*Journal of Korean*
*Data & Information Science Society*
*2003, Vol. 14, No.2 pp. 177~186*

# An application to Multivariate 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 correalated reponse variables are intrested, we have to extend the univariate zero-inflated regression model to multivariate model. In this paper, we study and simulate the multivariate zero-inflated regression model. A real example was applied to this model. Regression parameters are estimated by using MLE's. We also compare the fitness of multivariate zero-inflated Poisson regression model with the decision tree model.

**주제어**: 영과잉-포아송 분포, 다변량 영과잉-포아송 회귀모형, 의사결정나무모형

## 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 반응값이 영인 경우에 확률을 임의로 주는 모형을 생각했다. 영이 되는 확률이 표준 포

1) The present research was supported by the research fund of Daegu University in 2000.
2) Professor, Department of Statistics, Daegu University, Daegu, Korea.
E-mail: kmkim@daegu.ac.kr

아송분포보다 적게 되게되도록 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 distribution)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 다변량 영과잉-포아송 회귀모형의 적용사례로서 백화점 고객들의 상품구입회수에 관한 실제자료 예를 들어 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. 이는 많은 공변량들에 의해 종속된다고 생각할 수 있다. 본 논문에서 다루고 있는 공변량들로는 고객들의 연령, 성별 그리고 결혼여부이다. 이 자료를 관찰해보면 반응변수가 과잉으로 영이 관측되는 것을 알 수 있다. 반응변수가 계수형자료(count data)이므로 일반적인 다중회귀모형에 적합시키기는 어렵다. 또한 포아송 회귀모형에 적합시키는것도 반응값이 영이 과잉으로 관측되기 때문에 적용하기 힘들 것이다. 본 논문은 다변량 영과잉-포아송 회귀모형을 소개하고 실제자료를 이용하여 회귀계수들을 추정하고 모형의 적합성을 의사결정나무모형과 비교하여 알아보려고 한다.

## 2. 다변량 영과잉 포아송 회귀모형

다변량 영과잉-포아송 회귀모형을 소개하기 위하여 일변량 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. 먼저 일변량 영과잉-포아송분포(Zero-inflated Poisson Distribution, 이후 ZIP로 표기함)는 다음과 같이 정의 된다.

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

$$\begin{aligned} Y &\sim 0, && p \text{ 의 확률로} \\ &\sim Poisson(\lambda), && 1-p \text{의 확률,} \end{aligned}$$

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

$$\begin{aligned} P(Y=k) &= p+(1-p)e^{-\lambda}, && k=0 \\ &= (1-p)\,\lambda^k e^{-\lambda}/k!, && k=1,2,\cdots. \end{aligned}$$

Lambert(1992)가 제시한 일변량 ZIP분포를 이용한 ZIP 회귀모형을 다변량 ZIP(이후 MZIP로 표기함) 회귀모형으로 확장시키기 위하여, 본 논문은 Kim 등(1999)이 제시한 MZIP분포를 이용하기로 한다.

관측자료수가 $n$인 $m$차원 확률변수 $(Y_1, Y_2, \cdots, Y_m)$는 서로 종속관계가 있는 반응변수이고 다음과 같은 $m$차원 MZIP분포를 따른다 하자.

$$\begin{array}{ll}(Y_1, Y_2, \cdots, Y_m) & \\ \sim (0,0,\cdots,0), & p_0\text{의 확률로}, \\ \sim (Poisson(\lambda_1),0,\cdots,0), & p_1\text{의 확률로}, \\ \sim (0, Poisson(\lambda_2),0,\cdots,0), & p_2\text{의 확률로}, \\ \cdots & \\ \sim (0,0,\cdots, Poisson(\lambda_m)), & p_m\text{의 확률로}, \\ \sim \text{다변량} Poisson(\lambda_{10},\lambda_{20},\cdots,\lambda_{m0},\lambda_{00}), & p_{11}\text{의 확률로}, \end{array}$$

여기에서 $p_0+p_1+\cdots+p_m+p_{11}=1$ 그리고 $\lambda_{10},\lambda_{20},\cdots,\lambda_{m0}$ 들은 각각 $Y_1, Y_2, \cdots, Y_m$들이 일변량 포아송분포를 따를 때 이의 평균들을 의미한다. 또한 $\lambda_j \equiv \lambda_{j0}+\lambda_{00}, j=1,2,\cdots,m$ 이다. 위 $m$차원 $MZIP$분포가 포함하고 있는 모수들은 $2(m+1)$개이다. 왜냐하면 포아송 평균 모수들은 $\lambda_{10},\lambda_{20},\cdots,\lambda_{m0},\lambda_{00}$들로 총 개수는 $(m+1)$ 개이고 각 구간에 속할 확률 $p_{0,}p_{1,\cdots,}p_m$ 는 $(m+1)$ 개 이다. 그러므로 총 모수의 수는 $2(m+1)$이다. 이러한 $MZIP$분포의 pmf는 많은 모수를 포함하고 있는 매우 복잡한 형태를 지니게 되고 그 pmf는 다음과 같다.

$$\begin{aligned} &P(Y_1=0,\cdots,Y_m=0)=p_0+p_1e^{-\lambda_1}+\cdots+p_me^{-\lambda_m}+p_{11}e^{-\lambda}, \\ &P(Y_1=y_1,Y_2=0,\cdots,Y_m=0)=[p_1\lambda_1^{y_1}e^{-\lambda_1}+p_{11}\lambda_{10}^{y_1}e^{-\lambda}]/y_1!, \\ &P(Y_1=0,Y_2=y_2,Y_3=0,\cdots,Y_m=0)=[p_2\lambda_2^{y_2}e^{-\lambda_2}+p_{11}\lambda_{20}^{y_2}e^{-\lambda}]/y_2!, \\ &\cdots \\ &P(Y_1=0,\cdots,Y_{m-1}=0,Y_m=y_m)=[p_m\lambda_m^{y_m}e^{-\lambda_m}+p_{11}\lambda_{m0}^{y_m}e^{-\lambda}]/y_m!, \\ &P(Y_1=y_1,Y_2=y_2,\cdots,Y_m=y_m)=p_{11}\sum_{j=0}^{\min(y_1,y_2,\cdots,y_m)}\frac{\lambda_{10}^{y_1-j}\lambda_{20}^{y_2-j}\cdots\lambda_{m0}^{y_m-j}\lambda_{00}^{j}e^{-\lambda}}{(y_1-j)!(y_2-j)!\cdots(y_m-j)!j!}, \end{aligned} \tag{2.1}$$

여기에서 $\lambda \equiv \sum_{i=0}^{m}\lambda_{i0}$이다. 이러한 분포를 $MZIP(\lambda_{10},\lambda_{20},\cdots,\lambda_{m0},\lambda_{00}\ ;p_0,p_1,\cdots,p_m)$ 이라고 표시하기로 한다.

반응변수들이 MZIP분포를 따르고 몇 개의 공변량들에 의해 의존된다고 생각해보자. 반응변수의 평균과 영에 대한 확률은 공변량들의 회귀모형을 이룬다. 영의 값만 관측되는 경우의 확률은 로짓연결함수(logit link function)와 유사한 형태를 그리고 포아송분포의 평균은 로그연결함수를 이용하였다. 즉, $\log(\lambda)$와 $\log(p/(1-p))$ 가 공변량들의 선형함수로 표현되는 영과잉-포아송 회귀모형을 생각한다.

$n$개의 관측치를 갖는 $m$차원 반응변수들이 다변량 영과잉-포아송분포를 따를 때 다음과 같이 표시하기로 하자.

$$(Y_{i1}, Y_{i2}, \cdots, Y_{im}) \sim MZIP(\lambda_{i10},\lambda_{i20,\cdots,}\lambda_{im0},\lambda_{i00}\ ;p_{i0,}p_{i1,\cdots,}p_{im}), i=1,2,\cdots,n$$

이때 MZIP의 평균벡터 $\boldsymbol{\lambda}=(\boldsymbol{\lambda}_{10}, \boldsymbol{\lambda}_{20},\cdots, \boldsymbol{\lambda}_{m0}, \boldsymbol{\lambda}_{00})$ 그리고 영에 대한 확률벡터 $\boldsymbol{p}=(\boldsymbol{p}_0, \boldsymbol{p}_1,\cdots, \boldsymbol{p}_m, \boldsymbol{p}_{11})$는 다음 (2.2)식을 만족한다.

$$\begin{aligned} &\log(\boldsymbol{\lambda}_{h0})=\boldsymbol{B}\ \boldsymbol{\beta}_h, \\ &\boldsymbol{p}_h=\frac{\exp(\boldsymbol{G}\ \boldsymbol{\gamma}_h)}{1+\sum_{h=0}^{m}\exp(\boldsymbol{G}\ \boldsymbol{\gamma}_h)}, h=0,1,\cdots,m. \end{aligned} \tag{2.2}$$

여기에서 $\boldsymbol{B}$ 와 $\boldsymbol{G}$는 공변량들의 모형행렬이고 $\boldsymbol{\beta}_h$와 $\boldsymbol{\gamma}_h$는 회귀계수 벡터이다. 또한 $p_{i11}=1-p_{i0}-p_{i1}-\cdots-p_{im}$를 만족한다. *MZIP*에서는 $2(m+1)$개의 많은 모수들을 포함하고 있었는데, (2.2)식 모형의 모수 개수는 관측치수 $n$배 만큼 더 증가하게 되고 여기에 회귀계수 수만큼 더 추가된다. 즉, $2n(m+1)$개의 모수 개수에다 공변량들의 수와 수준에 따른 모수 수가 증가된다. 이러한 많은 모수들을 포함하는 회귀모형을 분석하기에는 많은 어려움이 따르게 된다. (2.2)식을 각각의 $i$번째 관측치별로 생각하면,

$$\lambda_{ih0}=e^{\boldsymbol{B}_i\boldsymbol{\beta}_h},$$
$$p_{ih}=\frac{e^{\boldsymbol{G}_i\boldsymbol{\gamma}_h}}{1+\sum_{h=0}^{m}e^{\boldsymbol{G}_i\boldsymbol{\gamma}_h}},\quad h=0,1,\cdots,m;\ i=1,2,\cdots,n. \tag{2.3}$$

(2.3)식에서 $\boldsymbol{B}_i$, $\boldsymbol{G}_i$는 각각 공변량 행렬에서 $i$번째 행을 의미하고, $\boldsymbol{\beta}_h$, $\boldsymbol{\gamma}_h$는 $h$차원 MZIP 회귀모형에서 회귀계수벡터를 의미한다. 또한 $\lambda_{ih}=\lambda_{ih0}+\lambda_{i00}$ 그리고 $p_{i11}=1-p_{i1}-p_{i2}-\cdots-p_{im}$이다. (2.4)식의 값을 MZIP 확률분포 (2.1)식에 대입하면, 회귀계수 벡터 $\boldsymbol{\beta}_h$와 $\boldsymbol{\gamma}_h$들의 로그-우도함수를 얻을 수 있다. 로그-우도함수 $L(\boldsymbol{\beta}_h, \boldsymbol{\gamma}_h; \boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_m)$이 매우 복잡한 형태를 지니게 되는데, $i$번째 반응변수 값 $(y_{i1}, y_{i2}, \cdots, y_{im})$ 모두가 0일 때, 단 한 개만 0이 아닐 때 그리고 두 개 이상이 0이 아닐 때로 나누어 진다.

$$\begin{aligned}
L&=\log\Pi_{i=1}^{n}P(Y_{i1}=y_{i1}, Y_{i2}=y_{i2}, \cdots, Y_{im}=y_{im})\\
&=\sum_{(\text{모든 } y_{ik}=0 \text{ 일때}, 1\le k\le m)}\log(p_{i0}+p_{i1}e^{-\lambda_{i1}}+\cdots+p_{im}e^{-\lambda_{im}}+p_{i11}e^{-\lambda_i})\\
&+\sum_{(\text{단 한개만 } y_{ik}>0 \text{ 일때}, 1\le k\le m)}\log[(p_{ik}\lambda_{ik}^{y_{ik}}e^{-\lambda_{ik}}+p_{i11}\lambda_{ik0}^{y_{ik}}e^{-\lambda_i})/y_{ik}!]\\
&+\sum_{(\text{두개이상 } y_{ij}>0, y_{ik}>0 \text{ 일때}, 1\le j\ne k\le m)}\log[p_{i11}\sum_{j=0}^{\min(y_{i1},\cdots,y_{im})}\frac{\lambda_{i10}^{y_{i1}-j}\lambda_{i20}^{y_{i2}-j}\cdots\lambda_{im0}^{y_{im}-j}\lambda_{i00}^{j}e^{-\lambda_i}}{(y_{i1}-j)!(y_{i2}-j)!\cdots(y_{im}-j)!j!}],
\end{aligned}$$

여기에서 모든 모수들은 (2.3)식을 만족하는 $\boldsymbol{\beta}_h$, $\boldsymbol{\gamma}_h$의 함수값으로 대입되어야 한다. 또한 최우추정량 $\hat{\boldsymbol{\beta}}_h$, $\hat{\tau}_h$을 이용한 적합도를 알아볼 수 있는 이탈도(deviance)는 점근적으로 $\chi^2_p$의 분포를 따르는 것으로 알려져 있다(참고문헌 6, 234쪽). 여기에서 $p$는 미지의 모수 개수이다.

## 3. 모의실험

MZIP 회귀모형을 따르는 공변량과 반응변수를 얻기 위하여 모의실험을 하였다. 가장 단순한 MZIP 회귀모형을 생각하기 위하여, 반응변수는 차원이 $m=2$ 이고, 공변

량은 한 개 그리고 표본크기는 $n=10$을 생각하였다. (2.3)식에서 모형행렬 $B, G$이 같고 공변량이 한 개라면 (2.3) 모형식은 다음과 같이 된다.

$$\begin{aligned}&(Y_{i1}, Y_{i2}) \sim MZIP(\lambda_{i10}, \lambda_{i20}, \lambda_{i00}; p_{i0}, p_{i1}, p_{i2}),\\ &\log(\lambda_{ih0}) = \alpha_{0h} + \alpha_{1h} x_i,\\ &\log\left(\frac{p_{ih}}{1-p_{ih}}\right) = \beta_{0h} + \beta_{1h} x_i,\ i=1,2,\cdots,10;\ h=0,1,2.\end{aligned} \tag{3.1}$$

덧붙여서 이 이외의 모수들은 다음과 같이 표현된다.

$$\lambda_i = \lambda_{i10} + \lambda_{i20} + \lambda_{i00},\ \lambda_{i1} = \lambda_{i10} + \lambda_{i00},\ \lambda_{i2} = \lambda_{i20} + \lambda_{i00},$$
$$p_{i11} = 1 - p_{i0} - p_{i1} - p_{i2}.$$

(3.1)식 모형은 평균 모수가 3개, 0에 대한 확률이 3개 그리고 회귀계수가 4개가 포함되어 있다. 따라서 관측치수 $n$, 차원 $h$ 모두를 생각한다면 $10(3+3)+3(2+2)=72$개의 많은 모수를 갖고 있다. 먼저 이변량 영과잉-포아송분포를 따르는 반응변수를 얻기 위하여 공변량과 그리고 회귀계수 값을 (3.2)식과 같이 가정하기로 하자. 공변량은 균일분포 $U_1$을 생각하였다.

$$\begin{aligned}&x_i \sim iid\ U_1(0,1), i=1,2,\cdots,10,\\ &(\alpha_{00}, \alpha_{10}) = (-1,1), (\alpha_{01}, \alpha_{11}) = (1,-1), (\alpha_{02}, \alpha_{12}) = (1,1),\\ &(\beta_{00}, \beta_{10}) = (1,-1), (\beta_{01}, \beta_{11}) = (-1,1), (\beta_{02}, \beta_{12}) = (-1,-1)\end{aligned} \tag{3.2}$$

한편 (3.1)식에서 로그함수를 없애면 (3.3)식을 얻을 수 있다.

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

| 관측치 $i$ | $x_i$ 공변량 | $U_{i2}$ | $\lambda_{i10}$ | $\lambda_{i20}$ | $\lambda_{i00}$ | $p_{i0}$ | $p_{i1}$ | $p_{i2}$ | $(y_{i1}, y_{i2})$ (반응값) | $\widehat{\lambda_{i10}}$ | $\widehat{\lambda_{i20}}$ | $\widehat{\lambda_{i00}}$ | $\widehat{p_{i0}}$ | $\widehat{p_{i1}}$ | $\widehat{p_{i2}}$ | $(\widehat{y_{i1}}, \widehat{y_{i2}})$ (예측값) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | .90 | .18 | .91 | 1.10 | 6.69 | .33 | .33 | .04 | (0,0) | .46 | 1.26 | 3.69 | .23 | .36 | .13 | (0.2) |
| 2 | .76 | .71 | .79 | 1.27 | 5.83 | .34 | .34 | .05 | (0,10) | .87 | 2.31 | 5.32 | .25 | .43 | .05 | (1,6) |
| 3 | .30 | .56 | .50 | 2.01 | 3.67 | .38 | .38 | .05 | (8,0) | .95 | 1.78 | 4.23 | .56 | .12 | .14 | (5,0) |
| 4 | .99 | .41 | .99 | 1.01 | 7.33 | .32 | .32 | .04 | (7,0) | .57 | 2.23 | 4.53 | .34 | .19 | .16 | (5,2) |
| 5 | .22 | .02 | .46 | 2.19 | 3.38 | .39 | .39 | .05 | (0,0) | .26 | 1.96 | 3.35 | .42 | .12 | .15 | (0,1) |
| 6 | .60 | .80 | .67 | 1.50 | 4.94 | .36 | .36 | .05 | (4,2) | .52 | 1.54 | 6.28 | .45 | .26 | .05 | (1,3) |
| 7 | .61 | .52 | .67 | 1.48 | 4.99 | .36 | .36 | .05 | (3,0) | .57 | 1.36 | 4.25 | .36 | .25 | .12 | (3,0) |
| 8 | .32 | .19 | .51 | 1.97 | 3.76 | .38 | .38 | .05 | (0,0) | .64 | 2.68 | 4.15 | .38 | .36 | .18 | (0,0) |
| 9 | .61 | .04 | .68 | 1.47 | 5.02 | .36 | .36 | .05 | (0,0) | .95 | 2.55 | 3.65 | .45 | .32 | .18 | (0,0) |
| 10 | .98 | .30 | .98 | 1.02 | 7.27 | .32 | .32 | .04 | (0,0) | .42 | 1.36 | 5.20 | .38 | .27 | .17 | (1,2) |

$$\begin{aligned}&\lambda_{ih0} = \exp(\alpha_{0h} + \alpha_{1h} x_i),\\ &p_{ih} = \frac{\exp(\beta_{0h} + \beta_{1h} x_i)}{1 + \exp[\sum_{h=0}^{m}(\beta_{0h} + \beta_{1h} x_i)]},\ i=1,2,\cdots,10, h=0,1,2.\end{aligned} \tag{3.3}$$

(3.3)식에 (3.2)식의 공변량 값과 모수값을 대입하면, 이변량 영과잉-포아송 분포를

따르는 반응변수 $(y_{i1}, y_{i2})$의 평균 $(\lambda_{i0}, \lambda_{i1}, \lambda_{i2})$와 확률 $(p_{i0}, p_{i1}, p_{i2})$를 얻을 수 있다. 이 값들이 얻어지면, 반응변수 $(y_{i1}, y_{i2})$는 또 다른 $(0,1)$ 균일변수 $U_2$를 이용하여,

$$
\begin{aligned}
(y_{i1}, y_{i2}) = & (0,0), \ u_{i2} < p_{i0}, \\
& \sim (Poisson(\lambda_{i1}), 0), \ p_{i0} \le u_{i2} < p_{i0} + p_{i1}, \\
& \sim (0, Poisson(\lambda_{i2})), \ p_{i0} + p_{i1} \le u_{i2} < p_{i0} + p_{i1} + p_{i2}, \\
& \sim \text{이변량 } Poisson(\lambda_{i10}, \lambda_{i20}, \lambda_{i00}), \ u_{i2} \ge p_{i0} + p_{i1} + p_{i2} (= p_{i11})
\end{aligned}
$$

이 되도록 한다. 이때 발생하는 난수들은 이변량 *ZIP* 분포를 따르게 된다. 반응값 $(y_{i1}, y_{i2})$가 얻어지면 로그-우도함수를 이용하여 회귀계수 $(\alpha_{0h}, \alpha_{1h})$와 $(\beta_{0h}, \beta_{1h})$를 추정한다. 추정하는 방법은 최우추정법이다. 로그-우도함수의 최대값을 구하기 위하여 Press(1992)의 부프로그램 'Powell'을 이용하였다. 이 방법은 미분없이 최대, 최소값을 찾는 방법이다. 추정된 회귀계수 최우추정치는 다음과 같다

$$
\begin{aligned}
&(\widehat{\alpha_{00}}, \widehat{\alpha_{10}}) = -(1.24, 2.21), (\widehat{\alpha_{01}}, \widehat{\alpha_{11}}) = 2.06, -2.56), (\widehat{\alpha_{02}}, \widehat{\alpha_{12}}) = 1.47, 2.89) \\
&(\widehat{\beta_{00}}, \widehat{\beta_{10}}) = 5.04, 0.12), (\widehat{\beta_{01}}, \widehat{\beta_{11}}) = -(3.06, 2.37), (\widehat{\beta_{02}}, \widehat{\beta_{12}}) = -(1.76, 0.08)
\end{aligned}
$$

회귀계수 추정값 $(\widehat{\alpha_{0h}}, \widehat{\alpha_{1h}})$와 $(\widehat{\beta_{0h}}, \widehat{\beta_{1h}})$는 실제값과 약간의 차이가 남을 알 수 있다. 추정된 회귀계수를 이용하여 모형의 예측값을 구하여 보기로 한다. 추정된 모수값을 (3.1)식에 대입하여 $(\widehat{\lambda_{i10}}, \widehat{\lambda_{i20}}, \widehat{\lambda_{i00}})$ 그리고 $(\widehat{p_{i0}}, \widehat{p_{i1}}, \widehat{p_{i2}})$를 구할 수 있고, 다시 이를 이변량 영과잉-포아송분포에 적용하여 적합된 예측값 $(\widehat{y_{i1}}, \widehat{y_{i2}})$을 얻을 수 있다. 공변량값 $x_i$와 반응값 $(y_{i1}, y_{i2})$ 그리고 예측값 $(\widehat{y_{i1}}, \widehat{y_{i2}})$가 <표3.1>에 나타나 있다. 실제 반응값이 $(0,0)$일때가 50%가 나타나는데 그중에서 예측값은 10번 중 2번 정확히 추정됨을 알 수 있다.

## 4. 사례연구

*MZIP* 회귀모형의 사례연구를 위하여 이용된 자료는 한 백화점의 광고인쇄물 발송(DM:Direct Mailing) 여부에 대한 자료이다. 만명의 고객을 대상으로 지난 24개월 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 이변량 반응변수를 $(BUY6, BUY18)$로 택하였다. 이 변수는 최근 6개월 그리고 18개월 간에 60$ 이상 상품을 구입한 회수이다. 이 변수를 반응변수로 택한 이유는 백화점 입장에서 보면 매출은 고객의 상품구입 회수에 직결 되기 때문이다. 또한 *MZIP* 회귀모형에서 반응변수는 단위당 나타나는 사건의 수이다. 특히 사건의 수가 0이 많이 나타나는 경우가 *ZIP*분포에 잘 적합된다. 고객이 최근 6개월 그리고 18개월(단위당) 안에 구입한 회수는 0을 많이 포함하고 있는 변수로 이를 이변량 반응

변수로 설정하였다. 두 변수의 공분산은 양의 값을 지니게 된다. 3차원 이상의 반응변수를 생각할 수 있겠으나 모수의 수를 줄이기 위하여 이변량분석을 하려고 한다.

반응변수에 영향을 미치는 공변량은 양적변수 1개 즉, 나이(AGE)와 질적변수 2개, 결혼여부(MARRIED), 성별(SEX)로 총 3개로 구성되어 있다. 이 공변량은 (2.2) 모형식에서 두 모형행렬에 이용될 것이다.

프로그램 수행을 원활하게 하기 위하여 10,000개의 관측치 중에서 결측치를 제외한 9766개에서 100개의 관측치를 단순임의표집하였다. 표집과정은 SAS 매크로 프로그램 'SRS'를 이용하였다. 표집된 100개의 표본들을 공변량별로 분류한 분할표가 <표4.2>이다. 이 표를 보면 40대 기혼이 45%로 많은 비중을 차지하고 있다. 또한 표집된 표본들 중에서 $(BUY6, BUY18) = (0,0)$인 경우가 67%를 차지하고 있다. 반응변수가 영과잉 관측된 자료로 이변량 $ZIP$ 회귀모형에 적합한 자료라 볼 수 있다. 이를 이변량 $ZIP$ 회귀모형으로 생각한다면 다음 모형식(4.1)이 된다. 모수 수가 많기 때문에 교호작용은 생각하지 않았다.

<표4.1> 사례연구에 이용된 자료의 변수설명

| 변수명 | 변수 내용 |
|---|---|
| AGE | 나이(년) |
| MARRIED | 1:결혼 0:미혼 |
| SEX | F:여자 M:남자 |
| BUY6 | 최근 6개월 간의 상품구입 횟수 |
| BUY18 | 최근 18개월 간의 상품구입 횟수 |

<표4.2> 공변량별 도수

| 공변량 | MARRIED | | | |
|---|---|---|---|---|
| | 미혼 | | 기혼 | |
| | SEX | | SEX | |
| AGE | 여 | 남 | 여 | 남 |
| 10대 | 0 | 0 | 0 | 1 |
| 20대 | 2 | 1 | 0 | 2 |
| 30대 | 9 | 5 | 2 | 0 |
| 40대 | 7 | 6 | 23 | 22 |
| 50대 | 3 | 3 | 4 | 5 |
| 60대 | 0 | 0 | 2 | 1 |
| 70대 | 0 | 0 | 1 | 1 |

$$
\begin{aligned}
&(BUY6_i, BUY18_i) \sim MZIP(\lambda_{i10}, \lambda_{i20}, \lambda_{i00}; p_{i0}, p_{i1}, p_{i2}),\\
&\log(\lambda_{ih0}) = \alpha_{0h} + \alpha_{1h} AGE_i + \alpha_{2h} SEX_i + \alpha_{3h} MARRIED_i, \qquad (4.1)\\
&p_{ih} = \frac{\beta_{0h} + \beta_{1h} AGE_i + \beta_{2h} SEX_i + \beta_{3h} MARRIED_i}{1 + \sum_{h=0}^{2} (\beta_{0h} + \beta_{1h} AGE_i + \beta_{2h} SEX_i + \beta_{3h} MARRIED_i)}, \; i = 1,2,\cdots,100; \; h = 0,1,2.
\end{aligned}
$$

(4.1) 모형식은 많은 모수들을 포함하고 있다. 관심이 있는 모수는 회귀계수가 $(4+4)\times 3 = 24$개이다. 또한 평균과 $0$에 대한 확률모수들이 $100(3+3) = 600$개다. 이러한 24개의 많은 회귀계수를 포함하고 있는 로그-우도함수 최대값을 구하기 위하여 Press 등(1992)의 부프로그램 Powell을 이용하였다. (4.1)식에서 추정된 회귀계수는 <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. 회귀분석의 주된 목적이 예측에 있기 때문에, 본 연구에서 제시된 이변량 영과잉-포아송 회귀모형의 예측을 알아보기로 한다. (4.2)식의 추정된 모수값

과 주어진 공변량들의 값을 이용해서 반응변수의 예측치를 구하여 보았다. 반응변수의 예측치 중 85%가 $(0,0)$, 4%가 $(0,1)$, 6%가 $(1,2)$ 그리고 5%가 2번 이상으로 나타났다. 실제자료(표본수=100)에서는 반응변수 값이 $(0,0)$인 경우가 67%, 관측치 만개의 모집단에서는 70%로 $(0,0)$이 과대적합 됐음을 알 수 있다. 한편 상품구입 회수에 큰 영향을 미치는 공변량 수준은 연령이 40대, 남자 기혼인 경우 그리고 여자 미혼인 경우로 나타났다.

<표4.2> 추정된 회귀계수 값

| $j$ | 0 | | 1 | | 2 | |
|---|---|---|---|---|---|---|
| $i$ | $\widehat{\alpha}_{ij}$ | $\widehat{\beta}_{ij}$ | $\widehat{\alpha}_{ij}$ | $\widehat{\beta}_{ij}$ | $\widehat{\alpha}_{ij}$ | $\widehat{\beta}_{ij}$ |
| 0 | 0.14 | -1.54 | -3.48 | -0.67 | 11.30 | 1.10 |
| 1 | 1.32 | 3.14 | 2.24 | 1.02 | -1.88 | 0.07 |
| 2 | 0.14 | -3.43 | 14.05 | -2.55 | 3.67 | 1.21 |
| 3 | 2.22 | 6.30 | -5.53 | 1.08 | 12.07 | -3.10 |

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 BUY6와 BUY18 그리고 입력변수(input)들은 자료의 공변량들로 설정하였다.

의사결정나무분석에서는 이변량 반응변수가 목표변수로 되지 못하므로 반응변수 각각을 일변량으로 처리하여 분석하였다. 의사결정나무분석에서 모든 옵션은 디폴트로 처리하였고 그 결과가 <그림4.1>에 나와 있다.

<그림4.1>에서 반응변수 BUY6은 AGE 한 개의 마디로 나무구조가 형성되었다. 나이가 39.5세 이상인 경우 구입회수가 1인 경우가 약 16.7%, 39.5세 미만이 4.5%로 나타난다. BUY18 반응변수는 MARRIED, AGE 두 개의 마디로 형성된다. 기혼인 경우가 미혼일 때보다도 한 번이상 구입회수는 약 16.9%로 정도 더 많다. 동시에 기혼인 경우 나이가 61세 미만인 경우가 이상인 경우보다 한번이상 구입횟수는 40.9% 많게 나타났다. 그러나 반응변수가 이변량이므로 의사결정나무분석으로 반응변수를 예측하기엔 어려울 것으로 판단된다.

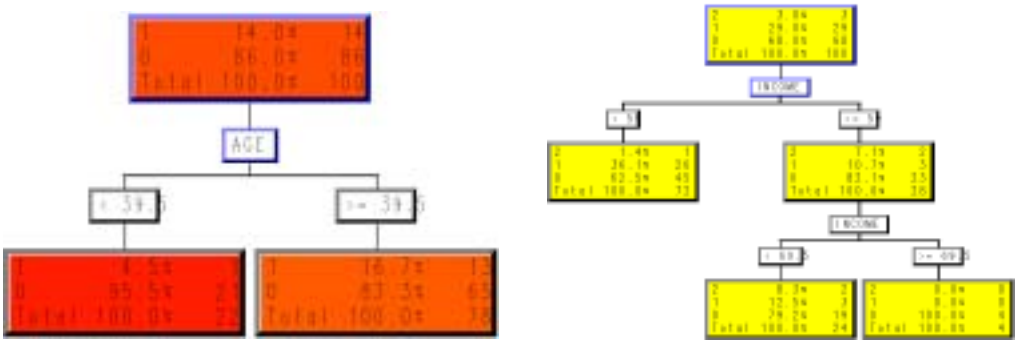
## 5. 결론

백화점에서 고객들의 관리차원을 생각한다면, 반응변수의 예측값을 크게하는 공변량들의 수준을 생각해야 할 것이다. 반응변수의 예측값을 크게하는 공변량들의 수준을 생각해야 할 것이다. 이변량 영과잉-포아송 회귀모형에서는 상품구입 회수에 큰 영향을 미치는 공변량 수준은 연령이 40대, 남자 기혼인 경우 그리고 여자 미혼인 경우로 나타났다.

나무구조모형에서는 최근 6개월 동안 상품구입회수는 39.5세 이상인 경우가 미만보다 많았다. 그리고 최근 18개월 동안 상품구입회수는 기혼인 경우가 미혼보다 그리고 기혼인 경우에도 61세 미만인 경우가 이상보다 많았다.

자료의 적합도를 두가지 모형으로 비교한다면, 나무구조모형은 반응변수가 이변량 변수인데도 불구하고 일변량으로 따로 분석해 보았지만, 결과가 예측하기 어려운 모형으로 나타났다. 따라서 이변량 영과잉-포아송 회귀모형을 이용하는 것이 바람직할 것으로 판단된다. 그러나 추정된 이변량 영과잉-포아송 회귀모형식을 해석하기엔 어려움이 따르게 되기 때문에 예측모형으로만 이용되어져야 할 것이다.

BUY6 BUY18




<그림4.1> 의사결정나무분석에 의한 나무구조모형

## 참고문헌

1. 강현철 외 4인, (1999). "SAS Enterprise Miner을 위한 데이터마이닝."
2. Lambert, D. (1992). Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing, *Technometrics*, 34, pp1-14.
3. Heilbron, 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.
4. Johnson, N.L., Kotz, S., (1969). *Distribtions in Statistics: Discrete Distributions*, Boston: Houghton Mifflin.
5. 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.
6. McCullah, P., Nelder, J.A., (1983). *Generalized Linear Models*, Chapman and Hall.
7. 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.

8. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., (1992). *Numerical Recipes in Fortran*, Cambridge.
9. Singh, S.N., (1963). A Note on Inflated Poisson Distribution, *Journal of the Indian Statistical Association*, 1, pp140-144.

[ 2003년 1월 접수, 2003년 4월 채택 ]