응용통계연구 제 11권 2호, 1998년, pp. 247-255

포아송 모형에서의 설명변수 선택문제
- 정규분포 설명변수하에서 -

박종선[1)]

요 약

일반선형모형의 하나인 포아송모형에서 설명변수들을 선택하는 문제를 고려하여 보았다. 설명변수들이 정규분포를 따르는 확률변수일 때 반응변수의 조건부 분포를 통하여 모형에 필요한 설명변수의 부분집합을 선택하는 방법을 제시하였다.

1. 서론

회귀분석에서 포아송 또는 이와 비슷한 과정에서 나타난 반응변수들은 0 또는 자연수의 값을 갖게 되는 데 일반선형모형(generalized linear model)의 하나인 포아송모형에서는 이러한 반응변수 y 에 대하여 평균이

$$\exp(\beta_0 + \beta^T x) \tag{1.1}$$

인 포아송 분포를 가정하게 된다.

본 논문에서는 포아송모형에서 설명변수가 확률변수일 때 모형에 필요한 설명변수의 부분집합을 찾는 문제를 고려하였다. 이 경우 문제는 설명변수 또는 설명변수의 선형결합이 주어졌을 때 반응변수의 조건부 분포, 즉 $F(y|x) = F(y|\beta^T x)$에 관한 문제가 되며 설명변수의 부분집합을 선택하는 문제는

$$F(y|x) \approx F(y|x_1) \tag{1.2}$$

을 만족하는 $p_1 \leq p$개의 이루어진 x 부분집합 x_1을 선택하는 문제로 귀착된다. 이러한 기준아래서 설명변수가 고정된(fixed) 값이 아닌 확률변수이며 특별히 정규분포를 따르거나 이에 유사할 때 이들 설명변수의 부분 또는 전체가 주어진 경우 반응변수의 조건부 분산을 이용하여 최적의 부분모형을 찾는 문제를 고려하였다.

2. 조건부 적률 및 계산

2.1. 일반적인 경우

일변량 반응변수 y와 p개의 설명변수 x를 가지고 있다고 가정하자. 이 때 $p_1 \leq p$개의 원소로 이루어진 x의 부분집합 x_1에 대하여 y의 조건부 분포는

1) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 조교수

$$
\begin{aligned}
f(y|x_1) &= \frac{f(y,x_1)}{f(x_1)} \\
&= \frac{\int_{X_2} f(y,x_1,x_2)dx_2}{\int_{X_2} f(x_1,x_2)dx_2} \\
&= \frac{\int_{X_2} f(y|x_1,x_2)f(x_1,x_2)dx_2}{\int_{X_2} f(x_1,x_2)dx_2}
\end{aligned} \tag{2.1}
$$

이 된다. 식 (2.1)에 포함되어 있는 분포들의 형태가 특수한 경우를 제외하면 일반적으로 계산이 불가능하다. 그러나 조건부분포 $f(y|x_1)$의 평균과 분산인 $\mathrm{E}(y|x_1)$과 $\mathrm{Var}(y|x_1)$은 설명변수가 정규분포를 따를 때 비교적 쉽게 계산이 가능하며 계산된 처음 두 적률에 의사우도함수접근법(quasi-likelihood approach; Wedderburn, 1974; McCullagh와 Nelder, 1989, 제9장)을 이용하여 여러 가지 분석을 행할 수 있다.

2.2. 정규설명변수를 가지는 포아송 모형의 경우

포아송모형의 경우 정칙연결함수(canonical link function)는 대수함수가 되며 이 경우 $f(y|x_1)$의 평균은 설명변수들이 정규분포를 따를 때

$$
\begin{aligned}
\mathrm{E}(y|x_1) &= \mathrm{E}(\mathrm{E}(y|x_1,x_2)|x_1) \\
&= \int_{X_2} \exp(\beta_0 + \beta_1^T x_1 + \beta_2^T x_2) f(x_2|x_1)dx_2 \\
&= \exp[.5\beta_2^T \mathrm{Var}(x_2|x_1)\beta_2 + \beta_2^T \mathrm{E}(x_2|x_1) + \beta_0 + \beta_1^T x_1] \\
&= \mu^*(x_1)
\end{aligned} \tag{2.2}
$$

이 되어 모수들은 다르지만 결국 x_1의 선형함수가 된다는 것에는 변화가 없다. 그리고 분산함수는

$$
\begin{aligned}
\mathrm{Var}(y|x_1) &= \mathrm{E}(y^2|x_1) - \mathrm{E}^2(y|x_1) \\
&= \mathrm{E}(\mu(x) + \mu(x)^2|x_1) - \mathrm{E}^2(\mu(x)|x_1) \\
&= \mathrm{E}(\mu(x)^2|x_1) + \mu^*(x_1)(1-\mu^*(x_1)) \\
&= \int_{X_2} \exp\{2(\beta_0 + \beta_1^T x_1 + \beta_2^T x_2)\} f(x_2|x_1)dx_1 + \mu^*(x_1)(1-\mu^*(x_1)) \\
&= \mu^*(x_1) + (D-1)(\mu^*(x_1))^2 \\
&= \mu^*(x_1) + (1/k)(\mu^*(x_1))^2
\end{aligned} \tag{2.3}
$$

이 된다. 여기서 $D = \exp(\beta_2^T \mathrm{Var}(x_2|x_1)\beta_2)$이고 $k = 1/(D-1)$이며 $\beta_2 = 0$일 때 $D = 1$이 되어 조건부 분산은 일반적인 포아송분포의 분산 $\mu^*(x_1)$이 되고 $D \neq 1$인 경우에 조건부 평균인 $\mu^*(x_1)$보다 큰 값을 갖게 되는 데 이 경우의 평균과 분산은 음이항(negative binomial)분포의 평균과 분산의 형태가 된다.

앞에서 구해진 평균과 분산에 의사우도함수기법과 Newton-Raphson방법을 적용하여 주어진 조건부모형의 모수 β를 추정할 수 있는 데 이 때 조건부 분산에 포함된 k 또한 동시에 추정되어야 한다. k를 추정하는 방법에는 다음의 두 가지 방법이 가능하나 시뮬레이션 결과 2.보다 1.이 효과적인 것으로 나타나 본 논문에서는 1.의 방법만을 사용하였다.

1. $k = 1/(D-1)$이고 $D = \exp(\beta_2^T \mathrm{Var}(x_2|x_1)\beta_2)$이므로 각 각의 이터레이션(iteration)에서 β_2의 추정치와 x의 주변분포로부터 $\mathrm{Var}(x_2|x_1)$의 추정치를 대입하여 k의 추정치를 구한다.

2. McCullagh와 Nelder(1989, p. 374)가 지적한 것처럼 다음과 같이 피어슨의 χ^2값과 기대값을 같게 놓고 이것을 k에 대하여 수치적 방법을 이용하여 푼다.

$$\sum \frac{(y-\hat{\mu^*})^2}{\hat{\mu^*} + (\hat{\mu^*})^2/k} - (n-p) = 0 \tag{2.4}$$

모든 계산은 Xlisp-Stat(Tierney, 1990)이 사용되었으며 Dec 및 Sun 워크스테이션에서 수행되었다.

3. 변수선택기준

Mallow's C_p와 비슷한 방법으로 유도된 변수선택기준을 적용하기 위하여 "표준제곱예측오차"를 다음과 같이 정의하였다.

$$C_1^* = \sum \frac{\mathrm{E}\{[\hat{y}(x_1) - \mathrm{E}(y|x_1)]^2|x_1\}}{\mathrm{Var}(y|x_1,x_2)} = \sum \frac{\mathrm{E}\{[\hat{\mu^*}(x_1) - \mu^*(x_1)]^2|x_1\}}{\mathrm{Var}(y|x_1,x_2)} \tag{3.1}$$

식의 분모에서 보듯이 항상 전체모형의 분산(부분모형 중에서 최소분산과 의미가 같다.)으로 표준화 하였다. 여기서 $y(\hat{x_1})$는 x_1만을 포함한 모형에서 반응변수에 대한 예측치로 $\hat{\mu^*}(x_1)$이 되며 분모의 분산을 v로 놓으면 식 (3.1)은

$$C_1^* = \sum \frac{1}{v} \mathrm{Var}(\hat{\mu^*}(x_1)|x_1) \tag{3.2}$$

가 된다. 여기에 의사우도함수기법을 이용한 추정치들을 대입하게 되는 데 구체적으로 β의 추정치들은 반복가중최소제곱법(iterated weighted least squares method)을 통하여 구할 수 있다. Williams(1987)에 의하면 점근적으로 $\mathrm{Var}(\hat{\mu^*}|x_1) = v^*h^*$가 되어

$$C_1^* \approx \sum \frac{v^*}{v} h^* \tag{3.3}$$

가 되며 여기서 $v^* = \mathrm{Var}(y|x_1)$이다. 설명변수 x에 대한 관측치를 모두 포함하는 X와 X_1은 크기가 각 각 $n \times (p+1)$, $n \times (p_1+1)$인 행렬이고 W^*가 X_1행렬의 대각 가중치 행렬일 때 h^*는 $(W^*)^{1/2}X_1(X_1^TW^*X_1)^{-1}X_1^T(W^*)^{1/2}$행렬의 대각원소들이 된다. 식 (3.3)에 v^*, v 그리고

h^*의 추정치를 대입하면 C_1^*의 추정치를 구할 수 있다. 물론 v^*, v 그리고 h^*의 추정치는 모두 β_0와 β_1의 함수이므로 이들의 추정치인 $\hat{\beta}_0$와 $\hat{\beta}_1$을 대입하면 된다.

이제 C_1^*의 추정치를 S_p라 하면 이 값은 항상 p_1+1보다 크거나 같아지게 된다. 만일 현재의 모형에서 사용된 설명변수들의 부분집합 x_1이 필요한 모든 설명변수를 포함한다면 $v^* \approx v$가 되고 $\sum h^* = p_1+1$ (Williams, 1987)이 되어 S_p가 대략 p_1+1이 되므로 Mallows의 기준과 비슷하게 $S_p \leq k$인 모형들을 최적모형의 대상으로 생각할 수 있다.

4. 모의실험 및 자료분석

4.1. 모의실험(SIMULATION) 비교

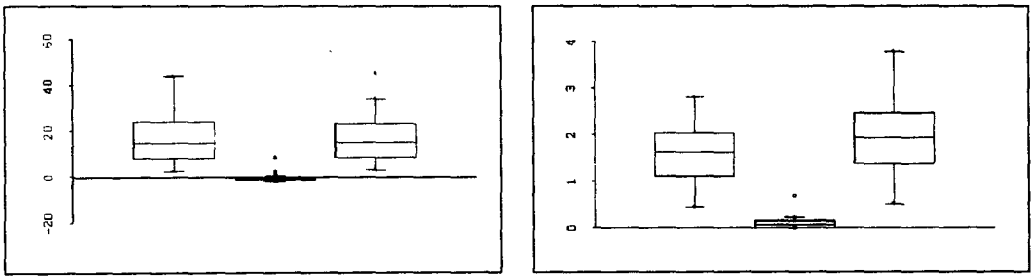
Mallows의 C_p방법은 설명변수들이 고정된 값이라는 가정에 바탕을 두고 있어 본 논문에서 제시한 방법과의 직접적인 비교는 불가능하다고 할 수 있으나 서로 유사한 점이 많아 모의실험을 통하여 이를 비교하여 보았다. 비교한 C_p방법은 포아송 회귀분석을 위하여 Hosmer와 동료들 (1989)이 선형회귀를 위한 C_p방법을 개조한 것이다. 비교를 위하여 사용된 모형에서 설명변수 $x=(x_1,x_2,x_3,x_4)^T$가 주어졌을 때 반응변수의 분포는 평균이

$$\mu(x) = \exp(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4) \tag{4.1}$$

인 포아송분포를 따르며 설명변수 x는 $N_4(0, I_4+\rho(J_4-I_4))$인 정규분포에서 추출되었다. ρ값은 독립인 경우, 상관관계가 큰 경우, 두 경우의 중간인 경우로 구분하여 0, 0.5, 0.95를 사용하였으며 회귀계수 β는 계수의 값이 다른 경우와 같은 경우를 고려하기 위하여

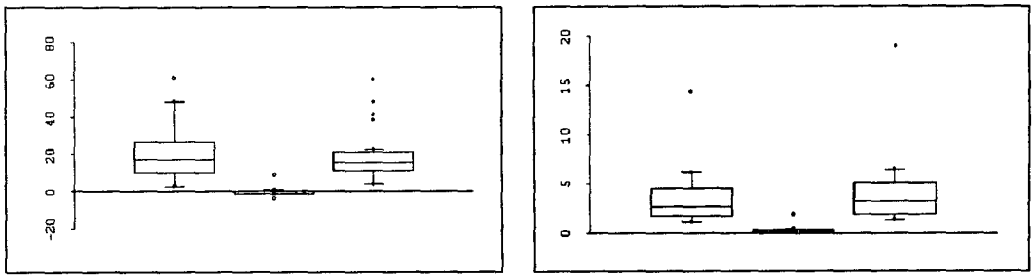
$$\beta=(\beta_0,\cdots,\beta_4)=\begin{cases}(1,1,0.5,0.2,0)\\(1,0.5,0.5,0,0)\end{cases} \tag{4.2}$$

를 사용하였다. 각각의 ρ와 β의 조합에 대하여 20번씩의 반복실험을 수행하였으며 각각의 실험에서 표본의 크기는 100으로 통일하였다. 연산은 Tierney(1990)의 Xlisp-Stat에 있는"glim prototype"(Tierney, 1991)을 수정하여 각 부분모형에 대한 정확한 분산(포아송모형의 분산과 다른)과 분산에 포함된 k값을 앞에서 설명한 두 가지 방법 중 1.을 이용하여 추정하고 끝으로 회귀계수들의 추정치는 뉴튼-랩손법(Newton-Raphson method)을 이용하여 추정하였으며 그 결과의 일부를 아래의 그림 4.1, 그림 4.2, 그림 4.3에 나타내었다.



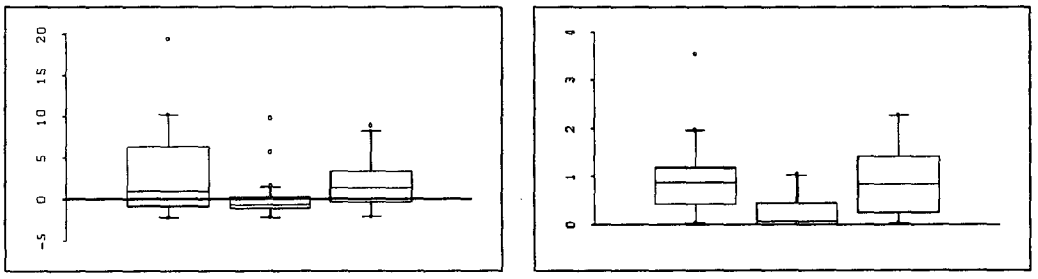
$C_p - (p_1 + 1)$ $S_p - (p_1 + 1)$

그림 4.1: $\beta = (1, 1, 0.5, 0.2, 0)$이고 $\rho = 0$인 경우 왼쪽 상자부터 $\{x_1, x_2\}$, $\{x_1, x_2, x_3\}$, $\{x_1, x_2, x_4\}$ 모형



$C_p - (p_1 + 1)$ $S_p - (p_1 + 1)$

그림 4.2: $\beta = (1, 1, 0.5, 0.2, 0)$이고 $\rho = 0.5$인 경우 왼쪽 상자부터 $\{x_1, x_2\}$, $\{x_1, x_2, x_3\}$, $\{x_1, x_2, x_4\}$ 모형



$C_p - (p_1 + 1)$ $S_p - (p_1 + 1)$

그림 4.3: $\beta = (1, 1, 0.5, 0.2, 0)$이고 $\rho = 0.95$인 경우 왼쪽 상자부터 $\{x_1, x_2\}$, $\{x_1, x_2, x_3\}$, $\{x_1, x_2, x_4\}$ 모형

표 4.1: 14개의 섬에 살고 있는 새의 종수 및 설명변수 AR, EL, DEc, DNI

섬의 이름	N	AR	EL	Dec	DNI
Chiles	36	0.33	1.26	36	14
Las Papas-Coconuco	30	0.50	1.17	234	13
Sumapaz	37	2.03	1.06	543	83
Tolima-Quindio	35	0.99	1.90	551	23
Paramillo	11	0.03	0.46	773	45
Cocuy	21	2.17	2.00	801	14
Pamplona	11	0.22	0.70	950	14
Cachira	13	0.14	0.74	958	5
Tama	17	0.05	0.61	995	29
Batallon	13	0.07	0.66	1065	55
Merida	29	1.08	1.50	1167	35
Perija	4	0.17	0.75	1182	75
Santa Marta	18	0.61	2.28	1238	75
Cende	15	0.07	0.55	1380	35

각 각의 그림에서 왼쪽은 $C_p-(p_1+1)$을 오른쪽은 $S_p-(p_1+1)$의 상자그림이며 각각의 상자들은 왼쪽부터 모형에 포함된 설명변수들이 $\{x_1,x_2\}$, $\{x_1,x_2,x_3\}$, $\{x_1,x_2,x_4\}$으로 실제의 β는 $\beta=(1,1,0.5,0.2,0)$인 경우이다. 전체적으로 C_p와 S_p는 비슷한 양상을 보이고 있으나 S_p가 C_p보다 변동이(분산이)적고 값 또한 작으며, 특히 설명변수들 간의 상관관계가 큰 경우에 S_p가 C_p보다 필요한 설명변수의 선택에 있어 우월하였다. 그러나 두 방법은 다른 가정하에서 구해진 것이기 때문에 직접적인 비교는 할 수 없으나 굳이 따지자면 설명변수들이 확률적인 경우에는 S_p를 이용한 방법이 더 적합하다고 할 수 있다.

4.2. 사례분석 (MANLY, 1991)

이 자료는 원래 Vuilleumier(1970)의 연구에서 비롯된 것으로 북부안데 (northern Andes)의 파라모(paramo)초원지역의 고립된 섬들에 살고있는 새의 종류를 조사한 자료이다. 이 연구의 목적은 새의 종류(N)가 4개의 설명변수인 AR ($1000km^2$으로 표시된 섬의 면적), EL (m로 표시된 섬의 고도), DEc (에쿠아도르에서 km로 표시된 거리), 그리고 DNI (km로 표시된 가장 가까운 섬까지의 거리)와 어떠한 관계가 있는 지를 알아보고자 한 것이다. 조사된 자료의 수(섬의 수)는 14개이며 자료는 표 4.1과 같다.

모의실험의 결과와 마찬가지로 C_p 및 S_p의 값이 전 모형에 걸쳐 비슷한 양상을 보였으나 AR과 DEc를 포함하는 모형에서 C_p는 1.84로 낮은 값을 보였다. C_p의 경우에는 AR, EL, DEc를 포함하는 모형과 AR, DEc, DNI를 포함하는 모형을 그리고 S_p의 측면에서는

표 4.2: 모형별 C_p 및 S_p 값

모형	1	2	3	4	1, 2	1, 3	1,4
C_p	13.0	14.7	8.1	21.7	13.3	1.8	13.3
S_p	8.4	10.6	6.1	12.0	11.7	3.7	11.4
모형	2, 3	2, 4	3, 4	1, 2, 3	1, 2, 4	1, 3, 4	2, 3, 4
C_p	5.0	17.4	9.3	3.0	14.2	3.9	6.3
S_p	4.8	13.7	8.3	4.0	13.3	4.9	6.0

AR, EL, DEc를 포함하는 모형이 최적에 가까운 것으로 보인다. 표 4.2에 모든 모형에 대한 C_p와 S_p값들을 나타내었다.

5. 결론

설명변수들이 정규분포를 따르거나 정규분포에 근사할 때 포아송모형에서 변수들의 부분집합을 선택하는 문제를 고려하여 보았다. 여기에서 사용된 기본적인 아이디어는 설명변수들이 확률변수일 때 주어진 정보의 양이 적을 때의 분산이 정보의 양이 많을 때보다 커진다는 점이다. 다시 말하면 포아송모형이 적합한 자료의 경우 주어진 모형이 필요한 모든 설명변수를 포함하지 못하면 그 분산은 평균보다 커지며 이 값은 모형에 포함되지 못한 변수들의 중요도에 따라 그 크기가 결정된다. 우리는 이 아이디어에 바탕을 두고 Mallows의 C_p와 비슷한 원리의 S_p라는 기준을 제시하였다.

참고문헌

[1] Hosmer, D.W., Jovanovic, B. and Lemeshow, S. (1989). Best subsets logistic regression, *Biometrics.* Vol. 45. 1265-70.

[2] Mallows, C.L. (1973). Some comments on C_p. *Technometrics.* Vol. 15. 661-75.

[3] McCullagh, P. (1983). Quasi-likelihood functions. *Annals of Statistics.* Vol. 11. 59-67.

[4] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* (2nd ed.). Chapman and Hall.

[5] Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A.* Vol. 135. 370-84.

[6] Park, C.S. (1995). Topics in Generalized Linear Models. University of Minnesota. unpublished Ph. D. thesis.

[7] Tierney, L. (1990). *Lisp-Stat: An Object-oriented Environment for Statistical Computing and Dynamic Graphics.* J. Wiley & Sons.

[8] Tierney, L. (1991). Generalized Linear Models in Lisp-Stat. *Technical report* No. 557. University of Minnesota.

[9] Wedderburn, R.W.M. (1974). Quasilikelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika.* Vol. 61. 439-47.

[10] Williams, D.A. (1987). Generalized linear model diagnostics using the deviance and single-case deletions. *Applied Statistics.* Vol.36. 181-91.

[1997년 4월 접수, 1998년 2월 최종수정]

Subset Selection in the Poisson Models - A Normal Predictors case -

Chongsun Park [1)]

ABSTRACT

In this paper, a new subset selection problem in the Poisson model is considered under the normal predictors. It turns out that the subset model has bigger variance than that of the Poisson model with random predictors and this has been used to derive new subset selection method similar to Mallows' C_p.

1) Department of Statistics, Sung Kyun Kwan University, Seoul, Korea.