Journal of the Korean Data & Information Science Society
2009, **20**(3), 515–526

한국데이터정보과학회지

다양한 모형화를 통한 자동차 보험가격 산출

김명준[1] · 김영화 [2]

[1]삼성화재해상보험주식회사 · [2]중앙대학교 자연과학대학 통계학과

접수 2009년 2월 16일, 수정 2009년 5월 20일, 게재확정 2009년 5월 23일

요 약

자동차 보험 산업에 있어 담보되는 위험도에 따른 적정 보험가격의 산출은 매우 중요하다. 본 논문에서는 자동차 보험 산업에서 보험가격 산출 방법이 어떻게 발전되어 왔는지에 대하여 고찰하고, 여러 통계적인 방법으로 산출한 보험가격과 실제 담보되는 위험도의 비교 분석을 통하여 보다 나은 통계적 보험가격 산출 방법을 제안하고자 한다. 그 중에서 일반선형모형을 중심으로 다루었으며, 오차항의 분포에 대한 다양한 가정을 통하여 최적의 접근 방법에 대한 논의를 하였다. 일반선형모형에 있어 오차항의 분포에 대한 적절한 가정은 모형의 최적화를 위한 중요한 가정이다. 본 연구에서는 일반적으로 널리 사용되지 않았음에도 불구하고 자동차 보험 사고 손해액과 매우 유사한 성격을 가지고 있는 트위디 분포를 오차항의 분포 가운데 하나로 적용하여 비교하였다. 실증자료 분석으로서 국내 자동차 보험사의 실제 자료를 통하여 여러 접근 방법에 대한 적정성 비교를 수행하였다.

주요용어: 보험가격, 일반선형모형, 자동차 보험, 트위디 분포.

1. 서론

1.1. 연구 목적

최근 국내 자동차 보험업계에서는 가격 경쟁이 갈수록 심화되고 있다. 고객의 입장에서 보면 일반적인 장기 보험의 경우는 계약 당시에만 보험료를 확인하고 여러 해를 해당 보험료로 담보 받는 반면, 자동차 보험의 경우 대부분 계약은 1년 단위로 체결되기 때문에 고객은 매년 새로운 보험료를 적용받게 된다. 또한 최근에는 인터넷의 발달에 따라 인터넷 상의 보험료 비교 견적을 통하여 여러 보험사의 보험료를 쉽게 비교해 볼 수도 있다. 이러한 비교 견적이 용이해짐과 더불어 온라인 보험사가 등장하면서, 보험모집인이나 보험설계사를 통하지 않고 인터넷이나 전화를 이용하여 상대적으로 저렴한 가격으로 보험가입이 가능해졌고 그 비중은 매년 증가하고 있는 추세이다.

이와 같은 이유로 인하여 자동차 보험회사는 경쟁사와 비교하여 보다 저렴한 가격을 제시함으로써 회사의 계약 규모를 늘려 성장하기를 원하고 또한, 이와 더불어 회사의 이익 창출을 동시에 기대하고 있다. 따라서 각 세부 계층별로 해당하는 위험도에 부합하는 적정한 보험료를 산출하는 것이 자동차 보험 산업 분야에서는 가장 중요한 요소 중의 하나로 부각되고 있으며, 자동차 보험 산업이 발달되어 있는 영국과 미국 등지에서도 이와 관련한 연구가 활발히 진행되어 왔고, 현재에도 계속하여 연구 개발 중에 있다.

대표적으로 Bailey와 Leroy (1960), Bailey (1963)는 전통적인 가격 산출 방법인 단변량 분석의 한계성을 인식하여 편의 (bias)를 최소화하는 방법을 소개하여 다변량 방법의 접근을 시도하였다. 이후 제

[1] (100-782) 서울특별시 중구 을지로 1가 87, 삼성화재해상보험주식회사, 자동차상품파트, 과장.
[2] 교신저자: (156-756) 서울특별시 동작구 흑석동 221, 중앙대학교 자연과학대학 통계학과, 부교수.
E-mail: gogators@cau.ac.kr

한적이었던 컴퓨팅 능력이 급속도로 발전하면서 다변량분석 방법 등 고급 통계를 활용하여 대용량 데이터를 분석하는 것이 가능해졌고, 이로 인하여 적정한 보험료를 산출하는 기법들이 급속하게 발전하는 토대가 마련되었다. 이에 따라, Jørgensen과 Paes de Souza (1994)는 GLM 방법 적용에 있어 자동차 보험 데이터에 사고 빈도와 심도에 각각 적합한 오차항의 분포를 가정하는 방법을 제안하였고, 이는 Smyth와 Jørgensen (2002)에 의하여 더욱 깊이있게 발전적으로 논의되었다.

본 논문에서는 자동차 보험이 가장 선진화되어 있는 영국과 미국 등에서 보험료를 산출하는 방법이 어떻게 발전되어 왔는지를 검토하여 보고, 상대적으로 보험료 산출 기법이 선진국에 비해 다소 뒤쳐져 있는 국내에서 어떠한 통계적 분석 방법의 활용이 가능한지와 어떤 방법을 통한 접근이 보다 효율적인지를 제시하고자 한다.

본 논문의 구성은 다음과 같다. 제1장에서는 자동차 보험료를 산출하는 다양한 방법에 대하여 살펴보고, 제2장에서는 Tweedie 분포의 특성과 이것이 자동차 보험료 산출에 어떻게 적용될 수 있는지를 논의한다. 제3장에서는 소지역 추정으로 보험금을 추정했던 Kim과 Kim (2009)에서 다루었던 국내 자동차 보험사의 실제 자료를 근거로 다양한 분석 방법을 통하여 얻어진 결과를 비교함으로써 어떠한 통계적 분석 방법이 국내 자동차 보험료 산출에 보다 효율적인지 제안한다.

1.2. 대표적인 자동차 보험료 산출 방법

1.2.1. 단변량 분석 방법

자동차 보험 산업에서 보험료를 산출하는데 있어 가장 오랫동안 사용되어온 가장 전통적인 방법으로 계약에 해당하는 요율요소에 대하여 하나의 변수마다 그 위험도를 측정하는 가장 단순한 방법이다. 사용하기 가장 간편하고 쉬운 장점이 있는 반면, 요율요소 사이에 상관관계가 존재할 때 산출된 상대도가 왜곡될 가능성이 있다는 단점을 가지고 있는 방법이다. 다음 표 1.1의 예를 통하여 간단한 보험료 산출 결과와 단변량 방법 적용 시 요율요소 사이의 상관관계로 인하여 발생 가능한 오류를 살펴보기로 한다.

표 1.1 성별, 결혼 여부에 따른 유효대수와 손해액

성 별	결혼여부	유효대수	손해액	사고건수
남	미혼	200대	800만원	80건
	기혼	100대	200만원	20건
여	미혼	100대	200만원	20건
	기혼	200대	200만원	20건
계		600대	1,400만원	140건

주) 심도(사고당 발생하는 손해액)는 동일한 것으로 가정.

단변량 방법을 적용하여 보험료를 산출하는 방법은 성별과 결혼유무를 따로 분리하여 보험료를 산출하는 방법이며, 표 1.1의 성별에 대하여 남성 300대의 유효대수에 전체 손해액은 1,000만원, 여성 300대의 유효대수에 400만원의 손해액으로 설정하는 것이다. 따라서 남자의 경우 평균 약 33,000(1,000만원/300대)원의 손해가, 여자의 경우 13,000원의 손해가 발생한다. 즉, 남성의 경우 여성에 비하여 2.5배의 높은 위험도를 가진 계층으로 분류되어 그 위험도에 비례하는 2.5배의 보험료가 부과되는 것이다. 동일한 방법으로 결혼유무에 대하여서도 미혼자가 기혼자보다 2.5배의 위험도를 나타내어 해당 위험도에 따른 보험료가 부과되게 된다. 이렇게 산출된 상대도를 바탕으로 '기혼여성'을 기준 계층으로 상대도 '1'을 부여하였을 때, '미혼남성'은 6.25배(2.5×2.5), '기혼남성'은 2.5배, '미혼여성'은 2.5배의 상대도가 적용되게 되는 것이다. 즉, 전체적인 손해를 보험사가 담보하면서 해당 계층의 위험도에 따라 위험도가 낮은 계층은 적은 보험료를 위험도가 높은 계층은 높은 보험료를 적용하여 전체 손해의 규모를 일치시키는 방법이다.

그러나 표 1.1에서 쉽게 확인이 가능한 것처럼, '미혼여성'의 경우 '기혼여성'에 비하여 2배의 위험도만을 가지고 있다 (200/100 : 200/200 = 2 : 1). 따라서 단변량으로 각 요율요소별 위험도를 측정하게 되는 경우, 어떤 계층은 실제 위험도보다 때로는 높은 보험료를, 때로는 낮은 보험료를 적용받아 실제 위험도에 따라 적절한 보험료가 적용되어야 하는 형평성의 원칙을 지키지 못하는 결과를 초래하게 되는 경우가 발생할 수 있는 것이다.

1.2.2. Bailey의 최소편의 절차

앞서 살펴본 전통적인 단변량 방법의 단점을 극복하기 위하여 Bailey에 의하여 제안된 것이 최소편의 절차 (minimum bias procedure)이다. 이 방법은 먼저 각 변수간의 상대도를 산출한 후, 산출된 상대도가 맞다는 가정하에서 다른 변수의 상대도를 산출하고, 다시 그 결과를 고정한 후 앞서 구한 상대도를 재산출하는 과정을 반복 실행하여 어느 지점에 수렴하게 될 때 각 변수별 상대도를 산출하게 되는 방법이다.

단변량 분석 방법에서 제시한 예제와 유사하게 두 개의 요율요소에 대하여 다음과 같은 구조하에서 보험료가 산출되는 방법에 대하여 살펴보기로 한다.

X_i : 첫 번째 요율요소의 수준i에서의 상대도$i = 1, 2, \ldots, A$

Y_j : 두 번째 요율요소의 수준j에서의 상대도$j = 1, 2, \ldots, B$

N_{ij} : 해당 셀의 유효 대수

L_{ij} : 해당 셀의 손해액

위와 같은 구조하에서, 발생된 손해액과 향후 적용되는 보험료가 동일하도록 설정하는 수지 균등의 원칙에 의한 보험료 산출식은 다음과 같다

$$\sum_j N_{ij} L_{ij} = \sum_j N_{ij} X_i Y_j. \tag{1.1}$$

즉, 이 방법은 좌변항에서 계산되는 손해액의 총합이 각 계약별로 해당 상대도를 적용했을 경우의 보험료의 합과 동일하도록 산출하는 방법인 것이다. 식 (1.1)을 통하여 X_i 에 대한 상대도를 다음과 같이 계산한다.

$$X_i = \frac{\sum_j N_{ij} L_{ij}}{\sum_j N_{ij} Y_j.} \tag{1.2}$$

식 (1.2)에서 미지수는 상대도 X_i, Y_j 두 개이고 식은 한 개이기 때문에, 상대도 Y_j 의 값으로 단변량 방법에서 산출된 값을 사용하고 X_i 값을 계산한다. 이렇게 계산된 X_i 값을 다시 식에 적용하여 Y_j 값을 다시 계산하고, X_i 와 Y_j 의 결과값이 수렴할 때까지 이러한 과정을 반복 수행하여 상대도의 최종값을 계산한다.

이 방법은 각 변수가 갖는 상대도의 효과를 사전에 가정하여 다른 변수의 상대도를 계산할 때 제거하였기 때문에 단변량에서 제기되었던 변수 사이의 상관관계를 어느 정도 고려할 수 있는 장점을 가지고 있다. 그러나 이 방법으로 인하여 단변량이 가지는 단점을 어느 정도 보완하기는 하였지만, 변수 사이에 존재하는 상관관계의 부분적인 반영이라는 점과 보험료 산출에 적용되는 변수의 수와 그 수준의 수가 증가함에 따라 산출되는 상대도가 수렴하지 않는 현상이 발생하는 제한성을 갖게 된다. 또한 여러 변수에 대한 적용방안은 Feldblum과 Brosius (2002)가 제시한 실제 사례를 통하여 확인할 수 있다.

1.2.3. GLM을 이용한 다변량 분석 방법

다변량 분석 방법 가운데 널리 사용되는 분석 방법인 GLM이 언제부터 자동차 보험에 실제로 적용되어 오고 있는지는 정확하게 판단하기는 어려우나, Jørgensen과 Paes de Souza (1994)에 의하면 1990년도 초반부터 이미 영국 등 보험 선진국이라 할 수 있는 유럽 국가들에서 이미 연구되어 적용되고 있음을 알 수 있다. 또한 Murphy 등 (2000)의 발표 자료를 통하여 선진 통계 기법의 적용에 대한 연구가 상당 부분 진전되어 있음을 알 수 있으며, 매년 미국에서 개최되는 위험도 산정 (rate making) 세미나 발표 자료를 통하여서도 자동차 보험료 산출에 있어 GLM이 미국에서도 1990년대 중반부터 이미 적용되기 시작한 것을 확인할 수 있다.

자동차 보험 가격 산출에 있어 GLM 방법이 급속도로 활용하기 시작된 현실적인 이유는 대용량 데이터를 다룰 수 있는 컴퓨터의 능력이 향상된 것이라 할 수 있다. 1990년대 이전에는 대용량 데이터의 활용이 필수 불가결한 자동차 보험료 산출에 계산이 복잡한 다변량 방법을 적용하는데 인프라가 부족한 것이 사실이었다.

또 다른 이유 중의 하나는, 자동차 보험 사고 데이터의 분포가 GLM에서 가정할 수 있는 오차항의 분포와 매우 흡사한 형태를 가지고 있다는 것이다. 먼저 사고의 심각성을 나타내는 심도 지표는 대부분의 데이터가 소액에 집중되어 있으면서 오른쪽으로 꼬리가 긴 분포의 형태를 가지게 되는데 이는 감마분포 가정을 통하여 적합한 모형을 설정할 수 있으며, 또한 사고의 빈번함을 나타내는 빈도 지표는 빈도의 정의가 일정 기간 동안 발생하는 사고 건수이므로 포아송분포의 형태로 모형을 적합시킬 수 있게 된다.

이와 같이 GLM 방법은 기술적인 발달과 학문적인 바탕이 결합되어 자동차 보험료 산출 분야에 1990년대 이후 급속도로 활용되기 시작한다. 여기서는 자동차 보험료 산출에 사용되는 GLM 방법의 기본 개념만 간단히 설명하고 다음 장에서 Tweedie 분포의 적용과 관련하여 자세하게 다루기로 한다.

Y 를 손해액이라 가정하고 X_1, X_2, X_3, X_4 를 현재 사용하고 있는 요율요소라 하면, $g^{-1}(\cdot)$ 를 연결함수 (link function)라 할 때 모형은 다음과 같이 정의된다.

$$g^{-1}(y) = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + error. \tag{1.3}$$

여기서 오차항의 분포를 감마분포로 가정하고 이 오차항을 최소화하는 a, b_1, b_2, b_3, b_4 를 추정하며, 예를 들어, 연결함수가 주로 사용되는 로그함수인 경우 산출된 결과는 다음과 같이 활용된다.

$$Y = e^a e^{b_1X_1} e^{b_2X_2} e^{b_3X_3} e^{b_4X_4}. \tag{1.4}$$

즉, e^a 가 Y 손해액의 기본보험료가 되고, 이를 각 요율요소별로 산출된 결과의 곱의 형태로 적용하게 되는 것이다.

GLM 방법을 이용하면 Bailey의 방법에서 충분하게 반영하지 못한 상관관계의 효과를 모형 내에 적용하여 반영하는 것이 가능하며, 또한 요율요소의 다양화에 따른 수렴 여부 문제 해결이 대부분 가능하다고 볼 수 있다.

1.2.4. 승법모형(Mutiplicative Model)을 이용한 방법

앞에서 언급한 것처럼 GLM 방법에서 정의되는 연결함수 (link function)인 $g^{-1}(\cdot)$ 가 자동차 보험료 산출 모형에 있어서 주로 로그 연결함수 (log-link function)을 사용하게 되는데, 그 이유는 다음과 같다.

$Log(Y) = a + b_1X_1 + b_2X_2 + \cdots$ 로 정의되는 경우, 지수함수를 양변에 취함으로 $Y = \exp\{a + b_1X_1 + b_2X_2 + \cdots\}$ 로 정의되게 되며 이는 지수함수의 특성으로 인하여 각 항의 곱으로 Y 값이 추정되게 되는 것이다. 즉, 모든 항의 곱으로 종속변수가 추정되는 승법모형 (multiplicative

model)을 통하여 보험료가 산출되는데, 이 방법 이외에 가법모형 (additive model), 혼합모형 (mixed model) 등 여러 방식이 가능하지만, 본 연구에서는 가장 설명이 용이한 승법모형을 사용한다.

위 식에서 설명된 각 항에서 절편에 해당하는 값은 각 요율요소의 특정 레벨이 기준이 되는 경우 평균 손해액이 되는 것이고, 이후 각 항에서 산출된 상대도-위험도의 상대적인 측정값이 되는 것이다. 따라서 각 요율요소별 해당 레벨에 대한 보험료는 절편과 산출하고자 하는 계층의 상대도의 곱으로 최종 가격이 산출되게 되는 구조를 가지게 된다.

2. Tweedie 분포

2.1. Tweedie 분포 개요

일반적인 지수족 (exponential family)의 확률밀도함수는 다음과 같다.

$$f(y;\theta,\psi)=\exp\left(\frac{y\theta-b(\theta)}{\psi}+c(y;\psi)\right). \tag{2.1}$$

Kaas 등 (2001)에 따르면, 위 식에서 평균은 $\mu=b\prime(\theta)$, 분산함수를 $V(\mu)$ 라 하면 분산은 $\sigma^2=\psi V(\mu)$ 로 표현할 수 있는데, $V(\mu)=\mu^p(1<p<2)$ 이 되도록 하는 함수 $b(\cdot)$, $c(\cdot;\cdot)$ 와 θ, ψ 에 대하여 Tweedie 분포족이라 하며 Tweedie (1984)에 의해 처음으로 그 형태가 제안되었고 Jørgensen (1987)에 의해 명명되었다.

Tweedie 분포함수는 정규분포, 감마분포와 같은 연속형 분포와 포아송분포와 같은 이산형 분포가 복합된 분포로서 지수족에 속하며, 점 '0'에서 양의 확률을 가지고 '0'보다 큰 부분에서 연속형 분포의 형태를 갖는다. Tweedie 분포의 평균과 분산은 각각 μ 와 $\psi\mu^p$ 이다. 여기서 $\psi(>0)$ 는 산포모수 (dispersion parameter)이며 p 는 Tweedie 분포족의 분포를 결정하는 지시모수 (index parameter)로서, $p=0$ 일 때 정규분포, $p=1,\psi=1$ 일 때 포아송분포, $p=2$ 일 때 감마분포가 된다. Tweedie 분포는 $0<p<1$ 을 제외한 모든 실수 값 p 에 대하여 존재하며, 앞서 열거한 특수한 경우 이외에는 확률밀도함수의 폐쇄형 (closed form)이 존재하지 않지만 컴퓨터 소프트웨어를 통하여 Tweedie 확률밀도함수의 정확한 값을 구할 수 있다.

확률변수 Y 가 보험료 청구액 (claim size)이 Gamma (α,β) 인 복합 (compound) Poisson (λ) 를 따른다고 할 때, 평균 $\mu=\lambda\alpha/\beta$, 분산 $\psi\mu^p=\lambda\alpha(\alpha+1)/\beta^2$ 이 되기 위해서는 다음과 같은 조건을 만족해야 한다.

$$\lambda=\frac{\mu^{2-p}}{\psi(2-p)},\quad \alpha=\frac{2-p}{p-1},\quad \frac{1}{\beta}=\psi(p-1)\mu^{p-1}. \tag{2.2}$$

식 (2.2)의 모수값들에 대하여, 연속형 $(y>0)$ 과 이산형 $(y=0)$ 이 혼합되어 있는 Y 의 확률밀도함수를 식 (2.1)과 같은 지수족의 형태로 표현하면 다음과 같다.

$$\begin{aligned} P(Y=0)&=e^{-}\lambda,\\ f(y)&=e^{-\beta y}e^{-\lambda}\sum_{n=1}^{\infty}\frac{\beta^{n\alpha}}{\Gamma(n\alpha)}y^{n\alpha-1}\frac{\lambda^n}{n!},\quad y>0. \end{aligned} \tag{2.3}$$

식 (2.2)에 따르면 $\lambda\beta^{\alpha}$ 은 μ 에 의존하지 않고 오직 ψ 와 상수 p 에만 의존한다. 따라서 식 (2.3)의 Σ 부분은 μ 에 의존하지 않고 오직 ψ 와 y 에 의존한다. 식 (2.1)을 식 (2.3)의 형태로 표현하기 위해서

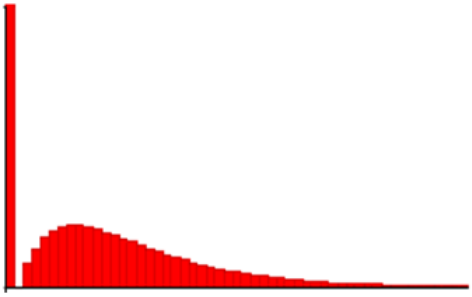
는 식 (2.1)의 $c(y;\psi)$ 를 식 (2.3)의 Σ 부분에 로그를 취한 것으로 정의하고 $c(0;\psi)=1$ 로 정의한 후, $-\beta=\theta(\psi), \lambda=b(\theta)/\psi$ 를 만족하는 θ 를 구하면 된다. 그 결과는 다음과 같다.

$$\theta=-\beta\psi=\frac{-1}{(p-1)\mu^{p-1}},\quad \mu(\theta)=(-\theta)p-1))^{-1/(p-1)},\quad b(\theta)=\lambda\psi=\frac{\mu^{2-p}}{2-p}.$$

2.2. 자동차 보험에서 Tweedie 분포의 적용

앞서 설명한 Tweedie 분포가 어떤 이유로 자동차 보험에 유용하게 사용될 수 있는지, 이 분포가 갖고 있는 특성과 자동차 보험 산업의 특성을 비교하여 그 활용의 적정성에 대하여 살펴보기로 한다.

자동차 보험 사고 손해액 데이터의 경우, 대부분 손해액이 '0', 즉 사고가 없는 차량이 대부분이 되고, 나머지 사고가 난 경우에 특정 손해액 범위에 많은 경우가 해당되고, 고액으로 갈수록 점점 줄어드는 것이 일반적이므로 다음과 같이 우측으로 꼬리가 긴 또 다른 분포함수의 형태를 가지게 된다.



지수족 함수의 경우, 사고가 발생한 손해액에 대한 분포는 감마분포 등의 분포로 가정할 수는 있으나, 일반적인 지수족 함수를 통하여서는 손해액이 '0'에서 대부분의 계약을 포함하는 형태를 설명해 낼 수가 없다. 이러한 특이한 양봉 (bimodal) 형태의 분포에 적합하도록 제안된 분포가 바로 Tweedie 분포인 것이다. 2.1절에서 소개한 분포함수에서 알 수 있는 것처럼 Tweedie 분포는 포아송분포와 감마분포를 합쳐 놓은 듯한 형태를 가지고 있는데, 이는 사고가 발생하지 않은, 즉 손해액이 '0'인 계약에 대하여서는 포아송분포의 형태로, 그리고 사고가 발생한 계약에 대하여서는 감마분포가 적용되는 형태인 것이다. 형태모수로 정의된 p 의 값이 1과 2의 사이에 있을 경우, y 변수가 '0'일 경우 높은 확률을 가지면서, 이후 감마분포와 유사한 형태를 가지게 되는 것이다. 참고로 $p\to 1$ 로 수렴하면 분포함수는 포아송분포에 근접하게 되고, $p\to 2$ 로 수렴하면 감마분포와 유사한 형태로 근접하는 특징이 있다. 자동차 보험료 산출에 있어서는 앞서 설명한 바와 같이 형태모수에 대하여 경험치를 통하여 확정할 수 있게 되는데, $p=1.5$ 를 사용하게 될 경우 심도와 빈도를 동일하게 간주하는 효과를 갖게 된다.

이와 같이 자동차 보험 산업에 아주 유용한 특성을 가진 함수이기는 하나, SAS 등을 통한 분포함수의 지원이 제한적이어서 MLE를 활용한 형태모수의 적용 등 아직은 광범위한 활용에는 한계가 있는 것이 현실이다. 따라서 다른 방법으로의 접근이 가능한데, GLM 방법을 사용하여 요율요소별 상대도를 산출하고 분류된 그룹에 대한 적정 보험료를 적용하는데 있어서 심도모형과 빈도모형을 나누어 적용하는 것이 일반적인 접근방법으로 알려져 있다. 즉, 사고건수를 대변하는 빈도모형의 경우 오차항을 포아송분포로 가정하여 산출하고, 사고의 경중을 대변하는 심도모형의 경우 오차항을 감마분포로 가정하여 산출한 후 두 모형을 곱하여 각 요율요소별 상대도 및 최종 보험료를 산출하게 되는 것이다.

이러한 방법의 장점은 특정 요율요소의 경우는 심도에만 영향을 미치게 되고 빈도에는 영향을 미치지 않는다고 가정할 때, 각 모형을 가장 잘 설명하는 변수를 각각 선택 할 수 있는 장점을 가진다고 할 수 있다. 그러나 이러한 방법도 심도모형과 빈도모형의 승산으로 이루어지고 있기에, 각 모형의 독립성을 가정해야 한다는 한계점을 가지고 있다. 따라서 지속적인 경험 데이터의 분석과 통계적 이론 연구를 통하여 최적의 산출 방법을 개발하여야 하며, 현재 컴퓨팅 능력에 대한 한계가 거의 없으므로 베이지안 방법의 적용 등에 대한 추가 연구가 고려되어야 할 필요성이 있다.

3. 실제 자료 분석

이 장에서는 실제 데이터를 통하여 앞에서 설명한 방법들을 실제 적용하고 그 결과의 비교를 통하여 자동차 보험료 산출에 있어 그 적정성을 비교해 보고자 한다.

각 해당 레벨 별 평균 손해액을 추정하고자 하는 모수의 참값이라고 가정하고, 각각 다른 보험료 산출 방법을 통하여 추정된 추정치를 모수와 비교하여 그 적정성을 비교하고자 한다. 적정성에 대한 평가기준은 다음과 같이 정의한다.

$$\sum(\text{모수의 참값} - \text{추정치})^2/\text{해당레벨구성합}$$

위와 같이 정의된 수식은, 각 해당 레벨 모수의 참값과 추정치의 차이에 대하여 해당 레벨의 빈도를 고려하여 측정하는 지표로서 그 값이 '0'에 가까울수록 해당 레벨 별 손해액을 정확하게 추정한 것을 의미하며, 그 값이 커질수록 정확도가 떨어지는 것을 의미한다. 3.1절에서는 분석에 사용된 실제 데이터와 모형에 대하여 설명하고, 3.2절에서 그 결과를 비교한다.

3.1. 실제 데이터 개요

보험료 산출 방법의 비교를 위한 분석에 사용된 데이터는 Kim과 Kim (2009)에서 다루었던 국내 자동차 보험사의 자동차 보험 대물 사고에 대한 실제 자료로서, 전체 데이터에서 10만개의 표본을 임의로 추출하여, 이 표본에서 얻어지는 값을 특성치를 모수의 참값으로 가정하였다.

위험도를 파악하는 요소로서는 연령, 성별, 차량형태, 운전자 범위, 운전경력 등 5개의 요소가 고려되었으며, 각 요소별 레벨의 정의는 다음과 같다

연령: 40세 이하, 41 ~ 50 세, 51세 이상 [3 레벨]
성별: 남성, 여성 [2 레벨]
차량형태: 소형, 중형, 대형, 다목적 차량 [4 레벨]
운전자 범위: 미제한, 가족, 부부, 본인 [4 레벨]
운전경력: 2년이하, 3년 이상 [2 레벨]

비교 분석 방법은 크게 나누어 단변량 방법과 GLM 방법이며, GLM 방법에서는 심도와 빈도를 분리하여 적용하는 방법과, 심도와 빈도를 동시에 고려하는 순보험료 모형을 비교 대상으로 하고자 한다.

심도와 빈도를 분리하여 적용하는 방법은 다음과 같다.

- **[심도모형]**
  $S(i) = \exp(x_i^T \beta_s)$

여기서 β_s 는 심도모형의 모수벡터이며, 이와 같은 심도모형의 오차항의 분포는 감마분포를 가정하게 되는데, 심도란 사고당 발생하는 손해액으로서, 그 특성상 소액에 다수가 분포되어 있고 금액이 고액으로 증가하면서 그 분포가 감소하는 특성을 가지고 있다.

- **[빈도모형]**
  $F(i) = \exp(x_i^T \beta_f)$

여기서 β_f 는 빈도모형의 모수벡터이며, 빈도모형은 앞서 언급한 심도모형과 같은 독립변수와 모형을 가지게 되며, 빈도는 해당 기간 내 발생하는 사고건수로 정의되며 그 특성상 포아송분포를 가정한다.

이렇게 구해진 심도모형과 빈도모형을 통하여 추정하고자 하는 보험료는 다음과 같이 산출되게 된다.

$$\begin{aligned} R(i) = S(i) * F(i) &= \exp(x_i^T \beta_s) * \exp(x_i^T \beta_f) \\ &= \exp\left\{ x_i^T (\beta_s + \beta_f) \right\}. \end{aligned}$$

즉, 심도와 빈도를 분리하여 적용하는 방법은 심도와 빈도가 독립이라는 가정하에 심도와 빈도의 특성에 맞는 위험도를 산출하여 두 모형에서 산출된 위험도의 곱을 통하여 추정하는 방법이다.

심도와 빈도를 동시에 고려하는 순보험료 모형은, 해당 셀의 손해액과 사고건수를 동시에 고려하여 보험료를 산출하는 방법이다. 이 방법을 사용하게 되는 경우 앞서 설명한 Tweedie 분포가 적절한 오차항의 분포로 가정되게 되는 것이다.

$$G(i) = \exp(x_i^t \beta_g).$$

여기서 β_g 는 모수벡터이며, 순보험료를 직접 추정하게 되는 경우 앞서 설명한 두 방법과는 달리 심도와 빈도가 동시에 고려되었기 때문에 하나의 모형의 통하여 위험도가 추정되고, 연속형 변수인 심도와 이산형 변수인 빈도가 동시에 고려되는 모형이다.

3.2. 분석 결과

단변량 방법으로 위험도를 산출하는 절차는 다음과 같이 비교적 단순하다.

$$a_{ij} = \sum_{j=1}^{n} \text{손해액}(j)/n.$$

여기서 i 는 해당 요율요소의 레벨 - 예를 들어, 연령 요소의 경우 $i = 1$ 은 40세 이하, $i = 2$ 는 41세 50세, $i = 3$ 은 50세 이상- 을 나타내며, j 는 해당 레벨에 존재하는 관측치의 수가 된다. 따라서 해당 레벨에 대한 각각의 평균을 구한 후 산출된 평균의 상대도 개념으로 위험도 정도를 비교해 나가는 방식이다. 즉, 위와 같은 방식으로 사용된 개의 변수 각 레벨 별 평균 손해액을 구한 후, 해당 변수의 특정 레벨을 기준으로 기타 레벨의 위험도의 상대적인 값 (상대도)을 구한다. 예를 들어, 연령 40대의 평균 손해액이 100이고, 40세 이하의 평균 손해액이 120일 경우, 40대를 기준으로 하여 40대의 상대도는 1이 되고, 30대의 상대도는 1.2가 되는 것이다. 이와 같은 개념으로 단변량 방법과 GLM 방법으로 구한 변수의 레벨 별 상대도를 구한 결과는 다음 표 3.1과 같다

참고로 순보험료 모형에서 감마분포를 가정하여 추정한 모형을 추가하여 비교함으로써 적합한 모형의 중요성을 언급하고자 하였다. 부적절한 분포가 가정되는 경우 위험도 정도의 오류 뿐 아니라 위험도의 방향성까지도 반대로 산출되는 오류를 범하게 됨을 확인할 수 있다. 예를 들어, 성별의 경우 여성의 위험도가 높으나, 감마분포를 가정하면 여성의 위험도가 더 낮게 추정된다. 위의 결과에서 확인할 수 있듯이 적절한 분포가 가정된 심도 빈도 모형과 Tweedie 순보험료 모형의 경우 추정된 위험도의 값이 유사

표 3.1 GLM을 사용하여 구한 변수의 레벨 별 상대도

구분	레벨	심/빈도 모형 (Gamma/Poisson)	순보험료 모형 (Tweedie)	순보험료 모형 (Gamma)	단변량 모형
연령	40세 이하	1.0000	1.0000	1.0000	1.0000
	41~50세	1.2205	1.2152	1.0900	1.4380
	50세이상	1.2711	1.2911	1.0991	1.2330
차종	대형	1.0082	1.0179	1.1284	1.0120
	중형	1.1336	1.1421	1.2031	1.1587
	다목적	1.3461	1.3508	1.1957	1.3195
	소형	1.0000	1.0000	1.0000	1.0000
운전 범위	미제한	1.0418	1.0366	1.1190	1.0773
	부부	1.0000	1.0000	1.0000	1.0000
	가족	1.3085	1.2959	1.0827	1.3956
	본인	1.0467	1.0275	1.0182	1.0368
운전 경력	2년 이하	1.0695	1.0657	0.8861	1.0242
	3년 이상	1.0000	1.0000	1.0000	1.0000
성별	여성	1.0408	1.0391	0.9971	1.0361
	남성	1.0000	1.0000	1.0000	1.0000

하게 산출됨이 확인되었으나, 단변량의 경우 요율요소 사이에 존재하는 상관관계 등으로 인하여 위험도 상대도의 값이 왜곡되어 산출되는 결과를 확인할 수 있다.

분석결과를 바탕으로 앞서 정의한 적정성 판단 기준에 의하여 그 수치를 비교하는 방법은 다음과 같다.

각 요소의 모든 조합이 본 연구에서 추정하고자 하는 최종 위험도의 값이 되므로, 연령(3)x성별(2)x차량형태(4)x운전범위(4)x운전경력(2) 의 조합인 192개의 그룹 (segment)이 존재하며, 각 셀 별로 모수와 추정치의 차의 제곱합을 통하여 그 적정성을 확인한다.

표 3.2~3.4는 분석 결과의 예시로서, 40세 이하 남성 소형차 소유주의 경우에 운전범위, 운전경력에 따른 결과이다.

표 3.2 40세 이하 남성 소형차 소유주의 운전범위, 운전경력에 따른 결과

운전	운전	심/빈도 모형 (Gamma/Poisson)			단변량 모형			순보험료 모형 (Gamma)		
범위	경력	운전범위	운전경력	결과	운전범위	운전경력	결과	운전범위	운전경력	결과
미제한	2년이하	1.0418	1.0695	1.1142	1.0773	1.0242	1.1034	1.1190	0.8861	0.9915
	3년이상	1.0418	1.0000	1.0418	1.0773	1.0000	1.0773	1.1190	1.0000	1.1190
가족	2년이하	1.3085	1.0695	1.3994	1.3956	1.0242	1.4294	1.0827	0.8861	0.9594
	3년이상	1.3085	1.0000	1.3085	1.3956	1.0000	1.3956	1.0827	1.0000	1.0827
부부	2년이하	1.0000	1.0695	1.0695	1.0000	1.0242	1.0242	1.0000	0.8861	0.8861
	3년이상	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
본인	2년이하	1.0467	1.0695	1.1194	1.0368	1.0242	1.0619	1.0182	0.8861	0.9023
	3년이상	1.0467	1.0000	1.0467	1.0368	1.0000	1.0368	1.0182	1.0000	1.0182

지면의 제한으로 192개의 조합 가운데 일부를 발췌하여 앞의 표와 같이 제시하여 비교하였으며, 최종 결과로서 각 모형에 따른 모수와 추정치의 차의 제곱합은 표 3.5와 같다.

결과에서 알 수 있는 것처럼 선진 통계 방식이 사용될수록 그 위험도를 추정하는데 있어 점점 정교화 되고 있음을 확인할 수 있다.

지금까지 위험도 평가 방식에 중심을 두어 비교하였고 이러한 평가 방식이 진화되면서 상관관계의 오류등과 같이 기존 방식이 가지고 있는 문제점이 어떻게 해결되는지 확인하였다. 본 논문에서는 모든 요

표 3.3 (계속)

운전 범위	운전 경력	순보험료 모형 (Tweedie) 운전범위	운전경력	결과	모집단 구성	전체 손해	순보험료 위험도	모집단
미제한	2년이하	1.0366	1.0657	1.1047	418	13,873,730	33,191	0.9413
	3년이상	1.0366	1.0000	1.0366	1,447	37,932,200	26,214	0.7434
가족	2년이하	1.2959	1.0657	1.3810	458	21,105,270	46,081	1.3068
	3년이상	1.2959	1.0000	1.2959	1,066	39,306,090	36,873	1.0457
부부	2년이하	1.0000	1.0657	1.0657	1,266	40,049,620	31,635	0.8971
	3년이상	1.0000	1.0000	1.0000	5,216	183,928,530	35,262	1.0000
본인	2년이하	1.0275	1.0657	1.0950	778	33,709,300	43,328	1.2287
	3년이상	1.0275	1.0000	1.0275	2,016	64,464,410	31,976	0.9068

표 3.4 (계속)

운전 범위	운전 경력	모수와 추정치의 차이 심/빈도 모형 (Gamma/Poisson)	단변량 모형	순보험료 모형 (Gamma)	순보험료 모형 (Tweedie)
미제한	2년이하	0.02991	0.0263	0.0025	0.0267
	3년이상	0.08904	0.1115	0.1411	0.0860
가족	2년이하	0.00857	0.0150	0.1207	0.0055
	3년이상	0.06908	0.1225	0.0014	0.0626
부부	2년이하	0.02971	0.0161	0.0001	0.0284
	3년이상	0.00000	0.0000	0.0000	0.0000
본인	2년이하	0.01195	0.0278	0.1066	0.0179
	3년이상	0.01957	0.0169	0.0124	0.0146

표 3.5 각 모형에 따른 모수와 추정치의 차의 제곱합

심/빈도 모형	단변량 모형	순보험료 모형 (Gamma)	순보험료 모형 (Tweedie)
170.10	184.91	177.97	170.06

소의 주효과만을 산출하여 비교하였다.

변수 사이에 존재하는 상관관계 (교호작용)에 대한 확인, 차원 결정 등에 대한 의사 결정 등 모형화 이전에 분석되어야 할 많은 요소들이 내포 되어 있으며, 또한 각 변수의 레벨을 정의, 분류하는 방법 등, 그리고 모형 내의 변수 간에 존재하는 상관관계를 어떻게 정의하여 모형에 반영할 것인지에 대한 연구 또한 지속적으로 연구되어야 한다. 본 연구에서는 이미 정의된 변수를 통하여 위험도의 적정한 평가를 논하였지만, 어떤 변수를 어떤 방식으로 정의하여 모형에 반영할 것인가에 대한 것 또한 최적의 위험도 평가 이전에 선행되어야 하는 필수불가결한 연구 분야이기 때문이다. 이러한 적절한 입력 (input) 방법에 대한 연구와 본 논문에서 다루고 있는 최적 결과 산출이 병행하여 발전한다면, 본 논문에서 추구하고자 하는 최상의 결과를 도출해 낼 수 있을 것이다.

4. 결 론

앞장에서 자동차 보험료 산출에 대하여 여러 가지 방법이 어떻게 발전되어 왔는지, 그리고 어떤 방법들이 적용 발전되고 있는지에 대하여 살펴보았고, 현재 국내에서 실제 발생된 자동차 보험 사고 데이터를 통하여 그 결과를 확인하여 보았다.

보험료를 통하여 보험사가 고객의 위험을 분담하여 준다는 측면에서 고객의 위험이 어느 정도인지를

정확하게 파악하는 것이 바로 적정 보험료 산출의 근거가 되는 것이라고 볼 수 있다. 따라서 보험사는 고객으로부터 담보하여야 할 위험도를 정확하게 산출하는데 지속적인 노력을 기울여야하는 것은 당연한 것이며, 그 산출 방법의 적정성 및 합리적 방법의 활용 등 다양한 노력을 기울어야 하겠다. 이러한 관점에서 GLM 방법은 가장 선진화된 통계 기법을 활용하는 방법 중의 하나이며, 본 논문에서 소개한 Tweedie 분포 또한 자동차 보험업의 특성상 그 현상을 잘 대변해 주는 분포로서 자동차 보험료 산출에 있어 가장 적합한 방법 중의 하나라고 할 수 있다.

현실적인 여건, 비용 대비 효과 등에 비추어 여러 가지 방법 중 특정 방법의 사용만을 주장하는 것은 무리이겠지만, 위에 열거된 여러 방법들을 적재적소에 잘 활용하여 적정보험료를 산출한다면 고객은 본인에게 가장 적합한 위험도에 맞는 보험료를 내게 되고 보험사 또한 해당 위험도에 따른 보험료 부과로 기업운영에 도움이 되는 윈-윈 효과를 얻을 수 있을 것으로 판단된다. 향후 여러 보험사에서 각 보험사가 쌓아 놓은 노하우에 본 논문에서 제시한 방법들에 대한 깊이있는 이해와 연구 및 적용을 통하여 국내 자동차 보험 산업 발전에 기여하기를 기대한다.

참고문헌

Bailey, R. A. and Leroy, J. S. (1960). Two studies in automobile insurance ratemaking. *Proceedings of the Casualty Actuarial Society*, **47**, 192-217.

Bailey, R. A. (1963). Insurance rates with minimum bias. *Proceedings of the Casualty Actuarial Society*, **50**, 4-11.

Feldblum, S. and Brosius, J. E. (2002). The minimum bias procedure - A partitioner's guide. *Proceedings of the Casualty Actuarial Society*.

Jørgensen, B. and Paes de Souza, M. C. P. (1994). Fitting Tweedie's compound model to insurance claims data. *Scandinavian Actuarial Journal*, **1**, 69-93.

Kaas, R., Goovaerts, M. J., Dhaene, J. and Denuit, M. (2001). *Modern actuarial risk theory*, Kluwer, Dordrecht.

Kim,Y-H. and Kim, Ki Su .(2009). Small area estimation of the insurance benefit for customer segmentation. *Journal of the Korean Data & Information Science Society*, **20**, 77-87.

Murphy, K. P., Brockman, M. J. and Lee, P. K. W. (2000). Using generalized linear models to build dynamic pricing systems. *Casualty Actuarial Forum*, Winter 2000.

Smyth, G. K. and Jørgensen, B. (2002). Fitting Tweedie's compound model to insurance claims data: Dispersion modelling. *ASTIN Bulletin*, **32**, 143-157.

Tweedie, M. C. K (1984). An index which distinguishes between some important exponential families in statistics applications and new directions. *Proceedings of the Indian Statistical Institute Golden Jubilee International Conference,* 579-604.

Journal of the Korean Data & Information Science Society
2009, **20**(3), 515–526

한국데이터정보과학회지

Various modeling approaches in auto insurance pricing

Myung Joon Kim[1] · Yeong-Hwa Kim[2]

[1]Automobile Insurance Product Pricing Dept., Samsung Fire & Marine Insurance Co.
[2]Department of Statistics, Chung-Ang University

Received 16 February 2009, revised 20 May 2009, accepted 23 May 2009

Abstract

Pricing based on proper risk has been one of main issues in auto insurance. In this paper, we review how the techniques of pricing in auto insurance have been developed and suggest a better approach which meets the existing risk statistically by comparison. The generalized linear model (GLM) method is discussed for pricing with different distributions. With GLM approach, the distribution of error assumed plays an main role for the best fit corresponding to the characteristics of dependent variables. Tweedie distribution is considered as one of error distributions in addition to widely used Gamma and Poisson distribution. With these different types of error assumption for estimating the proper premium in auto insurance, various modeling approaches are possible. In this paper, various modeling approaches with different assumptions for estimating proper risk is discussed and also real example is given by assuming different.

Keywords: Auto insurance, GLM, pricing, Tweedie distribution.

[1] Manager, Automobile Insurance Product Pricing Dept., Samsung Fire & Marine Insurance Co., Seoul, 100-782 Korea.
[2] Corresponding author: Associate Professor, Department of Statistics, Chung-Ang University, 221 Heuksuk-dong, Dongjak-gu, Seoul 155-756, Korea. E-mail: gogators@cau.ac.kr