두꺼운 분포의 극단 분위수를 추정해보았다. 모의실험과 손해율 데이터를 이용하여 분위수를 추정해 본 결과, 대부분의 경우에 POT 방법을 이용하여 추정한 분위수의 값이 Hill 추정치를 이용하여 추정한 분위수보다 정확도가 높다는 것을 알 수 있었다.

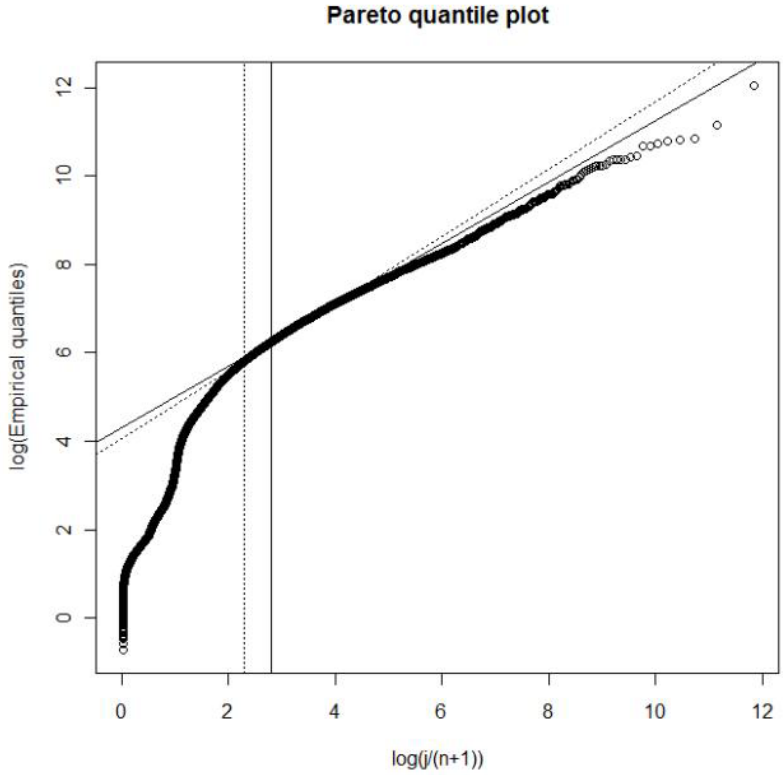

그림 **5.3.** 손해율의 Hill Plot

표 **5.4.** Hill 추정치를 이용한 분위수 추정: 절대오차(Absolute error)와 절대상대편향(Absolute relative bias; ARB)

u(분위수)	분위수			
	0.95	0.99	0.999	0.9999
90%	33.180 (0.0551)	178.873 (0.1014)	4336.93 (0.6254)	34069.7 (1.0888)
92%	24.443 (0.0406)	97.530 (0.0553)	3005.43 (0.4334)	21759.7 (0.6954)
94%	12.506 (0.0208)	44.874 (0.0254)	2068.74 (0.2983)	13498.6 (0.4314)
96%	-	2.283 (0.0013)	1008.89 (0.1455)	4506.8 (0.1440)
98%	-	8.386 (0.0048)	17.19 (0.0025)	3778.4 (0.1207)

POT 방법론을 사용하기 위해서는 GPD의 형태모수 ξ와 척도모수 σ의 추정이 우선적으로 이루어져야 한다. 우리는 다양한 모수추정 방법론을 이용하여 극단분위수를 추정하고 그 결과를 비교해보았다. 모의실험에서는 NLS-2, Zhang, MLE방법론의 결과가 가장 우수했고, 자동차 손해보험율의 추정에서는 임계점이 90%∼95% 분위수인 경우에는 Pickands와 Zhang 방법이, 96%∼98%를 임계점으로 사용한 경우에는 NLS-2와 MLE 방법이 가장 좋은 결과를 보여주었다. 즉, 이론적인 데이터에서와 실제 데이터에서의 가장 정확한 추정을 하는 방법이 다르게 나타났다. 이는 모의실험에서는 실제 단일한 GPD를 따르는 분포를 생성하여 분위수를 추정하였지만, 실제 손해율 데이터는 단일한 GPD분포를 따르는 것이 아니라 여러 가지의 분포가 섞여 있을 가능성이 크기 때문이라고 여겨진다. 그리고 추정 방법뿐만 아니라 임계점의 선택에 따라서도 결과가 달라지는 것도 확인할 수 있었다.

참고문헌

Balkema, A. and de Haan, L. (1974). Residual life time at great age, *Annals of Probability*, **2**, 792–804.

Beirlant, J., Teugels, J. and Vynckier, P. (1996). Practical analysis of extreme values, *Leuven University Press*, Leuven.

Emvrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modeling Extremal Events for Insurance and Finance*, Springer Verlag, Berlin.

Fisher, R. and Tippett, L. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample, *Proceedings of the Cambridge Philosophical Society*, **24**, 180–190.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution, *Annals of Statistics*, **3**, 1163–1174.

Hogg, R. and Klugman, S. (1984). *Loss Distributions*, Wiley, New York.

McNeil, A. J. and Saladin, T. (1997). The Peaks over thresholds method for estimating high quantiles of loss distribution, *Proceedings of 28^{th} international ASTIN Colloquium*.

Pickands, J. (1975). Statistical inference using extreme order statistics, *Annals of Statistics*, **3**, 119–131.

Song, J. and Song, S. (2011). A quantile estimation for massive data with generalized pareto distribution, *Computational Statistics and Data Analysis*, **56**, 143–150.

Weissman, I. (1978). Estimation of parameters and larger quantile based on the k largest observations, *Journal of the American Statistical Association*, **73**, 812–815.

Zhang, J. (2007). Likelihood moment estimation for the generalized pareto distribution, *Australian and New Zealand Journal of Statistics*, **49**, 69–77.

Zhang, J. (2010). Improving on estimation for the generalized pareto distribution, *Technometrics*, **52**, 335–339.

Estimation of Car Insurance Loss Ratio Using the Peaks over Threshold Method

S.Y. Kim[1] · J. Song[2]

[1]Department of Statistics, Ewha Womans University
[2]Department of Statistics, Ewha Womans University

(Received November 2011; Revise December 2011; Accepted December 2011)

Abstract

In car insurance, the loss ratio is the ratio of total losses paid out in claims divided by the total earned premiums. In order to minimize the loss to the insurance company, estimating extreme quantiles of loss ratio distribution is necessary because the loss ratio has essential profit and loss information. Like other types of insurance related datasets, the distribution of the loss ratio has heavy-tailed distribution. The Peaks over Threshold(POT) and the Hill estimator are commonly used to estimate extreme quantiles for heavy-tailed distribution. This article compares and analyzes the performances of various kinds of parameter estimating methods by using a simulation and the real loss ratio of car insurance data. In addition, we estimate extreme quantiles using the Hill estimator. As a result, the simulation and the loss ratio data applications demonstrate that the POT method estimates quantiles more accurately than the Hill estimation method in most cases. Moreover, MLE, Zhang, NLS-2 methods show the best performances among the methods of the GPD parameters estimation.

Keywords: Peaks Over Threshold(POT), Generalized Pareto distribution(GPD), Loss ratio, Hill estimator, extreme quantile estimation.

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2011-0005536).
[2]Corresponding author: Associate professor, Department of Statistics, Ewha Womans University, Seoul 120-750, Korea. E-mail: josong@ewha.ac.kr