이학석사 학위논문

Simple Compound Risk Model

with Dependent Structure

의존구조를 가진 단순복합위험모형

2016 년 8 월

서울대학교 대학원

통계학과

정 힘 찬

Simple Compound Risk Model

with Dependent Structure

by

Himchan Jeong

A Dissertation
submitted in fulfillment of the requirement
for the degree of
Master of Science
in
Statistics

The Department of Statistics
College of Natural Sciences
Seoul National University
August, 2016

Abstract

Himchan Jeong
The Department of Statistics
The Graduate School
Seoul National University

There have been fewer trials to address the claim severity in the development of optimal bonus-malus system (BMS), while the claim frequency has been dealt with a lot. In this article, the generalized linear mixed model (GLMM) was incorporated to address the severity, frequency, and their dependency simultaneously with 5 years insurance panel data. Also, estimated individual random effect coefficient from training set and past claim was utilized as a predictor of future loss. From the result of analysis, it was revealed that GLMM had the better fit than its alternatives including simple generalized linear model, dependency between the frequency and severity was significant, and estimated random effect coefficient predicted the future loss better as the length of training set increased. These results provide the rationale to reflect both the past frequency and past severity to construct the optimal BMS, and considering dependence between frequency and severity in the derivation of motor insurance premium.

Keywords : BMS, Random Effects Model, Severity, Prediction, Dependence, Compound Risk Model in Motor Insurance, GLMM.

Students Number : 2014-22361

Contents

List of Tables

List of Figures

1. Introduction

It is very important for auto insurers to examine factors that affect automobile accidents since automobile insurance is prevalent and compulsory in most countries. There are several classification variables to differentiate premiums. A priori rating variables, determined before the policyholder starts to drive, usually include age, gender, the occupation of the main driver, the residence area, and the type and the use of car. For experience rating, observation of past claims behavior of the driver is used as a posteriori rating variable. Bonus-malus system (BMS) penalizes the insureds responsible for one or more accidents by an additional premium or malus, and reward claim-free policyholders by awarding a discount or bonus. All BMS around the world except Korea penalize the number of reported claims, which is accident frequency, without considering claim amounts, accident severity (Lemaire, 1995).

In Korea, the severity of claims has been used for experience rating in addition to the frequency component. Korea plans to remove severity information in its BMS from 2018 following other bonus-malus systems throughout the world. There is empirical research, however, that unobserved heterogeneity of past claim severity can give information to predict future amount of losses, although it is less significant than that of past claim frequency (Pinquet, 1997). Moreover, there has been several theory and empirical research which suggest using both claim frequency and severity is optimal (Frangos & Vrontos, 2001; Mahmoudvand & Hassani, 2009; Mert & Saykan, 2005; Picard, 1976; Tzougas, Vrontos, & Frangos, 2014).

This study examines whether the past claim severity in addition to the claim

frequency can predict future losses. It shows the model designed to take both frequency and severity into consideration predicts future losses better than that designed to use only frequencies. Note that it is important to assess the individual effects for frequency and severity to predict future loss for each insured. Although there are a lot of models which captures individual effects, some relevant research suggests the random effects model as the most efficient one. (Boucher, Denuit, & Guillen, 2008; Pinquet, 1997).

These are hypotheses of this study. First, with respect to the model comparison criteria, random effects model shows the better fit than a model without random effects. Second, past claim frequency can predict the frequency of the future even after controlling a priori variables such as gender, age, territory, vehicle's age, size and model, and the same relation can be observed in average claim severity, respectively. Third, considering both past claim frequency and claim amount predicts claims better than considering either of them. In addition, using past 4 or 3 year data can predict claim better than using only past 2 year data.

We expect this study contributes to the literature as followings. First, this paper is the first paper to examine the power of claim frequency and severity to predict claims by empirical analysis. Even though the majority of BMS designed is based on the number of accidents without considering their severity, there is no empirical evidence showing the comparison of power of these components. Second, this paper uses unique individual data from one of the largest insurance company in Korea, which could show the results much more accurately. Third, by comparing the power of claim frequency and severity, this paper can suggest the direction, for

optimal bonus-malus system. Insurance companies still seek better ways to measure and predict risk. Especially in Korea of which auto insurers suffer losses in auto insurance, this study can help to improve the performance of insurers. For example, estimation period for experience rating and using both frequency and severity as factors could be one of the solutions for the insurers.

The rest of this paper proceeds as follows. In section 2, the related previous literatures are introduced. Then the sample data, variables, summary statistics and methodology are presented in section 3. Section 4 reveals that random effects model is more efficient than the model without that in the perspective of model comparison criteria. In section 5, I investigate the power of accident frequency and accident severity to predict auto accidents of the future. Section 6 tests whether considering both frequency and severity increases the prediction power. Section 7 concludes this paper and suggests future works.

2. Literature review

2.1. The Rationale of Bonus-Malus System in Auto Insurance

BMS was first introduced by the British in the 1910s. At that time, it was called NCD - no-claim discount; it literally meant "don't bother us at all, and you get the discount. You write to us for any reason, you don't get the discount". So any claim reported to the company, at-fault or not at-fault was penalized. But quickly, insurers moved away from that, and, when continental European insurers introduced BMS in the 1960s, only at-fault accidents were penalized.

The purpose of BMS is two folds. First is to reduce the information

asymmetry between policyholder and insurance company. By doing so, insurance company may charge more appropriate and fair rate for each driver's risk class. For instance, high-mileage users: if people do not have to report their mileage, insurers cannot charge mileage a priori, and so they compensate a little bit by charging a posteriori, because high-mileage users will have more accidents. Same with couples not telling that their kid is driving, so insurers cannot price the risk a priori accordingly, but they compensate somewhat with BMS, as the kid is going to be involved in more accidents. So the logic is that those who drive more or having a risky driving habit probably had more accident in the past, and they will have more accidents in the future. Therefore, the past accident history may have prediction power as long as the unobserved or omitted factors such as mileage driven is about the same in the past and near future.

The fact that all auto insurance BMS in force uses only the claim frequency implies that the future claim prediction mostly applies only to the claim frequency. The severity of claim, on the other hand, does not have much prediction power. For example, the one who had broken two legs will be penalized more than the driver who broke one leg from the past accident under the current Korean system. This can be justified only if the one who broke two legs is more likely to cause an accident with higher severity than the one who broke one leg. In other words, claim severity should also repeat. There may exists a little correlation between claim sizes – the driver who had severe accidents tend to have severe accidents in the future but the severity of accidents is mostly random outcome. We do not know who we end up hitting, which kind of car we hit.

The second rationale for auto insurance BMS is to control moral hazard. With an experience rating scheme like auto insurance BMS, policyholders who are aware of the fact that their premium will increase depending on the claim history will have incentive to reduce claim, which will reverse the moral hazard issue. In this regard, if policyholders are able to control the size of claim, using severity information in BMS has abundant justification. However, it is very hard to think that drivers can actually control the size of accidents, especially the degree of bodily injury in liability claims.

Note that those rationales for BMS can be supported in statistical manner. If we can detect heterogeneity of claim for each insured from observed data, then it will be valid to apply BMS for each insured. There have been some trials to detect heterogeneity in given data. Pinquet (1997) argues the sufficient condition for the existence of BMS based on a Bayesian model, and Jacqmin-Gadda and Commenges (1995) suggests two statistics to test the overdispersion and correlation between subject separately, which can be applied in a variety of distributions.

2.2. Designing optimal BMS with past frequency and severity

Dionne and Vanasse (1989) propose a bonus-malus system which integrates a priori and a posteriori information on an individual basis using data of 19,013 individuals in Quebec. In addition, Lemaire (1995) argues that the best predictor of the future number of claims is not the driver's age, sex or occupation but his past claims behavior and suggests the design of an optimal BMS based on the claim frequency. As an extension of these studies, using different distributions, there are

several studies considering claim frequency as a factor of auto accidents. (Coene & Doray, 1996; Tremblay, 1992; Walhin & Paris, 1999)

Pinquet (1997) designs an optimal BMS including the severity of the claims. From a rating model based on the analysis of number of claims and of costs of claims, two heterogeneity components are added. Frangos and Vrontos (2001) design an optimal BMS based only on a posteriori classification criteria and then generalize it in order to take under consideration both the a priori and the a posteriori classification criteria. In addition, Mahmoudvand and Hassani (2009) extend BMS model of Frangos and Vrontos (2001) which consider both a priori variable such as the age, the sex, the place of residence of the policyholder, type of the car, capacity of the car and a posteriori such as frequency and severity. Moreover, Mert and Saykan (2005), by designing an optimal BMS based only on the claim frequency and one based on both the claim frequency and the claim amount, suggest that it is fairer to charge policyholders premiums which consider both claim frequency and claim amount.

2.3. Individual effects and dependency between frequency and severity

To incorporate the frequency and severity of each insured simultaneously, there are two issues which have to deal with. First, once we can validate the existence of unobserved heterogeneity among each insured in a pricing system, we need to assess and reflect that numerically in our pricing system. Thus a variety of model has been suggested for that task. MOLENBERGHS (2005) suggests three categories of models which can be used at insurance panel data, conditional models, marginal models, and random effects model. And (Boucher et al. (2008)) expand the categories with Integer-valued autoregression models, copula models, and common shock

models. In their study, they conduct model comparison among these and conclude that the random effects model is the most useful model for insurance claim data with time dependence.

Second, we need to consider the dependency between the frequency and severity. Let we denote the claim count of i^{th} insured be n_i, while the j^{th} claim amount of i^{th} insured be y_{ij}. Historically, n_i and y_{ij} are assumed to be independent. For example, Jørgensen and Paes De Souza (1994) assume a Poisson distribution for n_i and a Tweedie distribution for y_{ij} and they are assumed to be independent. In particular, the independence assumption is convenient in various statistical computations such as maximum likelihood estimations (Jørgensen & Paes De Souza, 1994; Klugman, Panjer, & Willmot, 2012; Smyth & Jørgensen, 2002). Note that although the independence of the frequency and the individual severity is assumed, the dependence between n_i and m_i, the average severity of i^{th} insured can arise automatically in the model of Jørgensen and Paes De Souza (1994). Recently, evidences showing the necessity to model the dependence of m_i on n_i explicitly can be found in insurance literatures such as Boudreault, Cossette, Landriault, and Marceau (2006), Hernández-Bastida, Fernández-Sánchez, and Gómez-Déniz (2009), Czado, Kastenmeier, Brechmann, and Min (2012), and Shi, Feng, and Ivantsova (2015) introduce the dependence between n_i and m_i through a copula framework. However, the choice of right copula family is a difficult problem in practice and estimation for copula parameters is also not an easy task. Alternatively, Gschlößl and Czado (2007) and Shi et al. (2015) consider a conditional probability approach, which denotes using n_i as a covariate of the conditional mean model for m_i. It is the simplest, but flexible way of

considering the dependence between n_i and m_i explicitly. Shi et al. (2015) assume a zero-truncated negative binomial distribution for n_i and a generalized gamma distribution for m_i, using cross-sectional data. They considered regression models for location parameter only while the dispersion parameter remains constant.

3. Data and Methodology

3.1. Data Description

In this study, we used individual insured data who buy auto insurance from one of the largest insurance company in Korea. This insurance company's market share was about 30% in 2014. The sample period is from 2009 to 2013 and our sample includes most of rating variables age, gender, town, model of car, size of car, and complete claim information, and insurance premium charged.

The most important coverages in Korean motor insurance are bodily injury (BI), property damage (PD), and comprehensive and collision (CNC) coverages. BI and PD are liability coverages, in other words, they insure the risk of compensation to a third party, whereas CNC pays the repairing cost of the insured's own car due to the damage from collision and diverse reasons, such as theft, vandalism, and weather. In this research, we analyze the repetition of claim frequency and severity of liability coverages only as the BMS is mostly used for liability claims.

For this study, the insured do not have full continuous 5 year data in this company are excluded since I focus on the prediction power of the consecutive year. The original sample has 8,347,848 observations but after cleaning the total number of observations in our final 5 year annual panel data is 1,763,449. Table 1 shows the

claim distribution of our selected sample. The accident claim rate for selected sample is 12.8% and the rate for original sample is 13.6%, which are similar each other.

3.2. Proposed Model

Since it is not appropriate to assume that the frequency and severity of claim follows normal distribution, we may not apply simple linear model for parameter estimation and model fitting. Moreover, the model should reflect the heterogeneity of each insured. To handle these issues, we use generalized linear mixed model (GLMM) in this paper, which allows us to use non-normal distribution in model fitting as well as the variation between each insured. More specifically, we use the 'random-intercept mean model', which allows the intercept of mean can vary for each insured and the variation follows mean 0 normal distribution. If we denote the mean of i^{th} insured as μ_i, random-intercept mean model is given as following;

$$\log(\mu_i) = Z_i^{\mathrm{T}}\boldsymbol{\gamma} + E_i, \quad E_i \sim \mathrm{N}(0, \sigma^2), \quad \text{i.i.d.} \qquad (1)$$

Here $\boldsymbol{\gamma}$ is column fixed effect parameters, and $\{Z_i\}$ is covariate vector of i^{th} insured. Note that if we set $\sigma^2 = 0$, then it is just a generalized linear model (GLM).

For covariates, Let $\{\mathbf{X}_i\}$ be p -dimensional column vectors. Throughout the paper, we assume $\{\mathbf{X}_i\}$ are given, and all the statistical procedures are conditioned on $\{\mathbf{X}_i\}$. Also, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are p -dimensional column vectors of parameters, respectively. For convenience, we may define following set of index when $n_1, \cdots, n_k$

are given observations.

$$\mathcal{I}_k := \left\{ i \in \{1, \cdots, k\} \,\middle|\, n_i \neq 0 \right\}.$$

For modeling frequency, first we can regard Poisson distribution. Poisson distribution is generally used to model count data in various field. However, this distribution cannot capture the overdispersion of given data, which usually occurs in insurance claim count data. To handle this problem, some relevant studies such as Lemaire (1995) and Boucher et al. (2008) suggest negative binomial distribution as an alternative for Poisson. Thus, we use both model in frequency fitting and compare the efficiency of given distributions with respect to the model selection criteria, such as AIC or BIC. Thus, GLMM formula for mean of frequency is given as following;

$$\log(\theta_i) = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha} + R_i, \quad R_i \sim \mathrm{N}(0, \sigma_1^{\,2}), \quad \text{i.i.d.,} \quad \text{where} \quad E[N_i \mid \mathbf{X}_i, R_i] = \theta_i \qquad (2)$$

For modeling severity, we assume the gamma distribution, because we can get ease of analytic investigations as well as there are a lot of previous studies which assume this distribution for severity such as Pinquet (1997) , Smyth and Jørgensen (2002), and Shi et al. (2015). Note that we use different parameter for gamma distribution compared to standard one. We parameterize this distribution with ν and ξ, where mean is ξ and $1/\nu$ is called as a dispersion parameter. In this case, the density function which is denoted by $Y \sim Gamma(\xi, \nu)$ is given as following;

$$f_Y(y) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\xi} \right)^{\nu} y^{\nu-1} \exp\left(-\frac{y\nu}{\xi} \right)$$

Note that when $y_{i1}, \cdots, y_{in_i}$ are i.i.d. $Gamma(\xi_i, \nu)$, for given $n_i > 0$, their average

m_i follows $Gamma(\xi_i, n_i v)$. Hence, GLMM formula for mean of average severity is given as following;

$$\log(\xi_i) = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + U_i, \quad U_i \sim \mathrm{N}(0, \sigma_2{}^2), \quad \text{i.i.d., where} \quad E[M_i \mid \mathbf{X}_i, U_i, N_i] = \xi_i \qquad (3)$$

4. Model Comparison and Empirical Analysis

4.1. Frequency

As it is mentioned above, the existence of overdispersion and individual unobserved effects in given data has critical role in model assumption. Thus, these issues should be tested prior to empirical analysis. To examine whether the Poisson model is enough to capture the characteristic of given frequency data, we conducted Pearson Chi-square test. The results of test are summarized in following table.

# of accidents	Observed	Poisson Fitted	NB2 Fitted
0	1,538,416	1,534,024	1,538,786
1	206,248	213,906	204,913
2	17,033	14,989	18,164
3	1,518	705	1,411
4	189	0	176
5+	45	0	0
χ^2 statistics		5,543	468

Table 1 : Goodness-of-fit test of frequency models.

Note that as the value of χ^2 statistics is bigger, assumed model is the more inappropriate, and we can see that negative binomial distribution gives better fit then Poisson. The existence of individual unobserved heterogeneity can also be tested. If

we recall the GLMM framework for frequency, to test $H_0 : \sigma_1 = 0$ vs $H_1 : \sigma_1 \neq 0$ is equivalent to the test of unobserved heterogeneity. Hence, we conducted likelihood ratio test (LRT) for given H_0. With the result of below table, it is possible to reject H_0, in other words, we can assume the existence of individual unobserved heterogeneity in claim frequency.

Case	**2*Log-Likelihood**	**df**	**p-value**
Under Ω	-1,479,588	40	
Under H_0.	-1,480,405	39	
Difference	817	1	> **0.0001**

Table 2: LRT for the variance of random effects in frequency

After testing the overdispersion and unobserved heterogeneity of claim frequency, we fitted the data with three model; Poisson, NB without random effects, and NB with random effects. We used BIC as a model comparison criterion. Note that BIC equals to k*log(n) – 2*loglikelihood), where k is the number of parameter in given model, and n is the sample size. Thus, minimizing BIC is equivalent to maximize loglikelihood under the penalty of adding incremental parameter. Because it penalizes the more as the sample size grows up when we add parameter in a model, this criterion is appropriate in this study, of which sample size increases a lot as the period is lengthened. With the result of below table and figure, we may observe that fit of NB with random effects model is the best, and the improvement of BIC increases as we use the longer periods. The table for regression result of each frequency model using 2009-2013 years data is given in Appendix.

Years	Poisson	NB without random effects	NB with random effects
2009-2010	702,310	701,859	701,717
2009-2011	933,274	932,656	932,316
2009-2012	1,277,037	1,275,945	1,275,355
2009-2013	1,482,391	1,481,044	1,480,244
# of Parameters	38	39	40

Table 3: BIC values of each frequency model for various term

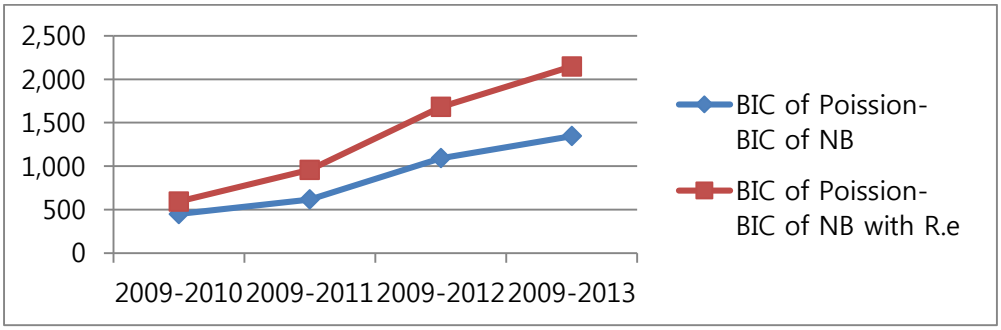

Figure 1: BIC differences among given frequency models according to the periods

With the regression result, we could set three risk profiles for calculation of a priori premium for claim frequency. A priori premium means the premium which is estimated with only observed characteristic of each contract and insured. The risk profiles and estimated a priori premium for each profile is presented in below.

Risk type	Sex	newcomer	Age	Car Size	Car Age	Driver Limit
Low	M	N	50s	Small	11~	One
Medium	M	N	40s	Medium	4~7	Couple
High	F	Y	30s	Medium	1~3	Couple

Table 4: Risk Profiles for claim frequency

Risk type	Poisson	NB
Low	0.10565	0.10548
Medium	0.13369	0.13360
High	0.16439	0.16441

Table 5: A priori premium for claim frequency

4.2. Severity

Since we assumed Gamma distribution, it automatically captures the overdispersion of given claim severity data. Thus we only need to test the existence of individual unobserved heterogeneity. Similar to heterogeneity test of frequency, it suffices to test $H_0 : \sigma_2 = 0$ vs $H_1 : \sigma_2 \neq 0$ using LRT. Again, with the result of below table, it is possible to reject H_0, in other words, we can assume the existence of individual unobserved heterogeneity in claim severity.

Case	**2*Log-Likelihood**	**Df**	**p-value**
Under Ω	-6,836,902	41	
Under H_0.	-6,837,474	40	
Difference	101,149	1	> **0.0001**

Table 6: LRT for the variance of random effects in severity

After testing the unobserved heterogeneity of claim severity, we fitted the data with two model; gamma without random effects, and gamma with random effects. We also used BIC as a model comparison criterion. With the result of below table and figure, we may observe that fit of the model with random effects is the better, and the improvement of BIC increases as we use the longer periods, respectively. The table for regression result of each severity model using 2009-2013 years data is given in Appendix.

Years	**Gamma without random effects**	**Gamma with random effects**
2009-2010	3,274,402	3,210,364
2009-2011	4,358,162	4,281,580
2009-2012	5,975,254	5,884,799
2009-2013	6,939,196	6,838,061
# of Parameters	40	41

Table 7: BIC values of each severity model for various term

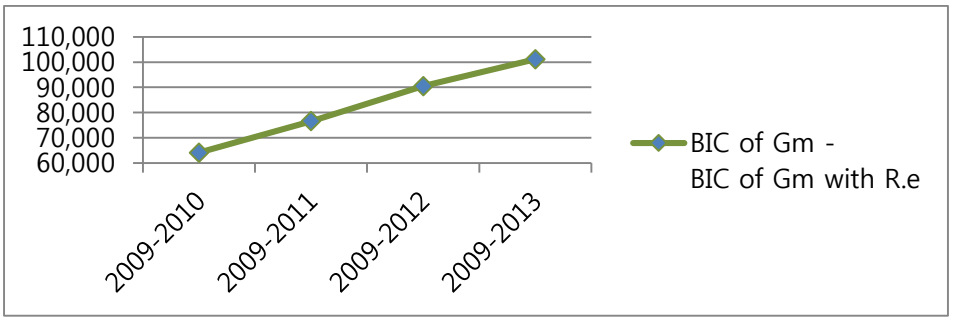

Figure 2: BIC difference between given severity models according to the periods

In case of average severity, we also could set three risk profiles for calculation of a priori premium. Note that the profiles differ from those of frequency. The risk profiles and estimated a priori premium for each profile is presented in below.

Risk type	Sex	New comer	Age	Car Size	Car Age	Driver Limit	Age Limit
Low	F	N	30s	Small	4~7	One	Y
Medium	M	Y	30s	Small	1~3	Couple	Y
High	M	N	50s	Medium	4~7	Family	N

Table 8: Risk Profiles for claim severity

Risk type	Gamma
Low	1,433,002
Medium	1,849,008
High	1,972,010

Table 9: A priori premium for average claim severity

5. Prediction

As it is mentioned, it is crucial to obtain optimal BMS for the whole motor insurance industry of a country. Random effects model has another benefit in addition to better fitting, which is providing the estimated random effect coefficient of each insured. Hence it is of interest whether we can use these coefficients as predictors of future claim. To examine this, we separated our sample into two categories, the

sample from 2011-2012 / 2010-2012 / 2009-2012 as the training sets, and the sample from 2013 as the test set. Our strategy consists of following two steps. First step is to estimate random effect coefficients E_i using the training sets under the GLMM, where $\log(\mu_i) = X_i^{\mathrm{T}}\boldsymbol{\gamma} + E_i$. And then, we fit the test set with ordinary GLM using E_i as an explanatory variable, so that $\log(\mu_i^*) = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma} + \eta E_i$. Thus, to test $H_0 : \eta = 0$ vs $H_1 : \eta \neq 0$ is equivalent to the test of the prediction power of random effect estimates.

5.1. Frequency

We can observe that the prediction power of random effect estimates from past frequency data is strongly significant, and the prediction power increases as the training period is lengthened. Note that individual random effect estimates are the residual portion of former claim, which is controlled with observable covariates. As Lemaire (1995) points out, since the purpose of BMS is to detect unobserved factor of each individual, it is reasonable to use such residual portion as a predictor for the pricing of next year, rather than merely former claim data. To validate this argument, we conducted another prediction analysis using mere past claim data, and in this case the GLMM is given as $\log(\mu_i^*) = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma} + \tau \log(n_i)$, where n_i is past claim frequency of given insured. In that case, the prediction power of mere past claim frequency is still significant, but is lower than that of random effect estimates.

Training Set	Coefficient	BIC	p-value
2009-2012	1.1341	205,242	**>0.0001**
2010-2012	1.1286	205,361	**>0.0001**
2011-2012	1.3767	205,322	**>0.0001**

Table 10: BIC values and estimated η using each training set (for frequency)

Training Set	Coefficient	BIC	p-value
2009-2012	0.1316	205,309	**>0.0001**
2010-2012	0.1436	205,355	**>0.0001**
2011-2012	0.1740	205,358	**>0.0001**

Table 11: BIC values and estimate τ using each training set (for frequency)

The trend of increasing prediction power along with the lengthened training periods can be tested in other way. Let we denote n_i be the predicted frequency from the model which is fitted with given training sets, then we may define mean absolute deviation (MAD) of claim frequency as $\sum | n_i - n_i | / N$, as a measure of prediction error. We derived this value from the training sets in below table, and it can be revealed that the longest training set gives the least value for prediction error.

Training Set	Sample Size	MAD
2009-2012	240,577	0.24148
2010-2012	240,577	0.24390
2011-2012	240,577	0.24455

Table 12: MAD using each training set (for frequency)

Moreover, GLMM can be applied to calculate a posteriori premium, which utilizes the past claim, as well as observed characteristic of each contract and insured.

In fact, since Poisson and gamma distributions are not conjugate with the distribution of random effects which is Normal, we cannot obtain a closed form of a posteriori premium, which is determined by the number of past claims and current observed characteristic. However, it is possible to get a conditional expectation of claim frequency, given the observed characteristic and random effects estimates. After that, we summarized those conditional expectations as mean values, with the number of past claims. The result is presented below.

Past Claim Freq	A Priori	0	1	2	3	4	5+
Low	0.1054	0.0986	0.1038	0.1091	0.1141	0.1179	0.1318
Medium	0.1336	0.1256	0.1319	0.1395	0.1449	0.1525	0.1662
High	0.1644	0.1514	0.1636	0.1720	0.1754	0.1865	0.1955

Table 13: Predictive premiums for claim frequency

Note that it is possible to compare the prediction power of predictive premiums from a priori model and a posteriori model. We summarized the predictive premiums of each contract with respect to the realized claim frequency in 2013. With the result, we can see that the predictive premiums from a posteriori model distinguish the potential risk better than a priori model.

2013 Claim Frequency	Premium Amounts		Premium Ratio	
	A Posteriori	A Priori	A Posteriori	A Priori
0	182,133	182,064	0.9944	0.9948
1	189,336	188,768	1.0337	1.0314
2	196,619	195,207	1.0735	1.0666
3	204,047	201,798	1.1140	1.1026
4+	205,116	201,942	1.1199	1.1034
Total	**183,159**	**183,014**	**1.0000**	**1.0000**

Table 14: Predictive premium amounts and ratio with the claim in 2013 frequency

5.2. Severity

For severity, there is an issue of missing data. If an insured maintains contract for consecutive 5 years, then we are able to observe the claim frequency for 5 years, which includes 0. However, if the claim frequency is equal to 0, then it is impossible to observe the average claim frequency. So we used only complete data for each term of training set. We can observe that the prediction power of random effect estimates from past severity data is lower than the frequency prediction and insignificant. And the table shows us that prediction power rapidly decreases as the training period is very short, with respect to the comparison of p-values. (Note that we cannot compare BIC of each model, because each term has different sample size.) In addition, we conducted another prediction analysis which is similar to frequency case, using mere past claim severity, and in this case the GLMM is given as $\log(\mu_i^*) = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\gamma} + \tau \log(m_i)$, where m_i is past claim frequency of given insured. In this case, we can see that the prediction power of mere past claim average severity data is less significant than that of random effect estimates.

Training Set	Coefficient	Sample Size	BIC	p-value
2009-2012	0.0769	21,615	668,339	0.1268
2010-2012	0.0711	18,520	572,795	0.1523
2011-2012	0.0249	13,577	420,045	0.6154

Table 15: BIC values and estimated η using each training set (for average severity)

Training Set	Coefficient	Sample Size	BIC	p-value
2009-2012	0.037106	21,615	668,347	0.1951
2010-2012	0.036568	18,520	572,802	0.2242
2011-2012	0.012809	13,577	420,046	0.7021

Table 16: BIC values and estimated τ using each training set (for average severity)

In case of average severity, we also can derive a posteriori premium for each risk profile, with the same logic which is used in claim frequency. The result is presented in below.

Past Claim Avg. Severity	A Priori	0~900,000	900,000~ 2,000,000	2,000,000~ 5,000,000
Low	1,433,002	570,095	965,234	1,576,614
Medium	1,849,008	681,446	1,319,219	2,197,386
High	1,972,010	1,107,249	1,618,586	3,172,474

Table 17: Predictive premiums for average claim severity

Although the prediction power of estimated random effects of claim average severity is less significant than that of claim frequency, we can still observe the trend of increasing prediction power along with the lengthened training periods. Let we denote m_i be the predicted average severity from the model which is fitted with given training sets, then we may define MAD of average severity, respectively. The result is presented below.

Training Set	Sample Size	MAD
2009-2012	21,615	1,826,411
2010-2012	18,520	1,883,300
2011-2012	13,577	1,978,701

Table 18: MAD using each training set (for average severity)

The predictive premiums of each contract with respect to the realized claim severity in 2013 can be analyzed, and the similar result can be observed with the case of claim frequency, respectively.

2013 Claim Severity	Premium Amounts		Premium Ratio	
	A Posteriori	A Priori	A Posteriori	A Priori
0~500,000	178,028	180,354	0.9720	0.9855
500,000~1,000,000	182,122	182,930	0.9943	0.9995
1,000,000~3,000,000	187,049	184,451	1.0212	1.0079
3,000,000~	189,818	186,472	1.0364	1.0189
Total	**183,159**	**183,014**	**1.0000**	**1.0000**

Table 19: Predictive premium amounts and ratio with the claim in 2013 severity

6. Dependency between the Frequency and Severity

Under GLMM framework, we may also test whether we may assume the independence between the frequency and severity. If we set a GLMM such as $\log(\xi_i) = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} + U_i + \nu n_i$, then to test $H_0 : \nu = 0$ vs $H_1 : \nu \neq 0$ is equivalent to the test of independence. Hence, we conducted likelihood ratio test (LRT) for given H_0. With the result of below table, it is possible to reject H_0, in other words, we can assume the existence of individual heterogeneity in claim frequency. With the result of below table and figure, we may observe the effect of observed frequency for average severity is significantly positive and the improvement of BIC by assuming the dependence increases as we use the longer periods, respectively.

Years	Coefficient	p-value
2009-2010	0.1573	> **0.0001**
2009-2011	0.1573	> **0.0001**
2009-2012	0.1390	> **0.0001**
2009-2013	0.1368	> **0.0001**

Table 20: Estimated ν and p-values for various term

Years	Under Ω	Under H_0.	Difference	**p-value**
2009-2010	-3,209,807	-3,209,415	392	> **0.0001**
2009-2011	-4,281,012	-4,280,540	472	> **0.0001**
2009-2012	-5,884,218	-5,883,709	509	> **0.0001**
2009-2013	-6,837,474	-6,836,902	572	> **0.0001**
df	42	41		

Table 21: 2*Log-Likelihood values of each severity model with/without frequency

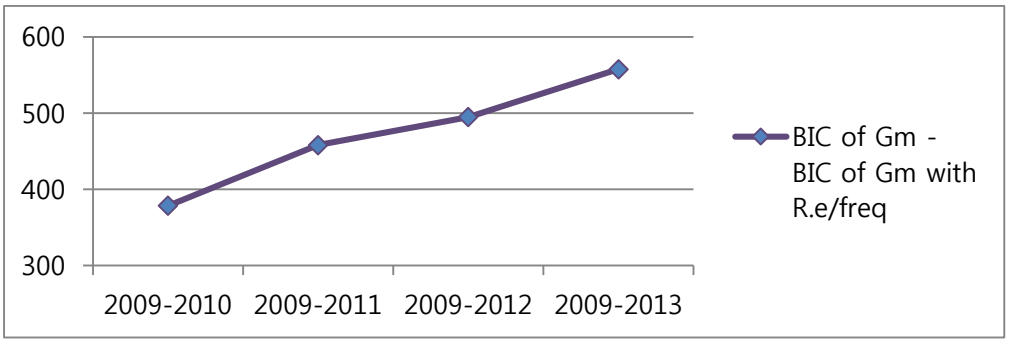

Figure 3: BIC difference between severity models with/without frequency according to the periods

7. Conclusion

This article tried to figure out the correlation matter in the context of auto insurance, such as correlation among the frequencies of each insured, correlation among the severities of each insured, and correlation between the frequencies and severities. First, we could suggest some model which can improve the fitting and prediction power for auto insurance claim. With regard to claim frequency, negative binomial with random effects model showed the best fit, which incorporates the overdispersion and unobserved heterogeneity among each insured simultaneously. And gamma with random effects model also showed better fit than gamma without random effects model.

And then, it could be shown that random effect estimates from past frequency data were very significant for the prediction of future claim frequency, more than mere past claim frequency. It was noted that prediction power of random effects estimates increased as the training period getting longer. Prediction of severity showed similar trends whereas random effects estimates were less significant for the prediction of future claim average severity, and prediction power has positive relationship with the length of training period, respectively.

Finally, this article could show that independence between frequency and severity is not ignorable. We could observe the positive and significant correlation between frequency and average severity. This can be a substantial contribution since we used longitudinal data which is more realistic and informative than cross-sectional to figure out this relationship, whereas previous studies focused on analyzing cross-sectional data.

Appendix

Variables	Definition
Male	A dummy variable that equals 1 when the owner of the car is male, otherwise it equals 0.
Domestic	A dummy variable that equals 1 when the car is domestic car otherwise it equals 0.
Renewal	A dummy variable that equals 1 when the insured was a customer of this company in last year, otherwise it equals 0.
Newcomer1	A dummy variable that equals 1 when the insured had driven less than 1 year, otherwise it equals 0.
Newcomer2	A dummy variable that equals 1 when the insured had driven between 2 and 3 years, otherwise it equals 0.
Sports	A dummy variable that equals 1 when the car is a sports car, otherwise it equals 0.
Yage_limit	A dummy variable that equals 1 when only the drivers older than 26 are insured, otherwise it equals 0.
Age20	A dummy variable that equals 1 when the insured is between the ages of 20 and 29, otherwise it equals 0.
Age30	A dummy variable that equals 1 when the insured is between the ages of 30 and 39, otherwise it equals 0.
Age40	A dummy variable that equals 1 when the insured is between the ages of 40 and 49, otherwise it equals 0.
Age50	A dummy variable that equals 1 when the insured is between the ages of 50 and 59, otherwise it equals 0.
Cartype1	A dummy variable that equals 1 when the insured car equals or is under 1600 c.c. except multi-purpose car, otherwise it equals 0.
Cartype2	A dummy variable that equals 1 when the insured is between 1600 c.c. and 2000 c.c. except multi-purpose car, otherwise it equals 0.
Cartype3	A dummy variable that equals 1 when the insured is over 2000 c.c. except multi-purpose car, otherwise it equals 0.
Carage1	A dummy variable that equals 1 when the car is under 3 years old, otherwise it equals 0.
Carage2	A dummy variable that equals 1 when the car is 4~7 years old, otherwise it equals 0.
Carage3	A dummy variable that equals 1 when the car is 7~10 years old, otherwise it equals 0.
Seoul	A dummy variable that equals 1 when the car is registered in Seoul, otherwise it equals 0.
Gyeongn	A dummy variable that equals 1 when the car is registered in

	Gyeongnam, otherwise it equals 0.
Gyeongg	A dummy variable that equals 1 when the car is registered in Gyeonggi, otherwise it equals 0.
Daej	A dummy variable that equals 1 when the car is registered in Daejeon, otherwise it equals 0.
Chungn	A dummy variable that equals 1 when the car is registered in Chungnam, otherwise it equals 0.
Jeonn	A dummy variable that equals 1 when the car is registered in Jeonnam, otherwise it equals 0.
Gyeongb	A dummy variable that equals 1 when the car is registered in Gyeongbuk, otherwise it equals 0.
Gangw	A dummy variable that equals 1 when the car is registered in Gangwon, otherwise it equals 0.
Jeju	A dummy variable that equals 1 when the car is registered in Jeju, otherwise it equals 0.
Jeonb	A dummy variable that equals 1 when the car is registered in Jeonbuk, otherwise it equals 0.
Busan	A dummy variable that equals 1 when the car is registered in Busan, otherwise it equals 0.
Chungb	A dummy variable that equals 1 when the car is registered in Chungbuk, otherwise it equals 0.
Ulsan	A dummy variable that equals 1 when the car is registered in Ulsan, otherwise it equals 0.
Gwangj	A dummy variable that equals 1 when the car is registered in Gwangju, otherwise it equals 0.
Incheon	A dummy variable that equals 1 when the car is registered in Incheon, otherwise it equals 0.
Daegu	A dummy variable that equals 1 when the car is registered in Daegu, otherwise it equals 0.
Limit_couple	A dummy variable that equals 1 when the insureds are specified to couple, otherwise it equals 0.
Limit_one	A dummy variable that equals 1 when the insured is specified to a person, otherwise it equals 0.
Limit_two	A dummy variable that equals 1 when the insureds are specified to two people, otherwise it equals 0.
Limit_family	A dummy variable that equals 1 when the insureds are specified to a family, otherwise it equals 0.

Table 22: Definitions of the Variables

Categories	Proportion	Accident Rate	Average Size(mil₩)
Age20	3.5%	15.3%	1.9
Age30	23.7%	12.1%	1.8
Age40	35.5%	12.6%	1.8
Age50	28.2%	13.2%	1.9
Others (Older than 60)	9.2%	12.6%	1.9
Cartype1	33.6%	12.2%	1.7
Cartype2	30.4%	13.0%	1.8
Cartype3	15.6%	12.6%	2.0
Others (Multi-Purpose)	20.5%	13.4%	1.9
Carage1	31.5%	12.9%	1.9
Carage2	29.3%	13.0%	1.9
Carage3	19.9%	13.3%	1.9
Others (Older than 11)	19.3%	11.6%	1.7
Domestic	68.5%	12.7%	2.6
Foreign	31.5%	12.9%	0.3
Normal	95.7%	12.8%	1.9
Sports	4.3%	11.2%	0.3
Yage_limit	95.3%	12.5%	1.8
Others (No Yage_limit)	20.5%	19.0%	2.0
Male	75.9%	12.3%	1.9
Female	24.1%	14.3%	1.8
Limit_one	32.6%	11.7%	1.9
Limit_two	2.9%	14.0%	1.9
Limit_family	18.6%	14.5%	2.0
Limit_couple	39.0%	12.4%	1.7
Others	6.9%	14.4%	2.3
Total	100.0%	12.8%	1.9

Table 23: summary for selected sample

Categories	Proportion	Accident Rate	Average Size(mil₩)
Age20	7.5%	18.0%	2.2
Age30	26.6%	12.9%	1.8
Age40	30.8%	13.0%	1.9
Age50	23.8%	13.7%	1.9
Others (Older than 60)	11.3%	13.4%	1.9
Cartype1	34.7%	13.3%	1.8
Cartype2	30.9%	13.8%	1.9
Cartype3	14.8%	13.1%	2.0
Others (Multi-Purpose)	19.6%	14.1%	2.0
Carage1	34.7%	13.9%	2.0
Carage2	27.8%	13.8%	1.9
Carage3	19.1%	14.0%	1.9
Others (Older than 11)	18.4%	12.1%	1.8
Domestic	95.5%	13.6%	1.9
Foreign	4.5%	12.0%	2.2
Normal	99.5%	13.6%	1.9
Sports	0.5%	12.9%	3.1
Yage_limit	95.3%	13.2%	1.9
Others (No Yage_limit)	4.7%	20.5%	2.3
Male	76.2%	13.0%	1.9
Female	23.8%	15.2%	1.8
Limit_one	36.0%	13.0%	2.0
Limit_two	3.1%	15.4%	2.0
Limit_family	18.7%	15.2%	2.0
Limit_couple	35.2%	12.9%	1.7
Others	6.9%	15.2%	2.1
Total	100.0%	13.6%	1.9

Table 24: summary for original sample

	Poisson		NB without random effects		NB with random effects	
Variables	Coeff	p-value	Coeff	p-value	Coeff	p-value
Intercept	-1.46718	> 0.0001	-1.46651	> 0.0001	-1.46631	> 0.0001
Male	-0.13474	> 0.0001	-0.13500	> 0.0001	-0.13521	> 0.0001
Domestic	0.18010	> 0.0001	0.18026	> 0.0001	0.17721	> 0.0001
Renewal	-0.11654	> 0.0001	-0.11656	> 0.0001	-0.11521	> 0.0001
Newcomer1	0.35127	> 0.0001	0.35181	> 0.0001	0.34998	> 0.0001
Newcomer2	0.19418	> 0.0001	0.19438	> 0.0001	0.19252	> 0.0001
Sports	-0.08032	0.0153	-0.08210	0.0150	-0.04389	0.1976
Yage_limit	-0.36983	> 0.0001	-0.37036	> 0.0001	-0.36933	> 0.0001
Age20	0.02019	0.1269	0.02046	0.1305	0.01825	0.1854
Age30	-0.08672	> 0.0001	-0.08695	> 0.0001	-0.08846	> 0.0001
Age40	-0.02270	0.0055	-0.02317	0.0055	-0.02505	0.0034
Age50	0.00430	0.5864	0.00398	0.6215	0.00104	0.8995
Cartype1	-0.13653	> 0.0001	-0.13693	> 0.0001	-0.13634	> 0.0001
Cartype2	-0.03896	0.0001	-0.03905	0.0001	-0.03979	0.0001
Cartype3	-0.05204	0.0001	-0.05187	0.0001	-0.05294	0.0001
Carage1	0.04391	0.0001	0.04423	0.0001	0.04412	0.0001
Carage2	0.09244	> 0.0001	0.09271	> 0.0001	0.09241	> 0.0001
Carage3	0.11934	> 0.0001	0.11947	> 0.0001	0.11869	> 0.0001
Seoul	-0.02130	0.0187	-0.02100	0.0231	-0.02127	0.0236
Gyeongn	-0.00078	0.9006	-0.00066	0.9175	-0.00050	0.9379
Gyeongg	0.00930	0.4591	0.00953	0.4571	0.01061	0.4144
Daej	-0.03219	0.0011	-0.03191	0.0016	-0.03180	0.0019
Chungn	-0.02434	0.0398	-0.02419	0.0451	-0.02490	0.0423
Jeonn	-0.02117	0.0338	-0.02101	0.0390	-0.02219	0.0319
Gyeongb	-0.11919	> 0.0001	-0.11899	> 0.0001	-0.12014	> 0.0001
Gangw	-0.09535	0.0003	-0.09493	0.0004	-0.09520	0.0005
Jeju	-0.00813	0.4684	-0.00771	0.5003	-0.00851	0.4648
Jeonb	0.00958	0.3571	0.00972	0.3599	0.00964	0.3708
Busan	-0.04721	0.0001	-0.04704	0.0002	-0.04646	0.0003
Chungb	0.08107	0.0001	0.08162	0.0001	0.08143	0.0001
Ulsan	0.11097	> 0.0001	0.11104	> 0.0001	0.11073	> 0.0001
Gwangj	0.06616	0.0001	0.06639	0.0001	0.06624	0.0001
Incheon	0.05154	0.0001	0.05162	0.0001	0.05056	0.0001
Daegu	-0.06606	0.1705	-0.06530	0.1836	-0.06637	0.1812
Limit_couple	-0.13791	> 0.0001	-0.13775	> 0.0001	-0.13911	> 0.0001
Limit_one	-0.20203	> 0.0001	-0.20239	> 0.0001	-0.20404	> 0.0001
Limit_two	-0.07318	0.0001	-0.07316	0.0001	-0.07410	0.0001
Limit_family	-0.06837	0.0001	-0.06844	0.0001	-0.06887	0.0001

Table 25: Regression results for 2009-2013 frequency data

	Gamma without random effects		Gamma with random effects	
Variables	Coeff	p-value	Coeff	p-value
Intercept	14.67596	> 0.0001	14.21960	> 0.0001
Male	0.04311	0.0667	0.00908	0.1166
Domestic	-0.13234	0.0180	-0.10317	0.0001
Renewal	-0.03476	0.1934	-0.00926	0.1269
Newcomer1	-0.06818	0.2515	0.00570	0.6753
Newcomer2	0.03737	0.3882	0.03053	0.0022
Sports	0.49753	0.0027	0.12661	0.0009
Yage_limit	-0.00308	0.9446	-0.06987	0.0001
Age20	0.04006	0.5408	-0.01299	0.4397
Age30	-0.00718	0.8653	-0.04715	0.0001
Age40	0.02365	0.5572	-0.01037	0.3337
Age50	0.03016	0.4408	0.02340	0.0204
Cartype1	-0.09852	0.0007	-0.09115	> 0.0001
Cartype2	-0.06017	0.0358	-0.05249	0.0001
Cartype3	0.01886	0.5923	-0.01070	0.1940
Carage1	0.09276	0.0026	0.08062	> 0.0001
Carage2	0.06669	0.0293	0.06560	> 0.0001
Carage3	0.05840	0.0765	0.05673	0.0001
Seoul	0.00015	0.9973	-0.17802	> 0.0001
Gyeongn	0.04905	0.1151	0.00641	0.3909
Gyeongg	0.05004	0.4228	0.00147	0.9230
Daej	0.12029	0.0144	-0.03891	0.0012
Chungn	0.16420	0.0052	-0.06442	0.0001
Jeonn	0.04971	0.3148	-0.13613	> 0.0001
Gyeongb	0.11875	0.0339	-0.06749	0.0001
Gangw	-0.21470	0.0978	-0.33082	> 0.0001
Jeju	0.15615	0.0051	-0.00627	0.6509
Jeonb	-0.17063	0.0010	-0.21366	> 0.0001
Busan	0.00870	0.8871	-0.02436	0.1074
Chungb	-0.14383	0.0434	-0.24832	> 0.0001
Ulsan	-0.00723	0.9072	-0.04496	0.0033
Gwangj	0.13140	0.0072	0.05293	0.0001
Incheon	-0.14416	0.0058	-0.17376	> 0.0001
Daegu	0.32931	0.1668	0.01793	0.7541
Limit_couple	-0.28871	0.0001	-0.13052	> 0.0001
Limit_one	-0.14692	0.0003	-0.05383	0.0001
Limit_two	-0.14243	0.0339	-0.01899	0.2190
Limit_family	-0.11897	0.0069	-0.04005	0.0001

Table 26: Regression results for 2009-2013 severity data

	2009-2012		2010-2012		2011-2012	
Variables	Coeff	p-value	Coeff	p-value	Coeff	p-value
Intercept	-1.37366	> 0.0001	-1.37133	> 0.0001	-1.37184	> 0.0001
R.ef. coeff	1.13414	> 0.0001	1.12856	> 0.0001	1.37667	> 0.0001
Male	-0.15048	> 0.0001	-0.14990	> 0.0001	-0.14986	> 0.0001
Domestic	0.19973	0.0001	0.19950	0.0001	0.19867	0.0001
Renewal	-0.10792	0.0001	-0.10900	0.0001	-0.10860	0.0001
Newcomer1	0.42714	> 0.0001	0.42598	> 0.0001	0.42694	> 0.0001
Newcomer2	0.21963	> 0.0001	0.21977	> 0.0001	0.22041	> 0.0001
Sports	-0.11838	0.2279	-0.12041	0.2202	-0.12238	0.2127
Yage_limit	-0.41093	> 0.0001	-0.41198	> 0.0001	-0.41158	> 0.0001
Age20	-0.01734	0.7094	-0.01807	0.6978	-0.01927	0.6788
Age30	-0.14325	0.0001	-0.14259	0.0001	-0.14300	0.0001
Age40	-0.06795	0.0009	-0.06792	0.0009	-0.06826	0.0009
Age50	-0.04043	0.0364	-0.03911	0.0431	-0.03930	0.0420
Cartype1	-0.13050	0.0001	-0.13035	0.0001	-0.13006	0.0001
Cartype2	-0.01957	0.2216	-0.01917	0.2313	-0.01885	0.2390
Cartype3	-0.03652	0.0550	-0.03623	0.0570	-0.03624	0.0569
Carage1	0.01630	0.3201	0.01651	0.3140	0.01674	0.3072
Carage2	0.03769	0.0184	0.03782	0.0180	0.03826	0.0167
Carage3	0.09439	0.0001	0.09449	0.0001	0.09469	0.0001
Seoul	0.03394	0.1684	0.03333	0.1763	0.03398	0.1680
Gyeongn	0.01405	0.4200	0.01383	0.4275	0.01395	0.4237
Gyeongg	0.00998	0.7710	0.00911	0.7904	0.01023	0.7654
Daej	0.00542	0.8416	0.00611	0.8218	0.00691	0.7990
Chungn	0.00038	0.9905	0.00025	0.9938	0.00111	0.9726
Jeonn	0.00706	0.7974	0.00795	0.7724	0.00740	0.7878
Gyeongb	-0.07231	0.0224	-0.07145	0.0241	-0.07064	0.0257
Gangw	-0.00030	0.9964	0.00044	0.9947	0.00191	0.9771
Jeju	0.05490	0.0699	0.05475	0.0707	0.05470	0.0710
Jeonb	0.06568	0.0199	0.06530	0.0207	0.06527	0.0208
Busan	-0.01687	0.6190	-0.01754	0.6052	-0.01666	0.6234
Chungb	0.14592	0.0002	0.14598	0.0002	0.14818	0.0002
Ulsan	0.15363	0.0001	0.15335	0.0001	0.15389	0.0001
Gwangj	0.02433	0.3826	0.02442	0.3809	0.02487	0.3722
Incheon	0.09710	0.0007	0.09695	0.0007	0.09728	0.0007
Daegu	-0.07566	0.5411	-0.07535	0.5428	-0.07277	0.5567
Limit_couple	-0.13344	0.0001	-0.13323	0.0001	-0.13352	0.0001
Limit_one	-0.21238	> 0.0001	-0.21165	> 0.0001	-0.21202	> 0.0001
Limit_two	-0.08600	0.0575	-0.08714	0.0543	-0.08812	0.0516
Limit_family	-0.06068	0.0126	-0.06167	0.0113	-0.06230	0.0105

Table 27: Predicting regression results frequency test data from various training set

	2009-2012		2010-2012		2011-2012	
Variables	Coeff	p-value	Coeff	p-value	Coeff	p-value
Intercept	-1.37366	> 0.0001	-1.37133	> 0.0001	-1.37184	> 0.0001
R.ef. coeff	1.13414	> 0.0001	1.12856	> 0.0001	1.37667	> 0.0001
Male	-0.15048	> 0.0001	-0.14990	> 0.0001	-0.14986	> 0.0001
Domestic	0.19973	0.0001	0.19950	0.0001	0.19867	0.0001
Renewal	-0.10792	0.0001	-0.10900	0.0001	-0.10860	0.0001
Newcomer1	0.42714	> 0.0001	0.42598	> 0.0001	0.42694	> 0.0001
Newcomer2	0.21963	> 0.0001	0.21977	> 0.0001	0.22041	> 0.0001
Sports	-0.11838	0.2279	-0.12041	0.2202	-0.12238	0.2127
Yage_limit	-0.41093	> 0.0001	-0.41198	> 0.0001	-0.41158	> 0.0001
Age20	-0.01734	0.7094	-0.01807	0.6978	-0.01927	0.6788
Age30	-0.14325	0.0001	-0.14259	0.0001	-0.14300	0.0001
Age40	-0.06795	0.0009	-0.06792	0.0009	-0.06826	0.0009
Age50	-0.04043	0.0364	-0.03911	0.0431	-0.03930	0.0420
Cartype1	-0.13050	0.0001	-0.13035	0.0001	-0.13006	0.0001
Cartype2	-0.01957	0.2216	-0.01917	0.2313	-0.01885	0.2390
Cartype3	-0.03652	0.0550	-0.03623	0.0570	-0.03624	0.0569
Carage1	0.01630	0.3201	0.01651	0.3140	0.01674	0.3072
Carage2	0.03769	0.0184	0.03782	0.0180	0.03826	0.0167
Carage3	0.09439	0.0001	0.09449	0.0001	0.09469	0.0001
Seoul	0.03394	0.1684	0.03333	0.1763	0.03398	0.1680
Gyeongn	0.01405	0.4200	0.01383	0.4275	0.01395	0.4237
Gyeongg	0.00998	0.7710	0.00911	0.7904	0.01023	0.7654
Daej	0.00542	0.8416	0.00611	0.8218	0.00691	0.7990
Chungn	0.00038	0.9905	0.00025	0.9938	0.00111	0.9726
Jeonn	0.00706	0.7974	0.00795	0.7724	0.00740	0.7878
Gyeongb	-0.07231	0.0224	-0.07145	0.0241	-0.07064	0.0257
Gangw	-0.00030	0.9964	0.00044	0.9947	0.00191	0.9771
Jeju	0.05490	0.0699	0.05475	0.0707	0.05470	0.0710
Jeonb	0.06568	0.0199	0.06530	0.0207	0.06527	0.0208
Busan	-0.01687	0.6190	-0.01754	0.6052	-0.01666	0.6234
Chungb	0.14592	0.0002	0.14598	0.0002	0.14818	0.0002
Ulsan	0.15363	0.0001	0.15335	0.0001	0.15389	0.0001
Gwangj	0.02433	0.3826	0.02442	0.3809	0.02487	0.3722
Incheon	0.09710	0.0007	0.09695	0.0007	0.09728	0.0007
Daegu	-0.07566	0.5411	-0.07535	0.5428	-0.07277	0.5567
Limit_couple	-0.13344	0.0001	-0.13323	0.0001	-0.13352	0.0001
Limit_one	-0.21238	> 0.0001	-0.21165	> 0.0001	-0.21202	> 0.0001
Limit_two	-0.08600	0.0575	-0.08714	0.0543	-0.08812	0.0516
Limit_family	-0.06068	0.0126	-0.06167	0.0113	-0.06230	0.0105

Table 28: Predicting regression results severity test data from various training set

	Gamma-GLMM Without Freq		Gamma-GLMM With Freq	
Variables	Coeff	p-value	Coeff	p-value
Intercept	14.21960	> 0.0001	14.05957	> 0.0001
Male	0.00908	0.1166	0.01106	0.0559
Domestic	-0.10317	0.0001	-0.10472	0.0001
Renewal	-0.00926	0.1269	-0.00719	0.2360
Newcomer1	0.00570	0.6753	-0.00439	0.7472
Newcomer2	0.03053	0.0022	0.02713	0.0065
Sports	0.12661	0.0009	0.12559	0.0010
Yage_limit	-0.06987	0.0001	-0.06536	0.0001
Age20	-0.01299	0.4397	-0.01081	0.5204
Age30	-0.04715	0.0001	-0.04484	0.0001
Age40	-0.01037	0.3337	-0.00967	0.3675
Age50	0.02340	0.0204	0.02353	0.0198
Cartype1	-0.09115	> 0.0001	-0.08937	> 0.0001
Cartype2	-0.05249	0.0001	-0.05220	0.0001
Cartype3	-0.01070	0.1940	-0.01042	0.2062
Carage1	0.08062	> 0.0001	0.08145	> 0.0001
Carage2	0.06560	> 0.0001	0.06559	> 0.0001
Carage3	0.05673	0.0001	0.05580	0.0001
Seoul	-0.17802	> 0.0001	-0.17661	> 0.0001
Gyeongn	0.00641	0.3909	0.00656	0.3802
Gyeongg	0.00147	0.9230	0.00135	0.9291
Daej	-0.03891	0.0012	-0.03818	0.0015
Chungn	-0.06442	0.0001	-0.06340	0.0001
Jeonn	-0.13613	> 0.0001	-0.13441	> 0.0001
Gyeongb	-0.06749	0.0001	-0.06505	0.0001
Gangw	-0.33082	> 0.0001	-0.32620	> 0.0001
Jeju	-0.00627	0.6509	-0.00537	0.6985
Jeonb	-0.21366	> 0.0001	-0.21300	> 0.0001
Busan	-0.02436	0.1074	-0.02381	0.1156
Chungb	-0.24832	> 0.0001	-0.24820	> 0.0001
Ulsan	-0.04496	0.0033	-0.04471	0.0035
Gwangj	0.05293	0.0001	0.05271	0.0001
Incheon	-0.17376	> 0.0001	-0.17351	> 0.0001
Daegu	0.01793	0.7541	0.01873	0.7436
Limit_couple	-0.13052	> 0.0001	-0.12753	> 0.0001
Limit_one	-0.05383	0.0001	-0.05149	0.0001
Limit_two	-0.01899	0.2190	-0.01749	0.2575
Limit_family	-0.04005	0.0001	-0.03814	0.0002
Frequency			0.13686	> 0.0001

Table 29: Gamma-GLMM regression results with/without frequency

References

Boucher, J.-P., Denuit, M., & Guillen, M. (2008). Models of insurance claim counts with time dependence based on generalization of Poisson and negative binomial distributions. *Variance, 2*(1), 135-162.

Boudreault, M., Cossette, H., Landriault, D., & Marceau, E. (2006). On a risk model with dependence between interclaim arrivals and claim sizes. *Scandinavian Actuarial Journal, 2006*(5), 265-285.

Coene, G., & Doray, L. G. (1996). A financially Balanced Bonus/Malus System. *Astin Bulletin, 26*(01), 107-116.

Czado, C., Kastenmeier, R., Brechmann, E. C., & Min, A. (2012). A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal, 2012*(4), 278-305.

Dionne, G., & Vanasse, C. (1989). A generalization of automobile insurance rating models: the negative binomial distribution with a regression component. *Astin Bulletin, 19*(2), 199-212.

Frangos, N. E., & Vrontos, S. D. (2001). Design of optimal bonus-malus systems with a frequency and a severity component on an individual basis in automobile insurance. *Astin Bulletin, 31*(01), 1-22.

Gschlößl, S., & Czado, C. (2007). Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal, 2007*(3), 202-225.

Hernández-Bastida, A., Fernández-Sánchez, M., & Gómez-Déniz, E. (2009). The net Bayes premium with dependence between the risk profiles. *Insurance: Mathematics and Economics, 45*(2), 247-254.

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

Jacqmin-Gadda, H., & Commenges, D. (1995). Tests of homogeneity for generalized linear models. *Journal of the American Statistical Association, 90*(432), 1237-1246.

Klugman, S. A., Panjer, H. H., & Willmot, G. E. (2012). *Loss models: from data to decisions* (Vol. 715): John Wiley & Sons.

Lemaire, J. (1995). *Bonus-Malus Systems in Automobile Insurance* (Vol. 19): Springer Science & Business Media.

Mahmoudvand, R., & Hassani, H. (2009). Generalized bonus-malus systems with a frequency and a severity component on an individual basis in automobile insurance. *Astin Bulletin, 39*(01), 307-315.

Mert, M., & Saykan, Y. (2005). On a bonus malus system where the claim frequency distribution is geometric and the claim severity distribution is pareto. *Hacettepe Journal of Mathematics and Statistics, 34*, 75-81.

MOLENBERGHS, G. V. G. (2005). Longitudinal and incomplete clinical studies. *Metron, 63*(2), 143-176.

Picard, P. (1976). Generalisation de l'etude sur la survenance des sinistres en assurance automobile. *Bulletin Trimestriel de l'Institut des Actuaires Français, 296*, 204-268.

Pinquet, J. (1997). Allowance for cost of claims in bonus-malus systems. *Astin Bulletin, 27*(01), 33-57.

Shi, P., Feng, X., & Ivantsova, A. (2015). Dependent frequency–severity modeling of insurance claims. *Insurance: Mathematics and Economics, 64*, 417-428.

Smyth, G. K., & Jørgensen, B. (2002). Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling. *Astin Bulletin, 32*(01), 143-157.

Tremblay, L. (1992). Using the Poisson inverse Gaussian in bonus-malus systems. *Astin Bulletin, 22*(01), 97-106.

Tzougas, G., Vrontos, S., & Frangos, N. (2014). Optimal bonus-malus systems using finite mixture models. *Astin Bulletin, 44*(02), 417-444.

Walhin, J., & Paris, J. (1999). Using mixed Poisson distributions in connection with Bonus-Malus systems. *Astin Bulletin, 29*(1), 81-99.

국문초록

자동차보험료 할인할증체계 수립에 있어서, 보험사고의 횟수에 비하여 보험사고의 크기는 상대적으로 적게 논의 되어왔다. 본 논문은 일반화혼합선형모형(generalized linear mixed model)을 한국 보험시장으로부터 얻은 5년간의 패널 데이터에 적용하여 사고 횟수와 크기에 영향을 미치는 독립변수들, 그리고 둘 사이의 의존성 여부를 조사했으며, 또한 모형에서 얻어진 개개인의 계수(individual random effect) 및 과거의 사고 정보를 미래의 사고 예측을 위하여 사용했다.

분석 결과에 따르면, 일반화혼합선형모형은 단순한 일반화선형모형(generalized linear model) 등 여타 다른 모형들과 비교하였을 때 가장 좋은 모형 적합도를 갖는 것으로 밝혀졌고, 사고 횟수와 사고 크기 간의 상관관계 역시 무시할 수 없었다. 또한 개개인의 계수 추출에 사용되는 데이터의 기간이 길어질수록, 미래의 사고에 대한 계수의 예측력은 더 좋아지는 것으로 나타났다. 이 결과들은 과거의 사고 횟수와 크기를 동시에 보험료 할인할증체계에 반영되어야 할 것과, 자동차보험료 계산에 있어서 사고 횟수와 크기 간의 의존성이 고려되어야 할 논리적 근거를 제공한다.

주요어 : 할인할증체계, 확률효과모형, 사고 크기, 보험료예측, 의존성, 자동차보험 복합위험모형, 일반화혼합선형모형

학 번 : 2014-22361