The Korean Journal of Applied Statistics
2021, Vol. 34, No. 5, 697–710

DOI: https://doi.org/10.5351/KJAS.2021.34.5.697
Print ISSN1225-066X / Online ISSN 2383-5818

Predicting claim size in the auto insurance with relative error: a panel data approach

Heungsun Park[1,a]

[a]Department of Statistics, Hankuk University of Foreign Studies

Abstract

Relative error prediction is preferred over ordinary prediction methods when relative/percentile errors are regarded as important, especially in econometrics, software engineering and government official statistics. The relative error prediction techniques have been developed in linear/nonlinear regression, nonparametric regression using kernel regression smoother, and stationary time series models. However, random effect models have not been used in relative error prediction. The purpose of this article is to extend relative error prediction to some of generalized linear mixed model (GLMM) with panel data, which is the random effect models based on gamma, lognormal, or inverse gaussian distribution. For better understanding, the real auto insurance data is used to predict the claim size, and the best predictor and the best relative error predictor are comparatively illustrated.

Keywords: best relative error predictor, best predictor, generalized linear mixed models, random effect, percentile error, panel data, claim size

1. 서론

예측(prediction)이란 '주어진 정보를 가지고 위험(risk)을 최소화하면서 임의의 확률변수를 추정하는 통계적 과정'이라고 할 수 있다. 만일 예측하고자 하는 확률변수가 Y이고, 독립변수가 X, 그리고 예측변수(predictor)가 $\tilde{Y}$라고 할 때, 일반적인 예측오차의 손실함수(loss function)는 $(Y-\tilde{Y})^2$이며, 이때 최량예측량(best predictor)은,

$$\mathrm{E}\left\{\left(Y-\tilde{Y}\right)^2\middle| X\right\} \tag{1.1}$$

을 최소화시키는 $\mathrm{E}(Y|X)$임을 알고 있다 (Bickel과 Docksum, 1977).

그러나 예측오차(prediction error), $(Y-\tilde{Y})$, 자체보다 미래값에 대한 상대오차(relative error), $(Y-\tilde{Y})/Y$, 에 더 관심이 있을 수 있다. 이는 상대오차가 단순히 척도와 무관하기 때문만이 아니라 (Chen 등, 2010), 어쩌면 행동과학(behavioral science)의 프로스펙트 이론(prospect theory)에서 말하듯이, 인간이 무의식적으로 사물을 절대적인 가치보다 상대적 가치에 의존해 의사결정을 하는 인지습성 때문일 수도 있다 (Kahneman과 Tversky, 1979). 이와 같은 상대오차는 이미 화학이나 공학분야에서 퍼센트오차(percent/percentage error),

$$\frac{|V_{\text{obs.}}-V_{\text{true}}|}{V_{\text{true}}}\times 100 \tag{1.2}$$

This work was supported by Hankuk University of Foreign Studies Research Fund of 2020.

[1] Department of Statistics, Hankuk University of Foreign Studies, 81, Oedae-ro, Mohyeon-eup, Cheoin-gu, Yongin-si, Gyeonggi-do 17035, Korea. E-mail: hspark@hufs.ac.kr

Published 31 October 2021 / journal homepage: http://kjas.or.kr

로서 의미있는 측정오차의 척도로 사용되어 왔고, 계량경제학이나 시계열자료의 모형평가에서도 모형적합도를 평가함에 있어서 흔히 사용하는 mean square error (MSE) 외에, mean squared percentage error (MSPE), mean absolute percentage error (MAPE)의 모습으로 사용되어 오고 있다 (Makridakis, 1984).

Park과 Stefanski (1998)는 예측량 관점에서 상대오차를 접근했는데, 그들은 Y가 반응변수이고 $\tilde{Y}$가 예측변수일 때, mean squared relative error (MSRE)인,

$$\mathrm{E}\left\{\left(\frac{Y-\tilde{Y}}{Y}\right)^2 \middle| X\right\} \tag{1.3}$$

를 최소화시키는 $\tilde{Y}$는 다름아닌,

$$\tilde{Y} = \frac{\mathrm{E}\left(Y^{-1}\middle|X\right)}{\mathrm{E}\left(Y^{-2}\middle|X\right)} = \frac{\mathrm{E}\left(Y^{-1}\middle|X\right)}{\mathrm{V}\left(Y^{-1}\middle|X\right)+\left[\mathrm{E}\left(Y^{-1}\middle|X\right)\right]^2} \tag{1.4}$$

임을 증명하였고, 이에 대한 추정량으로써

$$\mathrm{E}(Y^{-1}|X) = f(\boldsymbol{x},\boldsymbol{\beta}), \tag{1.5}$$

$$\mathrm{V}(Y^{-1}|X) = \sigma^2(f(\boldsymbol{x},\boldsymbol{\beta}))^{2\theta}, \tag{1.6}$$

의 가정 하에, 준가능도함수(quasi-likelihood)와 의사가능도함수(pseudo-likelihood)를 이용한 $(\hat{\boldsymbol{\beta}}, \hat{\sigma^2}, \hat{\theta})$을 구하거나, 적률추정법을 이용한 아래 값을 추천하였다.

$$\hat{\tilde{Y}} := \arg\min_{g} \sum_{i=1}^{n}\left(\frac{Y_i - g\left(\boldsymbol{x}_i,\hat{\boldsymbol{\beta}}\right)}{Y_i}\right)^2 \tag{1.7}$$

여기서 $g(\cdot)$ 함수는 $\boldsymbol{x}$의 일반적인 다항식이다.

그 후, 상대오차 예측은 정상시계열자료와 (Park과 Shin, 2005), 커널(kernel)추정을 이용한 비모수 회귀모형으로 확장되었다 (Jones 등, 2008). 또한, Chen 등 (2010, 2016)은 주식가격(stock price)을 예측함에 있어서, Y가 실제 거래가격이고 $\tilde{Y}$가 그 주식의 내재적 가치(예측값)이라고 했을 때 매입자(buyer)의 상대이득/손실은 $|Y-\tilde{Y}|/Y$이지만, 매도자(seller)의 상대손실/이득은 $|Y-\tilde{Y}|/\tilde{Y}$로 상대오차가 '비대칭적 (asymmetric)'인 점을 감안하여, least absolute relative error (LARE)

$$\mathrm{LARE}_n(\boldsymbol{\beta}) := \sum_{i=1}^{n}\left\{\left|\frac{Y_i - \exp\left(\boldsymbol{X}^T\boldsymbol{\beta}\right)}{Y_i}\right| + \left|\frac{Y_i - \exp\left(\boldsymbol{X}^T\boldsymbol{\beta}\right)}{\exp\left(\boldsymbol{X}^T\boldsymbol{\beta}\right)}\right|\right\}, \tag{1.8}$$

혹은 least product relative error (LPRE)

$$\mathrm{LPRE}_n(\boldsymbol{\beta}) := \sum_{i=1}^{n}\left\{\left|\frac{Y_i - \exp\left(\boldsymbol{X}^T\boldsymbol{\beta}\right)}{Y_i}\right| \times \left|\frac{Y_i - \exp\left(\boldsymbol{X}^T\boldsymbol{\beta}\right)}{\exp\left(\boldsymbol{X}^T\boldsymbol{\beta}\right)}\right|\right\}, \tag{1.9}$$

을 최소화시킬 것을 제안하고 있다. 그러나 이와 같은 LARE나 LPRE는 상대오차의 서로 다른 관점(매입자, 매도자)이 요구되거나, 생존분석의 가속수명시간(accelerated failure time, AFT)의 승법모형(multiplicative model)에 주로 적용되기 때문에 (Wang 등, 2018) 본 논문에서는 다루지 않게 될 것이다.

2. 사고 손해액 예측을 위한 상대오차예측량

자동차 보험업계에서 보험상품의 적정 손해율(손해액을 보험가액으로 나눈 값)을 예측하는 것은 매우 중요한 업무이다. 이때 순보험료(pure premium)는 과거 요율의 존재여부와 관계없이 요율을 새롭게 산출하는 방식으로, 사고발생 빈도(자주 발생하는 정도)와 심도(손해액 크기)가 일정기간 동안 변하지 않는다는 것을 전제함으로써, 동일 위험집단에 대한 정보를 가지고 손해빈도(Claim Frequency)와 손해심도 혹은 손해액(Claim Severity/Claim Size)을 이용해서

$$\text{Pure Premium} = (\text{Claim Frequency}) \times (\text{Claim Size}), \tag{2.1}$$

로 계산된다 (Werner와 Modlin, 2009; Frees, 2018). 여기서 손해빈도에 대한 추정은 포아송과정 (Huang 등, 2009)이나 확률과정론으로 이루어졌으며 (Wüthrich와 Merz, 2008), Jørgensen과 Souza (1994)는 감마확률변수의 포아송 합(poisson sum of gamma random variables)을 이용한 Tweedie 분포 (Tweedie, 1984)를 이용해서 손해빈도와 손해크기를 동시에 추정하였으나, 빈도와 심도에 끼치는 독립변수가 동일해야 하는 제한점 때문에 (Smyth와 Jørgensen, 2002) 실제로 보험업계에서 많이 사용되는 방법은 빈도와 심도를 따로 모형화시켜 대입하는 방법이다 (Kim과 Kim, 2009). 심도 추정에는 오른쪽으로 많이 기운 분포인 감마회귀모형(gamma regression), 로그정규회귀모형(lognormal regression), 역가우스회귀모형(inverse gaussian regression) 등이 사용되고 있다 (Jong과 Heller, 2008).

자동차 보험의 손해액을 예측함에 있어서 주의 깊게 고려되어야 할 사항 중 하나는 올바른 독립변수의 선택이라고 할 수 있다. 통상적으로 자동차 보험가입자의 손해액은 운전자의 인구통계학적인 특성은 물론, 차량의 연식, 차량의 종류(세단, 스포츠카, 트럭 등)와 사고의 노출정도(exposure)에 따라 다를 수 있으므로 이러한 독립변수를 손해액 추정모형에 고려해야 할 것이다. 통계학에서 동질집단은 흔히 군집(cluster) 혹은 코호트(cohort)로 분류되기 때문에, 같은 군집 내 자료 간에는 내재적 속성을 공유하는 상관관계를 고려한 패널자료(panel data)분석이 필수적이다.

또한, 손해보험회사는 과거의 손해자료를 이용해서 미래의 손해액을 예측하는 것을 중요한 분석과정 중 하나로 여기고 있다. Boland (2007)는 군집의 평균손해액 추정 외에도, 사고로부터 발생한 손해액 자료를 효과적으로 다루기 위해서는 보험사가 반드시 미래의 발생할 개별 보험상품의 손해액에 대한 예측(prediction)을 수행함으로써 그 위험에 대한 적절한 대비를 해야 한다고 주장한다. 정확한 예측방법은 곧 보험상품의 가격(pricing)과 회사의 지불능력(solvency)에 영향을 주기 때문에 보험분야에서는 여러 가지 예측방법에 대한 연구가 계속 진행되고 있다 (Hong과 Martin, 2019). 최근 손해보험과 생명보험에서는 뛰어난 컴퓨터의 발전과 방대한 자료를 이용한 '예측분석(predictive analytics)'을 강조하고 있으며, Telematics기법을 통해 블랙박스나 GPS를 통한 개인 운전자의 운전습성 정보를 이용해 손해액을 예측하는 시스템의 개발도 이루어지고 있다 (Boodhun과 Jayabalan, 2018). 이와 같이 보험가입자나 보험상품에 대한 손해액 예측은 보험회사의 비지니스 효율성(business efficacy) 증대를 위해 필수적인 요소로 여겨지며, 보험업자(underwriter)는 보험상품의 위험(risk)을 적절하게 평가하고, 보험회사가 잘 운영될 수 있도록 손해액(claim size)을 정확하게 예측해야 할 필요가 있다.

본 논문은 Park과 Stefanski (1998)의 상대오차예측법을 임의효과가 포함된 일반화선형혼합모형(generalized linear mixed model, GLMM)에 확장시키고, 자동차 보험가격 산정에서 손해액 추정에 널리 사용되는 감마회귀모형(gamma regression model), 로그정규회귀모형(lognormal regression), 역가우스회귀모형(inverse gaussian regression)에 대해 적용시키되, R 패키지 `insuranceData`에 저장되어 있는 `dataCar`라는 실제 자동차 보험자료를 사용하였다.

2.1. 일반화선형혼합모형

일반화선형모형(generalized linear model, GLM)에서 회귀계수가 임의의 확률분포에서 추출된 값이라고 가정할 때, 그 회귀계수는 임의효과(random effect)라 하고, 같은 회귀계수를 가진 군집 내 반응값들은 서로 상관관계가 존재하게 되며, 이를 일반화선형혼합모형(generalized linear mixed model, GLMM)이라고 한다. 다시 말하면, 독립변수 $\boldsymbol{z}$의 회귀계수 $\boldsymbol{\gamma}$가 $N(\mathbf{0}, G)$을 따른다고 할 때, 반응변수의 기대값 μ는 연결함수 $g(\cdot)$를 이용하여

$$g(\mu) = \boldsymbol{x}^T\boldsymbol{\beta} + \boldsymbol{z}^T\boldsymbol{\gamma}, \quad \boldsymbol{\gamma} \sim N(\mathbf{0}, G), \tag{2.2}$$

라고 할 수 있으며, 이를 간단히 예를 들어 설명하면, $\boldsymbol{x} = (1\ x)^T$와 $\boldsymbol{\beta} = (\beta_0\ \beta_1)^T$일 때, 임의계수모형(random coefficient model)은

$$g(\mu) = (\beta_0 + \gamma_0) + (\beta_1 + \gamma_1)x, \quad \boldsymbol{\gamma} = \begin{pmatrix} \gamma_0 \\ \gamma_1 \end{pmatrix} \sim N_2(0, G), \tag{2.3}$$

가 되며, $\boldsymbol{\gamma}_i = \gamma_{0i}$일 때 임의절편모형(random intercept model)이 된다 (Jong과 Heller, 2008).

이때, i번째 군집의 j번째 관측치를 y_{ij}라고 하면, $\boldsymbol{\gamma}_i$가 주어졌을 때 조건부 확률밀도함수 $f(y_{ij}|\boldsymbol{\gamma}_i)$가 지수족함수(exponential family)에 속하고 $\boldsymbol{\gamma}_i$가 다변량정규분포이고 $g(\cdot)$함수가 로그함수이면,

$$\mu\,|\,\boldsymbol{\gamma}_i = \exp\left(\boldsymbol{x}^T\boldsymbol{\beta} + \boldsymbol{z}^T\boldsymbol{\gamma}_i\right), \quad \boldsymbol{\gamma}_i \sim N(\mathbf{0}, G) \tag{2.4}$$

가 되고, 군집의 수가 a, 군집의 크기가 n_i인 경우의 결합가능도함수(joint likelihood function)는

$$L(\boldsymbol{\beta}) = \Pi_{i=1}^{a}\Pi_{j=1}^{n_i} \int f(y_{ij}|\boldsymbol{\gamma}_i)\, f(\boldsymbol{\gamma}_i)\, d\boldsymbol{\gamma}_i, \tag{2.5}$$

로 구현되지만 $L(\boldsymbol{\beta})$를 최대화시키기 위해선 그 가운데 적분값을 구하는 어려움이 존재한다. 이 적분값을 수치해석적으로 구하기 위한 대표적 연구는 Breslow와 Clayton (1993), Wolfingeer 와 O'Connell (1993) 등에 의해 SAS의 `proc glimmix`로 구현되었고, `proc nlmixed` 프로시져는 Gauss-Hermite Quadrature 알고리즘을 적용하여 이 적분문제를 해결하였다 (Golub과 Welsch, 1969; Davidian과 Giltinan, 1995; SAS/STAT® User's Guide, 2012). 또한, 임의효과가 비정규분포를 따른다면 다단계 일반화선형모형(hierarchical generalized linear model)을 사용할 수 있다 (Lee와 Nelder, 1996, 2001).

그리고 `proc glimmix`나 `proc nlmixed` 모두, 임의효과의 추정을 위해 사후확률인,

$$f(\boldsymbol{\gamma}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\gamma})\, f(\boldsymbol{\gamma})}{f(\boldsymbol{y})} \tag{2.6}$$

을 최대로 하는 $\hat{\boldsymbol{\gamma}}$를 선택하는 경험적 베이즈 추정량(empirical Bayes estimator)을 수치적 방법으로 제공하고 있다.

2.2. 감마회귀모형과 상대오차예측량

감마회귀모형은 독립변수 $\boldsymbol{x}$와 반응변수 y에 대한 일반화선형모형(generalized linear model)의 한 형태로서 다음과 같은 형태를 취하고 있다.

$$y \sim \text{Gamma}\,(\mu, \nu), \quad g(\mu) = \boldsymbol{x}^T\boldsymbol{\beta}. \tag{2.7}$$

여기서 $g(\cdot)$는 연결(link)함수이고 감마회귀모형의 정준연결함수는 역함수이지만, 편의상 로그연결함수를 사용하며, y의 확률밀도함수는 아래와 같이 정의되고,

$$f_Y(y) = \frac{y^{-1}}{\Gamma(\nu)} \left(\frac{\nu}{\mu} y\right)^{\nu} e^{-\frac{\nu}{\mu} y}, \quad y > 0, \tag{2.8}$$

평균과 분산은 다음과 같다.

$$\mathrm{E}(y) = \mu, \quad \mathrm{V}(y) = \frac{\mu^2}{\nu}. \tag{2.9}$$

윗 식에서 ν 값이 작을수록 분산은 커져서 꼬리가 길고 오른쪽으로 기운분포를 하게 된다. 따라서, 감마분포의 성질에 의해 $1/y$는 Inv.Gamma(μ, ν)인 역감마분포(inverse gamma distribution)를 따르며 적률값은,

$$\mathrm{E}\left(\frac{1}{y}\right) = \frac{\nu}{(\nu - 1)\mu}, \quad \nu > 1 \tag{2.10}$$

$$\mathrm{E}\left(\frac{1}{y^2}\right) = \frac{\nu^2}{(\nu - 1)(\nu - 2)\mu^2}, \quad \nu > 2 \tag{2.11}$$

가 되어, 감마회귀모형의 최량상대오차예측량(best relative error predictor, BREP)을 구하면

$$\mathrm{BREP} := \frac{(\nu - 2)}{\nu} \mu, \quad \nu > 2, \tag{2.12}$$

라고 계산된다. 따라서, 만일 $\mu = \exp(\boldsymbol{x}^T \boldsymbol{\beta})$이고 $\nu = 1/\sigma^2$ 이라면,

$$\mathrm{E}(y|\boldsymbol{x}) = e^{x^T \beta}, \tag{2.13}$$

$$\mathrm{V}(y|\boldsymbol{x}) = \sigma^2 e^{2x^T \beta}, \tag{2.14}$$

가 되고, 감마회귀모형의 최량상대오차예측량은 다음과 같다.

$$\mathrm{BREP}: := \frac{\mathrm{E}\left(y^{-1} \middle| \boldsymbol{x}\right)}{\mathrm{E}\left(y^{-2} \middle| \boldsymbol{x}\right)} = \left(1 - 2\sigma^2\right) e^{x^T \beta}, \quad \sigma^2 < \frac{1}{2}. \tag{2.15}$$

그러므로, 패널자료에서 감마회귀모형에 임의효과 $\boldsymbol{\gamma}$가 더해질 경우, 조건부 평균과 기대값은

$$\mathrm{E}(y|\boldsymbol{x}, \boldsymbol{\gamma}) = e^{x^T \beta + z^T \gamma}, \tag{2.16}$$

$$\mathrm{V}(y|\boldsymbol{x}, \boldsymbol{\gamma}) = \sigma^2 e^{2(x^T \beta + z^T \gamma)}, \tag{2.17}$$

가 되어, 감마분포의 패널자료에서 최량상대오차예측량은 $\boldsymbol{\gamma}$의 분포와 관계없이, 조건부 적률값을 이용하면

$$\mathrm{BREP}: = \frac{\mathrm{E}\left(y^{-1} \middle| \boldsymbol{x}, \boldsymbol{\gamma}\right)}{\mathrm{E}\left(y^{-2} \middle| \boldsymbol{x}, \boldsymbol{\gamma}\right)} = \left(1 - 2\sigma^2\right) e^{x^T \beta + z^T \gamma}, \quad \sigma^2 < \frac{1}{2}, \tag{2.18}$$

가 된다. 따라서 우리는 식 (2.4)–(2.6) 에서 구한 $\hat{\boldsymbol{\beta}}$, $\hat{\sigma^2}$와 $\hat{\boldsymbol{\gamma}}$를 대입하여 그 추정치를 구하게 된다. 그러나 만일 $\boldsymbol{\gamma}$가 정규분포가 아닌 다른 분포를 가진다면, Lee와 Nelder (1996, 2001)에 의한 다단계 일반화선형모형에서 구한 추정치를 여기에 넣을 수 있다. 본 논문의 자동차 보험의 손해액 예측자료에서는 `area`라는 변수가 정규분포에서 추출된 서로 다른 6개의 지역을 나타내는 임의효과로 가정함에 무리가 없기 때문에 `proc glimmix`의 추정값을 그대로 사용하였다.

2.3. 로그정규회귀모형과 상대오차예측량

감마분포와 같이 오른쪽으로 많이 기운 분포 중 대표적인 분포는 로그정규(lognormal)분포이다. 로그정규분포를 이용한 회귀모형도 보험의 손해액 추정에 많이 사용되고 있다 (Jong과 Heller, 2008). 만일 확률변수 y가 $Lognormal(\mu, \sigma^2)$을 따르면, 확률밀도함수는

$$f_Y(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log y - \mu)^2}{2\sigma^2}\right), \quad 0 < y < \infty, \tag{2.19}$$

이고, 적률값은

$$\mathrm{E}(y^r) = \mathrm{E}\left\{e^{r(\mu+\sigma Z)}\right\} = \exp\left(r\mu + \frac{r^2\sigma^2}{2}\right), \quad r = \pm 1, \pm 2, \ldots, \tag{2.20}$$

가 되어 로그연결함수를 이용해

$$\begin{aligned} \mathrm{E}(y|\boldsymbol{x}) &= e^{\boldsymbol{x}^T\boldsymbol{\beta}}, \\ \mathrm{V}(y|\boldsymbol{x}) &= \sigma^{*2} e^{2\boldsymbol{x}^T\boldsymbol{\beta}}, \end{aligned} \tag{2.21}$$

을 가정하면 이는

$$\begin{aligned} y|\boldsymbol{x} &\sim \text{Lognormal}\left(\mu, \sigma^2\right), \quad 0 < y < \infty \\ \mu &= \boldsymbol{x}^T\boldsymbol{\beta} - \frac{1}{2}\log\left(\sigma^{*2} + 1\right), \\ \sigma^2 &= \log\left(\sigma^{*2} + 1\right), \end{aligned} \tag{2.22}$$

에 해당되므로 최량상대오차예측량은,

$$\text{BREP} := \frac{\mathrm{E}\left(y^{-1}\middle|\boldsymbol{x}\right)}{\mathrm{E}\left(y^{-2}\middle|\boldsymbol{x}\right)} = \left\{\frac{1}{(\sigma^{*2}+1)^2}\right\} e^{\boldsymbol{x}^T\boldsymbol{\beta}} \tag{2.23}$$

가 된다. 따라서, 패널자료에서 로그정규회귀모형에 임의효과 $\boldsymbol{\gamma}$가 더해질 경우, 조건부 평균과 기대값은

$$\mathrm{E}(y|\boldsymbol{x}, \boldsymbol{\gamma}) = e^{\boldsymbol{x}^T\boldsymbol{\beta}+\boldsymbol{z}^T\boldsymbol{\gamma}}, \tag{2.24}$$

$$\mathrm{V}(y|\boldsymbol{x}, \boldsymbol{\gamma}) = \sigma^{*2} e^{2(\boldsymbol{x}^T\boldsymbol{\beta}+\boldsymbol{z}^T\boldsymbol{\gamma})}, \tag{2.25}$$

가 되므로, 로그정규분포의 패널자료에서 최량상대오차예측량은

$$\text{BREP:} = \frac{\mathrm{E}\left(y^{-1}\middle|\boldsymbol{x}, \boldsymbol{\gamma}\right)}{\mathrm{E}\left(y^{-2}\middle|\boldsymbol{x}, \boldsymbol{\gamma}\right)} = \left\{\frac{1}{(\sigma^{*2}+1)^2}\right\} e^{\boldsymbol{x}^T\boldsymbol{\beta}+\boldsymbol{z}^T\boldsymbol{\gamma}} \tag{2.26}$$

가 되고, $\hat{\boldsymbol{\beta}}$, $\hat{\sigma^2}$와 $\hat{\boldsymbol{\gamma}}$은 식 (2.4)–(2.6)을 통해 구한다.

2.4. 역가우스회귀모형과 상대오차예측량

자동차 보험의 손해액을 추정할 때, 오른쪽으로 많이 기운 정도가 아주 심할 경우, 역가우스분포(inverse gaussian distribution)를 사용하고 있다 (Jong과 Heller, 2008). 역가우스분포는 양수의 값을 가지면서 감마분포나

로그정규분포와 비슷한 모습이지만, 오른쪽으로 더 많이 기운 형태이고 봉우리는 더 뾰족한 것이 특징이다. 역가우스분포의 확률밀도함수는,

$$f_Y(y) = \frac{1}{\sqrt{2\pi y^3\sigma^2}} \exp\left(-\frac{1}{2y}\left(\frac{y-\mu}{\mu\,\sigma}\right)^2\right), \quad y > 0 \tag{2.27}$$

라고 정의된며, $y \sim IG(\mu, \sigma^2)$라고 표기한다. '역가우스'라는 이름은 누적확률밀도함수가 정규부분포의 누적확률밀도함수와 역수(inverse)관계를 갖는데서 비롯되었다. 이때 평균과 분산은

$$\mathrm{E}(y) = \mu, \quad \mathrm{V}(y) = \sigma^2\mu^3, \tag{2.28}$$

가 되며, 역적률값(inverse moments)은

$$\begin{aligned} \mathrm{E}\left(\frac{1}{y}\right) &= \frac{\mathrm{V}(y)+\mu^2}{\mu^3} = \sigma^2 + \frac{1}{\mu}, \\ \mathrm{E}\left(\frac{1}{y^2}\right) &= \frac{\mathrm{E}\left(y^3\right)}{\mu^5} = \frac{\mu^3\left(1+3\mu\sigma^2+3\mu^2\sigma^4\right)}{\mu^5}, = 3\sigma^4 + \frac{3\sigma^2}{\mu} + \frac{1}{\mu^2} \end{aligned} \tag{2.29}$$

가 되어서 (Stoyanov, 1999; Chhikara와 Folks, 1989; Johnson과 Kotz, 1970), 만일 역가우스회귀모형에서 연결함수로 로그함수를 사용하면, $\mu = \exp(\boldsymbol{x}^T\boldsymbol{\beta})$이므로

$$\begin{aligned} \mathrm{E}(y|\boldsymbol{x}) &= e^{\boldsymbol{x}^T\boldsymbol{\beta}}, \\ \mathrm{V}(y|\boldsymbol{x}) &= \sigma^2 e^{3\boldsymbol{x}^T\boldsymbol{\beta}}, \end{aligned} \tag{2.30}$$

가 되어 최량상대오차예측량은

$$\mathrm{BREP}: = \frac{\mathrm{E}\left(y^{-1}\middle|\boldsymbol{x}\right)}{\mathrm{E}\left(y^{-2}\middle|\boldsymbol{x}\right)} = \left\{\frac{1}{3\sigma^2 e^{\boldsymbol{x}^T\boldsymbol{\beta}} + \left(\sigma^2 e^{\boldsymbol{x}^T\boldsymbol{\beta}}+1\right)^{-1}}\right\} e^{\boldsymbol{x}^T\boldsymbol{\beta}}, \tag{2.31}$$

가 된다. 따라서, 패널자료에서 역가우스회귀모형에 임의효과 $\boldsymbol{\gamma}$가 추가되면, 조건부 평균과 기대값은

$$\begin{aligned} \mathrm{E}(y|\boldsymbol{x},\,\boldsymbol{\gamma}) &= e^{\boldsymbol{x}^T\boldsymbol{\beta}+z^T\boldsymbol{\gamma}}, \\ \mathrm{V}(y|\boldsymbol{x},\,\boldsymbol{\gamma}) &= \sigma^2 e^{3\left(\boldsymbol{x}^T\boldsymbol{\beta}+z^T\boldsymbol{\gamma}\right)}, \end{aligned} \tag{2.32}$$

가 되어 역가우스분포의 패널모형에서 최량상대오차예측량은,

$$\mathrm{BREP}: = \frac{\mathrm{E}\left(y^{-1}\middle|\boldsymbol{x},\,\boldsymbol{\gamma}\right)}{\mathrm{E}\left(y^{-2}\middle|\boldsymbol{x},\,\boldsymbol{\gamma}\right)} = \left\{\frac{1}{3\sigma^2 e^{\boldsymbol{x}^T\boldsymbol{\beta}+z^T\boldsymbol{\gamma}} + \left(\sigma^2 e^{\boldsymbol{x}^T\boldsymbol{\beta}+z^T\boldsymbol{\gamma}}+1\right)^{-1}}\right\} e^{\boldsymbol{x}^T\boldsymbol{\beta}+z^T\boldsymbol{\gamma}} \tag{2.33}$$

가 되고, $\hat{\boldsymbol{\beta}}$, $\hat{\sigma^2}$와 $\hat{\boldsymbol{\gamma}}$은 식 (2.4)–(2.6)에서 구하게 된다.

3. 자동차보험 손해액 패널자료

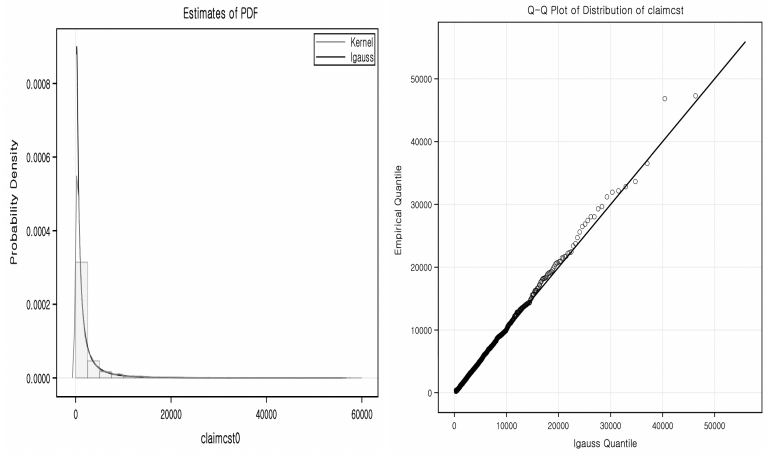
본 연구는 앞서 제안된 일반화선형혼합모형에서 상대오차를 이용한 손해액 예측방법을 2004년–2005년에 청구된 자동차 손해액 자료에 적용해 보았다. 이 자료는 통계패키지 R의 `insuranceData` 라이브러리에서 저장되어 있으며 Table 1은 이 자료에서 사용된 변수들을 나열한 것이다.

Table 1: The list of variables used from `datacar`

변수	내용
veh_value	차량의 가치($10,000)
numclaims	보험청구 횟수
claimcst0	손해액(claim amount 혹은 claim size)
veh_body	차량의 타입(SEDAN, TRUCK, BUS, . . . , 등)
veh_age	차량의 나이(1 [youngest], 2, 3, 4)
gender	male (M), female (F)
area	차량이 등록된 지역(A, B, C, . . . , F)
agecat	보험가입자의 나이 그룹(1 [youngest], 2, 3, 4, 5, 6)

Table 2: The summary of goodness of fit using `proc severity` for the claim size, `claimcst0`, and the column-wise optimals (*)

분포	All fit statistics					
	–2LogLikelihood	AIC	BIC	KS	AD	CvM
감마분포	78,939	78,943	78,956	10.172	384.586	29.096
로그정규분포	77,327	77,331	77,343	6.924	312.701*	10.836
역가우스분포	76,810*	76,814*	76,827*	6.438*	318.102	9.981*



(a) Histogram

(b) Q-Q plot for inverse gaussian distribution

Figure 1: *Goodness of fit for the inverse gaussian distribution on the claim size,* `claimcst0`.

이 분석에서 반응변수는 `claimcst0`이고, `veh_age`는 변수의 속성을 고려해 연속형 변수로 사용했고, `agecat`는 명목형 변수로, `area`는 차량이 등록된 지역의 우편번호를 중심으로 한 그룹변수로서, 본 연구에서는 임의효과(random effect)로 간주하였다. 원래 실제 자료는 67,856개로 구성되었지만, 그 중에 손해액이 0인 것, 중복된 레코드, 그리고 `veh_value`가 0이거나 $80,000을 초과한 것을 제외하고, 총 4,601개의 자료만을 사용하였다.

앞서 언급한대로, 손해액(claim size) 변수 `claimcst0`에 대해 자동차 보험업계에서 보편적으로 사용되

Table 3: The prediction of the random effect, area ($\hat{\gamma}_i$) using proc glimmix

Area	A	B	C	D	E	F
예측값	–0.063	–0.073	0.015	–0.079	0.048	0.135
(표준오차)	(0.061)	(0.063)	(0.056)	(0.076)	(0.070)	(0.095)

는 감마분포, 로그정규분포, 그리고 역가우스분포를 SAS의 proc severity를 이용해 적합해 보았고, 그 결과 Table 2와 Figure 1을 얻게 되었다. 여기서 우리는 이 손해액 자료를 설명하는데, 역가우스분포(inverse gaussian distribution)가 가장 적합하다는 것을 알 수 있었다.

이 자료를 가지고 log(numclaims)를 offset(오프셋) 으로 하는 역가우스회귀모형을 적합하였다. 여기서 offset은 사건이 노출된 시간이나 공간의 크기를 의미하며, 일반적으로 관측값의 관측된 시간/공간의 단위가 다른 경우에 이를 보정하기 위해 사용하고 있다. 이 경우, 보험청구 횟수(numclaims)를 $t_{x,\gamma}$라고 한다면, 사건당 청구된 평균 손해액은

$$\log\left(\frac{\mu_{y|x,\gamma}}{t_{x,\gamma}}\right) = \boldsymbol{x}^T\boldsymbol{\beta} + \boldsymbol{z}^T\boldsymbol{\gamma}, \tag{3.1}$$

가 되므로 log(numclaims)를 오프셋으로 간주하여 proc glimmix를 이용해서 적합해 보았는데, 그 SAS 코드는 아래와 같다.

```
proc glimmix data=a method=quad
    class agecat (ref="5") gender (ref="F") area (ref="A");
         model claimcst0 = veh_age agecat gender
             / dist=invgauss link=log offset=lnumclaim solution;
    random int / subject=area solution;
    output out=outa pred=p;
    lsmeans agecat / plots = diffplot;
run;
```

그 결과, 관측개수가 4,601개에 피어슨 카이제곱 통계량이 7.53으로 역가우스회귀모형은 적합한 것으로 판단되며, 따라서 area가 γ_i인 경우, 손해액에 대한 최량예측량(best predictor, BP)은 식 (2.32)에 의해

$$\begin{aligned}\widehat{\text{Claim Size}}_i &= \exp(\hat{\eta}_i),\\ \hat{\eta}_i &= 7.208 + 0.057 \times \texttt{veh\_age} + 0.400 \times \texttt{I(agecat = 1)} + 0.214 \times \texttt{I(agecat = 2)}\\ &\quad + 0.107 \times \texttt{I(agecat = 3)} + 0.126 \times \texttt{I(agecat = 4)} + 0.060 \times \texttt{I(agecat = 6)}\\ &\quad + 0.147 \times \texttt{I(gender = M)} + \hat{\gamma}_i,\end{aligned} \tag{3.2}$$

로 표현된다 ($\hat{\sigma^2} = 0.00137$). 따라서 자동차 보험의 손해액은 차량이 노후화될수록(veh_age) 증가하며 ($p = 0.012$), 운전자의 나이 수준(agecat)이 '5'일때 가장 낮고, '1(youngest)'일때 가장 높게 예측되고 ($p = 0.003$), 여성보단 남성이 손해액이 높게 예측되고 있다 ($p = 0.003$). 그리고, 각 지역(차량의 등록지, area)의 추정값은 Table 3에서 보여지듯이 A 지역에서 제일 낮고, F 지역에서 높은 예측값을 보여준다.

그러므로, 역가우스회귀모형에서 상대오차(혹은 퍼센트오차)를 최소로 하는 최량상대오차예측량(BREP)은 식 (2.33)에 의해 (i-번째 area에 대해), 식 (3.2)의 $\hat{\eta}_i$를 이용하여

$$\widehat{\text{BREP}}_i := \frac{\text{E}\left(\widehat{y^{-1}\big|\boldsymbol{x},\,\gamma_i}\right)}{\text{E}\left(\widehat{y^{-2}\big|\boldsymbol{x},\,\gamma_i}\right)} = \left\{\frac{1}{3 \times 0.00137\, e^{\hat{\eta}_i} + (0.00137\, e^{\hat{\eta}_i} + 1)^{-1}}\right\} e^{\hat{\eta}_i} \tag{3.3}$$

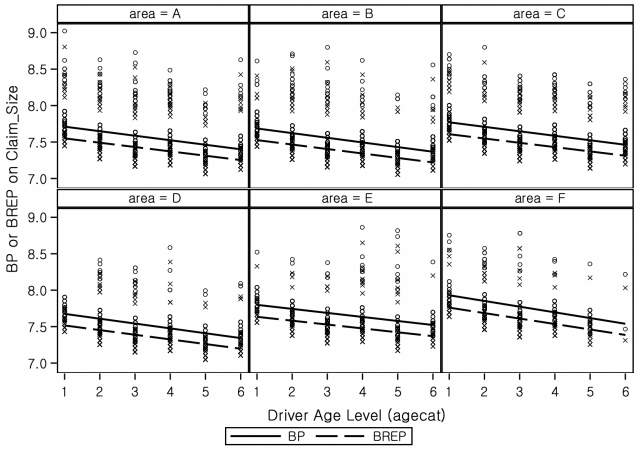


Figure 2: *The predicted claim sizes (*`claimcst0`*) with respect to area (*`area`*) and driver age (*`agecat`*): best predictor (BP: o) and best relative error predictor (BREP: x).*

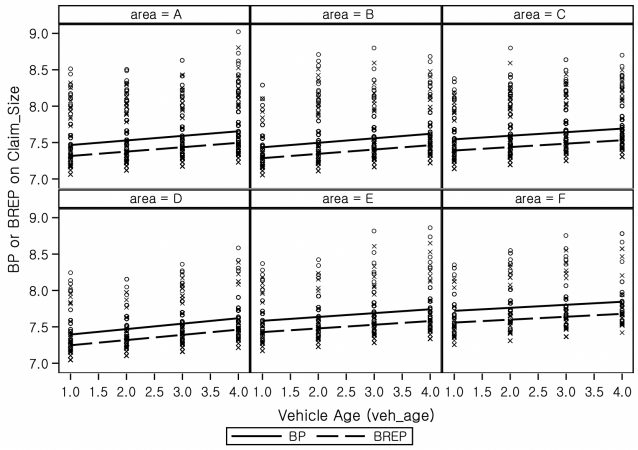


Figure 3: *The predicted claim sizes (*`claimcst0`*) with respect to area (*`area`*) and vehicle age (*`veh_age`*): best predictor (BP: o) and best relative error predictor (BREP: x).*

가 된다. 이를 지역별(`area`) 패널그림을 그려보면 Figures 2–4와 같다.

자동차 보험의 손해액 예측에 있어서, 자료의 분포가 역가우스분포를 따를 경우에, 기존의 MSE를 최소화시키는 최량예측량(BP)과 MSRE를 최소화시키는 최량상대오차예측량(BREP)을 비교하면, BP나 BREP 모두 운전자의 나이가 적을수록, 차량의 연식이 오래될수록, 그리고 여성보단 남성의 손해액이 높게 예측되었다. 또한, BREP는 BP보다 항상 낮게 예측(under-predict)하고 있는데, 이는 y^{-2}와 y가 음의 상관관계($y > 0$인 경우)를 가지므로, 공분산 공식에 의해

$$\mathrm{BREP}_i = \frac{\mathrm{E}\left(y^{-1} \middle| \boldsymbol{x}, \gamma_i\right)}{\mathrm{E}\left(y^{-2} \middle| \boldsymbol{x}, \gamma_i\right)} < \mathrm{E}\left(y | \boldsymbol{x}, \gamma_i\right) = BP_i \tag{3.4}$$

이기 때문이다 (Park과 Stefanski, 1998).

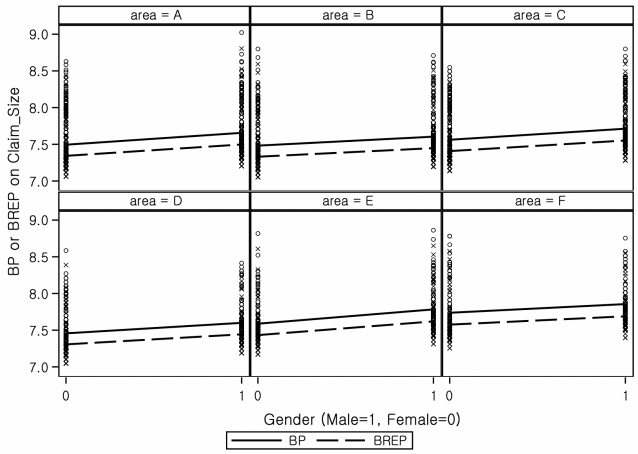


Figure 4: *The predicted claim sizes (*`claimcst0`*) with respect to area (*`area`*) and gender (*`gender`*): best predictor (BP: o) and best relative error predictor (BREP: x).*

4. 결론

상대오차를 최소화시키는 예측법은 계량경제학, 공학 및 생존분석에 이르기까지 연구가 활발해져 왔다. 예측오차의 절대적인 크기보다 상대적 크기에 대한 관심이 많아지고, 예측방법에 대한 다양화가 요구되면서 여러 분야 별로 모수적 방법 혹은 비모수적 방법으로 접근이 이루어져 왔다. 본 논문은 최량상대오차예측량(BREP)을 일반화선형혼합모형에 확장시켜 보았고, 그 예측방법을 자동차 보험분야의 손해액 예측에 적용하였다.

그리고 보험 손해액 예측자료에 상대오차를 사용함에 있어서 Chen 등 (2010, 2016)의 대칭적 상대오차의 개념은 사용되지 않았는데, 이는 주식의 가격결정과 달리 보험 손해액의 예측은 보험가입자 쪽이 아닌 보험회사 쪽에서만 적정 손해율 예측을 위해 이루어지기에 대칭적 상대오차 개념이 필요없기 때문이다.

자동차 보험의 손해율을 추정에 대해서는, 그 추정방벙에 대한 여러 연구가 이루어져 왔지만, 최근들어 각 보험상품에 대한 미래의 손해액 예측은 보험회사의 지불능력과 보험가격의 적절성을 높이기 위한 방법으로 빅데이터와 GPS를 통한 개인 운전기록 등을 통해 보험업계의 예측분석(predictive analytics)의 발전분야로 떠오르고 있다.

본 연구의 결과는 보험회사가 손해액을 예측하면서 상대적 예측오차를 줄이는 방법으로 활용될 수 있으며, 뿐만 아니라 일반화선형혼합모형을 사용하는 여러 분야에서도 상대오차예측량을 사용하는데 참고가 될 수 있을 것으로 사료된다.

References

Bickel PJ and Doksum KA (1977). *Mathematical Statistics*, Holden-Day Inc., Oakland, CA.

Boland PJ (2007). *Statistical and Probabilistic Methods in Actuarial Science*, Boca Raton: Chapman & Hall/CRC.

Boodhun N and Jayabalan M (2018). Risk prediction in life insurance industry using supervised learning algorithm, *Complex & Intelligent Systems*, **4**, 145–154.

Breslow NE and Clayton DG (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association*, **88**, 9–25.

Chen K, Guo S, Lin Y, and Ying Z (2010). Least absolute relative error estimation, *Journal of the American Statistical Association*, **105**, 1104–1112.

Chen K, Lin Y, Wang Z, and Ying Z (2016). Least product relative error estimation, *Journal of Multivariate Analysis*, **144**, 91–98.

Chhikara RS and Folks JL (1989). *The Inverse Gaussian Distribuition*, Marcel Dekker, New York.

Davidian M and Giltinan DM (1995).*Nonlinear Models for Repeated Measurement Data*, Boca Raton: Chapman & Hall/CRC.

Frees E (2018). *Loss Data Analytics*, an open text authored by the Actuarial Community.

Golub GH and Welsch JH (1969). Calculation of gaussian auadrature rules, *Mathematical Computing*, **23**, 221–230.

Hong L and Martin R (2019). *Valid Model-Free Prediction of Future Insurance Claims*, Retrieved October 12, 2019, from SSRN: https://ssrn.com/abstract=3468969 or http://dx.doi.org/10.2139/ssrn.3468969

Huang T, Zhao R, and Tang W (2009). Risk model with fuzzy random individual claim amount, *European Journal of Operational Research*, **192**, 879–890.

Johnson NL and Kotz S (1970). *Continuous Univariate Distributions: Distributions in Statistics*, John Wiley & Sons, New York.

Jones MC, Park H, Shin KI, Vines SK, and Jeong SO (2008). Relative error prediction via kernel regression smoothers, *Journal of Statistical Planning and Inference*, **138**, 2887–2898.

Jong Piet de and Heller GZ (2008). *Generalized Linear Models for Insurance Data*, international series on actuarial science, Cambridge University Press.

Jørgensen B and Souza MCP de (1994). Fitting Tweedie's compound poisson model to insurance claims data, *Scandinavian Actuarial Journal*, **1**, 69–93.

Kahneman D and Tversky A (1979). Prospect theory: an analysis of decision under risk, *Econometrica*, **47**, 263–291.

Kim MJ and Kim YH (2009). Various modeling approaches in auto insurance pricing, *Journal of the Korean Data & Information Science Society*, **20**, 515–526.

Lee Y and Nelder JA (1996). Hierarchical generalized linear models, *Journal of Royal Statistical Society: Series B (Methodological)*, **58**, 619–656.

Lee Y and Nelder JA (2001). Hierarchical generalized linear models: a synthesis of generalized linear models, random-effect models and structured dispersions, *Biometrika*, **88**, 987–1006.

Makridakis SG (1984). *The Forecasting Accuracy of Major Time Series Methods*, Wiley, New York.

Park H and Shin KI (2005). A shrinked forecast in stationary processes favoring percentage error, *Journal of Time Series Analysis*, **27**, 129–139.

Park H and Stefanski LA (1998). Relative-error prediction, *Statistics & Probability Letters*, **40**, 227–236.

SAS/STAT® User's Guide (2012). *The GLIMMIX Procedure*, SAS Institute Inc., Cary, NC, USA.

Smyth GK and Jørgensen B (2002). Fitting tweedie's compound poisson model to insurance claims data: dispersion modeling, *Actuarial Studies in Non-life insurance (ASTIN) bulletin*, **32**, 143–157.

Stoyanov J (1999). Inverse Gaussian distribution and the moment problem, *Journal of Applied Statistical Science*, **9**, 61–71.

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

Wang Z, Chen Z, and Chen Z (2018). H-relative error estimation approach for multiplicative regression model with random effect, *Computational Statistics*, **33**, 623–638.

Werner G and Modlin C (2009). *Basic Ratemaking Workshop*, The Casualty Actuarial Society.

Wolfinger RD and O'Connell M (1993). Generalized linear mixed models: a pseudo-likelihood approach, *Journal of Statistical Computation and Simulation*, **40**, 233–243.

Wüthrich MV and Merz M (2008). *Stochastic Claims Reserving Methods in Insurance*, West Sussex: John Wiley & Sons, England.

Received June 9, 2021; Revised July 1, 2021; Accepted July 2, 2021

상대오차예측을 이용한 자동차 보험의 손해액 예측: 패널자료를 이용한 연구

박흥선[1,a]

[a]한국외국어대학교 통계학과

요 약

상대오차를 이용한 예측법은 상대오차(혹은 퍼센트오차)가 중요시되는 분야, 특히 계량경제학이나 소프트웨어 엔지니어링, 또는 정부기관 공식통계 부분에서 기존 예측방법 외에 선호되는 예측방법이다. 그 동안 상대오차를 이용한 예측법은 선형 혹은 비선형 회귀분석 뿐 아니라, 커널회귀를 이용한 비모수 회귀모형, 그리고 정상시계열분석에 이르기까지 그 범위가 확장되어 왔다. 그러나, 지금까지의 분석은 고정효과(fixed effect)만을 고려한 것이어서 임의효과(random effect)에 관한 상대오차 예측법에 대한 확장이 필요하였다. 본 논문의 목적은 상대오차예측법을 일반화선형혼합모형(GLMM)에 속한 감마회귀(gamma regression), 로그정규회귀(lognormal regression), 그리고 역가우스회귀(inverse gaussian regression)의 패널자료(panel data)에 적용시키는데 있다. 이를 위해 실제 자동차 보험회사의 손해액 자료를 사용하였고, 최량예측량과 최량상대오차예측량을 각각 적용-비교해 보았다.

주요용어: 최량예측량, 최량상대오차예측량, 상대오차, 퍼센트오차, 임의효과, 손해액, 자동차 보험

이 연구는 2020 학년도 한국외국어대학교 교내학술연구비의 지원에 의하여 이루어진 것임.

[1](17035) 경기도 용인시 처인구 모현면 외대로 81, 한국외국어대학교 통계학과. E-mail: hspark@hufs.ac.kr