응용통계연구 (2011)
24(3), 485–494
DOI: 10.5351/KJAS.2011.24.3.485

제로팽창 모형을 이용한 보험데이터 분석

최종후[1] · 고인미[2] · 전수영[3]

[1]고려대학교 정보통계학과, [2]고려대학교 정보통계학과, [3]고려대학교 정보통계학과

(2011년 4월 접수, 2011년 5월 채택)

요 약

계수(Count) 데이터는 반응변수가 음이 아닌 계수로, 자동차 사고건수나 지진이 일어난 횟수, 보험처리 발생건수 등을 말한다. 이런 경우에는 주로 포아송 회귀모형을 사용하지만, 평균과 분산이 동일한 경우만 이용될 수 있다는 제약이 따른다. 실증적 자료에서는 그룹 간 이질성으로 인해 분산이 매우 큰 과대산포(Overdispersion) 현상을 볼 수 있는데, 이를 무시할 경우 회귀계수나 표준오차가 편의되는 현상이 발생한다.

보험은 보장성 개념이 강하기 때문에 실제로 보험처리가 발생하지 않는 경우가 많아, 보험처리 건수에 '0'값이 있을 수 있다. 본 논문에서는 '0'값이 많은 자료의 분석을 위해 제로팽창 모형(Zero-Inflated Model)을 고려하고, 여러 모형들의 효율성을 실증자료를 통하여 비교하였다. 실증 자료 분석 결과, 과대산포와 제로팽창 현상이 존재하는 자료에서 제로팽창 음이항 모형(Zero-Inflated Negative Binomial Regression Model)이 가장 효율적인 모형임을 보여 주었다.

주요용어: 계수데이터, 과대산포, 제로팽창 모형, 보험보장.

1. 서론

보험연구원 (변혜원과 박정희, 2010)에 의하면, 현재 우리나라의 자동차 보유율은 가구기준 79.2%로, 10가구 중 약 8가구가 자동차를 소유하고 있다고 한다. 이에 따라 자동차 보험도 수입보험료가 연간 7조원을 넘는 매우 큰 규모의 시장이 되었다. 보험사들은 다양한 채널을 통해 자사의 고객을 유치하는 한편, 다양한 보험 상품개발을 통해 이들의 요구를 신속히 반영하는 등 치열한 고객유치 경쟁을 벌이고 있다. 가입자가 보험 상품을 가입할 때 가장 중요하게 고려하는 부분은 보장내용일 것이다. 계약자는 보험 가입 후에도 보장내용이 충분하지 않을 때는 언제든지 그 계약내용을 변경할 수 있는데, 그 변경내용이 대부분 특약추가나 가입자의 보장을 강화하는 것이므로 이는 곧 보험회사의 수익과도 직결된다고 볼 수 있다. 그러므로 보험회사에서는 계약자들의 특성이나 행동들에 대해서 주시하고 분석할 필요가 있다.

해외에서는 보험 분야의 고객관리강화를 위해 다양한 통계기법을 도입하고 있는데, 국내에서도 요율산출부분이나 보험 사고빈도, 사고심도 등 보험통계분석 (기승도와 김대환, 2009)과 보험가격 산출 방법 (김명준과 김영화, 2009)에 일반화선형모형(Generalized Linear Model; GLM)을 활용하고 있다 (전희주 등, 2009). 특히, 자동차보험 사고건수와 같이 반응변수가 건수인 자료는 포아송 분포나 음이항 분포를 따르는 것으로 분석되고 있다 (Cameron과 Trivedi, 1998).

[3]교신저자: (339-700) 충남 연기군 조치원읍 서창리 208, 고려대학교 정보통계학과, 조교수.
E-mail: scheon@korea.ac.kr

포아송 회귀모형은 포아송 분포에서 나온 모형으로 평균과 분산이 동일한 경우를 가정한다. 그러나 실제로 많은 데이터에서 분산이 평균보다 매우 큰 과대산포(Overdispersion) 현상을 볼 수 있는데, 이러한 과대산포를 고려하지 않는다면, 모수 추정치와 표본오차는 비효율성을 가지게 될 것이다 (Cox, 1983; Grogger과 Carson, 1991). 음이항 회귀모형은 이러한 과대산포를 고려한다는 장점이 있다. 또한 보험자료의 특성상 반응변수가 건수인 경우에는 0이 많이 포함되어 있다. 이러한 성격의 자료에 대해서는 0값 과다분포 현상을 고려한 분석이 요구된다. 따라서 과대산포와 제로팽창 현상이 존재할 경우 제로팽창 음이항 모형(Zero-Inflated Negative Binomial Regression Model)이 이러한 효과를 모형에서 반영할 수 있기 때문에 많이 이용되고 있다 (박상일, 2009). Dean과 Lawless (1989), Gurmu (1991), Jung 등 (2006) 그리고 Ridout 등 (2001) 등이 이러한 여러가지 포아송 회귀모형에서의 과대산포에 대한 연구를 진행해 왔다.

본 연구에서는 계수데이터에서 나타날 수 있는 과대산포와 '0'값의 과다포함 문제가 있는 데이터를 이용하여 보험계약내용 변경건수를 반응변수로 두고 설명변수들과 관계를 살펴보기 위해, 포아송 회귀모형(Poisson Regression Model; P), 음이항 회귀모형(Negative Binomial Regression Model; NB), 제로팽창 포아송 회귀모형(Zero-Inflated Poisson Regression Model; ZIP), 제로팽창 음이항 회귀모형(Zero-Inflated Negative Binomial Regression Model; ZINB)들의 적용 결과를 비교해 보았다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 비교 모형으로 포아송, 음이항, 제로팽창모형을 소개하고, 과대산포와 제로팽창에 대한 가설 검정와 모형평가방법에 대해 서술한다. 3장에서는 실증 자료를 가지고 과대산포와 제로팽창에 대한 가설 검정과 비교 모형들을 평가해 본다. 마지막으로 4장에서는 결론을 다룬다.

2. 추정을 위한 기본개념

2.1. 포아송 회귀모형

포아송 확률변수 Y_i는 모수 μ_i에 의존하며 평균과 분산은 $E(Y_i|X_i) = \mathrm{Var}(Y_i|X_i) = \mu_i$로 같다. 로그-선형 형태에서 포아송 모형의 평균은 식 (2.1)과 같이 나타낼 수 있다.

$$\begin{aligned} E(Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i) &= \mu_i \\ &= \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \\ &= \exp(x_{i1}\beta_{i1})\exp(x_{i2}\beta_{i2})\cdots\exp(x_{ik}\beta_{ik}). \end{aligned} \tag{2.1}$$

식 (2.1)에서 $\boldsymbol{x}_i = (x_{i1}, x_{i1}, \ldots, x_{ik})'$는 설명변수의 벡터이고 $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)'$는 추정계수이다. 로그-선형 형태 변환 시 가장 많이 쓰이는 연결함수는 항등연결함수와 로그연결함수인데, 위 식에서는 로그연결함수를 이용하였다. 포아송 회귀모형에서 회귀계수 $\boldsymbol{\beta}$에 대한 추정은 최대우도추정(Maximum Likelihood Estimation)방법을 이용한다. 포아송 모형의 로그우도함수는 다음과 같이 나타낼 수 있다.

$$\begin{aligned} \ln L(\boldsymbol{\beta}) &= \sum_{i=1}^{n}(-\mu_i + y_i \ln \mu_i - \ln y_i!) \\ &= \sum_{i=1}^{n}\left(-\exp(\boldsymbol{x}_i'\boldsymbol{\beta}) + y_i\boldsymbol{x}_i'\boldsymbol{\beta} - \ln y_i!\right). \end{aligned} \tag{2.2}$$

이 때 식 (2.2)에서 로그우도함수의 회귀계수 $\boldsymbol{\beta}$의 추정은 수치적 반복에 의해 구해질 수 있는데, 그 대표적인 방법은 뉴튼-랍슨방법(Newton-Raphson Method)이다 (Cameron과 Trivedi, 1998).

2.2. 음이항 회귀모형

음이항 분포는 포아송 분포와 감마분포의 혼합 분포로, 과대산포를 고려한 분포이다. 음이항 분포는 분산이 평균보다 큰 값을 가질 수 있도록 모수를 추가적으로 하나 더 갖는다. 음이항 모형에서 y_i에 대한 분산을 식 (2.3)과 같이 정의한다. 고정된 p에 대해서,

$$\begin{aligned} w_i &= \mathrm{Var}(Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i) \\ &= w(\mu_i, \alpha) \\ &= \mu_i + \alpha\mu_i^p, \quad \text{for} \;\; 0 < p < 1. \end{aligned} \tag{2.3}$$

식 (2.3)에서 p와 α는 각각 상수 파라미터와 과대산포 파라미터이다. 가장 일반적인 모형은 분산($p = 2$)이 $w_i = \mu_i + \alpha\mu_i^2$ 형태의 음이항 분포(NB2)이다.

음이항 회귀모형은 포아송 분포와 감마분포가 결합된 음이항 분포을 가정한 모형으로, 평균이 $\mu_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$인 NB2 모형의 로그우도함수는 다음과 같이 나타낼 수 있다.

$$\ln L(\alpha\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(\sum_{j=0}^{y_i-1}\ln\left(j+\alpha^{-1}\right) - \ln y_i! - \left(y_i+\alpha^{-1}\right)\ln\left(1+\alpha\exp\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)\right) + y_i\ln\alpha + y_i\boldsymbol{x}_i'\boldsymbol{\beta}\right). \tag{2.4}$$

음이항 분포의 과대산포 α와 회귀계수 $\boldsymbol{\beta}$에 대한 추정방법 또한 수치적 반복에 의해 MLE를 구하는 방법인 뉴튼-랍슨방법을 이용한다 (Cameron과 Trivedi, 1998).

2.3. 제로팽창 회귀모형

계수데이터에서는 사건이 발생하는 경우보다 발생하지 않는 경우가 대부분이다. 그러므로 반응변수가 건수인 경우에 0값의 분포가 크다. 이러한 특수한 성격의 자료에 대해서는 0값의 과다분포 현상을 고려한 분석이 요구된다. 제로팽창 모형은 위의 효과를 모형에서 반영할 수 있기 때문에 제로팽창이 존재하는 자료에 많이 이용되고 있다.

y_i에 대하여 두가지 가능한 분포를 고려한다. 베르누이 시행에 대한 결과는 y_i값이 어느 분포를 따르는지 결정하는데 사용된다.

$$y_i = \begin{cases} 0, & \text{with probability} \;\; \phi_i, \\ g(y_i), & \text{with probability} \;\; 1-\phi_i. \end{cases} \tag{2.5}$$

본 연구에서는 포아송 분포와 음이항 분포 각각을 제로팽창 모형과 혼합한 제로팽창 포아송 회귀모형과 제로팽창 음이항 회귀모형을 이용한다. 제로팽창 모형에 대한 확률질량함수는 다음과 같다.

$$f(y_i|x_i) = \begin{cases} \phi_i + (1-\phi_i)g(0), & \text{for} \;\; y_i = 0, \\ (1-\phi_i)g(y_i), & \text{for} \;\; y_i = 1, 2, \ldots, \end{cases} \tag{2.6}$$

여기서 $g(y_i)$는 포아송 또는 음이항 분포를 따르며, $\phi_i (0 < \phi_i < 1)$는 0에서의 팽창확률을 나타낸다.

먼저, 제로팽창 포아송 회귀모형(ZIP(μ_i))에서 확률질량함수는 식 (2.7)과 같이 정의할 수 있다.

$$f(y_i|x_i) = \begin{cases} \phi_i + (1-\phi_i)e^{-\mu_i}, & \text{for} \;\; y_i = 0, \\ (1-\phi_i)\dfrac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, & \text{for} \;\; y_i = 1, 2, \ldots. \end{cases} \tag{2.7}$$

식 (2.7)의 제로팽창 확률 ϕ_i를 로지스틱 연결함수에 의해 식 (2.8)과 같이 $\boldsymbol{Z}_i'\boldsymbol{\gamma}$로 변환하여 모형화 할 수 있다.

$$\phi_i = \frac{\exp(\boldsymbol{Z}_i'\boldsymbol{\gamma})}{1+\exp(\boldsymbol{Z}_i'\boldsymbol{\gamma})}. \tag{2.8}$$

여기서, $\boldsymbol{Z}_i = (z_{i1}, z_{i1}, \ldots, z_{ik})'$, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_k)'$이다.

$\boldsymbol{\beta}$와 $\boldsymbol{\gamma}$에 대한 MLE를 구하는 데 필요한 결합 로그우도함수는 다음과 같이 구해진다.

$$\begin{aligned}\ln L(\boldsymbol{\beta},\boldsymbol{\gamma}|y_i) = &\sum_{i=1}^{n} I(y_i=0)\ln\left[\exp(\boldsymbol{Z}_i'\boldsymbol{\gamma}) + \exp(-\exp(\boldsymbol{x}_i'\boldsymbol{\beta}))\right] \\ &+ \sum_{i=1}^{n}(1-I(y_i=0))\left(y_i\boldsymbol{x}_i'\boldsymbol{\beta} - \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right) \\ &- \sum_{i=1}^{n}\ln\left(1+\exp(\boldsymbol{Z}_i'\boldsymbol{\gamma})\right).\end{aligned} \tag{2.9}$$

여기서, $I(y_i=0)$은 $y_i=0$일 때 1의 값을 가지는 더미변수이다.

다음으로 제로팽창 음이항 회귀모형에서 확률질량함수는 식 (2.10)과 같이 정의할 수 있다.

$$f(y_i|x_i) = \begin{cases} \phi_i + (1-\phi_i)(1+\alpha\mu_i)^{-\alpha^{-1}}, & \text{for } \ y_i = 0, \\ (1-\phi_i)\dfrac{\Gamma(y_i+\alpha^{-1})}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}\left(\dfrac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\dfrac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i}, & \text{for } \ y_i = 1,2,\ldots. \end{cases} \tag{2.10}$$

식 (2.10)과 같은 분포를 갖는 확률변수를 ZINB(μ_i,α)라고 정의하고, 제로팽창 확률 ϕ_i에 대해서도 로지스틱 연결함수에 의해 표현될 수 있다. ZINB 모형에서 $\boldsymbol{\beta}$와 $\boldsymbol{\gamma}$에 대한 MLE를 구하는 데 필요한 로그우도함수는 다음과 같이 구해진다 (Cameron과 Trivedi, 1998).

$$\begin{aligned}&\ln L(\boldsymbol{\beta},\boldsymbol{\gamma},\alpha|y_i) \\ &= \sum_{i=1}^{n} I(y_i=0)\left\{\frac{\exp(\boldsymbol{Z}_i'\boldsymbol{\gamma})}{1+\exp(\boldsymbol{Z}_i'\boldsymbol{\gamma})} + \frac{1}{1+\exp(\boldsymbol{Z}_i'\boldsymbol{\gamma})}(1+\alpha\mu_i)^{-\alpha^{-1}}\right\} \\ &\quad + \sum_{i=1}^{n}(1-I(y_i=0))\left(\sum_{j=1}^{y_i}\log(y_i\alpha+1-\alpha j) - (\alpha^{-1}+y_i)\log(1+\alpha\mu_i) + y_i\log\mu_i - \log y_i!\right) \\ &\quad - \sum_{i=1}^{n}\ln(1+\exp(\boldsymbol{Z}_i'\boldsymbol{\gamma})).\end{aligned} \tag{2.11}$$

로그우도 값을 최대화하는 $\boldsymbol{\beta}$와 $\boldsymbol{\gamma}$에 대한 MLE를 구하기 위해 여러 방법이 이용될 수 있는데, SAS 프로그램의 PROC COUNTREG 프로시저에서는 비선형 모형에 대한 최적화 방법으로 Quasi-Newton, Newton-Raphson, Trust region 세 가지 방법을 이용하고 있다.

2.4. 과대산포와 제로팽창에 대한 가설검정

설명변수 유의성에 대한 검정방법에는 우도비 검정, Wald검정, 스코어 검정 등이 있다. 우도비 검정(Likelihood Ratio; LR)은 선형모형에서 설명변수 유의성에 대한 검정방법으로 우도 또는 로그우도를 고려한 것이다. $\hat{\boldsymbol{\beta}}$를 $\boldsymbol{\beta}$의 비제한(unconditional) MLE라 하고, '설명변수들이 모두 0은 아니다'라는

표 **3.1.** 보험계약 변경횟수 분포

계약내용 변경건수	빈도	백분율	누적 백분율
0 건	7,821	76.68	76.68
1 건	757	7.42	84.10
2 건	1,032	10.12	94.22
3 건	207	2.03	96.25
4 건	203	1.99	98.24
5 건	60	0.59	98.82
6 건	50	0.49	99.31
7 건	26	0.25	99.57
8 건	18	0.18	99.75
9 건	6	0.06	99.80
10 건이상	20	0.20	100.00
합계	10,200	100.00	100.00

제약조건 하에서 로그우도함수 $l(=\log L)$이 최대화가 될 때 $\boldsymbol{\beta}$의 MLE를 $\tilde{\boldsymbol{\beta}}$라 하자. 그리고 $\hat{L}$를 $\hat{\boldsymbol{\beta}}$에서의 L의 값이라 하고, $\tilde{L}$를 $\tilde{\boldsymbol{\beta}}$에서의 L의 값이라고 하자. 우도비 검정에서는 $\hat{L}$과 $\tilde{L}$이 비교된다. $\hat{l}$과 $\tilde{l}$를 각각 모형의 로그우도라 할 때, 우도비와 검정통계량은 각각 식 (2.12)와 식 (2.13)으로 정의된다 (Piet와 Gillian, 2008).

$$\lambda = \frac{\hat{L}}{\tilde{L}}, \tag{2.12}$$

$$2\ln\lambda = 2\left(\hat{l} - \tilde{l}\right). \tag{2.13}$$

2.5. 모형평가

모형의 설명력은 모형에 대한 측도인 결정계수(R^2)나 수정결정계수(Adjusted R^2), Mallow C_p 및 AIC(Akaike Information Criterion), BIC(Bayesian Information Criterion) 등의 통계량을 통해 측정될 수 있다 (Berry와 Linoff, 1997; 강현철 등, 2001). 또한 모형적합을 판단할 때 AIC 또는 BIC 중 어떤 통계량을 쓸 것인지를 고민해야 한다. BIC는 모수의 수에 보다 많은 제약을 둠으로써 AIC에 비해 모수가 작은 모형을 고르는 경향이 있다. 자료의 수가 큰 보험 데이터의 경우에는 BIC가 너무 단순하다고 느껴지는 모형을 선택하는 경향이 있다. 이 경우에는 AIC가 선호된다 (Piet와 Gillian, 2008).

3. 사례연구

3.1. 분석자료

본 연구를 위하여 사용된 자료는 1996년 4월부터 1998년 10월까지 A보험회사의 DB에 축적된 고객 데이터로, 보험계약내용 변경횟수를 비롯한 여러 고객 정보들로 이루어져 있다. 데이터 정제 후, 경제활동 가능인구인 20세 이상 60세 미만의 보험가입자들을 중점적으로 분석하기 위해 최종적으로 10,200명의 데이터를 분석에 사용하였다. 반응변수는 보험계약내용 변경횟수이고, 분포는 다음과 같다. 표 3.1에서 반응변수 값의 평균은 0.5건이고 분산은 1.6으로, 해당기간 내에 보험계약내용을 변경하지 않은 가입자는 76.7%이다. 설명변수별 반응변수의 평균 및 분산을 살펴보면 표 3.2와 같다. 표 3.2와 같이 성별로 보았을 때 여성보다 남성에서 분산/평균의 차이가 더 크고, 연령별 비교에서는 50–59세에서 분산/평균이 가장 큰 차이를 보이고 있다. 전체적으로 볼 때 분산은 평균보다 약 3배 정도 크다. 보험계약내용 변

표 **3.2.** 연령·성별에 따른 계약내용 변경횟수 평균 및 분산

	전체	연령			성별	
		0∼39	40∼49	50∼59	남	여
평균	0.536	0.463	0.566	0.533	0.542	0.510
분산	1.622	1.333	1.678	1.694	1.663	1.412
분산/평균	3.026	2.879	2.964	3.178	3.068	2.768
관측치	10,200	1,842	4,925	3,433	8,522	1,678

표 **3.3.** 분석자료의 변수

변수	변수명	변수 설명	구분
ID	고객번호	고객 일련번호	
NY	보험기간	총 보험금 납입기간	3/5/7/10년
HP	합계보험료	총 보험료	10만원 미만/10만∼50만원/ 50만원∼100만원/100만원 이상
BG	변경코드	계약변경 내용에 따라 코드로 구분	내용변경/해약/이재통보/중도금
AGE	계약자 연령	연령	20대/30대/40대/50대
SEX	계약자 성별	성별	남자/여자
JBHS	증권 발행횟수	보험증권을 고객에게 발행한 횟수	0∼26건
IBHS	입금횟수	보험금 총 납입횟수	0∼95건
IJHS	이재횟수	사고로 재난이나 재해를 당한 횟수	0∼7건
CHHOI	계약내용변경횟수	초기 보험계약내용 변경 횟수	0∼29건

경횟수는 대표적인 계수형 자료로 주로 포아송 분포를 따르지만, 본 자료에서와 같이 평균에 비해 분산이 큰 경우에는 다른 분포를 고려해야 한다.

또한, 표 3.1에서 반응변수의 분포를 살펴보면 '0'이 차지하는 비율이 76.7%로 매우 높다. 실제로 많은 자료들에서 '0'값 과다분포 현상이 존재하므로, 이러한 효과를 반영할 수 있는 모형을 고려해야 한다. 본 연구에서는 보험데이터 특성과 함께 반응변수인 보험계약내용 변경건수에 영향을 미치는 설명변수를 자료에 적합한 회귀모형을 이용하여 알아보고자 한다. 최종적으로 모형개발에 이용된 변수는 표 3.3과 같다.

3.2. 과대산포·제로팽창 검정

과대산포 유무에 대한 가설은 다음과 같이 표현할 수 있다. α를 과대산포 파라미터라고 하면,

$$H_0 : \alpha = 0 \quad vs. \quad H_1 : \alpha > 0. \tag{3.1}$$

이 경우 귀무가설이 참이면 과대산포가 존재하지 않고, 귀무가설이 거짓이면 과대산포가 존재한다. 본 자료에서는 SAS 프로그램을 이용한 분석 결과 유의수준 0.05하에서 과대산포($\alpha = 4.04$)가 존재함을 알 수 있다. 또한 우도비 검정을 통해 과대산포 존재여부를 살펴보면, 포아송 모형과 음이항 모형의 LR 검정통계량은 자유도 1인 카이제곱분포를 따르며 검정통계량 값은 3456으로 유의수준 0.05에서 기각역에 속한다. 따라서 본 연구에 사용된 데이터에 과대산포가 존재함을 알 수 있다. 제로팽창 포아송 회귀모형과 제로팽창 음이항 회귀모형의 LR 검정통계량도 580으로 유의수준 0.05하에서 과대산포가 존재한다.

제로팽창 확률을 ϕ_i라고 할 때, 제로팽창 유무에 대한 가설은 다음과 같이 표현할 수 있다.

$$H_0 : \phi_i = 0 \quad vs. \quad H_1 : \phi_i > 0. \tag{3.2}$$

표 **3.4.** 모형 비교

모형	$\ln L$	AIC	BIC	Dispersion Parameter [α(p-value)]
P(포아송회귀모형)	−10679	21389	21497	
NB(음이항회귀모형)	−8951	17934	18050	4.08 (<0.001)
ZIP(제로팽창포아송회귀모형)	−9173	18381	18503	
ZINB(제로팽창음이항회귀모형)	−8884	17807	17952	0.71 (<0.001)
|P-ZINB|		3582	3545	

우도비 검정을 통해 제로팽창 존재여부를 살펴보면, 포아송 회귀모형과 제로팽창 포아송 회귀모형의 LR 검정통계량은 자유도 1인 카이제곱분포를 따르며 검정통계량 값은 3012로 유의수준 0.05에서 기각역에 속하므로 귀무가설을 기각한다. 또, 음이항 회귀모형과 제로팽창 음이항 회귀모형의 LR 검정통계량도 134로 유의수준 0.05하에서 귀무가설을 기각한다. 그러므로 본 연구에 사용된 데이터에 제로팽창이 존재함을 알 수 있다.

3.3. 모형평가

3.1절 기초통계량과 3.2절을 통해 본 연구에 사용된 데이터는 과대산포와 제로팽창이 존재함을 알 수 있었다. 이제 반응변수인 보험계약내용 변경건수에 영향을 미치는 요인을 계수자료에 적합한 회귀모형을 비교한 후, 가장 적합한 모형을 선택하고자 한다. 연결함수는 로짓모형을 이용하였고, 회귀모형의 $\boldsymbol{\beta}$값에 대한 MLE는 뉴튼-랍슨방법을 이용해 구하였다. 최종적으로 적합된 4가지 모형은 모형 선호 기준통계량인 로그우도($\ln L$), AIC, BIC로 평가하였다. 모형평가 시 일반적으로 AIC와 BIC 값이 작은 모형을 선택하는데 그 결과는 표 3.4와 같다.

일반적으로 로그우도 값은 크고, AIC 값과 BIC 값은 작을수록 우수한 모형으로 판단한다. 표 3.4에서 포아송 회귀모형(P)과 음이항 회귀모형(NB)의 $\ln L$, AIC, BIC 값을 비교해 본 결과, NB 모형이 더 우수한 것으로 판단된다. 또, 포아송 회귀모형과 제로팽창 포아송 회귀모형(ZIP)의 $\ln L$, AIC, BIC 값을 비교했을 때, ZIP 모형이 더 우수한 모형으로 나타났다.

구간별 비교결과 계약 미변경은 7821건으로, NB 모형이 실제값과 가장 근사한 7717건으로 예측되었다(표 3.5). 계약변경 건수가 0∼2건은 대체적으로 ZIP, ZINB 모형이 실제값과 근사하게 예측되었고, 3∼5건은 NB 모형이 실제 관측값과 가장 근사하게 예측되는 것으로 나타났다. 6건 이상일 때는 ZINB 모형이 실제값과 가장 근사하게 예측되었다. 구간별로 우수한 모형은 다르지만 종합적으로 보았을 때, 구간별 모형성능을 비교한 그림 3.1에 의해 모형의 우수성 정도는 ZINB > NB > ZIP > P로 나타났다.

3.4. 결과및 해석

연령(AGE)별로 살펴보면, 34세 미만 보험자에 비해 35∼44세 보험자의 보험계약 변경횟수는 $e^{0.203} = 1.225$로 22% 더 많고, 45∼54세 운전자는 $e^{0.120} = 1.127$로 13% 더 많다. 또, 55세 이상 운전자는 34세 미만 보험자에 비해 보험계약 변경횟수가 $e^{0.036} = 1.036$로 4%정도 높다. 즉, 34세 미만의 젊은 운전자들에 비해 35세 이상 운전자들의 계약변경횟수가 많음을 알 수 있다. 35∼44세 보험계약자들은 활발한 사회활동, 결혼으로 인한 가족 수의 증가 등 변화가 많은 시기이므로 이런 결과가 도출된 것으로 보인다.

성별(SEX)로 살펴보면, 여자가 계약을 변경한 횟수는 $e^{-0.129} = 0.878$로 남자에 비해 약 12% 적은 것으로 나타났다. 이는 여성운전자의 수에 비해 남성운전자들이 수가 많고, 그만큼 사고위험성도 크기 때문으로 생각된다.

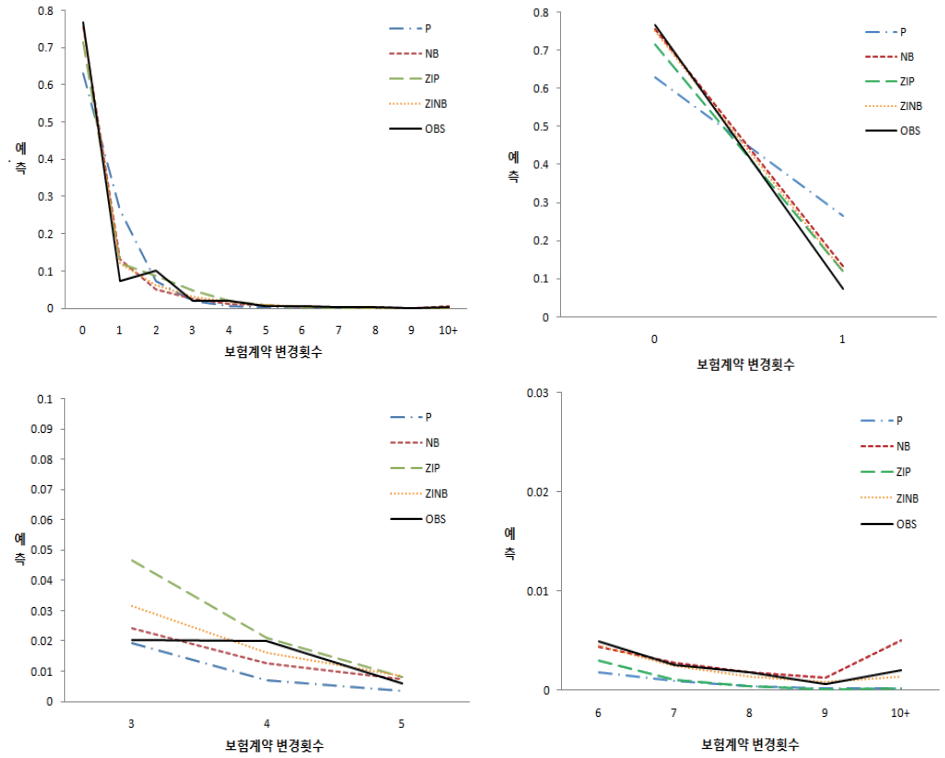

그림 **3.1.** 반응변수에 대한 모형비교

표 **3.5.** 모형별 예측값과 MSE

반응변수	Observed		Predicted				MSE			
	빈도	백분율	P	NB	ZIP	ZINB	P	NB	ZIP	ZINB
0	7821	76.67	6422	7717	7288	7662	24.0	0.1	3.5	0.3
1	757	7.42	2700	1357	1220	1225	479.1	45.7	27.2	27.8
2	1032	10.11	742	520	874	635	7.8	24.4	2.3	14.7
3	207	2.02	196	246	475	322	0.1	0.7	33.2	6.2
4	203	1.99	71	130	213	165	8.3	2.5	0.1	0.7
5	60	0.58	35	74	84	86	1.0	0.3	0.9	1.1
6	50	0.49	18	45	30	45	1.9	0.1	0.8	0.0
7	26	0.25	9	28	10	25	1.0	0.0	0.9	0.0
8	18	0.17	4	19	3	14	1.0	0.0	1.1	0.1
9	6	0.05	2	13	1	8	0.3	0.8	0.4	0.1
10+	20	0.10	2	51	2	14	1.6	0.9	1.7	0.3
합계	10200	100.00	10200	10200	10200	10200	526.2	75.5	72.1	51.2

합계보험료(HP)별로 보면, 10만원 미만 납부자들에 비해 10~50만원 납부자들의 계약변경 횟수가 $e^{0.200} = 1.221$로 22% 더 많고, 50~100만원 납부자들은 $e^{0.638} = 1.892$로 약 90%정도 계약변경횟수가 높다. 또, 100만원 이상 납부자들은 $e^{1.231} = 3.424$로 10만원 미만 납부자들에 비해 240% 더 많은 것으

로 나타났다. 2010 보험소비자 설문조사에서 연간 자동차 보험료 수준은 41~70만원이 가장 많은 것으로 나타났는데, 평균이상으로 많은 보험금을 지불하는 운전자 일수록 더 많은 보험혜택을 요구하는 것으로 생각된다.

4. 결론

본 연구에서는 자동차 보험 자료를 이용하여 보험계약내용 변경횟수를 반응변수로 두고, 자료에 적합한 회귀모형에 대해 살펴보았다. 고려한 4가지 회귀모형은 포아송 회귀모형, 음이항 회귀모형, 제로팽창 포아송 모형, 제로팽창 음이항 모형으로, 본 연구에 사용된 데이터와 같이 과대산포와 제로팽창 현상이 존재하는 자료에서는 제로팽창 음이항 모형이 가장 좋은 모형이라 할 수 있다.

본 데이터는 계약내용 변경횟수가 1건보다 2건이 더 많은 특징을 가지는데, 이는 구간별로 우수한 모형이 각각 다르게 선택된 원인 중 하나로 생각된다. 즉, 0건 그룹에 대해서만 고려할 것이 아니라 2건 그룹에 대해서도 고려할 필요가 있다. 본 연구에서 제안한 제로팽창 모형은 0건과 그 외, 즉 두 그룹으로만 나누어 고려하는 모형이다. 따라서 제 3그룹에 대해서도 같이 고려할 수 있는 새로운 모형을 개발한다면 데이터를 더욱 잘 설명할 수 있을 것이다. 뿐만 아니라, 계수데이터 분석에 이용되는 모형 중 하나인 Hurdle모형 분석을 추가하여 제로팽창 모형과 비교해 보는 것도 좋을 것이다.

참고문헌

강현철, 최종후, 한상태 (2001). <데이터마이닝: 방법론 및 활용>(제3판), 자유아카데미, 서울.

기승도, 김대환 (2009). 일반화선형모형(GLM)을 이용한 자동차보험 요율상대도 산출방법 연구, <보험연구원>.

김명준, 김영화 (2009). 다양한 모형화를 통한 자동차 보험가격 산출, *Journal of the Korean Data & Information Science Society*, **20**, 515–526.

박상일 (2009). <제로팽창 음이항 회귀모형을 이용한 MMS 사용횟수에 대한 분석>, 서울시립대학교 대학원 통계학과 석사학위 논문.

변혜원, 박정희 (2010). 2010 보험소비자 설문조사, <보험연구원>.

전희주, 최용석, 최종후, 기승도, 김은석 (2009). <보험자료를 활용한 일반화 선형모형>, 사이플러스, 서울.

Berry, M. J. A. and Linoff, G. (1997). *Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley & Sons, New York.

Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis of Count Data*, 7th, Cambridge University press, New York.

Cox, D. R. (1983). Some remarks on overdispersion, *Biometrika*, **70**, 269–274.

Dean, C. and Lawless, F. (1989). Tests for detecting overdispersion in Poisson regression models, *Journal of the American Statistical Association*, **84**, 467–472.

Grogger, J. and Carson, R. (1991). Models for truncated counts, *Journal of Applied Econometrics*, **6**, 225–238.

Gurmu, S. (1991). Tests for detecting overdispersion in the positive Poisson regression model, *Journal of Business and Economic Statistics*, **9**, 215–222.

Jung, B. C., Jhun, M. and Song, S. H. (2006). Testing for overdispersion in a censored Poisson regression model, *Statistics*, **40**, 533–544.

Piet, D. J. and Gillian, Z. H. (2008). *Generalized Linear Models for Insurance Data*, Cambridge University Press, New York.

Ridout, M., Hinde, J. and Demetrio, C. G. B. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives, *Biometrics*, **57**, 219–223.

A Zero-Inflated Model for Insurance Data

Jong-Hoo Choi[1] · In-Mi Ko[2] · Sooyoung Cheon[3]

[1]Department of Informational Statistics, Korea University
[2]Department of Informational Statistics, Korea University
[3]Department of Informational Statistics, Korea University

(Received April 2011; accepted May 2011)

Abstract

When the observations can take only the non-negative integer values, it is called the count data such as the numbers of car accidents, earthquakes, or insurance coverage. In general, the Poisson regression model has been used to model these count data; however, this model has a weakness in that it is restricted by the equality of the mean and the variance. On the other hand, the count data often tend to be too dispersed to allow the use of the Poisson model in practice because the variance of data is significantly larger than its mean due to heterogeneity within groups. When overdispersion is not taken into account, it is expected that the resulting parameter estimates or standard errors will be inefficient.

Since coverage is the main issue for insurance, some accidents may not be covered by insurance, and the number covered by insurance may be zero. This paper considers the zero-inflated model for the count data including many zeros. The performance of this model has been investigated by using of real data with overdispersion and many zeros. The results indicate that the Zero-Inflated Negative Binomial Regression Model performs the best for model evaluation.

Keywords: Count data, overdispersion, zero-inflated model, insurance coverage.

[3]Corresponding author: Assistant Professor, Department of Informational Statistics, Korea University, Jochiwon 339-700, Korea. E-mail: scheon@korea.ac.kr