연구보고서 2009-05

일반화선형모형(GLM)을 이용한

자동차보험 요율상대도 산출방법 연구

2009. 8

기승도 · 김대환

보험연구원

머 리 말

보험에서는 통계가 매우 중요하다. 보험회사의 모든 활동은 통계자료에 근거하여 이루어진다. 보험료 산출, 언더라이팅 활동 및 향후 보험시장 예측 등 보험의 모든 분야에서 회사에 집적된 통계 및 외부통계가 활용된다. 이런 통계의 중요성 때문에 보험회사들은 많은 자원을 투입하여 통계집적 전산설비를 갖추고, 통계를 분석할 수 있는 다양한 방법을 연구하고 있다.

통계를 집적·분석하여 보험회사를 운영하는 활동은 모든 보험종목에서 동일하게 이루어지고 있으나, 보험의 성격 때문에 통계의 집적측면에서 보험종목별로 다소 차이가 있다. 개인보험 성격의 보험종목은 통계집적이 원활히 이루어지고 있으며, 잘 사용되고 있다. 특히 자동차보험은 개인보험 성격이 있으면서 의무보험이므로 통계집적이 매우 잘되고 있는 보험종목 중 하나이다. 자동차보험에서는 집적된 통계를 활용하여 자동차보험 여러 분야에 적용하는 활동이 과거부터 활발히 이루어져 왔다.

자동차보험 통계를 사용하여 적정 지급준비금 수준을 산출하거나, 시장세분화를 통한 시장별 자동차보험 요율 상대도를 산출하고, 기본보험료 산출 및 언더라이팅 기준마련 등 많은 분야에 자동차보험 통계가 활용되고 있다. 이러한 통계모형은 일반적으로 통계활동에 사용되므로 전통적인 회귀분석 방법이나 과거부터 관례적으로 사용되던 기계적 계산모형이었다. 기존의 모형들은 과거 60년대 전후에 개발된 것으로 사용의 편리성이 있지만 위험도 평가에 사용되는 독립변수의 통계적 유의성을 파악하기 어렵거나, 종속변수(주로 사고발생률 및 1사고당 손해액)에 대한 분포가정이 비현실적인 등의 문제점이 있다.

이러한 전통적 방법의 단점 때문에 최근에는 자동차보험 통계분석에 최신의 통계모형을 활용하고 있다. 최근에 Data mining 분야에서 Neural Network 모델이나 Logistic 회귀분석 등을 활용하여 언더라이팅 기준을 설정하는 등의 현상은 최근 통계모형을 활용한 자료 분석 추세를 반영하는 것이다. Datamining에서 사용되는 Logistic 회귀모형은 일반화선형모형의 일종인데, 일반화선형모형이 특히 자동차보험 분야에서 널리 활용되고 있

다. 자동차보험 요율상대도 산출, 언더라이팅 기준 작성 및 준비금 평가 등 자동차보험 여러 분야에서 일반화선형모형이 널리 활용되고 있다.

그런데 우리나라 자동차보험 분야에서는 일반화선형모형 활용이 매우 미미한 실정이다. 일반화선형모형을 언더라이팅 기준 마련을 위해서 사용하는 일부회사가 있으나, 자동차보험 산업전체적으로는 널리 사용되지 않고 있다.

자동차보험 분야는 경쟁이 매우 치열한 보험산업이다. 치열한 경쟁환경에서 경영효율화를 달성하고, 수익을 내기 위해서는 경쟁보험회사보다 더 나은 통계분석능력을 갖추는 것이 매우 중요하다. 정교하고 효율적인 통계분석을 위해서 중요한 수단으로 활용될 수 있는 통계모형중 하나가 최근에 널리 활용되고 있는 일반화선형모형이다. 따라서 자동차보험산업의 경쟁력 향상을 위해서는 각 보험회사가 일반화선형모형의 적용방법을 익히고 사용하는 능력을 제고할 필요가 있을 것이다.

이에 본 연구원에서는 최신 통계모형의 활용 필요성이 점증하는 상황에 대응하여 '일반화선형모형을 이용한 자동차보험 요율상대도 산출방법'이라는 제목으로 일반화선형모형을 적용하는 방법을 연구하였다. 본 연구에서는 자동차보험의 요율산출 분야로 한정하여 연구하였지만, 일반화선형모형의 적용방법은 자동차보험 여러 분야로 확장할 수 있다. 아무쪼록 본 보고서가 자동차보험 산업의 선진화에 기여하고, 자동차보험 산업의 발달에 기여할 수 있기를 기대한다.

마지막으로 본 보고서에 수록된 내용은 연구자 개인의 의견이며, 우리원의 공식 견해가 아님을 밝혀둔다.

2009년 8월

보 험 연 구 원

원 장 나 동 민

목 차

요 약 **1**

Ⅰ. 서론 **11**
1. 연구배경 및 목적 11
2. 연구범위 13

Ⅱ. 자동차보험 요율산출 방법 **15**
1. 자동차보험 요율산출의 의미 15
2. 자동차보험 요율산출 방법 20

Ⅲ. 요율상대도 산출시 일반화선형모형 활용방법 **41**
1. 선행연구 41
2. 일반화선형모형 이론 44
3. 자동차보험에서 일반화선형모형의 적용방법 70

Ⅳ. 일반화선형모형을 이용한 실증분석 **80**
1. 모형적용기준 80
2. 통계자료 및 통계모형 90
3. 실증분석결과 94
4. 추가 논의사항 127

Ⅴ. 결론 및 시사점 **146**

참고문헌 **149**

< 표 차례 >

<표 Ⅱ-1> 요율상대도 자료형태 ··········· 30
<표 Ⅲ-1> canonical 연결(link) 및 분산(variance) 함수 ··········· 48
<표 Ⅲ-2> 사고빈도 ··········· 57
<표 Ⅲ-3> 기타 일반화선형모형 ··········· 61
<표 Ⅲ-4> AIC와 BIC 값의 비교와 모델의 선택 ··········· 70
<표 Ⅲ-5> 보험에서 일반화선형모형(GLM)을 적용하는 일반적 모형 · 75
<표 Ⅳ-1> 개인용 자동차보험 가입경력요율제도 ··········· 84
<표 Ⅳ-2> 교통법규위반 경력요율 적용기준 ··········· 85
<표 Ⅳ-3> 개인용자동차보험(플러스 개인용 포함) 기명피보험자 연령요율 구분 · 86
<표 Ⅳ-4> 지역범주 구분 기준 ··········· 90
<표 Ⅳ-5> 통계자료의 기준 ··········· 91
<표 Ⅳ-6> 과대산포 테스트 ··········· 96
<표 Ⅳ-7> Vuong 테스트: Excessive Zero 테스트(대인배상I) ··········· 96
<표 Ⅳ-8> AIC, BIC를 이용한 일반화선형모델 비교: 사고빈도 ··········· 97
<표 Ⅳ-9> AIC, BIC를 이용한 일반화선형모델 비교: 사고심도 ··········· 104
<표 Ⅳ-10> 일반화선형모델을 이용한 회귀분석 결과 ··········· 105
<표 Ⅳ-11> 가입경력과 위험도 ··········· 108
<표 Ⅳ-12> 일반화선형모형으로 추정된 모수들의 발생률 (Incidence Rate Ratio)로의 전환: 상대위험도 ··········· 110
<표 Ⅳ-13> 자동차 연식에 따른 교통사고 빈도 및 기술적 통계 ··········· 114
<표 Ⅳ-14> Vuong 테스트를 이용한: 사고빈도 ··········· 114
<표 Ⅳ-15> AIC, BIC를 이용한 일반화선형모델 비교: 사고심도 ··········· 116
<표 Ⅳ-16> 자차담보 분석: ZINB와 Gamma 적용 ··········· 117
<표 Ⅳ-17> 대인배상Ⅱ: NB와 Gamma모델 적용 ··········· 119
<표 Ⅳ-18> 대물배상: NB와 Lognormal모델 적용 ··········· 121

<표 Ⅳ-19> 자기신체사고: NB와 Lognormal 모델 적용 ······················ 123
<표 Ⅳ-20> 무보험차상해: NB와 Gamma 모델 적용 ···························· 125
<표 Ⅳ-21> 모형별 요율상대도 비교(대인배상Ⅰ, 사고빈도) ················ 129
<표 Ⅳ-22> 대인배상Ⅰ: 일반화선형모델을 이용한 회귀분석 결과 ···· 130
<표 Ⅳ-23> 일반화선형모형으로 추정된 모수들의 발생률
(Incidence Rate Ratio)로의 전환: 상대위험도 ···················· 132
<표 Ⅳ-24> 성별과 지역의 포함여부와 차종의 계수 비교 ·················· 134
<표 Ⅳ-25> 성별과 차종을 이용한 교호작용: 사고빈도 ························ 138
<표 Ⅳ-26> 성별과 차종을 이용한 교호작용: 사고심도 ························ 138
<표 Ⅳ-27> 성별과 차종의 교호작용 적용여부 검증: 사고빈도 ·········· 139
<표 Ⅳ-28> 성별과 연령을 이용한 교호작용: 사고빈도 ························ 139
<표 Ⅳ-29> 성별과 가입경력을 이용한 교호작용: 사고빈도 ················ 140
<표 Ⅳ-30> 성별과 가입경력을 이용한 교호작용: 사고심도 ················ 140
<표 Ⅳ-31> 교호작용을 이용한 일반화선형모형 ···································· 141
<표 Ⅳ-32> 성별과 가입경력에 따른 요율분석: 교호작용 적용 vs. 교호작용 비적용 ···· 143
<표 Ⅳ-33> 교호작용 적용의 정당성 검증 ·· 144
<표 Ⅴ-1> 요율분석에 적용된 일반화선형모형(GLM) ························ 146

< 그림 차례 >

<그림 Ⅱ-1> 대인배상I의 사고빈도 ······ 38
<그림 Ⅱ-2> 대물배상의 사고빈도 ······ 39
<그림 Ⅱ-3> 대인배상I의 사고심도 ······ 39
<그림 Ⅱ-4> 대물배상의 사고심도 ······ 40
<그림 Ⅲ-1> 종속변수의 연속성 여부에 따른 GLM의 분류 ······ 49
<그림 Ⅲ-2> 대인I 사고심도: 로그값 변환 전 ······ 64
<그림 Ⅲ-3> 대인I 사고심도: 로그값 전환 후 ······ 65
<그림 Ⅳ-1> 운전자 연령과 교통사고빈도: 대인배상Ⅰ ······ 88
<그림 Ⅳ-2> 사고빈도의 경우 계량모델 ······ 95
<그림 Ⅳ-3> 가우시안 모델 검정: Pearson 잔차항 히스토그램 ······ 98
<그림 Ⅳ-4> 가우시안 모델 검정: Pearson 잔차항과 임의데이터 비교 ·· 99
<그림 Ⅳ-5> Log-Linked 모델 검정: Pearson 잔차항 히스토그램 ······ 100
<그림 Ⅳ-6> Lognormal 모델 검정: Pearson 잔차항 히스토그램 ······ 101
<그림 Ⅳ-7> Lognormal 모델 검정: Pearson 잔차항과 임의데이터 비교 ·· 102
<그림 Ⅳ-8> Lognormal 모델 검정: Pearson 잔차항과 추정값 관계 ·· 103
<그림 Ⅳ-9> 자동차 연식에 따른 교통사고 빈도 ······ 115

요 약

Ⅰ. 요율상대도 산출시 일반화선형모형 활용방법

1. 일반화 선형모형

□ 일반화선형모형은 랜덤성분(Random component), 체계적성분(Systematic Component), 연결함수(Link Function)로 구성되어 있음.

○ 랜덤성분(Random Component)은 종속변수(또는 반응변수) Y의 확률분포를 규정하는 성분임.

○ 체계적성분(Systematic Component)은 모형의 예측변수로 사용되는 설명변수를 규정하여 확률분포의 평균인 Y의 기댓값$E(Y)=\mu$을 나타냄.

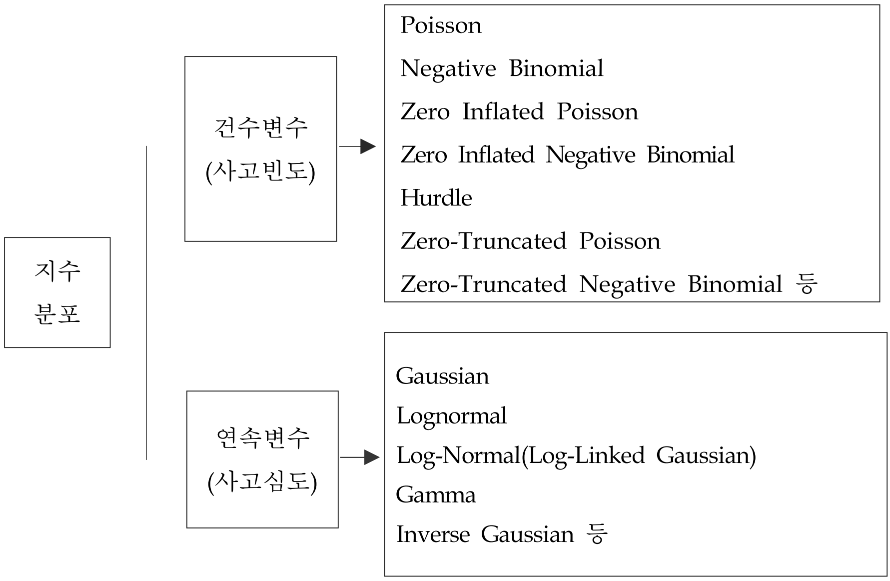
○ 연결함수(Link Function)는 첫 번째 성분인 랜덤성분($E(Y)=\mu$)과 두 번째 성분인 체계적성분($X\beta$)을 연결함.

□ 일반화선형모형은 랜덤성분인 Y가 지수족(exponential family)분포를 따르는 자료들에 적용 가능하며 링크함수로는 모든 구간에서 미분 가능한 단조증가 함수가 모두 사용될 수 있음.

$$g(\mu)=a+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\ldots\ldots+\beta_k x_k \qquad (\text{Ⅱ-1})$$

□ 일반화선형모형은 다음 그림과 같이 종속변수의 분포 종류에 따라 다양한 형태가 있음.

<요약 그림> 종속변수의 연속성 여부에 따른 GLM의 분류



2. 일반화선형모형의 자동차보험 요율산출에 적용방법

□ 자동차보험 요율산출에 회귀분석을 적용하는 것은 첫째 '통계정리', 둘째 '상대도 모형설정', 셋째 '모형적합', 넷째 '결과해석'의 절차를 따름.

□ 통계정리

○ 통계모형에 적용되는 독립변수 자료는 범주형 변수와 연속변수가 혼합되어 있는데, 자동차보험 요율은 범주형 기준에 따라 산출·적용되므로 연속형 독립변수 자료는 범주형으로 변환되어야 함.

○ 자동차보험에서 범주형 자료의 예를 들면, 연속변수인 가입경력을 최초가입자, 1년 이상 2년 미만, 2년 이상 3년 미만 등등으로 구분하는 것임.

□ 모형설정

○ 모형설정은 종속변수에 해당하는 자료에 적합한 분포선택과 가장 부합된 연결함수(link function)를 결정하는 것임.

○ 먼저 종속변수에 해당하는 자료에 부합되는 분포결정을 보면,

– 종속변수의 통계적 특성을 반영하는 분포를 선택하는 단계임.

– 사고빈도에서는 포아송분포(Poisson distribution), 음이항분포(Negative binomial distribution) 등이 있음.

– 사고심도에서는 감마분포(Gamma distribution), 로그노말(Lognormal)이 있으며, 순보험료 기준으로는 트위디분포(Tweedie distribution)가 있음.

○ 연결함수(Link Function)를 결정하는 단계를 보면,

– 여러 연결함수 중에서 통계모형과 자료를 가장 잘 적합시킬 수 있는 연결함수를 선택하는 것이 중요함.

– 현행 자동차보험 요율체계는 승산모형으로 구축되어 있으므로, 이러한 자동차보험 요율체계를 가장 잘 설명할 수 있는 연결함수는 로그(Log) 연결함수임.

□ 모형적합

○ 모형에 포함되는 변수선택은 우선 자동차보험 요율체계에 사용되는 모든 변수를 모형에 포함시키는 것이 원칙임.

– 그러나 현재 자동차보험 요율체계에 포함되어 있다 하더라도, 통계적으로 의미가 없는 변수가 있을 수 있으며, 이러한 변수를 찾는 방법은 Deviance 분석등 통계적인 분석과정을 거쳐 선별함.

○ 분석에 적용될 변수를 선택한 이후에는 종속변수 또는 오차항(error term)을 잘 설명할 수 있는 분포 및 연결함수를 선택하여 모형을 적

합 시키는 단계로 넘어감.

- 모형을 자료에 적용한 이후에 해당 모형이 적정한지 확인을 하여야 하는데, 확인방법은 잔차항(Residuals)을 이용한 여러 가지 통계분석을 활용하는 것임.

□ 결과의 해석

○ 산출된 결과를 독립변수의 기준그룹(Base Group)을 기준으로 추정된 위험변수의 계수값의 수치를 IRR(Incidence Rate Ratio)로 전환하여 ‘요율상대도’값으로 해석함(자세한 해석방법은 본문 참조).

Ⅱ. 일반화선형모형을 이용한 실증분석

1. 모형적용기준

□ 본 연구의 일반화선형모형에 포함되는 변수 및 적용보험료 계산식은 다음과 같음(곱셈방식의 적용보험료 계산식은 모든 담보 및 차종과 동일함).

<요약 표> 본 연구의 적용모형

적용 보험료	=	기본 보험료	×	가입경력 요율	×	교통법규위반 경력요율	×	우량할인·불량 할증요율

×	기명피보험자 연령요율	×	성	×	지역

□ 본 분석에서 독립변수의 기준수준(base level)은 다음과 같이 정함.

○ 연령은 '31~39세', 성은 '여자', 차종은 '중형', 지역은 '서울', 할인할증요율은 '100', 가입경력은 '4년 이상', 교통법규위반은 '할인계층'임.

2. 통계자료 및 통계모형

□ 통계자료는 업계 전체자료 중 일부를 표본 추출한 것임.

○ 보험개발원에 집적된 FY2005의 총 자료 중에서 표본추출(Random Sampling)한 자료(FY2005에 약 50만개)를 사용함.

<요약 표> 통계자료의 기준

기 준		세부분류
독립변수	성	남, 여
	연령	실제 기명피보험자 연령
	가입경력	최초가입자~1년 미만은 '1', 1년 이상 2년 미만은 '2', 2년 이상 3년 미만은 '3', 3년 이상 4년 미만은 '4', 4년 이상 '5'
	지역	'광역시도' 대분류 기준
	법규위반	현행기준 참조순보험료 요율서 기준
	할인할증률	40 ~ 200%
종속변수	대수	통계기간 중 평균유효대수로 추출
	사고건수	담보별 사고건수
	보험료	수입보험료
	손해액	담보별 손해액, 2008년 3월 말 현재 평가 OS(미지급 보험금)

주 : 1) 개인용 및 플러스 개인용 자동차보험(모든 담보)이 대상이다.
2) 할인할증률은 45%를 제외하고 5%단위에서 절하 하였다.
3) 연령은 연구자가 임의로 동일위험 집단 기준으로 묶어서 범주형 자료로 변환한 것이다.

□ 본 연구의 분석에 사용된 모형인 일반화선형모형에서 반응변수 Y의 기본적인 확률분포인 지수족(Exponential Family)의 일반 형태는 다음과 같음.

$$Y = X^T\beta + \epsilon,\ \epsilon \sim N(0, \sigma^2)$$

○ 여기서 Y 는 사고발생률(이하 "사고빈도"라고 한다) 또는 1사고당손해액(이하 "사고심도"라고 한다)이며, θ는 추정해야할 모수, ϕ은 표준오차(Standard Error)를 계산할 때 필요한 가중치(일반화선형모형에 대한 자세한 이론적 논의는 본문 참조)이고, 독립변수는 앞의 통계추출 기존의 독립변수 항목임.

3. 실증분석결과

□ 본 요약에서는 대인배상 I 의 분석결과만 제시하였으며, 기타 담보는 본문을 참조하기 바람.

<요약 표> 일반화선형모형으로 추정된 모수들의 발생률(Incidence Rate Ratio)로의 전환: 상대위험도

구 분		사고빈도		사고심도	
		Coefficient	Std. Dev.	Coefficient	Std. Dev.
연령	25세이하	1.408	0.042***	0.991	0.040
	26~30세	1.149	0.025***	1.013	0.024
	40~64세	1.186	0.015***	1.028	0.015*
	65세이상	1.140	0.038***	1.044	0.036
성	남자	0.931	0.016***	1.041	0.015***
차종	소형A	0.843	0.029***	0.941	0.028**
	소형B	0.924	0.017***	0.946	0.016***
	대형	0.918	0.025***	1.054	0.024**
	승합1	0.671	0.123***	1.063	0.115

	승합2	1.133	0.018***	1.053	0.017***
지역	부산	0.831	0.031***	0.879	0.030***
	경기	0.929	0.020***	1.079	0.019***
	강원	0.812	0.042***	1.179	0.040***
	충북	0.784	0.043***	0.976	0.040
	충남	0.798	0.038***	1.039	0.036
	전북	1.018	0.037	1.039	0.035
	전남	0.724	0.045***	1.047	0.043
	경북	0.768	0.034***	0.931	0.032**
	경남	0.724	0.033***	0.979	0.031
	제주	0.549	0.081***	0.694	0.077***
	대구	0.884	0.032***	0.803	0.030***
	인천	1.162	0.029***	1.133	0.028***
	광주	0.896	0.040***	0.962	0.038
	대전	0.959	0.036	0.981	0.034
	울산	0.782	0.047***	0.906	0.045**
할인 할증	할인40%	0.612	0.031***	0.952	0.029*
	할인45%	0.770	0.035***	0.949	0.033
	할인50%	0.809	0.034***	0.959	0.033
	할인60%	0.869	0.035***	1.001	0.033
	할인70%	0.889	0.034***	0.991	0.033
	할인80%	0.914	0.034***	1.020	0.032
	할인90%	0.977	0.032	1.018	0.031
	할증110%	1.120	0.046**	0.952	0.044
	할증120%	1.152	0.053***	1.096	0.050*
	할증130%	1.349	0.062***	1.058	0.058
	할증140%	1.239	0.090**	1.209	0.085**
	할증150%	1.599	0.107***	1.124	0.102
	할증160%	1.465	0.147***	1.018	0.143
	할증170%	1.872	0.195***	1.091	0.190
	할증180%	1.399	0.277	1.165	0.263
	할증190%	1.632	0.328	1.048	0.334

	할증200%	2.279	0.273***	1.066	0.299
가입경력	1년미만	1.355	0.038***	1.068	0.036*
	1~2년	1.142	0.035***	1.015	0.033
	2~3년	1.058	0.032*	0.968	0.031
	3~4년	0.988	0.026	1.009	0.025
법규위반	할증1그룹1	0.939	0.588	1.059	0.544
	할증1그룹2	1.093	0.032***	0.995	0.030
	할증2그룹	1.124	0.109	1.075	0.106
	기본그룹	1.130	0.016***	0.987	0.015

주 : Robust 표준오차에 기준하여 *은 유의수준 10%, **은 유의수준 5%, ***은 유의수준 1% 수준에서 통계적 유의성을 나타낸다.

4. 추가 논의사항

가. 적정 분석방법(또는 분석모형) 선택

□ 자동차보험 요율상대도 계산은 단변량방법 보다는 다변량 방법이 더 적정한 것으로 보임.

○ 다변량 방법은 특정 위험집단에 대한 위험도 평가를 정확히 할 수 있으며, 컴퓨터의 발달로 계산 비용도 많이 소요되지 않음.

□ 기계적 요율상대도 산출방법보다는 통계적 모형, 그중에서도 일반화선형모형을 이용하는 것이 더 좋은 선택인 것으로 판단됨.

○ 모형의 안정성, 모형의 적용적합성 및 발전단계를 볼 때 일반화선형모형이 가장 적합한 모형으로 판단됨.

나. 변수선택 범위

□ 사용되고 있는 변수를 사용하여 자동차보험 요율상대도를 산출하는 방법으로 회사간 일관성을 유지하는 것이 바람직한 것으로 판단됨.

○ 변수의 범위에 대한 시장참여자의 수용가능성을 볼 때, 사회적, 법률적 측면에서 허용되는 변수로 변수의 범위를 정하는 것이 적합한 것으로 보임.

다. 산출계수의 사회적 수용성

□ 사회적으로 합의된 요율계수가 포함된 일반화 선형모형으로 요율상대도 값을 산출하더라도, 요율계수의 수준 등에서 사회적 및 법률적 수용성을 고려하여야 함.

○ 예를 들면, 현재 자동차보험 할인할증률의 한도는 40%~200%이므로, 통계모형으로 산출된 할인할증 계수 중 200%를 초과하는 계수는 200%한도로 제한하는 것 등임.

라. 교호작용 효과 분석

□ 요율산출에 많은 변수가 사용되고, 변수별 위험집단을 더미변수로 구분하여 분석하는 경우에는 모든 교호작용의 효과를 분석하는 것은 비경제적인 것임.

□ 따라서 변수간 교호작용이 있는 변수를 찾아내고, 교호작용이 뚜렷하게 나타난 경우에 이를 감안하여 분석하는 것이 적합한 것으로 판단됨(세부적인 분석방법은 본문을 참조).

마. 기존 변수의 일부 상대도 유지시 분석방법(Offset 활용)

□ 요율상대도 계산시에 요율변수 중 일부 변수의 요율상대도는 현행을 유지하면서 기타 변수의 요율상대도를 산출할 때 사용하는 방법이 Offset을 활용하는 것임.

□ 구체적인 산출방법은 Data set에서 기존 요율상대도를 유지하려는 변수의 X변수의 값을 현행 상대도 값으로 변환하고, 모형설정시 현행 상대도 값을 유지하려는 변수를 Offset으로 설정하여 분석하면 됨.

Ⅲ. 결론 및 시사점

□ 우리나라 자동차보험 통계에서 담보별로 최적인 모형을 찾아낸 결과는 다음과 같음.

<요약 표> 요율분석에 적용된 일반화선형모형(GLM)

담 보	적용모델		
	사고빈도 GLM	사고심도	
		GLM	링크함수(Link)
대인배상Ⅰ	음이항분포	Lognormal	항등(identity)
대인배상Ⅱ	음이항분포	감마(Gamma)	로그(Log)
자기차량	ZINB	감마(Gamma)	로그(Log)
대물배상	음이항분포	Lognormal	항등(identity)
자기신체사고	음이항분포	Lognormal	항등(identity)
무보험차상해	음이항분포	감마(Gamma)	로그(Log)

Ⅰ. 서 론

1. 연구배경 및 목적

현재 우리나라 자동차보험 산업에서 요율산출자(actuary)는 자동차보험 요율을 산출할 때 전통적인 방법을 주로 사용하고 있다. 자동차보험 요율산출의 전통적 방법은 손해율법과 순보험료법을 사용하여 단변량 또는 산술적 방식으로 요율을 산출하는 것이다. 이러한 전통적인 방식은 요율산출방식에 대한 노하우(know-how)가 수십 년 동안 집적되어 있고, 요율산출자(actuary)가 이러한 방식으로 산출된 결과를 용이하게 해석할 수 있다는 장점이 있다. 이러한 이유로 전통적인 방법의 요율산출 방법은 과거에서부터 현재까지 계속적으로 연구되고 있다. 그리고 이러한 방식으로 요율을 산출하는 교육을 받은 요율산출자도 다수 배출되어 있다.

단변량 기준으로 요율을 산출하는 단순한 작업은 전산프로그램으로 가능하다. 하지만 전산의 자료 처리(data handling)에 의존한 요율산출과 위험도 분석은 업무의 효율성을 떨어뜨린다. 실제 요율산출작업, 특히 위험집단을 세분화하거나, 세부 집단별 상대위험도를 측정하는 작업을 전산 프로그램으로 처리하려면 많은 시간과 자원이 소요된다. 전산작업으로 복잡한 요율산출 결과가 나왔다고 하더라도, 그 결과는 통계적 유의성을 확인할 수 없는 평균값에 지나지 않는다.

요율산출자는 위험집단을 세분화하는 끊임없는 작업을 통하여 경쟁회사보다 더 나은 요율경쟁력을 확보하기 위해 노력한다. 이를 위해서 요율산출자는 잘 구축된 전산을 이용하여 세부 위험요소별로 손해율을 관찰하여 위험도가 변하는 집단을 계속적으로 주시한다. 이러한 작업은 주로 각 변수별로 진행된다. 즉 단변량 작업으로 위험도를 관찰하는 작업이 진행되는 것이다. 일반적으로, 대부분의 통계는 단변량으로 분석하더라도 위험도 변화추이를 어느 정도 확인할 수 있다. 그러므로 단변량 통계를 이용하여 위험집단별 손해율로 위험도를 계속 추적하는 것은 위험도가 변하는 집단을 추적

하는데 유효한 방법이 될 수 있다.

그러나 위험집단별로 위험도를 평가는 단계에 들어가면 단변량 위주의 분석은 여러 가지 문제에 직면하게 된다. 변수들 사이에는 통계적으로 상호 연관성이 존재하기 때문이다. 이것을 우리는 통계적으로 상관관계라고 하는데, 상관관계가 있는 변수들을 각 변수별 단변량으로 위험도를 산출하면 정확한 위험도가 산출되지 않는다. 이러한 문제를 해결하기 위해서 과거 요율산출방법을 연구하는 사람들은 다양한 다변량적 요율산출방법을 고안하여 제시하였다. 예를 들면 Baily and Simon method, Baily's method, Minimum bias method 등이 그것이다. 이러한 방법들은 실제값과 계산된 상대도 값으로 표현된 예측값의 차이가 최소화되도록 상대도 값을 계산하는 방식이다. 이러한 방법들은 구분된 집단사이의 가중치를 어떠한 방식으로 하느냐에 따라 차이가 있을 뿐이며 방법상의 차이는 크지 않다. 최적값을 계산하는 방법도 기계적인 Iterative Method를 사용한다.

이러한 전통적인 방법은 기계적인 계산방식으로 최적값을 구하는 것이므로, 어떠한 상황에서도 값은 도출된다. 그러나 이러한 방식들은 자료의 분포에 대한 고민이 반영되지 않았다. 사고발생률 또는 1사고당 손해액 등의 자료는 담보 종류에 따라 통계분포에 차이가 있다. 그러므로 통계분포를 반영하여 상대도 값을 계산할 경우 보다 정확하고 이상적인 상대위험도 계산이 가능하게 된다.

전통적인 요율산출방법을 사용하더라도 요율을 산출할 수 있으나, 전통적인 방법은 요율산출의 효율성, 통계적으로 유의한 변수를 확인하는 방법, 변수 간 상호관련성 분석 등에서 통계적인 방법보다 유용하지 못하다. 따라서 본 연구에서는 이러한 전통적 방식에서 고려하지 않은 통계분포, 즉 통계모형을 활용한 자동차보험 요율산출 방법을 제시하고자 하였다.

통계모형을 활용하는 방법으로는 최소자승법(OLS)을 사용하는 전통적 회귀분석이 있다. 그러나 최소자승법은 종속변수에 대한 가정이 실제자료와 일치하지 않는 경우가 자주 발생한다. 이러한 전통적 회귀분석 방법의 문제점을 극복하기 위해서 고안된 방법이 일반화선형모형이다. 자료와 분석모형의 적합성 때문에 현재 영국 등 유럽에서는 자동차보험 요율산출에 일반화

선형모형을 널리 활용하고 있고, 미국에서도 점차적으로 동 통계모형을 활용하는 추세이다. 그럼에도 불구하고 우리나라의 경우, 통계모형을 활용한 요율산출이 거의 이루어지지 않고 있다.

따라서 본 연구에서는 우리나라 자동차보험 요율산출 방법의 선진화를 유도하는 단초를 제공하기 위하여 일반화선형모형의 이론적 배경을 소개한다. 또한 자동차보험 실제 통계를 사용하여 자동차보험 요율산출(특히 요율상대도 산출)에 일반화선형모형을 활용하는 방법을 연구・제시하고자 한다.

2. 연구범위

본 연구의 범위는 통계적 모형, 특히 일반화선형모형(Generalized Linear Model)을 활용하여 보험료를 산출하는 방법을 찾는 것이다. 일반화 선형모형을 적용할 수 있는 분야는 다양한데, 특히 자동차보험 요율산출 방법을 그 대상으로 하였다. 자동차보험을 연구대상으로 한 이유는 자동차보험이 요율통계가 풍부하고, 요율체계가 잘 갖추어져 있으며, 타인에 대한 배상 및 자신에 대한 보상을 포함하는 종합적 성격을 가진 보험이기 때문이다.

연구는 다음의 단계별로 진행되었다. 우선 자동차보험의 요율특성을 소개하였다. 그리고 자동차보험 요율산출방법에 대한 선행연구를 살펴보았다. 요율산출방법에서는 요율산출방법의 일반이론 및 다변량 요율산출방법을 살펴보았다. 다변량 요율산출 방법에서는 기존의 기계적인 요율산출 방법과 통계적 모형을 활용한 요율 산출방법을 모두 살펴보았다. 마지막으로 실제 자동차보험 통계자료를 활용하여 일반화선형모형을 요율산출방법에 적용하여 보았다.

자동차보험 요율산출 실무업무에 대해 이해하고 있으면서, 계량경제학 또는 통계학에 대한 어느 정도 지식이 있는 독자가 이해할 수 있는 수준으로 본 연구보고서를 작성하였다. 일반화선형모형을 이해하기 위해서는 최소자승법(OLS)에 의한 전통적 회귀분석 방법을 알고 있어야 할뿐 아니라, 범주형 자료분석 및 분포에 대한 지식과 분포와 요율산출과의 관계를 이해할

수 있어야 한다. 따라서 본 연구보고서는 실제 자동차보험 요율산출 업무를 시행한 경험이 있거나, 자동차보험 요율산출 방법에 대하여 연구경험이 있는 사람에게 도움이 될 것으로 생각된다.

본 연구는 총 5장으로 구성되어 있다. 제1장은 서론 및 연구의 범위이다. 제2장은 자동차보험 요율산출의 전통적인 방법을 소개하였다. 제3장은 기존 자동차보험 요율산출방법의 문제점을 개선하기 위한 일반화선형모형의 선행연구를 소개하고 일반화선형모형에 대한 이론적 배경 및 적용방법을 구체적으로 살펴보았다. 제4장에서는 일반화선형모형을 자동차보험 통계에 적용하여 요율상대도를 산출하였다. 마지막으로 제5장에서는 본 연구의 내용을 정리 요약하고 한계점을 제시하였다.

Ⅱ. 자동차보험 요율산출 방법[1)]

1. 자동차보험 요율산출의 의미

가. 계리학적 의미[2)]

특정 집단에 해당하는 자동차 보험료를 산출한다는 것은 해당 집단의 사고빈도분포와 사고심도분포를 모두 감안한 평균 위험보험료를 산출한다는 것을 의미한다.

보험기간이 1년인 집단위험모델(collective model)을 가정해보자. 그리고 현재 자동차보험 요율체계가 가입경력, 할인할증, 성, 연령등과 같은 요율구분요소로 나누어지지 않은 단일요율체계라고 가정하자. 이 경우 자동차보험에서 지급되는 총 보험금은 다음과 같은 식(Ⅱ-1)로 나타낼 수 있다.

$$S = X_1 + ... + X_N \tag{Ⅱ-1}$$

이 식은 사고건수 변수 N과 사고건수별 지급된 보험금 X의 복합분포(compound distribution)식이다. 식(Ⅱ-1)에서 $X_{1,}X_{2,}......., X_N$이 사고확률변수인 N과 독립적인 iid확률변수라고 가정하고, 사고확률변수 N이 포아송 분포를 따른다고 가정해보자. 이 경우 식(Ⅱ-1)은 복합포아송 분포가 된다. 사고확률변수 N은 포아송 분포이외에 음이항 분포 등을 따를 수 있다.

이때 식(Ⅱ-1)의 기본 성질, 즉 평균과 분산 등은 다음과 같이 계산될 수 있다.

---

1) 본 장의 내용은 저자의 자동차보험 요율산출 경험을 바탕으로 서술된 것이다.
2) “Boland,P.J.(1993), 『Statistical and Probabilistic Method in Actuarial Science』, Chapman & Hall/CRC, 2007 및 Bjorn Sundt, 『An Introduction to Non-Life Insurance Mathematics』, Karlsruhe”의 내용을 참조하여 서술하였다.

$$
\begin{aligned}
E(S) &= E_N(E(S/N)) \\
&= \sum_{n=0}^{\infty} E(X_1 + \dots + X_n / N = n) P(N = n) \\
&= \sum_{n=0}^{\infty} [nE(X)] P(N = n) \\
&= E(X) \sum_{n=0}^{\infty} nP(N = n) \\
&= E(X)E(N) \qquad (\text{Ⅱ-2})
\end{aligned}
$$

식(Ⅱ-2)에서처럼 집단위험(collective risk)은 사고빈도($E(N)$)와 사고심도($E(X)$)의 함수로 구성된다는 것을 알 수 있다.

즉, 일반적으로 자동차보험 요율산출은 식(Ⅱ-2)를 이용하여 평균위험보험료를 산출하는 과정이다. 그런데 자동차보험의 위험보험료를 식(Ⅱ-2) 만으로 확정할 수 있는가에 대해서는 의문이 생긴다. 왜냐하면 평균 위험보험료 값은 동일하지만 위험집단들 간에는 분산에 차이가 발생하기 때문이다. 위험집단의 분산이 크다는 것은 위험도가 높다는 의미인데, 식(Ⅱ-2)로 산출한 평균값으로 보험료를 결정하는 것은 요율결정상 위험이 내재되는 것이다. 이러한 문제를 해결하기 위해서 다음의 식(Ⅱ-3)과 같이 분산을 산출하여, 요율산출에 적용한다.

$$
\begin{aligned}
Var(S) &= Var_N(E(S/N)) + E_N(V(S/N)) \\
&= Var_N(E(X) \cdot N) + E_N(N \cdot Var(X)) \\
&= E^2(X)\,Var(N) + Var(X)E(N) \qquad (\text{Ⅱ-3})
\end{aligned}
$$

식(Ⅱ-2)와 식(Ⅱ-3)으로 산출된 평균과 분산을 이용하여 자동차보험 요율은 다음 3가지 원칙중 하나로 산출된다.

$H(X)$를 비음의 순보험료라고 한다면,

첫째, ‘평균값의 원칙’에 따라 $H_1(S) = (1+a)E(S)$이고, 여기서 $a > 0$이다.

둘째, ‘표준분산원칙’에 따라 $H_2(S) = E(S) + b\sqrt{Var(S)}$이고, 여기서

$b > 0$이다.

셋째, '분산원칙'에 따라 $H_3(S) = E(S) + c\,Var(S)$이고, 여기서 $c > 0$이다.

첫 번째 평균값의 원칙은 위험분포는 단지 평균이라는 하나의 모수만 필요하다는 단순한 개념이다. 그래서 이 원칙은 동일한 평균을 갖는 집단은 동일한 보험료가 적용된다는 단점이 있다. 즉 앞서 언급한 위험, 즉 분산을 감안하지 않는 요율산출 원칙이다.

반면에 둘째와 셋째 요율산출원칙은 보험료에 위험이 반영되어진 것이며, 분산이 크면 위험도가 더 크다는 의미이므로 동일한 평균값이라 하더라도 그 위험도만큼 다른 보험료가 적용되어야 한다는 것을 의미한다.

현재 우리나라 자동차보험산업에서 사용되는 요율산출 방법은 순보험료 산출의 3가지 원칙 중 첫 번째인 '평균값의 원칙'이 적용되고 있다.

나. 실무적 의미

앞서 가.절의 계리학적 의미에서 살펴본 바와 같이 자동차보험요율산출은 특정위험집단별 사고빈도와 사고심도를 이용한 평균위험보험료를 산출하는 과정이다. 평균위험보험료를 산출하는 방법은 식(Ⅱ-1)에 따라 사고빈도 및 사고심도의 통계분포를 사용하여 수리적으로 계산할 수 도 있으나, 실제 자동차보험 통계자료는 이러한 분포로 설명하지 못하는 경우가 발생한다. 또한 분포를 사용하여 요율을 산출하는 방법은 너무 복잡하여 통계 모형을 이해하는 일부 사람만이 가능한 것이다. 따라서 이러한 문제를 해결하기 위하여 실무적으로 개발된 요율산출방법이 순보험료법, 손해율법이다. 이 방법은 실무적으로 사용이 용이할 뿐 아니라, 분포를 가정하지 않았지만 앞서 분포를 이용하여 유도된 요율산출방법과 동일한 비모수적 방법이다.

손해보험 실무에서 널리 활용되고 있는 요율산출방법은 판단법, 손해율법, 순보험료법 3가지가 있다. 이중 손해율법 및 순보험료법이 요율산출방법으로 가장 널리 이용된다. 손해율법의 요율산출결과와 순보험료의 요율산출결과는 동일[3]하므로, 요율을 산출・조정하는 상황에 부합되게 손해율법

3) Casualty Actuarial Society(2001),pp.90~91

과 순보험료법을 번갈아가면서 적용한다.

그리고 보험료 산출분야는 크게 '기본보험료를 산출하는 분야'와 '요율상대도(위험상대도)를 산출하는 분야'로 나눌 수 있다. 손해율법과 순보험료법 두 가지 요율산출분야에 모두 적용될 수 있는 방법이다. 그러므로 요율산출자는 모든 요율산출분야에서 두 가지 방법을 상황에 맞도록 변형하여 적용하고 있다.

손해율법과 순보험료법은 기본적으로 분모와 분자의 비율관계로 요율을 조정하는 방법이다. 두 방법은 분자는 동일한데 분모가 서로 다른 비율이다. 따라서 비율의 특성을 인식하고 산출하고자하는 요율의 종류에 따라 동 방법 중 하나를 선택하여 사용하면 된다. 즉, 자동차보험의 예를 보면, 두 방법의 분자는 손해액으로 동일하지만, 손해율법에서 분모는 경과보험료인데 반해 순보험료법에서는 평균유효대수이다. 경과보험료는 각 계약자의 특성에 맞게 적용된 보험료이다. 즉 계약자의 모든 특성이 보험료에 녹아있는 개념이다. 따라서 손해율법에서는 계약자의 특성이 포함된 손해액의 특성에 부합되고, 총경과보험료에서 손해액이 차지하는 비율이 일정수준을 유지되도록 경과보험료를 조정함으로써 요율 조정요인을 산출한다. 반면에 순보험료법에서 분모는 평균유효대수이다. 평균유효대수는 각 자동차의 대수를 의미하므로, 이 자료에는 계약자의 위험특성이 없다. 따라서 순보험료법으로 요율을 산출하는 방법은 계약자의 특성별로 만들어진 손해액을 해당 위험특성에 포함된 자동차대수, 즉 평균유효대수로 나누어 계약자 특성별로 위

◦ 손해율 방법 : $R=\dfrac{L(1+G)}{E(1-V-Q)}$

- $P=\dfrac{L}{E}$=> $L=EP$이고, $G=\dfrac{EF}{L}=\dfrac{F}{P}$ 이다.
- 여기서, R은 산출하고자 하는 요율, L은 경험손해액, E는 경험기간의 노출위험단위, G는 보험료에 연동되지 않는 손해액대비 비용 비율을 말한다.

◦ 순보험료 방법 : $R=\dfrac{EP[1+(F/P)]}{E(1-V-Q)}$ 또는 $R=\dfrac{P+F}{1-V-Q}$가 된다.

- 순보험료 방법의 식은 P와 G를 손해율 방법의 식에 대입하여 유도한 것이다.
- 여기서, R은 위험단위당 요율, P는 순보험료, F는 위험단위당 고정사업비, V는 변동사업비율, Q는 이익률이다.

험도를 산출하는 것이다.

이러한 손해율법과 순보험료법의 특성으로 인하여 손해율법은 기존보험료가 주어진 상황에서 기본보험료 등을 어느 정도 조정할 것인가를 판단하는데 사용된다. 순보험료법은 기존보험료를 새로 산출하거나 전혀 새로운 위험에 대한 보험료를 산출하는데 사용된다.

손해율법과 순보험료법의 실무적 산출계산식을 보면 다음과 같다. 손해율법은 실제손해율과 예정손해율을 비교하여 차이가 있는 경우 그 차이만큼 요율을 조정하는 방법인데, 이때 손해율을 결정할 때 사용되는 변수는 손해액과 경과보험료 이외에 예정사업비율, 대리점수수료율 및 이익률 등이다. 이러한 요인들을 어떤 조합으로 반영할 것인가에 따라 손해율법의 요율조정요인이 변한다.

이 방법은 새로운 보험요율을 산출하는 방법이 아니며, 기존 보험료를 현재의 실적손해율에 맞도록 조정하는 방법이다. 손해율법에 의한 요율조정요인 산식은 다음 3가지 방식이 있다.

$$\text{i) 조정요인} = \frac{\text{실적손해율}}{\text{예정손해율}} - 1 = \frac{\text{실적손해율}}{1 - \text{예정사업비율} - \text{예정이익률}} - 1 \qquad (\text{II-4})$$

예정손해율에 비하여 실적손해율이 변화된 정도를 조정요인으로 요율을 조정하는 방법이며, 사업비를 적용보험료의 일정비율로 항상 유지시키는 방법이다. 즉 사업비율이 30%라면 적용보험료의 30%가 사업비로 항상 사용되는 방법이다.

$$\begin{aligned}\text{ii) 조정요인} &= \frac{\text{실적손해율} + \text{실적사업비율}}{\text{예정손해율} + \text{예정사업비율}} - 1 \\ &= \frac{\text{실적손해율} + \text{실적사업비율}}{1 - \text{예정이익률}} - 1 \qquad (\text{II-5})\end{aligned}$$

실적손해율과 실적사업비율을 예정손해율과 예정사업비율로 나눈 값으로 실적사업비규모를 반영하는 계산식으로, 이익률은 영업보험료의 일정 비율

로 유지된다.

$$\text{iii) 조정요인} = \frac{\text{실적손해율} + \text{고정사업비율}}{1 - \text{변동사업비율} - \text{이익율}} - 1 \qquad (\text{Ⅱ-6})$$

조정인 iii)방식은 사업비율 중 변동사업비(대리점수수료율과 같이 영업보험료로서 일정비율로 나타나는 사업비)와 이익률은 영업보험료의 일정비율로 유지하고 고정사업비는 실제금액을 반영하는 산식이다.

손해율법이 기존 보험요율을 현재 손해율 수준으로 조정하는 방법인 반면 순보험료법은 현재의 위험도에 맞는 요율을 새로 산출하는 방법이다. 순보험료는 아래의 산식에서와 같이 전체손해액을 위험단위로 나누어 산출하며, 자동차보험에서 위험단위는 일반적으로 평균유효대수를 사용하고 있다.

$$\text{순보험료} = \frac{\text{손해액}}{\text{위험단위수}} \qquad (\text{Ⅱ7})$$

순보험료법을 분해해 보면, 사고빈도(사고건수÷평균유효대수)와 사고심도(손해액÷사고건수)로 나누어진다. 이 방식은 앞서 식(Ⅱ-2)에서 유도한 평균보험료 산출식과 동일한 방법이다.

2. 자동차보험 요율산출 방법

가. 요율산출절차

실무적인 측면에서 보면 자동차보험에서 요율산출은 1) 통계의 집적 및 추출, 2) 기본보험료 산출, 2) 요율상대도 산출, 4) 보상한도 확대 등 제도를 반영하는 절차에 따라 이루어진다.

요율산출 절차 중에서 기본보험료 산출은 보험종목별, 담보별, 차종별로 대당보험료를 산출 또는 적용하는 과정이다. 새로운 위험요인의 보험료를

산출하는 것이라면 순보험료법을 적용하지만, 기존의 위험요인의 기본보험료를 조정하는 것이라면 손해율법 및 순보험료법을 모두 적용할 수 있다. 앞의 ‘손해율법과 순보험료법’에서 설명한 요율조정요인 산출 계산식에 따라 보험종목별, 담보별, 차종별로 요율조정요인을 산출하고, 산출된 조정요인을 기존의 기본보험료에 적용하는 방법으로 기본보험료를 조정한다. 기본보험료 조정은 ‘첫째, 통계자료의 추출 및 정리’, ‘둘째, 기초자료에 손해액진전계수(LDF)반영’, ‘셋째, 기초자료에 각종 추세(사고발생률, 1사고당 손해액, 가입경력요율, 할인할증률, 기타 특약적용률)를 반영’, ‘넷째, 지급기준 변경 등 제도변화 반영’, ‘다섯째, 요율조정요인 산출’의 단계를 거쳐 이루어진다.

요율상대도 산출은 기본보험료가 조정이 된 이후, 위험집단별로 위험을 세분화하는 과정이다. 위험의 세분화는 위험도에 영향을 주는 각종 요인인 성별(현재 미적용), 연령, 지역(현재 미적용), 가입경력, 할인할증률 등의 기준에 따라 이루어진다. 이들 요율상대도 산출에 적용되는 변수이외에도 기본보험료 산출에 적용되는 변수 중에서 차종, 담보 등을 요율상대도 산출의 변수로 포함시킬 수 있다. 요율상대도 산출에 사용되는 변수의 범위는 요율산출자(actuary)가 임의로 정하여 사용할 수 있다. 그러나 기본보험료산출의 변수범위를 보험종목, 담보, 차종 수준에서 산출하는 것이 현재 우리나라의 요율산출 관례이다. 위험세분화 방법은 단변량방법과 다변량방법이 있다. 단변량방법은 예를 들면 성과 같이 단일변수별 순보험료 및 손해율로 요율상대도를 산출하는 것이다. 이 방법은 단일 변수만으로 위험도를 산출하는 것이므로 변수들 사이의 상호작용 효과를 측정할 수 없다는 단점이 있다. 이러한 단변량 방법의 문제점을 극복하기 위하여 다변량 방법의 요율상대도 산출방법이 개발되어 왔다. 전통적으로 사용되고 있는 다변량 방법으로는 Bailey and Simon Method, Bailey’s Minimum Bias Method, Least Square Method 등이 있다. 이들 방법은 단변량 방법보다 더 정교한 요율상대도를 산출할 수 있지만, 다소 기계적인 방법이다. 이러한 방법을 보완하기 위하여 개발된 것이 앞으로 소개할 통계적 모형(회귀분석 또는 일반화선형모형 등)을 활용하는 방법이다.

요율산출방법의 마지막 단계는 보상한도 확대 등 제도개선 내용을 요율에 반영하는 것이다. 보상한도 확대는 대인배상Ⅰ의 보상한도(현재는 사망기준으로 1억으로 되어 있음)를 법률에 따라 기존 보다 더 높이는 것을 의미한다. 대인배상Ⅰ의 보상한도를 기존보다 높인다는 것은 대인배상Ⅰ의 보상한도를 높이는 수준만큼 기존 대인배상Ⅰ의 보험료를 인상한다는 의미이고, 또 이에 비례하여 대인배상Ⅱ의 보험료를 인하한다는 것을 의미한다. 보상한도확대를 보험료에 반영하는 방법으로는 통계분포(로그노말분포, 감마분포, 파레토분포 등)를 사용하여 보상한도확대 효과를 추정하는 방법이 있고, 한도타절이라는 비모수적이면서 실무적인 방법이 있다.

요율조정에 최종적으로 반영해야 하는 기타 제도 변경으로는 지급기준[4]의 항목 변경에 따른 효과를 기본보험료를 조정한 이후에 반영하는 것이다. 이와 더불어 앞 단계의 요율상대도 산출로 전체보험료 수준의 변화가 발생할 경우, 변화가 되기 이전단계로 환산하는 작업을 해야 한다. 이 작업을 요율산출에서는 Off-Balance조정이라고 한다. Off-Balance 조정효과도 기본보험료가 산출된 이후 최종적인 단계에 반영된다.

이상의 요율산출 단계에서 일반화선형모형(GLM)을 적용할 수 있는 분야는 주로 요율상대도 산출 단계이다. 이외에도 일반화선형모형(GLM)을 적용할 수 있는 요율산출 분야는 손해액 진전계수(Loss Development Factor)를 구하는 곳이다. 손해액 진전계수(LDF)는 기본보험료 산출에서 정확한 지급준비금을 추정하는데 사용된다. 손해액 진전계수(LDF)산출에서 독립변수는 사고발생기간과 진전기간으로 하고 종속변수를 지급보험금으로 하여 일반화선형모형을 적용할 수 있다(wright, 1993). 이외에도 일반화 선형모형에서 요율상대도 값이 아닌 위험세분 집단별 사고발생률 추정값과 1사고당손해액 추정값을 산출할 수 있다. 위험세분 집단별 사고발생률 추정값과 1사고당 손해액 추정값을 곱하여 위험세분집단별 위험보험료를 산출할 수 있다.

---

4) 일부 지급기준 항목은 기본보험료 조정단계에 반영되기도 하고, 일부는 기본보험료를 조정한 이후 최종적으로 반영되기도 한다. 그러나 지급기준 변경은 향후 자동차사고 시 반드시 지급되어야 하는 보험금 변경항목이므로 의미적으로 볼 때 기본보험료를 조정한 이후 최종적으로 반영하는 것이 요율산출의 의미에 더 적합한 것으로 판단된다.

이처럼 일반화 선형모형은 자동차보험의 여러 분야에 다양하게 적용될 수 있다. 본 연구보고서에서는 일반화선형모형을 적용할 수 있는 자동차보험 여러 요율산출 분야 중에서 가장 중요한 분야인 요율변수들의 위험상대도 산출 분야로 연구대상을 한정하였다. 일반화 선형모형을 적용할 수 있는 모든 분야를 대상으로 연구를 범위를 확대하면 연구의 범위가 너무 넓어져 연구내용이 중복되는 결과가 발생할 수 있기 때문이다. 또한 요율상대도 분야로 연구범위를 한정하더라도, 동 방법을 충분히 이해할 경우 다른 분야로 연구결과를 쉽게 확장할 수 있을 수 있을 것으로 생각된다.

나. 통계의 집적 및 추출

1) 통계수집방법

자동차보험 요율산출에서 가장 중요한 것 중의 하나는 좋은 자료를 확보하는 것이다. 즉 쓸모없는 자료를 이용하면 아무리 정교한 수리모델을 이용하더라도 그 결과는 의미가 없기 때문이다.

자동차보험 요율산출에 사용되는 자동차보험의 자료는 크게 계약자료와 사고자료로 나눌 수 있다. 계약자료에는 현행 자동차보험 요율체계에 따른 계약자의 정보가 기록되어 있다. 즉 피보험자의 주민등록번호, 성별, 차종, 특약의 가입여부 등 가능한 모든 계약정보가 등록되어 있다. 사고자료에는 피해자, 피해물건, 지급보험금 및 지급준비금 등에 대한 정보가 기록되어 있다.

자동차보험 자료의 대부분은 계약체결 시 또는 사고 시 결정된 것이지만 지급준비금과 같이 향후에 결정되는 것도 있다. 피보험자 또는 보험가입자의 위험도 평가에서 중요한 종속변수인 손해액에 영향을 미치는 지급보험금 및 지급준비금은 담보종목에 따라 사고당시 결정되는 것과 시간이 경과함에 따라 점진적으로 결정되는 것이 있다. 그러므로 담보별 속성에 대한 충분한 이해가 필요하다.

자동차보험 통계자료는 계약자료와 사고자료의 통계기간의 일치여부, 사고

및 계약의 발생기간 등에 따라 사고년도기준 자료(Accident Year Data), 보험년도기준 자료(Policy Year Data) 및 역년 기준자료로 구분된다. 이중 자동차보험 요율 산출 및 분석에 이용되는 자료는 사고년도기준 자료(Accident Year Data), 보험년도기준 자료(Policy Year Data)이며 자료의 집적을 위하여 널리 사용되는 자료는 역년 기준 자료(Calendar Year Data)이다.

사고년도기준 자료는 통계기간을 1년 즉, 12개월이라고 하면, 동 기간 안에 발생한 계약통계(보험료)와 이 기간에 발생한 사고건수 및 해당 사고로 지급된 보험금 통계이다. 다시 말하면 사고년도기준 자료의 집적방법은 특정한 기간에 발생한 사고와 그 사고로 인한 손해액 및 보험료를 결정하는 방법이다. 이 사고년도 기준자료(Accident Year Data)는 손해액과 보험료를 일치시키는데 적합한 방법이다. 즉, 사고년도기준 자료(Accident Year Data)는 사고년도에 수입된 보험료로 그 기간 동안 발생한 사고에 대하여 보험금을 지급하는 것으로 계약과 사고의 위험노출을 정확히 일치시키는 방법으로 집적된 것이기 때문이다. 이때 사고년도(Accident Year)의 손해액을 경과보험료로 나눈 값을 Accident Year Loss Ratio라고 한다.

반면에 보험년도 기준자료(Policy Year Data)는 통계기간 1년 즉, 12개월 단위기간동안 체결된 보험계약의 보험료와 그 보험계약에 따라 지급된 보험금 및 위험단위가 기록된 것이다. 예를 들면 2008년 보험년도 기준자료(Policy Year Data)는 2008년 1월 1일부터 2008년 12월 31일 사이에 유효한 보험계약의 손해액, 보험료 및 평균유효대수 등이다. 이 방식은 특정기간동안에 체결된 보험계약에 주안점을 두어 이들 계약의 손해액과 보험료에 대한 자료를 수집하는 것이다. 즉, 해당 통계기간에 계약을 체결하고, 그 체결된 계약으로부터 발생한 사고자료를 수집하는 것이 보험년도 기준(Policy Year Basis) 자료수집 방법이다. 그러므로 이 방식으로 모든 통계 결과를 집적하기 위해서는 보험년도가 통계기간을 1년이라고 가정하면 총 24개월의 보험기간이 필요하다. 이 방식은 계약 자료와 손해액 자료가 정확히 일치한다는 장점은 있으나, 통계기간을 1년으로 가정하면 과거 2년 전부터 1년 전까지 체결된 과거 계약의 통계라는 점이 단점이다.

역년기준자료(Calendar Year Data)는 회계적 기준에 의한 자료수집방법

이다. 보험계약의 유효기간 또는 사고 날짜에 관계없이 통계기간을 1년이라고 가정하면, 통계기간인 12개월 동안 발생한 보험료와 손해액 자료를 집적하는 방법이다. 그러므로 역년기준(Calendar Year)보험료는 역년기준 동안 수입보험료에서 미실현지급보험금의 변동액을 제외한 금액이다. 이 역년기준자료(Calendar Year Data)는 과거에 발생한 사고인데 보험금지급이 지연되어 집적하고자 하는 통계기간 중에 보험금이 지급된 경우의 자료도 포함된다. 반면에 보험료통계는 통계기간 중 실현된 통계가 된다. 따라서 동 통계수집방법은 보험료와 손해액이 일치하지 않는다. 이 방법은 회계적 관점에서 보험료규모와 보험금 수준을 평가하는데 유용한 것이지만, '보험료와 손해액이 일치되어야 정확한 보험료를 산출할 수 있다'는 요율산출자의 입장에서는 적합한 통계가 아니다.

이러한 통계자료의 특성 때문에, 요율산출자 입장에서는 Calendar Year 자료 보다는 사고년도기준 자료(Accident Year Data) 또는 보험년도기준 자료(Policy Year Data)자료를 사용한다.

2) 자동차보험 통계집적

보험회사에 집적된 자동차보험 통계는 요율요소 구분기준에 따라 집적되어 있다. 즉 보험종목은 개인용(플러스 개인용 포함), 업무용(플러스 업무용 포함), 영업용으로 구분된다. 담보는 각 보험종목별로 대인배상 I 및 대인배상 II, 대물배상, 자기신체사고, 무보험차상해, 자기차량손해로 구분된다. 차종은 '자동차관리법에 의한 자동차', '군수품관리법에 의한 차량', '원동기장치자전거', '건설기계관리법의 적용을 받는 건설기계', '농업기계화촉진법에 의한 농업기계'로 구분된다. 이러한 큰 분류기준에 따라 계약관련 통계 및 사고관련 통계가 집적된다.

계약관련 통계에는 자동차보험계약 가입 시 적용되는 요소인 차종, 피보험자, 가입경력요율, 할인할증적용률[5] 각종 특약 및 특별요율의 부가여부

---

5) 2007년 9월1일 부터는 할인할증 적용률 대신 할인할증 등급기준으로 자료가 관리되고 있다.

등에 관한 분류기준 및 평균유효대수, 수입보험료, 경과보험료 등의 자료가 포함된다. 사고관련 통계에는 사고건수, 지급보험금, 지급준비금, 사망 및 부상여부, 전손 및 분손 여부 등의 자료가 포함된다.

3) 분석통계 추출

분석통계는 앞서 살펴본 통계특성을 이해한 이후 분석하고자 하는 목적에 맞게 통계를 추출하여야 한다. 예를 들면 기본보험료를 산출하는 경우에는 보험종목, 담보별, 차종별로 계약자료와 손해액 자료를 추출한다. 통계추출의 기준은 Calendar Accident Year Basis이다. 요율상대도 산출은 보험종목별, 담보별, 차종별 계약자료와 손해자료를 추출하되, 계약자료에는 각종 계약조건이 포함된다. 즉 성, 연령, 할인할증, 가입경력, 운전자 한정특약에 가입여부 등이 포함된다. 통계추출 기준은 Calendar Accident Year Basis나 보험년도기준 자료(Policy Year Data) 두 가지 모두 가능하다. 제도개선을 반영하는 경우에는 제도개선의 효과가 반영되도록 통계를 추출하여야 한다. 보상한도확대를 예로 들면, 일정 통계기간 중에 발생한 사고자료 중 완전히 종결된 자료를 대상으로 통계를 추출한다. 추출 기준은 지급된 보험금기준으로 일정금액 이하 및 초과 보험금과 해당 사고건수 자료를 추출한다.

다. 기본보험료 산출

‘요율산출에 부합된 통계 추출’, ‘통계를 현행 수준으로 조정(On- Level)’, ‘손해액 진전계수(Loss Development Factor) 반영’, ‘사고빈도 및 심도 추이 산출·반영’, ‘손해율법 또는 순보험료법으로 요율조정요인 산출 또는 보험료 산출’의 과정을 밟아 기본보험료가 산출(또는 조정)된다.

기본보험료 산출의 첫째 단계는 ‘요율산출에 부합된 통계추출’이다. 기본보험료산출에 사용되는 통계는 사고년도기준 자료(Accident Year Data)이다. 가장 최근년도가 포함된 1년간의 통계기간으로 사고년도(Accident Year)기준으로 통계를 추출한다. 그런데 자동차보험은 향후 특정한 시점부터 1년간 적용되는데, 추출된 통계는 과거 1년 또는 2내지 3년의 통계이다.

즉 추출통계와 적용기간이 일치하지 않게 된다. 따라서 추출된 통계를 향후 적용되어야할 보험기간의 제도에 부합하도록 통계를 조정해주는 작업(On-Level)을 해야 한다. 통계를 적정수준으로 반영하도로 조정해주는 작업은 두 가지 방향에서 이루어진다. 하나는 보험요율제도 및 지급기준과 같은 보험금제도의 변화를 반영하는 것이다. 두 번째는 정확한 지급준비금을 추정하여 반영하는 것이다.

기본보험료 산출의 두 번째 단계는 '통계를 현행 수준으로 조정'하는 것이다. '보험요율제도 및 지급기준을 반영하는 방법'은 보험기간과 보험금지급흐름, 보험기간과 보험료수입흐름을 각 변으로 하는 사다리꼴 도형(Parallelogram Method)을 이용하여 수리적으로 계산하는 방법이 많이 활용된다. 이것은 과거 제도변화가 있었던 시점과 제도변화로 인한 효과가 사다리꼴 자료도형의 어느 부분에 영향을 주었는지 파악하고, 해당 영향정도가 향후 요율적용기간에 어떻게 영향을 줄 것인지 면적개념으로 계산하는 방법이다. 동 방법은 사다리꼴 면적을 이용하여 근사적으로 계산하는 방법이므로, 정확성은 다소 떨어지는 방법이다. 오늘날에는 컴퓨터가 발달하여 전산으로 개별 계약 및 사고건에 대하여 제도변화요인을 반영하는 보다 정확한 방법이 일부 활용되고 있다.

기본보험료산출의 세 번째 단계는 '손해액 진전계수(LDF)'의 산출(또는 반영)이다. 보험료를 산출할 때 손해액 추정에 필요한 손해액 진전계수를 산출하기 위해서는 지급준비금 추정이 필요하다. 이 지급준비금은 향후 지급될 보험금을 미리 적립한 것이다. 지급준비금에는 개별추산 지급준비금과 통계적 추산 지급준비금 두 가지가 있다. 개별추산 지급준비금은 손해사정사가 개별적 속성을 파악하여 적립하는 것이므로, 회사의 정책적 판단이 개입되며 사고특성에 적합한 준비금이 적립되지 않을 수 있다. 이러한 개별추산 지급준비금 적립방식을 보완하기 위하여 개발된 것이 통계적 추산 방법의 지급준비금 적립방식이다. 현재까지 통계적 추산 방식 지급준비금 방식으로는 사다리꼴방법(Chain Ladder Method, Average Payment Method 등)이 많이 사용되고 있다. 사다리꼴방법은 사고발생년도와 보험금이 지급된 진전기간별로 지급된 보험금 변화의 흐름을 파악하는 방법이다. 지급준

비금을 추정하는 사다리꼴 방법은 손해보험 분야에서 준비금을 파악하는데 널리 활용되고 있는 방법이다.

기본보험료 산출의 네 번째 단계는 '사고빈도 및 사고심도추이를 산출 및 적용'하는 것이다. 향후 자동차보험이 적용될 보험기간의 제도수준으로 통계가 조정이 된 이후, 기본보험료 산출은 미래의 손해액을 추정하기 위해 추세를 감안하여야 한다. 추세를 반영하는 방법으로는 사고빈도 및 사고심도의 추세를 고려하여 향후 적용될 보험기간 동안 사고빈도 및 사고심도가 어떤 수준이 될지를 판단하기 위해 단순 회귀식이 많이 사용된다. 이러한 작업은 보험종목, 담보 및 차종별로 이루어진다. 과거 추출된 통계를 미래 적용될 통계수준으로 조정(On-Level)하고, 사고빈도 및 사고심도 등 보험금의 수준에 영향을 주는 추세요인을 반영하면 기본보험료를 산출할 바탕이 마련된 것이다.

기본보험료산출의 마지막 단계는 '순보험료법 또는 손해율법으로 요율조정요인을 산출하거나 순보험료를 산출'하는 것이다. 자동차보험에서는 차종별 기본보험료에 각종 요율요소를 곱하는 방법으로 요율이 적용된다. 이러한 요율적용의 행태로 인해 자동차보험 요율산출의 기준은 기본보험료 산출이다. 기본보험료는 손해율법 또는 순보험료법으로 보험종목별, 담보별, 차종별로 산출되고 있다.

라. 요율상대도 산출

현재 자동차보험에서는 기명피보험자 연령그룹, 가입경력, 할인할증요율, 교통법규위반 경력요율, 24세운전자연령특약과 같은 운전자 연령한정특약 및 가족운전한정특약과 같은 운전자 한정특약 등을 포함하여 요율상대도 산출방법으로 보험료를 산출하고 있다.

자동차보험에서 이들 요율변수별 요율상대도는 주로 곱셈의 원칙[6]에 따

6) 현재 우리나라 자동차보험제도에서는 대부분의 요율변수가 곱(×)의 법칙으로 적용되지만, 가입자특성요율인 보험가입경력요율과 교통법규위반 경력요율의 경우와 우량할인·불량할증요율과 특별할증은'합(+)'의 법칙에 따라 산출·적용되고 있다.'Ⅳ. 일반화선형모형을 이용한 실증분석의 1. 모형적용기준'에서 요율변수

라 산출되고 실제 적용되고 있다. 즉 모든 요율요소를 곱하면 세부위험집단별 위험도에 맞는 보험료가 결정되는 구조이다. 이러한 곱셈의 원칙은 아래에서 설명한 자료 분류와 이들 자료를 설명하는 방식에 대한 철학에 근거하여 만들어진 것이다. 즉 우리나라 자동차보험은 위험집단들이 상호 작용(교차분류)하는 부분의 종속변수(사고빈도 및 사고심도)를 위험집단의 분류기준의 곱으로 설명할 수 있다는 철학에 따라 요율제도가 만들어진 것이다. 이러한 우리나라 자동차보험 요율체계를 볼 때 아래에서 소개되는 요율상대도 산출방식 중 우리나라의 요율체계에 부합하는 모형은 '승산모형'이다.

승산모형으로 우리나라 요율체계를 분리하여 보면, 요율상대도 산출방법이 제한적으로 적용된다는 것을 알 수 있다. 즉 연령그룹 및 가입경력과 같이 일부 요율요소에 제한적으로 상대도 산출방법이 적용되고 있는 것이다. 이처럼 제한적으로 요율상대도 산출 방법을 적용하고 있는 이유는 '과거 요율산출의 관성', '모든 변수를 포함하여 요율상대도를 산출할 수 있는 컴퓨터의 물리적 한계' 때문이다. 이러한 한계를 극복한다면, 현재 우리나라 요율체계에서 대부분의 요인을 단일 통계모형에 포함시켜 요율상대도를 산출할 수 있다. 즉 보험종목, 담보, 차종, 연령, 가입경력, 할인할증, 운전자한정특약 등의 대부분의 요율요소를 모두 모형에 포함시켜 요율상대도를 산출할 수 있다. 일반화선형모형(GLM)으로는 대부분의 변수를 하나의 모형에 포함시켜 요율상대도를 산출할 수 있으므로, 과거 요율상대도산출방법의 한계를 극복할 수 있는 통계모형이라고 말할 수 있다.

본 장에서는 과거 전통적으로 사용되었던 요율상대도의 개념과 방법들에 대하여 설명하고, 일반화선형모형을 활용한 요율산출방법은 다음 장에서 보다 구체적으로 설명하도록 하겠다.

1) 요율상대도 자료형태

변수의 선택과 위험집단 분류에 따라 자료를 분류하면 아래 <표Ⅱ-1>과 같은 교차분류자료(Cross-classified Data)의 형태가 된다. 예시는 변수가 2

적용방법을 보다 세부적으로 다룰 것이다.

개인 경우의 교차분류자료이며, 변수의 수가 늘어나면 교차분류자료의 형태도 2차, 3차…의 형태가 된다.

교차분류자료에서 A와 B는 앞의 위험변수 선택방법에 의해 선택된 변수이며, 분류 I와 J는 위험집단분류방법에 의하여 분류한 위험집단의 개수이다. 그리고 셀 (i, j)은 두 변수 A와 B의 위험집단 i, j에 의하여 결정된 보험금관련 자료이며, 이 자료는 승산모형과 가산모형으로 각각 표현될 수 있다.

<표 II-1> 요율상대도 자료형태

B \ A	1	…	j	…	J
1					
:					
i			(i, j)		
:					
I					

자료 : Casualty Actuarial Society(1990), pp.89~113

위험도의 산출은 보험금관련자료 (i, j)를 설명할 수 있도록 각 변수 A, B의 각 집단 I 및 J의 상대적 위험도를 산출하는 것이다. 이때 변수의 개수에 따라 단순(One-Way)위험상대도, 다중(Multiple)위험상대도로 구분할 수 있다.

단순(One-way)위험상대도의 계산은 위험구분자료(손해율 또는 위험보험료)를 이용하여 해당 위험집단의 위험계수를 전체대비 상대값으로 산출하는 것으로 산출방법이 용이하나, 보험금관련 자료가 단순히 1개 변수의 작용으로 나온 것이므로 설명력이 떨어지는 단점을 가지고 있다. 따라서 실제 위험상대도를 계산할 때에는 다중(Multiple)위험상대도 계산법을 주로 이용하

고 있다.

위험계수의 공정한 배분을 위하여 이용되는 상대도 분석은 단순(One-Way)위험상대도 계산과 다중(Multiple)위험상대도 계산으로 구분될 수 있다. 단순위험상대도 계산은 독립변수들의 선형결합으로 또는 곱셈으로 종속변수를 설명하는 것이 아니라, 각각의 독립변수별로 위험도를 계산하는 방법이다. 각 변수별로 위험도를 계산하기 때문에 이러한 방법은 '단변량 위험상대도 산출'이라고도 한다. 단순(One-Way)위험상대도 계산은 분석절차가 간편하나 결과가 크게 왜곡될 수 있어서, 미국 ISO의 경우 등 대부분의 나라에서는 다중(Multiple)위험상대도 계산방법을 사용하고 있다.

이에 반하여 다중위험상대도 산출방법은 종속변수(사고발생률 또는 1사고당 손해액)를 독립변수의 선형결합으로 설명하는 것, 즉 여러 독립변수를 동시에 고려하여 위험상대도를 산출하는 방법이다. 여러 독립변수를 동시에 고려하기 때문에 독립변수 간 상호작용효과 등을 모두 파악할 수 있는 장점이 있다. 이러한 다중위험상대도 산출방법으로는 전통적인 방법과 최근에 자동차보험 요율산출에 사용되고 있는 통계적 방법이 있다. 전통적인 다중위험상대도 산출방법으로는 Bailey and Simon Method, Least Square Method, Marginal Total, Maximum Likelihood Method 등이 있다. 이러한 방법들은 독립변수를 다중으로 고려하여 요율을 산출한다는 점에 있어서 단순위험상대도 산출방법보다 진일보된 방법이다. 그러나 이러한 방법은 독립변수와 종속변수의 관계를 계산식에 따라 산술적, 기계적으로 위험상대도를 산출하는 방법이라는 한계점이 있다. 즉 각 변수의 통계적 유의성이 감안되어 있지 않다. 예를 들면 전통적인 방법으로 가입경력이 1년차인 그룹의 위험도가 1.3으로 산출되었다고 하자. 그런데 1년차인 그룹의 자동차대수가 통계적으로 유의하지 않은 수준이라면, 이 집단의 위험도를 1.3이라고 인정하기 어렵다. 그런데 전통적인 방법으로는 기계적으로 해당 그룹의 위험도가 계산되기 때문에 계산된 값을 그대로 인정할 수밖에 없다. 이러한 전통적 방법의 단점에도 불구하고, 현재 우리나라 자동차보험의 요율산출에서는 전통적인 방법이 주로 사용되고 있다.

2) 전통적인 요율상대도 산출방법

전통적인 다중요율상대도 산출방법은 ‘최소 편의 방법(Minimum Bias Procedure)’이라고 한다. 이 방법은 1960년대에 개발되었으며, 현재까지 지속적으로 개선·연구되어 통계학 분야의 한 부분으로 발전하였다. 그런데 최소편의 방법 중 많은 것은 일반화선형모형(Generalized Linear Model)의 특별한 형태에 지나지 않는다.

일반화선형모형(GLM)은 다음 장에서 상세히 설명하도록 하고 본 장에서는 다중 요율상대도 모형의 개념을 이해하기 위해 최소편의 방법 중 널리 알려져 있는 'Bailey and Simon Method', 'Bailey' Minimum Bias Method', 'Least Square Method'등 일부만 소개하였다(Casualty Actuarial Society(1990)).

<Bailey and Simon Method>

전체위험도와 각 변수별 위험도의 총 값은 유사할 것이라는 가정 하에서 전체위험도와 모형에 의한 변수별 위험도 총 값의 편차(분산)를 최소화하도록 각 변수별 위험상대도를 계산하는 방법이다.

Bailey and Simon Method도 보험금관련 변수값이 어떻게 구성되어 있는지에 대한 가정에 따라 승산모형과 가산모형으로 구분된다. Bailey and Simon Method의 기본 가정을 3개 변수 x, y, z의 모형으로 표현하고, 상대도값 계산식을 보면 다음과 같다.

승산모형의 경우 :

$$\text{minimize} \quad Q = \sum_{i,j,k} \frac{n_{ijk}}{x_i y_j z_k} (\gamma_{ijk} - x_i y_j z_k)^2 \qquad (\text{Ⅱ-8})$$

가산모형의 경우 :

$$\text{minimize} \quad Q = \sum_{i,j,k} \frac{n_{ijk}}{x_i + y_j + z_k} (\gamma_{ijk} - (x_i + y_j + z_k))^2 \qquad (\text{Ⅱ-9})$$

즉, 실제 위험값(보험금관련 자료)과 이론적 가치값의 차이가 최소화되도록 하기 위한 이론적 가치들을 찾기 위하여 위의 통계모형을 각 변수 x, y, z에 대하여 미분을 하면, 다음과 같은 이론적 요율상대도 계산식이 산출된다.

승산모형의 경우 :

$$x_i = \sqrt{\frac{\sum_{jk} \frac{n_{ijk}\gamma_{ijk}^2}{y_j z_k}}{\sum_{jk} n_{ijk} y_j z_k}}, \quad y_j = \sqrt{\frac{\sum_{ik} \frac{n_{ijk}\gamma_{ijk}^2}{x_i z_k}}{\sum_{ik} n_{ijk} x_i z_k}}, \quad z_k = \sqrt{\frac{\sum_{ij} \frac{n_{ijk}\gamma_{ijk}^2}{x_i y_j}}{\sum_{ij} n_{ijk} x_i y_j}} \qquad (\text{Ⅱ-10})$$

가산모형의 경우 : 모형의 미분을 통하여 산출한 값을 Newton- Raphson 의 Iterative Process과정을 통하여 산출한다.

$$x_{i+[1]} = x_{i+[0]} + \frac{\sum_{jk} n_{ijk} (\frac{\gamma_{ijk}}{x_i + y_j + z_k})^2 - \sum_{jk} n_{ijk}}{2\sum_{jk} n_{ijk} (\frac{\gamma_{ijk}}{x_i + y_j + z_k})^2 (\frac{1}{x_i + y_j + z_k})} \qquad (\text{Ⅱ-11})$$

$$y_{i+[1]} = y_{i+[0]} + \frac{\sum_{jk} n_{ijk} (\frac{\gamma_{ijk}}{x_i + y_j + z_k})^2 - \sum_{jk} n_{ijk}}{2\sum_{jk} n_{ijk} (\frac{\gamma_{ijk}}{x_i + y_j + z_k})^2 (\frac{1}{x_i + y_j + z_k})} \qquad (\text{Ⅱ-12})$$

$$z_{i+[1]} = z_{i+[0]} + \frac{\sum_{jk} n_{ijk} (\frac{\gamma_{ijk}}{x_i + y_j + z_k})^2 - \sum_{jk} n_{ijk}}{2\sum_{jk} n_{ijk} (\frac{\gamma_{ijk}}{x_i + y_j + z_k})^2 (\frac{1}{x_i + y_j + z_k})} \qquad (\text{Ⅱ-13})$$

위험상대도의 계산은 도출된 위험상대도 산식에 따라 이루어지고, 최적의 위험상대도는 값이 균형을 이루어 더 이상 변화가 없는 시점까지 위험상대도 계산과정을 계속적으로 반복함으로써 산출된다. 우선기초자료로 x_i, y_j, z_k의 1단계 상대도 값을 구한다. 이후 승산모형 또는 가산모형의 x_i, y_j, z_k의 계산공식에 따라 각 변수의 상대도 값이 수렴할 수 있게 계산이 반복되도록 한다. 이러한 계산의 반복과정을 거쳐서 x_i, y_j, z_k의 상대도 값이 수렴하는 수준을 최적 상대도 값으로 정한다. 이처럼 단계적으로 상대도의 최적값을 계산하는 과정 때문에 이러한 종류의 상대도 계산방식은 Iterative method라고 불린다.

<Bailey's Minimum Bias Method>

이 방법은 전체위험도는 각 변수별 위험도로 구성되어 있으므로, 전체위험도에서 모형에 의한 각 변수별 위험도를 차감한 값이 "0" 이 되도록 각 변수에 대한 위험상대도를 구하는 방법이다. 이때 전체위험도를 구성하는 변수들의 위험도가 '승산으로 이루어져 있느냐' 또는 '가산으로 이루어져 있느냐'에 따라 승산모형과 가산모형으로 구분할 수 있다.

Bailey' Minimum Bias Model의 가정을 모형화하면 다음과 같다.

승산모형의 경우 :

$$\sum_{ik} n_{ijk}(\gamma_{ijk} - x_i \times y_j \times z_k) = 0, \ \sum_{jk} n_{ijk}(\gamma_{ijk} - x_i \times y_j \times z_k) = 0,$$

$$\sum_{ij} n_{ijk}(\gamma_{ijk} - x_i \times y_j \times z_k) = 0 \qquad (\text{Ⅱ-14})$$

가산모형의 경우 :

$$\sum_{ik} n_{ijk}(\gamma_{ijk}-x_i-y_j-z_k)=0,\ \sum_{jk} n_{ijk}(\gamma_{ijk}-x_i-y_j-z_k)=0,$$
$$\sum_{ij} n_{ijk}(\gamma_{ijk}-x_i-y_j-z_k)=0 \qquad (\text{Ⅱ-15})$$

3개의 식을 각 변수에 대하여 정리하면, 다음과 같은 위험상대도값의 계산식이 얻어진다. Baily and Simon Method에서와 같이 최적 해는 계산식에 따라 위험 상대도 값이 균형이 되도록 계산을 반복함으로써 구해진다.

승산모형 :

$$x_i=\frac{\sum_{jk} n_{ijk}(\gamma_{ijk})}{\sum_{jk} n_{ijk}\times y_j\times z_k},\quad y_j=\frac{\sum_{ik} n_{ijk}(\gamma_{ijk})}{\sum_{ik} n_{ijk}\times x_i\times z_k},\quad z_k=\frac{\sum_{ij} n_{ijk}(\gamma_{ijk})}{\sum_{ij} n_{ijk}\times x_i\times y_j}$$
$$(\text{Ⅱ-16})$$

가산모형 :

$$x_i=\sum_{jk}\frac{n_{ijk}(\gamma_{ijk}-y_j-z_k)}{\sum_{jk} n_{ijk}},\ y_j=\sum_{ik}\frac{n_{ijk}(\gamma_{ijk}-x_i-z_k)}{\sum_{ik} n_{ijk}},$$
$$z_k=\sum_{ij}\frac{n_{ijk}(\gamma_{ijk}-x_i-y_j)}{\sum_{ij} n_{ijk}} \qquad (\text{Ⅱ-17})$$

<Least Square Method>

‘전체위험도는 각 변수에 대한 위험도 값과 에러 값으로 구성되어 있다’는 가정 하에서 에러 값을 최소화하여 변수별 위험상대도를 구하는 방법이다. 이 방법은 앞서 살펴본 Bailey and Simon Method와 이론적 가정에 큰 차이가 없고, 에러 값에 곱하는 가중치 차이만 있을 뿐이다. 따라서 이론적 가정을 모형화하여 보면 다음과 같다.

승산모형의 경우 :

$$\text{Minimize } SSE = \sum_{ijk} n_{ijk} e_{ijk}^2 = \sum_{ijk} n_{ijk} (\gamma_{ijk} - x_i y_j z_k)^2 \qquad (\text{Ⅱ-18})$$

가산모형의 경우 :

$$\text{Minimize } SSE = \sum_{ijk} n_{ijk} e_{ijk}^2 = \sum_{ijk} n_{ijk} (\gamma_{ijk} - x_i - y_j - z_k)^2 \qquad (\text{Ⅱ-19})$$

모형의 각 변수에 대하여 미분하여 '0'로 놓아 구한 각 변수에 대한 해가 위험상대도가 되며, 그 계산결과는 다음과 같다.

승산모형의 경우 :

$$x_i = \frac{\sum_{jk} n_{ijk} \gamma_{ijk} y_j z_k}{\sum_{jk} n_{ijk} (y_j)^2 (z_k)^2}, \quad y_i = \frac{\sum_{ik} n_{ijk} \gamma_{ijk} x_i z_k}{\sum_{ik} n_{ijk} (x_i)^2 (z_k)^2},$$

$$z_k = \frac{\sum_{ij} n_{ijk} \gamma_{ijk} x_i y_j}{\sum_{ij} n_{ijk} (x_i)^2 (y_j)^2} \qquad (\text{Ⅱ-20})$$

가산모형의 경우 :

$$x_i = \sum_i \frac{n_{ijk} (\gamma_{ijk} - y_j - z_k)}{\sum_{jk} n_{ijk}}, \quad y_j = \sum_j \frac{n_{ijk} (\gamma_{ijk} - x_i - z_k)}{\sum_{ik} n_{ijk}},$$

$$z_k = \sum_k \frac{n_{ijk} (\gamma_{ijk} - x_i - y_j)}{\sum_{ij} n_{ijk}} \qquad (\text{Ⅱ-20})$$

Least Square Method의 경우에도 이전의 모형과 동일하게 각 변수에 대한 최적 위험상대도는 에러 값의 통계량이 최소가 되는 시점, 또는 위험상대도가 균형에 도달하는 시점까지의 계산식에 따른 계산을 계속적으로 반복함으로써 이루어진다. 앞서 소개된 상대도 계산방식에서와 동일하게 Iterative한 절차에 따라 상대도 값이 계산된다.

3) 전통적 요율상대도 산출방법의 문제점

전통적 요율상대도 산출방법은 자료의 분포와 관계없이 요율상대도를 산출한다는 것이다. 즉 기계적인 요율산출방법이 되므로, 모형이 적합 되었는지 여부, 자료와 통계분포의 적합성 여부, 변수 및 모형의 통계적 유의성 등을 확인하지 못한다. 그러므로 전통적인 방법으로 요율상대도를 산출한 값이 통계적으로 진정한 의미의 요율상대도인지 확신하기 어렵다.

과거에 사용되던 기계적 방식의 요율상대도 산출방식의 단점을 극복하기 위해 전통적 회귀분석(OLS) 방법을 사용할 수 있다. 전통적 회귀분석방법은 종속변수의 분포를 정규분포로 가정한 방법이다. 이 방식에서는 종속변수로 사고발생률과 1사고당 손해액을 사용한다. 사고발생률과 1사고당 손해액이 정규분포를 따른다는 가정 하에서 모형을 설정하는 것이다. 전통적 회귀분석방식을 사용하는 경우에는 과거의 기계적 방식의 요율산출방법에서 지적된 ‘모형의 적합성’, ‘변수의 유의성’ 등을 통계적으로 확인하지 못한다는 단점을 극복할 수 있다.

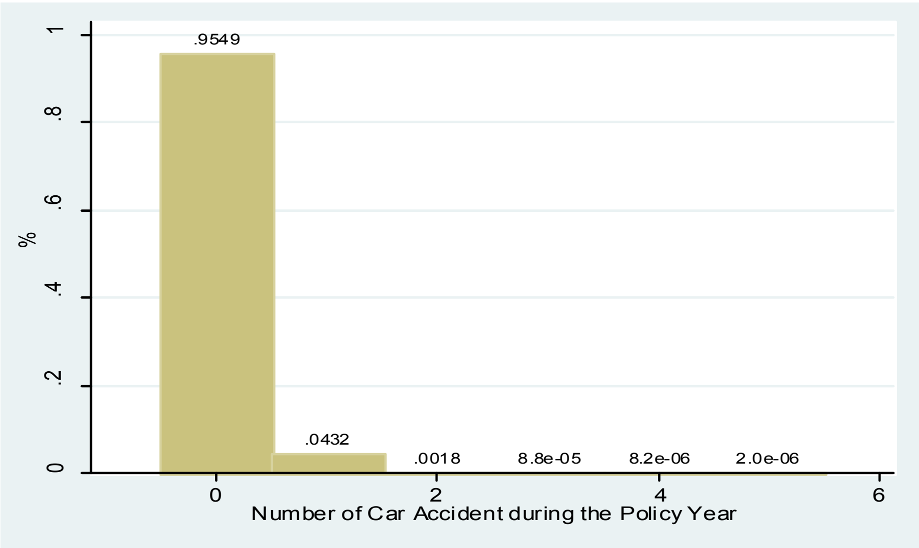
그러나 이러한 전통적 회귀분석을 사용하는 경우 생각해 보아야 할 문제점이 있다. 즉, 기존의 회귀분석 방법의 경우에도 자료의 특성과 모형을 일치시키는데 다소 문제가 있다.

<그림 Ⅱ-1>~<그림 Ⅱ-4>에서 보듯이 자동차사고 빈도와 심도는 지수형태의 분포를 따르기 때문에 정규분포를 가정한 최소자승법은 위험요소를 판별할 수 있는 분석방법은 될 수 있겠지만 위험요소별 상대위험도를 측정하여 정확한 보험요율을 결정해야 하는 데에는 적절한 분석방법이 될 수 없다.

자동차보험에서는 사고 종류별로 발생하는 인적비용과 물적비용을 보험으로 담보하기 위하여 대인배상I, 대인배상Ⅱ, 대물배상, 자기신체사고, 자기차량손해 및 무보험차상해 담보를 만들었다. 이들 담보의 사고분포를 보면, <그림Ⅱ-1>은 대인배상I에 해당하는 사고빈도이며 <그림Ⅱ-2>는 대물배상담보에 해당하는 사고빈도를 보여주는 히스토그램이다. 그리고 <그림Ⅱ-3>와 <그림Ⅱ-4>는 각각 대인배상I에 해당하는 사고심도와 대물배상담보에 해당하는 사고심도를 보여준다.[7]

이들 히스토그램을 볼 때 자동차보험 사고담보의 모든 분포가 정규분포를 따르지 않는다는 것을 쉽게 알 수 있다. 이러한 결과는 전통적 회귀분석 방법을 사용하여 위험상대도를 산출할 경우 부정확한 요율상대도값으로 귀결될 수 있다는 것을 입증한다.

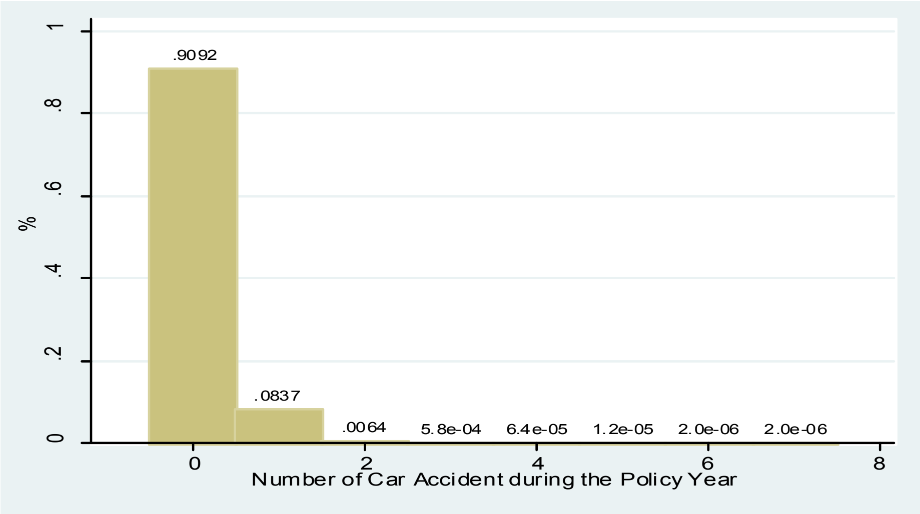
<그림 Ⅱ-1> 대인배상Ⅰ의 사고빈도



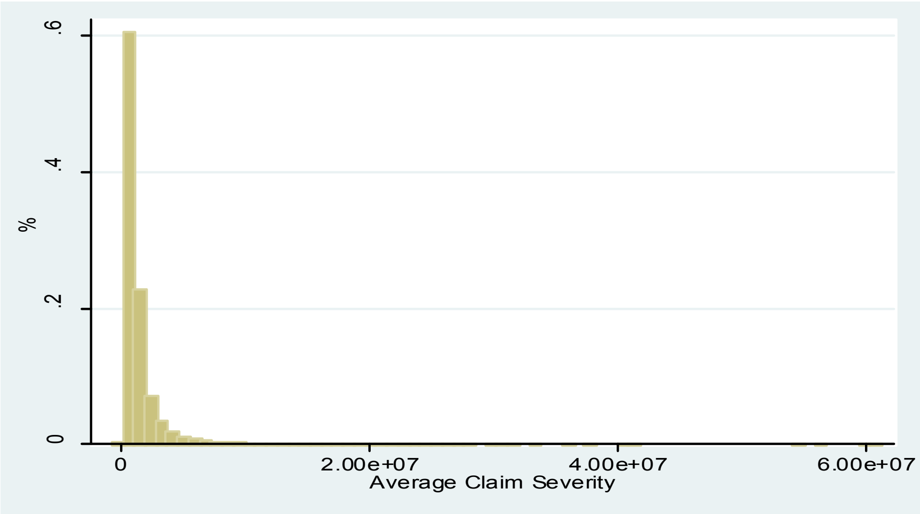
---

7) 편의상 네 개만을 본 연구에 제시하였으며 다른 담보에 해당하는 사고빈도 및 사고심도의 히스토그램은 요청에 따라 이용 가능하다.

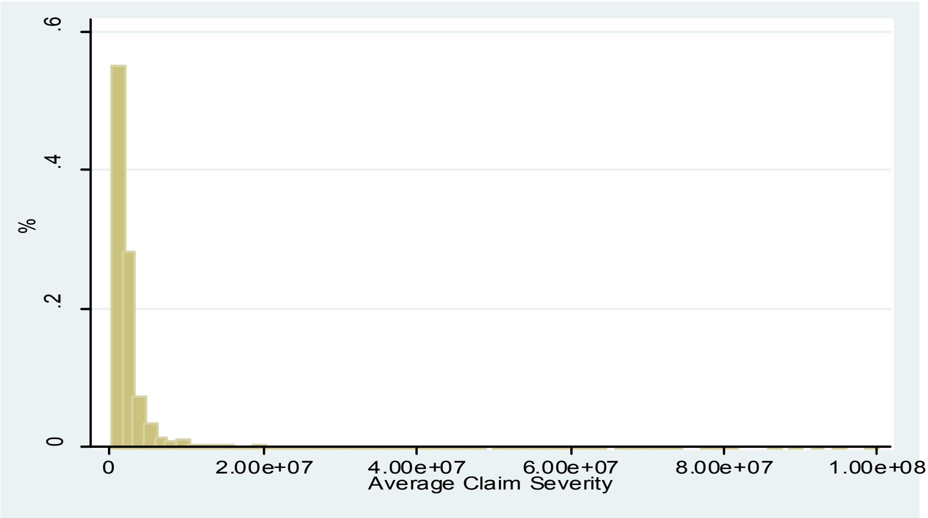
<그림 Ⅱ-2> 대물배상의 사고빈도



<그림 Ⅱ-3> 대인배상I의 사고심도



<그림 Ⅱ-4> 대물배상의 사고심도



일반화선형모형(GLM)은 종속변수의 분포를 정규분포 하나로 가정하고 위험변수들의 상대위험도를 추정하는 최소자승법과 달리 종속변수의 분포가 지수집단(Exponential Family) 위주의 다양한 확률분포를 다루기 때문에 자동차 보험요율분석에 적합한 실증분석 모델이라고 하겠다. 자동차보험 연구에 종속변수로써 가장 빈번히 사용되는 변수인 사고빈도와 사고심도는 <그림II-1>~<그림II-4>에서 보는 바와 같이 지수분포 형태를 취한다.

사고빈도와 사고심도 변수는 서로 상이한 특성을 가지고 있기도 하지만 비슷한 특성을 공유하기도 한다. 공통점은 <그림II-1>~<그림II-4>에서처럼 두 종속변수가 지수분포라는 큰 틀에서 같은 형태의 확률분포를 가지고 있다는 것이며 다른 점은 두 종속변수의 연속성 여부이다. 사고빈도는 연속변수가 아닌 건수변수(Count Variable)이고 사고심도는 연속변수(Continuous Variable)이며, 변수의 연속성 여부에 따라 적용되는 일반화선형모델들이 다르다.

Ⅲ. 요율상대도 산출시 일반화선형모형 활용방법

1. 선행연구

Nelder와 Wedderburn(1972)은 처음으로 일반화선형모형을 소개하였다. McCullagh와 Nelder(1983)는 다시 일반화선형모형에 대한 의미 있는 논문을 발표하였고 일반화선형모형을 적용할 수 있는 몇 가지 예를 제시하였으며, 그 예 중 하나는 자동차보험 자료를 활용하는 것이었다. 즉 동 논문은 자동차보험의 평균손해액(Average Claim Costs)자료를 일반화 선형모형으로 적합할 수 있다는 것을 보여주었다.

Haberman과 Renshaw(1988)는 일반화선형모형이 통계학 분야에서 소개된 이후 보험의 요율산출분야에 적용할 수 있는 방법을 발표하였다. Haberman과 Renshaw(1988)는 동 논문에서 일반화선형모형이 자동차보험 분야에만 적용되는 것이 아니라 생명보험분야에도 적용될 수 있다는 것을 제시하였다. 즉 Herberman과 Renshaw는 손해보험의 손해액분포를 일반화 선형모형으로 적합 시키는 방법, 생명보험에서 사력변화(the Variation in Force of Mortality)를 나타내는 방법 등을 소개하였다.

Holler외 3인(1999)은 요율상대도 산출방법에 일반화 선형모형을 적용하는 방법을 연구·제시하였다. Keith외 3인은 단변량을 사용하여 상대도를 산출할 때 발생할 수 있는 오류를 제거하기 위하여 다변량을 사용하여 상대도를 산출하는 수학적인 전통적인 방법을 소개하였다. 그리고 이 전통적인 방법을 대체할 수 있는 방법으로 일반화선형모형(Generalized Linear Model)을 소개하고, 동 통계모형의 장점을 소개, 증명하였다. 그리고 Keith외 3인은 신용보험의 예로 일반화 선형모형을 적용하여 요율상대도를 산출하는 방법을 적용하였다. Keith외 3인은 일반화선형모형이 개인보험(Personal Lines)에 주로 적용된다는 점을 지적하였다. 그리고 일반적으로 일반화선형모형이 적용되는 방법으로, 사고데이터를 사고빈도와 사고심도로 분리해 일반화선형모형을 적용하여 요율상대도를 산출하고, 사고빈도는 포

아송 또는 음이항 분포의 구조가 주로 적용되고, 사고심도는 감마 및 로그노말(Lognormal)분포의 구조가 적용된다고 설명하였다. 또한 상대도 산출모델의 구조는 곱셈구조가 적합하다고 하였다. 그러나 이와 같이 일반적으로 사용되는 일반화선형모형 적용방법분야에서 '모형구조를 곱셈으로 하는 것 또는 덧셈구조로 하는 것이 적당한가' 및 '사고빈도와 심도로 나누어 분석하는 것이 좋은지 아니면 사고심도와 빈도를 모두 포함한 위험보험료 단위로 분석하는 것이 좋은 지'등에 대하여 아직도 논란이 있다고 지적하였다.

Murphy외 3인(2000)은 일반화 선형모형을 개인보험에 적용하는 방법을 소개하였다. Murphy외 3인은 Brockman과 Wright가 일반화선형모형을 개인보험에 적용할 수 있는 방법을 제시한 내용을 근거로 실무적으로 적용하는 방법을 설명하고 있다. 이를 살펴보면 증권별 위험보험료를 평가하는데 일반화선형모형이 어떻게 사용될 수 있는지, 일반화선형모형을 사용하여 가격탄력성 커브를 평가하는 방법, 비용과 수요측면에서 가격을 결정하는 방법을 자세하게 설명하고 있다.

Anderson외 5인(2003)은 보험통계를 이용하여 일반화선형모형을 적용하는 방법을 실제 요율산출 종사자들에게 설명하기 위한 자료를 발표하였다. Duncan외 5인은 동 자료에서 일반화선형모형의 일반이론, 일반화선형모형을 적용하기 위한 자료의 종류, 모델 선택, 모델 정제화, 해석방법 등을 설명하고 있다.

또한 일반화선형모형을 적용할 때 분포가정에 대한 연구는 사고심도와 사고빈도로 나누어 시행하였다. 사고심도에 관해서는 일반화선형모형을 심도에 적용할 때 일반적으로 감마분포와 정규분포를 사용하는 것이 적정하다는 연구들이 있다. 이들 분포 외에도 로그노말(Lognormal)도 많이 사용되고 있다. 심도자료에 대하여 로그를 취하여 자료를 변환시켜 일반화선형모형(정규분포)을 적용하면, 심도자료가 정규분포를 따르게 되고 모형식이 선형으로 된다. 즉 심도자료에 로그를 취하고 일반화선형모형을 취하는 것은 심도자료에 로그노말 일반화선형모형을 적용하는 것과 같다. 이러한 모형적용방식은 널리 사용되고 있으며, 이와 같은 방식으로 심도자료에 일반화선형모형을 적용하는 것의 타당성에 대하여 많은 연구결과들이 발표되었다.

Luyang과 Moncher(2000)는 심도자료를 이용하여 일반화선형모형으로 요율상대도를 산출할 때 감마분포가 로그노말보다 더 정확성과 효율성이 뛰어나다고 주장하였다. Lyuang와 Moncher(2000)는 몬테칼로 시뮬레이션 방법으로 자신의 주장을 입증하였다. 그리고 일반화선형모형을 사고빈도에 적용할 때에는 포아송분포(Poisson Distribution)와 음이항분포(Negative Binomial Distribution)가 유용하다는 연구를 입증하였다(Apostolaskis 1998).

이들 연구이외에 자동차보험통계로 일반화선형모형을 사용하여 요율상대도를 산출할 때 발행하는 과대산포(Overdispersion)를 해결하는 방법(Ismail 및 Jemain 2007), 일반화선형모형과 신뢰도의 관계(Ohlsson 및 Johansson, 2004, Nelder 및 Verrall, 1997), 일반화선형모형을 준비금에 적용(Mack 1991 및 1994; Renshaw Arthor E 1995; Verrall, Richard J 2000; Wright 1990 및 1993)등과 같이 일반화선형모형을 적용할 수 있는 분야에 대한 연구가 점점 확대되고 있다.

이처럼 외국의 요율산출분야에서는 일반화선형모형이 처음 발표된 이후 지금까지 일반화선형모형의 적용방법에 대하여 많은 연구와 실제 적용방법이 발표되어 왔다. 그러나 우리나라에서는 외국에서 활발하게 사용되고 있는 일반화선형모형을 활용한 요율산출방법에 대한 연구가 미미한 실정이다. 우리나라의 경우에는 최우석 외 1인(2008)이 자동차보험에 일반화선형모형(GLM)을 적용한 분석결과를 발표한 적이 있다. 최우석 외 1인(2008)은 Smyth-Jorgenson(2002)이 적용한 바 있는 Tweedie's Compound 포아송 분포 가정하에 평균과 분산을 함께 고려할 수 있는 DGLM을 우리나라와 호주의 자동차보험 통계에 적용하였다. 위의 분석결과에 따르면 특정 공변량들의 경우 청구빈도와 규모에 동일한 또는 반대의 영향을 미치고 있음을 실증적으로 확인하였고, 요율요소(특히 할인할증 제도)간 위험도 편차가 있는 것을 발견하였다. 지역별로는 수도권과 지방간 보험료 양극화 문제가 심각하지 않은 것으로 분석하였다. 그러나 최우석외 1인의 분석은 Jong 과 Helle(2008)이 책자에서 제공한 호주 자료를 활용한데 지니지 않고, 분석에 사용한 우리나라 자동차보험 통계도 사고빈도(사고발생률) 자료만으로 한정되었다. 또한 최우석외 1인의 연구결과는 자동차보험료 산출방법의 전반에

대한 것이 아닌 DGLM을 적용하는데 국한된 연구결과이다.

이상의 선행 연구결과를 볼 때 자동차보험 요율산출에서 일반화선형모형을 적용한 연구는 우리나라에서 거의 수행되지 않았다. 현재 외국에서 자동차보험 요율산출에 일반화선형모형을 적용하는 추세 등을 볼 때 자동차보험 요율산출 시에 일반화선형모형을 적용하는 방법의 기준이 되는 연구가 필요한 시점이다. 일반화선형모형을 자동차보험 요율산출에 적용할 때 고려해야할 많은 문제들, 그리고 외국에서 논의되었고 현재 논의되고 있는 많은 이슈들에 대한 해답을 우리나라 자동차보험 통계를 활용하여 확인하고, 그 결과를 제시하는 것은 현재 시점에서 중요한 의미가 있을 것으로 판단된다.

따라서 본 연구에서는 선행연구결과를 바탕으로 일반화선형모형의 일반적인 이론을 우선 살펴보고, '요율산출에 실질적으로 일반화선형모형(GLM)을 적용할 수 있는 방법', '사고빈도와 사고심도로 분리하여 분석하는 경우의 타당한 모델', '사고심도와 사고빈도 모형결정 시 적합한 연결함수(Link Function) 선택문제', '모형을 적합 시키는 방법', '적합된 모형의 해석 문제' 등에 대한 기준을 제시하고자 한다.

2. 일반화선형모형 이론

가. 일반화선형모형의 구성

일반화선형모형이란 종속변수에 영향을 주는 1개 이상의 독립변수의 효과를 측정하는 기존의 선형모형을 일반화시킨 모형이다. 좀 더 명확한 설명을 위해 일반화선형모형을 구성하는 3가지 기본성분을 살펴 볼 필요가 있겠다. 첫 번째로 랜덤성분(Random Component)인데 이것은 종속변수(또는 반응변수) Y의 확률분포를 규정하는 성분이다. 즉, 표본의 크기가 N이며 종속변수의 관측 값을 Y_i라고 할 때 일반화선형모형은 N개의 Y들이 서로 독립임을 가정하며 일반화선형모형의 랜덤성분이 지수족(Exponential Family)에 한해 Y의 확률분포를 결정하게 되는 것이다.

두 번째 성분인 체계적성분(Systematic Component)은 모형의 예측변수로 사용되는 설명변수를 규정하여 확률분포의 평균인 Y의 기댓값 $E(Y)=\mu$을 나타낸다. 즉 Y는 설명변수들(X_i)의 수준에 따라 다양하게 나타나게 되는 것이다.

일반화선형모형의 세 번째 성분은 연결함수(Link Function)로 첫 번째 성분인 랜덤성분($E(Y)=\mu$)과 두 번째 성분인 체계적성분($X\beta$)을 연결하며 또한 두 성분이 어떻게 연결되어 있는지 설명하는 역할을 한다. 각각의 일반화선형모형들은 연결함수로써 항등함수(Identity), 로그(Log)함수 등을 포함하여 다양한 연결함수의 적용이 가능하다. <표 Ⅲ-1>는 대표적인 일반화선형모형에서 사용되고 있는 일반적인 연결함수들 이다.

기존의 선형모형은 구성성분인 랜덤성분에 대해 정규분포만을 가정하였고, 세 번째 구성요소인 링크함수를 아래와 같은 항등함수(Identity Function)만을 허용한 제한적인 모델이다.

$$g(\mu)=a+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\ldots\ldots+\beta_k x_k \tag{Ⅲ-1}$$

하지만 일반화선형모형의 경우는 이러한 제약 없이 랜덤성분인 Y가 지수분포인 여러 가지 분포들에 적용이 가능하며 링크함수 역시 전 구간에서 미분 가능한 단조증가 함수를 모두 사용할 수 있다. 또한 일반적으로 사용되어온 선형모델은 동분산(Homoskedasticity)을 가정하는 반면 일반화선형모형에서는 분산이 평균의 함수형태로 나타내어진다.

$$V(y)=a(\phi)\,V(\mu)^{8)} \tag{Ⅲ-2}$$

이러한 특성을 반영한 일반화선형모형에서 Y의 기본적인 확률분포인 지수족(Exponential Family)의 일반 형태는 다음과 같다.

---

8) $a(\phi)$은 가중치이며 포아송, Binomial, Negative Binomial등의 모델에서는 1이 사용된다.

$$f_y(y;\theta,\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\} \tag{Ⅲ-3}$$

이때 θ는 추정해야할 모수이며 ϕ은 가우시안(Gaussian), 감마(Gamma), 역가우시안(Inverse Gaussian) 등에서 표준오차(Standard Error)를 계산할 때 필요한 가중치이다. 각 관측치 y_i가 서로 독립임을 가정하기 때문에 샘플의 관측치y_i의 결합분포(Joint Density of the Sample of Observations y_i)는 다음과 같다.

$$f_{y_1 y_2,\ldots,y_n}(y_1,y_2,\ldots,y_n;\theta,\phi) = \prod_{i=1}^{n}\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\} \tag{Ⅲ-4}$$

또는 우도함수(Likelihood Function)의 형태로 다음과 같이 나타낼 수 있다.

$$L(\theta,\phi;y_1,y_2,\ldots,y_n) = \prod_{i=1}^{n}\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\} \tag{Ⅲ-5}$$

마지막으로 계산상 편의를 위해 식(III-5)의 우도함수에 로그(Log)를 취한 뒤에 로그우도함수(Log Likelihood Function)를 최대화 시키는 θ과 ϕ를 측정하게 된다.[9)]

$$\ell(\theta,\phi;y_1,y_2,\ldots,y_n) = \sum_{i=1}^{n}\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\} \tag{Ⅲ-6}$$

일반적으로 로그우도함수를 극대화 시키는 θ과 ϕ값을 계산하기 위한 알고리즘(Algorithm)으로 Newton-Raphson 또는 IRLS 알고리즘이 이용 된다.[10)]

---

9) 식(Ⅲ-6)을 최대화시키는 θ과 ϕ값은 식(Ⅲ-5) 또한 최대화시키기 때문에 식(Ⅲ-5)의 식에 로그를 취한 뒤에 계산이 간편한 식(Ⅲ-6)을 이용하여 θ과 ϕ값을 측정하게 된다.

10) Newton-Raphson과 IRLS 알고리즘에 대한 논의는 본 연구보고서의 목적과 직

지수족 분포의 평균과 분산은 다음과 같이 계산된다. 분포함수가 지수족 분포이므로, 분포의 정의에 따라 동 분포를 전 영역에 대하여 적분(또는 합산)할 경우 그 값은 1이 된다. 즉, $\int f(y,\theta,\psi)dy=1$ 이다.

이 식을 θ에 대하여 미분하여 0으로 놓으면 다음의 식과 같이 된다.

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int f(y,\theta,\psi)dy = \int \frac{\partial}{\partial\theta} f(y,\theta,\psi)dy \\ &= \int \left\{\frac{\partial}{\partial\theta} log f(y,\theta,\psi)\right\} f(y,\theta,\psi)dy = E\left\{\frac{\partial}{\partial\theta}\ell(y,\theta,\psi)\right\} \text{ 이고,} \end{aligned} \quad (\text{Ⅲ-7})$$

여기서 $\ell(y,\theta,\psi)$는 $\ell(y,\theta,\psi)=\log f(y,\theta,\psi)$인 log-likelihood이다. 그리고 $\frac{\partial}{\partial\theta}\ell(y,\theta,\psi)$는score이며, $E\left\{\frac{\partial^2}{\partial\theta^2}\ell(y,\theta,\psi)\right\}=-E\left\{\frac{\partial}{\partial\theta}\ell(y,\theta,\psi)\right\}^2$이다.

$$\ell(y,\theta,\psi)=\log f(y,\theta,\psi) \quad (\text{Ⅲ-8})$$

이 식에 앞의 지수족 일반식 식(Ⅲ-3)을 대입하면, 다음과 같이 평균과 분산을 계산할 수 있다.

$$0=E\left\{\frac{Y-b'(\theta)}{a(\psi)}\right\},\ E\left\{\frac{-b''(\theta)}{a(\psi)}\right\}=-E\left\{\frac{Y-b'(\theta)}{a(\psi)}\right\}^2 \text{이다.} \quad (\text{Ⅲ-9})$$

따라서 $E(Y)=\mu=b'(\theta)$이고, $Var(Y)=V(\mu)a(\psi)=b''(\theta)a(\psi)$이 된다. 이 식에 따라 지수족의 분산을 계산해보면 다음의 <표 Ⅲ-1>과 같다.

결되지 않으며 일반화선형모형 이외의 모델에서도 사용되는 최적화방법의 이론적 소개에 불과하여 생략하기로 한다. 무엇보다 Newton-Raphson과 IRLS 알고리즘에 대한 논의는 대학원 수준 이상의 거의 모든 계량경제학 책에서 다루고 있다. 특히 Newton-Raphson에 대해서는 Cameron 과 Trivedi(2005)에서 자세히 다루고 있다.

<표 Ⅲ-1> canonical 연결(link) 및 분산(variance) 함수

Family	canonical link	이 름	분 산
Binomial	$\log(\mu/(1-\mu))$	logit	$\mu(1-\mu)$
Gamma	$-1/\mu$	inverse	μ^2
Gaussian	μ	identity	1
Inverse-Gaussian	$-2/\mu^2$	1/mu^2	μ^3
Poisson	$\log\mu$	log	μ
Negative Binomial	$\log\mu$	log	$\mu_i(1+\alpha\mu)$

자료 : Venables,W.N.,Ripley,B.D.(2002), p185

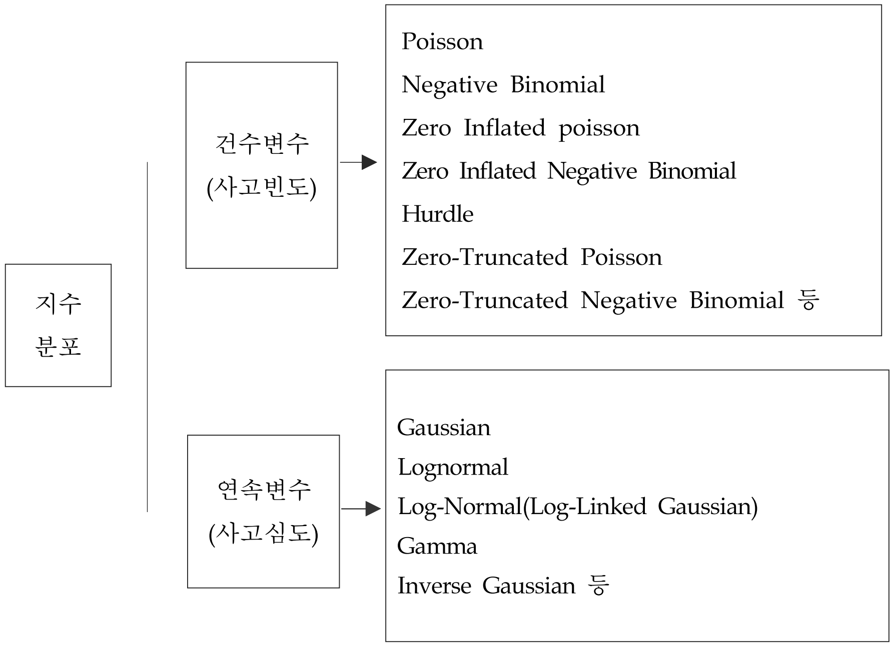
나. 일반화선형모형의 종류

일반화선형모형에 해당하는 분석모델들이 매우 많은 게 사실이며[11], 다양한 일반화선형모형들을 크게 두 분류로 나눌 수 있는 가장 기본이 되는 구분기준은 종속변수의 연속성이다. 자동차사고 시 사고액(사고심도)과 같은 연속변수(Continuous Variable)와 교통사고 빈도처럼 연속적이지 않은 이산변수(Discrete Variable)가 기준이 되는 것이다. 비연속변수 중 특히 교통사고나 병원방문 횟수와 같이 한번, 두 번, 세 번, 네 번…(0,1,2,3,4,…)처럼 셀 수 있는, 즉 종속변수가 0과 양의정수의 범위에 속하는 경우를 건수변수(Count Variable)[12]라고 한다.

11) 많은 연구자들에 의해 사용되고 있는 로짓(Logit) 또는 프로빗(Probit)도 GLM이라고 인식되지 않고 있지만 이론적으로 GLM의 범주에 속한다.

12) 건수변수의 이론적 그리고 실증적 분석의 전반적인 이해를 위해 Cameron and Trivedi (1986)의 계량경제학 서적을 추천한다.

<그림 III-1> 종속변수의 연속성 여부에 따른 GLM의 분류



건수변수를 분석하는 적합한 GLM 분석모형으로 가장 기본적인 모델로 포아송(Poisson)이 있으며 이밖에 음이항 분포(NB: Negative Binomial), Zero Inflated poisson(ZIP), Zero Inflated Negative Binomial(ZINB), Hurdle, Zero-Truncated Poisson(ZTP), Zero-Truncated Negative Binomial(ZTNB) 등이 있다. 이 외에도 계량경제학자들에 의해서 새로운 모델들이 개발 중에 있다.

지수분포 형태의 연속변수를 분석하기 위한 일반화선형모형(GLM)의 분석모형은 가우시안(Gaussian), 로그노말(Lognormal), Log-Normal(또는 Log- Linked Gaussian), 감마(Gamma), 역가우시안(Inverse Gaussian) 등이 있다.

1) 건수변수(Count Variable)를 위한 일반화선형모형

가) 포아송(Poisson)

1년 동안 자동차 사고빈도뿐만 아니라 병원 방문 수, 입원 수, 어느 지역에서 24시간 동안 발생하는 교통사고 사망자 수, 백과사전 한 페이지에 나타난 한자의 수, 창문 유리 한 장에 나 있는 긁힌 자국의 수 등과 같이 희귀한 사건의 발생횟수에 관심을 갖고 있는 경우를 생각해 볼 수 있다. 이와 같이 시간적인 또는 공간적인 단위구간 내에서 '성공'의 출현 횟수를 Y로 정의하면, 확률변수 Y는 다음의 조건들이 만족될 경우 포아송 분포(Poisson Distribution)를 따른다고 한다.

어떤 시간 또는 공간(또는 물질의 부피)으로 표시되는 구간에 있어서 특정 확률사상이 발생한 건수를 Y라고 하면 Y가 일어날 확률은 다음과 같은 특성을 지니고 있다.

(1) 독립성: 겹치지 않는 구간 내에서의 '성공' 횟수들은 서로 독립적이다. 즉 어떤 공간이나 시간을 단위형태로 갖는 구간에서 어떤 사건이 일어나는 것은 동일 구간이나 다른 구간에서 일어나게 되는 또 다른 발생에는 전혀 영향을 미치지 못한다.

(2) 비군집성(Lack of Clustering): 극히 작은 구간 내에서 둘 이상의 성공이 일어날 확률은 0으로 간주한다. 즉, 구간의 무한히 짧은 부분에 그 사건이 한 번 이상 일어날 확률은 무시된다.

(3) 비례성(Proportion): 충분히 짧은 구간 내에서 정확히 1번의 '성공'이 발생할 확률은 그 구간의 길이에 비례한다.

(4) 이론적으로는 그 구간에서는 관찰대상 사건이 무한히 많이 발생할 수 있다.

이상의 조건이 만족될 때 확률변수 Y의 밀도함수는 다음과 같다.

$$f(y:\mu)=\frac{e^{-\mu}\mu^{y}}{y!}$$ 13) (Ⅲ-10)

또는 지수분포 형태로 다음과 같이 나타낼 수 있다.

$$f(y:\mu)=\exp\{y\ln(\mu)-\mu-\ln\Gamma(y+1)\}$$ (Ⅲ-11)

y는 0, 1, 2, 3...의 성공 횟수를 나타내며 μ는 단위구간 당 ‘성공’의 평균 출현횟수로 아래와 같다.

$$\mu=\exp(x\beta)$$ (Ⅲ-12)

포아송 분포의 가장 흥미로운 특징은 평균과 분산이 같다는 가정에 기반을 두어 만들어졌다는 것이다.

$$E(y:\mu)=Var(y:\mu)=\exp(x\beta)=\mu$$ (Ⅲ-13)

포아송 모델은 일반적으로 최우추정방법을 이용하여 (ML: Maximum Likelihood)추정하며, 최우추정함수는 다음과 같다.

$$\ln L_{Possion}=\sum_{i=1}^{n}\{y\ln(\mu)-\mu-\ln\Gamma(y+1)\}$$ (Ⅲ-14)

포아송 모델은 건수변수를 분석하기 위해 만들어진 가장 초기적인 모델이면서 여전히 실증분석에서 가장 빈번히 사용되는 모델이기도 하다. 하지만 평균과 분산이 같다는 다소 제약적인 가정 하에 만들어진 포아송 모형은 실제로 많은 실증분석을 하기에는 한계가 있다. 실제 데이터를 분석하다보면

---

13) 모든 y값에 대하여 $f(y)\geq 0$이고 $\sum f(y)=1$이므로 이것은 확률분포의 조건에 부합된다.

분산이 평균보다 작거나 큰 경우가 나타날 수 있으며 분산이 평균보다 작아서 생기는 문제를 저산포(Underdispersion)라 하고 반대의 경우를 과대산포(Overdispersion)라 한다. 일반적으로 분산이 평균보다 큰 경우가 대부분이며 과대산포의 경우 포아송 모델 대신 음이항분포(Negative Binomial)모델을 이용하는 것이 적당하다. 참고로 관측되지 않은 이질성 때문에 과대산포의 문제가 발생할 수 있으며, 또한 계량모델에 중요변수를 생략했을 때도 과대산포의 문제가 발생할 수 있다. 이러한 논리에 기반을 둘 때, 과대산포의 경우 음이항분포를 사용하는 대신 충분히 설명력이 있는 중요 변수들을 모델에 포함시키고 포아송을 사용해도 무방할 수 있다는 주장이 설득력을 얻을 수 있다. 하지만 분석모델에 포함된 변수들이 충분히 관측되지 않은 이질성을 감소시키고 있는지에 대한 테스트는 여전히 필요하다고 하겠다. 그리고 과대산포의 원인으로 연구자가 확인해야 할 부분 중 하나는 이상치(Outlier)들이다. 과대산포는 이상치들 때문에도 생길 수 있기 때문에 일반화 선형모형을 적용하기 전에 이상치들을 충분히 제거하는 것이 최적의 모델을 선택하기위해 선행되어야 한다. 이상치들을 제거하는 방법에 대해서는 5장에서 자세히 설명하였다. 이밖에도 과대산포는 중요한 교호작용 변수(Interaction Term)들을 분석모델에서 포함시키지 않았을 경우에도 발생할 수 있으며 앞서 서술한 바와 같이 과대산포의 원인은 매우 다양하다.

나) 음이항분포(Negative Binomial)

평균과 분산이 같다는 제약적인 포아송분포의 단점을 보완하기 위해 만들어진 모델이 음이항분포(NB: Negative Binomial) 모형이다. 즉, 음이항분포 모형은 과대산포의 경우 부적합한 포아송모델 대신 포아송과 감마(Gamma)분포를 혼합하여 만들어진 확률분포이다.

$$f(y:x) = \int_0^\infty \frac{e^{-(\lambda)}(\lambda_i)^y}{y!} \frac{\delta^\mu}{\Gamma(\mu)} \lambda^{\mu-1} e^{-\lambda\delta} d\lambda$$

$$= \frac{\delta^{\mu_i}}{\Gamma(y_i+1)\Gamma(\mu_i)} \int_0^\infty \lambda_i^{(y_i+\mu_i)-1} e_i^{-\lambda_i(\delta+1)} d\lambda_i$$

$$= \frac{\delta^{\mu}}{\Gamma(y+1)\Gamma(\mu)} \frac{\Gamma(y+\mu)}{(\delta+1)^{y+\mu}} \int_0^\infty \frac{(\delta+1)^{y+\mu}}{\Gamma(y+\mu)} \lambda^{(y+\mu)-1} e^{-\lambda(\delta+1)} d\lambda$$

$$= \frac{\delta^{\mu}}{\Gamma(\mu+1)\Gamma(\mu)} \frac{\Gamma(y+\mu)}{(\delta+1)^{y+\mu}}$$

$$= \frac{\Gamma(y+\mu)}{\Gamma(\mu)\Gamma(y+1)} \left(\frac{\delta}{\delta+1}\right)^{\mu} \left(\frac{1}{1+\delta}\right)^{y} \qquad (Ⅲ\text{-}15)$$

음이항분포의 평균과 분산은 다음과 같다.

$$E(y) = \frac{e^{x\beta + offset}}{\delta} \qquad (Ⅲ\text{-}16)$$

$$V(y) = \frac{e^{x\beta + offset}}{\delta} \frac{1}{1+\delta} \qquad (Ⅲ\text{-}17)$$

즉 음이항분포의 분산은 평균에 $\frac{1}{1+\delta}$를 곱한 값이다. $\frac{1}{1+\delta}$에서 δ를 α로 치환한 다음 식(III-15)을 다시 쓰면 아래와 같다.

$$f(y:x) = \frac{\Gamma(y+\mu)}{\Gamma(\mu)\Gamma(y+1)} \left(\frac{\alpha}{1+\alpha}\right)^{\mu} \left(\frac{1}{1+\alpha}\right)^{y} \qquad (Ⅲ\text{-}18)$$

$$E(y) = \mu \qquad (Ⅲ\text{-}19)$$

$$V(y) = \mu(1+\alpha\mu), \quad \alpha > 0 \qquad (Ⅲ\text{-}20)$$

음이항분포(Negative Binomial)의 최우추정함수는 다음과 같다.

$$\ln L_{NB} = \sum_{i=1}^{n}\left\{ y\ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \frac{1}{\alpha}ln(1+\alpha\mu_i)\ln\Gamma(y_i+\frac{1}{\alpha}) - \ln\Gamma(y_i+1) - \ln\Gamma(\frac{1}{\alpha})\right\}$$

(Ⅲ-21)

식(III-19)과 식(III-20)에서 보듯이 음이항분포는 분산이 평균보다 항상 크고(과대산포) 과대산포의 정도는 α가 커질수록 같이 커진다는 것을 알 수 있다. 또한 분산의 크기가 일정하지 않고 평균의 함수임을 알 수 있다.

과대산포의 문제는 관측되지 않은 이질성(Heterogeneity)이나 발생 사건들이 서로 독립적이지 않을 때 발생하게 되는데 이 경우 α가 0보다 큰 값을 가지게 된다. 과대산포의 경우 표준오차(Standard Error)가 실제보다 작게 측정되며(Underestimated) 결과적으로 유의하지 않은 위험변수도 유의한 결과가 나올 수 있는 오류를 범하게 된다. 일반적으로 많은 연구자들이 표본(Sample)에서 Y의 평균과 분산의 크기를 비교하여 분산이 평균보다 클 경우 과대산포가 존재한다고 결정하고 포아송 대신 음이항분포모형을 사용하게 되는데 이때, 좀 더 정교한 테스트가 요구된다. 즉, 본 연구의 실증분석 자료를 볼 때 대인배상I의 경우 평균은 0.047이며 표준편차는 0.222이므로 분산이 0.049로 평균보다 크다는 것을 알 수 있다. 하지만 많은 연구자들이 분산 0.049가 평균 0.047보다 크다고 판단하고 쉽게 음이항분포(Negative Binomial)모델을 적용하곤 하는데 매우 부적절한 결정일 수 있다. 포아송과 음이항분포(Negative Binomial)모델 중 적합한 모델을 선택하기 위해선 주의 깊은 테스트가 필요하다.

실제로 모델을 분석하는 경우 과대산포 테스트가 아닌 ‘조건부 과대산포(Conditional Overdispersion)’ 테스트를 해야 한다.

$$\hat{z} = \rho\hat{\mu} + \epsilon \quad (Ⅲ\text{-}22)$$

$$\hat{z} = \frac{(Y-\hat{\mu})^2 - Y}{\hat{\mu}\sqrt{2}} \quad (Ⅲ\text{-}23)$$

포아송 모델을 이용해 각 관측치에 대해 $\exp(x\beta)$의 측정치인 $\hat{\mu}$를 계산하고 식(Ⅲ-23)에 의해 $\hat{z}$를 계산한다. 그 다음 식(Ⅲ-22)을 최소자승법(OLS)으로 추정하고 $H_0 : \hat{\rho} = 0$을 테스트 하는 것이다. 많은 통계프로그램에서는 과대산포 테스트의 결과가 식(Ⅲ-20)에 있는 α와 0을 비교하는 테스트로 이루어진다. 즉 $H: \alpha > 0$ 인지를 테스트 하게 되는데 이 논리는 위에서 언급한 과대산포 테스트와 같으며 결과의 해석도 같다. 즉, 식(Ⅲ-20)에서 α가 0보다 크면 음이항분포를 사용해야 하고 그렇지 않은 경우 포아송 모델을 이용하는 것이 적합하다. 이는 식(Ⅲ-18)의 α에 0을 대입할 때 식(Ⅲ-18)이 포아송 분포 식(Ⅲ-10)과 같아지는 것을 보면 이해가 쉬울 것이다.

2) Continuous Parameter Binomial(CPB)

실제로 여러 종류의 데이터를 다루다 보면 과대산포 때문에 평균과 분산이 같다는 제약적인 가정에 기반한 포아송 모델 대신 음이항분포를 사용하게 된다. 하지만 저산포의 문제를 가지고 있는 데이터도 있으며 이럴 경우에는 과대산포와 반대로 표준오차(Standard Error)를 과대 측정하게 되고 결국 유의한 위험변수를 유의하지 않게 해석하는 오류를 범하게 된다. 그러므로 저산포의 경우 포아송이 아닌 CPB(Continuous Parameter Binomial) 모델을 적용하는 것이 바람직하다.

$$f(y:x) = \frac{\dfrac{\Gamma\left(\dfrac{-\mu}{\alpha-1}+1\right)}{Y!\,\Gamma\left(\dfrac{-\mu}{\alpha-1}-Y+1\right)}(1-\alpha)^{y}(\alpha)^{\frac{-\mu}{\alpha-1}-y}}{D} \qquad (Ⅲ\text{-}24)$$

$$E(y) = \mu \qquad (Ⅲ\text{-}25)$$

$$V(y) = \mu\alpha, \quad 0 < \alpha < 1 \qquad (Ⅲ\text{-}26)$$

D는 단순한 Scaling 요소(Factor)이며, CPB의 경우 분산이 평균보다 작다는 것을 알 수 있다. 이는 식(III-26)에서 α가 0보다 작으면 CPB를 사용해야 하고 그렇지 않은 경우 포아송 모델을 이용하는 것이 적당하다는 것을 보여준다. CPB의 최우추정함수는 다음과 같다.

$$\ln L_{CPB} = \sum_{i=1}^{n} \left\{ \begin{array}{l} \ln\Gamma(\frac{-\mu_i}{\alpha-1}+1) - \ln\Gamma(\frac{-\mu_i}{\alpha-1} - Y_i + 1) \\ + Y_i \ln(\alpha-1) + (\frac{-\mu_i}{1-\alpha} - Y_i)\ln(\alpha) - \ln(D_i) \end{array} \right\}$$

(Ⅲ-27)

다) Zero-Inflated Negative Binomial(ZINB)와 Zero-Inflated Poisson

0을 포함하는 건수변수는 과대산포 또는 저산포의 문제 이외에도 Excessive Zero의 문제가 있을 수 있다. 건수변수를 다루는 일반화선형모형들은 주로 보건경제학에서 다루어져 왔으며 Excessive Zero의 이해를 돕기 위해 보건경제학에서 흔히 이용되는 예를 이용해 쉽게 설명할 수 있다. 종속변수가 개별 학생들의 학교 결석일수라고 가정해 보자. 일정 기간 동안 결석을 단 하루도 하지 않은 학생들이 대부분일 것이며 데이터 상에 0으로 기입될 것이다. 하지만 결석을 하지 않는 학생들은 여러 타입으로 그룹화되어질 수 있다. 예를 들어, 결석하지 않은 두 명의 학생이 있다고 가정할 경우 첫 번째 학생은 아프면 결석을 하는 성향을 가지고 있지만 데이터를 집적하는 일정 기간 동안 아프지 않아서 0이라는 종속변수로 귀결될 수 있다. 두 번째 학생은 아파도 부모님의 강한 권유로 인해 결석하지 않는 성향을 가지고 있어 항상 0이라는 값으로 귀결될 것이다. 즉 두 번째 학생과 같은 성향을 가진 학생들로 인해 종속변수에 0의 개수가 많아진다는 것이 바로 Excessive Zero 또는 0이 부풀려졌다는 의미로 Inflated Zero라고 한다. 결국 Excessive Zero가 의미하는 것은 수많은 종속변수 0이라는 것들이 수학적으로 0이라는 같은 의미를 전달하지만 경제학적인 의미는 다를 수 있으며 이런 차이가 계량분석 모델에 반영되어져야 한다는 것이다.

포아송과 음이항분포 모형은 학생들의 다른 성향으로 인한 Excessive Zero를 반영하지 못하고 있으므로 이러한 문제점을 보완하기 위해 만들어진 모델이 ZINB와 Hurdle 모형이라고 하겠다. <표 III-2>에서처럼 자동차 488,139명의 표본 운전자 중 FY2005에 사고를 내지 않은 운전자는 95.47%에 해당한다. 본 연구에서 정의한 사고빈도는 실제 교통사고 건수가 아닌 접수된 사고건수(Claims)이다.

<표 III-2> 사고빈도

Accident	Freq.	Percent
0	466,040	95.47
1	21,168	4.34
2	883	0.18
3	43	0.01
4	4	0
5	1	0
Total	488,139	100

<표 III-2>에서도 Excessive Zero현상이 나타날 수 있는 이유는 여러 가지가 있을 수 있다. 운전자 중에 한명은 실제로 자동차 보험에 가입해서 운전을 하는 동안 교통사고를 일으키지 않았기 때문에 0을 보고하게 되는 것이지만 다른 운전자는 보험에만 가입해 있고 운전을 하지 않았기 때문에 항상 무사고로 인지될 수도 있을 것이다. 요약하면, Excessive Zero의 문제가 있을 때 확률분포를 가정하고 있는 포아송과 음이항분포(Negative Binomial)로 측정되어지는 무사고의 건수가 너무 적어 실제 데이터를 충분히 설명할 수 없어 위험변수 계수가 정확히 측정되지 못하고 계수의 유의성 검사의 신뢰성도 떨어져 ZINB를 적용해야 한다는 것이다.

$$\Pr(y=0) = B(0) + \{(1-B(0)\}\Pr(0) \qquad (\text{III-28})$$

$$\Pr(y=k;k>0)=\{1-B(0)\}\Pr(k) \tag{Ⅲ-29}$$

식(III-28)은 운전자가 0의 사고빈도를 보고(Claim)할 확률로써, 위 예의 두 번째 성향의 운전자로 항상 0을 보고 할 확률($B(0)$)에 첫 번째 성향의 운전자($(1-B(0))$이지만 사고를 내지 않을 확률($\Pr(0)$)을 더해주는 것이다. 반면 식(III-29)은 운전자가 첫 번째 성향의 운전자($(1-B(0))$이면서 사고를 낼 확률($\Pr(k)$)을 계산한다.

ZINB의 경우 $\Pr(0)$와 $\Pr(k)$ 부분에 음이항분포를 적용하여 만들어진 확률분포이다. 반면 ZIP의 경우 $\Pr(0)$와 $\Pr(k)$ 부분에 포아송 분포를 적용하여 만들어진 확률분포이다. 먼저 ZINB의 식(III-28)과 식(III-29)에 음이항분포를 이용해 아래와 같이 유도되어질 수 있음을 쉽게 알 수 있을 것이다.

$$f(y_i:x_i)=p_i+(1-p_i)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}} \qquad \text{if}\, y_i=0 \tag{Ⅲ-30}$$

$$f(y_i:x_i)=(1-p_i)\frac{\Gamma(y_i+\mu_i)}{\Gamma(\mu_i)\Gamma(y_i+1)}\left(\frac{1}{1+\alpha}\right)^{\mu_i}\left(\frac{\alpha}{1+\alpha}\right)^{y_i} \qquad \text{if}\, y_i>0 \tag{Ⅲ-31}$$

ZIP의 경우에는 $\Pr(0)$와 $\Pr(k)$ 부분에 포아송분포를 이용하여 만들어진 분포이지만 ZINB의 식(III-30)과 식(III-31)의 α에 0을 대입하여 쉽게 유도되어질 수 있음을 알 수 있다.

$$f(y_i:x_i)=p_i+(1-p_i)e^{-\mu} \qquad \text{if}\, y_i=0 \tag{Ⅲ-32}$$

$$f(y_i:x_i)=(1-p_i)\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \qquad \text{if}\, y_i>0 \tag{Ⅲ-33}$$

과대산포(Overdispersion)를 검증할 때 Excessive Zero를 연구자가 주관

적인 판단으로 검증하는 것은 매우 위험하다. Excessive Zero는 Vuong Test로 검증이 가능하며 Vuong Test의 귀무가설이 기각될 경우에는 Excessive Zero의 문제가 있다는 것을 의미하며 이것은 음이항분포가 아닌 ZIP이나 ZINB가 적합하다는 것이다.

ZIP 확률분포의 α에 0을 대입하면 포아송이 유도되며 ZINB의 α와 p에 0을 대입하면 포아송으로 회귀되는 것을 알 수 있듯이 건수변수를 분석하는 일반화선형모형은 포아송분포를 기반으로 개발된 것으로 분포의 적합성에 문제가 제기될 때마다 한계점을 해결하고 보완되고 있는 분석모형이다.

라) 건수변수를 위한 기타 일반화선형모형

건수변수를 분석하기 위해 지금까지 언급한 일반화선형모형 이외에도 수많은 일반화선형모델들이 계량경제학자들에 의해 개발되어지고 있으며 모델들도 복잡해지고 있다. 예를 들어, 보건경제학자들은 환자가 입원했을 경우 얼마나 오랫동안 입원할 지에 대한 예측을 위해 일반화선형모형을 사용하게 되었는데 이 경우 종속변수는 환자 당 입원일수가 될 것이다. 환자가 입원하는 순간 입원일수는 1부터 시작하게 될 것이며 자연히 건수 종속변수에서 0이 제외된다. 예를 들어, 포아송의 경우 0이 관측될 확률이 $\exp(-\mu)$에 의해 계산되기 때문에 0이 없는 건수변수를 분석하기 위해 포아송을 적용하는 것은 바람직하지 못한 결정이라고 하겠다. 이 경우 Zero-Truncated Poisson(ZTP), 또는 Zero-Truncated Negative Binomial(ZTNB)을 이용하게 된다.

ZTP의 확률분포는 포아송 분포가 0이라는 건수변수를 측정하고 있기 때문에 이 부분을 조절한 분포이다. 즉 0이 관측될 확률이 $\exp(-\mu)$이기 때문에 0이 아닌 기타 건수변수가 관측될 확률을 1- $\exp(-\mu)$로 정의하고 포아송 분포에 적용하는 것이다.

$$f(y:\mu|y>0)=\frac{e^{-\mu}\mu^{y}}{\{1-e^{(-\mu)}\}y!} \qquad (Ⅲ\text{-}34)$$

ZTP의 최우추정함수는 다음과 같다.

$$\ln L_{ZTP} = \sum_{i=1}^{n} \left\{ y(\mu) - e^{\mu} - \ln \Gamma(y+1) - \ln[1 - e^{-e^{(\mu)}}] \right\} \tag{Ⅲ-35}$$

식(III-34)에서 보듯이 ZTP는 포아송을 '1-exp$(-\mu)$'로 나누어준 개념이다. ZTNB도 같은 논리에 의해 음이항분포(Negative Binomial)를 0이상의 건수 변수를 볼 확률로 나누어준 개념이다. ZTP와 ZTNB는 자동차 보험에서 사용될 가능성이 희박하여 이상의 구체적인 설명은 다른 기회를 통해 다루기로 하겠다.

Zero Excessive의 경우 ZIP나 ZINB 이외에 Hurdle 모델들을 사용하기도 한다. Hurdle 모형들은 Zero-Alternated Poisson(ZAP)과 Zero-Alternated Negative Binomial(ZANB)로 불리기도 하는데, Hurdle 모형들이 ZIP나 ZINB와 가장 다른 점은 사고를 낼 확률부분인 Binary 부분(y=0)이 건수 부분(y>0)과 별도로 추정되어진다는 것이다. 다시 말해, ZIP나 ZINB는 Binary부분을 Probit으로 추정하고 건수부분에 적용되어지지만 Hurdle의 경우 Binary부분이 Logit이나 Probit으로 추정되고 건수 부분이 별도로 ZTP나 ZTNB로 추정되어 두 부분을 한 번에 추정할 수 있는 편리함이 있다. 하지만 이는 이론상으로 발전된 모형이 아니다.

이 밖에도 건수변수를 이용한 다른 다양한 일반화선형모형들이 개발되고 있다. <표 Ⅲ-3>에서 보이는 것처럼 많은 일반화선형모형들이 이론적으로는 개발되어져 있지만 많은 모델들이 실제 데이터 분석에서 거의 적용되어지지 않고 있으며 이러한 이유로 관련 통계프로그램 개발도 활발하지 못한 실정이다.

<표 III-3> 기타 일반화선형모형

데이터 종류	기타 일반화선형모형의 종류
Cross Section 데이터	Zero-Truncated Poisson
	Zero-Truncated Negative Binomial
	Truncated Poisson
	Truncated Negative Binomial
	Hurdle Models
	Censored Poisson
	Censored Negative Binomial
	Poisson Selection Models
	Negative Selection Models
	Negative Binomial with Endogenous Stratification
	Generalized Poisson Model
	Generalized Negative Binomial 등
패널 데이터	Poisson with Gamma-Distributed Random Effect
	Poisson with Gaussian-Distributed Random Effect
	Negative Binomial with Beta-Distributed Random Effect
	Negative Binomial with Gaussian-Distributed Random Effect 등

2) 연속변수(Continuous Variable)를 위한 일반화선형모형

가) 가우시안(Gaussian)

가우시안 모델은 1800년대에 Karl Friendrich Gauss에 의해 만들어진 모델로 표준분포(Normal Distribution)의 다른 이름이다. 또한 가우시안 모델인 표준분표 모델은 일반적으로 최소자승법(OLS)로 불린다. 즉 일반화선형모형은 최소자승법(OLS)을 포함하는 광의의 모델이다. 이는 가우시안 확률분포를 식(Ⅲ-3)과 같은 일반화선형모형에서 Y의 기본 확률분포인 지수족

(Exponential Family)의 일반 형태로 나타낼 수 있다는 것을 보임으로써 증명할 수 있다.

$$f_y(y;\theta,\phi) = \frac{1}{\sqrt{2\pi\sigma}} exp\frac{-(y-\mu)^2}{2\sigma^2} \tag{Ⅲ-36}$$

즉, 위의 가우시안 확률분포는 다음과 같이 정의할 수 있다.

$$f_y(y;\theta,\phi) = \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2} - \frac{1}{2}ln(2\pi\sigma^2)\right\} \tag{Ⅲ-37}$$

$$f_y(y;\theta,\phi) = \exp\left\{\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}ln(2\pi\sigma^2)\right\} \tag{Ⅲ-38}$$

식(Ⅲ-3)과 식(Ⅲ-38)을 비교할 때 가우시안은 다음과 같은 특성을 지닌 일반화선형모형의 한 종류이다.

$$\theta = \mu = x\beta \tag{Ⅲ-39}$$

$$b(\theta) = \frac{\theta^2}{2} = \frac{(x\beta)^2}{2} \tag{Ⅲ-40}$$

$$b'(\theta) = x\beta \tag{Ⅲ-41}$$

$$b''(\theta) = 1 \tag{Ⅲ-42}$$

$$\alpha(\phi) = \sigma^2 \tag{Ⅲ-43}$$

마지막으로, 가우시안의 최우추정함수는 다음과 같다.

$$\ln L_{Gaussian} = \sum_{i=1}^{n}\left\{\frac{y_i x_i\beta - (x_i\beta)^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}ln(2\pi\sigma^2)\right\} \tag{Ⅲ-44}$$

하지만 가우시안 모델은 2장에서 증명했듯이 자동차 보험요율을 분석하는데 적정한 모델이라고 할 수 없다.

나) Log-Linked 가우시안(Log-Normal) 모델

가우시안 모델의 활용성에 대한 인지도는 매우 높으나 많은 경우 데이터의 분포가 표준분포를 따르지 않고 있어 이용에 한계가 있다. 데이터의 분포가 표준분포가 아님을 알면서도 많은 연구자들은 일반화선형모형이 난해하다는 이유와 통계프로그램의 제약으로 인해 여전히 가우시안 모델을 이용하고 있다. 하지만 이는 잘못된 모수와 표준오차(Standard Error)를 양산하게 된다. 잘못 추정된 모수와 표준오차(Standard Error)로 분석된 결과가 요율정책 등에 반영될 경우 사회적 효용은 감소하게 되기 때문에 적절한 계량모델을 선택하는 것이 매우 중요하다.

이러한 가우시안 모델의 약점을 보완하고자 개발된 모델이 Log- Normal 모델이다. Log-Normal 모델은 가우시안과 같지만 연결함수(Link Function)를 항등함수(Identity Function)가 아닌 로그함수를 사용한다는 점에서 두 모델은 차이가 있다.

로그 연결함수를 이용한 가우시안의 최우추정함수는 다음과 같다.

$$\ln L_{LLG} = \sum_{i=1}^{n}\left\{\frac{y_i \exp(x_i\beta) - \exp(x_i\beta)^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}ln(2\pi\sigma^2)\right\} \qquad (\text{Ⅲ-45})$$

Log-Normal 가우시안의 최우추정함수인 식(Ⅲ-45)은 가우시안 최우추정함수인 식(Ⅲ-44)의 $x_i\beta$ 대신 $\exp(x_i\beta)$이 이용된 것이다.

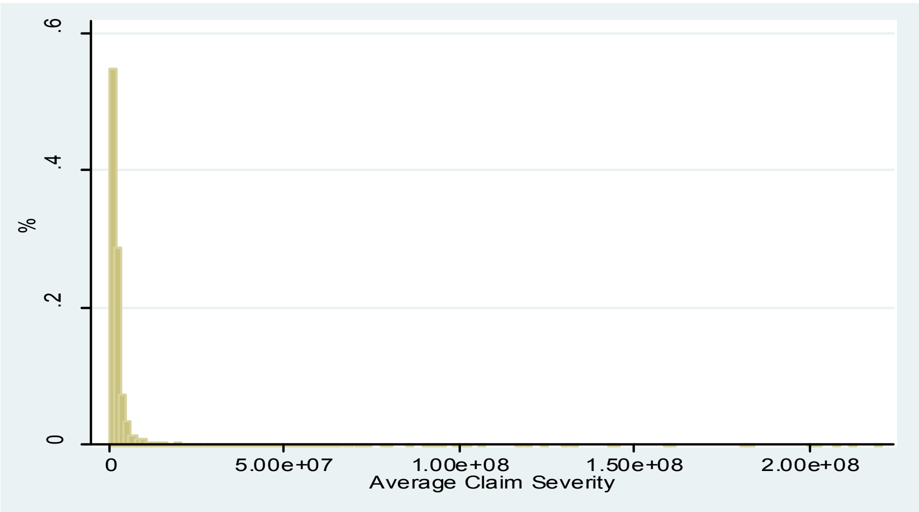
다) 로그노말(Lognormal) 모델

Log-Normal 또는 Log-Linked 가우시안 모델은 가우시안 모델에 연결함수를 항등함수가 아닌 로그함수를 이용한 모델이다. 반면 로그노말(Lognormal) 모델에서는 체계적 성분(Systematic Component)인 $x_i\beta$대신에 반응변수(Response Variable) y에 로그 값을 취한 뒤 가우시안 분석 방법을 이용한다. 즉 반응변수에 로그 값을 취하면 지수분포를 가진 데이터가 표준

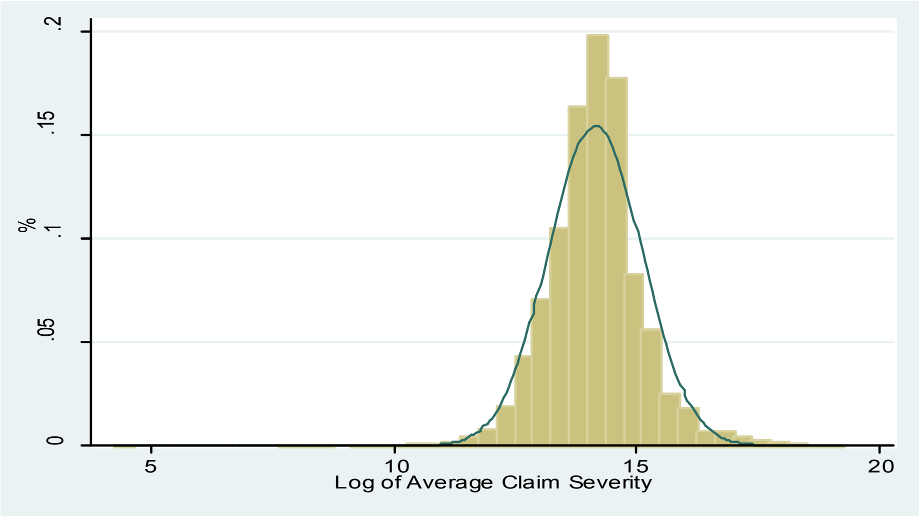
화되는 특성을 이용하여 분석이 매우 쉽다는 장점이 있다. 로그 값을 취한 반응변수가 표준분포라는 가정을 하기 때문에 연결함수는 항등함수를 사용하며 이로 인해 분석의 결과를 해석하는데도 매우 용이하다.

많은 경우 반응변수의 값을 로그 값으로 변환할 때 자료의 분포가 표준분포 모양일 수도 있기 때문에 주의하여야 한다. <그림 Ⅲ-2>는 로그 값으로 전환하기 전 대인 I의 사고심도 분포이다. 기울기가 매우 가파른 지수분포 형태임을 알 수 있다. 반면 <그림 Ⅲ-3>은 대인 I의 삼고심도 값에 로그변환 후의 분포이다. <그림 Ⅲ-3>에서 보이는 것처럼 사고심도에 로그 값을 취하면 표준분포에 매우 가까진다는 것을 볼 수 있다. 이것으로 로그노말(Lognormal) 모델이 유용한 분포라는 것을 알 수 있다.

<그림 Ⅲ-2> 대인I 사고심도: 로그값 변환 전



<그림 Ⅲ-3> 대인I 사고심도: 로그값 전환 후



라) 감마(Gamma) 모델

감마모델은 반응변수가 0 또는 양의 값을 가질 때만 적용 가능한 모델이다. 주로 반응변수가 연속 변수일 때 이용되지만 다양한 범위의 값을 가지는 건수변수에도 적용 가능한 모델이다.

$$f_y(y;\mu,\phi) = \frac{1}{y\Gamma(1/\phi)} \frac{y}{\mu\phi^{1/\phi}} exp(-\frac{y}{\mu\phi}) \quad (Ⅲ\text{-}46)$$

식(Ⅲ-46)은 지수함수 형태로 다음과 같이 나타낼 수 있다.

$$f_y(y;\mu,\phi) = \exp\left\{\frac{y/\mu-(\ln\mu)}{-\phi} + \frac{1-\phi}{\phi} lny - \frac{\ln\phi}{\phi} - \ln\Gamma(\frac{1}{\phi})\right\} \quad (Ⅲ\text{-}47)$$

감마모델에는 일반적으로 역 연결함수(Inverse Link Function 또는

Reciprocal Function) $\mu = \frac{1}{x_i\beta}$ 이 적용된다. 그래서 감마모델의 경우 역연결함수를 특별히 언급하지 않아도 감마모델의 최우추정함수는 역연결함수를 적용하여 다음과 같이 도출된다.

$$\ln L_G = \sum_{i=1}^{n}\left\{\frac{y_i/\mu_i - (-\ln\mu_i)}{-\phi} + \frac{1-\phi}{\phi}lny_i - \frac{\ln\phi}{\phi} - \ln\Gamma(\frac{1}{\phi})\right.$$

$$= \sum_{i=1}^{n}\left\{\frac{y_i x_i\beta - (-\ln\mu)}{-\phi} + \frac{1-\phi}{\phi}lny_i - \frac{\ln\phi}{\phi} - \ln\Gamma(\frac{1}{\phi})\right\} \quad (\text{Ⅲ-48})$$

감마모델에 로그 연결함수를 이용한 로그감마(Log-Gamma) 모델의 경우는 반응변수가 0보다 큰 경우에만 이용된다. 예를 들면, 환자의 입원을 전제하여 입원기간이 반응변수가 되는 경우이다. 역연결함수 대신에 로그연결함수가 사용되어 결과의 해석이 용이하기 때문에 역연결함수 보다는 로그연결함수를 이용한 로그감마모델이 실무에서 주로 이용되기도 한다.

$$\ln L_{Log-G} = \sum_{i=1}^{n}\left\{\frac{y_i\exp(x_i\beta) + x_i\beta}{-\phi} + \frac{\phi+1}{\phi}lny_i - \frac{\ln\phi}{\phi} - \ln\Gamma(\frac{1}{\phi})\right\} \quad (\text{Ⅲ-49})$$

마) 역가우시안(Inverse Gaussian) 모델

역가우시안 모델은 일반화선형모형(GLM) 중 실무에서 사용빈도가 가장 낮은 모델 중 하나이다. 역가우시안의 확률분포는 다음과 같다.

$$f_y(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi y^3\sigma^2}}exp(-\frac{(y-\mu)^2}{2(\mu\sigma)^2 y}) \quad (\text{III-50})$$

그리고 다음과 같이 지수형태로 다시 표현할 수 있다.

$$f_y(y;\mu,\sigma^2) = \exp\left\{\frac{(y-\mu)^2}{2y(\mu\sigma)^2} - \frac{1}{2}ln(2\pi y^3\sigma^2)\right\} \tag{III-51}$$

마지막으로 역가우시안의 최우추정함수는 다음과 같다.

$$\ln L_{IG} = \sum_{i=1}^{n}\left\{\frac{y_i/(2\mu^2) - 1/\mu}{-\sigma^2} + \frac{1}{-2y_i\sigma^2} - \frac{1}{2}ln(2\pi y^3\sigma^2)\right\} \tag{III-52}$$

다. 일반화선형모형의 기타 내용

1) Deviance 및 Residual

Deviance는 모델에 어떤 변수를 포함시킬지의 여부와 모델간 우위를 판정하기 위해서 사용되는 통계량이다. 어떤 변수를 포함시키는 것이 적합한지에 대한 판정은 포화모델(saturated model)과 하부모델(sub-model)의 Deviance차이를 비교하여 이루어진다.

이렇게 변수선택과 모델선택을 위해서는 두 모델 간에 통계량의 차이가 어떠한 분포인지를 아는 것이 중요하다. 두 모델 차이 통계량은 χ^2분포를 따르는 것으로 알려져 있으며 다음과 같은 방법으로 유도되어진다.

먼저 $y_1, y_2, y_{3,} ...., y_n$인 n개의 변수를 가정하자. 각 변수의 모수를 θ_i라고 하면 변수 y_i는 $\mu(\hat{\theta_i}) = y_i,\ \ i = 1,2,3,......., n$ 로 표시할 수 있다. 그리고 모수 추정값 $\hat{\theta_i}$는 canonical 연결함수인 G를 이용하여 $\hat{\theta_i} = G(y_i)$로 나타낼 수 있다.

이때 이들 변수의 음지수우도함수(minus-log-likelihood)의 값은 다음과 같이 표시할 수 있다.

$$l_y(\widehat{\theta_{1,}}, ..., \widehat{\theta_{n,}}, \psi) = \sum_{i=1}^{n}\frac{b(\hat{\theta_i}) - y_i\hat{\theta_i}}{a_i(\psi)} + \sum_{i=1}^{n}c(y_{i,}\psi) \tag{Ⅲ-53}$$

식(Ⅲ-53)에서 $a_i(\psi)=\dfrac{\psi}{w_i}$로 대체하면, scaled deviance D_1^*는 식(Ⅲ-54)와 같이 표현할 수 있다.

$$\begin{aligned} D_1^* &= 2[l_y(\hat{\beta},\psi)-l_y(\widehat{\theta_1},\ldots,\widehat{\theta_n},\psi)] \\ &= \frac{1}{\psi}\sum_{i=1}^{n}2w_i[y_i(\hat{\theta_i}-\theta_i(\hat{\beta}))-b(\widehat{\theta_i})+b(\theta_i(\hat{\beta}))] \end{aligned} \quad (Ⅲ\text{-}54)$$

모델에 대한 deviance인 D_1은 ψD_1^*으로 정의할 수 있다. 이 scaled deviance는 자유도가 n-d-1인 χ^2분포를 따르는 것으로 알려져 있다.

이상의 방법으로 하부모델(submodel)의 deviance를 구하고, 이것을 D_2라고 한다면, 포화모델과 하부모델의 차이를 나타내는 식(Ⅲ-55)의 통계량은 $d-d'$의 자유도를 가진 χ^2분포를 따른다.

$$\frac{D_2-D_1}{\psi} \quad (Ⅲ\text{-}55)$$

따라서 식(Ⅲ-55)의 통계량을 이용하면 포화모델과 하부모델에 차이가 있는지의 여부를 통계적으로 확인할 수 있다.

잔차(residual)는 선택된 통계모형이 자료에 적합한지 여부 및 이상치(outlier)를 판정하는데 사용되는 통계량이다. 선형회귀모형에서와 같이 일반화 선형회귀분석에서도 잔차(residual)를 다음과 같이 정의할 수 있다. 즉, $y_i-\widehat{\mu_i}$, 여기서 $\widehat{\mu_i}=\mu(\theta_i(\hat{\beta}))$이다.

잔차를 분산함수로 정규화 시키면 Pearson residual이 된다. Pearson residual은 식(Ⅲ-56)과 같다.

$$\frac{y_i-\widehat{\mu_i}}{\sqrt{V(\theta_i(\hat{\beta}))}} \quad (Ⅲ\text{-}56)$$

식(Ⅲ-56)로 계산된 잔차(residual)나 일반적인 잔차값을 이용하여 모형의 적합성을 확인한다. 즉 일반화회귀분석으로 추정한 추정값(estimates)과 잔차의 그래프를 보고, 두 값으로 표현된 점들의 분포되는 모양에 따라 설정된 모형이 적합한지 여부를 판단한다.

2) AIC와 BIC

위에서 잔차값을 이용하여 특정 변수의 설명력을 판정하거나 적용 가능한 여러 가지 모델의 우위를 검증하기위해 활용되는 방법론에 대해 설명하였다. 하지만 잔차값을 직접적으로 이용하는 방법은 독립변수를 연구자가 인위적으로 추가 및 삭제하여 어느 정도 조절이 가능하다는 이유로 실증분석에서는 보다 객관적인 판단기준이 필요하게 된다. 이러한 문제를 해결하기 위해 AIC(Akaike's Information Criterion: Akaike 1974)와 BIC(Bayesian Information Criterion: Schwarz 1978)의 방법론이 주로 이용된다.[14)]

$$AIC = -2\ln L + 2q \qquad (Ⅲ\text{-}57)$$

$$BIC = -2\ln L + (\ln N)q \qquad (Ⅲ\text{-}58)$$

L은 최대우도값(Maximized Value of Likelihood Function) 이며 q는 독립변수의 개수로서 AIC와 BIC는 연구자가 임의로 독립변수를 증가시켜 잔차값을 감소시키려는 행위에 대해 패널티(Penalty)를 부과하고 있다. AIC와 BIC의 차이점은 AIC에 비해 BIC가 독립변수를 인위적으로 증가시키려는 행위에 대해 더 무거운 패널티(Penalty)를 부과하는 것이며, AIC와 BIC 모두 작은값의 모델이 더 작은 값인 모델이 선호된다는 것이다.

일반적으로 두 모델간에 AIC와 BIC값의 차이가 2와 6사이의 값이면 값이 작은 모델이 약간 선호된다. AIC와 BIC의 값 차이가 6과 10사이의 값이면 값이 작은 모델이 선호된다고 하고 값 차이가 10 이상이면 값이 작은

---

14) 일반적으로 AIC보다는 BIC가 주로 이용된다.

모델이 매우 선호된다.

<표 Ⅲ-4> AIC와 BIC 값의 비교와 모델의 선택

AIC와 BIC의 절대값 차이	모델의 선택
2 미만	차이 없음
2 초과 6 미만	약간 선호
6 초과 10 미만	선호
10 이상	매우 선호

3. 자동차보험에서 일반화선형모형의 적용방법

가. 자동차보험 통계의 특징

자동차보험 통계는 계약자료와 사고자료가 상호 연계된다. 보험계약자료는 보험계약자가 보험계약을 체결하면서 보험회사에 제공하는 정보와 보험계약자가 보험회사에 납입하는 보험료로 구성되며 보험계약 체결과 동시에 제공된다. 그리고 계약자료의 정보는 보험기간이 끝날 때까지 동일하다. 그러나 보험기간이 종료되고 갱신할 때에는 계약자료의 특징들이 변경된다. 즉 연령도 한살 더 많아지고, 과거에 사고가 없었다면 할인할증률은 1단계 내려가고, 사고가 있었다면 평가된 만큼 할인할증률이 올라간다. 또 자동차보험에 가입한 경력이 1년 지났으므로 가입경력기간도 1년 증가한다. 사고자료는 보험계약자가 계약을 체결한 후 피보험자가 보험기간 동안 사고가 났는지의 여부와 사고가 났다면 얼마의 보험금이 지급되는지의 여부가 기록된 자료이다. 자동차보험에서는 자동차보험 사고의 특징을 파악하기 위하여 계약자료와 사고자료를 서로 연계하여 분석한다.

계약자료와 사고자료를 연계할 때 위험도(사고빈도와 사고심도)를 파악하

기 위한 요소들은 범주형 자료(categorical data)이다. 예를 들어 가입경력을 보면, 자동차보험에 가입한 경력이 1년, 2년, 3년, 4년, 5년 이상 등으로 구분된다. 할인할증도 100%인 사람, 90%인 사람, 80%인 사람 등으로 구분된다. 연령의 경우는 연속형 자료이지만 자동차보험 요율산출에서는 유사한 위험을 가진 연령끼리 그룹을 만든다. 즉 저연령층인 20세 이하연령, 21세부터 30세 이하 연령, 30세 초과연령 등으로 구분하여 범주형 자료(categorical data)를 만들게 된다.

자동차보험 통계에서 통계모형의 독립변수에 해당하는 항목들이 범주형 자료이므로, 자동차보험 통계분석은 범주형 자료분석에 해당되는 방법이 적용된다.

자동차보험 통계모형에서 통계모형의 종속변수에 해당하는 자료에는 연속형 자료와 건수 자료(count data)가 모두 해당된다. 연속형 자료(continuous data)는 보험금 자료가 이에 해당된다. 즉 자동차사고가 나서 지급되는 보험금은 이론적으로 1원부터 무한대까지 가능하며 금액도 1원단위로 계산된다. 위험도의 정도를 나타내는 자동차보험 중에서 사고심도라는 변수가 있으며 사고심도에 해당되는 항목이 연속형 자료인 보험금이다. 건수 자료(count data)로는 사고발생 건수가 있다. 0, 1, 2, 3처럼 보험계약기간 중에 피보험자가 사고를 내지 않았다거나, 1건을 내었거나, 2건을 내었거나, 3건 이상을 내었다는 것을 나타내는 자료로 비연속변수이다.

종속변수 자료의 특성은 보험종목과 담보별로 차이가 있다. 종속변수 자료의 분포는 자료의 연속성과 같은 특성에 따라 차이가 발생한다. 사고발생건수 자료라면, 자료가 포아송분포 또는 음이항 분포의 특성을 보일 수 있다. 보험금 자료는 지수분포, 감마분포 또는 로그노말 등의 특성이 있을 수 있다. 이러한 종속변수의 분포 특성을 반영하여 통계분석을 할 때 자료의 특성을 정확히 파악할 수 있다. 자료의 특성이 파악된 통계모형을 가지고 요율을 산출할 때 보다 정확한 요율산출이 될 수 있다. 자동차보험 요율산출을 할 때 통계모형을 사용하는 것은 이러한 이유 때문이다.

나. 적용방법

본 절에서는 앞서 살펴본 일반화선형모형 이론에 따라 실제 어떠한 방법으로 일반화선형모형이 활용되는지 실무적인 측면에서 살펴보고자 한다.

자동차보험 요율산출에 회귀분석을 적용하는 것은 첫째 통계정리, 둘째 상대도 모형설정, 셋째 모형적합, 넷째 결과의 해석의 절차를 따른다. 통계정리 부분은 자동차보험통계의 특성을 반영하여 통계모형에 적합 시키기 전에 통계를 어떻게 만들 것인가에 대한 내용이며, 상대도 모형의 설정 부분은 요율상대도의 모형가정에 따라 어떠한 모형식이 적정한지에 대한 부분이다. 그리고 모형적합 부분은 ‘통계정리’ 단계 및 ‘상대도 모형설정’단계에서 결정된 내용에 따라 실제 자료를 모형에 적합하는 방법에 대한 내용이다. 자료적합 부분에서는 모형의 설명력을 높이기 위한 여러 가지 방법들에 대한 내용이 다루어진다. 마지막으로 결과의 해석부분은 앞의 ‘모형적합’ 부분에서 추정된 값 $\hat{\beta}$을 변수별 특성에 맞추어 해석하는 것이다.

1) 통계정리

자동차보험 통계자료는 종속변수 자료와 독립변수 자료가 있다. 이중 종속변수 자료는 사고건수와 같이 건수변수(count variable)이거나 1사고 당 보험금과 같이 연속변수(continuous variable)이다.

통계모형에 적용되는 독립변수 자료는 범주형 변수와 연속변수가 혼합되어 있다. 그런데 자동차보험 요율은 군집 또는 범주형 기준에 따라 산출·적용되므로 연속형 독립변수 자료는 범주형으로 변환되어야 한다. 자동차보험에서 범주형 자료의 예를 들면, 연속변수인 가입경력은 최초가입자, 1년 이상 2년 미만, 2년 이상 3년 미만, 3년 이상 4년 미만, 4년 이상으로 구분할 수 있다, 운전자 한정특약에서는 기본가입, 가족운전 한정특약가입, 부부운전 한정특약가입, 1인 운전 한정특약으로 구분된다. 차종의 경우 개인용 자동차보험에서는 소형A, 소형B, 중형, 대형, 다인승 1종, 다인승 2종 등으로 구분할 수 있다. 이처럼 본질적으로 범주형 자료의 특성을 갖는 독립변

수는 자료의 변환이 필요하지 않으나, 운전자 연령등과 같이 연속형 자료는 자동차보험의 운영특성에 맞는 범주형 자료로의 변환이 필요하다. 즉, 자동차보험에서는 자동차보험료가 독립변수의 기준에 따라 일정비율로 산출되므로, 독립변수가 연속형일 경우 명시적인 독립변수별 요율수준을 산출할 수 없는 단점이 있다. 따라서 자동차보험 요율결정시 독립변수 중에서 연속형 변수의 경우도 유사한 위험집단별로 그룹을 나누어 범주형 자료로 변환시키는 것이 유용하다. 범주형 자료로 변환하는 기준은 위험도가 유사한 그룹으로 구분하는 것이 가장 좋다. 그러나 보험회사에서는 자신의 영업측면을 고려하여 임의적으로 연령별 집단을 구분하는 경우도 있다. 위험도가 유사한 그룹으로 구분하는 방법은 연령별로 사고발생률 또는 1사고 당 손해액을 산출한 후 요율산출자가 자신의 판단으로 구분할 수 있다. 요율산출자의 자의성을 배제하기 위해서는 통계모형을 사용하여야 하는데, 사용할 수 있는 통계모형으로는 데이터 마이닝(data mining)에서 사용되는 Tree모델 등을 적용하는 것이다. 현재 우리나라 자동차보험에서 일반적으로 구분하고 있는 연령그룹의 예를 들면 다음과 같다. 동 연령그룹은 요율산출자의 임의적 판단으로 구분한 것이다. 기명피보험자 연령요율의 경우에 연령그룹을 20세 이하, 21세 이상 26세 이하, 27세 이상 50세 이하, 51세 초과 60세 이하, 61세 이상 등의 범주를 정하는 것이다. 본 연구의 목적이 Tree모형을 적용하여 유사위험집단을 구분하는 방법을 개발하는 것이 아니기 때문에 기술통계를 이용하여 연령그룹을 나누었으며, 연령그룹의 적정 위험도를 산출하는 방법을 연구하였다.

2) 모형설정

모형설정은 종속변수에 해당하는 자료에 적합한 분포를 선택하는 단계, 가장 부합된 연결함수(link function)를 결정하는 단계를 거쳐 이루어진다.

먼저 종속변수에 해당하는 자료에 부합되는 분포를 결정하는 단계를 보면, 종속변수를 사고빈도와 사고심도로 구분할지 아니면 사고빈도와 사고심도가 모두 포함된 순보험료법 기준으로 할지 여부에 따라 결정된다. 사고빈

도와 사고심도로 구분하는 경우에는 사고빈도에 부합된 분포와 사고심도에 부합되는 분포가 다르다. 따라서 정교한 분석을 위해서는 사고빈도와 사고심도로 나누어 모형을 설정하는 것이 적합한 것으로 판단된다. 선행연구(Ohlsson과 Johansson, 2004)[15]에서도 일반화선형모형(GLM)을 요율상대도 산출에 적용할 때 표준이 되는 방법은 사고빈도와 사고심도를 구분하는 것이라고 하고 있다.

앞장에서 살펴본 것처럼 사고자료의 특성에 따라 적용될 수 있는 분포는 매우 다양하다. 이중 자동차보험 요율산출에 가장 많이 사용되는 것이 사고빈도에서는 포아송분포(Poisson distribution), 음이항분포(Negative binomial distribution)이고, 사고심도에서는 감마분포(Gamma distribution), 로그노말(Lognormal)이다. 사고빈도와 사고심도를 모두 합한 순보험료 기준으로는 트위디분포(Tweedie distribution)가 있다.

자동차보험 요율산출에 이처럼 다양한 분포를 적용할 수 있지만, 종속변수 자료와 적용하려는 분포의 차이가 크지 않다면 분포차이로 인한 자동차보험 요율위험도 차이는 크지 않다. 이러한 이유 때문에 자동차보험 요율산출에 일반적으로 사용하는 분포는 <표 Ⅲ-5>와 같이 몇 가지 분포로 한정된다. 즉 사고빈도나 사고건수 분포로는 포아송 분포를 사용하고, 평균보험금(1사고 당 손해액) 분포로는 감마분포가 실무에서 주로 활용된다.

이처럼 실무에서는 한정된 분포가 사용되지만, 본 연구에서는 보다 적합한 위험도를 산출하기 위하여 적용가능한 모든 분포를 자료에 적합해보고 가장 부합된 분포모형으로 요율을 산출하였다. 2장에서 설명하였듯이 사고빈도를 설명하기 위해 사용되는 포아송 분포는 평균과 분산이 동일해야 한다는 조건이 충족되어야 한다. 그러나 현실 자료에서는 평균과 분산이 동일한 경우가 많지 않다. 이러한 포아송 분포와 현실자료 간에 발생하는 괴리 때문에 음이항분포가 사고빈도 자료를 설명하는데 더 적합한 분포가 될 수 있다. 그러므로 본 연구에서도 포아송 분포와 음이항 분포 등을 포함한 다양한 일반화선형모형을 모두 통계자료에 적합시키고 이중에서 가장 부합된 통계모형을 선택하여 요율산출에 적용하였다.

---

15) Ohlsson,E.,Johansson,B.(2004), p319

모형에 적합한 연결함수(link function)는 분포에 따라 다양하다. 포아송 분포를 보면 항등(identity), 로그(log), 제곱(square) 연결함수 등이 적용될 수 있다. 감마분포에는 항등(identity), 역(inverse), 로그(log) 연결함수가 적용될 수 있다.

그러므로 여러 연결함수 중에서 통계모형과 자료를 가장 잘 적합 시킬 수 있는 연결함수를 선택하는 것이 중요하다. 즉, 통계자료와 부합된 통계분포를 선택하고, 자료를 가장 잘 설명할 수 있는 연결함수를 최종 연결함수로 결정하는 것이 가장 바람직한 통계분석 방법이다.

그러나 요율산출에서는 이처럼 통계적 절차에 따른 연결함수 선택방법 이외에 고려해야할 요인이 있다. 즉 자동차보험 요율적용 체계이다. 현행 자동차보험 요율체계는 곱셈방식으로 이루어져 있다. 어떤 보험계약자가 자동차보험에 가입하고자 할 경우 보험계약자의 연령위험도, 운전경력, 할인할증 등 모든 위험요소별 상대위험도를 곱하여 보험계약자의 최종 보험료가 결정된다. 따라서 이러한 자동차보험 요율 적용체계를 감안할 때 자동차보험 요율산출, 특히 요율상대도 산출 시 모형은 승산모형이 가장 적합하다. 자동차보험 요율체계를 감안하여 연결함수를 선택할 경우 가장 부합된 연결함수는 로그(log) 연결함수이다. <표 Ⅲ-5>에서와 같이 각 종속변수의 분포에 적합한 연결함수로 가장 많이 사용되는 것이 로그(log) 연결함수이다.

<표 Ⅲ-5> 보험에서 일반화선형모형(GLM)을 적용하는 일반적 모형

종속변수(Y)	사고빈도	사고건수	평균보험금	확 률
연결함수	ln(x)	ln(x)	ln(x)	ln(x)
오 차	포아송분포	포아송분포	감마분포	이항분포
크기모수(Φ)	1	1	추정값	1
분산함수	x	x	x^2	$x(1-x)$ 1)
사전가중치	평대 (위험노출단위)	1	사고건수	1
Offset	0	ln(평대)	0	0

주 : 1)은 시도가 1인 경위며, 시도가 t 이면 동 식은 $x(t-x)/t$로 바뀜.
자료 : Anderson 외5인, p22

3) 모형적합

모형적합은 모형에 포함되는 변수선택, 선택된 변수로 적합시킨 모형의 적합성 확인의 단계로 이루어진다.

모형에 포함되는 변수선택은 우선 자동차보험 요율체계에 사용되는 모든 변수를 모형에 포함시키는 것이 원칙이다. 그러나 현재 자동차보험 요율체계에 포함되어 있음에도 통계적으로 의미가 없는 변수가 있을 수 있다. 이러한 변수를 찾는 방법은 편차치(deviance)를 이용할 수 있다.

앞의 식(Ⅲ-55)에 따르면 포화모델(자동차보험 요율체계에 포함된 모든 변수를 포함한 모델)과 하부모델(일부변수를 제외한 모델)의 각 편차치(deviance) 차이가 χ^2분포를 따르는 것으로 알려져 있다. 따라서 포화모델과 여러개의 하부모델을 만들어 두 모델 간에 deviance 차이의 통계량으로 포화모델과 하부모델 간에 통계적으로 차이가 없다는 귀무가설(H_0)을 검정한다. 만일 두 모델간 차이 통계량이 차이통계량의 분포의 기준 값을 초과하면 두 모델이 동일하다는 귀무가설을 기각하고, 초과하지 않으면 귀무가설을 채택하게 된다. 귀무가설이 채택되면 포화모델과 하부모델에 차이가 없다는 것을 의미하므로, 보다 단순한 모형인 하부모델 자료를 설명하는 모형으로 선택한다. 만일 차이가 있는 것으로 나타나면 하부모델에 추가된 변수가 의미 있는 변수라는 것을 의미하므로 포화모델을 선택하여 자료를 설명하는 것이 좋다. 이처럼 편차치(deviance)를 이용하여 변수를 선택할 수 있다.

편차치(deviance)를 사용하여 변수를 선택한 이후에는 종속변수 또는 Error Term을 잘 설명할 수 있는 분포 및 연결함수를 선택하여 모형을 적합 시키는 단계로 넘어간다. 모형을 자료에 적합한 이후에 해당 모형이 적정한지 확인을 하여야 하는데, 확인방법은 잔차(residuals)를 이용할 수 있다. 즉 잔차와 추정값의 그래프를 그려보고, 그래프 상에서 일정한 패턴이 존재하는지 여부를 파악하여 모형이 잘 적합된 것인지를 판단한다.

만일 추정값과 잔차값의 그래프에 일정한 패턴이 존재하면 해당모형을 수정해야 한다. 이는 일반화선형모형에서 종속변수를 설명하는 분포함수를

잘못 선택했다는 것을 의미한다. 그러므로 분포함수를 변경해야 한다. 즉 분포함수를 포아송함수로 하여 분석하고, 분석한 결과의 추정값과 잔차의 그래프를 그려봤을 때 일정한 패턴이 존재한다면, 분포함수를 음이항 분포로 변환하는 등, 분포함수 변경을 시도해야 한다. 추정값과 잔차값의 그래프의 점들이 일정한 패턴이 없이 무작위적으로 흩어진 상태가 모형이 가장 잘 적합 되었다는 것을 의미한다.

또한 추정값과 잔차의 그래프 점들에서 이상치(outlier)를 발견할 수 있다. 그래프에서 두드러지게 나타난 점은 해당 자료가 이상치(outlier)라는 것을 나타낸다. 이상치는 모형의 적합성을 떨어뜨린다. 이상치가 발견되면 원자료를 살펴서 진정한 이상치 인지, 아니면 이것이 실제로 의미가 있는 자료인지를 판단하여야 한다. 만일 실제자료가 모형의 설명력만 떨어뜨리는 오류자료로 판단되면 모형에서 해당 자료를 제외하고 분석하고, 이상치가 아닌 의미 있는 자료라고 판단되면 모형에 동 자료를 포함하여 분석하여야 한다.

4) 결과해석

현행 자동차보험 요율제도는 곱셈공식에 따라 이루어진다. 개인용 자동차를 예로 들면, 차종에 따른 기본보험료에 가입자가 선택한 각종 특약요율과 가입자의 사고경력에 따른 할인할증률, 가입자의 운전경력 요율이 곱해져 적용보험료가 산출된다.

따라서 이러한 자동차보험요율 적용방법을 반영한 일반화선형모형을 사용하여 요율상대도를 산출하는 것이 가장 좋다. 자동차보험 요율적용방법에 가장 용이한 모형적합 방법은 연결함수(link function)를 로그(log)함수로 정하는 것이다. 종속변수 자료의 형태에 따라 적용될 수 있는 연결함수의 종류는 다양하다. 연결함수를 어떠한 것으로 선택하느냐에 따라 모형의 적합정도에 다소 영향을 줄 수 있으나, 연결함수를 로그로 하는 것이 모형 적합 이후 해석이 가장 용이하다는 이유로 현재까지 요율산출에 일반화선형모형을 적용하는 경우에 로그연결함수를 일반적으로 사용한다.

로그연결함수를 사용하는 경우 해석방법은 다음과 같다. 자동차보험 요율산출에 일반화선형모형(GLM)을 적용하기 위해서는 모형의 독립변수가 범주형(categorical)자료 이어야 한다. 범주형 자료를 통계분석에서 factor자료라고 부른다. factor자료(범주형 자료)는 주요 기준(base)이 되는 범주를 '0'으로 하고 나머지 범주가 '1'이 되도록 더미(dummy) 변수를 이용한다.

만일 독립변수가 '성별'변수라고 가정하면, 분석모형에서 자료의 양이 가장 많은 남성을 기준(base)로 설정하게 되는데, 이 경우 남성은 '0'이 되고 여성은 '1'이 되는 더미(dummy)변수가 된다. 사고건수를 종속변수로 하고 더미(dummy)변수로 변환된 성별을 독립변수로 하여 회귀분석한 결과 여성의 회귀계수(coefficient)가 0.219라고 분석되었다고 가정하자. 이 경우를 식으로 표현하여 보면, $\ln y = \beta_0 + \beta_1 x(\text{성별})$이 된다. 이모형을 남성과 여성으로 나누어 표현해 보면, 남성의 경우는 $\ln y = \beta_0 + \beta_1 x(\text{남성} = 0)$이고, 여성의 경우는 $\ln y = \beta_0 + 0.219_1 x(\text{여성} = 1)$이 된다. 이 식들을 변환하여 성별 계수의 원래 값을 추정하면, 남성은 $y_{\text{남성}} = e^{\beta_0} e^{\beta_1 \times 0} = e^{\beta_0}$이고, 여성은 $y_{\text{여성}} = e^{\beta_0} e^{0.219 \times 1} = e^{\beta_0} e^{0.219}$ 가 된다.

따라서 상대도 값은 기준(base)대비 추정하고자 하는 변수의 상대적 값이므로, $\frac{y_{\text{여성}}}{y_{\text{남성}}} = \frac{e^{\beta_0} e^{0.219}}{e^{\beta_0}} = e^{0.219} = 1.244$가 된다. 이것은 여성의 사고발생률이 남성보다 24.4%높다는 의미이고, 여성의 보험료를 남성보다 24.4%더 높게 받아야한다는 것을 의미한다. 따라서 가정에 따른 일반화선형모형을 적용한 성별 변수의 요율상대도 값은 남성을 '1.000'으로 하고 여성을 '1.244'로 하면 된다.

앞서와 같이 연결함수를 로그함수로 하는 경우 이외에 항등함수를 사용하는 경우도 있다. 항등함수를 사용하는 경우에도 로그함수를 사용하는 경우와 같이 해석이 용이하다. 즉 앞서 예로 들은 모형에서 연결함수로 항등함수를 사용하는 경우의 모형은 $y = \beta_0 + \beta_1 x(\text{성별})$이다. 따라서 여성의 보험료 상대도 값은 $\frac{y_{\text{여성}}}{y_{\text{남성}}} = \frac{\beta_0 + \beta_1 \times 1(\text{여성})}{\beta_0 + \beta_1 \times 0(\text{남성})} = \frac{\beta_0 + \beta_1 \times 1(\text{여성})}{\beta_0}$이

된다.

단일 변수만을 생각할 경우에는 연결함수를 로그로 하거나, 항등함수로 하거나 큰 차이가 없다. 그러나 변수가 두개 이상인 경우에는 차이가 발생한다. 로그연결함수는 변수 간 관계 및 변수내의 관계가 곱(×)의 형태이므로 현행 자동차보험 요율체계와 동일하지만, 항등함수인 경우에는 변수 간 관계가 합(+)의 형태이므로 현행 자동차보험 요율체계와 다소 차이가 있다.

따라서 이상과 같이 일반화 선형모형의 해석은 어떤 연결함수를 선택하여 일반화 선형모형을 적합 시켰는가에 따라 달라진다. 분석결과를 연결함수와 연관시켜서 상대도 값을 산출하고 해석하여야 한다. 앞서 살펴본 바와 같이 log 연결함수를 사용하는 경우에는 변수간의 관계가 곱(×)의 형태라는 것을 의미하고, 항등함수를 연결함수로 사용하는 경우에는 변수사이의 관계가 덧셈(+)이라는 것을 의미한다.

Ⅳ. 일반화선형모형을 이용한 실증분석

1. 모형적용기준

가. 적용보험료 계산식[16]

현행 자동차보험 요율체계에서는 담보별로 적용보험료 계산식에 차이가 있다. 대인배상Ⅰ은 의무담보이므로, 요율요소 적용에 다소 제약이 있어서 기타 담보보다 제한적으로 적용되고 있다. 기타 담보는 임의담보(대물배상[17]은 예외)이므로 많은 적용되고 있기 때문이다.

대인배상Ⅰ과 임의담보의 차이점은 운전자 한정특약의 적용여부이다. 대인배상Ⅰ은 의무담보이므로 미가입자 발생가능성 때문에 운전자 한정특약이 적용되지 않는다. 여기서 운전자한정 특약에는 '1인 운전한정특약', '부부운전한정특약', '가족한정특약' 등 운전자의 범위를 제한하는 특약과 '21세 연령한정특약', '24세 연령한정특약', '26세 연령한정특약' 등 운전자의 연령을 한정하는 특약 등이 있다.

이러한 대인배상Ⅰ과 임의담보의 차이점에 따라 적용보험료 계산식은 담보별로 다음과 같다. 즉 대인배상Ⅰ식에서는 운전자 한정특약요소가 적용되어 있지 않지만, 임의담보의 식에서는 운전자 한정특약요소가 적용되어 있다. 운전자 한정특약 중에서 운전자 연령한정 특약은 기명피보험자의 연령요율과 중복되는 성격이다. 그런데 대인배상Ⅰ에는 운전자 연령한정특약이 적용되지 않으므로 기명피보험자 연령요율의 위험도가 온전하게 나타날 수 있으나 임의담보는 두 가지 변수가 모두 고려되므로 기명피보험자 연령요율에는 운전자 연령한정특약의 요율이 제외된 위험도만 나타난다.

---

16) 본 연구의 분석대상인 개인용자동차보험 및 플러스 개인용자동차보험에 한정하여 논의한다.

17) 자동차손해배상보장법에 따라 2005년 2월 22일부터 대물배상 담보가입(가입한도 1천만원)이 의무화되었다.

<대인배상 I >

적용 보험료	=	기본 보험료	×	관용자동차에 관한 특약요율	×	가입자특성요율 (가입경력요율± 교통법규위반 경력요율	×	(우량할인·불량 할증요율 + 특별할증률)

× 기명피보험자 연령요율 × 특별요율

<대인배상II, 대물배상, 자기신체사고, 자동차상해, 무보험차상해, 자기차량손해>

적용 보험료	=	기본 보험료	×	특약요율 등	×	가입자특성요율 (가입경력요율± 교통법규위반 경력요율	×	(우량할인 · 불량 할증요율 +특별할증률)

× 기명피보험자 연령요율 × 특별요율

적용보험료 계산식 방식은 ‘곱셈(×)’방식이 주를 이루고, 가입자특성요율의 구성항목인 가입경력요율과 교통법규위반 경력요율 및 우량할인 · 불량할증요율과 특별할증은 ‘덧셈(+)’ 방식으로 되어 있다.

현행 적용보험료 방식에 부합하게 일반화선형모형을 적용하기 위해서는 곱셈방식과 덧셈방식을 포함하도록 연결함수(link function)를 설정하여야 한다. 그런데 덧셈방식과 곱셈방식을 모두 포함하는 연결함수를 설정하기가 어렵고, 또한 두 가지 형태가 포함된 모형이 더 효율적이라고 말하기 어렵다. 따라서 본 연구에서는 현재 운영되고 있는 자동차보험 요율체계의 적용보험료 계산방식에도 불구하고 본 연구의 적용보험료 계산방식을 곱셈(×)방

식으로 통일하였다.

이외에도 현재 자동차보험 요율체계에서 사용되고 있는 요율요소에는 성, 지역 등의 요인은 포함되어 있지 않다. 정확한 요율 상대도를 산출하기 위해서는 설명력이 있는 독립변수를 포함하는 것이 중요하다. 따라서 현재 자동차보험 요율체계에서 사용하고 있는 변수 이외에도 보험회사에 집적된 통계 중에서 신뢰성이 확보된 성과 지역변수도 포함시켰다. 변수를 많이 넣을수록 사고발생률(사고건수 자료)을 일반화선형모형에 적합 시킬 때 나타날 수 있는 과대산포(overdispersion)문제도 완화시킬 수 있는 장점이 있다.

현행 자동차보험 요율체계에서 적용하고 있는 특약요율, 특별요율 및 특별할증은 본 연구에 포함시키지 않았다. 본 연구의 목적이 자동차보험 요율 산출법으로 통계모형-특히 일반화선형모형-을 적용하는 방법을 연구하는 것이므로 현재 적용하고 있는 요율체계의 모든 변수를 포함시킬 필요가 없었기 때문이다. 또한 본 연구에서 소개하고 있는 일반화선형모형의 적용방법을 단순히 확대하기만 하면 특약요율 등 기타 변수의 요율상대도도 쉽게 산출할 수 있는 이유도 있다. 특별요율의 경우는 보험종목별로 매우 다양한 특별요율이 있고, 모든 특별요율 종류를 요율산출에 적용하는 것은 매우 비경제적인 측면이 있다. 기존 산출방법에서와 동일하게 중요한 요율변수의 요율을 먼저 산출하고, 마지막에 특별요율을 산출하는 방법이 더 효율적이다. 특별할증은 중요한 할인할증제도의 일부이지만 할인할증제도를 보완하는 제도이므로 본 연구의 요율상대도 산출과정에서 변수로 포함시키지 않았다.

이상의 논의된 방식으로 본 연구의 요율상대도 모형에 사용되는 적용보험료 계산방식은 다음과 같다. 본 연구에서는 일반화선형모형에 포함되는 변수인 ‘가입경력요율제도’, ‘교통법규위반 경력제도’, ‘우량할인·불량할증제도’, ‘기명피보험자연령요율’, ‘성’, ‘지역’을 모두 곱셈(×)식으로 연결한 적용보험료 계산식을 적용한다. 곱셈방식의 적용보험료 계산식은 모든 담보 및 차종에 동일하다.

<연구에 사용되는 적용보험료 계산식 : 모든 담보 동일>

적용 보험료 = 기본 보험료 × 가입경력 요율 × 교통법규위반 경력요율 × 우량할인 · 불량 할증요율 × 기명피보험자 연령요율 × 성 × 지역

나. 요율변수별 범주구분(위험변수)

1) 가입경력제도

가입경력 구분은 회사별로 차이가 있으나, 기본적으로는 네 가지로 구분되어 있다. 즉, '최초가입자 또는 1년 미만', '1년 이상 2년 미만', '2년 이상 3년 미만', '3년 이상'으로 구분되어 있다. 가입경력을 나누는 기준은 현재 자유화가 되어 있으므로, 회사별로 차이가 있다. 따라서 일부 회사들은 3년 이상 가입경력을 더 세분화하여 요율을 산출하고 있다.

가입경력요율은 보험종목(개인용 · 플러스개인용, 업무용 · 플러스 업무용, 영업용 및 이륜차) 및 차종(업무용의 경우는 소유구분)별로 차이가 있다.

본 연구에서는 기존 참조순보험료 요율서의 구분기준과 유사하게 구분하였다. 참조순보험료 요율서에서는 '최초가입자 또는 1년 미만', '1년 이상 2년 미만', '2년 이상 3년 미만', '3년 이상'으로 구분하였지만, 본 연구보고서에서는 '3년 이상 4년 미만'을 추가하고 기준(base)이 되는 범주는 '4년 이상'을 하였다.

<표 Ⅳ-1> 개인용 자동차보험 가입경력요율제도

보험가입 경과기간	개인용·플러스 개인용		
	소형A, B	중·대형	다인승
최초가입자 또는 1년미만	-	-	-
1년 이상 2년 미만	-	-	-
2년 이상 3년 미만	-	-	-
3년 이상	-	-	-

주 : 요율은 회사별로 다르기 때문에 생략하였다.
자료 : 보험개발원, 『자동차보험 참조순보험요율서 신고서』, 2008

가입경력요율의 산출은 개인용자동차보험(플러스개인용 포함)의 담보별로 산출하였다. 기존 자동차보험 요율제도에서는 모든 담보에 공통으로 적용되는 평균적인 가입경력요율을 산출·적용하고 있다. 그러나 전산이 발달한 오늘날에는 담보별로 평균 산출한 결과를 가입경력요율로 사용할 필요가 없다. 평균적인 가입경력요율을 사용하면 계산이 용이하다는 장점이 있으나 위험도에 부합된 요율산출이 되지 않는다는 단점이 있다. 이러한 이유로 본 연구에서는 가입경력요율 산출을 보험종목별, 담보별로 실시하였다.

2) 교통법규위반경력요율제도

교통법규위반경력요율제도는 크게 '할증그룹', '기본그룹', '할인그룹'으로 나뉜다. 다시 '할증그룹'은 '할증 1그룹' 및 '할증 2그룹'으로 나뉜다. 이들 교통법규위반경력요율 적용구분 기준은 교통법규위반 항목 중에서 피해자에게 심각한 영향을 주거나, 위험도 수준에 따라 만들어졌다.

할증그룹에 포함되는 교통법규위반 항목은 '무면허운전 금지', '사고발생시 조치', '주취운전', '신호지시 준수의무', '중앙선 우측통행', '속도제한' 등 중대교통법규를 위반하는 경우에 해당된다. 그리고 기본그룹에는 할증그룹에 포함된 교통법규위반자 이외의 일반교통법규위반자들이 포함된다. 할인그룹은 할증그룹과 기본그룹이외의 모든 경우가 포함된다.

교통법규위반 경력요율은 현재 최고할증률이 약 20%수준이고, 기본그룹

은 0%이다. 할인그룹의 요율은 할증그룹에 교통법규위반 경력요율을 적용한 후 납입된 추가보험료를 할인그룹에 나누어주는 비율로 산출한다.

본 연구에서도 이러한 교통법규위반 경력요율제도의 구분기준과 동일하게 범주를 설정하였다. 즉 ‘할증 1그룹’, ‘할증 2그룹’, ‘기본그룹’, ‘할인그룹’의 네 가지로 구분하였다. 교통법규위반 경력요율의 네 가지 범주 중에서 ‘할인그룹’을 기준수준(base level)로 하여 일반화선형모형을 적용하였다.

<표 Ⅳ-2> 교통법규위반 경력요율 적용기준

구 분		적용대상 법규위반	경력요율
할증그룹	1그룹	1. 무면허운전금지 2. 사고발생시 조치	-
		3-1. 주취운전금지 1회, 3-2. 주취운전금지 2회 이상	-
	2그룹	4-1. 신호지시준수의무, 중앙선우측통행법, 속도제한을 항목구분없이 2회~3회 4-2. 신호지시준수의무, 중앙선우측통행, 속도제한을 항목구분없이 4회이상	-
기본그룹		기타 할증 그룹 및 일반교통법규위반항목, 자동차보험사고가 있는 경우	-
할인그룹		할증 및 기본그룹 이외의 경우	-

자료 : 보험개발원, 『자동차보험 참조순보험요율서 신고서』, 2008

3) 우량할인 · 불량할증제도

우량할인·불량할증요율은 과거 사고유무, 사고내용에 따른 할증점수로 산출된다. 우량할인·불량할증 대상사고는 자동차보험에서 보상책임이 발생하는 사고이고, 청구포기 사고는 제외된다. 평가대상기간은 약 3년으로 한다. 따라서 과거 3년간 자동차보험에서 보험금이 지급된 사고(청구포기 사고는 제외)가 있는 경우의 사고내용점수에 따라 할인할증률이 산출·적용된다.

2009년 현재 기준으로 할증그룹과 할인그룹은 약 24개 그룹으로 구분되어 있다. 그러나 본 연구에서 사용된 FY2005 자료에는 18개의 할인할증

그룹으로 나누어져 있다. 할인할증그룹이 18개에서 24개로 확대된 것은 할인할증그룹별로 위험도에 부합된 요율이 적용될 수 있도록 2007년 9월부터 할인할증요율 적용기준이 변경되었기 때문이다.

따라서 본 연구에서 사용된 통계가 FY2005 자료이므로, 우량할인·불량할증 구분기준도 FY2005 기준으로 하여 통계모형을 적용하였다. 본 연구에서 우량할인·불량할증 변수를 적용할 때 기준수준(base level)이 되는 범주는 할인할증률이 '100%'인 집단이다.

4) 기명피보험자 연령요율 제도

기명피보험자 연령요율은 18세부터 99세까지 연령별로 산출할 수 있다. 그러나 보험회사들은 각 연령별로 요율을 산출·적용하지 않고, 동일한 위험집단별로 범주화하여 요율을 산출·적용하고 있다.

현재 보험회사에서 구분하고 있는 연령그룹 범주는 회사의 보험통계 특성, 경영전략 등이 반영되어 있다. 그러므로 기명피보험자 연령그룹 범주는 회사별로 매우 다양하다.

기명피보험자 연령그룹 범주는 대인배상Ⅰ과 임의담보로 나뉘어 산출되고 있으며, 차종별로도 차이가 있다. 대인배상Ⅰ과 임의담보로 나뉘어 산출한 이유는 대인배상Ⅰ에는 운전자연령운전한정특약이 적용되지 않으나, 임의담보는 동 특약이 적용되기 때문이다. 기명피보험자 연령요율과 운전자연령운전한정특약은 위험도에서 trade-off의 성격이 있다. 즉 운전자연령운전한정특약이 적용되지 않는 담보에서는 기명피보험자 연령요율이 연령별로 큰 차이가 있으나, 동 특약이 적용되는 담보에서는 연령별로 요율의 차이가 크지 않다.

전체보험회사 통계로 산출된 참조순보험료 요율서 기준으로 기명피보험자 연령요율의 구분기준을 보면, '20세 이하', '21~23세', '24~25세', '26~35세', '36~40세', '41~45세', '46~65세', '66세 이상'으로 구분하고 있다. 또한 앞서 언급한 것처럼 운전자한정특약이 적용되었는지 여부에 따라 담보별로 구분하여 기명피보험자 연령요율을 산출하고 있다. 하지만 본 연구에서는

참조순보험료 요율서에서 나누고 있는 기명피보험자 연령요율구분과 달리 하였다.

연령에 따라 자동차 사고빈도가 다르게 나타나며 또한 사고 발생 시 사고심도도 달라질 수 있어 연령은 자동차 보험요율 분석에서 매우 중요한 변수이다.

<표 IV-3>는 보험개발원『2008 자동차보험 참조순보험요율서 신고서』를 참조하여 만든 개인용자동차보험(플러스 개인용 포함)의 기명피보험자 연령별 요율 구분이다.

<표 Ⅳ-3> 개인용자동차보험(플러스 개인용 포함) 기명피보험자 연령요율 구분

구 분		20세 이하	21~ 23세	24~ 25세	26~ 35세	36~ 40세	41~ 45세	46~ 65세	66세 이상
대인Ⅰ	소형A · B, 중형, 대형	-	-	-	-	-	-	-	-
	다인승1 · 2종	-	-	-	-	-	-	-	-
임의	소형A · B, 중형, 대형	-	-	-	-	-	-	-	-
	다인승1 · 2종	-	-	-	-	-	-	-	-

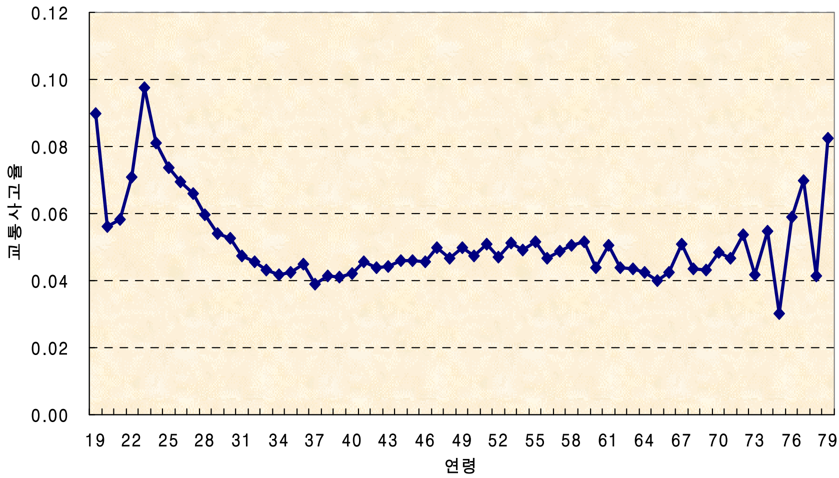
주 : 요율은 회사별로 다르기 때문에 생략하였다.
자료 : 보험개발원,『자동차보험 참조순보험료율서 신고서』, 2008

연령과 교통사고 빈도 및 심도에 관한 연구는 활발히 이루어지고 있다. 기존의 연구들은 대체적으로 젊을수록 사고빈도와 심도가 높다는 결론을 보여주고 있는 것처럼 현재 우리나라 보험사들이 적용하는 보험료는 나이가 들수록 감소한다. Laflamme 외 3인(2006)에 따르면 저연령 운전자는 상대적으로 위험한 운전습관으로 인해 사고빈도와 심도가 높을 수 있다고 지적하고 있다. 대체적으로 저연령 운전자들은 운전미숙, 높은 위험감수경향, 야간운전, 음주운전 등의 요인으로 인해 사고위험도가 높다고 증명되고 있다.

젊은 운전자들이 사고위험도가 상대적으로 높다(David 와 Morrisey, 2001)는 일관된 결론에 반해 고령 운전자의 사고위험도에 대한 의견들은 상이하다. 하지만 일반적으로 고령 운전자들이 중간연령층의 운전자보다 교통사고 위험도가 높다는 결론이다(Delucia 와 Mather, 2006; 기승도, 2009). 고령 운전자들과 높은 교통사고 위험도를 설명하는 가장 논리적인 이론은 외삽법 (Extrapolation)이다. 즉, 운전자는 운전 중 여러 가지 계기판을 확인하고 도로 상태를 통해 위험인지를 해야 하는데 이러한 능력들이 나이가 들수록 저하된다는 것이다. 또한 위험의 인지가 행동으로 전달되기까지의 시간은 나이와 비례하여 소요되는 것이다.

아래 <그림 Ⅳ-1>은 본 연구보고서에서 사용된 데이터를 이용해 연령과 교통사고율의 관계를 도시한 것이다.

<그림 Ⅳ-1> 운전자 연령과 교통사고빈도: 대인배상 I



<그림 Ⅳ-1>에서 보이는 것처럼 사고빈도는 젊은 운전자 층에서 매우 높게 나타나고 있으며 나이가 40세에 이르기까지 지속적으로 감소하는 것을 볼 수 있다. 이후 연령층에 해당하는 운전자들(40~50대)은 30~40세 운전자

보다는 다소 높은 사고빈도를 보이고 있지만 큰 변화 없이 완만한 사고빈도를 보이며 65세 정도부터는 다시 증가하는 모습을 보이고 있다. 그래프의 양 끝에 해당하는 젊은 층의 운전자(23세 이하)들과 노령층(75세 이상)의 운전자들은 다소 혼란스러운 그림을 보여주고 있다고 생각 할 수 있다. 그러나 이것은 해당 연령층에 포함된 표본수가 상대적으로 작거나 운전거리가 기타 연령대의 운전자에 비해 상대적으로 짧은 등의 이유로 생겨나는 현상일 수 있다. 무엇보다 <그림 Ⅳ-1>은 계량적인 모델 분석에서 나온 결과물이 아니라 단순한 통계치에 기반을 둔것으로 연령별 사고빈도를 나타내는 그림이다. 연령별 자동차사고에는 다른 다양한 위험요소의 영향이 내재되어 있을 수 있으므로 위험도에 영향을 줄 수 있는 다양한 정보가 동시에 검토되어야 한다.

이러한 연령별 위험도 특성을 감안하여, 본 연구에서는 연령구분을 다섯개의 그룹으로 나누었다. 즉 '25세 이하', '26세부터 29세', '30세부터 39세', '40세부터 64세', '65세 이상'으로 구분하였다. 보험회사의 마케팅 전략, 동질적 위험집단 등을 감안하여 기명피보험자 연령을 구분하는 것이 일반적이다. 자동차보험 요율상대도를 산출하는 법을 설명하는 것이 본 연구의 목적이므로 현재 보험회사들이 하고 있는 구분 방법대로 연령구분을 할 필요가 없었다. 다만, 앞서 살펴본 바와 같이 저연령자와 고연령자의 위험도가 주요 연령층에 비하여 높다는 연구결과들이 있으므로, 저연령, 고연령 및 주요 연령의 위험도 특징을 살펴보고 분석의 편리성을 확보하기 위하여 세대별 연령구분 방법을 본 연구에서는 사용하였다. 기명피보험자 연령변수에서는 30대(30세부터 39세)를 기본수준(base level)으로 설정하였다.

5) 성별 및 지역요율 제도

성별 기준과 지역구분 기준은 현재 자동차보험 제도에서는 사용되지 않고 있는 변수이다. 그러나 이들 기준에 따라 위험도를 평가해보면, 변수의 범주별 위험도 차이가 있는 것으로 알려져 있다(기승도, 2009). 위험도에 부합된 요율을 적용해야한다는 원칙에 따르면, 성과 지역요소도 요율산출의

변수로 포함시킬 수 있다.

성별 기준은 ‘남성’, ‘여성’으로 범주를 구분하였다. 성별 기준에서 기준(base)이 되는 범주는 ‘여성’으로 하였다.

지역기준은 ‘광역시’로 구분하였다. 지역별 범주구분은 ‘광역시 기준’뿐 아니라 ‘시군구별 기준’으로 나눌 수도 있다. 이들 범주를 나누는 기준은 동일한 위험도를 가진 집단별로 묶는 것이다. 본 연구는 지역별 요율차등화 제도를 도입하는 것을 검토하는 것이 아니라, 일반화선형모형(GLM)을 요율 산출에 적용하는 방법을 연구하는 것이다. 그러므로 지역별 범주구분을 광역시도 기준으로 광범위하게 설정하였다. 지역의 구분기준인 광역시의 구분 중에서 서울을 기준(base)으로 하여 일반화 선형모형을 적용하였다.

<표 Ⅳ-4> 지역범주 구분 기준

서울시, 부산시, 대구시, 대전시, 광주시, 강원도, 충청북도, 충청남도, 경상북도, 경상남도, 전라북도, 전라남도, 제주도

2. 통계자료 및 통계모형

가. 통계자료

통계자료는 업계 전체자료 중 일부를 표본 추출한 것이다. 보험개발원에 집적된 FY2005의 총 자료 중에서 임의적으로 표본추출(Random Sampling)한 자료(FY2005에 약 50만개)를 사용하였다. 3가지 통계 형태 중에서 계약자료와 사고자료가 가장 일치하는 증권년도기준(policy year basis)으로 자료를 추출하였다.

<표 Ⅳ-5> 통계자료의 기준

기 준		세부분류
독립변수	성	남, 여
	연령	실제 기명피보험자 연령
	가입경력	최초가입자~1년 미만은 '1', 1년 이상 2년 미만은 '2', 2년 이상 3년 미만은 '3', 3년 이상 4년 미만은 '4', 4년 이상 '5'
	지역	'광역시도' 대분류 기준
	법규위반	현행기준 참조순보험료 요율서 기준
	할인할증율	40% ~ 200%
종속변수	대수	통계기간 중 평균유효대수로 추출
	사고건수	담보별 사고건수
	보험료	수입보험료
	손해액	담보별 손해액, 2008년 3월 말 현재 평가 OS[18]

주 : 1) 개인용 및 플러스 개인용 자동차보험에 가입한 경우의 통계이다.
2) 할인할증률은 45%를 제외하고 5%단위에서 절하하였다.[19]
3) 연령은 연구자가 임의로 동일위험 집단 기준으로 묶어서 범주형 자료로 변환한 것이다.

통계추출 대상 보험계약은 개인용 및 플러스개인용으로 하였다. 개인용 및 플러스 개인용이 전체자동차보험에서 차지하는 비율이 평균유효대수 기준으로 70.4%로 매우 높고, 보험회사들이 자동차보험 요율상대도를 산출할

18) 손해액은 지급보험금과 지급준비금(OS)으로 구성되어 있다. 이중 지급준비금은 FY2005에 발생한 사고로 지급되어야할 보험금을 적립한 것이다. 사고발생당시 적립한 보험금은 이후 매년 치료비등으로 지급되기 때문에 지급준비금은 어느 시점에서 평가하느냐에 따라 자료가 달라진다. 시간의 경과에 따라 변하는 지급준비금의 특성 때문에 가급적 최근통계 기준으로 지급준비금을 평가하는 것이 최선의 적정지급준비금을 정하는 것이다. 이러한 이유로 지급준비금을 2008년 3월말 현재 기준으로 하였다.

19) 통계를 추출하던 당시(FY2005)의 할인할증요율제도는 최저 할인할증률 40%, 45%, 50%, 이후 10%씩 증가하여 최대 200%의 체계였다. 따라서 본 연구도 통계추출당시 요율체계를 따라서 45%를 제외하고 5%단위에서 절하하여 분석하였다.

때 가장 관심을 가지고 있는 보험종목이기 때문이다. 그리고 현재 자동차보험에서 운영되고 있는 모든 담보를 분석대상으로 하였다. 개인용 자동차보험에서 운영되고 있는 담보는 대인배상Ⅰ, 대인배상Ⅱ, 대물배상, 자기신체사고, 자기차량손해, 무보험차상해가 있다.

독립변수에 해당되는 요율변수는 현재 자동차보험에서 운영되는 변수 및 사용되고 있지는 않지만 분석에 필요한 것을 대상으로 하여 독립변수를 선택하였다. 즉, 분석에 사용되는 자료는 독립변수로 ‘성’, ‘연령’, ‘가입경력’, ‘지역‘, ’법규위반 유형‘, ’할인할증률‘이다. 이들 변수자료는 전산에 등록되어 있는 원 자료 수준으로 추출된 것이다. 이 자료는 일반화선형모형을 사용하여 요율상대도를 산출할 때 범주형으로 변환되어 사용된다. 이들 독립변수자료는 앞서 추출기준인 개인용 및 플러스 개인용과 담보별 구분에 따라 추출되었다.

종속변수는 2가지로 구분하여 추출하였다. 즉 종속변수는 사고빈도(사고발생률)와 사고심도(1사고 당 손해액)로 구성되어 있다. 사고빈도는 사고건수 자료이고, 사고심도자료는 1사고 당 지급된 보험금자료이다. 이들 종속변수자료도 앞의 독립변수처럼 개인용 및 플러스개인용과 담보별 기준에 따라 추출된 것이다.

앞의 기준으로 추출된 자료 중 논리적 모순이 있는 자료를 제거하는 방법으로 자료를 정리하였다. 예를 들면, ‘보험료가 없는 자료’, ‘대인배상Ⅰ의 보험금이 1억원을 초과하는 자료[20]’ 등을 분석 자료에서 배제하거나 논리에 맞도록 수정하였다.

나. 통계모형

본 연구의 분석에 사용된 모형은 다음과 같은 일반화선형모형이다. 일반화 선형모형은 선형모형을 확장한 것으로 종속변수는 독립변수들의 선형결합으로 설명할 수 있다는 가정에 따라 만들어진 것이다. 이것을 행렬식으로

---

20) 2005년 자동차보험 대인배상Ⅰ의 보상한도가 1억원이므로, 1억원을 초과한 손해액 자료는 오류가 있는 것이다.

표현하면 다음의 식 (Ⅳ-1)과 같다.

$$Y = X^T\beta + \epsilon,\ \epsilon \sim N(0, \sigma^2) \qquad (\text{Ⅳ-1})$$

여기서 Y 는 사고발생률(이하 '사고빈도'라고 한다) 또는 1사고당 손해액(이하 '사고심도'라고 한다)이다. β는 계수 벡터이다. X 는 주어진 보험제도 변수를 요소로 하는 벡터이다. X에는 앞에서 언급한 '나. 요율변수별 범주구분'에서 제시한 자동차보험 제도인 가입경력제도, 교통법규위반경력제도, 우량할인・불량할증제도, 기명피보험자의 연령요율제도, 성별 및 지역요율제도가 해당된다. X는 범주형 자료(Categorical Data)이어야 하기 때문에 모든 독립변수(행렬 X의 모든 열 항목)를 더미(Dummy)변수로 변환하여 분석하였다.

본 연구 모형을 분석하기 위해 사용한 통계분석 프로그램은 R[21]통계 언어 또는 STATA이다. 세부 통계분석은 '1단계는 선택된 변수로 모형을 적합시키는 과정', '2단계는 그래프 등[22]을 이용하여 모형의 적합성, 가정의 타당성 등을 확인하고 모형을 수정', 그리고 '3단계는 채택된 모형과 변수를 해석'하는 단계로 하였다.

---

21) R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

22) 그래프를 이용하여, Residuals와 Fitted값의 관계, Q-Q Plot을 이용한 Normality 및 Leverage 확인 등을 하였다.

3. 실증분석결과

가. 대인배상Ⅰ 담보

1) 모형선택

일반화선형모형을 이용하여 위험변수의 상대위험도를 계산하는데 있어서 중요한 단계중 하나는 데이터와 분포를 가장 잘 설명하는 모델을 선택하는 것이다. 예를 들어, 사고심도에 영향을 주는 위험변수의 상대위험도를 계산하기 위해 가우시안(Gaussian), 감마(Gamma), Log-normal 등 매우 다양한 모델을 이용할 수 있다. 많은 경우 모델들의 상대적 적합성을 충분히 분석하지 않고 분석자의 주관적인 판단으로 모델을 선택하는 경우가 있다. 그러나 잘못된 보험요율산출 및 적용으로 보험사는 경상수지 또는 손해율 악화에 직면할 수 있으며, 정확한 위험도에 따라 요율을 부과하지 않으면 소비자의 효용이 감소되고 사회적 비용이 증가되는 부정적 측면이 발생할 수 있다.

가) 사고빈도

사고빈도와 같은 지수형 분포를 가진 건수변수가 종속변수일 경우 포아송, 음이항분포, CPB, ZINB, ZIP 등 다양한 계량모델들을 적용하여 위험변수들의 상대위험도를 산출할 수 있다. 하지만 위험변수들의 정확한 상대위험도와 유의성을 측정하기 위해 가장 적절한 계량모델을 선택하는 것은 요율분석에서 매우 중요한 단계이다. 잘못된 모델 선택은 잘못된 요율부과를 양산하며 나아가 보험사의 수익악화와 보험계약자의 효용성 감소로 귀결된다.

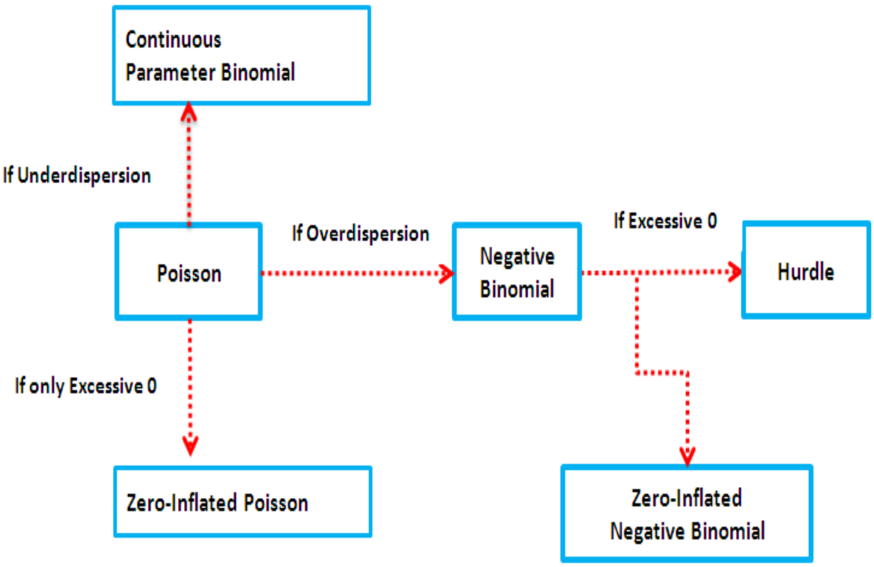
제 3장에서 언급한 것처럼 건수변수를 설명하는 최적의 계량모델을 선택하기위해 크게 두 가지 테스트를 해야 한다. 하나는 과대산포 또는 저산포 테스트이며 다른 하나는 Excessive Zero 테스트다.

먼저 포아송 모델을 기본으로 하여 과대산포의 문제가 존재하면 음이항분포를 선택하며 반대로 저산포의 문제가 있으면 CPB를 선택한다. 과대

산포의 문제와 함께 Excessive Zero의 문제가 있으면 ZINB나 Hurdle모델을 선택한다. 마지막으로 과대산포나 저 분산의 문제는 없으나 Excessive Zero의 문제가 있으면 ZIP모델 선택이 적합하다.

보통의 통계 프로그램들에서는 α의 로그값을 측정하고 이를 다시 α값으로 변환하여 귀무가설(Null Hypothesis) $H: \alpha = 0$을 테스트하도록 되어있다. <표 Ⅳ-6>는 과대산포 테스트 결과이며 귀무가설(Null Hypothesis) $H: \alpha = 0$이 기각되어 과대산포가 존재한다는 것을 보여준다. 이 테스트 결과로 인해 <그림 Ⅳ-2>에서 보이는 것처럼 ZIP와 CPB를 제외한 음이항분포, ZINB, 그리고 Hurdle 모델 중 하나를 선택하는 것이 적합하다.

<그림 Ⅳ-2> 사고빈도의 경우 계량모델



<표 Ⅳ-6> 과대산포 테스트

구 분	Coefficient	Std. Err	Z
lnalpha	-0.306	0.754	-
alpha	0.737	0.056	13.16
Likelihood-ratio test of alpha=0: chibar2(01)=262.06, Prob>=chibar2 = 0.000			

다음으로 Excessive Zero의 문제가 존재하지 않으면 음이항분포가 분석모델로써 적합하다고 판단되며, Excessive Zero문제가 존재하면 ZINB나 Hurdle모델 중 하나를 선택하게 되는 것이다. 위에서 언급한 것처럼 Excessive Zero가 존재할 경우에 포아송이나 음이항분포로 요율분석을 하게 되면 위험변수 계수와 표준오차(Standard Error)가 잘못 측정된다. Excessive Zero가 존재하는지의 여부는 Vuong 테스트를 통해 가능하다.

<표 Ⅳ-7> Vuong 테스트: Excessive Zero 테스트(대인배상I)

Vuong 테스트 적용변수	Vuong 테스트 결과 (ZINB vs. NB)
성별	z=0.46(Pr>z=0.32)
차종	z=0.99(Pr>z=0.16)
성별, 가입경력	z=1.38(Pr>z=0.08)
성별, 가입경력, 차종	z=1.39(Pr>z=0.08)
기타 변수들	z<1.96

Vuong 테스트의 결과는 Inflate 부분에 어떤 위험변수를 포함시키는지에 따라 다른 결과를 얻을 수 있어 다양한 ZINB의 모델형태(Specification)에 Vuong 테스트를 적용해야 한다. 다양한 위험변수의 조합을 이용하여 Vuong 테스트를 실시한 결과 Excessive Zero가 없었으며 이는 ZINB 보다

는 음이항분포가 더 적합한 모델이라는 것을 증명한다. 결과적으로, 위 두 가지의 테스트로 연구보고서에 사용된 대인배상 I 의 자동차 사고빈도에는 음이항분포가 가장 적합한 모델을 이라는 것을 알 수 있다.

이외에도 통계적으로 음이항 분포가 자동차사고 빈도를 설명하는 적합한 통계모형이라는 것을 확인하기 위해 AIC(Akaike's Information Criteria)와 BIC(Bayesian Information Criteria)를 계산하는 방법을 사용할 수 있다. AIC와 BIC를 이용할 때에는 두 계산된 값의 차이를 비교하는 방법이 이용된다. 즉 두 모델간 AIC 혹은 BIC의 절대값의 차이가 2 이내이면 두 모델이 큰 차이가 없다는 것을 의미한다.

<표 Ⅳ-8>는 건수 반응변수를 분석하기 위해 가장 기본이 되는 포아송, 음이항분포(NB), 그리고 ZINB의 AIC와 BIC를 비교한 것이다. 이 표에 따르면 AIC와 BIC의 값이 포아송이 음이항분포나 ZINB보다 크다는 것을 알 수 있다. 이것은 포아송 모델이 적절하지 못하다는 것을 의미한다. 음이항분포의 AIC 값은 ZINB보다 작지만 그 차이가 2내외여서 AIC로는 두 모델간 적합성을 평가하지 못한다. 그러나 음이항분포의 BIC는 ZINB보다 10 이상 작으므로 음이항분포가 더 적합한 모델이라는 것을 의미한다.

<표 Ⅳ-8> AIC, BIC를 이용한 일반화선형모델 비교: 사고빈도

	Poisson	NB(1)	ZINB(2)	(2)-(1)
AIC	185405.3	185153.0	185155.0	2
BIC	185971.3	185730.1	185743.2	13.1

나) 사고심도

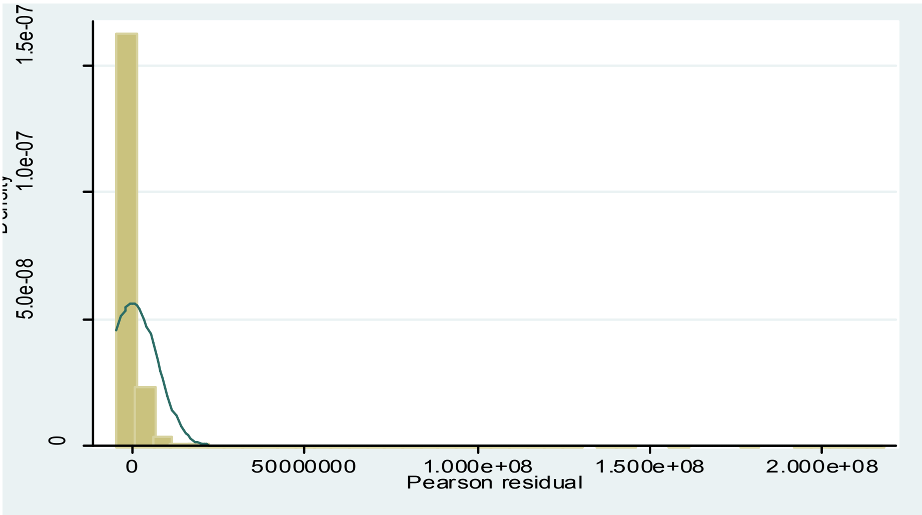
본 연구에서는 편의상 사고빈도를 위한 모델 선별작업을 먼저 소개했지만 실무에서는 사고심도를 분석하기위한 모델선별작업을 하면서 데이터를 먼저 정리하는 것이 바람직하다. 사고빈도의 경우 위험도에 영향을 주는 위험요소의 위험 상대도를 분석하기 위해 적합한 모델을 선별하기 위한

테스트들을 실시했다. 하지만 과대산포(Overdispersion)와 Excessive Zero를 테스트하여 사고빈도 분석을 실시하기 전에 사고심도에 영향을 주는 위험요소의 위험상대도를 추정하기 위한 모델선별 작업이 선행되어야 한다. 사고심도의 경우에는 모델 선별작업을 통해 이상치들을 제거할 수 있는 작업들이 동시에 이루어지는데 이상치들은 과대산포(Overdispersion)에 영향을 주기 때문이다.

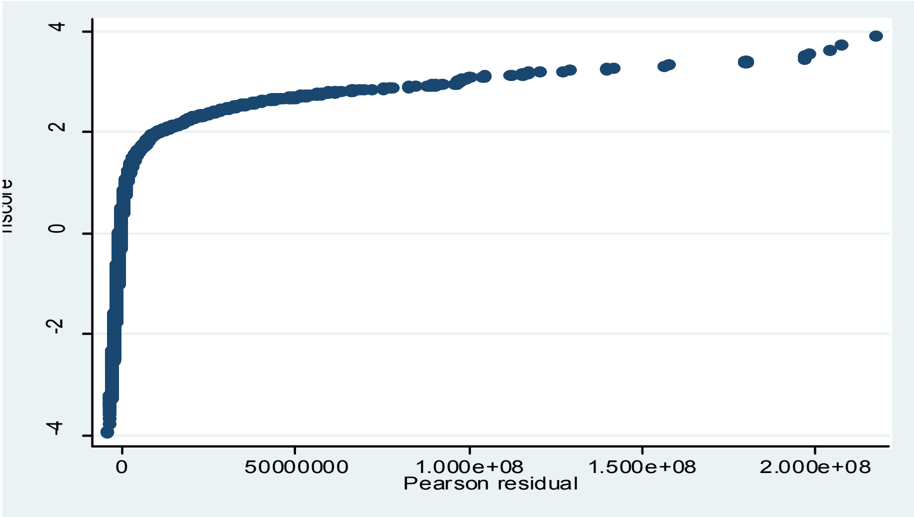
그리고 적합한 모델들을 선별하기 전에 개별적인 모델을 테스트하여야 한다. 예를 들어, 가우시안 모델이 왜 자동차보험 요율분석에 적합하지 않은지를 확인해 볼 필요가 있다. <그림 Ⅳ-3>는 가우시안을 이용한 일반화선형모형으로 회귀분석을 한 뒤 Pearson 잔차항을 이용해 그린 히스토그램이다. 반응변수의 정규분포를 가정하는 가우시안이 자동차사고 심도의 데이터를 잘 설명한다면 잔차항의 히스토그램은 정규분포에 가까워야 한다. 그렇지만 그림에서는 잔차항의 분포는 정규분포와 다르다.

정규성을 확인하기 위해서 임의 추출된 정규분포자료와 가우시안의 잔차를 비교하는 방법도 있다. 그 비교 결과는 <그림 Ⅳ-4>와 같다.

<그림 Ⅳ-3> 가우시안 모델 검정: Pearson 잔차항 히스토그램

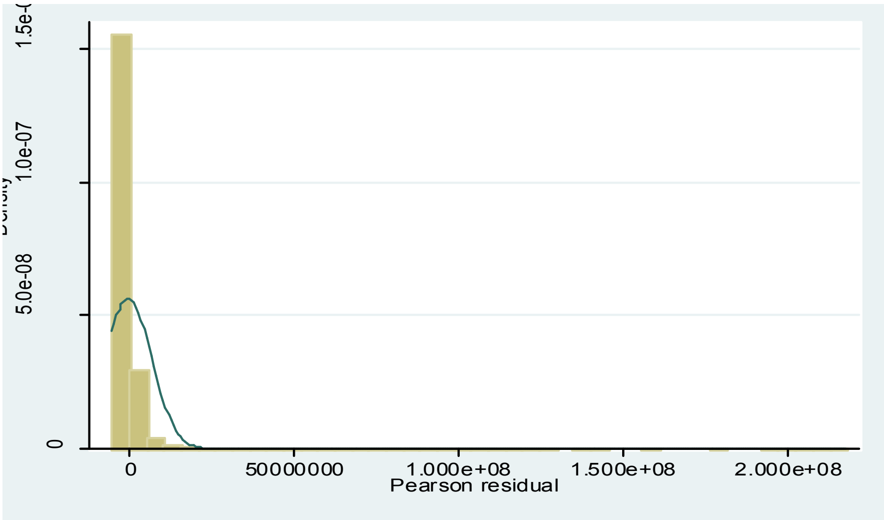


<그림 Ⅳ-4> 가우시안 모델 검정: Pearson 잔차항과 임의데이터 비교



약 회귀분석에 이용된 자동차사고 심도가 가우시안 분포를 따른다면 <그림 Ⅳ-4>의 점들이 대각선을 중심으로 분포되어야 한다. 그러나 그림에서 보듯이 점들이 대각선으로부터 상당한 거리를 두고 떨어져 있는 모습이다. 이것은 가우시안 모델이 자동차사고 심도를 분석하기에는 부적합하다는 것을 증명한다.

<그림 Ⅳ-5> Log-Linked 모델 검정: Pearson 잔차항 히스토그램

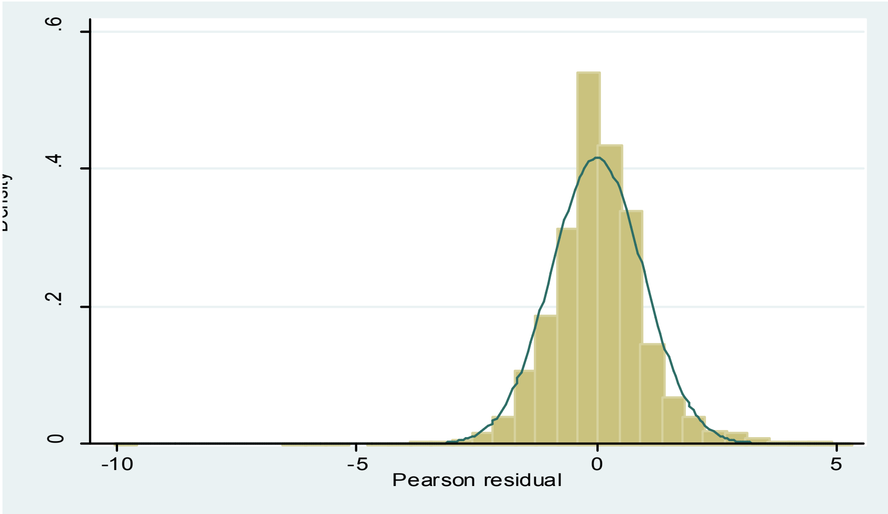


Log-Linked 모델의 경우에도 위 <그림 Ⅳ-5>에서처럼 가우시안의 잔차항과 거의 같은 분포이며, 정규분포와는 다른 모습이다. 이것으로부터 Log-Linked 모델 역시 자동차사고 심도를 연구하기에 적절하지 못한 모델이라는 것을 알 수 있다.

다음으로 Log-Linked 모델과 반대로 반응변수 값을 로그로 전환 한 뒤 가우시안과 같은 회귀분석을 하는 로그노말(Lognormal) 모델을 검증해 보았다. 위 두 모델에서처럼 먼저 로그노말(Lognormal) 모델을 이용해 회귀분석 한 뒤에 잔차항을 이용해 히스토그램을 그려보았다.

<그림 Ⅳ-6>에서는 보이는 것처럼 잔차항이 0을 중심으로 대칭이다. 이것은 로그노말(Lognormal) 모델이 자동차사고 심도를 잘 설명하는 모델일 수 있음을 방증하는 것이다.

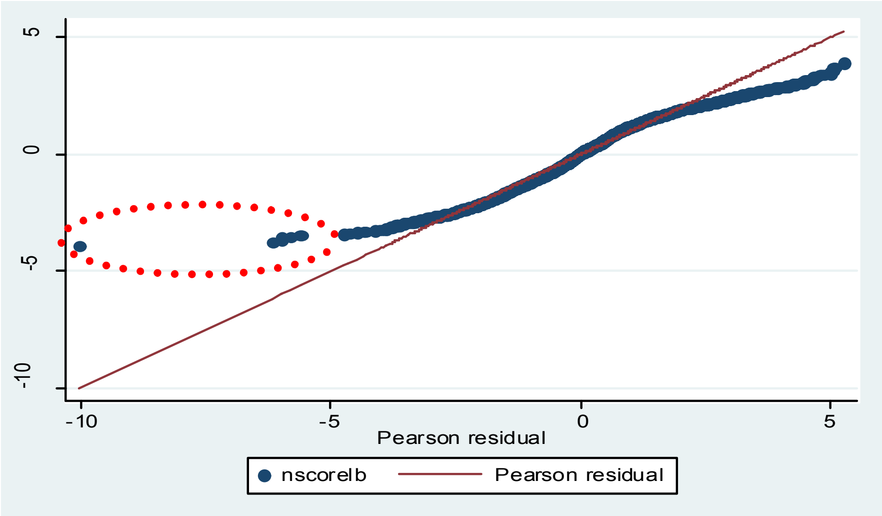
<그림 Ⅳ-6> Lognormal 모델 검정: Pearson 잔차항 히스토그램



가우시안 분포의 적합성 테스트에서 사용한 방법과 동일한 방식으로 좀 더 구체적인 테스트를 해 보았다. 즉 실제자료를 로그노말(Lognormal) 모델로 회귀분석하여 산출된 잔차항과 정규분포를 따르는 임의생성 자료를 비교하여 보았다. 이때 실제자료와 임의 추출 자료의 표본 수는 동일하다. 가우시안 분포의 테스트 결과와 다르게 대부분의 점들이 대각선을 중심으로 분포되어있다. 이것은 로그노말(Lognormal)을 대체할 만한 적정한 모델이 없을 경우 로그노말(Lognormal) 모델이 자동차사고 심도의 요율분석을 위해 매우 적정한 모델이 될 수 있다는 것을 의미한다.

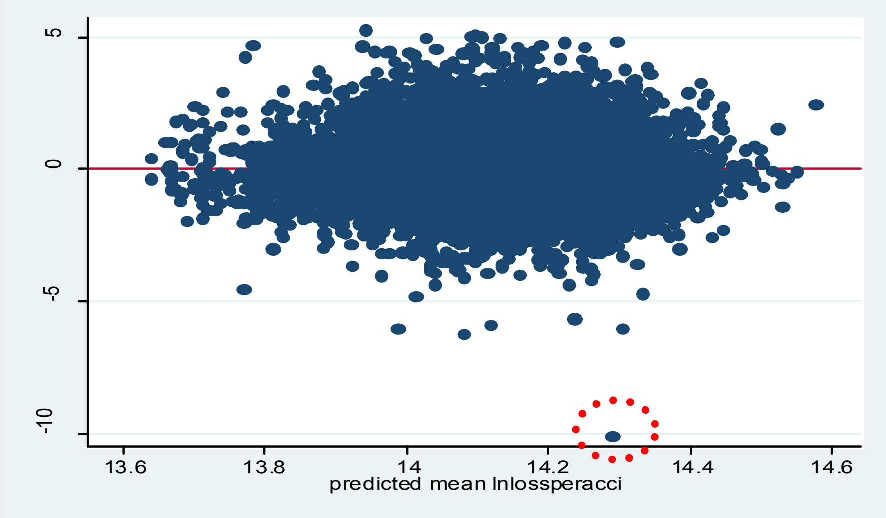
모델 선별작업을 하는 동안에도 이상치(Outlier)들을 찾아내고 삭제하는 작업이 병행되어야 한다. <그림 Ⅳ-7>에 나타난 것처럼 대각선 가장 아래쪽 원안에 위치한 다섯 개의 잔차항들은 이상치(outlier)이다. 그래프를 통해서 이상치를 확인한 후에는 통계모형에서 이상치를 제거하고 분석을 실시하여야 한다. 그런데 주의해야할 점은 그래프 상에서 두드러지게 벗어난 점들이 진짜 이상치인 지를 확인하고, 이상치라고 확신되는 경우에 그 값을 제거하고 분석하여야 한다.

<그림 IV-7> Lognormal 모델 검정: Pearson 잔차항과 임의데이터 비교



<그림 Ⅳ-8>은 잔차항과 추정값의 관계를 보여주는 그림이다. 이 그림을 통해서도 로그노말(Lognormal) 모델이 사고심도 분석에 적합하다는 것을 알 수 있다. 그러나 원안에 있는 데이터는 이상치(outlier)로 의심이 되므로 모델 확인 작업 중에 해당 값을 삭제하여야 할지 여부를 판단하여야 한다. 이처럼 모델 확인 및 모델 비교작업을 하는 동안 발견되는 이상치(outlier)들을 확인·삭제하는 과정이 계속 병행되어야 한다.

<그림 Ⅳ-8> Lognormal 모델 검정: Pearson 잔차항과 추정값 관계



지수분포를 가진 연속 반응변수를 분석하기 위해 주로 감마(Gamma)모델과 로그노말(Lognormal) 모델이 이용된다. 일반화선형모형 중에서 역가우시안(Inverse Gaussian)은 실무에서 또는 학계에서 실제 데이터에 적용되는 빈도가 가장 낮다. <표 Ⅳ-9>는 본 연구에서 사용된 자료를 활용하여 로그노말(Lognormal)과 감마(Gamma), 그리고 역가우시안(Inverse Gaussian) 모델의 AIC와 BIC를 계산한 값이다. 역가우시안(Inverse Gaussian)의 AIC와 BIC 두 값 모두 로그노말과 감마모델에 비해 크다. 로그노말 모델의 AIC와 BIC는 감마의 AIC와 BIC보다 값이 현저히 작다. 이상의 모델별 비교결과를 볼 때, 본 연구에서 사용된 자동차보험 데이터 대인배상I 분석에서는 로그노말(Lognormal)이 가장 적합한 모델이라고 할 수 있다.

<표 Ⅳ-9> AIC, BIC를 이용한 일반화선형모델 비교: 사고심도

구 분	Lognormal (1)	Gamma (2)	Inverse Gaussian	(2)-(1)
AIC	2.8	31.5	44.3	28.7
BIC	-200170	-193645.5	-119283.7	6524.5

다) 모형선택 과정에서 시사점

<표 Ⅲ-3>에서 제시된 바와 같이 자동차보험 요율상대도 산출에서 일반적으로 사용되는 일반화선형모형은 포아송분포(사고빈도)와 감마분포(사고심도)이다. 그런데 본 연구에서 자동차보험 자료를 가지고 모델선택과정을 거치면서 선택된 모델로는 사고빈도의 경우 음이항분포 그리고 사고심도의 경우는 로그노말(Lognormal) 분포이다. 이 결과는 자동차보험 요율산출에서 일반적으로 사용되는 모델과 차이가 난다.

이는 많은 요율산출 실무자들이 비연속성 지수분포의 반응변수를 분석하기 위해 일반적으로 포아송(Poisson)을 선택하고 연속성 지수분포의 반응변수를 분석하기 위해 감마(Gamma)모델을 선택하는 경향이 잘 못되었다는 것을 의미한다. 보험사마다 자사의 경험통계를 바탕으로 일반화선형모델들을 비교 분석하여 모델을 선택해야 하는데, 선행연구에서 추천한 결과를 단순히 추종하기만 할 경우에 나타날 수 있는 오류인 것이다.

따라서 각 회사는 일반화선형모형을 요율산출에 활용할 때 자기 회사의 자료에 부합된 모델을 선택하는 과정을 반드시 거치는 것이 필요하다. 각 보험사의 자동차보험에 가입된 모집단 데이터 또는 표본 데이터마다 가입자의 특성이 다르거나 데이터의 분포에서 차이가 있을 수 있기 때문에 적정 일반화모델을 선택하는 과정을 거쳐 위험상대도를 추정해야 한다.

2) 모형적합 : 회귀분석 결과

본 연구의 분석에 사용된 요율체계에는 연령, 성, 차종, 지역, 할인·할증, 가입경력, 교통법규위반경력이 포함되어 있다. <표 Ⅳ-10>은 본 연구에서 가정하고 있는 요율체계에 부합된 분석 결과이다.

<표 Ⅳ-10> 일반화선형모델을 이용한 회귀분석 결과

구 분		사고빈도		사고심도	
		Coefficient	Std. Dev.	Coefficient	Std. Dev.
연령	25세이하	0.342	0.042***	-0.009	0.040
	26~30세	0.139	0.025***	0.013	0.024
	40~64세	0.171	0.015***	0.028	0.015*
	65세이상	0.131	0.038***	0.043	0.036
성	남자	-0.071	0.016***	0.040	0.015***
차종	소형A	-0.170	0.029***	-0.061	0.028**
	소형B	-0.079	0.017***	-0.056	0.016***
	대형	-0.086	0.025***	0.053	0.024**
	승합1	-0.400	0.123***	0.061	0.115
	승합2	0.125	0.018***	0.052	0.017***
지역	부산	-0.185	0.031***	-0.129	0.030***
	경기	-0.074	0.020***	0.076	0.019***
	강원	-0.208	0.042***	0.165	0.040***
	충북	-0.243	0.043***	-0.024	0.040
	충남	-0.226	0.038***	0.038	0.036
	전북	0.018	0.037	0.038	0.035
	전남	-0.323	0.045***	0.046	0.043
	경북	-0.263	0.034***	-0.072	0.032**
	경남	-0.323	0.033***	-0.021	0.031
	제주	-0.601	0.081***	-0.366	0.077***
	대구	-0.123	0.032***	-0.220	0.030***
	인천	0.150	0.029***	0.125	0.028***
	광주	-0.110	0.040***	-0.039	0.038
	대전	-0.042	0.036	-0.019	0.034

	울산	-0.246	0.047***	-0.099	0.045**
할인할증	할인40%	-0.491	0.031***	-0.049	0.029*
	할인45%	-0.262	0.035***	-0.052	0.033
	할인50%	-0.213	0.034***	-0.042	0.033
	할인60%	-0.141	0.035***	0.001	0.033
	할인70%	-0.117	0.034***	-0.009	0.033
	할인80%	-0.090	0.034***	0.020	0.032
	할인90%	-0.023	0.032	0.018	0.031
	할증110%	0.114	0.046**	-0.049	0.044
	할증120%	0.141	0.053***	0.092	0.050*
	할증130%	0.299	0.062***	0.056	0.058
	할증140%	0.214	0.090**	0.190	0.085**
	할증150%	0.469	0.107***	0.117	0.102
	할증160%	0.382	0.147***	0.018	0.143
	할증170%	0.627	0.195***	0.087	0.190
	할증180%	0.336	0.277	0.153	0.263
	할증190%	0.490	0.328	0.047	0.334
	할증200%	0.824	0.273***	0.064	0.299
가입경력	1년미만	0.304	0.038***	0.066	0.036*
	1~2년미만	0.133	0.035***	0.015	0.033
	2~3년미만	0.057	0.032*	-0.033	0.031
	3~4년	-0.012	0.026	0.009	0.025
교통법규	할증1그룹1	-0.063	0.588	0.057	0.544
	할증1그룹2	0.089	0.032***	-0.005	0.030
	할증2그룹	0.117	0.109	0.072	0.106
	기본그룹	0.122	0.016***	-0.013	0.015

주 : 1) Robust 표준오차에 기준하여 *은 유의수준 10%, **은 유의수준 5%, ***은 유의수준 1% 수준에서 통계적 유의성을 나타낸다.

2) NB: LR test of alpha=0: chibar2(01)=251.5, Prob>chibar2=0.000, LR chi2(54) = 2323.81

3) Lognormal: AIC=2.69, BIC=-200788.1, Scale parameter=.862, Deviance=18940.4

일반화선형모델로 분석한 결과 자동차의 종류는 자동차사고 빈도와 심도 모두에 매우 큰 영향을 주는 위험요소이다. 사고빈도가 가장 낮은 차종은 2,000cc 이상의 대형차들이었으며 소형A, 소형B, 다인승 1종 차량들 모두 중형차에 비해 사고빈도가 낮은 차량들이었다. 차종 중에 다인승2종이 사고율이 가장 높았으며 중형차에 비해 12.5%가 높았다.

대형차와 다인승 2종은 사고가 났을 때 피해액수가 높았다. 하지만 소형A와 소형B는 중형차에 비해 사고빈도도 낮을 뿐만 아니라 교통사고 발생시 피해액도 적었다.

회귀분석결과 인적요인인 성, 연령, 가입경력 모두가 자동차사고에 영향을 주는 유효한 위험변수인 것으로 분석되었다. 모든 조건이 동일할 때 남자는 여성에 비해 약 7.1%정도 사고빈도가 낮게 나타났다. 하지만 남성은 여성에 비해 사고가 났을 경우 사고심도가 1.5%정도 높았다.

연령계층별 분석결과를 보면, 25세 이하의 저연령계층의 운전자들은 모든 연령계층에서 사고빈도가 가장 높은 연령층이었으며 30~40세 운전자에 비해 사고빈도가 무려 34.2%정도나 높았다. 65세 이상의 고연령 운전자들도 사고빈도가 높았다. 하지만 일반적으로 운전자의 연령층은 사고심도에 큰 영향을 주지 않은 변수로 분석되었다.

자동차보험 가입경력을 보면, 가입경력이 짧을수록 교통사고 발생 빈도가 높았으며 자동차보험에 가입한 경력이 1년이 지나지 않은 운전자는 가입경력이 4년 이상인 운전자 보다 무려 사고빈도가 30.4%정도나 높았다. 하지만 가입경력이 증가할수록 교통사고 발생빈도가 낮아지는 정도가 점차 줄어들었으며, 가입경력이 3년 이상이면 가입경력에 따른 교통사고 빈도의 변화가 크지 않았다. 가입경력이 1년 미만인 운전자의 사고피해액에 가장 높았으며 그 이후로는 큰 차이가 없었다. 가입경력 3~4년차 운전자의 위험도를 일반화할 수 있는지 확인하기 위하여 가입경력 4~5년차의 더미변수를 추가하여 추가로 분석하였다. 그 결과 운전자의 가입경력이 3~4년 이상이면 위험도가 높지 않은 것으로 나타났다(표 <Ⅳ-11> 참조).

<표 IV-11> 가입경력과 위험도

구 분		사고빈도		사고심도	
		Coefficient	Std. Dev.	Coefficient	Std. Dev.
가입경력	1년미만	0.298	0.039***	0.078	0.037**
	1~2년미만	0.129	0.035***	0.036	0.033
	2~3년미만	0.052	0.033	-0.023	0.031
	3~4년미만	-0.016	0.026	0.017	0.025
	4~5년	-0.021	0.024	0.023	0.022

주 : 1) Robust 표준오차에 기준하여 *은 유의수준 10%, **은 유의수준 5%, ***은 유의수준 1% 수준에서 통계적 유의성을 나타낸다.

일반화선형모형으로 분석한 결과 할인할증 역시 자동차 보험료의 요율분석의 결정요인으로 유효한 위험변수인 것으로 나타났다. 자동차보험 가입운전자 중 과거 자동차 사고경력에 따른 보험료할인 혜택이 가장 많은 그룹은 모든 조건이 동일할 때 60%의 자동차보험료 할인이 적용되는데 이 범위에 속한 운전자들은 자동차보험료 할인할증 그룹에 속하지 않는 운전자에 비해 무려 49.1%정도 교통사고 빈도가 낮았다. 자동차보험료 할인혜택이 높은 운전자 일수록 자동차사고 빈도가 낮아지는 것을 알 수 있다. 또한 자동차보험료가 할증되어 부과되는 즉 과거 교통사고 경력이 높은 운전자 일수록 자동차사고 빈도가 점차 높아지는 것을 알 수 있다. 특히 자동차 사고경력이 많아 자동차 보험료가 200% 할증되는 그룹의 운전자들은 할인할증 적용을 받지 않은 운전자들에 비해 무려 82.4%정도의 높은 교통사고 빈도를 나타내는 것으로 분석되었다.

교통법규위반 경력의 경우 할증정도가 가장 심한 할증1그룹1에 해당하는 운전자의 경우 교통사고빈도가 낮게 나타났다. 통계적으로 유의하지는 않지만 고위험군에 속한 운전자임에도 위험도가 낮게 나타난 것은 통계기간 중 운전면허 취소 등으로 운전기간이 절대적으로 짧아서 나타나는 현상으로 보인다. 이것은 할증1그룹1에 속한 운전자들이 무면허운전금지나 사고발생

시 조치의무 위반과 같이 운전면허 취소사유에 해당하는 법규를 위반한 운전자들이기 때문이다. 일반적으로 과거 교통사고 경력으로 인하여 할증을 받거나 할인을 받지 못하는 운전자들은 할인을 받는 운전자들보다 사고빈도가 높으며 사고액은 크게 차이가 없는 것으로 분석되었다.

전국을 운전할 수 있는 자동차의 특성을 무시하고 단지 자동차의 등록지를 기준으로 지역을 분류하여 자동차보험료에 차이를 두는 것이 형평성에 어긋난다는 논리로 인해 아직은 지역에 따른 자동차보험료 차등화는 적용되지 않고 있다. 하지만 일반화선형모형을 이용해 분석한 결과 모든 조건이 동일하더라도 지역은 자동차사고 위험도를 결정하는 주요 변수로 나타났다. 운전자의 운전경력이나 기타 조건들이 동일하더라도 안전시설 설치 여부와 교통위반 단속 등과 같은 지역별 차이로 인하여 자동차사고 상대 위험도가 지역별로 상이할 수 있다는 것이 본 연구의 결과이다. 지역별 자동차 보험료 차등화에 대해서는 본 연구보고서의 목적과 다소 거리가 있어 필요하면 차후에 다른 연구보고서를 통해 논의하겠다.

3) 분석결과의 해석

일반화선형모형으로 추정된 모수는 상대위험도를 나타내기는 하지만 정확한 값이 아니다. 그래서 지금까지 '약' 또는 '정도'와 같은 단어를 이용해 상대위험도를 설명하였다. 따라서 정확한 상대위험도를 계산하기 위해서는 추정된 모수를 발생률(Incidence Rate Ratio)로 전환하여야 한다. 계량분석에서 추정된 특정 모수는 자동차보험 사고빈도 및 심도에 영향을 줄 수 있는 다른 위험요소들에 변화가 없다고 가정할 때 특정 위험요소가 한 단위 증가할 경우 자동차보험 사고빈도 또는 심도의 증가정도를 의미한다. 특정 위험변수 x가 x_0에서 x_{0+1}로 한 단위 증가했다고 가정하자. 모든 다른 위험요소에 변화가 없다고 가정할 때 특정 위험요소 x가 한 단위 증가할 때 교통사고빈도 또는 심도의 변화분은 β가 될 것이며 식(Ⅳ-1)을 이용해 β를 다음과 같이 정의할 수 있다.

$$\beta = \log(\mu_{x_{0+1}}) - \log(\mu_{x_0}) \quad (\text{IV-1})$$

또한 로그값의 차이가 비율로 표현될 수 있는 로그의 특성을 이용하여 모수의 값을 다음과 같이 다시 표현할 수 있다.

$$\beta = \log(\frac{\mu_{x_{0+1}}}{\mu_{x_o}}) \quad (\text{IV-2})$$

<표 Ⅳ-12>는 음이항분포모델과 로그노말(Lognormal)을 이용해 추정된 모수를 상대적 위험도로 전환한 값이다.

<표 Ⅳ-12> 일반화선형모형으로 추정된 모수들의 발생률(Incidence Rate Ratio)로의 전환: 상대위험도

구 분		사고빈도		사고심도	
		Coefficient	Std. Dev.	Coefficient	Std. Dev.
연령	25세이하	1.408	0.042***	0.991	0.040
	26~30세	1.149	0.025***	1.013	0.024
	40~64세	1.186	0.015***	1.028	0.015*
	65세이상	1.140	0.038***	1.044	0.036
성	남자	0.931	0.016***	1.041	0.015***
차종	소형A	0.843	0.029***	0.941	0.028**
	소형B	0.924	0.017***	0.946	0.016***
	대형	0.918	0.025***	1.054	0.024**
	승합1	0.671	0.123***	1.063	0.115
	승합2	1.133	0.018***	1.053	0.017***
지역	부산	0.831	0.031***	0.879	0.030***
	경기	0.929	0.020***	1.079	0.019***
	강원	0.812	0.042***	1.179	0.040***
	충북	0.784	0.043***	0.976	0.040
	충남	0.798	0.038***	1.039	0.036

지역	전북	1.018	0.037	1.039	0.035
	전남	0.724	0.045***	1.047	0.043
	경북	0.768	0.034***	0.931	0.032**
	경남	0.724	0.033***	0.979	0.031
	제주	0.549	0.081***	0.694	0.077***
	대구	0.884	0.032***	0.803	0.030***
	인천	1.162	0.029***	1.133	0.028***
	광주	0.896	0.040***	0.962	0.038
	대전	0.959	0.036	0.981	0.034
	울산	0.782	0.047***	0.906	0.045**
할인할증	할인40%	0.612	0.031***	0.952	0.029*
	할인45%	0.770	0.035***	0.949	0.033
	할인50%	0.809	0.034***	0.959	0.033
	할인60%	0.869	0.035***	1.001	0.033
	할인70%	0.889	0.034***	0.991	0.033
	할인80%	0.914	0.034***	1.020	0.032
	할인90%	0.977	0.032	1.018	0.031
	할증110%	1.120	0.046**	0.952	0.044
	할증120%	1.152	0.053***	1.096	0.050*
	할증130%	1.349	0.062***	1.058	0.058
	할증140%	1.239	0.090**	1.209	0.085**
	할증150%	1.599	0.107***	1.124	0.102
	할증160%	1.465	0.147***	1.018	0.143
	할증170%	1.872	0.195***	1.091	0.190
	할증180%	1.399	0.277	1.165	0.263
	할증190%	1.632	0.328	1.048	0.334
	할증200%	2.279	0.273***	1.066	0.299
가입경력	1년미만	1.355	0.038***	1.068	0.036*
	1~2년	1.142	0.035***	1.015	0.033
	2~3년	1.058	0.032*	0.968	0.031
	3~4년	0.988	0.026	1.009	0.025

교통법규	할증1그룹1	0.939	0.588	1.059	0.544
	할증1그룹2	1.093	0.032***	0.995	0.030
	할증2그룹	1.124	0.109	1.075	0.106
	기본그룹	1.130	0.016***	0.987	0.015

주 : Robust 표준오차에 기준하여 *은 유의수준 10%, **은 유의수준 5%, ***은 유의수준 1% 수준에서 통계적 유의성을 나타낸다.

지금까지 자동차사고빈도 및 심도에 영향을 줄 수 있는 위험요소들의 영향을 일반화선형모델로 추정하였고 추정된 모수의 해석 방법을 알아보았다. 마지막으로 추정 후 비율로 변환된 모수를 이용해 보험요율에 적용하는 방법을 살펴보고자 한다.

앞서 여러번 언급한 것처럼 현행 자동차보험 요율제도는 곱셈공식에 따라 이루어진다. 즉 개인용 자동차를 예로 들면, 차종에 따른 기본보험료에 가입자가 선택한 각종 특약 요율과 가입자의 사고경력에 따른 할인할증률, 가입자의 운전경력 요율을 곱하여 적용 보험료가 산출된다. 기타 위험변수들에 대한 보험요율 적용은 이처럼 곱의 원리를 적용하면 된다. 이러한 원리에 따라 산출된 변수별 위험상대도 결과들의 해석은 아래와 같다.

차종에 따른 상대위험도의 결과를 보험요율에 적용하는 방법을 예로써 설명하면, 중형차를 기준으로 소형A는 사고빈도가 낮고 사고 시에도 피해액이 적다. 또한 이러한 차이가 통계적으로 유의하기 때문에 소형A를 운전하는 운전자에게는 중형차를 운전하는 운전자가 지불하는 보험료에 비해 0.793[23]배의 보험료를 부과해야 한다. 이에 반해 다인승 1종에 해당하는 차량을 소유한 운전자들은 중형차에 비해 사고빈도가 매우 낮고, 사고 시에는 피해액수에 큰 차이가 없어 중형차 운전자의 보험료에 비해 0.671배의 보험료를 부과시켜야 한다.

현재 자동차 보험요율에 적용하고 있지는 않지만 성별에 의해 자동차 보험료에 차등을 두는 경우를 보면 다음과 같다. 남성은 여성에 비해 교통사고 시에 피해액수가 높다. 하지만 교통사고 심도가 높음에도 불구하고 교통

23) 0.843×0.941=0.793

사고 빈도가 여성보다 현저히 낮기 때문에 모든 조건이 동일하다고 가정했을 때 남성은 여성에 비해 0.969배의 보험료를 지불하는 것이 형평성에 적합하다.

결과적으로 소형A를 운전하는 남성에게 중형을 운전하는 여성운전자에 비해 0.768[24]배, 즉 23.2% 낮은 보험료를 부과하는 것이 적합하다.

나. 차량담보

차량담보에 일반화 선형모형을 적용하는 방법도 대인배상 I 과 동일하다. 모형설정, 모형적합 및 결과의 해석의 단계에 따라 일반화 선형모형이 적용된다.

<표 Ⅳ-13>는 본 연구보고서 분석에 사용된 자동차 연식에 따른 기술적 통계(Descriptive Statistics)이다. 자기차량 담보에 가입된 차량은 27만 2,885개로 의무보험인 대인배상의 가입차량의 50%가 조금 넘는다. <표 Ⅳ-13>에서 눈에 띄는 특징은 연식이 7~9년 이상인 차량의 경우 연식이 증가할수록 자동차 사고빈도가 낮아지고 있다. 자동차 연식과 사고빈도의 관계를 좀 더 자세히 살펴보기 위해 두 변수의 관계를 <그림Ⅳ-9>로 나타냈다.

24) 0.793×0.969

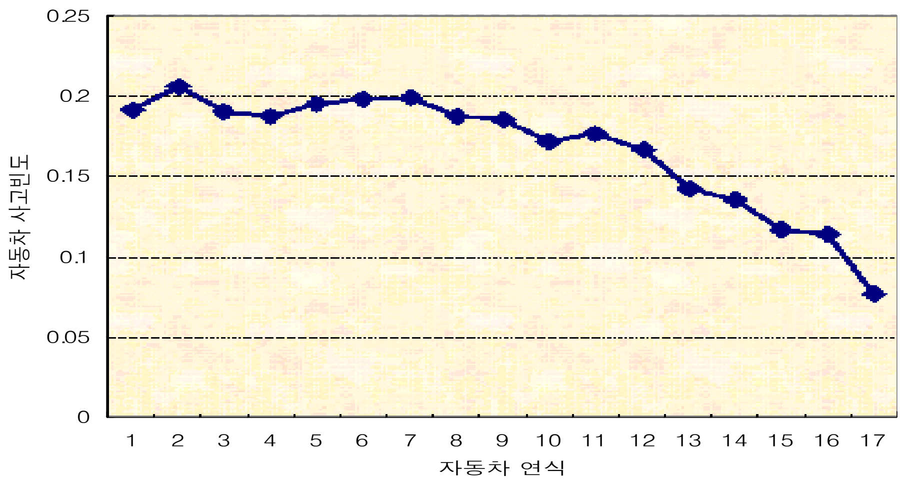
<표 Ⅳ-13> 자동차 연식에 따른 교통사고 빈도 및 기술적 통계

연 식	개체수	평균 (Mean)	표준67편차 (Std. Div.)	사고빈도 최솟값	사고빈도 최댓값
1	10,084	0.191	0.476	0	6
2	29,123	0.206	0.488	0	6
3	25,906	0.191	0.465	0	6
4	29,130	0.187	0.475	0	8
5	34,210	0.195	0.489	0	5
6	25,266	0.199	0.507	0	7
7	25,377	0.200	0.510	0	6
8	18,598	0.188	0.489	0	5
9	9,421	0.185	0.495	0	5
10	19,497	0.171	0.466	0	5
11	17,960	0.177	0.487	0	6
12	14,048	0.167	0.468	0	5
13	8,232	0.142	0.419	0	5
14	3,848	0.136	0.413	0	5
15	1,381	0.117	0.366	0	4
16	544	0.114	0.356	0	2
17	156	0.077	0.290	0	2
18	80	0.125	0.402	0	2
19	14	0.214	0.426	0	1
20	4	0	0	0	0
21	1	0	-	0	0
22	1	0	-	0	0
23	1	0	-	0	0
24	0	-	-	-	-
25	0	-	-	-	-
26	0	-	-	-	-
27	1	0	-	0	0
28	0	-	-	-	-
29	0	-	-	-	-
30	2	0	0	0	0

주 : 새차라 함은 최초신규등록년도가 보험책임시작년도와 같거나 보험책임개시일이 최초신규 등록일(외국산자동차의 경우 제작일)로부터 6개월 이내인 자동차를 말한다. 단, 제작일을 알 수 없는 외국산자동차의 경우 연식과 보험책임시작년도가 같은 자동차를 말한다.

<그림 Ⅳ-9>에서 보면 자동차 연식이 약 7~8년 정도일 때까지는 연식에 따른 사고빈도의 변화의 양상이 뚜렷하지 않으나, 그 이후에는 연식이 증가할수록 사고빈도가 감소한다. 이러한 관계는 자동차 연식이 일정 수준 이상 증가함에 따라 운행거리가 짧아 나타나는 현상일 수도 있다. 중요한 것은 <표 Ⅳ-13>과 <그림 Ⅳ-9>에서 나타나는 자동차 연식과 위험도의 관계는 단순히 기술적 통계에 불과하다는 것이다. 따라서 다른 위험변수들의 영향을 고려한 자동차 연식과 자동차사고의 관계를 분석해야 한다.

<그림 Ⅳ-9> 자동차 연식에 따른 교통사고 빈도



위에서 언급한 모델선별 테스트 결과를 보면, 자차담보의 사고빈도의 경우에는 ZINB, 사고심도의 경우에는 감마(Gamma)모델이 가장 적합한 모델인 것으로 나타났다. 감마모델의 경우 링크함수(Link Function)는 역링크(Reciprocal)보다는 로그링크(Log Link)함수가 더 적당한 것으로 분석되었다.

<표 Ⅳ-14> Vuong 테스트를 이용한: 사고빈도

Vuong 테스트
Vuong of zinb vs. negative binomial: z=2.02 Pr>z = 0.0219

주 : Vuong 테스트에 이용된 변수는 성별과 가입경력이다.

Vuong 테스트의 결과 ZINB가 음이항분포 모형보다 약간 선호된다고 분석되었으나 두 모델을 통해 실제로 추정된 계수를 비교했을 경우에는 거의 차이가 없었다.

<표 Ⅳ-15> AIC, BIC를 이용한 일반화선형모델 비교: 사고심도

	Lognormal	Gamma
AIC	2.6	7.2
BIC	-54514.1	-151727.8

<표 Ⅳ-16>는 더미변수를 이용하여 자동차 연식이 자동차사고에 끼치는 영향을 분석한 결과이다. 분석 결과에 따르면 자동차 연식이 증가함에 따라 자동차 사고빈도도 증가하는 양상이다. 새차의 경우 사고빈도가 가장 낮게 나타났지만 사고심도는 가장 높게 나타났다. 새 차의 경우 사고빈도가 가장 낮게 나타난 이유는 운전자가 새차를 조심해서 운전하는 성향 때문으로 판단되며, 사고심도가 가장 높게 나타난 것은 새 차의 부품가격이 상대적으로 비싸기 때문으로 판단된다.

기타 독립변수들 역시 교통사고 빈도 및 심도의 분석에 상당한 설명력이 있는 것으로 나타났다. 구체적인 해석상의 방법론은 이미 앞장에서 구체적으로 다루었기 때문에 본 장에서는 생략한다.

<표 IV-16> 자차담보 분석: ZINB와 Gamma 적용

구 분		사고빈도		사고심도	
		Coefficient	Std. Dev.	Coefficient	Std. Dev.
연식	연식1	0.539	0.178***	1.474	0.159**
	연식2	1.085	0.029***	0.960	0.025*
	연식3	1.082	0.030***	1.030	0.026
	연식4	1.075	0.030***	1.079	0.026***
	연식5	1.126	0.029***	1.070	0.025***
	연식6	1.156	0.030***	1.102	0.026***
	연식7	1.184	0.030***	1.084	0.026***
	연식8	1.131	0.032***	1.106	0.028***
	연식9	1.130	0.037***	1.088	0.032***
	9년이상	0.967	0.028	0.991	0.024
연령	25세이하	1.228	0.044***	1.110	0.038***
	26~30세	1.072	0.020***	1.020	0.017
	40~64세	1.061	0.011***	1.044	0.010***
	65세이상	1.013	0.028	1.060	0.025**
성	남자	0.904	0.012***	1.024	0.010**
차종	소형A	0.746	0.022***	0.789	0.019***
	소형B	0.879	0.014***	0.871	0.012***
	대형	1.124	0.017***	1.299	0.015***
	승합1	0.719	0.108***	0.920	0.100
	승합2	0.907	0.014***	0.974	0.012**
지역	부산	1.228	0.025***	0.917	0.022***
	경기	0.985	0.014	1.036	0.012***
	강원	0.993	0.032	1.018	0.027
	충북	0.937	0.032**	1.096	0.028***
	충남	0.993	0.028	1.076	0.024***
	전북	1.004	0.032	1.052	0.027*
	전남	1.337	0.032***	0.993	0.027

지역	경북	1.207	0.025***	0.976	0.021
	경남	1.208	0.022***	0.999	0.019
	제주	1.435	0.050***	0.840	0.044***
	대구	1.224	0.025***	0.965	0.021
	인천	1.138	0.022***	1.029	0.020
	광주	1.208	0.030***	0.896	0.025***
	대전	0.980	0.028	1.021	0.024
	울산	1.184	0.037***	0.845	0.032***
할인할증	할인40%	0.788	0.024***	0.755	0.022***
	할인45%	0.838	0.027***	0.810	0.024***
	할인50%	0.841	0.027***	0.863	0.024***
	할인60%	0.803	0.028***	0.916	0.025***
	할인70%	0.812	0.028***	0.935	0.025***
	할인80%	0.848	0.028***	0.965	0.025
	할인90%	0.872	0.028***	0.958	0.024*
	할증110%	0.969	0.039	0.980	0.034
	할증120%	1.021	0.047	1.012	0.041
	할증130%	1.095	0.056	1.090	0.050*
	할증140%	1.060	0.082	1.085	0.072
	할증150%	1.037	0.115	1.027	0.103
	할증160%	1.058	0.150	1.384	0.132**
	할증170%	1.507	0.197**	1.116	0.170
	할증180%	1.831	0.283**	1.709	0.244**
	할증190%	1.521	0.358	0.959	0.293
	할증200%	3.072	0.296***	1.123	0.265
가입경력	1년미만	1.213	0.033***	1.060	0.029**
	1~2년미만	1.101	0.030***	1.000	0.026
	2~3년미만	1.059	0.027**	1.009	0.024
	3~4년미만	1.005	0.020	0.991	0.018
교통	할증1그룹1	2.109	0.371**	1.103	0.293
	할증1그룹2	1.275	0.025***	1.001	0.022

법규	할증2그룹	1.126	0.089	1.011	0.077
	기본그룹	1.087	0.011***	0.993	0.010

주 : 1) Robust 표준오차에 기준하여 *은 유의수준 10%, **은 유의수준 5%, ***은 유의수준 1% 수준에서 통계적 유의성을 나타낸다.
2) ZINB: LR test of alpha=0: Prob>chibar2=0.000 Vuong test of zinb vs. standard negative binomial: z=3.5, Pr>z=0.000 LR chi2(60) = 2037.83
3) Vuong Test에 차종과 가입경력 적용되었다.
4) Gamma: AIC=7.169, BIC=-450278.4, Scale parameter=0.0043, Deviance=187.05
5) 사고빈도의 경우 Excessizve Zero가 있지만 실제로 추정된 계수의 비교에서 ZINB와 음이항분포 모형 간에 차이가 없어 ZINB 대신 음이항분포를 이용해도 무방하다고 분석된다.

다. 대인배상Ⅱ, 대물배상, 자기신체사고 및 무보험차상해

대인배상Ⅰ및 자기차량손해 이외에도 대인배상Ⅱ, 대물배상, 자기신체사고, 무보험차상해 담보별 분석방법은 앞에서 기술한 대인배상Ⅰ과 자기차량손해담보와 동일하다.

일반화선형모형을 자동차보험 요율상대도 산출에 적용하는 방법에 따라 대인배상Ⅱ, 대물배상, 자기신체사고 및 무보험차상해 담보로 분석한 결과는 다음의 표와 같다.

<표 Ⅳ-17> 대인배상Ⅱ: NB와 Gamma모델 적용

구 분		사고빈도		사고심도	
		Coefficient	Std. Dev.	Coefficient	Std. Dev.
연령	25세이하	1.562	0.106***	1.182	0.206
	26~30세	1.094	0.063	1.052	0.121
	40~64세	1.181	0.039***	1.082	0.075
	65세이상	1.299	0.090***	1.098	0.174
성	남자	0.973	0.042	1.054	0.079
차종	소형A	0.836	0.074**	1.094	0.142
	소형B	0.889	0.044***	0.989	0.085
	대형	0.977	0.063	1.082	0.120

차종	승합1	0.897	0.292	0.583	0.555
	승합2	1.150	0.046***	1.090	0.088
지역	부산	0.782	0.081***	1.653	0.153***
	경기	0.903	0.050**	1.376	0.095***
	강원	1.053	0.095	1.589	0.182***
	충북	0.654	0.117***	1.692	0.223**
	충남	0.888	0.092	1.563	0.175***
	전북	0.856	0.099	1.743	0.188***
	전남	0.693	0.117***	1.471	0.222*
	경북	0.780	0.086***	1.885	0.164***
	경남	0.783	0.079***	1.346	0.150**
	제주	0.767	0.179	1.241	0.339
	대구	0.705	0.088***	1.237	0.167
	인천	1.150	0.074*	1.210	0.140
	광주	0.772	0.109**	1.384	0.209
	대전	0.920	0.095	1.618	0.180***
	울산	0.855	0.114	1.699	0.217**
할인할증	할인40%	0.520	0.078***	1.089	0.149
	할인45%	0.656	0.087***	0.982	0.168
	할인50%	0.716	0.086***	0.942	0.165
	할인60%	0.701	0.089***	1.051	0.171
	할인70%	0.760	0.087***	1.105	0.168
	할인80%	0.872	0.085	1.066	0.161
	할인90%	0.960	0.083	1.012	0.158
	할증110%	0.865	0.125	1.132	0.241
	할증120%	1.042	0.138	1.337	0.261
	할증130%	1.555	0.144***	0.877	0.273
	할증140%	1.600	0.201***	1.144	0.380
	할증150%	1.810	0.254***	1.784	0.480
	할증160%	1.771	0.343*	0.490	0.647
	할증170%	1.389	0.586	0.563	1.106
	할증180%	0.741	1.007	1.363	1.965

	할증190%	8.11e-08	0.0003	0.000	0.000
	할증200%	4.932	0.521***	0.053	0.968***
가입경력	1년미만	1.384	0.096***	1.122	0.186
	1~2년미만	1.068	0.090	1.137	0.169
	2~3년미만	1.077	0.081	0.970	0.155
	3~4년미만	1.072	0.065	0.968	0.127
교통법규	할증1그룹1	1.07e-07	0.293	0.000	0.000
	할증1그룹2	1.194	0.081**	1.077	0.155
	할증2그룹	1.495	0.246	1.585	0.484
	기본그룹	1.068	0.040	1.022	0.078

주 : 1) Robust 표준오차에 기준하여 *은 유의수준 10%, **은 유의수준 5%, ***은 유의수준 1% 수준에서 통계적 유의성을 나타냄낸다.
2) N=433450
3) NB: LR test of alpha=0: chibar2(01)=14.61 Prob>=chibar2=0.000
4) Gamma: LR chi2(50)=498.44, AIC=2.76, BIC=-187270.9

<표 Ⅳ-18> 대물배상: NB와 Lognormal모델 적용

구 분		사고빈도		사고심도	
		Coefficient	Std. Dev.	Coefficient	Std. Dev.
연령	25세이하	1.171	0.032***	1.083	0.030***
	26~30세	1.079	0.018***	1.016	0.017
	40~64세	1.154	0.011***	1.039	0.010***
	65세이상	1.134	0.026***	1.079	0.024***
성	남자	0.896	0.012***	1.037	0.011***
차종	소형A	0.803	0.021***	0.869	0.019***
	소형B	0.879	0.012***	0.950	0.011***
	대형	0.935	0.018***	1.113	0.016***
	승합1	0.797	0.080***	1.061	0.076
	승합2	1.154	0.013***	1.085	0.012***
지역	부산	0.976	0.022	0.948	0.020***

지역	경기	0.961	0.014***	1.022	0.013
	강원	0.814	0.030***	0.953	0.028*
	충북	0.841	0.030***	1.051	0.028*
	충남	0.903	0.026***	1.000	0.024
	전북	0.929	0.028***	1.103	0.026***
	전남	0.895	0.030***	1.009	0.028
	경북	0.971	0.023	1.045	0.021**
	경남	0.965	0.021*	0.961	0.020**
	제주	0.887	0.047***	0.932	0.043
	대구	1.064	0.022***	0.983	0.020
	인천	1.028	0.022	1.025	0.021
	광주	0.993	0.028	0.986	0.026
	대전	0.925	0.027***	1.018	0.025
	울산	0.980	0.031	0.915	0.029***
할인할증	할인40%	0.731	0.022***	0.800	0.020***
	할인45%	0.852	0.025***	0.863	0.023***
	할인50%	0.885	0.024***	0.867	0.023***
	할인60%	0.888	0.025***	0.917	0.024***
	할인70%	0.919	0.025***	0.920	0.023***
	할인80%	0.921	0.025***	0.957	0.023*
	할인90%	0.976	0.024	0.958	0.022*
	할증110%	1.082	0.034**	0.921	0.032***
	할증120%	1.155	0.039***	0.968	0.037
	할증130%	1.199	0.048***	0.957	0.044
	할증140%	1.200	0.067***	0.946	0.062
	할증150%	1.548	0.081***	1.059	0.078
	할증160%	1.441	0.109***	0.987	0.105
	할증170%	1.769	0.149***	0.897	0.145
	할증180%	1.262	0.215	1.270	0.202
	할증190%	1.627	0.244**	1.391	0.247
	할증200%	1.824	0.224***	1.324	0.256

가입경력	1년미만	1.405	0.027***	1.024	0.026
	1~2년미만	1.111	0.026***	1.004	0.024
	2~3년미만	1.069	0.023***	1.018	0.022
	3~4년미만	0.959	0.019**	1.027	0.018
교통법규	할증1그룹1	0.954	0.422	1.481	0.413
	할증1그룹2	1.063	0.023***	1.027	0.021
	할증2그룹	1.190	0.077**	0.947	0.072
	기본그룹	1.137	0.011***	0.971	0.010***

주 : 1) Robust 표준오차에 기준하여 *은 유의수준 10%, **은 유의수준 5%, ***은 유의수준 1% 수준에서 통계적 유의성을 나타낸다.
2) N=424169
3) NB: LR test of alpha=0: chibar2(01)=905.27, Prob>=chibar2= 0.000
4) Lognormal: AIC=2.68, BIC=-435796.8, Scale Parameter=0.85, Deviance=37623.48

<표 Ⅳ-19> 자기신체사고: NB와 Lognormal 모델 적용

구 분		사고빈도		사고심도	
		Coefficient	Std. Dev.	Coefficient	Std. Dev.
연령	25세이하	1.343	0.120**	0.729	0.158**
	26~30세	0.952	0.073	1.112	0.097
	40~64세	1.148	0.045***	1.058	0.059
	65세이상	1.105	0.113	0.773	0.150*
성	남자	0.799	0.046***	0.871	0.060**
차종	소형A	1.202	0.074**	0.923	0.096
	소형B	1.020	0.050	1.004	0.066
	대형	0.821	0.080***	1.231	0.105**
	승합1	0.768	0.381	0.512	0.500
	승합2	0.971	0.056	1.019	0.074
지역	부산	1.149	0.101	0.866	0.132
	경기	1.491	0.065***	1.164	0.086*
	강원	1.557	0.117***	1.428	0.153**
	충북	1.732	0.113***	1.085	0.148

지역	충남	1.672	0.104***	1.225	0.136
	전북	2.349	0.097***	1.165	0.126
	전남	1.949	0.110***	1.537	0.143***
	경북	1.498	0.097***	1.298	0.127**
	경남	1.311	0.094***	1.119	0.123
	제주	1.059	0.221	0.448	0.287***
	대구	0.956	0.113	0.938	0.147
	인천	1.655	0.094***	1.210	0.123
	광주	1.622	0.116***	1.049	0.152
	대전	1.751	0.106***	1.261	0.138*
	울산	1.008	0.153	1.087	0.200
할인할증	할인40%	0.427	0.086***	0.921	0.112
	할인45%	0.487	0.100***	0.913	0.132
	할인50%	0.572	0.097***	0.847	0.127
	할인60%	0.590	0.100***	0.869	0.130
	할인70%	0.733	0.095***	0.928	0.125
	할인80%	0.702	0.095***	0.841	0.123
	할인90%	0.803	0.091**	0.829	0.119
	할증110%	0.880	0.132	0.902	0.174
	할증120%	1.017	0.148	0.771	0.192
	할증130%	1.092	0.178	1.062	0.235
	할증140%	1.216	0.242	0.859	0.314
	할증150%	1.827	0.271**	0.956	0.350
	할증160%	1.812	0.366	0.805	0.471
	할증170%	1.604	0.591	1.782	0.759
	할증180%	0.876	1.012	0.422	1.312
	할증190%	1.538	1.015	0.414	1.309
	할증200%	7.856	0.480***	1.907	0.590
가입경력	1년미만	1.171	0.107	0.742	0.138**
	1~2년미만	1.148	0.099	0.950	0.128
	2~3년미만	1.085	0.091	0.966	0.117
	3~4년미만	0.914	0.077	1.012	0.102

교통법규	할증1그룹1	7.113	0.738***	0.595	0.925
	할증1그룹2	1.274	0.093	0.922	0.122
	할증2그룹	0.974	0.357	1.428	0.465
	기본그룹	1.137	0.046***	1.040	0.060

주 : 1) Robust 표준오차에 기준하여 *은 유의수준 10%, **은 유의수준 5%, ***은 유의수준 1% 수준에서 통계적 유의성을 나타낸다.
2) N=429211
3) NB: LR test of alpha=0: chibar2(01)=16.73 Prob>=chibar2 =0.000, LR chi2(50)=584.9
4) Lognormal: AIC=3.38, BIC=-15843.7, Scale Parameter=1.69, Deviance=4341.28

<표 Ⅳ-20> 무보험차상해 : NB와 Gamma 모델 적용

구 분		사고빈도		사고심도	
		Coefficient	Std. Dev.	Coefficient	Std. Dev.
연령	25세이하	1.264	0.396	1.153	0.861
	26~30세	0.976	0.208	0.773	0.487
	40~64세	0.858	0.124	1.235	0.289
	65세이상	0.482	0.428*	4.336	0.887
성	남자	1.072	0.141	1.937	0.328**
차종	소형A	1.321	0.216	1.945	0.487
	소형B	1.087	0.145	2.566	0.339***
	대형	0.786	0.237	1.357	0.676
	승합1	0.837	1.049	0.838	1.999
	승합2	1.199	0.155	1.974	0.344**
지역	부산	0.875	0.273	0.666	0.581
	경기	0.914	0.176	1.213	0.370
	강원	1.455	0.293	0.702	0.556
	충북	1.076	0.332	2.638	0.690
	충남	0.885	0.319	0.510	0.694
	전북	1.432	0.285	8.100	0.765***
	전남	1.391	0.306	2.722	0.698

지역	경북	0.861	0.286	0.837	0.571
	경남	1.255	0.232	1.137	0.473
	제주	1.041	0.544	1.140	1.094
	대구	0.653	0.315	0.433	0.661
	인천	1.221	0.258	0.716	0.504
	광주	0.997	0.343	8.369	0.743***
	대전	1.328	0.291	0.454	0.591
	울산	0.623	0.442	13.640	0.937***
할인할증	할인40%	0.471	0.241***	1.299	0.493
	할인45%	0.593	0.271*	2.762	0.580*
	할인50%	0.575	0.272***	3.425	0.604**
	할인60%	0.640	0.275	1.785	0.583
	할인70%	0.560	0.282**	2.025	0.621
	할인80%	0.686	0.268	1.251	0.544
	할인90%	0.879	0.259	1.255	0.573
	할증110%	0.331	0.551**	0.100	1.178*
	할증120%	0.512	0.558	1.106	1.048
	할증130%	0.856	0.572	0.505	1.091
	할증140%	1.009	0.792	0.218	1.439
	할증150%~200%	0.443	0.469	0.553	1.944
가입경력	1년미만	0.581	0.350	3.941	0.684**
	1~2년미만	0.881	0.300	0.928	0.656
	2~3년미만	1.014	0.266	2.623	0.551*
	3~4년미만	0.903	0.227	1.332	0.579
교통법규	할증1그룹	1.409	0.344	1.243	0.627
	할증2그룹	0.897	1.059	0.211	2.157
	기본그룹	0.954	0.134	0.729	0.324

주 : 1) Robust 표준오차에 기준하여 *은 유의수준 10%, **은 유의수준 5%, ***은 유의수준 1% 수준에서 통계적 유의성을 나타낸다.

2) NB: LR test of alpha=0: chibar2(01)=152.8, Prob>=chibar2= 0.000, LR chi2(50)=53.9

3) Lognormal: AIC=33.1, BIC=-752.7, Scale Parameter=3.46, Deviance=973.52

4) 할증150~200% 그리고 할증1그룹1과 2의 위험변수에 해당 개체수의 부족으로 그룹화하여 분석한다.

4. 추가 논의사항

가. 적정 분석방법(또는 분석모형) 선택

선행연구에서 논의된 바와 같이 현재까지 자동차보험 요율상대도 산출에 사용되는 모형으로 개발된 것은 크게 단변량방법과 다변량방법으로 구분할 수 있다. 단변량 방법을 사용하면 특정 위험집단이 과대평가되거나 과소평가될 수 있는 단점이 있는 반면, 사용이 편리하다는 장점이 있다.

다변량방법은 특정위험집단에 대한 위험도 평가를 정확히 할 수 있다는 장점이 있으나, 계산과 해석이 복잡한 단점이 있다. 컴퓨터의 발달로 계산이 복잡한 다변량방법의 단점은 많이 희석된 상태이다. 다변량방법은 전통적인 기계적 방법과 통계적 방법이 있다. 기계적 방법은 통계적 방법이 도입되기 이전에 개발된 것으로 다변량의 요율상대도 값을 계산식에 따라 기계적으로 계산할 수 있다.

그러나 이 방법은 변수의 통계적 유의성 등을 판단할 수 없다는 단점이 있다. 이러한 단점을 극복하고 나온 방법이 통계적 방법이다. 통계적 방법도 많은 변화가 있었는데, 처음에는 전통적 회귀분석(OLS)을 이용하는 방법이 개발되었다. 그러나 전통적 회귀분석은 종속변수(사고빈도 및 사고심도)의 분포에 대한 부정확한 가정으로 분석을 하는 방법이다. 전통적 회귀분석을 시행할 때, 종속변수가 전통적 회귀분석에서 정한 가정에 부합하지 않은 분포인 경우 분석결과를 신뢰하기 어렵다. 이러한 전통적 회귀분석방법의 단점을 극복하기 위해 개발된 것이 일반화선형모형이다. 자동차보험 요율산출에서 일반화 선형모형을 적용한다는 것은 통계학의 발전단계에서 최신의 통계모형을 적용한다는 의미이다.

이러한 모형의 발전단계를 볼 때, 비교적 최근에 개발되었고 많은 계량경제학자와 통계학자에 의해 검증되어 모형이 안정화된 일반화선형모형을 사용하는 것이 자동차보험 요율상대도를 산출하기 위해 가장 효율성과 정확도가 높다고 할 수 있다. 통계모형을 사용하기 위해서는 앞에서 설명한 바와 같이 복잡한 통계모형의 적합과정을 거쳐야 한다. 통계학적으로 적합한

모형을 찾아내지 못한 상태에서 통계모형을 사용하는 것은 이전의 단변량 방법이나 전통적인 기계적 방법을 사용하는 것과 큰 차이가 없을 수 있으므로, 자동차보험 요율 상대도 산출시 이점을 유의하여야 한다.

참고로 자동차보험 요율산출에서 과거 단변량방법부터 일반화선형모형까지 요율상대도 산출결과 어떻게 변화되는지를 살펴보았다<표 Ⅳ-21>. 모형비교 분석에서는 논의의 이해를 돕기 위해서 통계모형 분석에 모든 위험변수가 포함되었음에도 불구하고 성과 가입연령에 해당하는 계수만을 도시하고 비교하였다. 분석대상 담보는 대인배상 I 이며, 종속변수는 사고빈도 이다.

전통적 회귀분석(OLS)의 경우에는 앞에서 설명하였듯이 사고빈도에는 적용하기에는 많은 단점이 있다. 사고빈도가 정규분포가 아니기 때문에 전통적 회귀분석은 자동차사고 위험도를 설명하기에 부적합하고, 이외에도 일반화선형모형과 단변량 분석과의 비교를 위해 빈도값을 로그로 전환해야 하는데 무사고를 뜻하는 0에는 로그값이 정의되지 않는 문제가 발생한다. 그럼에도 불구하고 단지 모델간의 비교를 위해 분석결과를 표에 삽입하였다. 이렇게 최소자승법(OLS)의 부적합성 때문에 구체적인 논의는 생략하기로 한다. 단변량으로 분석한 계수값과 일반화선형모형으로 분석한 계수값이 증가하거나 감소하는 흐름은 비슷하나 각 위험변수에 해당하는 계수값은 다소 차이를 보인다. 이것은 앞에서 설명했듯이 단변량 분석의 경우 하나의 위험변수의 계수값에 다른 위험변수들의 영향이 감안되어진 값이기 때문이다. 즉, 변수간에 상호작용(Correlation) 존재하기 때문이다. 예를 들면, 성별과 관련된 다른 위험변수의 영향이 남자라는 위험변수의 계수 0.850에 포함되어져 있다는 것이다.

<표 Ⅳ-21> 모형별 요율상대도 비교(대인배상Ⅰ, 사고빈도)

구 분		단변량	OLS	GLM(음이항분포)
성	남자	0.850	1.000	0.931
가입경력	1년미만	1.730	1.007	1.355
	1~2년미만	1.456	0.998	1.142
	2~3년미만	1.320	0.996	1.058
	3~4년미만	1.152	1.004	0.988

나. 변수선택 범위

일반화 선형모형을 사용하여 자동차보험 요율상대도를 산출하는데 있어서 어떤 변수를 모형에 포함시켜야할 지에 대한 논의가 필요하다. 즉 자동차보험 통계에는 성, 지역, 직업 등과 같이 실제 요율산출에 사용되지 않는 변수가 많이 있다. 즉, 자동차보험 요율상대도 산출 통계모형을 만들 때 이들 변수를 모두 포함시켜서 분석하는 것이 적합한가에 대한 검토가 필요하다.

자동차보험 요율을 결정할 때 판단기준은 ‘계리학적 기준’, ‘사회적 합의 기준’, ‘법률적 기준’등 여러 가지가 있다. 계리학적 기준으로는 요율산출에 모든 변수를 포함시키는 것이 가장 합리적이다. 자동차보험 요율제도는 법률적, 사회적 합의에 따라 운영되는 것이므로, 계리학적으로 합리적이라 하더라도 요율요소에 포함시킬 수 없는 변수들이 있다. 현재 우리나라 자동차보험 제도에서는 성, 지역, 직업 등이 요율요소로 도입되지 않고 있다. 그러므로 계리적 기준으로 이들 요소를 모두 포함시키는 것이 적합한 것으로 판단되더라도 이들 요소를 요율적용에서 제외하여야 한다.

그런데 사회적으로 미 합의된 변수라 하더라도 통계모형분석에는 이들 변수를 포함시켜야하지 않는가에 대한 논의가 필요하다. 먼저 사회적으로 합의되지 않은 변수(이하 ‘적용 불가능한 변수’라 한다)를 통계모형에 포함시키고, 실제 적용할 때에는 사회적으로 합의된 변수(이하 ‘적용 가능한 변수’라 한다)만의 요율상대도를 적용하는 경우를 생각할 수 있다. 또한 통계

모형에 포함시킬 변수로 적용 불가능한 변수를 제외하고 분석하는 방법을 생각할 수 있다. 앞서 분석방법은 full 모델이 되며, 뒤의 분석방법은 제한 모델이 된다. 통계모형을 적합 시킨 결과 full모델과 제한모델이 모두 적합한 것으로 나타났고, 두 모형에 대하여 ANOVA test를 해본 결과 full모델을 적용하는 것이 더 나은 모델로 나타나는 경우가 있다. 즉, ‘통계적으로 더 나은 모델인 full 모델- 즉, 사회적으로 합의되지 않은 변수를 포함한 모델-로 분석한 결과에서 사회적으로 합의된 변수의 값만을 적용하는 것이 더 나은 적용 방법인가’하는 논란의 소지가 있다.

이와 관련하여 두 모형의 차이를 비교하기 위하여 앞장에서 제시한 full 모델이외에 다음의 <표 Ⅳ-22>와 <표 Ⅳ-23>에서 성, 지역변수를 제외한 분석결과를 예시하였다. 여기서 <표 Ⅳ-22>는 모델을 실행한 결과이고, <표 Ⅳ-23>은 <표 Ⅳ-22>에서 분석된 결과를 상대도 값으로 변환한 표이다.

<표 Ⅳ-22> 대인배상Ⅰ: 일반화선형모델을 이용한 회귀분석 결과

구 분		사고빈도		사고심도	
		Coefficient	Std. Dev.	Coefficient	Std. Dev.
연령	25세이하	0.303	0.056***	-0.020	0.040
	26~30세	0.121	0.028***	0.011	0.024
	40~64세	0.176	0.018***	0.022	0.014
	65세이상	0.127	0.043***	0.044	0.035
차종	소형A	-0.189	0.024***	-0.072	0.027***
	소형B	-0.084	0.016***	-0.059	0.016***
	대형	-0.079	0.023***	0.048	0.024**
	승합1	-0.374	0.084***	0.093	0.114
	승합2	0.124	0.021***	0.062	0.017***
할인할증	할인40%	-0.512	0.018***	-0.053	0.029*
	할인45%	-0.280	0.026***	-0.055	0.033*
	할인50%	-0.225	0.027***	-0.041	0.032
	할인60%	-0.147	0.030***	0.011	0.033

할인할증	할인70%	-0.119	0.031***	-0.001	0.032
	할인80%	-0.089	0.031***	0.023	0.032
	할인90%	-0.023	0.032	0.018	0.031
	할증110%	0.131	0.053***	-0.048	0.043
	할증120%	0.164	0.063***	0.098	0.050**
	할증130%	0.328	0.086***	0.080	0.057
	할증140%	0.243	0.114***	0.221	0.085***
	할증150%	0.512	0.179***	0.144	0.101
	할증160%	0.422	0.224***	0.041	0.141
	할증170%	0.678	0.385***	0.131	0.188
	할증180%	0.375	0.403	0.188	0.260
	할증190%	0.534	0.561	0.115	0.331
	할증200%	0.878	0.658***	0.114	0.296
가입경력	1년미만	0.323	0.052***	0.061	0.036*
	1~2년미만	0.149	0.040***	0.019	0.033
	2~3년미만	0.070	0.035**	-0.036	0.030
	3~4년미만	-0.008	0.026	0.004	0.025
교통법규	할증1그룹1	-0.108	0.529	0.104	0.539
	할증1그룹2	0.053	0.031*	-0.020	0.030
	할증2그룹	0.116	0.123	0.100	0.105
	기본그룹	0.115	0.018***	-0.017	0.015

주 : 1) Robust 표준오차에 기준하여 *은 유의수준 10%, **은 유의수준 5%, ***은 유의수준 1% 수준에서 통계적 유의성을 나타낸다.

2) NB: LR test of alpha=0: chibar2(01)=268.03, Prob>chibar2 =0.000, LR chi2(54) = 1919.43

3) Lognormal: AIC=2.70, BIC=-200749.9, Scale parameter= .870, Deviance=19148.66

<표 Ⅳ-23> 일반화선형모형으로 추정된 모수들의 발생률(Incidence Rate Ratio)로의 전환: 상대위험도

구 분		사고빈도		사고심도	
		Coefficient	Std. Dev.	Coefficient	Std. Dev.
연령	25세이하	1.354	0.056***	0.981	0.040
	26~30세	1.129	0.028***	1.011	0.024
	40~64세	1.192	0.018***	1.022	0.014
	65세이상	1.136	0.043***	1.045	0.035
차종	소형A	0.828	0.024***	0.930	0.027***
	소형B	0.919	0.016***	0.943	0.016***
	대형	0.924	0.023***	1.049	0.024**
	승합1	0.688	0.084***	1.097	0.114
	승합2	1.132	0.021***	1.064	0.017***
할인할증	할인40%	0.599	0.018***	0.948	0.029*
	할인45%	0.756	0.026***	0.947	0.033*
	할인50%	0.799	0.027***	0.959	0.032
	할인60%	0.863	0.030***	1.011	0.033
	할인70%	0.888	0.031***	0.999	0.032
	할인80%	0.914	0.031***	1.023	0.032
	할인90%	0.977	0.032	1.018	0.031
	할증110%	1.140	0.053***	0.954	0.043
	할증120%	1.178	0.063***	1.103	0.050**
	할증130%	1.388	0.086***	1.084	0.057
	할증140%	1.275	0.114***	1.247	0.085***
	할증150%	1.668	0.179***	1.155	0.101
	할증160%	1.524	0.224***	1.042	0.141
	할증170%	1.969	0.385***	1.141	0.188
	할증180%	1.455	0.403	1.207	0.260
	할증190%	1.705	0.561	1.122	0.331
	할증200%	2.406	0.658***	1.121	0.296

가입경력	1년미만	1.381	0.052***	1.063	0.036*
	1~2년	1.161	0.040***	1.019	0.033
	2~3년	1.073	0.035**	0.965	0.030
	3~4년	0.992	0.026	1.004	0.025
교통법규	할증1그룹1	0.898	0.529	1.109	0.539
	할증1그룹2	1.054	0.031*	0.980	0.030
	할증2그룹	1.123	0.123	1.105	0.105
	기본그룹	1.122	0.018***	0.983	0.015

주 : Robust 표준오차에 기준하여 *은 유의수준 10%, **은 유의수준 5%, ***은 유의수준 1% 수준에서 통계적 유의성을 나타낸다.

성별과 지역을 포함한 일반화선형모형의 결과를 도시한 <표 IV-12> 와 성별과 지역을 제외한 분석결과를 도시한 <표 Ⅳ-23>를 비교해 보면 일부 위험변수의 계수값은 비슷하지만 일부 위험변수의 경우는 계수값에 상당한 차이가 있다. 성별, 지역과 같은 미적용 위험변수를 포함한 모형과 제외한 모형에서 도출된 차종의 추정계수를 비교하여 보면 <표 IV-24>에서와 같이 성별과 지역의 포함여부는 차종의 계수에 큰 영향을 미치지 않는다. 반면 차종 이외의 가입경력 및 연령등과 같은 위험변수들의 계수값들은 <표 IV-12>와 <표 IV-23>에서 확인할 수 있는 것처럼 운전자의 성별과 지역의 포함여부에 따라 변한다.

<표 Ⅳ-24> 성별과 지역의 포함여부와 차종의 계수 비교

구 분		사고빈도		사고빈도	
		Coefficient	Std. Dev.	Coefficient	Std. Dev.
		성별과 지역 포함한 모형		성별과 지역 제외한 모형	
차종	소형A	0.843	0.029***	0.828	0.024***
	소형B	0.924	0.017***	0.919	0.016***
	대형	0.918	0.025***	0.924	0.023***
	승합1	0.671	0.123***	0.688	0.084***
	승합2	1.133	0.018***	1.132	0.021***

주 : Robust 표준오차에 기준하여 *은 유의수준 10%, **은 유의수준 5%, ***은 유의수준 1% 수준에서 통계적 유의성을 나타낸다.

이것은 성별과 지역 등의 위험변수를 모델에 포함하지 않을 경우에 성별과 지역의 위험변수 값은 오차항(Error Term)에 포함된다는 것을 의미한다. 이렇게 위험변수를 오차항에 포함시켜 일반화선형모형을 분석할 경우는 성별과 지역의 영향을 모델에서 독립(Control)시키지 않는다는 뜻이 된다. 이는 성별과 지역을 오차항에 포함시킬 경우 계수가 변하는 위험변수들은 오차항에 포함된 성별 또는 지역과 상관관계가 존재한다는 뜻이다. 반대로 차종처럼 계수가 변하지 않는 위험변수들은 성별과 지역변수와는 상관관계가 존재하지 않는다는 것을 의미한다. 즉, 성별과 지역을 오차항에 포함시킨다고 해도 이들의 상대위험도가 어느 정도는 다른 위험변수들(계수가 변하는 위험변수들)에 반영되어진다는 뜻이다.

이상의 사실을 바탕으로 통계적 또는 계량경제학적으로 full 모델이 제한모델보다 더 나은 모델인 것으로 해석될 수 있으나, 요율적용이라는 측면에서는 어떤 방법이 더 나은 것이라고 판단하기가 어렵다는 것을 알 수 있다. 본 분석에서 사용된 지역과 성별이외의 자료가 다시 분석에 사용된다면, 현재의 full모델보다 더 나은 모델이 나타난다. 이 경우 새로 분석된 모델의 변수의 상대도 값들은 다시 변하게 된다. 이러한 것은 변수가 새로 추가될 때마다 새로운 모델이 만들어지고, 새로운 상대도값(coefficients)이 산출된

다는 의미이다. 즉, 통계적 또는 계량경제학적으로 유의한 무한이 많은 모델들이 개발될 수 있다는 것이다.

자동차보험 요율상대도 값을 산출하는데 통계적, 계량경제학적 측면만을 생각하여 변수를 추가함으로써 만들어지는 무수히 많은 모델들을 모두 인정하는 것은 자동차보험 요율적용의 혼란만을 야기할 수 있다. 따라서 자동차보험 요율상대도 산출에 일반화선형모형을 적용하는 데 변수선택의 범위는 일정한 수준으로 하고 정한 범위 내에서 모든 보험회사들이 요율을 산출·적용하는 것이 요율적용 측면에서 합리적일 것으로 판단된다. 변수선택의 범위는 이제까지 자동차보험 요율산출분야에서 해왔던 것처럼 사회적, 법률적 기준에 따라 정해지는 것이 이해관계자가 수용하기에 용이한 것으로 보인다.

다만, 언더라이팅 분야는 적용 불가능한 변수를 포함하여 모형을 적용할 수 있기 때문에, 통계모형을 활용하여 언더라이팅 기준을 마련하는데 적용불가능한 변수를 포함시킬 수 있을 것이다. 이 경우 회사는 이익이 나는 계층(즉, 손해율이 낮은 계층)에 영업을 집중할 수 있으므로 더 나은 이익을 실현할 수 있을 것이다. 새로운 변수를 발굴하고, 이를 언더라이팅에 적용하는 이러한 능력은 보험회사의 몫이 될 것이다.

변수의 범위에 대한 이상의 논란의 소지에도 불구하고 본 연구에서는 현재 자동차보험 제도에서는 사용되지 않고 있는 변수인 성, 지역변수를 포함시켰다. 향후 동 변수가 자동차보험 요율제도로 포함될 수도 있다고 판단되었고, 본 연구의 목적이 자동차보험 요율상대도 산출에 일반화선형모형을 적용하는 방법을 연구하는 것이므로 분석에 사용된 변수를 제약할 필요가 없다고 판단되었기 때문이다.

다. 산출계수의 사회적 수용성

사회적으로 합의된 요율계수가 포함된 일반화 선형모형으로 요율상대도 값을 산출하더라도, 요율계수의 수준 등에서 사회적 및 법률적 수용성을 고려하여야 한다. 이러한 측면에서 본 연구의 분석결과를 보면, 통계모형으로

산출된 할인할증요율 계수의 범위가 최대한도 200%를 초과하는 경우도 있었다. 현재 자동차보험 할인할증률의 한도는 40%~200%이므로, 통계모형으로 산출된 할인할증 계수 중 200%를 초과하는 계수는 200%한도로 제한하여야 한다.

차종별 위험상대도를 고려할 때, 대체적으로 대형차로 갈수록 사고위험도(사고빈도×사고심도)가 큰 것으로 나타났다. 그런데 통계의 종류에 따라서 대형자동차보다 작은 자동차가 사고위험도가 더 높을 수 있다. 이 경우는 소형차가 사고위험도가 더 낮을 것이라고 생각하는 사회적 정서에 반하는 결과이므로, 자동차보험 요율의 사회적 수용성측면에서는 요율 수준을 어떻게 결정할 지에 대한 고민이 필요하다.

라. 교호작용(Interaction) 효과 분석[25]

교호작용 변수들을 적용하는 방법에는 크게 더미변수들 간의 교호작용 변수들을 만들어 적용하는 방법과 더미변수와 연속변수간 교호작용 변수를 만들어 적용하는 방법이 있다. 후자는 전자에 비해 비교적 모델설정과 해석에서 복잡하다. 하지만 자동차 보험 요율분석은 위험변수들을 비슷한 위험군으로 나누고 각 위험군을 대표하는 더미변수를 이용하여 이루어지기 때문에 후자는 본 보고서의 논의대상에서 제외하였다. 교호작용 변수들을 고려하고 분석하는 것은 일반화선형모형에서 매우 중요한 부분이다. 적합한 교호작용 변수들을 고려하지 않고 모델을 분석했을 경우에는 과대산포의 문제가 없는데도 있는 것처럼 과대산포 테스트의 결과가 반대로 나올 수 있기 때문이다. 이밖에도 교호작용 변수들은 연구자로 하여금 좀 더 현실적인 분석모델을 구축할 수 있도록 한다.

그러나 일반적으로 자동차 요율분석의 경우처럼 대부분의 위험변수들이 비슷한 위험도에 따라 세분화되어 각 세분화된 위험군을 대표하는 수많은 더미변수를 이용하여 분석되어지는 경우에는 교호작용을 적용할 필요성이 떨어지게 된다. 예를 들어, 가입경력과 사고경력에 따른 할인할증의 교호작

25) 대인배상 I 통계를 사용하여 설명하였다.

용을 분석한다고 가정했을 경우 추가해야할 교호작용 변수만도 약 100여개에 달한다. 이렇게 많은 수의 변수를 통계모형에 적용하는 것이 현실적으로 쉽지 않으며 적용한다고 해도 이렇게까지 세분화된 요율적용이 현실적으로 불필요하기 때문이다. 또한, 이미 세분화된 더미변수들이 요율에 적용되고 있는 경우에 유의한 교호작용을 찾아냈다고 하더라도 대부분의 유의한 교호작용에서 구해지는 위험상대도는 교호작용 변수에 이용된 본래의 위험변수의 유의성을 변화시키거나 추정된 계수의 값을 변화시켜 나오기 때문에 교호작용의 적용이 불필요한 경우가 대부분이다.

하지만 성별처럼 세분화될 수 없는 위험변수의 경우에는 교호작용을 분석해야할 필요성이 있다. 예를 들면, 지금까지 분석한 모델은 자동차 가입경력의 차이 때문에 생기는 자동차사고 위험도가 성별에 따라 차이가 없다는 가정 하에 만들어진 모델이다. 하지만, 자동차보험 최초 가입자의 경우 남자운전자와 여자운전자 사이에 자동차 사고 상대위험도가 다를 수 있다. 비슷한 예로, 지금까지 분석한 모델은 차종의 차이 때문에 생기는 자동차사고 위험도가 성별에 따라 차이가 없다는 가정 하에 만들어진 모델이다. 하지만, 같은 초보 운전자라고 하더라도 또는 같은 대형차를 운전하더라도 남자가 여자보다 공간 지각 능력이 높아 자동차사고 위험도가 낮을 수도 있다는 것이다. 이러한 현실적인 논리의 검증을 위해 교호작용 변수가 필요한 것이다.

<표 IV-25> 및 <표 Ⅳ-26>에 보는 바와 같이 남성운전자와 여성운전자간 차종에 따른 위험도 차이는 없는 것으로 나타났으며, 사고심도 분석에서도 성별과 차종의 교호작용은 없는 것으로 나타났다.

<표 Ⅳ-25> 성별과 차종을 이용한 교호작용: 사고빈도

구 분	Coefficient	Std. Dev.
남성이면서 소형A 운전자	0.057	0.060
남성이면서 소형B 운전자	0.080	0.062
남성이면서 대형 운전자	-0.054	0.040
남성이면서 승합1 운전자	-0.050	0.045
남성이면서 승합2 운전자	0.232	0.326
다른 모든 변수들	생략	생략

주 : 1) Robust 표준오차에 기준하여 *은 유의수준 10%, **은 유의수준 5%, ***은 유의수준 1% 수준에서 통계적 유의성을 나타낸다.
2) 결과는 모든 다른 변수들을 포함하여 분석한 결과이며 교호작용 변수 이외의 변수에 해당하는 값들은 편의상 생략하였다.

<표 Ⅳ-26> 성별과 차종을 이용한 교호작용: 사고심도

구 분	Coefficient	Std. Dev.
남성이면서 소형A 운전자	0.027	0.040
남성이면서 소형B 운전자	0.004	0.026
남성이면서 대형 운전자	-0.062	0.035*
남성이면서 승합1 운전자	-0.160	0.299
남성이면서 승합2 운전자	-0.024	0.028
다른 모든 변수들	생략	생략

주 : <표 Ⅳ-24>와 동일하다.

이렇게 성별과 차종 간에 교호작용이 없을 경우 교호작용 변수들을 통계분석 모델에 포함시키는 것은 의미가 없으며, 무리하게 이러한 교호작용 변수들을 모델에 포함시킬 경우 <표 Ⅳ-27>에서처럼 통계모델의 설명력이 오히려 떨어진다.

<표 Ⅳ-27> 성별과 차종의 교호작용 적용여부 검증: 사고빈도

구　　분	AIC	BIC
교호작용 변수들을 적용하지 않은 모델(A)	185441.1	186062.6
교호작용 변수들을 적용한 모델(B)	185442.2	186074.8
차　이(B-A)	1.1	12.2

연령에 따른 성별 간에도 위험도 차이가 있을 수 있다는 가정을 검증하기 위해 연령과 성별을 이용한 교호작용을 이용해 분석결과를 <표 Ⅳ-28>에 표시하였다. 나이를 기준으로 같은 위험군에 속하는 남성과 여성 운전자 사이에 위험도 차이는 나타나지 않았다.

<표 Ⅳ-28> 성별과 연령을 이용한 교호작용: 사고빈도

구　분	Coefficient	Std. Dev.
남성이면서 25세 이하	0.119	0.085
남성이면서 26~30세	0.042	0.054
남성이면서 40~64세	0.008	0.035
남성이면서 65세 이상	0.170	0.104
다른 모든 변수들	생략	생략

주 : 1) Robust 표준오차에 기준하여 *은 유의수준 10%, **은 유의수준 5%, ***은 유의수준 1% 수준에서 통계적 유의성을 나타낸다.
2) 결과는 모든 다른 변수들을 포함하여 분석한 결과이며 교호작용 변수 이외의 변수에 해당하는 값들은 편의상 생략한다.

이러한 방법으로 성별과 기타 위험변수들을 이용하여 교호작용을 분석하였다. 대부분의 요율분석 모델에서 교호작용 부분이 필요 없는 것으로 나타났다. 다만 성별과 가입경력간에는 교호작용 효과가 있는 것으로 나타났으며, 그 결과를 <표 Ⅳ-29>와 <표 Ⅳ-30>에 표시하였다.

<표 Ⅳ-29> 성별과 가입경력을 이용한 교호작용: 사고빈도

구 분	Coefficient	Std. Dev.
남성이면서 가입경력 1년미만	0.118	0.053***
남성이면서 가입경력 1~2년미만	0.089	0.060
남성이면서 가입경력 2~3년미만	0.035	0.060
남성이면서 가입경력 3~4년미만	0.086	0.054
다른 모든 변수들	생략	생략

주 : 1) Robust 표준오차에 기준하여 *은 유의수준 10%, **은 유의수준 5%, ***은 유의수준 1% 수준에서 통계적 유의성을 나타낸다.
2) 결과는 모든 다른 변수들을 포함하여 분석한 결과이며 교호작용 변수 이외의 변수에 해당하는 값들은 편의상 생략한다.

최초가입자의 경우 남성운전자의 위험도가 여성운전자의 위험도보다 사고빈도 및 사고심도 모두에서 높게 나타났으며 통계적으로도 유의하였다. 하지만 가입경력이 1년 이상인 경우에는 성별에 따른 위험도 차이는 나타나지 않았다. 그래서 남성이면서 가입경력이 1년 미만인 운전자의 교호작용만을 이용하여 사고빈도와 사고심도의 일반화선형모형을 분석하였다.

<표 Ⅳ-30> 성별과 가입경력을 이용한 교호작용: 사고심도

구 분	Coefficient	Std. Dev.
남성이면서 가입경력 1년미만	0.117	0.039***
남성이면서 가입경력 1~2년	0.053	0.044
남성이면서 가입경력 2~3년	0.035	0.042
남성이면서 가입경력 3~4년	0.030	0.036
다른 모든 변수들	생략	생략

주 : 1) Robust 표준오차에 기준하여 *은 유의수준 10%, **은 유의수준 5%, ***은 유의수준 1% 수준에서 통계적 유의성을 나타낸다.
2) 결과는 모든 다른 변수들을 포함하여 분석한 결과이며 교호작용 변수 이외의 변수에 해당하는 값들은 편의상 생략한다.

이상의 교호작용 효과를 감안한 일반화 선형모형의 분석결과는 다음 <표 Ⅳ-31>과 같다.

<표 Ⅳ-31> 교호작용을 이용한 일반화선형모형

구 분		사고빈도		사고심도	
		Coefficient	Std. Dev.	Coefficient	Std. Dev.
연령	25세이하	1.398	0.059***	1.107	0.038***
	26~30세	1.145	0.029***	1.017	0.018
	40~64세	1.186	0.018***	1.044	0.010***
	65세이상	1.142	0.043***	1.063	0.025**
성	남자	0.922	0.016***	1.014	0.011
차종	소형A	0.843	0.025***	0.811	0.019***
	소형B	0.924	0.016***	0.883	0.012***
	대형	0.918	0.023***	1.307	0.015***
	승합1	0.671	0.082***	0.912	0.100
	승합2	1.133	0.021***	1.009	0.011
지역	부산	0.832	0.026***	0.913	0.022***
	경기	0.929	0.019***	1.034	0.012***
	강원	0.813	0.034***	1.014	0.027
	충북	0.785	0.034***	1.085	0.027***
	충남	0.798	0.030***	1.071	0.024***
	전북	1.019	0.038	1.053	0.027*
	전남	0.725	0.033***	0.991	0.027
	경북	0.765	0.026***	0.969	0.021
	경남	0.724	0.024***	0.998	0.019
	제주	0.546	0.044***	0.839	0.043***
	대구	0.884	0.028***	0.961	0.021*
	인천	1.163	0.034***	1.028	0.020
	광주	0.897	0.036***	0.896	0.025***
	대전	0.957	0.035	1.019	0.024

	울산	0.779	0.037***	0.837	0.032***
할인할증	할인40%	0.612	0.019***	0.757	0.022***
	할인45%	0.771	0.027***	0.814	0.023***
	할인50%	0.810	0.028***	0.868	0.024***
	할인60%	0.870	0.030***	0.920	0.025***
	할인70%	0.891	0.031***	0.942	0.025**
	할인80%	0.915	0.031***	0.972	0.025
	할인90%	0.976	0.032	0.964	0.024
	할증110%	1.120	0.052**	0.987	0.034
	할증120%	1.149	0.061***	1.023	0.041
	할증130%	1.351	0.083***	1.098	0.050*
	할증140%	1.240	0.111**	1.096	0.072
	할증150%	1.601	0.172***	1.029	0.103
	할증160%	1.468	0.216***	1.397	0.132***
	할증170%	1.875	0.366***	1.141	0.171
	할증180%	1.402	0.388	1.716	0.245**
	할증190%	1.635	0.537	0.973	0.294
	할증200%	2.281	0.623***	1.141	0.266
가입경력	1년미만	1.282	0.062***	0.981	0.035
	1~2년미만	1.144	0.040***	0.986	0.026
	2~3년미만	1.058	0.034*	1.009	0.023
	3~4년미만	0.988	0.026	0.987	0.018
교통법규	할증1그룹1	0.938	0.552	1.084	0.294
	할증1그룹2	1.091	0.035***	1.002	0.022
	할증2그룹	1.125	0.123	1.010	0.077
	기본그룹	1.129	0.018***	0.995	0.010
남성*1년미만		1.101	0.056*	1.113	0.039***

주 : 1) Robust 표준오차에 기준하여 *은 유의수준 10%, **은 유의수준 5%, ***은 유의수준 1% 수준에서 통계적 유의성을 나타낸다.
2) NB: LR test of alpha=0: chibar2(01)=251.25, Prob>chibar2=0.000, LR chi2(54) = 2319.56
3) Lognormal: AIC=2.69, BIC=-200787.31, Scale parameter=.862, Deviance=18941.27

교호작용 효과를 감안한 분석결과와 감안하지 않은 분석결과가 실제 요율적용에 차이가 있는가 하는 문제에 대해서는 논란의 여지가 있으나, 요율적용 측면에서 그 차이는 아래와 같이 미미한 것으로 분석되었다.

<표 IV-32> 성별과 가입경력에 따른 요율분석: 교호작용 적용 vs. 교호작용 비적용

구 분	위험상대도	
	교호작용 비적용	교호작용 적용
남성이면서 가입경력 1년미만	1.313(A)	1.316(B)
차이(B-A)	0.003	

주 : 1) 유의수준 5%에서 통계적으로 유의한 추정계수를 적용하여 계산된다.
2) 1.313=0.931×1.041×1.355
3) 1.316=0.922×1.282×1.113

먼저 남성이면서 가입경력이 1년 미만인 운전자의 상대위험도를 교호작용을 적용하지 않은 일반화선형모형으로 분석하였다. 다음에는 교호작용이 포함된 일반화선형모형을 분석하였다. 그 결과 <표 IV-32>에서처럼 교호작용을 이용하여 계산된 ‘남성이면서 가입경력이 1년 미만인 운전자’의 상대위험도는 교호작용을 이용하지 않은 모델에서 계산된 상대위험도보다 불과 0.003이 높은 것에 불과하였다. 즉, 거의 차이가 없는 것으로 분석되었다. 이는 성별 그리고 가입경력 1년 미만에 해당하는 교호작용 추정계수가 교호작용을 이용하지 않은 일반화선형모형의 성별 및 가입경력 요율로 이전되었기 때문이다.

교호작용변수의 적용여부를 확인하기 위해 두 모델의 회귀분석을 실시한 뒤 다양한 값을 계산하여 비교한 결과가 <표 IV-33>에 나타나 있다. 사고빈도의 LR 테스트와 사고심도의 Deviance 및 AIC 검증을 통해서는 교호작용을 적용한 모델과 그러하지 않은 모델사이에 차이가 거의 없었다. 그리고 BIC를 비교한 결과에서는 교호작용을 적용하지 않은 모델이 더 적합한 것으로 나타났다.

<표 Ⅳ-33> 교호작용 적용의 정당성 검증

구 분	교호작용 비적용	교호작용 적용
사고빈도 LR chi2(50)	2316.1	2319.6
사고심도 Deviance	18941.2	18941.3
사고빈도 AIC	185440.9	185439.4
사고빈도 BIC	186018.0	186027.6
사고심도 AIC	59284.2	59286.1
사고심도 BIC	59692.2	59702.1

지금까지 교호작용을 분석한 결과를 종합하여 볼 때, 일반화선형모형에서 교호작용이 이론적으로는 매우 중요하지만 자동차 요율분석처럼 대부분의 위험변수들이 비슷한 위험도에 따라 세분화되어 있는 경우에는 교호작용의 적용이 큰 의미가 없을 수도 있다. 또한 통계적으로 유의한 교호작용이 발견되었을 경우에도 사회·문화적 통념상 현실적으로 적용하기 어려운 부분이 있을 수 있으며 무엇보다 교호작용은 분석자의 임의성이 많이 반영될 수 있다는 단점이 있다. 그러므로 교호작용은 분석자의 주의가 필요한 부분이며 감독당국에서도 명확한 기준을 마련해야 할 부분이다.

마. 일부 변수의 기존 상대도 유지 시 분석방법(Offset활용)

기존의 전통적 분석방법에 비해 일반화선형모형이 자동차사고 위험도를 예측하는데 더 뛰어난 이론이며 정확한 실증방법임에는 의심의 여지가 없다. 이러한 이슈에 대해서는 이미 많은 학자들에 의해서 증명돼 왔으며 주요국에서는 일반화선형모형을 이용해 보험료율을 차등 적용하는 것은 더 이상 새로운 것이 아닌 당연한 것이 되었다. 이미 보험요율분석에 일반화선형모형을 이용하고 있는 주요국처럼 자동차보험을 시작으로 일반화선형모형을 도입하는 것이 시급한 상황이다. 하지만 어떻게 도입을 하는가에 대해서는 논의가 필요하다.

보험사는 이미 전통적인 분석방법으로 계산한 보험료를 부과하고 있고 소비자는 역시 그 보험료를 지불하고 있다. <표 IV-21>에서 보았듯이 일반화선형모형에 의해 추정된 위험변수의 계수는 전통적인 분석방법에 의한 계수와 상이하다. 즉, 일반화선형모형을 이용해 분석한 계수를 이용해 보험료를 부과할 경우 경제학적으로는 보험사와 소비자 모두 잉여(Surplus)가 증가하게 된다. 하지만 일반화선형모델을 통해 분석한 보험료를 일괄적으로 부과할 경우 일부 소비자는 기존에 지불하고 있는 보험료와 새로운 보험료가 매우 상이하여 불만을 가질 수 있으며, 이 경우 보험사로써는 그러한 소비자들에게 일반화선형모형을 통한 새로운 기준의 보험료에 대해 이해시켜야하는 어려움을 겪을 수 있다.

이러한 현실적인 문제들 때문에 이론적으로 더욱 적합한 일반화선형모형이라 할지라도 새로운 분석모델을 현실 적용단계에서는 점차적으로 적용해야 한다. 즉 보험료를 계산할 때 일반적인 위험변수에 해당하는 상대위험도를 위해 기존 값을 그대로 유지·적용하면서 특정 위험변수만을 일반화선형모형에 우선 적용하여 상대위험도를 계산한다. 그리고 점차 기존 값을 그대로 유지·적용하는 부분을 줄이고 특정 위험변수를 증가시켜 일반화선형모형의 적용을 확대시켜나가는 방법이 현실적으로 적합하다. 이러한 현실적인 문제를 고려한 것이 일반화선형모형의 ‘Offset’ 부분이다. 일반화선형모형의 한 부분인 체계적 성분(Systematic Factor)인 $X\beta$를 기존 값을 그대로 유지·적용하는 부분에 해당하는 위험변수와 일반화선형모형을 적용할 특정 위험변수로 구분한다. 그리고 일반화선형모형을 적용할 특정 위험변수 이외의 변수에는 기존 값을 그대로 적용시켜주는데 이 부분을 Offset이라고 한다. 비확률부분(Nonstochastic)인 이 Offset부분을 제외한 체계적 성분에 해당하는 일부 위험변수만을 일반화선형모형에 적용하여 상대위험도를 추정하고 이후 수정하고자 하는 변수를 추가함으로써 Offset 부분을 점차 줄여나간다. 포아송을 예로 들면, 식(III-12)의 평균 $\mu=\exp(x\beta)$에 Offset을 감안할 경우 $\mu=\exp^{x\beta+offset}$ 과 같이 된다.

Ⅴ. 결론 및 시사점

자동차보험을 통해 운전자의 리스크를 인수한 보험회사의 안정적인 경영과 보험가입 고객의 형평성 제고를 위해 정확한 요율상대도 분석은 아무리 강조해도 지나치지 않다.

본 연구에서는 단변량 또는 기계적인 다변량 요율산출 방법을 주로 사용하는 국내의 요율산출 시스템에 일반화선형모형이라는 통계적 분석모형의 이론과 실증분석 방법을 소개하고, 이미 일반화선형모형을 자동차보험에 적용중인 주요국의 문제점을 살펴보았다. 그리고 지수형태의 분포를 가지는 반응변수(Response Variable)를 건수변수(Count Variable)와 연속변수(Continuous Variable)로 구분하여 두형태의 반응변수에 여러 가지 일반화선형모형을 적용하여 보았다. 우리나라 자동차보험 통계에서 담보별로 최적인 모형을 찾아낸 결과는 <표 Ⅴ-1>와 같다.

<표 Ⅴ-1> 요율분석에 적용된 일반화선형모형(GLM)

담보	적용모델		
	사고빈도 GLM	사고심도	
		GLM	링크함수(Link)
대인배상I	음이항분포	Lognormal	항등(identity)
대인배상Ⅱ	음이항분포	감마(Gamma)	로그(Log)
자기차량	ZINB	감마(Gamma)	로그(Log)
대물배상	음이항분포	Lognormal	항등(identity)
자기신체사고	음이항분포	Lognormal	항등(identity)
무보험차상해	음이항분포	감마(Gamma)	로그(Log)

미국, 영국, 프랑스와 같은 주요국에는 일반화선형모형을 자동차요율분석에 적용하고 있지만 다양한 일반화선형모형을 비교·분석하여 최적의 분석

모델을 선택하기 보다는 사고빈도 부분에서는 포아송(Poisson)모델을 그리고 사고 심도에는 감마(Gamma)를 이용하여 일괄적으로 적용하고 있다. 하지만 분석결과 사고빈도에는 담보에 상관없이 포아송보다는 음이항분포가 최적의 일반화선형모형이었으며 사고심도에는 감마(Gamma)와 로그노말(Lognormal)이 적합하였다. 특히 감마(Gamma) 모델에는 역연결함수(Inverse Link Function)보다는 로그연결함수(Log Link Function)가 가장 적절한 모형으로 판명되었다.

그러므로 보험사가 일반화 선형모형을 자동차보험 요율산출에 적용할 때, 이미 선행 연구결과를 단순히 사용하는 것보다는 자사의 자료와 가장 일치하는 일반화선형모형을 찾아내는 것이 바람직하다. 또한 이론적으로는 일반화선형모형에서 교호작용 적용의 여부가 매우 중요할 수도 있겠으나 실증분석의 결과를 볼 때 교호작용을 요율차별화에 적용하는 것은 추가적인 분석이 필요하며 감독당국과의 협의도 필요한 부분이라고 판단된다.

사회·문화적 이유로 인해 아직 자동차보험요율 차등화에 적용되지 않고 있는 성별과 지역은 경제학적인 그리고 통계학적인 기준으로는 운전자의 상대위험도를 좀 더 정확하게 분석할 수 있는 중요한 위험변수인 것으로 분석되었다. 성별과 지역에 대한 위험상대도를 자동차보험 요율차등화에 반영하는 방안이 그동안 많은 연구물을 통해 논의되어 왔지만 본 연구의 결과는 기존 연구에 적용된 모델들과는 달리 데이터의 성격이 잘 반영된 분석모델을 사용한 것이므로, 그 분석결과가 의미하는 바가 크다고 할 수 있겠다. 물론 통계적 또는 계량경제학적인 기준으로 분석되어진 각 변수의 위험상대도를 요율차등화에 그대로 적용하기는 쉽지 않다. 특히 자동차보험에는 의무담보(대인배상Ⅰ 및 대물배상)가 포함되어 있으므로, 모든 운전자는 자동차보험 제도의 변화에 큰 영향을 받는다. 따라서 자동차보험의 요율체계의 변화는 사회적·문화적 그리고 정치적으로도 이해당사자간의 공감대 형성이 중요하다.

그리고 통계적·계량경제학적 분석을 통해 제시되는 위험변수의 요율상대도는 시장상황 또는 보험사간 경쟁상황에 따라 임의로 조정될 수 있지만, 그 중심에는 정확한 요율분석이라는 철학이 자리하여야 한다.

본 연구에서는 통계의 이론적 측면 뿐 아니라 실무적 측면에서 일반화선형모형을 요율산출분야에 적용하는 방법을 다양하게 살펴보았다. 그러나 요율산출분야이외에 준비금 산출 분야, 신뢰도 이론 분야 등에도 일반화선형모형을 적용할 수 있는데, 연구범위가 요율산출분야만으로 한정한 점은 다소 아쉬운 부분이다. 그리고 본 연구에서는 Poisson, Negative Binomial, Zero Inflated Possion, Zero Inflated Negative Binomial Hurdle, Zero-Truncated Poisson, Zero-Truncated Negative Binomial, Zero-Inflated Negative Binomial, Zero-Inflated Poisson, Gaussian, Lognormal, Log-Normal Gamma, Inverse Gaussian, Gamma, Log-Normal 등에 한정하여 보험통계에 적합한 분포를 확인하였다. 보험통계에 적합한 분포는 이들 분포이외에 더 많이 있고, 보다 적합한 분포를 더 개발할 수도 있을 것이다. 본 연구가 자동차보험 요율산출에 일반화선형모형을 적용하는 방법에 한정되어 진행하였기 때문에, 추가 모형을 개발하는데 다소 소홀한 점이 있는 것은 본 연구의 한계점이다.

따라서 향후에는 일반화선형모형을 적용할 수 있는 통계적, 실무적 방법에 대한 연구가 요율산출분야 이외로 확대되고, 자동차보험 통계에 더욱 부합된 분포(distribution)를 찾아서 이 분포를 사용한 일반화선형모형을 개발하는 것까지 연구가 확장될 필요가 있다고 생각한다.

참고문헌

기승도, 「자동차사고의 사회적 비용 최소화 방안」, 보험연구원, 연구보고서 2009-2, p107, 2009, p111

보험개발원, 『자동차보험 참조순보험료율서 신고서』, 2008.3

최우석・한상일, 「이중일반선형모형(DGLM)을 이용한 자동차보험 요율 추정」, 보험개발연구 제19권 제3호, 2008, pp.37~57

Akaike, H. , "A new look at the statistical model identification", IEEE Transactions on Automatic Control 19 (6): 716723, 1974

Anderson,D 외5인, "A Practitioner's Guide to Generalized Linear Model', *Casualty Acturial Society*, p5

Apostolaskis, N., *Tariff classification in auto insurance*, M.Sc. Thesis, Department of Statistics, Athens University of Economics, ISBN : 960-7929-00-4 ,1998

Bjorn Sundt, 『An Introduction to Non-Life Insurance Mathematics』, Karlsruhe

Boland,P.J.(1993), 『Statistical and Probabilistic Method in Actural Science』, Chapman & Hall/CRC, 2007

Cameron A.C., Trivedi P.K., "Regression Analysis of Count Data", Economic Society Monograph No. 30, Cambridge University Press, Cambridge, October, 1998

______, "Micro Econometrics: Methods and Applications", Cambridge University Press, New York, May 2005

Casualty Actuarial Society, *Pricing*, Volume I , 1990, pp.89~113

______, *Foundations of Casualty Actuarial Science*, Fourth Edition, 2001, pp.90~91

David,C.G. and Morrisey,M.,A,"The Effect of State Regulations on Motor Vehicle Fatalities for Younger and Older Drivers : A

Review and Analysis", *The Milbank Quarterly*, vol 79, 2001

Delucia P. and Mather R. D., "Motion extrapolation of car-following scenes in younger and older drivers," *Human Factors; Winter2006,* Vol. 48 Issue 4, 2006, p666-674

Gelman,A. and Hill,J., *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge University press, 2007

Heberman,S. and Renshaw,A.E., "Generalised Linear Models in Actuarial Work", *Staple Inn Actuarial Society*, 1988

Holler, K.D. and Sommer,D. and Trahair,G., "Something Old, Something New in Classification Ratemaking With a Novel Use of GLMs for Credit Insurance", *CAS*, 1999

Ismail,N. and Jemain,A.A.,,"Handling Over dispersion with Negative Binomial and Generalized Poisson Regression Models", *Casualty Actuarial Society Forum, Winter 2007*, pp.103~158

James,W.H. and Hilbe,J.M., *Generalized Linear Models and Extensions*, STATA Press, Second Edition, 2007

Jong,P.D. and Heller,G.Z., *Generalized Linear Model for Insurance Data*, Cambridge, 2008

Laflamme, L. and Hasselberg, M., Kullgren, A., Vaez, M., Delucia P., Mather R. D.,"MFirst car-to-car crashes involving young adult drivers: main patterns and their relation to car and driver characteristics.," *International Journal of Injury Control & Safety Promotion; Sep2006, Vol. 13 Issue 3,* 2006,p179-186

Luyang F. and Moncher,R.B., "Severity Distributions for GLMS : Gamma or Lognormal? Evidence from Monte Carlo Simulations", *CAS*, pp.149~230

Mack,T. "A simple parametric model for rating automobile insurance or estimating IBNR claims reserves", *ASTIN Bulletin 21(1)*, 1991, pp.93~109

Mack,T.,"Which stochastic model is underlying the chain ladder method?", *Insurance ; Mathematics and Economics 15,* 1994, pp.133~138

McCullagh,P. and Nelder,J.A.,*Generalized linear models*, Monographs on statistics and applied probability, Chapman and Hall, New York, 1983

Murphy,K.P. and Brockman, M.J. and Peter, K. and Lee, W., "Using Generalized Linear Models to Build Dynamic Pricing Systems", *CAS*, 2000

Nelder,J.A. and Verrall,R.J., "Credibility Theory and Generalized Linear Models", *Austin Bulletin, vol 27, No1*, 1997, pp.71~82

Nelder,J.A. and Wedder Burn, R.W.M "Generalized Linear Models", Journal of Royal Statistical Society, Ser. A, P8, 1972, 135~370

Ohlsson,E. and Johansson,B.,"Combining Credibility and GLM for Rating of Multi-Level Factors", *Casualty Actuarial Society Discussion Paper Program Casualty Actuarial Society - Arlington,* Virginia, 2004 , p319

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Renshaw,A.E., "Claim reserving by jount modeling“, *Proceeding XXVI ASTIN Colloquium*, Cannes, 1995

Renshaw, A.E. and Verall,R.J.,"A stochastic model undelying the chain-ladder technique", *Proceedings ASTIN Colloquium*, 1994

Schwarz,G.E., "Estimating the dimension of a model", *Annals of Statistics* 6 (2): 461464, 1978

Verrall,R.J. "An investigation into stochastic claims reserving models and the chain-ladder technique", *Insurance : Mathmatics and Economics* 26(1), 2000, pp.91~99

Venables,W.N. and Repley,B.D., *Modern Applied Statistics with S*, Springer, 2002, p184

Wood,S.N., *Generalized Additive Models*, Chapman & Hall/CRC, 2006

Wright,T.S.,"Stochastic method for claims reserving in general insurance" *Journal of th Institute of Actuaries* 117, 1993, pp.677~731

http://www.quantlet.com

보험연구원(KIRI) 발간물 안내

■연구보고서

2006-1	보험회사의 은행업 진출 방안 / 류근옥 2006.1
2006-2	보험시장의 퇴출 분석과 규제개선방향 / 김헌수 2006.3
2006-3	보험지주회사제도 도입 및 활용방안 / 안철경, 이상우 2006.8
2006-4	보험회사의 리스크공시체계에 관한 연구 / 류건식, 이경희 2006.12
2007-1	국제보험회계기준도입에 따른 영향 및 대응방안 / 이장희, 김동겸 2007.1
2007-2	민영건강보험료율 결정요인 분석 / 조용운, 기승도 2007.3
2007-3	퇴직연금 손·익 위험 관리전략에 관한 연구 / 성주호 2007.3
2007-4	확률적 프런티어 방법론을 이용한 손해보험사의 기술효율성 측정 / 지홍민 2007.3
2007-5	금융겸업화에 대응한 보험회사의 채널전략 / 안철경, 기승도 2008.1
2008-1	보험회사의 리스크 중심 경영전략에 관한 연구 / 최영목, 장동식, 김동겸 2008.1
2008-2	한국 보험시장과 공정거래법 / 정호열 2008.3
2008-3	확정급여형 퇴직연금의 자산운용 / 류건식, 이경희, 김동겸 2008.3
2009-1	보험설계사의 특성분석과 고능률화 방안/ 안철경, 권오경 2009.1
2009-2	자동차사고의 사회적 비용 최소화 방안 / 기승도 2009.1
2009-3	우리나라 가계부채 문제의 진단과 평가 / 유경원, 이혜은 2009.3

■ 조사보고서

2006-1	2006년도 보험소비자 설문조사 / 김세환, 조재현, 박정희 2006.3
2006-2	주요국 방카슈랑스의 운용사례 및 시사점 / 류건식, 김석영, 이상우, 박정희, 김동겸 2006.7
2007-1	보험회사 경영성과 분석모형에 관한 비교연구 / 류건식, 장이규, 이경희, 김동겸 2007.3
2007-2	보험회사 브랜드 전략의 필요성 및 시사점 / 최영목, 박정희 2007.3
2007-3	2007년 보험소비자 설문조사 / 안철경, 기승도, 오승철 2007.3
2007-4	주요국의 퇴직연금개혁 특징과 시사점 / 류건식, 이상우 2007.4
2007-5	지적재산권 리스크 관리를 위한 보험제도 활용방안 / 이기형 2007.10
2008-1	보험회사 글로벌화를 위한 해외보험시장 조사 / 양성문, 김진억, 지재원, 박정희, 김세중 2008.2
2008-2	노인장기요양보험 제도 도입에 대응한 장기간병보험 운영 방안 / 오영수 2008.3
2008-3	2008년 보험소비자 설문조사 / 안철경, 기승도, 이상우 2008.4
2008-4	주요국의 보험상품 판매권유 규제 / 이상우 2008.3
2009-1	2009년 보험소비자 설문조사 / 안철경, 이상우, 권오경 2009.3
2009-2	Solvency II의 리스크평가모형 및 측정방법 연구 / 장동식 2009.3
2009-3	이슬람 보험시장 진출방안 / 이진면, 이정환, 최이섭, 정중영, 최태영 2009.3
2009-4	미국 생명보험 정산거래의 현황과 시사점 / 김해식 2009.3
2009-6	복합금융 그룹의 리스크와 감독 / 이민환, 전선애, 최원 2009.4
2009-7	보험산업 글로벌화를 위한 정책적 지원방안 / 서대교, 오영수, 김영진 2009.4
2009-8	구조화금융 관점에서 본 금융위기 분석 및 시사점 / 임준환, 이민환, 윤건용, 최원 2009.7
2009-9	보험리스크 측정 및 평가 방법에 관한 연구 / 조용운, 김세환, 김세중 2009.7

■ 정책보고서

2006-1	2007년도 보험산업 전망과 과제 / 동향분석팀 2006.12
2006-2	의료리스크 관리의 선진화를 위한 의료배상보험에 대한 연구 / 차일권, 오승철 2006.12
2007-1	퇴직연금 수탁자리스크 감독방안 / 류건식, 이경희 2007.2
2007-2	보험상품의 불완전판매 개선방안 / 차일권, 이상우 2007.3
2007-3	퇴직연금 지급보증제도의 요율체계에 관한 연구:미국과 영국을 중심으로/ 이봉주 2007.3
2007-4	보험고객정보의 이용과 프라이버시 보호의 상충문제 해소방안 / 김성태 2007.3
2007-5	방카슈랑스가 보험산업에 미치는 영향 분석 / 안철경, 기승도, 이경희 2007.4
2007-6	2008년도 보험산업 전망과 과제 / 양성문, 김진억, 지재원, 박정희, 김세중 2007.12
2008-1	민영건강보험 운영체계 개선방안 연구 / 조용운, 김세환 2008.3
2008-2	환경오염리스크관리를 위한 보험제도 활용방안 / 이기형 2008.3
2008-3	금융상품의 정의 및 분류에 관한 연구 / 유지호, 최원 2008.3
2008-4	2009년도 보험산업 전망과 과제 / 이진면, 이태열, 신종협, 황진태, 유진아, 김세환, 이정환, 박정희, 김세중, 최이섭 2008.11
2009-1	현 금융위기 진단과 위기극복을 위한 정책제언 / 진익, 이민환, 유경원, 최영목, 최형선, 최원, 이경아, 이혜은 2009.2
2009-2	퇴직연금의 급여 지급 방식 다양화 방안 / 이경희 2009.3
2009-3	보험분쟁의 재판외적 해결 활성화 방안 / 오영수, 김경환, 이종욱 2009.3

■ 경영보고서

2009-1 기업휴지보험 활성화 방안 연구 / 이기형, 한상용 2009.3

2009-2 자산관리서비스 활성화 방안 / 진익 2009.3

■ 연구논문집

1호	보험산업의 규제와 감독제도의 미래 / Harold D. Skipper, Robert W. Klein, Martin F. Grace 1997.6
2호	세계보험시장의 변화와 대응방안 / D. Farny, 전천관, J. E. Johnson, 조해균 1998.3
3호	제1회 전국대학생 보험현상논문집 1998.11
4호	제2회 전국대학생 보험현상논문집 1999.12

■ 영문발간물

1호	Environment Changes in the Korean Insurance Industry in Recent Years : Institutional Improvement, Deregulation and Liberalization / Hokyung Kim, Sango Park, 1995.5
2호	Korean Insurance Industry 2000 / Insurance Research Center, 2001.4
3호	Korean Insurance Industry 2001 / Insurance Research Center, 2002.2
4호	Korean Insurance Industry 2002 / Insurance Research Center, 2003.2
5호	Korean Insurance Industry 2003 / Insurance Research Center, 2004.2
6호	Korean Insurance Industry 2004 / Insurance Research Center, 2005.2
7호	Korean Insurance Industry 2005 / Insurance Research Center, 2005.8
8호	Korean Insurance Industry 2006 / Insurance Research Center, 2006.10
9호	Korean Insurance Industry 2007 / Insurance Research Center, 2007.9
10호	Korean Insurance Industry 2008 / Korea Insurance Research Institute, 2008.9

■ Insurance Business Report

20호	선진 보험사 재무공시 특징 및 트렌드(유럽 및 캐나다를 중심으로) / 장이규 2006.11
21호	지급여력 평가모형 트렌드 및 국제비교 / 류건식, 장이규 2006.11
22호	선진보험그룹 글로벌화 추세와 시사점 / 안철경, 오승철 2006.12
23호	미국과 영국의 손해보험 직판시장 동향분석 및 시사점 / 안철경, 기승도 2007.7
24호	보험회사의 자본비용 추정과 활용: 손해보험회사를 중심으로 / 이경희 2007.7
25호	영국손해보험의 행위규제 적용과 영향 / 이기형, 박정희 2007.9
26호	퇴직연금 중심의 근로자 노후소득보장 과제 / 류건식, 김동겸 2008.2
27호	보험부채의 리스크마진 측정 및 적용 사례 / 이경희 2008.6
28호	일본 금융상품판매법의 주요내용과 보험산업에 대한 영향 / 이기형 2008.6
29호	보험회사의 노인장기요양 사업 진출 방안 / 오영수 2008.6
30호	교차모집제도의 활용의향 분석 / 안철경, 권오경 2008.7
31호	퇴직연금 국제회계기준의 도입영향과 대응과제 / 류건식, 김동겸 2008.7
32호	보험회사의 헤지펀드 활용방안 / 진익 2008.7
33호	연금보험의 확대와 보험회사의 대응과제 / 이경희, 서성민 2008.9

■ CEO Report

2006-1 생보사 개인연금보험 생존리스크 분석 및 시사점 / 생명보험본부 2006.1

2006-2 보험회사의 퇴직연금 운용전략 / 보험연구소 2006.1

2006-3 생보사 FY2006 손익 전망 및 분석 / 생명보험본부 2006.2

2006-4 의무보험제도의 현황과 과제 / 손해보험본부 2006.2

2006-5 자동차보험 지급준비금 분석 및 과제 / 자동차보험본부 2006.3

2006-6 보험사기 관리실태와 대응전략 / 정보통계본부 2006.3

2006-7 자동차보험 의료비 지급 적정화 방안 / 자동차보험본부 2006.3

2006-8 자동차보험시장 동향 및 전망 / 자동차보험본부 2006.4

2006-9 날씨위험에 대한 손해보험회사의 역할 강화 방안 / 손해보험본부 2006.4

2006-10 장기손해보험 상품운용전략 -손익관리를 중심으로- / 손해보험본부 2006.5

2006-11 자동차 중고부품 활성화 방안 / 자동차기술연구소 2006.5

2006-12 장기간병보험시장의 활성화를 위한 상품개발 방향 / 보험연구소 2006.6

2006-13 보험산업 소액지급결제시스템 참여방안 / 보험연구소 2006.7

2006-14 생명보험 가입형태별 위험수준 분석 / 리스크·통계관리실 2006.8

2006-15「민영의료보험법」제정(안)에 대한 검토 / 보험연구소 2006.9

2006-16 모기지보험의 시장규모 및 운영방안 / 손해보험본부 2006.9

2006-17 생명보험 상품별 가입 현황 분석 / 생명보험본부 2006.10

2006-18 자동차보험 온라인시장의 성장 및 시사점 / 자동차보험본부 2006.10

2007-1	퇴직연금제 시행 1년 평가 및 보험회사 대응과제 / 보험연구소 2007.4
2007-2	외국의 협력정비공장제도 운영현황과 전략적 시사점 / 자동차기술연구소 2007.4
2007-3	예금보험제 개선안의 문제점 및 과제 / 보험연구소 2007.6
2007-4	자본시장통합법 이후 보험산업의 진로 / 보험연구소 2007.7
2007-5	방카슈랑스 확대 시행과 관련한 주요 이슈 검토 / 보험연구소 2007.11
2007-6	자동차보험 시장변화와 전략적 시사점 / 자동차보험본부 2007.11
2008-1	자동차보험 물적담보 손해율 관리 방안 / 기승도 2008.6
2008-2	보험산업 소액지급결제시스템 참여 관련 주요 이슈 / 이태열 2008.6
2008-3	FY2008 수입보험료 전망 / 동향분석실 2008.8
2008-4	퇴직급여보장법 개정안의 영향과 보험회사 대응과제 / 류건식, 서성민 2008.12
2009-1	FY2009 보험산업 수정전망과 대응과제 / 동향분석실 2009.2
2009-2	퇴직연금 예금보험요율 적용의 타당성 검토 / 류건식, 김동겸 2009.3
2009-3	퇴직연금 사업자 관련규제의 적정성 검토 / 류건식, 이상우 2009.6

정기간행물

■ 계간

- ○ 보험동향
- ○ 보험금융연구
- ○ 보험회사 재무분석

『도서회원 가입안내』

회원 및 제공자료

<table>
<tr><td></td><td>법인회원</td><td>특별회원</td><td>개인회원</td><td>연속간행물
구독회원</td></tr>
<tr><td>연회비</td><td>₩ 300,000원</td><td>₩ 150,000원</td><td>₩ 150,000원</td><td>간행물별로 다름</td></tr>
<tr><td rowspan="2">제공자료</td><td>- 연구보고서
- 정책/경영보고서
- 조사보고서
- 기타보고서

-연속간행물
· 보험금융연구
· 보험동향(계간)</td><td>- 연구보고서
- 정책/경영보고서
- 조사보고서
- 기타보고서

-연속간행물
· 보험금융연구
· 보험동향(계간)</td><td>- 연구보고서
- 정책/경영보고서
- 조사보고서
- 기타보고서

-연속간행물
· 보험금융연구
· 보험동향(계간)</td><td rowspan="2">-보험금융연구
(년3회
₩ 30,000)

-보험통계월보
(월간
₩ 50,000)

-보험동향
(계간
₩ 20,000)</td></tr>
<tr><td>-본원 주최 각종 세미나 및 공청회 자료
-보험통계월보
-손해보험통계연보</td><td>-보험통계월보
-손해보험통계연보</td><td>-</td></tr>
</table>

※ 특별회원 가입대상 : 도서관 및 독서진흥법에 의하여 설립된 공공도서관 및 대학도서관

가입문의

보험연구원 도서회원 담당
전화 : (02)3775-9115, 9080 팩스 : (02)3775-9102

회비납입방법

- 무통장입금 : 국민은행 (400401-01-125198)
  예금주 : 보험연구원
- 지로번호 : 6360647

가입절차

보험연구원 홈페이지(www.kiri.or.kr)에 접속 후 도서회원가입신청서를 작성·등록 후 회비입금을 하시면 확인 후 1년간 회원자격이 주어집니다.

자료구입처

서울 : 보험연구원 보험자료실, 교보문고, 영풍문고, 반디앤루니스
부산 : 영광도서

저 자 약 력

기 승 도

한국외국어대학교 통계학 박사 수료
현 보험연구원 전문연구위원
(E-mail : kaebi@kiri.or.kr)

김 대 환

University of California, Davis, 경제학박사
현 보험연구원 부연구위원
(E-mail : dhkim@kiri.or.kr)

연구보고서 2009-5

일반화선형모형을 이용한 자동차보험 요율상대도 산출방법 연구

발 행 일	2009년 8월 일
발 행 인	나 동 민
발 행 처	보 험 연 구 원
	서울특별시 영등포구 여의도동 35-4
	대표전화 (02) 3775-9000

ISBN 978-89-5710-090-5 정가 10,000원