응용통계연구 제 18권 1호, 2005년, pp. 173–182

순서범주형자료 분석를 위한 베이지안 분계점 모형*

최병수[1] 이승천[2]

요 약

순서를 갖는 범주형자료의 분석을 위한 중요한 통계적 방법인 순위로짓모형의 대안으로 무정보 사전분포에 의한 베이지안 분계점 모형을 정의하고, 실증 자료분석을 통해 베이지안 모형의 유용성을 살펴보았다.

주요용어: 베이지안 분계점 모형, 깁스표본, 순위로짓모형, 로지스틱모형

1. 서론

자료분석에서 중요한 의미를 갖는 많은 변수들이 흔히 이진 또는 순서척도에 의해 측정되는 범주형인 경우가 많다. 이진 또는 순서척도를 갖는 범주형 변수들에 대한 전통적인 통계적 분석은 로지스틱회귀모형과 순위로짓(ordered logit) 회귀모형에 의존하는 바가 크다고 하겠다. 그러나 이러한 모형에서 모수 추정은 최대우도법에 의존하는 바, Griffiths (1987) 등에 의하면 표본크기가 작을 경우 추정값의 편향이 큰 것으로 알려져 있다.

Albert와 Chib (1993)은 로지스틱 또는 순위로짓모형에서 전통적인 추정방법에 대한 문제를 제기하고 범주형 자료에 대한 베이지안 모형을 제시하였다. Albert와 Chib (1993)의 베이지안 모형은 선형회귀모형을 따르는 잠재변수와 가상의 분계점을 가정하고 잠재변수와 분계점의 값에 의해 범주를 구별하게 된다. 즉, U가 순서범주형 변수 Y에 대응되는 잠재변수라고 할 때, $t_{\ell-1} < U \leq t_\ell$ 이면 $Y = \ell$이 된다고 가정한다. 이때 분계점인 t들은 모형에서 추정하여야 할 미지의 모수이다. 이하에서 이러한 가정에 따르는 베이지안 모형을 베이지안 분계점 모형으로 부르기로 하자.

잠재변수와 분계점에 의한 순서범주형 변수의 표현은 매우 오랜 기원을 갖는 것으로 Wright (1934)에서 시작되었다고 한다. 또 Bliss (1935)도 독성실험에서 이와 같은 개념을 사용하였다고 하며, 이후 Gianola (1982), Foulley (1987), Wang (1994), Lee와 Lee (2002) 등 주로 육종학 분야에서 많은 실험 데이터들이 잠재변수와 분계점 개념에 의해 분석되었다.

육종학 분야에서 특히 이러한 모형에 관심을 갖는 이유 중 하나는 분산요소의 추정과 매우 밀접한 관계가 있다. 즉, 육종학에서 분산요소는 유전성을 측정하는 중요한 모수인데 범주형 자료에 의해 분산요소를 추정하는 것은 매우 어려운 문제이다. 그러나 연속형 변수인 잠재변수를 이용하면 보다 쉽게 문제를 해결할 수 있다. 이와 같이 잠재변수를 가정한

* 본 연구는 2004학년도 한성대학교 교내연구비 지원과제임

1) (136-793) 서울시 성북구 삼선동3가 389. 한성대학교 컴퓨터공학부, 교수
E-mail: cbs@hansung.ac.kr
2) (447-791) 경기도 오산시 양산동 411, 한신대학교 정보통계학과, 교수
E-mail: seung@hanshin.ac.kr

베이지안 모형은 여러 응용 분야에 사용될 수 있는 유연성을 갖고 있다. 또 Albert와 Chib (1993)은 다른 측면에서 베이지안 모형의 유연성을 설명하기도 하였다.

Albert와 Chib (1993)은 계층적 사전 공액분포에 의해 계층적 분계점 베이지안 모형을 설정한 후, 모수들의 완전 조건부 분포(full conditional distribution)를 유도하여 깁스표본에 의한 추정 방법을 제시하였다. 그러나 그들의 모형은 전통적인 입장에서 보자면 분산성분모형으로 일반적인 순위로짓모형의 가정과는 일치하지 않는다. 이것은 사전 공액 분포의 특성으로 인한 것으로 본 논문에서는 Soresen (1995) 등과 Lee와 Lee (2002)에서 연구된 무정보 사전분포에 의한 베이지안 분계점 모형을 순위로짓모형의 대안으로 제시하고 실증자료분석을 통하여 설정된 베이지안 모형의 유용성을 입증하려고 한다.

2. 베이지안 분계점 모형

$Y_i, i = 1, 2, \ldots, n$은 c 개의 범주를 갖는 순서범주형 확률변수로서, Y_i의 값은 잠재 확률변수 U_i와 분계점 $\mathbf{t} = (t_1, \ldots, t_{c-1})$에 의해

$$t_{\ell-1} < U_i \leq t_\ell \quad \text{이면} \quad Y_i = \ell$$

와 같이 결정된다고 가정한다. 이때 $-\infty = t_0 < t_1 < \ldots < t_{c-1} < t_c = \infty$ 이다. 또한 잠재 확률벡터 $\mathbf{u} = (U_1, \ldots, U_n)'$는 $\mathbf{X}, \boldsymbol{\theta}, \epsilon$가 각각 $n \times p$인 계획행렬, $p \times 1$ 회귀계수벡터, $n \times 1$ 오차벡터라고 할 때,

$$\mathbf{u} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \tag{2.1}$$

와 같은 선형회귀모형을 가정한다.

Albert와 Chib (1993)은 오차항 분포가 정규분포인 프로빗 연결함수와 t 연결함수를 고려하였다. 로지스틱 연결함수는 자유도 9인 t 연결함수와 매우 유사한 형태를 갖는다고 하는데 여기에서는 모형의 간편성을 위하여 프로빗 연결함수를 고려하기로 한다. 즉,

$$\mathbf{u}|\boldsymbol{\theta} \sim N(\mathbf{X}\boldsymbol{\theta}, \mathrm{I}) \tag{2.2}$$

을 가정한다.

잠재변수는 가상적인 것이므로 정규성을 가정하였을 때 임의의 위치 및 척도를 갖을 수 있다. 이러한 임의성은 모수의 추정에서 식별성의 문제를 야기하게 되므로 위치와 척도에 대한 제약조건을 가하여야 한다. 일반적으로 제약조건은 오차항의 분산값과 하나의 분계점 값을 지정하거나 또는 두 개의 분계점 값을 지정한다. 실제로 두 종류의 제약조건은 모수 해석상의 차이가 있을 뿐 사실상 모형상에는 차이가 없다. 그러나 분산값을 지정하면 오차항 분산에 대한 사전분포를 필요로 하지 않게되므로 여기에서는 오차항 분산값을 1로 설정한다.

$\mathbf{x}_{i\cdot}'$를 $\mathbf{X}$의 i-번째 행이라고 하면 주어진 가정하에 Y_i의 조건부 확률함수는

$$\begin{aligned}\Pr[Y_i = \ell|\boldsymbol{\theta}, \mathbf{t}] &= \Pr[t_{\ell-1} < U_i \leq t_\ell|\boldsymbol{\theta}, \mathbf{t}] \\ &= \Phi(t_\ell - \mathbf{x}_{i\cdot}'\boldsymbol{\theta}) - \Phi(t_{\ell-1} - \mathbf{x}_{i\cdot}'\boldsymbol{\theta})\end{aligned}$$

와 같이 유도된다. 그러므로 U_i의 완전조건부 분포는 다음과 같은 절단 정규분포를 따르게 된다.

$$f(u_i|\theta,\mathrm{t},\mathbf{y}) = \frac{\phi(\mathbf{x}_i'.\theta,1)}{\Phi(t_\ell - \mathbf{x}_i'.\theta) - \Phi(t_{\ell-1} - \mathbf{x}_i'.\theta)}\mathbf{1}(t_{\ell-1} < u_i \le t_\ell), \quad \text{단 } Y_i = \ell \tag{2.3}$$

여기서 $\Phi(\cdot)$는 표준정규분포의 분포함수이고, $\phi(\mu,\sigma)$는 평균과 분산이 각각 μ와 σ^2인 정규분포의 밀도함수를 나타낸다. 한편 $\mathbf{y}$와 Y_i의 완전조건부 분포는 다음과 같은 퇴화분포가 된다.

$$p(\mathbf{y}|\theta,\mathrm{t},\mathrm{u}) = \prod_{i=1}^{n}\left\{\sum_{\ell=1}^{c}\mathbf{1}(t_{\ell-1} < u_i \le t_\ell)\mathbf{1}(Y_i = \ell)\right\} \tag{2.4}$$

$$\Pr[Y_i = \ell|\theta,\mathrm{t},\mathrm{u}] = 1(t_{\ell-1} < u_i \le t_\ell) \tag{2.5}$$

회귀계수 θ와 분계점 t의 사전분포에 대해 알아보기로 하자. 전통적 회귀모형에서 회귀계수가 모수인자임을 가정하다면 베이지안 모형에서 θ의 무정보 사전분포는

$$f(\theta) \propto \text{constant}$$

와 같이 설정할 수 있다. 한편 (2.1)이 절편을 포함한 회귀모형이라고 하면 식별성을 위하여 일반적으로 $t_1 = 0$이 주어진다. 그러므로 단지 $t_2,\ldots,t_{c-1}$에 대한 사전분포만을 필요로 하게 되는데, 이들은 $0 < t_2 < \cdots < t_{c-1}$와 같은 순서를 갖고 있어 서로 독립이 아님을 알 수 있다. 이와 같이 순서를 갖는 모수의 무정보 사전분포로서 균일분포에서 추출된 순서통계량의 분포를 설정할 수 있다. 즉 분계점의 무정보 사전분포는

$$f(t_2,\ldots,t_{c-1}) = (c-2)!\left(\frac{1}{t_{\max}}\right)^{c-2}\mathbf{1}(0 < t_2 < \ldots < t_{c-1} < t_{\max})$$

와 같다.

θ와 t가 서로 독립임을 가정할 때, $\mathbf{y}$ 조건부 $\theta,\mathrm{t},\mathrm{u}$의 결합확률분포는

$$\begin{aligned} f(\theta,\mathrm{t},\mathrm{u}|\mathbf{y}) &\propto f(\theta)f(\mathrm{t})f(\mathrm{u}|\theta,\mathrm{t})p(\mathbf{y}|\theta,\mathrm{t},\mathrm{u}) \\ &= f(\theta)f(\mathrm{t})f(\mathrm{u}|\theta)p(\mathbf{y}|\mathrm{t},\mathrm{u}) \end{aligned} \tag{2.6}$$

인 관계를 갖는다. 이 식에서 $f(\mathrm{t})$와 $f(\mathbf{y}|\mathrm{t},\mathrm{u})$는 θ를 포함하고 있지 않으므로 θ의 완전 조건부 분포는

$$f(\theta|\mathrm{t},\mathrm{u},\mathbf{y}) \propto f(\theta)f(\mathrm{u}|\theta) \propto f(\mathrm{u}|\theta)$$

가 성립되어

$$\theta|\mathrm{t},\mathrm{u},\mathbf{y} \sim N\left((\mathbb{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u},(\mathbf{X}'\mathbf{X})^{-1}\right)$$

임을 알 수 있으며 다음과 같은 정리가 유도된다.

정리 2.1 $\theta_j, j = 1, 2, \ldots, p$의 완전조건부 분포는

$$\theta_j|\boldsymbol{\theta}_{-j}, \mathbf{t}, \mathbf{u}, \mathbf{y} \sim N\left((\mathbf{x}'_{\cdot j}\mathbf{x}_{\cdot j})^{-1}\mathbf{x}'_{\cdot j}\left(\mathbf{u} - \mathbf{X}_{-j}\boldsymbol{\theta}_{-j}\right), (\mathbf{x}'_{\cdot j}\mathbf{x}_{\cdot j})^{-1}\right) \tag{2.7}$$

와 같다.

정리 2.1에서 $\boldsymbol{\theta}_{-j}$와 $\mathbf{X}_{-j}$는 각각 $\boldsymbol{\theta}$와 $\mathbf{X}$에서 j-번째 요소와 j-번째 열을 제외한 나머지를 나타내고, $\mathbf{x}_{\cdot j}$는 $\mathbf{X}$의 j-번째 열이다. 이하 "$-$" 첨자는 벡터 또는 행렬에서 해당 요소 또는 벡터를 제외한 나머지 부분을 나타내기로 한다.

마지막으로 분계점 $\mathbf{t}$의 완전 조건부분포는 (2.6)에서

$$f(\mathbf{t}|\boldsymbol{\theta}, \mathbf{u}, \mathbf{y}) \propto f(\mathbf{t})p(\mathbf{y}|\mathbf{t}, \mathbf{u}) \propto p(\mathbf{y}|\mathbf{t}, \mathbf{u})$$

이 성립되는데, (2.4)를 t_ℓ의 함수로 본다면 다음과 같은 정리가 성립된다.

정리 2.2 $t_\ell|\mathbf{t}_{-\ell}, \boldsymbol{\theta}, \mathbf{u}, \mathbf{y}$는 구간 $(\max\{\max(\mathbf{u}|Y_i = \ell), t_{\ell-1}\}, \min\{\min(\mathbf{u}|Y_i = \ell+1), t_{\ell+1}\})$에서 균일분포를 따른다.

$$t_\ell|\mathbf{t}_{-\ell}, \boldsymbol{\theta}, \mathbf{u}, \mathbf{y} \sim U((\max\{\max(\mathbf{u}|Y_i = \ell), t_{\ell-1}\}, \min\{\min(\mathbf{u}|Y_i = \ell+1), t_{\ell+1}\}) \tag{2.8}$$

이제 (2.3), (2.7), (2.8)으로부터 반복적으로 깁스표본을 추출한다. 초기의 깁스표본은 안정화되어 있지 않으므로 초기에 얻어진 깁스표본은 사용하지 않는다. 일정 횟수의 반복 이후에 얻어진 깁스표본을 $Z_1, Z_2, \ldots$ 라고 하자. 일반적으로 Z_i들은 매우 강한 양의 상관관계를 갖게 되는데, 실제 모수 추정에는 자기상관이 0이 되는 시차의 간격을 두고 얻어진 깁스표본을 이용한다. 즉, 시차 간격을 l이라고 하면 사후 평균의 추정에 사용되는 깁스표본은 $X_i = Z_{i\times l}, i = 1, 2, \ldots, m$이 된다. 이렇게 얻어진 깁스표본을 이용하여 사후평균과 분산은 각각

$$\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} X_i$$

$$\hat{\sigma}^2_m = \frac{1}{m}\sum_{i=1}^{m}(X_i - \hat{\mu}_m)^2$$

으로 추정될 수 있다. 실제 추정에 있어서는 추정값이 실제값에 수렴하기 위한 표본크기 m의 값을 알아야 하는데, 이 문제에 대해서는 Raftery와 Lewis (1992)를 참조할 수 있다.

(2.8)의 완전 조건부 분포는 범주의 수가 3 이상인 경우에 해당되는 것으로 이진변수의 경우는 해당되지 않는다. 즉, 이진변수의 경우 하나의 분계점 $t_1 = 0$ 만이 필요하므로 분계점에 대한 사전분포를 요구하지 않는다. 또 이 경우 (2.3)는

$$f(u_i|\boldsymbol{\theta}, \mathbf{y}) = \begin{cases} \frac{\phi(\mathbf{x}'_{i\cdot}\boldsymbol{\theta}, 1)}{\Phi(\mathbf{x}'_{i\cdot}\boldsymbol{\theta}, 1)}\mathbf{1}(u_i > 0), & Y_i = 1\text{일 때} \\ \frac{\phi(\mathbf{x}'_{i\cdot}\boldsymbol{\theta}, 1)}{1-\Phi(\mathbf{x}'_{i\cdot}\boldsymbol{\theta}, 1)}\mathbf{1}(u_i \le 0), & Y_i = 0\text{일 때} \end{cases} \tag{2.9}$$

와 같이 수정되어야 하며, 깁스표본은 (2.7)와 (2.9)로부터 얻어지게 된다.

3. 사례분석

3.1. Morz 데이터

Morz (1987)은 기혼 여성이 직업을 갖을 것인지에 대한 의사결정에 영향을 미치는 여러 가지 사회적 환경 요인들에 대해 연구를 하였다. 이러한 연구의 일환으로 그는 백인 기혼 여성 753명에 대해 직업을 갖을 것인지에 대한 의사를 나타내는 변수 `infl` (1, 직업을 갖음, 0, 갖지 않음)와 `faminc` (총 가계소득, 달러), `wage` (예상 임금, 달러/시간), `hours`(일한 시간), `educ` (총 교육연수), `expr` (경력 연수), `age` (나이), `kidslt6` (6세 미만의 자녀의 수), `kidsgt6` (6세 이상 18세 미만의 자녀의 수) 등과 같은 환경적 변수들을 측정하였다.

Morz의 데이터는 여러 논문에서 인용된 것으로 Lee와 Huh (2003)은 기혼 여성의 직업 참여에 대한 의사결정모형으로 종속변수 `infl`에 대해 `nwifeinc, edu, expr, exprsq, age, kidslt6, kidsge6`의 7개 독립변수를 갖는 로지스틱 회귀모형을 고려하였다. 여기서 `nwifeinc`와 `exprsq`는 각각

$$\texttt{nwifeinc} = (\texttt{faminc} - \texttt{wage} \times \texttt{hours})/1000$$

$$\texttt{exprsq} = \texttt{expr} \times \texttt{expr}$$

와 같이 산출된 변수들로서 특히 `nwifeinc`는 총 가계소득 중에서 여성의 소득을 따로 산출한 것이다. 이렇게 설정된 로지스틱 회귀모형의 적합 결과는 표 3.1과 같다. 적합된 회귀모형에서 절편과 `kidsge6`는 유의하지 않은 것으로 나타 났으며, 총 753개의 데이터에서 199개를 오분류하여 오분류율이 26.43%인 것으로 나타났다.

한편 같은 모형에 대해 베이지안 분계점 모형을 적합시킨 결과는 표 3.2와 같다. 표에 나타난 회귀계수 추정은 2000번의 반복 후, 매 10회의 반복에서 한번씩 총 4000번회 깁스 표본의 표본평균으로 부터 구하여진 결과이다. 따라서 추정값을 구하기 위해 총 42000회의 반복을 실시하였다. 또 깁스표본의 수렴을 확인하기 위해 비교적 큰 분산을 갖는 `kidge6`의

표 3.1: Morz 데이터의 로지스틱 회귀모형

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.425452   0.860365   0.495  0.62095
nwifeinc    -0.021345   0.008421  -2.535  0.01126 *
educ         0.221170   0.043439   5.091  3.55e-07 ***
expr         0.205870   0.032057   6.422  1.34e-10 ***
exprsq      -0.003154   0.001016  -3.104  0.00191 **
age         -0.088024   0.014573  -6.040  1.54e-09 ***
kidslt6     -1.443354   0.203583  -7.090  1.34e-12 ***
kidsge6      0.060112   0.074789   0.804  0.42154
---

Residual deviance: 803.5303 on 745 degrees of freedom
AIC: 819.5303
```

```
Confusion matrix

          Predicted

Actual Frequency|   0|   1| Total
       ---------+----+----+
              0 | 207| 118|   325
       ---------+----+----+
              1 |  81| 347|   428
       ---------+----+----+
       Total      288  465    753
```

표 3.2: Morz 데이터의 계층적 베이지안 회귀모형

```
Coefficients:
             Estimate Std. Error z value  Pr(>|z|)
Intercept  0.241739100  0.522643   0.463  0.64370
nwifeinc  -0.012535900  0.004812  -2.605  0.00918 **
educ       0.134406800  0.025780   5.214  1.85e-07***
expr       0.117072400  0.018558   6.309  2.82e-10***
exprsq    -0.001702628  0.000597  -2.853  0.00433 **
age       -0.052632050  0.008389  -6.274  3.51e-10***
kidslt6   -0.883725000  0.117229  -7.538  4.75e-14***
kidsge6    0.034927780  0.044021   0.793  0.42753
---

Residual deviance: 803.0034 on 745 degrees of freedom
AIC: 819.0034
```

```
Confusion matrix

Actual     Predicted

A  Frequency|    0|    1| Total
c  ---------+-----+-----+
t         0 |  214|  111|   325
u  ---------+-----+-----+
a         1 |   81|  347|   428
l  ---------+-----+-----+
   Total       288   465    753
```

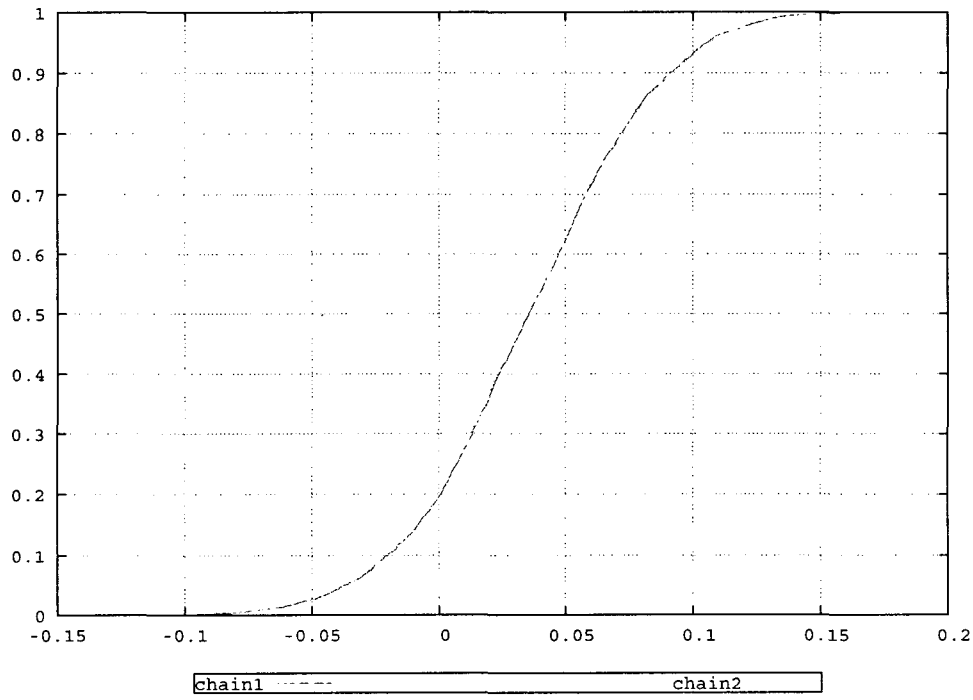

그림 3.1: 두 개의 깁스 체인에서 구한 kidge6 회귀계수의 경험적 분포함수

회귀계수에 대해 2회 깁스 체인을 실행하였으며 두 개의 체인에서 구한 경험적 분포함수는 그림 3.1과 같이 나타나 두 개의 체인이 수렴하고 있음을 알 수 있다. 다른 회귀계수에서도

모두 깁스표본의 수렴을 확인할 수 있었다.

베이지안 모형에서 얻어진 회귀계수의 추정값을 로지스틱 회귀모형에서 구한 회귀계수의 추정값과 직접 비교하기는 곤란하지만 두 모형에서 모두 같은 부호로 추정되어 각 변수들이 inlf에 대해 같은 방향으로 영향을 미치고 있다. 특히 베이지안 모형에서 AIC 정보량은 819.0034로서 로지스틱 회귀모형과 비교하여 근소하나마 선호되는 모형으로 판단할 수 있다. 이 결과는 베이지안 모형의 오분류표에서도 확인할 수 있다. 즉 베이지안 모형의 오분류율 약 25.50%로 로지스틱 회귀모형과 비교하여 근소하게나마 작은 오분류율을 나타내고 있다. 그러므로 Morz 데이터에서 기혼 여성의 직업 참여에 대한 의사결정모형으로 베이지안 분계점 모형이 로지스틱 회귀모형보다는 선호된다고 하겠다.

3.2. 와인 데이터

와인 데이터는 이태리의 한 지역에서 숙성된 세 품종의 와인에 대한 13 가지 화학적 성분과 품종을 나타내는 14개의 변수로 구성된 데이터로서 http://www.ics.uci.edu/~mlearn/MLRepository.html에 수록되어 있다. 이 데이터는 기계학습 분야에서는 잘알려진 것으로 판별분석에서 분류자(classifier)의 성능을 평가하는데 많이 이용되고 있다. 현재까지 알려진 바에 의하면 오직 RDA 방법에 의해서만 100% 정확하게 품종을 구별할 수 있었다고 한다. 그 외에 LDA와 QDA는 각각 98.9%와 99.4%의 정분류율을 갖는다고 한다.

와인데이터의 클래스변수는 품종을 나타내는 명목척도의 성질을 갖는 변수이기는 하지

표 3.3: Wine 데이터의 순위로짓모형

Coefficients:

	Estimate	Std. Error	Chi-Square	Pr > ChiSq
Inter1	-52.2628	71.0927	0.5404	0.4623
inter2	73.0859	101.4	0.5199	0.4709
X1	7.5959	8.5299	0.7930	0.3732
X2	3.4277	7.3637	0.2167	0.6416
X3	7.9879	10.2260	0.6102	0.4347
X4	-31.4736	44.2822	0.5052	0.4772
X5	1.1667	1.4877	0.6151	0.4329
X6	-8.3937	14.8717	0.3186	0.5725
X7	52.0534	77.0694	0.4562	0.4994
X8	11.2979	16.0164	0.4976	0.4806
X9	-8.9861	16.1263	0.3105	0.5774
X10	-33.8693	50.0522	0.4579	0.4986
X11	10.3689	16.9060	0.3762	0.5397
X12	19.9907	26.5423	0.5673	0.4514
X13	48.4759	75.3111	0.4143	0.5198

-2 Log L : 18.997
AIC: 48.997

Confusion matrix

		Predicted			
	Frequency	1	2	3	Total
Actual	1	58	1	0	59
	2	0	70	1	71
	3	0	2	46	48
Total		58	73	47	178

표 3.4: Wine 데이터의 베이지안 분계점 회귀모형

```
Coefficients:    Thresholds = 0, 50.90356

        Estimate Std. Error z value  Pr(>|z|)
Inter 21.0360300  2.9519220   7.126  1.032e-12 ***
X1    -3.4150420  1.5858156  -2.153  0.03128   *
X2    -1.2410890  0.8458893  -1.467  0.14232
X3    -3.2132480  0.9350378  -3.436  5.893e-04 ***
X4    12.4261500  1.8241565   6.812  9.625e-12 ***
X5    -0.4667664  0.4794010  -0.974  0.30233
X6     3.3799400  1.5277505   2.212  0.02694   *
X7   -20.1032300  2.4847615  -8.091  6.661e-16 ***
X8    -4.2143320  0.7499593  -5.619  1.916e-08 ***
X9     2.8871830  1.4214603   2.031  0.04224   *
X10   12.2466300  2.0710943   5.913  3.357e-09 ***
X11   -4.3306880  1.3193161  -3.283  0.00103   **
X12   -8.5971400  1.5873075  -5.416  6.088e-08 ***
X13  -18.2947100  3.1312700  -5.843  5.140e-09 ***
---

-2 Log L : 20.55785
     AIC: 48.55785
```

Confusion matrix

Actual \ Predicted	1	2	3	Total
1	58	1	0	59
2	0	71	0	71
3	0	2	46	48
Total	58	74	46	178

만 순위로짓모형도 표 3.3에서 보듯이 약 97.75%의 정분류율을 보이고 있어 클래스 변수는 품종에 따라 우열을 가릴 수 있는 순위척도의 성질을 갖고 있는 것으로 판단된다.

각 독립변수들을 표준화하여 순위로짓모형을 적합시킨 결과가 표 3.3에 나타나 있다. 특이한 것은 순위로짓모형이 데이터를 매우 잘 적합시키고 있으나 회귀계수는 모두 유의하지 않은 것으로 나타났다. 한편 10만 회의 반복에서 처음 2만 회의 반복 결과를 버린 후 매 20회 시차간격을 두고 구한 총 4000 회의 깁스표본을 이용하여 추정된 베이즈 모형의 결과는 표 3.4와 같다. 순위로짓모형과는 달리 대부분의 회귀계수가 유의하였다. 또 AIC 기준에 의해 두 모형을 비교한다면 베이지안 모형이 근소하게 선호된다고 하겠다.

이상의 두 예제에서 계층적 베이지안 모형은 이진 또는 순위를 갖는 범주형 데이터에 대해 로지스틱 회귀모형보다 우수한 적합 능력을 갖을 수 있음을 살펴보았다. 상기된 두 개의 데이터 이외에도 몇 개의 데이터에서 계층적 베이지안 모형과 로지스틱 모형을 비교하였는데 데이터에 따라 로지스틱 모형이 더 잘 적합되는 경우도 있어 어떤 모형이 일률적으로 우수하다고 판단하기는 어렵다. 그러나 많은 데이터에서 AIC 기준으로 베이지안 모형이 선호되는 것으로 나타났다.

4. 결론

본 연구는 순서척도에 의해 측정된 범주형 자료를 분석하기 위하여 베이지안 모형을 적용하였으며, 선형회귀모형을 따르는 잠재변수와 분계점 개념을 설정한 후 무정보 사전분포를 이용한 완전 조건부 분포를 유도하고 깁스표본에 의한 추정방법을 제시하였다.

본 연구에서 제안된 베이지안 분계점 모형은 몇 개의 알려진 자료에 대하여 분석한 결과 순위로짓모형에 비해 우수한 적합능력을 갖는 것으로 나타났다. 그러므로 본 연구에서 제안된 베이지안 분계점 모형을 순위로짓모형과 같이 범주형 데이터의 적합에 사용하여도 좋을 것으로 보인다.

앞으로의 연구방향은 다양한 시뮬레이션 자료에 대해 두 방법의 효율성을 비교하는 것이며, 깁스표본 이외의 계산방법에 대한 연구가 추가되어야 할 것이다.

참고문헌

Albert, J. H., and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association,* **88**, 669–679.

Bliss, C. I. (1935). The calculation of the dosage-mortality curve, *Annals of Application Biology,* **22**, 134 167.

Foulley, J. L., Im, S. Gianola, D., and Höschele, I. (1987). Empirical Bayes estimation of parameters for n polygenic binary traits, *Genetics Selection Evolution,* **23**, 309–338.

Gianola, D. (1982). Theory and analysis of threshold characters, *Journal of Animal Science,* **54**, 1079-1096.

Griffiths, W. E., Hill, R. C., and Pope, P. J. (1987). Small sample properties of probit model estimators, *Journal of the American Statistical Association,* **82**, 929–937.

Morz, T. A. (1987). The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions, *Econometrica,* **55** 765–799.

Lee, S.-C., and Huh, M. (2003). A measure of association for complex data, *Computational Statistics & Data Analysis,* **44**. 211–222.

Lee, S.-C., and Lee, D. (2002). Bayesian analysis of multivariate threshold animal models using Gibbs sampling, *Journal of the Korean Statistical Society,* **31**. 177–198.

Soresen, D. A., Anderson, S., Gianola, D., and Korsgaard, I. (1995). Bayesian inference in threshold models using Gibbs sampling, *Genetics Selection Evolution,* **27**, 229–249.

Wang, C. S., Rutledhe, J. J., and Gianola, D. (1994). Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs, *Genetics Selection Evolution,* **26**, 91–115.

Wright, S. (1934). An analysis of variability in number of digits in an inbred strain of guinea pigs, *Genetics,* **19**, 506–536.

[2004년 7월 접수, 2004년 11월 채택]

A Bayesian Threshold Model for Ordered Categorical Traits*

Byongsu Choi[1] Seung-Chun Lee[2]

ABSTRACT

A Bayesian threshold model is considered to analyze binary or ordered categorical traits. Gibbs sampler for making full Bayesian inferences about the category probability as well as the regression coefficients is described. The model can be regarded as an alternative to the ordered logit regression model. Numerical examples are shown to demonstrate the efficiency of the model.

Keywords: Bayesian threshold model, Gibbs sampling, Ordered logit regression model, Logistic regression model

* This Research was financially supported by Hansung University in the year of 2004.

1) Professor, School of Computer Engineering, Hansung University, 3-389, Samsung-Dong, Seoul, 136-793
E-mail: cbs@hansung.ac.kr

2) Professor, Dept. of Statistics, Hanshin University, 411 Yangsan-Dong, Osan, Kyunggi-Do, 447-791
E-mail: seung@hanshin.ac.kr