이중일반선형모형(DGLM)을 이용한 자동차 보험요율 추정*

Estimating the Rate of Motor Insurance Premium by Double Generalized Linear Model

최우석**
Choi Woo-Suk

한상일***
Han Sang-Il

본 연구는 손실함수가 Tweedie's Compound 포아송 분포를 따르는 경우, 적정 자동차 보험료를 추정함에 있어 DGLM을 이용하여 평균뿐만 아니라 평균의 분산까지 모델링함으로써 보험요율 추정의 정확도를 향상시킬 수 있는 대안에 대하여 고찰해 보았다. 또한 한국과 호주의 제3자배상 자동차보험 포트폴리오를 DGLM에 적용하여 추정된 실증분석 결과를 통해 특정 공변량들의 경우 청구빈도와 규모에 동일한 또는 반대의 영향을 미치고 있음을 실증적으로 확인하였다. 이러한 결과로부터 보험금 청구 데이터의 분산을 일정한 상수로 가정하는 모형보다 분산의 동시 모델링을 통해 보험료 추정치의 정확도를 향상시킬 수 있음을 알 수 있었다.

한편 보험료 할인(할증) 위험요인들 간에도 상당한 편차가 존재함을 알 수 있었다. 실증분석에 사용된 여러 위험요인 중 차종의 최대할인율차이가 가장 크게 나타났으며, 운전자의 거주 지역변수 역시 운전자의 연령에 못지않게 중요한 위험 할인(할증) 요인인 것으로 나타났다. 그러나 본 논문의 실증분석 결과에 따르면 지역별 자동차보험료 차등화 방안에 있어 그동안 논란의 중심이었던 수도권과 지방간 보험료 양극화 우려의 문제는 심각하게 발생하지 않을 것으로 사료된다.

국문색인어 : 이중일반선형모형(DGLM), 자동차보험, Compound 포아송분포
학술진흥재단 분류 연구분야 코드: B051605

* 본 논문에 대해서 유익한 논평을 해주신 익명의 두 분 심사자께 감사드립니다.
** 성공회대학교 유통정보학과 교수(wschoi@skhu.ac.kr)
*** 한국기술교육대학교 산업경영학과 교수(sihan@kut.ac.kr)
논문 투고일: 2008. 08. 07, 논문 게재 확정일: 2008. 11. 21

I. 서론

자동차 보험요율 자유화 이후 손해보험사는 손해율과 인수조건에 따라 각기 상이한 보험료를 책정하고 있으며, 이에 따라 적절한 위험요인의 탐색과 보험요율 추정 모형의 정교화에 대한 관심이 날로 높아지고 있다. 특히 OECD 회원국가 중 인구 10만명당 교통사고 사망자수가 가장 높은 국내 여건을 감안할 때, 위험요인에 따른 정교한 보험요율 추정 모형의 구축 및 개선은 손해보험사의 사활을 좌지우지할 수 있는 중요한 분야로 대두되고 있다. Jorgensen-Souza(1994)는 청구빈도가 포아송분포, 개별 청구건에 대한 지급금액이 감마분포를 따른다는 가정하에, 계약건당 보험금의 기대값을 추정하였다. 상기 두 가정은 일정기간 동안의 총 청구금액 즉, 손실함수가 Tweedie Compound 포아송 과정을 따른다는 것을 의미하며, 이에 따른 계약건당 평균 손실금액을 일반선형모형(Generalized Linear Model: GLM)을 이용하여 직접 모델링하였다[1]. Jorgensen-Souza(1994)의 방법론은 청구리스크에 영향을 미치는 변수들(지역, 나이, 차종, 성별, 무사고경력 등)이 청구빈도와 청구규모를 동시에 증가 또는 감소시킴으로써 총청구금액(보험금)의 기대값에 영향을 미침을 암묵적으로 가정하고 있다. 그러나 현실은 어떤 변수는 청구빈도 보다 규모에 더 영향을 미치는 반면 다른 변수는 청구규모보다 빈도에 더 영향을 미칠 수 있으며, 어떤 변수는 청구빈도와 규모에 서로 반대 방향으로 영향을 미칠 가능성도 존재한다. 이러한 가능성은 보험금 청구 데이터의 분산이 일정한 상수가 아닐 수 있음을 의미하며, 따라서 보험료의 모델링에 있어서 추정치의 정확도를 향상시키기 위해서는 평균뿐만 아니라 분산을 모델링할 필요가 있음을 암시한다.

이에 따라 Smyth-Jorgensen(2002)은 보험료를 모델링함에 있어 평균뿐만 아니라 분산에 대한 모델링이 필요함을 주장하였다. 그들은 보험료를 추정함에 있어,

1) GLM과 보험수리적 응용에 관한 연구로는 Renshaw(1994), Haberman-Renshaw(1998), Millenhall(1999), Murphy-Brockman-Lee(2000) 등이 있으며, 이 중 Millenhall(1999)은 보험수리 분야에 있어서 Compound 포아송과정의 활용 사례를 풍부하게 제시하고 있다. 한편, Compound 포아송분포에 대한 수학적인 상세 기술은 Jorgensen(1997), Rolski-Schmidli-Schmidt-Teugels(1999)를 참조하기 바란다.

Nelder-Pregibon(1987), Smyth(1989), Smyth-Verbyla(1999) 등이 제안한 이중일반선형모형(Double GLM: DGLM)을 이용하였다. DGLM을 이용하면 평균과 분산을 동시에 모델링할 수 있으며, 이들은 분산모형이 평균 추정치의 정확도를 향상시키고 있음을 보고하였다[2]. Smyth-Jorgensen(2002)은 DGLM을 이용해 청구건수 데이터 없이 총청구금액 데이터만으로도 보험료 모델링이 가능함을 보였다. 한편, 분산 모델링에 있어 평균에 영향을 미치는 모수의 수가 표본의 크기에 비례하여 커지는 경우 최대우도추정법(MLE)은 분산 추정치에 편의(Bias)를 발생시키게 된다. Lee-Nelder(1998), Smyth-Verbyla(1999), Smyth-Huele-Verbyla(2001) 등은 DGLM에서의 잔차최대우도추정(Residual Maximum Likelihood: REML) 문제를 연구하였다. REML에서의 초점은 평균값 추정을 위한 분산 서브모형의 추정방법 조정이며, REML이 분산의 불편 추정치를 제공하고 있음을 보고하였다.

본 논문에서는 자동차 보험료를 추정함에 있어 Smyth-Jorgensen(2002)이 적용한 바 있는 Tweedie's Compound 포아송 분포 가정하에 평균과 분산을 함께 고려할 수 있는 DGLM을 사용하였으며, 모수 추정방법은 분산의 불편 추정치를 제공하는 REML을 사용하였다. II장에서는 손실함수로서 Compound 포아송모형과 DGLM 방법론에 대해 기술하였다. III장에서는 한국과 호주의 제3자배상 자동차보험(책임보험) 포트폴리오 데이터에 DGLM을 직접 적용하여 위험요인들에 따른 보험료의 할인(할증)의 차이에 대해 분석하였다. 마지막으로 IV장에서는 결론 및 추후 연구과제에 대해 정리하였다.

2) Nelder-Pregibon(1987), Smyth(1989), Smyth-Verbyla(1999) 등은 다른 분야 데이터의 평균-분산 모델링에 DGLM을 이용한 바 있다.

II. 평균-분산 보험료 추정 모형

1. Compound 포아송모형과 DGLM

본 절에서는 보험료 추정에 있어서의 DGLM 방법론 적용과정을 소개하고자 한다. 먼저 N_i를 i번째 분류범주에서의 청구건수(i=1,,m), Z_i를 i번째 분류범주에서의 총청구금액(보험금 지급금액),ω_i를 계약연도(policy year) 기준으로 측정된 계약건수라고 정의하자. 그러면 $Y_i = Z_i/\omega_i$는 i번째 분류범주에서의 보험계약 한건당 평균 지급금액 즉, 평균 손실금액이 된다. 한편 청구건수(N_i)가 평균이 $\lambda_i\omega_i$인 포아송분포를 따른다고 가정하자. 그리고 손실의 심도(Severity)인 청구건당 손실규모를 평균이 T_i, 형상모수(shape parameter)가 α인 감마분포라고 가정하자. 그러면 Z_i는 Compound 포아송분포를 따르게 된다. 또한 $\Pr(N_i=0) = \Pr(Y_i=0) = e^{-\lambda_i\omega_i}$이 되며, Y_i는 '0'인 경우를 제외하면 항상 연속이고 陽의 값을 갖게 된다. 또한 개별 청구건이 독립적으로 발생한다고 가정하면, N_i가 양수로 주어진 상황에서 Y_i의 조건부 분포는 평균이 $N_i T_i/\omega_i$인 감마분포를 따르게 된다. 따라서 Y_i역시 Compound 포아송과정을 따르게 된다. 또한 본 논문에서는 분류범주 i에 대해 청구 건수와 규모 둘 다 관찰가능하며, (n_i, y_i)가 독립적으로 관찰가능하다고 가정한다.

한편, Jorgensen-Souza(1994)는 보험료 산출문제에 있어 특정범주 i에서 계약건당 평균 손실금액(μ_i)이 $\mu_i = E(Y_i) = \lambda_i T_i$임을 보였다. 또한 Jorgensen(1987, 1997)은 평균 손실금액의 분산이 $Var(Y_i) = \phi_i \mu_i^p/\omega_i$(단, $p=(\alpha+2)/(\alpha+1)$, ϕ_i는 분산모수)임을 보였다. 즉, N_i와 Y_i의 결합밀도함수를 Y_i의 평균과 분산을 설명하고 있는 μ_i, ϕ_i, p 3가지로 모수화(parameterization)할 수 있다. 그런데 여기서 ϕ_i와 p가 μ_i과 각각 통계적으로 직교(orthogonal)하고 있으므로[3], μ_i, ϕ_i, p를 모수로 추정하는 것이 λ_i, T_i, α를 추정하는 것보다 추정방법이 간편해진다. 한편, $Var(Y_i)$를 N_i가 주어진 상황에서 Y_i의 조건부 분포(감마분포)를 이용해 구할 수 있다.

3) 이것은 공분산행렬(Fisher information matrix)에서 非대각선원소가 '0'을 의미한다.

$$Var(Y_i) = E_{N_i} Var(Y_i/N_i) + Var_{N_i} E(Y_i/N_i)$$
$$= \lambda_i T_i^2/(\alpha\omega_i) + \lambda_i T_i^2/\omega_i = (\frac{1}{\alpha} + 1)\lambda_i T_i^2/\omega_i$$

위식을 $Var(Y_i) = \varnothing_i \mu_i^p/\omega_i$와 연결하여 정리하면 분산모수는 다음과 같다.

$$\varnothing_i = \omega_i var(Y_i)/\mu_i^p = \lambda_i^{1-p} T_i^{2-p}$$

여기서 λ_i의 지수(1-p)가 음수이므로[4] 사고건당 청구규모(심도)의 평균(T_i)에 영향을 주지 않으면서 청구빈도(λ_i)를 증가시키는 위험요인은 계약건당 손실함수의 평균(μ_i)을 증가시키는 반면 분산($\varnothing_i$)은 감소시키게 된다. 반면, T_i의 지수(2-p)는 양수이므로 청구빈도(λ_i)의 증가 없이 청구심도의 평균(T_i)를 증가시키는 요인은 μ_i와 $\varnothing_i$를 동시에 증가시킨다. 따라서, 이 경우 평균(μ_i)과 분산($\varnothing_i$)이 동시에 공변량(covariates)에 영향을 받는 모형을 수립하는 것이 평균 추정치의 정확도를 향상시키게 된다. 따라서 다음과 같은 DGLM을 설정하였다.

범주 i에 대한 평균 청구금액(μ_i) :

$$g(\mu_i) = x_i^T \beta \quad (1)$$

g : 단조연결함수(monotonic link function)
x_i: 공변량 벡터
β : 회귀계수 벡터

범주 i에 대한 평균 청구금액의 분산($\varnothing_i$) :

$$g_d(\varnothing_i) = z_i^T \gamma \quad (2)$$

g_d : 단조연결함수
z_i : 분산에 대한 공변량 벡터
γ : 회귀계수 벡터

4) Tweedie 분포에서 $1 < p < 2$

g와 g_d 를 로그함수라 하면 식(1)과 (2)는 청구빈도(λ_i)와 청구규모(T_i)에 대해서 대수선형(log-linear) 모형이 되어, 위의 DGLM 모형은 청구빈도와 규모를 설명변수들(공변량)과 각각 대수선형 모형을 수립한 것과 동일한 구조를 갖게 된다.

2. 결합우도함수

본 절에서는 편의상 특별한 경우를 제외하고 분류범주 기호 i를 생략하기로 한다. N과 Y의 결합밀도함수는 다음과 같다.

$$f(n,y:\mu,\phi/\omega,p)=a(n,y,\phi/\omega,p)\exp\{\frac{\omega}{\phi}t(y,\mu,p)\}$$

$$a(n,y,\phi/\omega,p)=\{\frac{(\omega/\phi)^{\alpha+1}y^{\alpha}}{(p-1)^{\alpha}(2-p)}\}^{n}\frac{1}{n!\Gamma(n\alpha)y}$$

$$t(y,\mu,p)=y\frac{\mu^{1-p}}{1-p}-\frac{\mu^{2-p}}{2-p}$$

또한 DGLM의 모수 β, γ, p에 대한 로그-우도함수와 μ 와 ϕ에 대한 로그-밀도함수의 도함수는 다음과 같다.

$$l(n,y;\beta,\gamma,p)=\sum_{i=1}^{m}\log f(n_i,y_i;\mu_i,\phi_i/\omega_i,p)$$

$$\frac{\partial logf(n,y;\mu,\phi/\omega,p)}{\partial\mu}=\frac{\omega}{\phi}\frac{\partial t(y,\mu,p)}{\partial\mu}=\frac{\omega}{\phi}\frac{y-\mu}{\mu^{p}} \tag{3}$$

$$\frac{\partial logf(n,y;\mu,\phi/\omega,p)}{\partial\phi}=-\frac{n}{(p-1)\phi}-\frac{\omega}{\phi^{2}}t(y,\mu,p) \tag{4}$$

$$\frac{\partial^{2}logf(n,y;\mu,\phi/\omega,p)}{\partial\phi^{2}}=-\frac{n}{(p-1)\phi^{2}}+\frac{2\omega}{\phi^{3}}t(y,\mu,p)$$

$$E[t(Y,\mu,p)] = \frac{\mu^{2-p}}{(1-p)(2-p)}$$

$$E(N) = \frac{\omega}{\phi}\,\frac{\mu^{2-p}}{2-p}$$

여기서 분산 가중치 ω_d와 분산반응 d를 다음과 같이 정의하자[5].

$$\omega_d = \frac{2\omega\mu^{2-p}}{(2-p)(p-1)\phi}$$

$$d = \frac{2\phi^2}{\omega_d}\,\frac{\partial logf}{\partial\phi} + \phi = -\frac{2}{\omega_d}\left(\frac{n\phi}{p-1} + \omega t\right) + \phi$$

그러면 ϕ에 대한 로그-밀도함수의 1차 도함수인 식(4)는 다음과 같아지며,

$$\frac{\partial logf(n,y;\mu,\phi/\omega,p)}{\partial\phi} = \frac{\omega_d(d-\phi)}{2\phi^2}$$

$\hat{\gamma}$의 표준오차는 다음과 같은 ϕ에 대한 피셔 정보행렬(Information Matrix)의 역행렬로부터 얻어진다[6].

$$\mathfrak{I}_d = diag\left\{\frac{\omega_{d_i}}{2V_d(\phi_i)}\right\}$$

5) d는 식(1)의 GLM에서 평균으로부터의 유닛편차(unit deviation)이며 식(2)의 GLM에서 반응(response)이 된다. Nelder-Pregibon(1987), Jorgensen(1997), Smyth-Verbyla(1999)는 안장점근사(saddlepoint approximation 또는 steepest descent method)를 통해 d의 분포가 근사적으로 ϕx^2_1을 따르고 있음을 보였다.

6) 자세한 유도과정은 Smyth(1989)를 참조하기 바람.

3. ML 추정방법

평균과 분산에 대한 회귀계수 β와 γ의 MLE는 식(1)과 (2)의 GLM으로부터 차례로 얻어진다. β에 대한 피셔 점수방정식(Fisher Scoring Equation)은 다음과 같다.

$$\beta^{k+1} = (X^T W X)^{-1} X^T W z \tag{5}$$

여기서 W는 다음과 같은 가중치 대각행렬이고,

$$W = diag\{[\frac{\partial g(\mu_i)}{\partial \mu}]^{-2} \frac{\omega_i}{\phi_i V(\mu_i)}\}$$

분산함수 , $V(\mu) = \mu^p$, $z_i = \frac{\partial g(\mu_i)}{\partial \mu}(y_i - \mu_i) + g(\mu_i)$이다.

한편, $\hat{\beta}$의 표준오차는 다음과 같은 피셔 정보행렬의 역행렬로부터 얻어진다.

$$\mathfrak{I}_\beta = X^T W X$$

γ 에 대한 피셔 점수방정식은 다음과 같다.

$$\gamma^{k+1} = (Z^T W_d Z)^{-1} Z^T W_d z_d \tag{6}$$

여기서 W_d는 다음과 같은 가중치 대각행렬이고,

$$W_d = diag\{[\frac{\partial g_d(\phi_i)}{\partial \phi}]^{-2} \frac{\omega_{d_i}}{2V_d(\phi_i)}\}$$

분산함수 , $V_d(\phi) = \phi^2$, $z_{d_i} = \frac{\partial g_d(\phi_i)}{\partial \phi}(d_i - \phi_i) + g_d(\phi_i)$이다.

한편, $\hat{\gamma}$의 표준오차는 다음과 같은 피셔 정보행렬의 역행렬로부터 얻어진다.

$$\mathfrak{I}_{\gamma} = Z^{T} W_{d} Z$$

여기서 Z는 평균 청구비용의 분산(ϕ)에 대한 GLM인 식(2)의 공변량행렬이고, W_d는 분산 가중치(ω_d) 대각행렬이다. 앞절의 식(3)을 ϕ 또는 p에 대해서 미분하면 도함수가 '0'이 된다. 즉 μ는 ϕ ,p 모두와 직교하고 있다. μ가 ϕ와 p에 직교하므로, β도 γ와 p에 직교한다. β와 γ가 서로 직교하므로 식(5)과 (6)를 교대로 반복하는 알고리즘은 빠르게 수렴하게 된다(Smyth (1996))[7].

4. REML 추정방법

선형회귀모형에서 모수의 수가 표본의 수보다 큰 경우 분산에 대한 MLE는 하향편의(偏倚)성을 나타내며, 이는 DGLM에서도 유사하게 나타난다. 즉, 평균에 영향을 미치는 모수(μ_i, ϕ_i, p)의 수가 분류범주의 수보다 많기 때문에 $\hat{\phi}_i$의 MLE는 하향편의되었으며, 추정된 분산 $\hat{\phi}_i \hat{\mu}_i^{\hat{p}} / \omega_i$도 역시 과소추정되는 경향이 있다. 이러한 경우, REML(Residual ML 또는 Restricted ML)은 분산 추정치를 조정함으로써 추정된 평균값이 진정한(true) 평균값에 보다 근접하게 만들며, 분산에 대한 불편추정치를 제공하게 된다[8].

REML에서 β에 대한 조정된 피셔 점수방정식은 다음과 같다.

$$\beta^{k+1} = W^{1/2} X (X^{T} W X)^{-1} X^{T} W^{1/2} z$$

여기서 가중치 대각행렬 W, 분산함수 $V(\mu)$, z_i는 ML에서와 동일하다.

---

7) 반복의 초기값은 , $\mu_i = y_i$, $\phi_i = 1$로 설정

8) REML의 장점과 β와 γ의 피셔 점수방정식 조정에 대한 자세한 설명은 Lee-Nelder(1998), Smyth-Verbyla(1999), Smyth-Huele-Verbyla(2001)을 참조하기 바람.

$\emptyset_i$의 근사 불편추정치는 γ에 대한 피셔 점수방정식을 다음과 같은 방식으로 수정하여 구할 수 있다.

$$\gamma^{K+1} = (Z^T W_d^* Z)^{-1} Z^T W_d^* z_d^*$$

여기서 W_d^*는 다음과 같은 가중치 대각행렬이고,

$$W_d^* = diag\left(\frac{\partial g_d(\emptyset_i)}{\partial \emptyset}\right)^{-2} \frac{|\omega d_i - h_i|_+}{2V_d(\emptyset_i)}$$

$$|\omega d_i - h_i|_+ = Max[\omega_{d_i} - h_i, 0]$$

분산함수 $V_d(\emptyset_i) = \emptyset_i^2$, $z_{d_i}^* = \left\{\frac{\partial g_d(\emptyset_i)}{\partial \emptyset}\right\}(d_i^* - \emptyset_i) + g_d(\emptyset_i)$이며, d_i는 다음과 같이 조정한다.

$$d_i^* = \frac{\omega_{d_i}}{\omega d_i - h_i} d_i$$

한편, $\hat{\gamma}$의 표준오차는 다음과 같은 피셔 정보행렬의 역행렬로부터 얻어진다.

$$\mathfrak{I}_\gamma = Z^T W_d^* Z$$

III. 실증분석

1. 자료

본 연구에서 DGLM 모형에 사용한 자동차보험 포트폴리오는 2007년 한국의 제3자배상 자동차보험(책임보험) 자료와 2004년 호주 뉴사우스웨일즈州의 제3자배상 자동차보험(책임보험) 포트폴리오 자료이다. 국내 자료는 위험할인(증) 변수로 성별, 운전자 연령, 차종, 거주지역에 관한 정보를 제공하고 있다. 또한 각 범주별로 계약건수(보험계약연도 기준), 청구건수, 총청구금액(보험금 지급금액)을 포함하고 있다. 운전자 연령은 20대, 30대, 40대, 50대, 60대이상 총 5개 범주로 구분하였다[9].

차종은 승용차, 승합차, 화물차, 건설기계, 특수차, 이륜차 총 6개 범주로 분류된다. 마지막으로 운전자의 거주지역 정보는 9개 도(道)와 7개 광역시를 합해 총 16개 지역으로 구분되어 있다. 사용된 자료의 특성을 요약하면 총계약건수는 15,551,662개이고, 총청구건수는 923,381개이다. 분류범주는 총 960개(=성별(2개)×연령(5개)×차종(6개)×지역(16개))이며, 이중에서 청구건수가 "0"인 범주는 53개이다.

호주 자료의 경우, 위험할인(증) 변수로 성별, 차령(車齡, vehicle age), 운전자 연령, 거주지역을 사용하였다. 이 자료 역시 각 범주별로 계약건수, 청구건수, 총청구금액을 포함하고 있다. 차령은 총 4개 범주로 구분되어 있으며, 1년, 2년, 3년, 4년 이상으로 구분되어 있다. 운전자 연령은 10대, 20대, 30대, 40대, 50대, 60대 이상 총 6개 범주로 분류하였다. 마지막으로 거주지역 정보는 총 6개 지역으로 구분된다. 사용된 자료의 특성을 요약하면 총계약건수는 65,535개이고, 총청구건수는 4,762개이다. 분류범주는 총 288개(=성별(2개)×차령(4개)×연령(6개)×지

9) 국내의 경우 외국과 달리 10대 운전자의 보험계약건수가 매우 적어 분석대상에서 제외하였다.

역(6개))이며, 이중에서 청구건수가 "0"인 범주는 8개이다. 본 논문에서는 총청구비용뿐만 아니라 각 범주별 청구건수 정보를 사용하여 식 (1)과 (2) 구조의 DGLM 모형을 구성하여 REML 방식으로 평균과 평균의 분산 모수들을 추정하였다. 공변량 벡터는 각 범주들의 조합으로 구성되어 있으며[10], 평균과 분산 모두 대수선형 모형을 적용하였다.

2. 추정 결과

〈표 1〉은 국내 자료를 사용한 DGLM에서 추정된 보험료의 평균과 분산 회귀계수들, 그리고 표준보험료(Base Rate)와 각 범주별 위험할인(증) 요소를 보여준다[11]. 운전자 연령은 20대를 제외하면 보험료 평균 기대값(μ_i)을 증가시키는 반면 기대값의 분산을 감소시키고 있음을 알 수 있다. 이 공변량은 특정구간, 즉 20대를 제외하면 보험료와 단조함수적인 관계를 나타내고 있다. 이는 운전자의 연령이 증가할수록 평균 청구빈도는 많아지나 건당 청구규모(severity)는 감소함을 의미한다. 이는 보험료 추정시 평균뿐만 아니라 분산의 모델링이 중요함을 확인시켜준다. 차종의 경우에서 분산의 모델링이 더 중요함을 알 수 있다. 이륜차를 제외한 모든 차종이 승용차에 비해 보험료 평균 기대값을 증가시키고 있으나, 분산의 경우 건설기계와 특수차는 증가시키는 반면 승합차와 화물차는 분산을 감소시키는 요인으로 작용하고 있다. 이러한 실증결과는 청구빈도와 규모가 동시에 증가 또는 감소하지 않는다는 것을 보여주며, 보험금 청구 데이터의 분산이 일정한 상수가 아님을 의미한다. 이는 곧 보험료 모델링에 있어 평균뿐만 아니라 분산을 함께 고려하는 것이 추정치의 정확도를 향상시키는데 중요한 역할을 하고 있

10) 공변량 벡터는 한국 자료의 경우 성별, 운전자 연령, 차종, 지역으로 구성되며, 호주 자료의 경우는 성별, 차령(車齡), 운전자 연령, 지역이다.

11) 청구빈도와 총 청구비용 자료를 모두 사용하였으며, REML 방식으로 모수를 추정하였다. 한편 모든 회귀계수들이 5% 유의수준하에서 유의한 결과를 나타내어 표에서는 별도의 표시를 하지 않았다.

음을 알려준다. 성별의 경우, 여성이 남성보다 보험료 평균의 기대값은 높지만 분산은 작게하는 요인으로 나타났으나 그 차이는 크지 않았다. 운전자의 거주지역별 보험료 할인(할증)을 살펴보면 동일범주 내에서의 최대 차이가 0.468로 성별에 비해 훨씬 큰 격차를 보이고 있어, 최근 이슈가 되고 있는 지역별 보험료 차등제 도입의 필요성이 학술적인 측면에서 일부 인정된다고 할 수 있다. 하지만 DGLM을 이용한 본 논문의 실증분석 결과는 그동안 논란의 핵심이었던 수도권과 지방으로의 양극화된 차등과는 다른 양상을 보여주고 있다. 보험금 청구 데이터의 평균과 분산을 동시에 고려할 경우 경기도, 전라북도, 광주광역시, 서울특별시의 할인(할증)율이 거의 비슷한 반면, 경기도, 서울특별시와 같은 수도권인 인천광역시의 할인(할증)율이 상당부분 차이가 난다는 것이다. 또한 전반적으로 보면 지방에 비해 수도권의 할증율이 더 높게 추정되었다 [12]. 강원도 거주 30대 남성의 승용차에 대한 표준보험료는 116,283.6원으로 추정되었으며, 다른 특정 범주에 속하는 물건에 대한 보험료는 표준보험료에 각 범주별 위험할인(증) 요소를 곱하여 산출하게 된다[13].

〈표 2〉는 호주 자료를 사용한 DGLM에서 추정된 보험료의 평균과 분산 회귀계수들, 그리고 표준보험료와 각 범주별 위험할인(증) 요소를 보여준다[14]. 연령의 경우 나이가 증가함에 따라 보험료 평균 기대값(μ_i)을 감소시키는 요인으로 작용하고 있으며, 이 공변량은 특정구간, 즉 50대를 제외하면 보험료와 단조함수적인 관계를 나타내고 있다. 한편 연령은 기대값의 분산을 증가시키고 있음을 알 수 있다. 이는 연령이 증가할수록 평균 청구건수는 작아지나 건당 청구금액은 증가함을 의미하여 한국의 경우와 반대의 현상을 보여주고 있다. 성별의 경우 역시, 한국과는 달리 여

12) 그러나 본 논문에 사용된 자료는 종합보험료가 아닌 책임보험료로 한정되어 있으며, 지역별 보험료 차등제의 실행은 지역 간 교통시설 인프라 차이 등을 종합적으로 고려해야 할 사안임을 밝혀둔다.

13) 이는 위험설명변수 4개를 고려한 경우이며, 실제 보험료 산출 시에는 보다 많은 위험 설명변수들을 사용하게 된다.

14) 역시 청구빈도와 총 청구비용 자료를 모두 사용하였으며, REML 방식으로 모수를 추정하였다. 한편 모든 회귀계수들이 5% 유의수준하에서 결과를 나타내어 표에서는 별도의 표시를 하지 않았다.

성 보다 남성 운전자일수록 보험료 평균의 기대값이 높고, 기대값의 분산 역시 크게 나타났다. 이는 남성의 청구건수와 건당 청구금액이 모두 여성 운전자에 비해 크기 때문인 것으로 해석된다. 이러한 실증결과 역시 보험료 추정 시 평균뿐만 아니라 분산의 모델링이 중요함을 재차 확인시켜 주고 있다. 차령과 지역의 경우는 보험료와 특별한 단조함수 관계를 형성하고 있지는 않지만 1년 이내와 D지역의 평균 청구금액이 가장 작은 것을 알 수 있다. 한편 동일범주 내에서의 최대할인율 차이를 살펴보면, 지역과 연령이 차령과 성별보다 더 중요한 위험할인(증) 요소임을 알 수 있다. 호주 자료의 제3자배상 자동차보험 표준보험료는 104.9 호주달러(AUD)로 추정되었으며, 차령이 1년 이내이고 D지역에 거주하는 50대 여성운전자의 보험료가 가장 저렴하게 책정될 수 있는 것으로 나타났다.

IV. 결론

본 논문은 손실함수가 Compound 포아송 분포를 따르는 경우 적정 자동차 보험료를 추정함에 있어 DGLM을 이용하여 평균뿐만 아니라 평균의 분산까지 모델링함으로써 보험요율 추정의 정확도를 향상시킬 수 있는 대안에 대하여 고찰해 보았다. 또한 한국과 호주의 제3자배상 자동차보험 포트폴리오를 이용하여 추정된 실증분석 결과를 통해 특정 공변량들의 경우 청구빈도와 규모에 동일한 또는 반대의 영향을 미치고 있음을 확인하였다. 또한 실증분석 결과는 동일한 분류 범주 내에서도 평균 기대값에 미치는 영향과 기대값의 분산에 미치는 영향이 상이할 수 있음을 보여 주었다. 한편 연령과 성별에서 알 수 있었듯이 동일한 위험할인(할증) 요인일 지라도 국가 간에 차이가 존재하고 있었다. 이러한 결과로부터 보험금 청구 데이터의 분산을 일정한 상수로 가정하는 모형보다 분산을 동시에 고려한 모형이 보험료 추정치의 정확도를 향상시키는데 중요한 역할을 하고 있음을 알 수 있다.

한편 보험료 할인(할증) 위험요인들 간에도 상당한 편차가 존재함을 알 수 있었다. 실증분석에 사용된 여러 위험요인 중 차종의 최대할인율차이가 가장 크게 나타났으며, 지역변수 역시 운전자의 연령에 못지않게 중요한 위험 할인(할증) 요인인

것으로 나타났다. 본 논문의 실증분석 결과에 따르면 지역별 자동차보험료 차등화 방안에 있어 그동안 논란의 중심이었던 수도권과 지방간 보험료 양극화 우려의 문제는 심각하게 발생하지 않을 것으로 예상된다. 아직까지 국내에서는 지역에 따른 보험료 차등제도가 시행되고 있지 않지만, 보험요율 자유화의 취지를 살리기 위해서는 보다 다양한 위험할인(증) 요인들을 실증분석에 기반한 유효성 검증을 통해 반영할 필요가 있다. 이를 위해서는 국내의 자동차 종합보험 포트폴리오를 이용하여 보다 다양한 위험요인들을 반영한 실증분석이 이루어져야 할 것으로 사료된다. 여타 다른 추정 모형들과의 비교분석이 이루어지지 못한 부분은 본 논문의 한계점으로 밝혀두며, 이 역시 추후의 연구과제로 돌리고자 한다.

〈표 1〉 한국 제3자배상 자동차보험료 추정결과

분류 범주	회귀 계수		표준보험료와	동일범주內
	평균(β)	분산(γ)	위험할인(증)	최대할인율차이
표준보험료	-	-	116,283.6(원)	-
상수항	11.786	7.566	-	-
성별: 남성	0.000	0.000	1.000	0.091
성별: 여성	0.087	-0.121	1.091	(여성 - 남성)
연령: 20대	0.539	-0.327	1.715	
연령: 30대	0.000	0.000	1.000	0.715
연령: 40대	0.052	-0.031	1.054	(20대 - 30대)
연령: 50대	0.115	-0.040	1.122	
연령: 60대 이상	0.209	-0.102	1.234	
차종: 승용차	0.000	0.000	1.000	
차종: 승합차	0.618	-0.137	1.856	
차종: 화물차	0.285	-0.015	1.329	2.003
차종: 건설기계	0.972	0.037	2.643	(건설기계 - 이륜차)
차종: 특수차	0.122	0.429	1.129	
차종: 이륜차	-0.446	0.173	0.640	

주: 1) 총청구비용과 청구빈도 자료를 함께 고려했음.
2) 성별은 여성, 연령은 40대, 차종은 승용차를 기준(base)으로 하였음.

분류 범주	회귀 계수		표준보험료와 위험할인(증)	동일범주內 최대할인율차이
	평균(β)	분산(γ)		
표준보험료	-	-	116,283.6(원)	-
상수항	11.786	7.566	-	-
지역: 강원	0.000	0.000	1.000	0.468 (인천 - 제주)
지역: 경기	0.103	-0.246	1.108	
지역: 경남	-0.121	0.041	0.886	
지역: 경북	-0.088	0.075	0.916	
지역: 광주	0.104	-0.222	1.110	
지역: 대구	-0.116	-0.198	0.889	
지역: 대전	0.041	-0.263	1.042	
지역: 부산	-0.084	-0.184	0.919	
지역: 서울	0.105	-0.322	1.111	
지역: 울산	-0.196	-0.115	0.822	
지역: 인천	0.212	-0.386	1.236	
지역: 전남	0.092	0.118	1.096	
지역: 전북	0.101	-0.094	1.106	
지역: 제주	-0.264	0.193	0.768	
지역: 충남	0.068	0.003	1.071	
지역: 충북	-0.017	-0.041	0.983	

주: 1) 총청구비용과 청구빈도 자료를 함께 고려했음.
2) 성별은 여성, 연령은 40대, 차종은 승용차를 기준(base)으로 하였음.

〈표 2〉 호주 뉴사우스 웨일즈州 제3자배상 자동차보험료 추정결과

분류 범주	회귀 계수		표준보험료와 위험할인(증)	동일범주內 최대할인율차이
	평균(β)	분산(γ)		
표준보험료	-	-	104.9	-
상수항	4.654	5.204	-	-
성별: 여성	0.000	0.000	1.000	0.183
성별: 남성	0.168	0.042	1.183	(남성 - 여성)
차령(車齡): 1년	0.000	0.000	1.000	
차령(車齡): 2년	0.178	-0.065	1.195	0.195
차령(車齡): 3년	0.088	0.024	1.092	(2년 - 1년)
차령(車齡): 4년 이상	0.111	0.089	1.118	
연령: 10대	0.487	-0.069	1.627	
연령: 20대	0.129	-0.005	1.137	
연령: 30대	0.036	-0.012	1.036	0.898
연령: 40대	0.000	0.000	1.000	(10대 - 50대)
연령: 50대	-0.315	0.118	0.729	
연령: 60대 이상	-0.221	0.147	0.802	
지역: A	0.000	0.000	1.000	
지역: B	0.022	-0.041	1.022	
지역: C	0.076	0.013	1.079	0.675
지역: D	-0.101	0.076	0.904	(F지역 - D지역)
지역: E	0.137	0.063	1.147	
지역: F	0.457	0.021	1.579	

주: 1) 총청구비용과 청구빈도 자료를 함께 고려했음.
2) 차령은 1년 이하, 성별은 여성, 연령은 40대를 기준(base)으로 하였음.
3) 표준보험료는 호주달러('08년 10월 1일 기준 1AUD=953.36원)

참 고 문 헌

Andrews, D.E, and Herzberg, A.M., "Data: A collection of problems from many fields for the student and research worker" Springer-Vedag, New York, 1985, pp. 413~421.

Hallin, M., and Ingenbleek, J.E, "The Swedish automobile portfolio in 1977: A statistical study" *Scandinavian Actuarial Journal*, 1983, pp. 49~64.

Jorgensen, B., "Exponential dispersion models" *Journal of Royal Statistical Society B 49*, 1987, pp. 127~162.

_____, "Theory of Dispersion Models" Chapman & Hall, London, 1997.

Jorgensen, B. and De Souza, M.C.P., "Fitting Tweedie's compound Poisson model to insurance claims data" *Scandinavian Actuarial Journal*, 1994, pp. 69~93.

Lee, Y. and Nelder, J.A., "Generalized linear models for the analysis of quality-improvement experiments" *Canadian Journal of Statistics 26*, 1998, pp. 95~105.

Millenhall, S.J., "A systematic relationship between minimum bias and generalized linear models" *Proceedings of the Casualty Actuarial Society 86*, 1999, pp. 393~487.

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

Nelder, J.A., and Lee, Y., "Likelihood, quasi-likelihood and pseudo-likelihood: some comparisons" *Journal of Royal Statistical Society B 54*, 1992, pp. 273~284.

Nelder, J.A. and Pregibon, O., "An extended quasi-likelihood function" *Biometrik 74*, 1987, pp. 221~231.

Renshow, A.E., "Modelling the claims process in the presence of covariates" *ASTIN Bulletin 24*, 1994, pp. 265~286.

Rolski, T., Schmidli, n., Schmidt, V., and Teugels, J., "Stochastic Processes for

Insurance and Finance" John Wiley & Sons, Chichester, 1999.

Smyth, G.K., "Generalized linear models with varying dispersion" *Journal of Royal Statistical Society B 51*, 1989, pp. 47~60.

_____, "Partitioned algorithms for maximum likelihood and other nonlinear estimation" *Statistics and Computing 6*, 1996, pp. 201~216.

Smyth, G.K., Huele, F., and Verbyla, A.P., "Exact and approximate REML for heteroscedastic regressionn" *Statistical Modelling 1*, 2001, pp. 161~175.

Smyth, G.K. and Jorgensen, B., "Fitting Tweedie's Compound Poisson Model to Insurance Claims data : Dispersion Modelling" *ASTIN Bulletin 32*, 2002, pp. 143~157.

Smyth, G.K., and Verbyla, A.P., "Adjusted likelihood methods for modelling dispersion in generalized linear models" *Environments 10*, 1999, pp. 696~709.

Abstract

This study estimates the rate of motor insurance premium by using Double Generalized Linear Model(DGLM) under Tweedie's compound Poisson loss function. Our model used can improve accuracy of estimating fair motor insurance premium by modeling dispersion of the insurance claims as well as their mean. We empirically estimate the rates of premium by using third party motor insurance data for Korea and Australia. The results provide some covariates which explain risk factors simultaneously affect same or opposite direction to the frequency and size of claims. This implies that the DGLM can improve the precision of estimates for car insurance premium rather than assuming constant dispersion.

We also observe that a meaningful difference between risk factors of premium in the extent of maximum deviation within the category. The kind of vehicle body displays the largest deviation in discounting(or raising) premium across the risk factors used. The empirical results also shows that the driver's area of residence is also important risk factor as good as the driver's age. From our empirical results, a system which discriminates insurance premium by driver's area of residence could not cause serious polarization problem widening premium gaps between metropolitan areas and provinces.

※ Key Words: Compound Poisson Distribution, Double Generalized Linear Model(DGLM), Motor Insurance Premium