Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.1 pp. 45~53

An application to Zero-Inflated Poisson Regression Model[1)]

Kyung Moo Kim[2)]

Abstract

The Zero-Inflated Poisson regression is a model for count data with exess zeros. When the reponse variables have excess zeros, it is not easy to apply the Poisson regression model. In this paper, we study and simulate the zero-inflated Poisson regression model. An real example was applied to this model. Regression parameters are estimated by using MLE's. We also compare the fitness of zero-inflated Poisson model with the Poisson regression and decision tree model.

주제어: 영과잉-포아송 분포, 영과잉-포아송 회귀모형, 포아송 회귀모형

1. 서론

영과잉-포아송분포라 함은 이산형확률분포에 있어서 정상적인 포아송 확률분포보다 영의 값이 과잉관측되는 분포를 말한다. 포아송분포가 생산공정단계에서 발생하는 불량품 수에 관한 확률분포로서 지금까지 중요한 분포로 이용되어 왔다. 그러나 현대문명의 발달과 제품을 만들어내는 기술의 고급화로 인하여 불량률은 현저하게 감소되어 가고 있다. 반응변수가 영이 과잉 관측되는 경우 기존의 포아송 분포에 적용시켜 통계적인 추론을 한다면 이는 제3종의 오류를 범하는 결과를 초래할 것이다.

이러한 공변량이 없는 일변량 영과잉-포아송분포는 Singh(1963)와 Johnson-Kotz(1969)에 의해 소개되었으나 수학적인 모형으로만 인식되어 응용분야가 다양하지 못했다. 그 이후 Heilbron(1989)는 영변경(zero altered)-포아송 음이항 회귀모형을 이용하여 위험요소가 많은 사람들의 행동과학에 대하여 연구하였다. 그는 반

1) The present research was supported by the research fund of Daegu University in 2001.
2) Professor, Department of Statistics, Daegu University, Daegu, Korea.
E-mail: kmkim@daegu.ac.kr

응값이 영인 경우에 확률을 임의로 주는 모형을 생각했다. 영이 되는 확률이 표준 포아송분포보다 적게 되도록 양의 포아송분포를 생각하였다. 그 이후 영변경-포아송 음이항 회귀모형과 유사한 모형을 Lambert(1992)는 제시하였다. 그는 공변량에 의존되는 반응변수가 영과잉-포아송 분포를 따르는 영과잉-포아송분포(zero-inflated Poisson regression model)를 이용한 회귀모형을 소개하였다. 그는 반도체 부품들을 몇 가지 요인으로 나누어 각 경우마다 나타나는 불량개수를 관측한 실제자료에 적용하였다. 회귀계수들은 최우추정법을 이용하여 추정하였고 공변량(covariates)들의 효과를 분석하였다. 그 이후 공변량이 없는 영과잉-포아송분포를 Li 등(1999)은 다변량 영과잉-포아송분포로 확장시켰다. 다변량 영과잉-포아송분포는 많은 모수를 포함하고 있는데 이들의 적률추정량, 최우추정량들과 분포의 성질들을 연구하였다.

본 논문은 영과잉-포아송 회귀모형의 적용사례로서 백화점 고객들의 상품구입횟수에 관한 실제자료를 분석하려고 한다. 백화점 입장에서 보면 백화점의 매출은 고객들의 상품구입 회수에 직결되기 때문에 고개관리 차원에서 본다면 매우 중요한 일일 것이다. 반응변수는 고객이 최근 18개월 안에 구입한 상품구입 회수이다. 반응변수에 영향을 미치는 설명변수 혹은 요인(facter)을 공변량이라고 하자. 고객들의 상품구입 회수는 많은 공변량들에 의해 종속된다고 생각할 수 있다. 본 논문에서 다루고 있는 공변량들로는 고객들의 연령, 수입, 성별, 결혼여부 그리고 집소유 여부이다. 이 자료를 관찰해보면 반응변수가 과잉으로 영이 관측되는 것을 알 수 있다. 반응변수가 계수형 자료(count data)이므로 일반적인 다중회귀모형에 적합시키기는 어렵다. 또한 포아송 회귀모형에 적합시키는것도 반응값이 영이 과잉으로 관측되기 때문에 적용하기 힘들것이다. 본 논문은 영과잉-포아송 회귀모형을 소개하고 실제자료를 이용하여 회귀계수들을 추정하고 모형의 적합성을 포아송 회귀모형 그리고 의사결정나무모형과 비교하여 알아보려고 한다.

2. 영과잉 포아송 회귀모형

영과잉-포아송 회귀모형을 소개하기 위하여 영과잉-포아송 분포를 먼저 설명하기로 한다. 영과잉-포아송분포는 포아송분포와 베르누이분포와의 혼합모형(mixed model)으로 볼 수 있다. 포아송분포에서 0이 과잉 관측되는 경우로 생각할 수 있다.

확률변수 Y는 일정 단위당 나타나는 계수형 자료(count data)로서 영만 나타나는 상태(perfect state)의 확률값이 따로 정해진다. 즉,

$$Y \sim 0, \quad p \text{ 의 확률로}$$

$$\sim Poisson(\lambda),\ 1-p \text{의 확률},$$

여기에서 $0 \le p \le 1$ 는 영의 값에서 주어지는 임의의 확률이며 $\lambda > 0$는 포아송분포의 평균이다. 이때 확률질량함수(pmf)는 아래와 같이 된다.

$$P(Y=k) = p+(1-p)e^{-\lambda}, \quad k=0$$

$$= (1-p)\ \lambda^k e^{-\lambda}/k!,\ k=1,2,\cdots.$$

반응변수가 영과잉-포아송분포를 따르고 몇 개의 공변량들에 의해 의존된다고 생각해보자. 반응변수의 평균 λ와 영에 대한 확률 p는 공변량들의 회귀모형을 이룬다. 영의 값만 관측되는 경우의 확률 p는 로짓연결함수(logit link function)를 그리고 포아송분포의 평균 λ는 로그연결함수를 이용하였다. 즉, $\log(\lambda)$와 $\log(p/(1-p))$ 가 공변량들의 선형함수로 표현되는 영과잉-포아송 회귀모형을 생각한다.

반응벡터 $\boldsymbol{Y}=(Y_1, Y_2, \cdots, Y_n)'$ 들이 독립이고 그리고

$$Y_i \sim 0, \quad p_i \text{ 의 확률로,}$$

$$\sim Poisson(\lambda_i),\ 1-p_i \text{의 확률}$$

의 영과잉-포아송분포를 따른다고 하자. 이때 모수벡터 $\boldsymbol{\lambda}=(\lambda_1, \lambda_2, \cdots, \lambda_n)'$ 그리고 영에대한 확률벡터 $\boldsymbol{p}=(p_1, \cdots, p_n)'$는 다음을 만족한다.

$$\log(\boldsymbol{\lambda}) = \boldsymbol{B}\,\boldsymbol{\beta}, \qquad \log\left(\frac{\boldsymbol{p}}{1-\boldsymbol{p}}\right) = \boldsymbol{G}\,\boldsymbol{\gamma}, \tag{2.1}$$

여기에서 $\boldsymbol{B}$ 와 $\boldsymbol{G}$는 공변량들의 모형행렬(model matrix)이다. 위 모형은 많은 모수들을 포함하기 때문에 다음과 같이 몇 가지로 분류하여 생각할 수 있다.

첫째, $\boldsymbol{\lambda}$와 $\boldsymbol{p}$가 함수적인 관계가 없고 두 공변량이 같을 때 즉, $\boldsymbol{B}$ 와 $\boldsymbol{G}$가 같을 때이다. 이때는 모형에 포함되는 모수의 수가 포아송 회귀모형보다 두배 많게 된다. 둘째는 $\boldsymbol{p}$가 공변량에 의존되지 않을 때이다. 이때는 $\boldsymbol{G}$가 원소가 1인 벡터가 되어서 포아송 회귀모형보다 모수가 단 한 개 많게 된다. 마지막으로는 같은 공변량 즉, $\boldsymbol{B}$ 와 $\boldsymbol{G}$가 같을 때 그리고 $\boldsymbol{\lambda}$와 $\boldsymbol{p}$가 함수관계가 있을 때이다. 이때는 모수의 수가 많이 감소하게 된다. 일반적으로 0만이 완전하게 나타날(perfect state) 확률 $\boldsymbol{p}$는 포아송 평균인 $\boldsymbol{\lambda}$에 반비례한다. $\boldsymbol{\lambda}$가 커질수록 $\boldsymbol{p}$가 작아지는 것은 함수는 여러 가지로 생각할 수 있으나, 두 모수에 대한 사전정보가 $p_i = 1/(1+\lambda_i^{\tau})$ 같이 알려져 있다면, (2.1)식에서

$$\frac{\log(\boldsymbol{p}/(1-\boldsymbol{p}))}{\log\ \boldsymbol{\lambda}} = \frac{\boldsymbol{B\gamma}}{\boldsymbol{B\beta}} = -\tau$$

가 되므로 $\boldsymbol{B\gamma} = -\tau\,\boldsymbol{B\beta}$ 가 된다. 여기에서 τ는 형태모수(shape parameter)로서 이 값이 커지면 $\boldsymbol{p}$는 기하급수적으로 감소하게 된다. 위 (2.1)식 영과잉 포아송 회귀모형은 다음과 같이 된다.

$$\log(\boldsymbol{\lambda}) = \boldsymbol{B}\,\boldsymbol{\beta}, \qquad \log\left(\frac{\boldsymbol{p}}{1-\boldsymbol{p}}\right) = -\tau\,\boldsymbol{B}\ \ \boldsymbol{\beta}. \tag{2.2}$$

위 모형은 일반화 선형모형을 만들기 위하여 포아송 평균의 로그 연결함수 그리고

베르누이분포의 성공의 확률에 대한 로짓연결함수가 이용되었다. 한편 영과잉-포아송 회귀모형은 많은 회귀계수들을 포함하고 있어 회귀계수들을 추정하는데 많은 어려움이 있다. 따라서 본 논문에서는 모수들의 수가 비교적 적은 (2.2)식의 모형을 다루려고 한다.

회귀계수 벡터 $\boldsymbol{\beta}$와 형태모수 τ에 대한 로그-우도함수는

$$\begin{aligned} L(\boldsymbol{\beta}, \tau, \boldsymbol{y}) &= \sum_{y_i=0} \log(e^{-\tau \boldsymbol{B}_i\boldsymbol{\beta}} + \exp(-e^{\boldsymbol{B}_i\boldsymbol{\beta}})) \\ &\quad + \sum_{y_i>0}(y_i \boldsymbol{B}_i\boldsymbol{\beta} - e^{\boldsymbol{B}_i\boldsymbol{\beta}}) - \sum_{i=1}^{n} \log(1+e^{-\tau \boldsymbol{B}_i\boldsymbol{\beta}}) \end{aligned} \tag{2.3}$$

와 같이 된다. 또한 최우추정량 $\hat{\boldsymbol{\beta}}, \hat{\tau}$을 이용한 적합도를 알아볼 수 있는 이탈도(deviance)는 $2[(L(\boldsymbol{y}; \boldsymbol{y}) - L(\hat{\boldsymbol{\beta}}, \hat{\tau}; \boldsymbol{y})]$으로 이는 점근적으로 χ^2_p의 분포를 따르는 것으로 알려져 있다. 여기에서 p는 미지의 모수 개수이다.

3. 모의실험

영과잉-포아송 회귀모형을 따르는 공변량과 반응변수를 얻기위하여 모의실험을 하였다. 가장 단순한 영과잉-포아송 회귀모형을 생각하기 위하여 공변량은 한 개로 설정하고 표본크기는 $n=10$을 생각하였다. 공변량이 한 개일때의 (2.2)식 모형은 다음과 같이 된다.

$$\begin{aligned} &\log(\lambda_i) = \beta_0 + \beta_1 x_i, \\ &\log\left(\frac{p_i}{1-p_i}\right) = -\tau(\beta_0 + \beta_1 x_i), i = 1, 2, \cdots, 10. \end{aligned} \tag{3.1}$$

공변량 x_i는 $(0,1)$ 균일분포(uniform distribution)를 따르는 난수들로 구성되어 있고, 회귀계수벡터는 $\boldsymbol{\beta} = (\beta_0, \beta_1)' = (-1, 1)'$ 그리고 $\tau = 1$로 설정하였다. 그러면 (3.1)식은

$$\lambda_i = e^{x_i - 1}, \quad p_i = e^{1-x_i}/(1+e^{1-x_i}), \; i = 1, 2, \cdots, 10. \tag{3.2}$$

가 된다. (3.2)식에 공변량 값을 대입하면 영과잉-포아송 분포를 따르는 반응변수 y_i의 평균 λ_i와 확률 p_i를 얻을 수 있다. 반응변수 y_i는 또 다른 $(0,1)$ 균일변수 U_i를 이용하여 $U_i < p_i$ 이면 $y_i = 0$, 그렇지 않으면 $y_i \sim Poisson(\lambda_i)$이 되도록 난수를 발생한다. 반응값 y_i가 얻어지면 주어진 공변량 x_i를 영과잉-포아송 회귀모형 (3.1)식을 이용하여 회귀계수 β_0, β_1과 형태모수 τ를 추정한다. 추정하는 방법은 최우추정법이다. (2.3)식의 우도함수의 최대값을 구하기 위하여 Press 등(1992)의 부프로그램 Powell을 이용하였다. 이 방법은 미분없이 최대, 최소값을 찾는 방법이다. 추정된 회귀계수 최우추정치는 $(\hat{\beta}_0, \hat{\beta}_1) = (-1.26, 0.74)$ 이고 형태모수 $\hat{\tau} = -0.26$ 으로 나타났다. 회귀계수 추정값은 실제값에 근접하지만 형태모수 τ는 차이가 많이 남을 알 수 있다.

추정된 회귀계수와 형태모수를 이용하여 모형의 예측값을 구하여 보기로 한다. 추

정된 모수값을 (3.1)식에 대입하여 $\hat{\lambda}_i$ 그리고 $\hat{p}_i$을 구할 수 있고, 다시 이를 영과잉-포아송분포에 적용하여 적합된 예측값 $\hat{y}_i$을 얻을 수 있다. 공변량값 x_i와 반응값 y_i 그리고 예측값 $\hat{y}_i$가 <표3.1>에 나타나 있다. <표3.1>에서 'Poisson'으로 표기된 부분은 영과잉이 아니라 포아송 회귀모형에 적용했을 때의 실제값과 예측값이다.

한편 (2.3)식의 로그-우도함수 값은 $L(\hat{\beta}_0, \hat{\beta}_1, \hat{\tau}\ ; \boldsymbol{y}) = -11.561$으로 계산되었고 이때의 이탈도는 0.031로 나타났다. 표본크기가 10이라 크지는 않지만, 이탈도가 점근적으로 χ^2_3의 분포를 따르기 때문에 점근적 이론을 적용하면, 귀무가설 $H_0 : \beta_0 = \hat{\beta}_0\ , \beta_1 = \hat{\beta}_1,\ \tau = \hat{\tau}$ 에서의 유의확률은 0.998로서 귀무가설을 채택하게 된다.

<표3.1> 모의실험을 통하여 얻어진 공변량과 반응값 그리고 추정된 예측값

n_i	x_i (공변량)	U_i	λ_i	p_i	Poisson	y_i (반응값)	$\hat{\lambda}_i$	$\hat{p}_i$	Poisson (예측값)	$\hat{y}_i$
1	.726	.322	.760	.568	1	0	.485	.453	0	0
2	.562	.207	.645	.608	1	0	.430	.445	1	0
3	.151	.745	.428	.700	0	0	.317	.426	0	0
4	.632	.449	.692	.591	2	0	.453	.449	0	0
5	.986	.616	.986	.503	0	0	.589	.466	1	1
6	.558	.015	.643	.609	0	0	.429	.445	0	0
7	.587	.872	.662	.602	2	2	.438	.447	0	0
8	.786	.870	.807	.553	0	0	.508	.456	0	0
9	.195	.415	.447	.691	0	0	.328	.428	0	0
10	.106	.173	.409	.710	0	0	.307	.424	1	0
평균	0.58	0.50	0.67	0.60	0.67	0.22	0.44	0.45	0.22	0.11

<표3.1>을 살펴보면 반응값 y_i는 0이 아닌 경우가 단 한 개 나타나므로 90%가 영과잉으로 되어있다. 영과잉-포아송 회귀모형에 적합해 본 결과 7번째 관측치에서는 0으로 예측되지는 않았지만 90%의 영과잉으로 나타나게 되어 잘 적합함을 알 수 있다.

4. 사례연구

영과잉-포아송 회귀모형의 사례연구를 위하여 이용된 자료는 미국의 어떤 백화점에서 만명의 고객을 대상으로 지난 18개월 동안 고객들이 구입한 상품구입 회수 및 고객들에 대한 정보이다. 이는 강현철 등(1999)에 첨부된 파일 중 'BUYTEST.SD2' 데이터이다. 이 데이터는 총 26개의 변수와 관측치 수는 10,000개이다. 많은 변수들 중에서 본 연구에 도움이 되는 몇 개의 변수만 활용하였다. 변수명과 변수내용이 다음 <표4.1>에 나타나 있다. 이들 변수 중에서 관심이 있는 반응변수를 BUY18로 택하였다. 이 변수는 최근 18개월 간에 60$ 이상 상품을 구입한 횟수이다. 이 변수를 반응변

수로 택한 이유는 백화점 입장에서 보면 백화점의 매출은 고객의 상품구입 횟수에 직결 되기 때문이다. 또한 영과잉-포아송 회귀모형에서 반응변수는 단위당 나타나는 사건의 수이다. 특히 사건의 수가 0이 많이 나타나는 경우가 영과잉-포아송분포에 잘 적합된다. 고객이 최근 18개월(단위당) 안에 구입한 횟수는 0을 많이 포함하고 있는 변수로 이를 반응변수로 설정하였다. 반응변수에 영향을 미치는 공변량은 양적변수 2개 즉, 나이(AGE), 수입(INCOME)와 질적변수 3개, 결혼여부(MARRIED), 성별(SEX) 그리고 집소유 여부(OWNHOME)로 총 5개로 구성되어 있다. 이를 영과잉-포아송 회귀모형으로 생각한다면 다음 모형식이 된다.

$$(BUY18)_i \sim 0,\ p_i \text{ 의 확률로}$$

$$\sim Poisson(\lambda_i),\ 1-p_i \text{의 확률},$$

$$\log(\lambda_i)=\beta_0+\beta_1*AGE_i+\beta_2*INCOME_i+\beta_3*SEX_i+\beta_4*MARRIED_i+\beta_5*OWNHOME_i\ ,$$

$$\log\left(\frac{p_i}{1-p_i}\right)=-\tau*\log(\lambda_i),\ \ i=1,2,\cdots,100.$$

프로그램 수행을 원활하게 하기 위하여 10,000개의 관측치 중에서 결측치를 제외한 9766개에서 100개의 관측치를 단순임의표집하였다. 표집과정은 SAS 매크로 프로그램 'SRS'를 이용하였다. 표집된 100개의 표본들에서 반응변수를 살펴보면 구입회수(BUY18)가 0인 경우가 67%, 1번 구입이 31% 그리고 2번 이상 구입이 2%로 나타났다. 반응변수가 영과잉 관측된 자료로 볼 수 있다. 표집된 표본들의 반응변수를 질적변수별로 보면 다음 <표4.2>와 같다.

흥미있는 점은 결혼한 여자인 경우 집을 소유했건 안했건 백화점에서 상품을 구입한 회수가 제일 많았다. 그리고 미혼 무주택 남자인 경우는 구입횟수가 전혀 없는 것으로 나타났다. 주어진 자료를 한번은 영과잉-포아송 회귀모형에 다른 한번은 포아송 회귀에 적합시켰다. 두 모형의 적합성을 알아보기 위하여 우도비를 계산하고 회귀계수와 형태모수를 구하였다. 그 결과가 <표4.3>에 나타나 있다.

일반적으로 영과잉-포아송 회귀모형에서 양적변수가 아닌 질적변수들의 추정된 회귀계수의 의미를 해석하기는 용이하지 않다. 그리고 공변량들의 교호작용들도 생각할 수 있겠으나 회귀계수들이 너무 많아지기 때문에 최우추정치를 찾기가 쉽지 않으므로 생략하기로 한다.

<표4.1> 사례연구에 이용된 자료의 변수설명

변수명	변수 내용
AGE	나이(년)
INCOME	년 수입(단위: 천달러)
MARRIED	1:결혼 0:미혼
SEX	F:여자 M:남자
OWNHOME	집 소유 여부(1:소유, 0:미소유)
BUY18	최근 18개월 간의 상품구입 횟수

<표4.2> 질적변수별 상품구입 회수(반응변수)의 평균

성별 \ 결혼	미혼		기혼	
집	소유	미소유	소유	미소유
남	0.33	0.00	0.26	0.38
여	0.10	0.40	0.50	0.50

영과잉-포아송 회귀에 적합시켰을 때의 로그-우도비 값은 -309.263으로 나타났고, 포아송 회귀에 적합시켰을 때는 -614.534로 나타났다. 서로 다른 모형의 우도비를 비교하는 것은 의미가 없으나 포아송-회귀모형의 우도비가 크게 나타났다. 한편 적합된 영과잉-포아송 회귀모형을 보면, 나이와 수입이 많아질 수 록 반응변수 BUY18의 평균 즉, $\hat{\lambda}$는 증가함을 알 수 있다. 반면 포아송 회귀모형은 나이가 증가하면 BUY18의 평균, $\hat{\lambda}$이 감소하나 수입은 반대현상을 나타낸다.

<표4.3> 영과잉-포아송회귀와 포아송 회귀에서 추정된 모수값

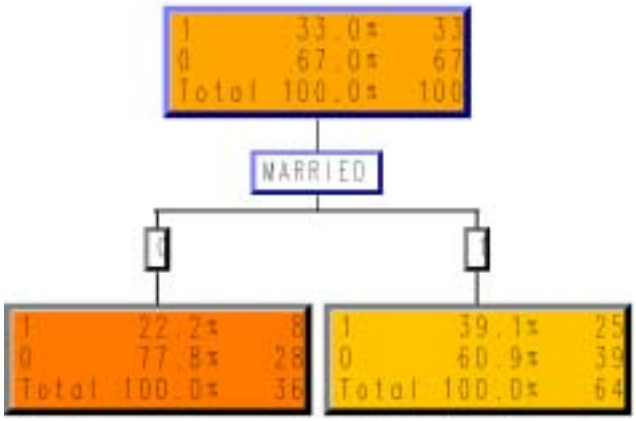
모형	공변량과 모수	추정된 모수값	로그-우도함수
영과잉-포아송 회귀	절편	-1.20	-309.263
	AGE	3.02	
	INCOME	3.03	
	SEX	1.69	
	MARRIED	4.19	
	OWNHOME	0.60	
	τ	1.20	
포아송 회귀	절편	-0.53	-614.534
	AGE	-1.28	
	INCOME	1.03	
	SEX	-3.26	
	MARRIED	0.34	
	OWNHOME	1.25	
	τ	*	

회귀분석의 주된 목적이 예측에 있기 때문에, 본 연구에서 제시된 영과잉-포아송 회귀모형의 예측을 알아보기로 한다. <표4.3>의 추정된 모수값과 주어진 공변량들의

값을 이용해서 반응변수의 예측치를 구하여 보았다. 반응변수의 예측치 중 82%가 0, 13%가 1 그리고 5%가 2번이상으로 나타났다. 실제자료(표본수=100)에서는 반응변수 값이 0인 경우가 67%, 관측치 만개의 모집단에서는 70%로 0이 과대적합 됐음을 알 수 있다. 포아송 회귀모형에 적합시켰을 때의 예측치는 0인 경우가 37% 1인 경우가 58% 그리고 2 이상인 경우가 5%로 영과소 적합됨을 알 수 있다. 두 모형을 비교해 본다면 실제자료에 더 잘 적합된 것은 영과잉-포아송 모형이라 할 수 있겠다.

다음으로 사례연구에 이용된 자료를 의사결정나무(decision tree)방법으로 분석해 보고자 한다. 의사결정나무분석은 반응변수에 영향을 주는 공변량들을 의사결정규칙에 의하여 나무구조로 도표화하여 분류, 예측을 수행하는 방법이다. 'SAS의 E-miner'를

이용하여 목표변수(target)는 BUY18 그리고 입력변수(input)들은 자료의 공변량들로 설정하였다. BUY18는 1 이상인 경우를 1로 변환하여 이진변수로 변환하고 분석하였다. 그 이유는 의사결정나무분석은 목표변수가 양적변수일 경우는 예측하기 힘든 단점이 있기 때문이다. 의사결정나무 분석에서 모든 옵션은 디폴트로 처리하였고 그 결과가 <그림4.1>에 나와 있다.



<그림4.1> 의사결정나무분석에 의한 나무구조모형

<그림4.1>에서 반응변수에 영향을 주는 공변량은 MARRID(결혼여부) 한 개로 나타났다. 기혼인 경우 구입회수가 0인 경우가 약 61%, 1번 이상은 39%로 나타난다. 기혼인 경우가 미혼일 때보다도 구입회수는 약 17%로 정도 더 많다. 그러나 선택된 한 개의 공변량으로 반응변수를 예측하기엔 어려울 것으로 판단된다.

5. 결론

백화점에서 고객들의 관리차원을 생각한다면, 반응변수의 예측값을 크게하는 공변량들의 수준을 생각해야 할 것이다. 영과잉-포아송 모형에서는 고객들의 상품구입 회수가 많을 때의 공변량 수준은 주효과 별로 연령은 40대, 수입은 2만 달러, 성별은 남자, 기혼 그리고 무주택자로 나타났다. 포아송 회귀모형에서는 40대, 4만 달러, 여자, 기혼, 무주택자가 최적의 수준이다. 나무구조모형에서는 반응변수에 영향을 미치는 공변량이 결혼여부(MARRIED) 한 개로 선택되어 졌다. 기혼인 경우가 미혼일 때보다도 구입회수가 많게 된다.

자료의 적합도를 세 모형으로 비교한다면, 나무구조모형은 반응변수가 양적변수라서 이를 이진변수로 변환하여 분석해 보았지만, 결과가 예측하기 어려운 모형으로 나타났다. 영과잉-포아송 회귀모형은 포아송 회귀모형보다 잘 적합되므로 추정된 영과잉-포아송 회귀모형을 이용하는 것이 바람직할 것으로 판단된다. 그러나 추정된 영과잉-포아송 회귀모형식을 해석하기엔 어려움이 따르게 되기 때문에 예측모형으로만 이용되어져야 할 것이다.

참고문헌

1. 강현철 외 4인, (1999). "SAS Enterprise Miner을 위한 데이터마이닝."
2. Cohen, A.C., (1963). Estimation in Mixtures of Discrete distributions, *Proceedings of the International Symposium on discrete Distributions*, Montreal, pp373-378.
3. Diane Lambert, (1992). Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing, *Technometrics*, 34, pp1-14.
4. Farewll, V.T., (1986). Mixture Models in Survival Analysis: Are They Worth the Risk?, *Canadian Journal of Statistics*,14, pp257-262.
5. Heilborn, D.C., (1989). Generalized Linear Models for Altered Zero Probabilities and Overdispersion in Count Data, *unpublished technical report, University of California*, San Francisco, Dept. of Epidemiology and Biostatistics.
6. Johnson, N.L., Kotz, S., (1969). *Distribtions in Statistics: Discrete Distributions*, Boston: Houghton Mifflin.
7. 'Li, C.H, Lu, J.C., Park, J.H., Kim, K.M, Brinkly, P.A., Peterson, J.P., (1999). Multivariate Zero-Inflated Poisson Models and Their Applications, *Technometrics*, 41, 1, pp.29-38.
8. Powell, M.J.D., (1964). An Efficient Method for Finding the Minimum of a Function of Several Variables without Calculating Dervatives, *Computer Journal*, 7, pp155-162.
9. 'Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., (1992). *Numerical Recipes in Fortran*, Cambridge.
10. Singh, S.N., (1963). A Note on Inflated Poisson Distribution, *Journal of the Indian Statistical Association*, 1, pp140-144.

[2002년 12월 접수, 2003년 2월 채택]