응용통계연구 (2009)
22(6), 1215–1227

언더라이팅 시스템 구축을 위한 일반화가법부분선형모형의 활용

기승도[1] · 강기훈[2]

[1]한국외국어대학교 정보통계학과, 보험연구원; [2]한국외국어대학교 정보통계학과

(2009년 11월 접수, 2009년 12월 채택)

요 약

보험회사가 보험가입자의 정확한 위험도를 측정하여, 현재 보험료 수준으로 해당 가입자를 보험에 가입하도록 허용하는 것이 보험회사에게 손해인지 여부를 판정하는 보험회사의 활동을 언더라이팅이라 한다. 언더라이팅 시스템을 구축하는 방법으로는 기존 전통적 방법과 통계모형을 활용하는 방법이 있다. 기존의 요율산출방법에 따라 위험집단의 위험도 수준을 정하고, 해당 위험집단에 속한 가입자의 위험도를 기계적으로 계산하는 전통적인 방법은 모형의 이해가 용이하고, 사용이 편리하나 통계적으로 부합된 모형이라고 할 수는 없다. 본 연구에서는 우리나라 자동차보험 분야에서 언더라이팅 기준을 구축하기 위해 통계모형을 활용하는 방법으로 일반화가법모형을 활용하는 방안을 제시하고 분석하였다. 본 연구의 결과는 현재 자동차보험 요율산출에 사용되고 있는 변수들의 유의성을 판단하는 데에도 활용될 수 있을 것이다.

주요용어: 일반화가법모형, 자동차보험, 연결함수.

1. 서론

보험 분야에서는 다양한 통계모형이 사용되고 있는데 특히 많이 활용되는 분야는 언더라이팅(underwriting)이다. 언더라이팅은 보험가입자의 정확한 위험도를 측정하여, 현재 보험료 수준으로 해당 가입자가 보험에 가입하도록 허용하더라도 보험회사에게 손해인지 여부를 판정하는 보험회사의 활동이다. 언더라이팅을 얼마나 정교하고, 정확히 하는 지에 따라 보험회사의 영업수지는 큰 영향을 받는다. 즉, 위험도가 높은 가입자가 보험에 가입하도록 허용하면 손해를 보고, 위험도가 낮은 가입자를 잘 선택하여 가입하도록 하면 이익을 볼 수 있는 것이다. 이처럼 언더라이팅이 보험회사의 손익에 큰 영향을 미치므로 보험회사는 보다 정교한 언더라이팅 시스템을 구축하기 위해서 많은 노력을 한다.

언더라이팅 시스템을 구축하는 방법으로는 전통적 방법과 통계모형을 활용하는 방법이 있다. 전통적인 방법은 정해진 요율산출 방법에 따라 위험집단의 위험도 수준을 정하고, 해당 위험집단에 속한 가입자의 위험도를 계산하는 방식이다. 전통적인 방법에서는 통계적 방법이 아닌 기계적 계산방식인 위험도 평가모형이 사용된다. 이들 전통모형은 모형의 이해가 용이하고, 사용이 편리하다는 장점이 있으나 통계적으로 부합된 모형이라고 보기는 어렵다. 특히, 오늘날처럼 다양한 통계모형이 개발되고, 통계모형을 구축할 수 있는 전산시스템이 잘 갖추어진 때에는 전통적인 언더라이팅 구축시스템이 효율적인지 의문이 든다.

이 논문은 2009년 한국외국어대학교 학술연구비 지원에 의해 이루어졌음.
[2]교신저자: (449-791) 경기도 용인시 처인구 모현면, 한국외국어대학교 정보통계학과, 부교수.
E-mail: khkang@hufs.ac.kr

이러한 전통적 방법의 문제점을 해결하기 위해서 사용되는 언더라이팅 구축방법이 통계모형을 활용하는 것이다. 언더라이팅에서 통계모형을 활용할 때 주로 사용되는 방법은 일반화선형모형(Generalized Linear Model)이다. 실제 보험통계의 사고자료 분포들이 정규분포를 따르지 않고 다양한 분포를 따른다는 점을 감안할 때, 일반화선형모형을 활용하여 언더라이팅 시스템을 구축하는 것은 보다 정교한 방법이라 할 수 있다. 예를 들어, 자동차보험에서 사고건수 자료에 적합한 분포로는 포아송분포, 음이항분포, 역감마분포 등등 다양한데, 이들 분포는 모두 지수족(exponential family)에 속하므로, 이들 분포를 가정한 일반화선형모형으로 정확한 분석이 가능해진다. 이러한 이유로 언더라이팅 시스템 구축에서 일반화선형모형이 널리 활용되고 있는 것이다. 일반화선형모형은 반응변수의 평균과 설명변수들의 선형결합을 적절한 연결함수(link function)로 모형화하게 된다. 이 때, 설명변수의 선형결합으로 반응변수와의 관계를 설명하기 어려운 경우에는 문제가 된다. 이러한 문제를 해결하기 위해서 최근에는 일반화선형모형을 확정한 일반화가법모형(Generalized Additive Model)이 개발되었다. 일반화가법모형에서 반응변수에 대한 가정은 일반화선형모형과 같이 지수족을 가정한다.

일반화가법모형은 Hastie와 Tibshirani (1986)가 기존의 일반화선형모형과 가법모형(additive)을 결합하여 처음으로 개발 제시하였다. 일반화선형모형은 Nelder와 Wedderburn (1972)가 처음 제시하였으며, 가법모형은 Stone (1985)이 차원의 저주(curse of dimensionality) 문제를 해결하기 위하여 제시한 모형이다. Hastie와 Tibshirani (1986) 이후 일반화가법모형의 연구는 활용도를 개선하는 방향으로 이루어졌다. 기존의 일반화가법모형에서는 독립변수의 비모수적 함수 형태들로만 종속변수의 기댓값을 설명하고자 했다. 그러나, 독립변수가 함수로만 되어 있을 경우 종속변수와 독립변수의 관계를 특정한 모수(parameter)로 표현할 수 없어서, 일반화가법모형으로 분석한 결과를 해석하기 용이하지 않다는 단점이 있다. 이런 단점을 해결하기 위해 개발된 것이 일반화가법부분선형모형(Generalized Additive Partial Linear Models)이다. 이 모델은 독립변수의 비모수적 함수와 모수를 이용한 선형함수가 모두 포함된 모형이다. Speckman (1988)은 이 모델에 사용되는 모수 β가 $\sqrt{n}$-비율로 추정될 수 있다는 것을 보였고, Fan 등 (1998)은 이 모델에서 해를 구하는 방법을 제시하였다. Lin과 Zhang (1999)는 고정효과(fixed effect)와 랜덤효과(random effect)를 설명할 수 있는 모형인 일반화가법혼합모형(Generalized Additive Mixed Models)을 제안하였다. Wood (2000, 2003, 2004, 2006)는 일반화가법모형에 대한 연구를 바탕으로 2006년에 일반화가법모형의 전반적인 내용을 설명하는 책자를 발간하였다.

일반화가법모형의 이론을 개발하는 연구는 활발히 진행되었으나, 이를 활용하는 분야에 대한 연구는 제한적이었다고 할 수 있다. 특히 일반화가법모형이 보험분야에 적용된 경우에 대한 연구는 외국 및 우리나라에서 거의 없는 편이다. 최근 Jong과 Heller (2008)가 자동차보험 요율산출 분야에 활용하여 통계모형을 구축할 때, 독립변수와 종속변수의 관계를 규명하는데 일반화가법모형을 활용할 수 있다고 설명한 것이 그나마 관련이 있다고 할 수 있다. 따라서 일반화가법모형으로 우리나라 자동차보험 자료를 분석하고, 그 결과를 활용하여 자동차보험 언더라이팅 시스템을 구축하는 방법을 제시하고자 하는 본 연구는 일반화가법모형 활용성 증대 및 보험분야의 언더라이팅 정교화 측면에서 큰 의미가 있다고 판단된다.

본 논문은 총 5절로 구성되어 있다. 2절에서는 분석대상 자료 및 변수 선정에 관해 서술하였으며, 3절은 본 논문에서 고려하는 모델에 대해 설명하였다. 4장에서는 자동차보험 자료를 이용한 실증분석 결과를 제시하였으며, 5장은 분석한 결과를 바탕으로 결론을 도출하고 토의하였다.

2. 실증분석을 위한 자료 및 변수 선정

2.1. 통계자료

분석을 위해 이용한 통계자료는 보험개발원에 집적된 회계년도 2005년 (FY2005)의 총 자료 중에서 임

표 **2.1.** 개인용 및 플러스 개인용 자동차보험 통계자료

구분	변수	세부분류
독립변수	성별	남, 여
	차종	소형, 중형, 대형, 다인승
	연령	기명피보험자 연령 (범주형)
	가입경력	자동차보험 가입경력기간 1~6년
	법규위반	현행기준 참조 순보험료 요율서 기준
	할인할증률	40%~200% (10% 단위)
종속변수	사고빈도	담보별 사고건수
	사고심도	1사고당 지급된 보험금

의표본추출한 약 50만개를 대상으로 하였다. 보험계약이 체결되고, 체결된 계약에서 발생한 사고기록 자료인 증권년도기준(policy year basis) 통계를 이용하였다. 증권년도기준 통계는 계약자료와 사고자료가 정확하게 일치하므로, 사고원인 특성을 정확하게 분석할 수 있는 통계추출 방식이다.

본 연구의 목적은 일반화가법모형을 자동차보험분야에 적용할 수 있는 가능성을 확인해보는 것이다. 따라서 본 연구에서는 자동차보험의 통계특성을 잘 반영할 수 있는 보험종목과 담보 및 사고 자료로 한정하여 연구를 진행하였다. 통계추출 대상 보험종목은 개인용 및 플러스개인용으로 하였다. 개인용 및 플러스 개인용은 전체자동차보험에서 차지하는 비율이 평균 유효대수 기준으로 70.4%로 매우 중요한 보험종목이고, 개인용 자동차보험은 운전자가 개별적으로 가입하는 보험종목이므로 통계적으로 유의한 운전자의 특성을 파악할 수 있는 대상이다. 반면에 업무용자동차보험 및 영업용자동차보험은 가입대상이 대부분 법인이므로, 운전자의 개별특성을 파악하기 어려워 분석대상에서 배제하였다. 그리고 본 연구에서는 분석대상 담보를 대인배상 I 과 대물배상 등 의무가입대상 담보로 하였다. 자동차손해배상 보장법에는 보험가입자가 의무적으로 이 담보를 가입하여야 하고, 특히 대인배상 I 의 경우는 보험회사가 동 담보가입을 거절할 수 없도록 되어 있다. 대인배상 I 과 대물배상은 인적사고와 물적사고를 대표하는 담보로 타인 또는 남의 물건을 파손할 때 지급되는 배상책임담보이다. 이 담보의 사고기록으로 실제 우리나라 자동차사고의 대표적 특성을 찾아낼 수 있다고 보기에 본 연구에서는 대인배상 I 과 대물배상 담보를 분석대상으로 한 것이다.

2.2. 변수의 선정

이상에서 언급된 본 연구의 대상 자료에 대한 요약이 표 2.1에 주어져 있다. 독립변수는 현재 자동차보험에서 사용되고 있거나 분석에 필요한 것으로 '성별', '차종', '연령', '가입경력', '법규위반 유형', '할인할증률' 등을 고려하였다. 이들 변수 자료는 전산에 등록되어 있는 원 자료 수준으로 추출된 것이다. 이 자료 중의 일부는 범주형으로 변환되고 일부 변수는 함수형태로 사용된다. 자동차사고에서 보상은 사고빈도와 사고심도와 밀접하기 때문에 종속변수는 이 둘을 고려하였다. 사고빈도는 사고건수에 해당되는 것이고, 사고심도는 1사고 당 지급된 보험금자료에 해당된다. 사고빈도와 사고심도가 모두 고려되어야 최종 손해액을 추정할 수 있으며, 추정된 최종손해액으로 추정손해율을 추정하여 언더라이팅 기준을 만들 수 있다. 각 변수별로 자세히 설명해 보겠다.

1) **성별 및 차종:** 성별 변수는 남성과 여성으로 구분된 범주형 자료이고, 차종구분은 보험개발원(2008)에서 정의된 개인용 및 플러스 개인용자동차보험의 차종구분을 따라 6개로 나누었다. 6개의 차종 구분의 기준은 배기량(cc)과 탑승자 인원 등으로 표 2.2에 주어진 바와 같다.
2) **연령:** 기명피보험자 연령은 18세부터 99세까지 이다. 보험회사들은 각 연령별로 요율을 산출·적용

표 **2.2.** 차종 구분 기준

차종	구분 기준
소형A	법정승차정원 6인 이하의 일반형승용자동차, 승용겸화물 자동차, 짚형자동차 및 구급용자동차로서 배기량이 1,000cc이하인 자동차
소형B	법정승차정원 6인 이하의 일반형승용자동차, 승용겸화물 자동차, 짚형자동차 및 구급용자동차로서 배기량이 1,000cc초과 1,600cc 이하인 자동차
중형	법정승차정원 6인 이하의 일반형승용자동차, 승용겸화물 자동차, 짚형자동차 및 구급용자동차로서 배기량이 1,600cc초과 2,000cc 이하인 자동차
대형	법정승차정원 6인 이하의 일반형승용자동차, 승용겸화물 자동차, 짚형자동차 및 구급용자동차로서 배기량이 2,000cc초과
다인승 I	법정승차정원 7인 이상 10인 이하의 일반형자동차(승합포함), 구급용자동차 및 기타자동차 중 전방조정 자동차
다인승 II	법정승차정원 7인 이상 10인 이하의 일반형자동차(승합포함), 구급용자동차 및 기타자동차 중 전방조정 자동차가 아닌 자동차

하지 않고, 동일한 위험집단별로 범주화하여 요율을 산출·적용하고 있다. 본 연구에서는 연령을 1년 단위로 운전가능 연령인 18세 이상과 80세까지로 하였다. 80세 이상 연령은 실질적으로 운전을 하기 어렵기 때문에 80세로 한정하였다.

3) **가입경력:** 가입경력 구분은 회사별로 차이가 있으나, 일반적으로 '최초가입자 또는 1년 미만', '1년 이상 2년 미만', '2년 이상 3년 미만' 및 '3년 이상'으로 구분되어 있다. 가입경력을 나누는 기준은 현재 자율화가 되어 있으므로 회사별로 차이가 있다. 따라서 일부 회사들은 3년 이상 가입경력을 더 세분화하고 있다. 본 연구에서는 가입경력을 1년 단위로 6등분하여 분석하였다.

4) **법규위반:** 교통법규 위반경력 요율제도는 크게 할증그룹, 기본그룹 그리고 할인그룹으로 나뉜다. 다시 할증그룹은 할증 1그룹 및 할증 2그룹으로 나뉜다. 이들 교통법규 위반경력 요율 적용구분 기준은 교통법규위반 항목 중에서 피해자에게 심각한 영향을 주거나, 위험도 수준에 따라 만들어졌다. 할증그룹에 포함되는 교통법규위반 항목은 무면허운전, 주취운전과 사고발생시 조치, 신호지시 준수의무, 중앙선 우측통행, 속도제한 등 중대교통법규를 위반하는 경우에 해당된다. 그리고 기본그룹에는 할증그룹에 포함된 교통법규 위반자 이외의 일반교통법규 위반자들이 포함된다. 할인그룹은 할증그룹과 기본그룹이외의 모든 경우가 포함된다.

5) **할인할증률:** 할인할증률은 과거 사고유무, 사고내용에 따른 할증점수에 따라 평가대상기간을 약 3년으로 하여 산출된다. 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 FY2005년도 자료에는 18개의 할인할증 그룹으로 나누어져 있다. 할인할증그룹이 18개에서 24개로 확대된 것은 할인할증그룹별로 위험도에 부합된 요율이 적용될 수 있도록 2007년 9월부터 할인할증요율 적용기준이 변경되었기 때문이다. 본 연구에서 사용된 통계가 FY2005년 자료이므로, 본 연구에서도 우량할인·불량할증 구분기준을 FY2005년 기준으로 하여 통계모형을 적용하였다.

3. 통계모형

초기에 개발되고 사용되었던 가법모형(Addtive Model)은 종속변수를 독립변수의 함수들의 선형결합으로 설명하고자 하는 식 (3.1)과 같은 것이다. 초기의 가법모형에서는 종속변수에 대한 가정이 정규분포

인 경우로 한정되어 있고, 독립변수도 비모수함수만으로 구성되어 있어서 사용분야가 한정될 수 밖에 없었다.

$$Y = c + \sum_{\alpha=1}^{d} g_\alpha(x_\alpha) + \varepsilon. \tag{3.1}$$

이에 반해 일반화가법모형(Generalized Additive Model)은 가법모형을 확장한 것으로 다음 식 (3.2)와 같이 지수족 분포를 따르는 종속변수의 기댓값이 각 독립변수들의 비모수적 함수의 결합으로 연결함수(link function) G를 사용하여 설명할 수 있다는 가정에 따라 만들어진 것이다. 최근에는 독립변수의 비모수적 요소와 모수족 요소가 모두 포함된 일반화가법모형이 개발되어 여러분야에서 활용될 수 있는 모델로 지속적으로 개선되고 있다.

$$E(Y|X) = G\left(c + \sum_{\alpha=1}^{d} g_\alpha(x_\alpha)\right). \tag{3.2}$$

보험 자료는 독립변수가 연속형인 자료와 범주형 자료로 구성되어 있다. 그러므로 독립변수가 연속형인 경우만 분석이 가능한 기존의 일반화가법모형으로는 보험통계를 분석할 수 없다. 이에 본 연구에서는 독립변수가 연속형 자료인 경우 및 범주형 자료인 경우에도 적용할 수 있는 일반화가법모형을 분석모형으로 정하였다. 이러한 일반화가법모형을 일반화가법부분선형모형(Generalized Additive Partial Linear Model; GAPLM)이라고 한다. GAPLM을 수식으로 표현하면 식 (3.3)과 같다.

$$E(Y|\mathbf{U},\mathbf{T}) = G\left(c + \mathbf{U}^t\beta + \sum_{\alpha}^{q} g_\alpha(T_\alpha)\right), \tag{3.3}$$

여기서, Y는 종속변수로 본 논문에서는 사고빈도 또는 사고심도에 해당된다. $\mathbf{U}$는 보험제도 변수를 나타내는 범주형 변수이며, β는 이에 대한 계수 벡터이다.

이러한 GAPLM으로 통계분석을 시행할 때에는 종속변수인 사고빈도 및 사고심도 자료를 잘 설명할 수 있는 분포를 찾는 것과 연결함수를 선택하는 것을 고려해야 한다. 따라서 본 연구에서는 종속변수에 부합한 분포를 찾는 방법을 우선적으로 시행하였고, 사용되는 모형이 일반화가법모형이므로 종속변수의 분포 범위를 지수족으로 한정하였다. 연결함수 선택문제에서는 모형적합이 뛰어난 연결함수가 사용되도록 하였다. GAPLM은 일반화선형모형처럼 자동차보험 요율산출분야, 자동차보험 언더라이팅 모델 구축분야, 지급준비금 추정분야 등에서 활용될 수 있을 것으로 기대된다.

식 (3.3)의 모형 적합을 위해서는 모수적 요소인 β와 비모수적 함수인 g_α를 추정해야 한다. β의 추정을 위해 사용되는 추정방법은 backfitting과 local scoring algorithm이며, 자세한 내용은 Härdle 등(2004a)의 9장 2절을 참조하면 된다. 또한, g_α를 추정하는데에는 다음의 국소가능도함수를 변형한 프로파일(profile) 가능도함수를 이용한다.

$$l_{h,\mathbf{H}}(Y, \mu_{m(\mathbf{T})}, \phi) = \sum_{i=1}^{n} K_h(t_\alpha - \mathbf{T}_\alpha)\mathcal{K}_{\mathbf{H}}(\mathbf{t}_{\underline{\alpha}} - \mathbf{T}_{\underline{\alpha}}) l\left(Y_i, G\left\{\mathbf{U}_i^t\beta + m(t_\alpha, \mathbf{t}_{\underline{\alpha}})\right\}, \phi\right).$$

이 식에서 해를 찾기 위해서는 Fan 등 (1998)에 의해 제안된 marginal integration method를 적용하고, 그 최적 추정치는 다음 식 (3.4)와 같다. 경계지역 또는 자료가 드문 부분에서는 계산상의 문제를 피하기 위해서 평균계산에 적절한 가중치가 적용되어야 한다.

$$\hat{g}_\alpha(t_\alpha) = \frac{1}{n}\sum_{l=1}^{n} \hat{m}(t_\alpha, \mathbf{T}_{l\underline{\alpha}}) - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n}\sum_{l=1}^{n} \hat{m}(T_{i\alpha}, \mathbf{T}_{l\underline{\alpha}}). \tag{3.4}$$

표 **4.1.** 범주형 자료 요약

독립변수					
차종 (qcar)		성별 (qgender)		법규위반 (qbvio)	
소형A	36,562	남자	386,677	중대교통법규 I	60
소형B	152,179	여자	105,516	중대교통법규 II	22,984
중형	152,433			중대교통법규 III	1,616
대형	48,259			일반교통법규	142,437
다인승 I	2,097			무위반	325,096
다인승 II	100,663				

끝으로, 상수항 c는 다음 식 (3.5)로 추정된다.

$$\hat{c} = \frac{1}{n}\sum_{i=1}^{n} \hat{m}(\mathbf{T}_i)\frac{1}{n}\sum_{i=1}^{n} \hat{m}(T_{i\alpha}, \mathbf{T}_{i\underline{\alpha}}). \tag{3.5}$$

이상에서 설명한 방법에 따라 GAPLM의 추정치가 만들어지고, 이 추정치로 본 연구의 분석이 이루어진다. 알고리즘에 대한 부가적인 세부 설명은 Hastie와 Tibshirani (1990)와 Härdle 등 (2004b)도 참고하면 도움이 될 것이다. 본 논문에서는 공개 통계언어인 R을 이용해 프로그램을 작성하고 수행하였다.

4. 자료분석 및 언더라이팅 모형구축 방법

본 논문에서 일반화가법모형을 이용한 분석에 사용된 자료의 특성을 요약해 보고자 한다. 먼저 종속변수를 보면 사고빈도에 대해서는 최대 사고건수가 5건이고, 최소가 0건이다. 사고심도는 대인배상의 경우 최저값이 70원이고 최대값이 1억원이며, 대물배상의 경우는 최소 7,700원에서 최대 6천1백만원 수준이다.

연속형 변수로 취급한 연령, 할인할증 및 가입경력에 관한 자료의 특성을 보면 다음과 같다. 연령은 최소가 18세이며, 최대가 80세이다. 현재 자동차를 운전할 수 있는 법정 연령이 18세 이상이므로, 연령변수의 연령한도는 분석에 적합한 것으로 판단된다. 할인할증률은 현재 자동차보험 할인할증제도의 최소한도가 40이고 최고한도가 200이다. 분석에 사용된 통계자료의 할인할증률의 범위도 이들 한도 내에 있으므로, 분석에 문제가 없다. 가입경력요율은 최소가 1, 최대가 6으로 되어 있으며, 평균 가입경력이 5.1년으로 나타났다. 가입경력은 6년 이상이 있을 수도 있으나, 현재 자동차보험의 가입경력요율 제도는 가입경력 4년 이상을 기본으로 하여 이보다 가입경력자이 적은 사람에게 보험료를 할증하는 구조로 되어 있다. 이러한 자동차보험의 특성을 반영하여 통계자료를 최대 6년 이상으로 수정하였다. 현재 자동차보험에서는 가입경력 자료를 약 8년 수준까지 관리하고 있으므로, 6년이상으로 가입경력을 한정하여 통계를 분석하여도 충분히 가입경력으로 사고자료를 설명할 수 있을 것으로 생각된다.

범주형 변수의 특성을 보면 표 4.1과 같다. 차종의 경우는 소형B, 중형, 다인승II의 비중이 가장 큰 것으로 나타났다. 다인승 자동차는 운전대의 위치에 따라 전방에 있는 경우는 다인승 I , 후방에 있는 경우는 다인승II로 분류된다. 여기서 다인승II는 최근에 많이 판매되고 있는 SUV 자동차가 포함되며, 다인승 I 에는 과거 봉고자동차와 같은 것이다. 현재 우리나라에서 판매되는 자동차들을 보면, 소형B, 중형 자동차 및 다인승II 자동차가 가장 많이 판매되고 있는 것으로 알려져 있다. 이처럼 판매되는 자동차의 분포의 특성과 본 연구통계의 특성을 비교하여보면 두 분포가 거의 유사한 것으로 보인다. 따라서 본 연구에 사용된 차종 변수는 충분히 자동차보험 사고 특성을 설명할 수 있을 것으로 생각된다. 성별의 경우는 남자가 여자보다 약 3배정도 많은 것으로 나타났다. 교통법규 위반은 중대교통법규위반 I 은 '무면허운전', '뺑소니' 등의 중대법규위반을 한 경우이며, 중대교통법규위반II는 '음주운전'에 해당된다. 중대

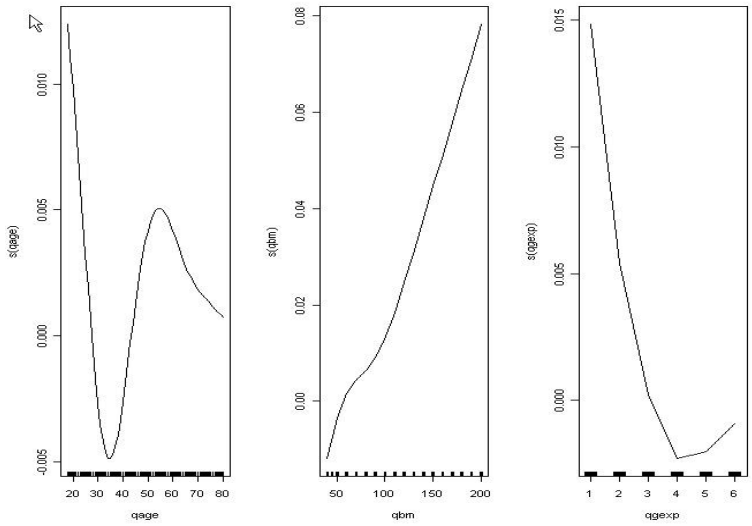

그림 4.1. 대인배상 I 사고빈도의 독립변수 함수형태

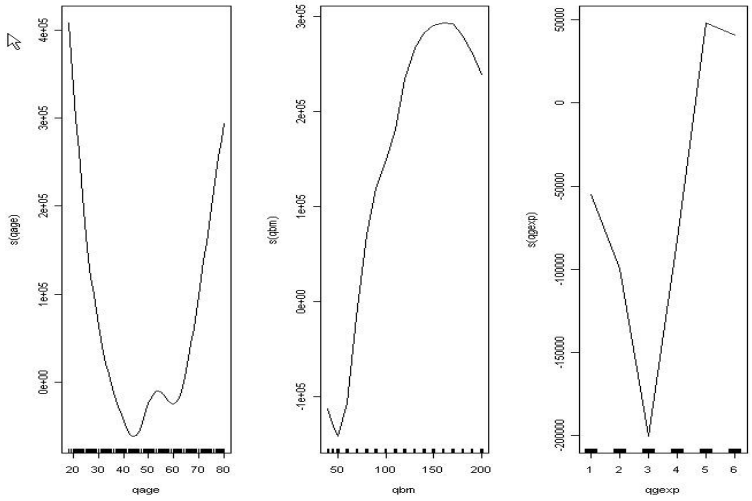

그림 4.2. 대인배상 I 사고심도의 독립변수 함수형태

교통법규위반Ⅲ은 '중앙선침범' 등의 교통법규위반이다. 일반교통법규위반자는 앞서 중대교통법규이외의 일반적인 경미한 법규위반인 경우이다.

4.1. 대인배상 I 분석 결과

종속변수(사건빈도 또는 사고심도)의 가정은 각각 음이항분포와 로그정규분포로 하였다. 이것은 대인배상Ⅰ 종속변수를 실제 분포로 적합시켜 본 결과 이러한 분포에 적합되었기 때문이다. 적합도를 가장 높일 수 있도록 일반화가법모형에 적용하는 변수의 유형을 정하였다. 이를 위해 우선 사고빈도 및 사고심도와 독립변수의 관계를 그래프로 그림 4.1과 4.2에 그려보았다. 할인할증(그림의 가운데)은 다소 선형

표 **4.2.** 대인배상 I 추정값 및 환산값

	사고빈도		사고심도	
계수	추정값	환산값	추정값	환산값
(Intercept)	−3.8798*	0.0207	14.0996*	1,328,551
qcar21	0.1058*	1.1116	0.0027	0.9973
qcar22	0.1938*	1.2138	0.0495*	1.0508
qcar23	0.1155*	1.1225	0.0995*	1.1046
qcar74	−0.1861	0.8302	0.1464	1.1577
qcar75	0.3188*	1.3755	0.1156*	1.1226
cs(qage)	0.0043*	1.0043	0.0010	1.0010
qgender2	0.0757*	1.0786	−0.0471*	0.9540
cs(qbm)	0.0079*	1.0079	0.0013*	1.0013
cs(qgexp)	−0.0449*	0.9561	−0.0052*	0.9948
qbvio1	0.1895	1.2086	−0.1375	0.8715
qbvio2	0.2340	1.2637	−0.0014	0.9986
qbvio3	0.2341	1.2638	−0.1203	0.8866
qbvio4	0.1039	1.1095	−0.0994	0.9054

∗: 유의수준 0.05에서 유의함

에 가까운 모습이다. 연령(가장 왼쪽 그림)과 가입경력(가장 오른쪽 그림)은 선형보다는 비선형 함수의 모습이다. 이에 따라 일반화가법부분선형모형의 적합 과정에서 처음에 범주형 변수인 성별, 차종, 법규위반은 모수적 요소로 하고 연령, 할인할증과 가입경력은 비모수적 요소로 모형을 설정하고 분석하였다. 연결함수로는 사고빈도인 음이항분포의 경우에는 로그함수(log)를 사용하였고, 사고심도의 분포인 로그정규분포의 경우에는 항등함수(idendity)를 사용하였다.

이상의 결과에 따라 최종적 일반화가법부분선형모형으로 분석한 결과가 표 4.2에 주어져 있다. 계수 추정값을 보면, 사고빈도에서는 차종, 연령, 성별, 할인할증, 가입경력은 통계적으로 유의한 것으로 분석되었다. 분석결과 대부분의 독립변수는 사고건수와 양(+)의 상관관계를 가지고 종속변수를 설명하는 것으로 나타났는데, 차종 74(qcar74)와 가입경력(cs(qgexp))은 음의 상관관계를 가지는 것으로 나타났다. 차종 74는 기준(base)이 되는 차종20(qcar20, 소형A)대비 사고발생률이 떨어진다는 것을 의미하고, 가입경력은 가입경력이 오래될수록 사고발생률이 감소한다는 것을 의미한다. 이러한 분석결과는 논리적으로 합당한 것으로 판단된다.

종속변수가 사고심도인 경우에는 차종, 성별, 할인할증 및 가입경력이 설명변수로서 통계적으로 유의한 것으로 분석되었다. 사고빈도의 경우와 차이점은 연령 변수가 통계적으로 유의하지 않다는 점이다. 각 범주에 속한 변수들의 위험도 수준을 알아보기 위해서 추정된 독립변수 계수값을 환산한 결과도 표 4.2에 제시하였다. 결과에서 사고빈도의 경우를 보면, 차종의 경우는 소형A(qcar20)에 비하여 소형B(qcar21)는 사고위험도가 11.2%, 중형(qcar22)은 21.4%, 대형(qcar23)은 12.3%, 다인승 Ⅱ(qcar75)은 37.6% 더 높은 것으로 분석되었다. 이러한 결과는 대형자동차가 자동차가 크므로 사고위험도가 더 클 것으로 예상할 수 있으나, 대형자동차의 경우는 전문 기사와 같이 운전경험이 많은 사람들이 운전하는 경우가 있을 수 있으므로 사고발생률이 중형 등에 비하여 낮게 나온 것이 올바른 것으로 판단된다. 성별은 남자보다는 여자(qgender2)가 사고발생률이 7.9% 더 높은 것으로 나타났다. 교통법규위반 유형의 경우는 중대교통법규위반인 경우 더 사고발생위험도가 더 높은 것으로 분석되었다.

사고심도의 환산결과를 보면 다음과 같다. 차종의 경우 소형A 대비 기타 차종의 위험도가 증가하는 모습이고, 법규위반은 중대교통법규 I 대비 다른 법규위반의 위험도가 낮은 것으로 분석되었다. 이러한 분석결과는 논리적으로 합당한 결과인 것으로 보인다. 분석자료가 대인배상 I 자료이므로 자동차

가 클수록 상대편 자동차에 큰 타격을 주므로 차종이 대형으로 갈수록 사고심도가 커지는 것이다. 교통법규위반의 경우도 중대교통법규위반 유형이 더 심각할수록 사고심도가 커질 개연성이 큰데, 분석결과는 이러한 모습을 보여주고 있다. 그러면 실제 분석결과를 해석할 수 있도록 계수를 환산한 값을 살펴보도록 하자. 차종의 경우 소형A(qcar20)보다 중형(qcar22)이 5.1%, 대형(qcar23)이 10.4%, 다인승 Ⅱ(qcar75)가 12.3% 더 1사고당 손해액이 더 많이 지급된 것으로 분석되었다. 성별로는 여성(qgender2)이 남성보다 약 5% 더 위험도가 낮은 것으로 분석되었다. 앞서 사고빈도 분석에서는 여성의 위험도가 더 높은 것으로 나타났다. 이러한 결과를 볼 때 여성의 경우 소액사고는 자주 내지만, 보험금 지급이 큰 대형사고는 적게 나는 것으로 해석될 수 있는 것이다. 교통법규위반 분석결과를 보면, 가장 심각한 중대교통법규위반인 무면허 및 뺑소니 유형보다 기타 중대교통법규위반, 일반교통법규위반 및 무위반자의 위험도가 약 10%(p)수준 낮은 것으로 분석되었다. 이러한 결과는 논리적으로 합당한 것으로 판단된다.

4.2. 대물배상 분석 결과

종속변수의 분포 선택방법은 대인배상Ⅰ과 동일한 절차를 거쳤다. 즉, 대물배상 종속변수(사고자료)를 가장 잘 설명할 수 있는 분포를 선택하여 분포의 적합성 테스트, 일반화가법모형의 적합성 수준 비교의 과정을 거쳤다. 이중 가장 모형 적합력이 큰 분포를 종속변수의 최종모형 분포로 하였다. 사고빈도의 경우에는 포아송분포와 음이항분포가 적합되었는데 이 중에 음이항분포를 적합시킨 결과가 더 좋았기 때문에 음이항분포를 가정하였다. 또한 사고심도에 대해서도 비슷한 방법으로 적합해본 결과 로그정규분포가 가장 적합한 것으로 파악되었다.

대인배상Ⅰ에서와 같이 대물배상에서도 일반화가법부분선형모형에 적용하는 변수의 유형도 모형의 적합도가 가장 높일 수 있는 경우로 정하였다. 이를 위해 우선 사고빈도 및 사고심도와 독립변수의 관계를 그래프로 그려본 결과는 대인배상Ⅰ의 경우와 유사하였고 여기서는 생략하였다. 적합성이 좋은 모형을 위해 범주형 변수인 차종, 성별, 법규위반 유형은 모수적 변수로 하고 연령, 할인할증 및 가입경력은 비모수적 요소로 분석하였다. 연결함수도 대인배상Ⅰ에서와 같이 선행연구의 결과, 분석결과의 설명 용이성을 감안하여 사고빈도는 로그함수로 하고, 사고심도는 항등함수로 하였다.

이상의 방법으로 대물배상 자료를 가장 잘 설명할 수 있는 모형을 설정하고, 일반화가법부분선형모형으로 자료를 분석한 결과가 표 4.3에 제시되어 있다. 대물배상에 대해서도 대인배상에서와 같이 거의 모든 독립변수가 통계적으로 유의하게 설명하는 것으로 분석되었다. 다인승Ⅱ와 가입경력이 대물배상 사고빈도과 음의 상관관계를 가지고 설명력이 있다는 점도 대인배상Ⅰ의 경우와 동일하다. 각 범주에 속한 변수들의 위험도 수준을 알아보기 위해서 최종 일반화 가법모형의 계수를 환산하여 보았다. 사고빈도의 경우를 보면, 차종에서는 소형A(qcar20)에 비하여 소형B(qcar21)는 사고위험도가 9.6%, 중형(qcar22)은 24.8%, 대형(qcar23)은 16.8%, 다인승Ⅱ(qcar75)는 43.7%더 높은 것으로 분석되었다. 이러한 결과는 앞서 대인배상Ⅰ의 경우와 유사한 결과이다. 다만, 대인배상Ⅰ에서 보다 사고빈도가 대형차 및 다인승Ⅱ가 더 높은 것으로 분석된 것은 차이점이다. 성별은 남자보다는 여자(qgender2)가 사고발생률이 11.1%더 높은 것으로 나타났다. 교통법규위반 유형의 경우는 중대교통법규위반인 경우 더 사고발생위험도가 높은 것으로 분석되었다.

사고심도의 경우에도 사고빈도와 유사한 모습을 보여주고 있다. 즉 차종이 클 수로 사고심도의 위험도가 큰 것으로 나타났다. 소형A(qcar20)대비 중형(qcar22), 대형(qcar23) 및 다인승Ⅱ(qcar75)의 위험도는 각각 15.1%, 28.0%, 24.8% 높은 것으로 분석되었다. 성별은 여성(qgender2)이 남성보다 대물배상 사고심도의 위험도가 약 6% 낮은 것으로 분석되었다. 교통법규위반의 경우는 가장 심각한 중대교통법규위반 보다 위험도가 평균 30%씩 낮은 것으로 분석되었다. 이 경우는 가장 심각한 중대교통법규위

표 **4.3.** 대물배상 추정값 및 환산값

	사고빈도		사고심도	
계수	추정값	환산값	추정값	환산값
(Intercept)	−2.9610*	0.0518	13.6615*	857,240
qcar21	0.0912*	1.0955	0.0862*	1.0900
qcar22	0.2212*	1.2475	0.1409*	1.1513
qcar23	0.1551*	1.1678	0.2471*	1.2803
qcar74	−0.0097	0.9904	0.2127*	1.2370
qcar75	0.3627*	1.4372	0.2216*	1.2481
cs(qage)	0.0050*	1.0050	0.0023*	1.0023
qgender2	0.1112*	1.1176	−0.0393*	0.9615
cs(qbm)	0.0050*	1.0050	0.0025*	1.0025
cs(qgexp)	−0.0440*	0.9565	−0.0182*	0.9820
qbvio1	0.1365	1.1462	−0.3536	0.7021
qbvio2	0.2341	1.2638	−0.4197	0.6573
qbvio3	0.2020	1.2238	−0.4109	0.6631
qbvio4	0.0613	1.0632	−0.3839	0.6812

∗: 유의수준 0.05에서 유의함

반인 무면허 및 뺑소니의 경우 자료의 양이 매우 적어, 상대적 위험도를 신뢰하기는 좀 어려운 것으로 판단된다.

4.3. 언더라이팅 시스템 구축

언더라이팅 담당자가 언더라이팅 의사결정을 하는데 도움이 되는 통계모형을 구축하는 것은 보험회사 입장에서도 매우 필요한 것이다. 통계모형으로 언더라이팅 시스템을 구축하는 절차는 첫째로 과거 또는 현재의 자료를 가장 잘 설명할 수 있는 통계모형을 찾는 것이다. 둘째는 앞서 찾은 통계모형으로 예측시스템을 구축하여 최적 통계모형으로 향후 가입하고자 하는 계약자의 위험도를 분석하는 시스템을 구축하는 것이다.

본 연구에서는 언더라이팅 기준을 구축하기 위해서 사고빈도와 사고심도로 나누어 분석한 앞의 통계분석 결과를 사용하였다. 소비자의 가입조건은 앞서 통계분석에 사용된 독립변수으로 지정된다. 독립변수와 종속변수의 관계를 규명하는 모형이 만들어지면, 사고빈도 모형과 사고심도 모형으로 예측된 값(사고빈도의 경우 사고발생률, 사고심도의 경우 1사고당 손해액)을 곱하여 최종 위험보험료를 만들 수 있다. 즉, 어떤 소비자가 자동차보험에 가입하고자 한다면, 앞서 분석한 사고빈도 모형과 사고심도 모형에 가입자의 조건을 넣어 해당 가입자의 위험보험료를 계산한다. 그리고 계산된 위험보험료를 보험회사가 설정하고 있는 위험보험료 기준과 비고하여 해당 가입자에게 보험가입의 허용여부를 결정한다. 기준보다 낮은 위험보험료가 추정되면 가입을 허용하고, 높은 위험보험료가 추정되면 거절하는 것이다.

예를 들면 표 4.4에 차종, 연령, 성별, 할인할증, 가입경력 및 법규위반 유형 등 가입자가 자동차보험에 가입하려고 할 때 조건을 예시하였다. 법규위반에서 0은 무면허 및 뺑소니에 해당하고, 3은 일반법규위반을 한 경우, 4는 법규위반을 하지 않은 경우이다. 이러한 조건을 가진 사람들이 어떤 보험회사에 자동차보험 가입신청을 한다고 가정할 때, 보험회사는 GAPLM으로 분석한 위험도 평가모형에 가입자의 조건을 입력하여 가입자의 향후 사고발생률(A)과 1사고당 손해액(B)을 추정한다. 추정된 사고발생률과 1사고당 손해액을 곱한 위험보험료(C)를 평균위험보험료인 판단기준(D)과 비교하여 인수여부를 판단한다. 이러한 기준에 따라 본 예시의 가입자들 중에서 언더라이팅 판단 의사결정을 한 결과가 표 4.5에 제시되어 있고, 소비자 6과 소비자 16이 인수거절 대상이 된다.

표 **4.4.** 언더라이팅 판단을 위한 예시

	가입조건					
소비자	차종	연령	성별	할인할증	가입경력	법규위반
1	대형	38	여자	40	6년이상	4
2	대형	39	남자	40	6년이상	4
3	중형	34	남자	45	6년이상	4
4	중형	33	남자	40	6년이상	4
5	중형	65	남자	40	3년	4
6	소형B	37	남자	70	6년이상	0
7	소형B	36	남자	40	6년이상	4
8	소형B	36	남자	40	6년이상	4
9	다인승Ⅱ	39	남자	40	4년	4
10	소형A	32	여자	50	4년	4
11	소형B	38	남자	40	3년	3
12	소형B	52	여자	40	6년이상	4
13	소형B	39	남자	40	6년이상	4
14	소형B	41	남자	40	6년이상	4
15	중형	39	남자	40	6년이상	3
16	다인승Ⅱ	34	남자	70	4년	4
17	소형B	40	남자	40	6년이상	4
18	다인승Ⅱ	39	남자	40	6년이상	4
19	소형B	35	남자	40	3년	4
20	소형B	38	남자	40	6년이상	3
21	대형	39	남자	40	6년이상	4

표 **4.5.** 언더라이팅 판단 의사결정표

	위험평가				
소비자	사고빈도 (A)	사고심도 (B)	위험보험료 (C)	판단기준 (D)	인수여부 (E)
1	0.0776	1,312,717	101,921	128,287	인수
2	0.0701	1,369,319	96,005	128,287	인수
3	0.0773	1,250,002	96,671	128,287	인수
4	0.0731	1,203,293	87,966	128,287	인수
5	0.0825	1,411,375	116,387	128,287	인수
6	0.0743	2,115,467	157,120	128,287	**거절**
7	0.0644	1,138,039	73,242	128,287	인수
8	0.0644	1,138,039	73,242	128,287	인수
9	0.0849	1,453,472	123,379	128,287	인수
10	0.0719	1,157,015	83,227	128,287	인수
11	0.0727	1,197,923	87,080	128,287	인수
12	0.0853	1,171,980	99,923	128,287	인수
13	0.0658	1,152,015	75,772	128,287	인수
14	0.0673	1,161,336	78,144	128,287	인수
15	0.0862	1,206,366	104,006	128,287	인수
16	0.1010	1,655,768	167,178	128,287	**거절**
17	0.0665	1,156,792	76,909	128,287	인수
18	0.0863	1,341,385	115,745	128,287	인수
19	0.0622	1,205,023	74,913	128,287	인수
20	0.0750	1,127,644	84,594	128,287	인수
21	0.0701	1,369,319	96,005	128,287	인수

5. 결론 및 시사점

본 연구에서는 FY2005의 우리나라 자동차보험 전체통계를 활용하여 일반화가법부분선형모형을 적용하여 보았다. 그 결과 담보별로 사고빈도와 사고심도에 적합한 일반화가법모형은 대인배상 I 의 경우 각각 음이항분포와 로그정규분포 가정이 적합하였고, 대물배상의 경우에도 동일하였다. 이와 더불어 모형에 포함된 변수들인 차종, 성별, 연령, 가입경력, 할인할증, 법규위반 경력 중에서 법규위반을 제외한 모든 변수가 자동차보험 사고를 설명할 수 있는 의미있는 변수인 것으로 분석되었다. 다만, 법규위반 항목별 위험도를 보면, 교통법규를 위반한 자가 위반하지 않은 자 보다 사고위험도가 약10%(p) 높은 것으로 나타났다. 따라서 교통법규위반 유형에 대하여는 추가 분석이 필요한 것으로 보인다. 자동차보험의 보험료가 사고위험을 줄이는 유인요소로 활용될 수 있으므로, 사고위험과 밀접한 관계가 있는 교통법규위반 경력 자동차보험 제도를 개선하는 방법을 추가 연구를 통해 찾을 필요가 있는 것으로 판단된다.

본 연구 결과는 자동차보험 요율산출분야, 언더라이팅 시스템 구축분야 등과 같이 자동차보험의 여러 분야에 활용될 수 있을 것이다. 또한 현재 자동차보험 요율에 사용되고 있는 변수의 유의성을 판단하는 데에도 활용될 수 있을 것이다. 따라서 일반화가법모형이 본 연구에서 살펴본 언더라이팅 분야 이외에도 다양한 분야에 활용될 수 있는 방안에 대한 추가 연구가 필요하다고 생각된다.

참고문헌

보험개발원 (2008). 자동차보험 참조순보험료 요율.

Fan, J., Hardle, W. and Mammen, E. (1998). Direct estimation of low dimensional components in additive models, *Annals of Statistics*, **26**, 943–971.

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004a). *Nonparametric and Semiparametric Models*, Springer, Heidelberg.

Härdle, W., Huet, S., Mammen, E. and Sperlich, S. (2004b). Bootstrap inference in semiparametric generalized additive models, *Econometric Theory*, **20**, 265–300.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models (with discussion). *Statistical Science*, **1**, 297–318.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*, Vol 43 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

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

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models using smoothing splines. *Journal of the Royal Statistical Society. Series B,* **61**, 381–400.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A*, **135**, 370–384.

Speckman, P. E. (1988). Regression analysis for partially linear models. *Journal of the Royal Statistical Society. Series B*, **50**, 413–436.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, **13**, 689–705.

Wood, S. N. (2000). Modeling and Smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society. Series B,* **62**, 413–428.

Wood, S. N. (2003). Thin plate regression splines. *Journal of The Royal Statistical Society. Series B,* **65**, 95–114.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statical Association*, **99**, 673–686.

Wood. S. N. (2006). *Generalized Additive Models : An Introduction with R*, Chapman Hall, London.

Using Generalized Addtive Partial Linear Model for Constructing Underwriting System

Seungdo Ki[1] · Kee-Hoon Kang[2]

[1]Department of Statistics, Hankuk University of Foreign Studies, Korea Insurance Research Institute
[2]Department of Statistics, Hankuk University of Foreign Studies

(Received November 2009; accepted December 2009)

Abstract

Underwriting refers to the process that the insurance company measures the potential risk of the future clients and decide whether insuring them with current premium. Although the traditional underwriting system used in Korean automobile insurance market is easy to understand, it is not based on a reliable statistical procedure. In this paper, we propose to apply the generalized additive model into construction of underwriting system, which is based on statistical analysis. We use automobile insurance data in Korea and apply our approach to the data. The results from the empirical analysis would be useful even for determining the significance of each variable in calculating automobile insurance premium.

Keywords: Generalized additive model, automobile insurance, link function.

This research was supported by the research fund of Hankuk University of Foreign Studies, 2009.
[2]Corresponding author: Associate Professor, Department of Statistics, Hankuk University of Foreign Studies, Mohyun, Cheoin-Koo, Yongin 449-791, Korea. E-mail: khkang@hufs.ac.kr