순서범주형자료 분석을 위한 베이지안 분계점 모형*

최병수1) 이승천2)

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, 김스표본, 순위로짓모형, 로지스틱모형

Ⅱ. 서론

자료분석에서 중요한 의미를 갖는 많은 변수들이 흔히 이진 또는 순서척도에 의해 측정되는 범주형 경우가 많다. 이진 또는 순서척도를 갖는 범주형 변수들에 대한 통계적 분석은 로지스틱회귀모형과 순위로짓(ordered logit) 회귀모형에 의존하는 바가 크다고 했다. 그러나 이러한 모형에서 모수 추정은 최대우도법에 의존하는 바, Griffiths (1987) 등에 의하면 표본크기가 작을 경우 추정값의 정확성이 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 잔재변수와 가상의 분계점을 가정하고 잔재변수와 분계점의 값에 의해 범주를 구별하게 된다. 즉, \( U \)가 순서범주형 변수 \( Y \)에 대응되는 잔재변수라 할 때, \( t_{l-1} < U \leq t_l \) 이면 \( Y = l \)이 된다고 가정한다. 이때 분계점인 \( t \)들은 모형에서 추정하여야 할 미지의 모수이다. 이하에서 이러한 가정에 따라 베이지안 모형을 베이지안 분계점 모형으로 부르기로 하자.


육종학 분야에서 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 잔재변수를 이용하면 보다 쉽게 문제를 해결할 수 있다. 이와 같이 잔재변수를 가정한

* 관련 연구는 2004학년도 한성대학교 컴퓨터공학부, 교수
1) (136-793) 서울시 서초구 삼성동3가 389. 한성대학교 컴퓨터공학부, 교수
E-mail: cbs@hansung.ac.kr
2) (447-791) 경기도 오산시 양산동 411, 한성대학교 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.


2. 베이지안 분계점 모형

\( Y_i, i = 1, 2, \ldots, n \)은 \( c \) 개의 범주를 갖는 순서변수형 확률변수로서, \( Y_i \)의 값은 잠재 확률 변수 \( U_i \)와 분계점 \( t = (t_1, \ldots, t_{c-1}) \)에 의해

\[ t_{c-1} < U_i \leq t_c \quad \text{이면} \quad Y_i = \ell \]

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 \( \mathbf{X}, \mathbf{\theta}, \epsilon \)가 각각 \( n \times p \)인 계획행렬, \( p \times 1 \) 회귀계수벡터, \( n \times 1 \) 오차벡터라고 할 때,

\[ \mathbf{u} = \mathbf{X}\mathbf{\theta} + \epsilon \quad (2.1) \]

와 같은 선형회귀모형을 가정한다.

Albert와 Chib (1993)은 오차항 분포가 정규분포인 프로빗 연결함수와 \( t \) 연결함수를 고려하였다. 로지스틱 연결함수는 자유도 0인 \( t \) 연결함수와 매우 유사한 형태를 갖는다고 하는데 여기에서는 모형의 간편성을 위하여 프로빗 연결함수를 고려하기로 한다. 즉,

\[ u|\mathbf{\theta} \sim N(\mathbf{X}\mathbf{\theta}, 1) \quad (2.2) \]

을 가정한다.

잠재변수는 가상적인 것이므로 정규성을 가정하였을 때 임의의 위치 및 척도를 갖을 수 있다. 이러한 임의성은 모수의 추정에서 식별성을 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.

\( x_i \)'를 \( X \)'의 \( i \)-번째 행이라고 하면 주어진 가정하에 \( Y_i \)의 조건부 확률함수는

\[
\text{Pr}[Y_i = \ell|\mathbf{\theta}, t] = \text{Pr}[t_{c-1} < U_i \leq t_c|\mathbf{\theta}, t] \\
= \Phi(t_c - x_i'|\mathbf{\theta}) - \Phi(t_{c-1} - x_i'|\mathbf{\theta})
\]
와 같이 유도된다. 그리므로 \(U_t\)의 완전조건부 분포는 다음과 같은 절단 정규분포를 따르게 된다.
\[
f(u_t|\theta, t, y) = \frac{\phi(x_{t_t}^\top \theta, 1)}{\Phi(t_{\ell} - x_{t_t}^\top \theta) - \Phi(t_{\ell-1} - x_{t_t}^\top \theta)}^1 t_{\ell-1} < u_t \leq t_{\ell}. \quad \text{단 } Y_t = \ell \tag{2.3}
\]
여기서 \(\Phi(\cdot)\)는 표준정규분포의 분포함수이고, \(\phi(\mu, \sigma)\)는 평균과 분산이 각각 \(\mu\)와 \(\sigma^2\)인 정규 분포의 밀도함수를 나타낸다. 한편 \(y\)와 \(Y_t\)의 완전조건부 분포는 다음과 같은 로절환분포가 된다.
\[
p(y|\theta, t, u) = \prod_{t=1}^{n} \left\{ \sum_{\ell=1}^{c} 1(t_{\ell-1} < u_t \leq t_{\ell}) I(Y_t = \ell) \right\} \tag{2.4}
\]
\[
\Pr[Y_t = \ell|\theta, t, u] = 1(t_{\ell-1} < u_t \leq t_{\ell}) \tag{2.5}
\]
회귀계수 \(\theta\)와 분별점 \(t\)의 사전분포에 대해 알아보기로 하자. 전통적 회귀모형에서는 회귀 계수가 모수인자임을 가정한다면 베이지안 모형에서 \(\theta\)의 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 \(t_c, \ldots, t_{c-1}\)에 대하여 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 순서통 계량의 분포를 설정할 수 있다. 즉 분별점의 무정보 사전분포는
\[
f(t_2, \ldots, t_{c-1}) = (c - 2)! \left( \frac{1}{t_{\text{max}}} \right)^{c-2} 1(0 < t_2 < \cdots < t_{c-1} < t_{\text{max}})
\]
와 같다.
\(\theta\)와 \(t\)가 서로 독립임을 가정할 때, \(y\) 조건부 \(\theta, t, u\)의 결합확률분포는
\[
f(\theta, t, u|y) \propto f(\theta)f(t)f(u|\theta, t)p(y|\theta, t, u) = f(\theta)f(t)f(u|\theta)p(y|t, u) \tag{2.6}
\]
인 관계를 갖는다. 이 식에서 \(f(t)\)와 \(f(y|\cdot, u)\)는 \(\theta\)를 포함하고 있지 않으므로 \(\theta\)의 완전 조건부 분포는
\[
f(\theta|t, u, y) \propto f(\theta)f(u|\theta) \propto f(u|\theta)
\]
가 성립되어
\[
\theta|t, u, y \sim N((X'X)^{-1}X'u, (X'X)^{-1})
\]
임을 알 수 있으며 다음과 같은 정리가 유도된다.
정리 2.1 \( \theta, j = 1, 2, \ldots, p \)의 완전조건부 분포는
\[
\theta_j | \theta_{-j}, t, u, y \sim N((x'_j x_j)^{-1} x'_j (u - X_{-j} \theta_{-j}),(x'_j x_j)^{-1})
\] 
(2.7)
와 같다.

정리 2.1에서 \( \theta_{-j} \)와 \( X_{-j} \)는 각각 \( \theta \)와 \( X \)에서 \( j \)-번째 요소와 \( j \)-번째 열을 제외한 나머지를 나타내고, \( x_j \)는 \( X \)의 \( j \)-번째 열이다. 이하 "-" 절차는 벡터 또는 행렬에서 해당 요소 또는 벡터를 제외한 나머지 부분을 나타내기로 한다.

마지막으로 분산값 \( t \)의 완전 조건분포는 (2.6)에서
\[
f(t|\theta, u, y) \propto f(t)p(y|t, u) \propto p(y|t, u)
\]
이 성립되는데, (2.4)를 \( t \)의 함수로 본다면 다음과 같은 정리가 성립된다.

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

\[
t_{\ell} | t_{\ell-1}, \theta, u, y \sim U(\{ \max \{ \text{max}(u|Y_i = \ell), t_{\ell-1} \}, \min \{ \text{min}(u|Y_i = \ell + 1), t_{\ell+1} \} \}) 
\] 
(2.8)

이제 (2.3), (2.7), (2.8)로부터 반복적으로 심스분포를 추출한다. 초기의 심스분포는 안정화되어 있지 않으므로 초기에 얻어진 심스분포는 사용하지 않는다. 일정 횟수의 반복 이후에 얻어진 심스분포를 \( Z_1, Z_2, \ldots \) 하고 하자. 일반적으로 \( Z_i \)들은 매우 강한 양의 상관관계를 갖게 되는데, 실제 모수 추정에는 자기상관이 0이 되는 시차 간격을 두고 얻어진 심스분포를 이용한다. 즉, 시차 간격을 \( l \)이라고 하면 사후 평균의 추정에 사용되는 심스분포는 \( X_i = Z_{i+l}, i = 1, 2, \ldots, m \)이다. 이렇게 얻어진 심스분포를 이용하여 사후평균과 분산은 각각
\[
\hat{\mu}_m = \frac{1}{m} \sum_{i=1}^{m} X_i \\
\hat{\sigma}_m^2 = \frac{1}{m} \sum_{i=1}^{m} (X_i - \hat{\mu}_m)^2
\]
으로 추정될 수 있다. 실제 추정에 있어서는 추정값이 실제값에 수렴하기 위한 표본크기 \( m \)의 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 이전변수의 경우는 해당되지 않는다. 즉, 이전변수의 경우 하나의 분산점 \( t_1 = 0 \) 만이 필요하므로 분산점에 대한 사전분포를 요구하지 않는다. 또 이 경우 (2.3)는
\[
f(u_i|\theta, y) = \begin{cases} 
\phi(x'_i \theta, 1) \mathbf{1}(u_i > 0), & Y_i = 1 \text{일 때} \\
\phi(x'_i \theta, 1) / (1 - \phi(x'_i \theta, 1)) \mathbf{1}(u_i \leq 0), & Y_i = 0 \text{일 때}
\end{cases}
\] 
(2.9)
와 같이 수정되어야 하며, 심스분포는 (2.7)와 (2.9)로부터 얻어지게 된다.
3. 사례분석

3.1. Morz 데이터

Morz (1987)은 기혼 여성의 직업을 갖는 것인지에 대한 의사결정에 영향을 미치는 여려 가지 사회적 환경 요인들에 대해 연구를 하였다. 이러한 연구의 일환으로 그는 백인 기혼 여성 753명에 대해 직업을 갖는 것인지에 대한 의사결정 나타내는 변수 infr (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 변수들로 측정하였다.

Morz의 데이터는 여러 논문에서 인용된 것으로 Lee와 Huh (2003)은 기혼 여성의 직업 참여에 대한 의사결정모형으로 종속 변수 infr에 대해 nwifeinc, edu, expr, exprsq, age, kidslt6, kidsge6의 7개 독립변수를 갖는 로지스틱 회귀모형을 고려하였다. 여기서 nwifeinc와 exprsq는 각각

\[
\text{nwifeinc} = \frac{(\text{faminc} - \text{wage} \times \text{hours})}{1000}
\]
\[
\text{exprsq} = \text{expr} \times \text{expr}
\]

와 같이 산출된 변수들로서 특히 nwifeinc는 총 가계소득 중에서 여성의 소득을 따로 산출할 것이다. 이정계 설정된 로지스틱 회귀모형의 적합 결과는 표 3.1과 같다. 적합된 회귀모형에서 절편과 kidsge6은 유의하지 않은 것으로 나타났으며, 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 4000번의 검스표본의 표본평균으로부터 구하여진 결과이다. 따라서 추정값을 구하기 위해 총 42000회의 반복을 실시하였다. 또 검스표본의 수렴을 확인하기 위해 비교적 큰 분산을 갖는 kidsge6의

| Coefficients: | Estimate | Std. Error | z value | Pr(>|z|) | Confusion matrix |
|---------------|----------|------------|---------|----------|-----------------|
| (Intercept)   | 0.425452 | 0.860365   | 0.495   | 0.62095  |                 |
| nwifeinc      | -0.021345| 0.008421   | -2.535  | 0.01125  |                 |
| educ          | 0.221170 | 0.043439   | 5.091   | 3.55e-07 |                 |
| expr          | 0.205670 | 0.032057   | 6.422   | 1.34e-10 |                 |
| exprsq        | -0.003154| 0.001016   | -3.104  | 0.00191  |                 |
| age           | -0.088024| 0.014573   | -6.400  | 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: 819.6303

<table>
<thead>
<tr>
<th>Frequency</th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted</td>
<td>---------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>207</td>
<td>207</td>
</tr>
<tr>
<td>t</td>
<td>1</td>
<td>34</td>
<td>34</td>
</tr>
<tr>
<td>a</td>
<td>1</td>
<td>181</td>
<td>181</td>
</tr>
<tr>
<td>l</td>
<td>1</td>
<td>42</td>
<td>42</td>
</tr>
</tbody>
</table>

Total 288 465 753
표 3.2: Morz 데이터의 계층적 베이지안 회귀모형

| Coefficients: | Estimate  | Std. Error | z value | Pr(>|z|)  | Confusion matrix |
|--------------|-----------|------------|---------|-----------|------------------|
| 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 *** | Actual          |
| expr         | 0.117072400 | 0.018558 | 6.309   | 2.82e-10 *** | Predicted       |
| exprsq       | -0.001702628 | 0.000597 | -2.853  | 0.00433 ** |                  |
| age          | -0.052632050 | 0.008389 | -6.274  | 3.51e-10 *** |                  |
| kids1t6      | -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

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

회귀계수에 대해 2회 김스 체인을 시행하였으며 두 개의 체인에서 구한 경험적 분포함수는 그림 3.1과 같이 나타나 두 개의 체인이 수렴하고 있음을 알 수 있다. 다른 회귀계수에서도
모두 갑스표본의 수렴을 확인할 수 있었다.
베이지안 모형에서 얻어진 회귀계수의 추정값을 로지스틱 회귀모형에서 구한 회귀계수의 추정값과 직접 비교하는 것은 근본하지만 두 모형에서 모두 같은 부호로 추정되어 각 변수들이 lnlf에 대해 같은 방향으로 영향을 미치고 있다. 특히 베이지안 모형에서 AIC 정보량은 819.0034로서 로지스틱 회귀모형과 비교하여 근소하다마 선호되는 모형으로 판단할 수 있다. 이 결과는 베이지안 모형의 오분류율에서도 확인할 수 있다. 즉 베이지안 모형의 오분류율은 약 25.50%로 로지스틱 회귀모형과 비교하여 근소하게나마 작은 오분류율을 나타내고 있다. 그러므로 Morz 데이터에서 기존 여성의 직접 참여에 대한 의사결정모형으로 베이지안 분세점 모형이 로지스틱 회귀모형보다는 선호된다고 하겠다.

3.2. 와인 데이터
와인 데이터는 이태리의 한 지역에서 수확된 세 품종의 와인의 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 잘 알려진 것으로 분별분석에서 분류자(classifier)의 성능을 평가하는데 많이 이용되고 있다. 현재까지 알려진 바에 의하면, 적 RDA 방법에 의해서만 100% 정확하게 품종을 구별할 수 있었다고 한다. 그 외에 LDA와 QDA는 각각 98.9%와 99.4%의 성능을 갖는다고 한다.
와인데이터의 클래스번수는 품종을 나타내는 병목적도의 성질을 갖는 변수이기 때문에

표 3.3: Wine 데이터의 순위로짓모형

<table>
<thead>
<tr>
<th>Coefficients:</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter1</td>
<td>-52.2628</td>
<td>71.0927</td>
<td>0.5404</td>
<td>0.4623</td>
</tr>
<tr>
<td>inter2</td>
<td>73.0859</td>
<td>101.4</td>
<td>0.8199</td>
<td>0.4709</td>
</tr>
<tr>
<td>X1</td>
<td>7.5959</td>
<td>8.5299</td>
<td>0.7930</td>
<td>0.3732</td>
</tr>
<tr>
<td>X2</td>
<td>3.4277</td>
<td>7.3637</td>
<td>0.2167</td>
<td>0.6416</td>
</tr>
<tr>
<td>X3</td>
<td>7.9879</td>
<td>10.2550</td>
<td>0.6102</td>
<td>0.4347</td>
</tr>
<tr>
<td>X4</td>
<td>-31.4736</td>
<td>44.2822</td>
<td>0.5052</td>
<td>0.4772</td>
</tr>
<tr>
<td>X5</td>
<td>1.1667</td>
<td>1.4877</td>
<td>0.6151</td>
<td>0.4329</td>
</tr>
<tr>
<td>X6</td>
<td>-8.3937</td>
<td>14.8717</td>
<td>0.3186</td>
<td>0.5725</td>
</tr>
<tr>
<td>X7</td>
<td>52.0534</td>
<td>77.0694</td>
<td>0.4562</td>
<td>0.4994</td>
</tr>
<tr>
<td>X8</td>
<td>11.2979</td>
<td>16.0164</td>
<td>0.4976</td>
<td>0.4806</td>
</tr>
<tr>
<td>X9</td>
<td>-8.9861</td>
<td>16.1263</td>
<td>0.3105</td>
<td>0.5774</td>
</tr>
<tr>
<td>X10</td>
<td>-33.8693</td>
<td>50.0522</td>
<td>0.4579</td>
<td>0.4986</td>
</tr>
<tr>
<td>X11</td>
<td>10.3689</td>
<td>16.9069</td>
<td>0.3762</td>
<td>0.5397</td>
</tr>
<tr>
<td>X12</td>
<td>19.9907</td>
<td>26.5423</td>
<td>0.5673</td>
<td>0.4614</td>
</tr>
<tr>
<td>X13</td>
<td>48.4759</td>
<td>75.3111</td>
<td>0.4143</td>
<td>0.5198</td>
</tr>
</tbody>
</table>

Confusion matrix

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Frequency</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td></td>
<td>a</td>
<td>b</td>
<td>c</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>58</td>
<td>1</td>
<td>0</td>
<td>59</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>0</td>
<td>70</td>
<td>1</td>
<td>71</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>0</td>
<td>46</td>
<td>1</td>
<td>48</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>58</td>
<td>73</td>
<td>47</td>
<td>178</td>
</tr>
</tbody>
</table>

\[-2\text{Log L : 18.997}]
\[\text{AIC : 48.997} \]
표 3.4: Wine 데이터의 베이지안 분계점 회귀모형

<table>
<thead>
<tr>
<th>Coefficients:</th>
<th>Thresholds = 0, 50.90356</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>Std. Error</td>
</tr>
<tr>
<td>Inter</td>
<td>21.0360300</td>
</tr>
<tr>
<td>X1</td>
<td>-3.4150420</td>
</tr>
<tr>
<td>X2</td>
<td>-1.2410890</td>
</tr>
<tr>
<td>X3</td>
<td>-3.2132480</td>
</tr>
<tr>
<td>X4</td>
<td>12.4261500</td>
</tr>
<tr>
<td>X5</td>
<td>-0.4667664</td>
</tr>
<tr>
<td>X6</td>
<td>3.3799400</td>
</tr>
<tr>
<td>X7</td>
<td>-20.1032390</td>
</tr>
<tr>
<td>X8</td>
<td>-6.2143320</td>
</tr>
<tr>
<td>X9</td>
<td>2.8871830</td>
</tr>
<tr>
<td>X10</td>
<td>12.2466300</td>
</tr>
<tr>
<td>X11</td>
<td>-4.3306880</td>
</tr>
<tr>
<td>X12</td>
<td>-8.5971400</td>
</tr>
<tr>
<td>X13</td>
<td>-18.2947100</td>
</tr>
</tbody>
</table>

Predicted Frequencies: 1 | 2 | 3 | Total
A |-------------------------------
1 | 58 | 1 | 0 | 59

-2 Log L : 20.55785
AIC: 48.55785

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

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

이상의 두 예제에서 계층적 베이지안 모형은 이전 또는 순위를 갖는 범주형 데이터에 대해 로지스틱 회귀모형보다 우수한 적합 능력을 갖고 있을 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 의한 측정된 범주형 자료를 분석하기 위하여 베이지안 모형을 적용하였으며, 신형회귀모형을 따르는 잔대변수와 분계점 개념을 설정한 후 무정 보 사전분포를 이용한 완전 조건부 분포를 유도하고 깊이표본에 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 모형은 순위로짓모형과 같이 범주형 데이터의 적합에 사용하여도 좋을 것으로 보인다.

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

참고문헌


[ 2004년 7월 검수, 2004년 11월 체택 ]
A Bayesian Threshold Model for Ordered Categorical Traits*

Byongsu Choi¹)  Seung-Chun Lee²)

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