

베이지안 분계점 모형에 의한 순서 범주형 변수의 대체*

이승천¹⁾

요 약

대개의 표본조사에서 무응답은 필연적으로 발생되고 있고, 직접 표본조사에 참가하지 않은 데이터의 사용자는 무응답의 원인을 알 수 없는 것이 일반적이므로 데이터 분석에 어려움을 갖는다. 또 대부분의 통계분석 방법은 무응답을 전제하지 않고 있어 무응답이 있는 항목은 데이터 분석의 걸림돌이 된다고 하겠다. 최근 무응답에 대해 대체법이 하나의 표준적인 처리 방법이 되고 있어 현재까지 대체법에 대한 많은 연구가 있었으나 대부분의 대체법은 정규성 등을 가정한 연속형 변수의 대체법에 대한 것이었다. 그러나 표본조사에서 많은 중요한 항목들이 순서범주에 의해 측정되는 경우가 많으므로 범주형 변수의 대체법에 대한 연구가 필요하며, 본 연구에서는 보조변수가 있는 경우 Bayesian 모형에 의한 순서범주형 항목의 대체법에 대해 알아본다.

주요용어: 대체법, 순위범주형변수, 베이지안 분계점 모형, 순위 로짓회귀모형

1. 서론

표본조사에서는 표본설계의 효율성 여부와는 관계없이 필연적으로 무응답이 존재하기 마련이고, 데이터의 분석에서 무응답은 자료의 일부분이 없어져 일반적인 자료구조의 변형을 초래하여 데이터 분석을 어렵게 만든다. 예를 들어 상관분석에서 각 표본단위에서 측정된 두 항목 중 하나에서 무응답이 발생하게 되면 해당 표본단위에서는 상관관계를 측정하기 곤란하여 측정된 항목의 값도 사용하지 않는 일이 발생하게 된다. 이 경우 정보의 손실을 예상할 수 있으므로 무응답 항목에 대한 대체를 고려하게 된다. 무응답의 원인을 알 수 없는 일반적인 데이터의 사용자에게는 특히 대체법의 필요성이 부각된다고 하겠다.

최근 대체법에 대해 Rubin (1987), Meeden (2000), Paddock (2002), Lee (2004) 등 일일이 나열할 수 없을 정도로 많은 연구가 있어 왔다. 그러나 이 결과들은 정규성 가정에 의한 연속형 변수들의 대체법에 대한 것으로 범주형 항목의 대체를 위해 사용하기는 곤란하므로 범주형 항목의 대체법에 대한 연구가 필요하다고 하겠다.

일반적으로 빈도학적 사고에 의한 무응답의 처리는 매우 어렵다. 그러나 베이지안의 입장에서 무응답은 이론적으로 매우 직관적이다(Meeden, 2000.) 즉, 베이지안적 사고에 의하면, 무응답은 추정하여야 할 하나의 모수(parameter)로서 대체법은 일반적인 모수 추정과 다를 바가 없다. 한편 빈도학파에서 대체법에 대한 방법론은 정리되지 않았지만 대체는 예측방법에 의존하고 있다. 예를 들어 연속형 변수와 범주형 변수들의 무응답 대체값으로

* 이 논문은 2005년도 한신대학교 학술연구비 지원에 의하여 연구되었음

1) (447-791) 경기도 오산시 양산동 411, 한신대학교 정보통계학과, 교수

E-mail: seung@hanshin.ac.kr

많이 사용되고 있는 표본평균과 최빈값은 빈도학파적 사고에 의한 대체로 파악할 수 있다. 표본평균의 경우는 후에 베이지안의 대체법으로도 알려져 있지만 대체에 대한 사고는 두 학파가 일치하지 않는다. 즉, iid로 가정되는 표본에서는 독립성으로 인하여 관측된 데이터와 관측되지 않은 데이터에 대한 연결고리가 없다. 다시 말하면 관측된 데이터는 무응답 데이터에 대한 아무런 정보를 갖고 있지 못하다. 단지 두 종류의 데이터가 같은 분포를 갖는다는 가정에 따라 관측된 데이터로부터 모집단의 평균을 표본평균으로 추정하고 이를 이용하여 모형에 따라 무응답 항목의 값 $y_i = \mu + \epsilon_i$ 를 표본평균으로 예측하여 대체값을 구한다. 한편 베이지안 입장에서는 두 종류의 표본은 사전분포에 의해 연결되어 있고, 무응답을 하나의 모수로 파악하여 무응답을 추정한다. 이 경우, 손실함수를 평균제곱오차를 가정하면 무응답 y_i 의 베이지안 추정량은

$$E(y_i | \text{관측데이터}) \quad (1.1)$$

이며, 이 값을 대체값으로 사용하게 된다.

Ghosh와 Meeden (1998)은 무정보 사전분포를 가정하여 구한 “Polya posterior”에서 (1.1)이 표본평균이라는 것을 밝혀냈다. 대체법에 대한 두 학파의 입장이 어떠한 실재에 있어서는 어느 대체법이 효율성이 있는 지가 문제일 것이다. 특히 표본평균에 의한 대체는 두 학파 모두에게서 인정될 수 있는 대체값이기는 하지만 무응답이 여럿이 있는 경우, 하나의 값에 의한 대체는 대체 이후 데이터 분석에서 통계적 추론의 정도를 과소평가하게 되는 문제가 발생한다. 이에 대한 대응 방법으로 대체에 의한 데이터에 대해 다른 가중치를 적용한다거나 (Nusser, Carriquiry, Dodd, and Fuller, 1996,) jackknife 유형의 방법을 이용하여 분산을 추정 (Rao and Shao, 1992) 하는 등 여러 가지 방법들이 제안되었지만 이러한 방법들은 대체에 의해 완전한 형태의 데이터만을 볼 수 있어 어느 데이터가 대체에 의한 것인지를 알 수 없는 경우에는 적용될 수 없는 것으로, 현재에 있어서는 다중대체법이 일반적인 방법이라고 하겠다.

Meeden (2000)은 연속형 변수에 대한 흥미로운 베이지안 다중대체법을 소개하였고, Lee (2004)도 매우 효율적인 다중대체법을 제안하였다. 여기서 사용된 접근 방식은 직관적이고 베이지안 및 빈도학과 모두에게 만족스럽지만 범주형 변수의 대체에 적용시키기는 여러 가지 어려움이 있다. 특히 보조변수가 없을 경우 만족스러운 다중대체 방법을 찾기는 어렵다고 판단된다. 그러므로 본고에서는 보조변수가 있는 경우에 순서범주형 변수의 대체 방법에 대해 알아 보려고 한다.

앞에서도 언급한 바와 같이 대체는 예측 또는 모수 추정의 문제로 파악될 수 있다. 또 Rubin (1987)은 모형에 기반을 둔 통계분석의 중요성을 언급하였다. 즉, 적절한 모형에 기반을 둔 추정 또는 예측에 의해서만이 효과적인 대체를 기대할 수 있다. 순위로짓모형은 순서범주형 데이터의 일반적인 통계적 모형으로 이미 잘 알려져 있다. 그러므로 순위로짓모형에 위한 예측 방법은 순서범주형 변수의 대체법으로 효율성을 기대할 수 있다고 하겠다. 한편 베이지안 사고에 의한 대체법 모형으로 Lee와 Lee (2002), 최병수와 이승천(2005)에서 언급된 베이지안 분계점 모형과 이와 유사한 베이지안 모형으로 Chib (2000), 또는 Chen과 Dey (2000)의 모형을 염두에 둘 수 있다. 그러나 뒷에 언급된 두 개의 모형은 분산성분을 포함하는 선형모형으로 일반적인 순위로짓모형의 가정과는 다르므로 여기에서는 배제

하기로 한다. 이하 2 절에서는 무응답을 포함한 데이터의 베이저안 분계점 모형에 대해 살펴보고, 3 절에서 두 모형에서 구한 대체법에 대해 모의실험을 통해 효율성을 비교하기로 한다.

2. 무응답을 포함한 베이저안 분계점 모형

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

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

와 같이 결정된다고 가정한다. 이때 $-\infty = t_0 < t_1 < \dots < t_{\ell-1} < t_\ell = \infty$ 이다. 여기서 Y_i 들의 벡터 $\mathbf{y} = (Y_1, \dots, Y_n)$ 는 무응답을 포함하고 있는데 마지막 n_m 개의 데이터가 무응답이라고 가정하자. 그러므로 처음 $n^o = n - n_m$ 개의 \mathbf{y} 요소는 관측된 것으로 가정하고, 관측된 요소들의 벡터를 \mathbf{y}^o 라고 나타낸다. 또한 잠재 확률벡터 $\mathbf{u} = (U_1, \dots, U_n)$ 는 $\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\epsilon}$ 가 각각 $n \times p$ 계측행렬, $p \times 1$ 회귀계수벡터와 $n \times 1$ 오차벡터라고 할 때,

$$\mathbf{u} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \quad (2.1)$$

와 같은 선형회귀모형을 가정한다.

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

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

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 않게 되므로 여기에서는 오차항 분산값을 1로 설정하였다.

(2.2)의 모형에서 \mathbf{x}_i 를 \mathbf{X} 의 i 번째 행이라고 하면 Y_i 의 조건부 확률함수는 $\phi(x)$ 와 $\Phi(x)$ 를 표준정규분포의 확률밀도함수와 분포함수라고 할 때,

$$\begin{aligned} \Pr[Y_i = \ell | \boldsymbol{\theta}, \mathbf{t}] &= \Pr[t_{\ell-1} < U_i \leq t_\ell | \boldsymbol{\theta}, \mathbf{t}] = \int_{t_{\ell-1} - \mathbf{x}'_i \boldsymbol{\theta}}^{t_\ell - \mathbf{x}'_i \boldsymbol{\theta}} \phi(x) dx \\ &= \Phi(t_\ell - \mathbf{x}'_i \boldsymbol{\theta}) - \Phi(t_{\ell-1} - \mathbf{x}'_i \boldsymbol{\theta}) \end{aligned}$$

와 같이 유도된다. 또한 U_i 의 완전조건부 분포는 Y_i 가 무응답인지 또는 그렇지 않은지에 따라 다음과 같은 확률밀도함수를 갖게 된다.

$$f(u_i|y^o, \theta, t) = \begin{cases} \frac{\phi_{U_i}(x'_i, \theta, 1)}{\Phi(t_\ell - x'_i, \theta) - \Phi(t_{\ell-1} - x'_i, \theta)} \mathbf{1}(u_i \in A_\ell), & Y_i \in y^o, Y_i = \ell \\ \phi_{U_i}(x'_i, \theta, 1), & Y_i \notin y^o \end{cases} \quad (2.3)$$

여기서 $\phi(\mu, \sigma)$ 는 평균과 분산이 각각 μ 와 σ^2 인 정규분포의 확률밀도함수를 나타내고, $\mathbf{1}(x \in A)$ 은 표시함수(indicator function), $A_\ell = (t_{\ell-1}, t_\ell]$ 이다. 이때 y^o 의 완전조건부 확률분포는

$$f(y^o|u, t, \theta) = \prod_{i=1}^{n^o} \left\{ \sum_{\ell=1}^k \mathbf{1}(u_i \in A_\ell) \mathbf{1}(y_i = \ell) \right\} \quad (2.4)$$

와 같은 퇴화분포가 된다.

전통적 회귀모형에서 회귀계수가 모수인자임을 가정하다면 베이지안 모형에서 θ 의 무정보 사전분포는

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

와 같이 설정할 수 있다. 한편 (2.1)가 절편을 포함한 회귀모형이라고 하면 식별성을 위하여 일반적으로 $t_1 = 0$ 이 주어진다. 그러므로 단지 t_2, \dots, t_{c-1} 에 대한 사전분포만을 필요로 하게 되는데, 이들은 $0 = t_1 < t_2 < \dots < t_{c-1}$ 와 같은 순서를 갖고 있어 서로 독립이 아님을 알 수 있으며 최병수와 이승천(2005)에서와 같이 무정보 사전분포로서 균일분포에서 추출된 순서통계량의 분포를 설정할 수 있다. 즉, 분계점의 무정보 사전분포는

$$f(t) = (k-2)! \left(\frac{1}{t_m} \right)^{k-2} \mathbf{1}(t \in T)$$

와 같이 설정한다. 여기서 $T = \{(t_1, \dots, t_{c-1}); 0 = t_1 < t_2 < \dots < t_{c-1} < t_m\}$ 이며 t_m 은 임의의 큰 값이다.

t 와 θ 가 서로 독립임을 가정하면, y^o 가 주어졌을 때, u, t, θ 의 결합확률분포는

$$\begin{aligned} f(u, t, \theta|y^o) &\propto f(\theta)f(t)f(u|\theta, t)f(y^o|u, t, \theta) \\ &= f(\theta)f(t)f(u|\theta)f(y^o|u, t) \end{aligned} \quad (2.5)$$

와 같이 구할 수 있는데, 이 식에서 $f(t)$ 와 $f(y^o|u, t)$ 는 θ 를 포함하고 있지 않으므로 θ 의 완전조건부 분포는

$$f(\theta|u, t, y^o) \propto \exp \left[-\frac{1}{2} (\theta - (X'X)^{-1}X'U)' (X'X) (\theta - (X'X)^{-1}X'U) \right]$$

가 성립되어

$$\theta|u, t, y^o \sim N((X'X)^{-1}X'U, (X'X)^{-1})$$

임을 알 수 있다. 그러므로 $\theta_j, j = 1, 2, \dots, p$ 의 완전조건부 분포에 대해 다음과 같은 결론을 얻을 수 있다.

$$\theta_j|\theta_{-j}, u, t, y^o \sim N((x'_j x_j)^{-1}x'_j(u - X_{-j}\theta_{-j}), (x'_j x_j)^{-1}) \quad (2.6)$$

θ_{-j} 와 \mathbf{X}_{-j} 는 각각 θ 와 \mathbf{X} 에서 j 번째 요소와 j 번째 열을 제외한 나머지를 나타내고, \mathbf{x}_{-j} 는 \mathbf{X} 의 j 번째 열을 나타낸다. 이하 “-” 첨자는 벡터 또는 행렬에서 해당 요소 또는 벡터를 제외한 나머지 부분을 나타내기로 한다.

마지막으로 분계점 \mathbf{t} 의 완전 조건부 분포는 (2.5)에서 \mathbf{t} 을 포함하고 있는 항은 $f(\mathbf{t})$ 와 $f(\mathbf{y}^o|\mathbf{u}, \mathbf{t})$ 뿐이므로 다음과 같은 식이 성립된다.

$$\begin{aligned} f(\mathbf{t}|\theta, \mathbf{u}, \mathbf{y}^o) &\propto f(\mathbf{t})f(\mathbf{y}^o|\mathbf{u}, \mathbf{t}) \\ &\propto \mathbf{1}(\mathbf{t} \in T)f(\mathbf{y}^o|\mathbf{u}, \mathbf{t}) \end{aligned}$$

그러므로 $t_\ell, \ell = 2, 3, \dots, c-1$ 의 완전조건부 분포는 구간 $(t_\ell^{\max}, t_{\ell+1}^{\min})$ 에서 균일분포를 따르게 된다. 즉,

$$f(t_\ell|\mathbf{t}_{-\ell}, \theta, \mathbf{u}, \mathbf{y}^o) = \frac{1}{t_{\ell+1}^{\min} - t_\ell^{\max}} \mathbf{1}(t_\ell^{\max} < t_\ell \leq t_{\ell+1}^{\min}) \quad (2.7)$$

이 된다. 단 t_ℓ^{\max} 과 $t_{\ell+1}^{\min}$ 는 $I_\ell = \{i : Y_i = \ell\}$ 이라고 할 때 각각

$$t_\ell^{\max} = \max \left\{ \max_{i \in I_\ell} \{u_i\}, t_{\ell-1} \right\}, \quad t_{\ell+1}^{\min} = \min \left\{ \min_{i \in I_{\ell+1}} \{u_i\}, t_{\ell+1} \right\}$$

와 같이 정의된다.

이제 (2.3), (2.6) 그리고 (2.7)으로부터 반복적으로 깃스표본을 추출한다. 초기의 깃스표본은 안정화되어 있지 않으므로 초기에 얻어진 깃스표본은 사용하지 않는다. 일정 횟수의 반복 이후에 얻어진 깃스표본들을 Z_1, Z_2, \dots 라고 하자. 일반적으로 Z_i 들은 매우 강한 양의 상관관계를 갖게 되는데, 실제 모수 추정에는 자기상관이 0이 되는 시차의 간격을 두고 얻어진 깃스표본을 이용하는 것이 일반적이다. 즉, 시차 간격을 r 이라고 하면 사후 평균의 추정에 사용되는 깃스표본은 $X_i = Z_{i \times r}, i = 1, 2, \dots, m$ 이 된다. 이렇게 얻어진 깃스표본을 이용하여 사후평균과 분산은 각각 X_i 들의 표본평균과 표본분산에 의해 추정된다. 실제 추정에 있어서는 추정값이 실제값에 수렴하기 위한 표본크기 m 의 값을 알아야 하는데, 이 문제에 대해서는 Raftery와 Lewis (1992)를 참조할 수 있다.

앞에서 언급한 바와 같이 무응답 Y_i 의 대체값을 구하기 위해서 $E(Y_i|\mathbf{y}^o)$ 를 구하여야 하지만 일반적으로 이 값은 Y_i 들의 모수공간 $\{1, 2, \dots, c\}$ 에 속하지 않는다. 즉, 행동공간(action space)과 모수공간이 일치하지 않는다. 일반적인 추정문제에 있어서는 이것이 큰 문제가 되지 않으나 대체에 있어서는 두 공간이 같아야 하므로 여기에서는 잠재변수 U_i , 분계점 \mathbf{t} 와 Y_i 의 관계를 이용하여 대체값을 구한다. 즉, 깃스표본의 의해 추정된 잠재변수와 분계점의 사후평균을 각각 $\tilde{U}_i, \tilde{t}_\ell$ 이라고 하면 무응답 Y_i 의 대체값은

$$\tilde{Y}_i = \ell \quad \text{if } \tilde{t}_{\ell-1} < \tilde{U}_i \leq \tilde{t}_\ell$$

에 의해 구하게 된다. 이 과정은 범주확률(class probability)이 가장 크게 추정된 범주로 할당하는 것과 같다.

(2.3)의 완전 조건부 분포는 범주의 수가 3 이상인 경우에만 해당되는 것으로 이진변수의 경우는 여기에 해당되지 않는다. 즉, 이진변수의 경우 하나의 분계점 $t_1 = 0$ 만이 필요

하므로 분계점에 대한 사전분포를 요구하지 않는다. 또 이 경우 (2.3)은

$$f(u_i|y^o, \theta) = \begin{cases} \frac{\phi_{U_i}(x'_i, \theta, 1)}{\Phi(x'_i, \theta)} 1(u_i > 0), & Y_i \in y^o, Y_i = 1 \\ \frac{\phi_{U_i}(x'_i, \theta, 1)}{1 - \Phi(x'_i, \theta)} 1(u_i \leq 0), & Y_i \in y^o, Y_i = 0 \\ \phi_{U_i}(x'_i, \theta, 1), & Y_i \notin y^o \end{cases} \quad (2.8)$$

와 같이 수정되어야 하며, 깁스표본은 (2.6)과 (2.8)으로부터 얻어지게 된다.

3. 대체법에 대한 모의실험

모의실험은 유한모집단에서 추출된 표본에서 무응답을 가정하여 순위로짓모형, 순위프로빗모형 및 베이지안분계점 모형에 의해 무응답을 대체한 후, 각 범주의 확률을 추정하여 추정값의 효율성을 모의실험하였다. 이를 위해 이변량 정규분포에서 상관계수의 값이 각각 $\rho = 0.95, 0.90, 0.8, 0.70$ 일때, 500개의 난수를 발생하여 4개의 가상 유한모집단 (U, X)를 구하였다. 사용된 이변량 정규분포의 평균은 (0,0)이고 분산은 (2,3)이다. 이렇게 얻어진 모집단에서 이진변수 Y 는 임의적인 분계값에 의해

$$Y_i = \begin{cases} 1, & U_i > 0.5 \\ 0, & U_i \leq 0.5 \end{cases}$$

와 같이 구하였고, Y_i 가 구해진 후 U_i 는 모두 삭제하였다. 즉, 4개의 유한모집단은 이진변수 Y 와 보조변수 X 로 구성된다. 이렇게 얻어진 4개의 모집단에서 각각 50개의 표본을 추출하고 (All 으로 표현) 이중 5개 또는 10개의 표본을 임의로 선택하여 Y 의 값을 무응답으로 처리 (Obser로 표시) 한 후, 각각 로짓회귀모형, 프로빗 회귀모형 및 베이지안 분계점 모형에 의해 대체값을 구하였다. 즉, 하나의 모집단에서 추출된 50개의 표본으로부터 All, Obser, 그리고 대체값에 의한 3개의 표본, logis, probit, Gibbs, 총 5개의 표본이 구해지게 된다. 이렇게 얻어진 5개의 표본에서 각각 표본비율에 의해 $p = \Pr[Y = 1]$ 를 추정한다. 이러한 과정을 1000회에 걸쳐 반복한 후, 추정값과 모집단에서 구한 값을 이용하여 평균제곱오차 등을 구하였다. 그 결과는 표 3.1과 같다.

표 3.1에 나타난 결과를 요약하면 베이지안 분계점 모형에 의한 대체를 하였을 때, 다른 두 모형에 의해 대체하였을 경우와 비교하여 모비율의 추정에서 일률적으로 작은 평균제곱오차값을 갖고 있으며, $n_m = 10, \rho = 0.95$ 인 경우를 제외하면 베이지안 분계점 모형은 모두 가장 작은 평균절대오차값을 갖고 있었다. 또 상관계수가 0.90 이상인 경우 세 가지 대체법은 모두 대체를 하지 않고 무응답을 버린 경우보다 평균제곱오차 및 평균절대오차의 기준에서 보다 좋은 추정을 할 수 있다는 것으로 보여주고 있다. 그러나 보조변수와 잠재변수의 상관계수가 그리 크지 않은 경우 ($\rho \leq 0.8$)에는 대체법들이 무응답을 버린 경우보다 오히려 추정의 정도를 낮추고 있음을 알 수 있었다. 한편 대체법에 의해 구하여진 신뢰구간의 포함확률은 신뢰계수보다 작은 값을 갖고 있어 적절한 길이의 신뢰구간을 구하여 주지 못하고 있다. 이는 대체를 하였을 경우, 분산이 과소추정되어 신뢰구간이 길이가 짧아진 것으로 대체법의 일반적인 특징인 분산의 과소추정 문제가 발생되고 있음을 알 수 있다.

표 3.1: 1000회 반복에서 구한 모비율의 추정값, 95%신뢰구간의 포함비율, 평균제곱오차 $\times 1000$, 평균절대값오차 $\times 100$

모집단 유형	표본 유형	포함 확률	평균 넓이	평균	MSE	MAE	포함 확률	평균 넓이	평균	MSE	MAE
$n_m = 5$							$n_m = 10$				
$\rho = 0.95$ $p = 0.338$	All	0.942	0.2488	0.3385	4.0672	5.1536	0.942	0.2488	0.3385	4.0672	5.1536
	Obser	0.954	0.2636	0.3388	4.6463	5.4818	0.937	0.2808	0.3384	5.5906	6.0716
	Gibbs	0.926	0.2484	0.3376	4.3992	5.3440	0.902	0.2481	0.3375	4.8926	5.6560
	Logis	0.923	0.2483	0.3372	4.4506	5.3704	0.896	0.2478	0.3364	4.9526	5.6728
	Probit	0.925	0.2483	0.3373	4.4301	5.3584	0.902	0.2479	0.3369	4.9101	5.6528
$\rho = 0.90$ $p = 0.408$	All	0.956	0.2589	0.4090	4.0848	5.0832	0.956	0.2589	0.4090	40.848	5.0832
	Obser	0.953	0.2744	0.4089	4.5351	5.2985	0.959	0.2926	0.4091	52.359	5.6816
	Gibbs	0.950	0.2587	0.4078	4.2618	5.1760	0.935	0.2584	0.4078	46.023	5.3700
	Logis	0.950	0.2586	0.4076	4.2945	5.1916	0.935	0.2584	0.4075	46.372	5.4020
	Probit	0.950	0.2586	0.4076	4.2945	5.1972	0.935	0.2584	0.4076	46.176	5.3880
$\rho = 0.80$ $p = 0.382$	All	0.934	0.2551	0.3795	4.4586	5.3708	0.934	0.2551	0.3795	44.586	5.3708
	Obser	0.939	0.2705	0.3799	4.8835	5.6495	0.935	0.2886	0.3807	54.549	5.9270
	Gibbs	0.907	0.2544	0.3766	5.0243	5.7096	0.882	0.2535	0.3727	57.491	6.0520
	Logis	0.907	0.2544	0.3763	5.0443	5.7168	0.884	0.2534	0.3720	57.664	6.0764
	Probit	0.906	0.2544	0.3763	5.0478	5.7184	0.881	0.2534	0.3722	57.858	6.0768
$\rho = 0.70$ $p = 0.338$	All	0.956	0.2560	0.3835	3.8418	5.0128	0.956	0.2561	0.3835	38.418	5.0128
	Obser	0.964	0.2714	0.3834	4.2900	5.3191	0.947	0.2893	0.3832	50.593	5.7284
	Gibbs	0.924	0.2553	0.3807	4.4878	5.3728	0.892	0.2542	0.3760	53.782	5.8924
	Logis	0.923	0.2553	0.3805	4.5087	5.3852	0.888	0.2540	0.3751	54.761	5.9532
	Probit	0.924	0.2553	0.3806	4.5040	5.3864	0.889	0.2540	0.3754	54.480	5.9384

두 번째 모의실험은 <http://www.ics.uci.edu/mlearn/MLRepository.html>에 수록된 와인 데이터를 이용하였다. 와인 데이터는 이탈리아의 한 지역에서 숙성된 세 품종의 와인에 대한 13 가지 화학적 성분과 품종을 나타내는 14개의 변수로 구성된 178개의 데이터로서 기계학습 분야에서는 판별분석에서 분류자(classifier)의 성능을 평가하는데 많이 이용되고 있다. 현재까지 알려진 바에 의하면 오직 RDA 방법에 의해서만 100% 정확하게 품종을 구별할 수 있었다고 한다. 와인데이터의 클래스변수는 품종을 나타내는 명목척도의 성질을 갖는 변수이기는 하지만 최병수와 이승천 (2005)에 의하면 순위로짓모형과 베이지안 분계점 모형이 모두 97.75%의 정분류율을 보이고 있어 클래스 변수는 품종에 따라 우열을 가릴 수 있는 순위척도의 성질을 갖고 있는 것으로 판단된다.

모의실험은 이전과 비슷한 방법에 의해 실행되었다. 즉, 178개의 데이터를 유한모집단으로 간주하여 이 모집단에서 50개의 표본을 추출한다(Total). 추출된 50개의 표본에서 5개 또는 10개를 무응답으로 처리하고 (Observed), 무응답은 순위로짓모형(Logistic)과 베이지

표 3.2: 1000회 반복에서 구한 범주비율의 추정값, 평균제곱오차 $\times 1000$, 평균절대값오차 $\times 100$

n_m	표본 유형	범주 1 ($p = 0.33146$)			범주 2 ($p = 0.39888$)			범주 3 ($p = 0.26966$)			총	총
		평균	MSE	MAE	평균	MSE	MAE	평균	MSE	MAE	MSE	MAE
5	Total	0.330	3.342	4.645	0.399	3.507	4.698	0.271	2.860	4.287	9.710	13.630
	Obser	0.330	3.810	4.936	0.399	4.131	5.084	0.272	3.308	4.581	11.249	14.600
	Gibbs	0.330	3.342	4.645	0.399	3.728	4.806	0.266	3.012	4.430	10.082	13.881
	Logis	0.332	3.477	4.712	0.395	3.729	4.879	0.273	2.967	4.369	10.173	13.959
10	Total	0.330	3.342	4.645	0.399	3.507	4.698	0.271	2.861	4.287	9.710	13.630
	Obser	0.331	4.436	5.325	0.399	4.677	5.419	0.270	3.868	4.998	12.981	15.742
	Gibbs	0.330	3.342	4.645	0.419	4.154	5.159	0.262	3.338	4.661	10.835	14.465
	Logis	0.335	3.603	4.783	0.389	4.141	5.142	0.276	3.210	4.524	10.954	14.448

안 분계점 모형 (Gibbs)에 의해 대체값을 구한다. 이렇게 구하여진 4 개의 표본에서 각각 클래스 확률을 추정한다. 이러한 과정을 1000회 반복하여 추정값들의 정도를 비교하였다. 그 결과는 표 3.2와 같다.

표에서 확인할 수 있듯이 순위로짓모형과 베이지안 분계점 모형에 의한 대체는 클래스 확률의 추정에 있어 무응답을 제외한 경우보다 정도 높은 추정을 가능하게 하였다. 베이지안 분계점 모형에 의한 대체는 순위로짓모형에 의한 대체보다 작은 총 오차제곱평균값을 갖고 있어 근소하게나마 선호되는 대체법이라고 할 수 있다. 또 두 모형은 모두 Obser보다 작은 총 평균제곱오차 및 총 평균절대오차 값을 가지고 있어 모두 적절한 대체라고 판단할 수 있다.

4. 결론

모의실험 결과 보조변수가 순위범주형 변수에 대해 어느 정도 설명력이 있을 경우 여러 회귀모형들에 의한 대체는 매우 유의한 것으로 판단되고 있으며 특히 베이지안 분계점 모형에 의한 대체는 모수추정에 있어 평균제곱오차 및 평균절대오차의 기준에서 다른 모형보다 우수한 대체법이 될 수 있다는 것을 살펴보았다. 또한 베이지안에 의한 대체는 예측에 의한 대체보다 이론적으로 매우 명확하여 선호될 수 있는 것으로 판단된다. 그러나 대부분의 대체법이 그러하듯이 분산의 과소추정문제가 발생되고 있어 이에 대한 해결 방법이 모색되어야 할 것이다.

참고문헌

- 최병수, 이승천 (2005). 순서범주형자료 분석을 위한 베이지안 분계점 모형, <응용통계연구>, **18**, 173-182.
- Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association*, **88**, 669-679.
- Chen, M. H. and Dey, D. K. (2000). Bayesian analysis for correlated ordinal data models. In: Generalized Linear models: A Bayesian Perspective, (D. K. Dey, S. K. Ghosh, B. K. Mallick, eds), Marcel Dekker, Inc. New York.
- Chib, S. (2000). Bayesian methods for correlated binary data. In: Generalized Linear models: A Bayesian Perspective, (D. K. Dey, S. K. Ghosh, B. K. Mallick, eds), Marcel Dekker, Inc. New York.
- Ghosh, M. and Meeden, G. (1998). *Bayesian Methods for Finite Population Sampling*, Chapman & Hall, London.
- Lee, S.-C. (2004). A naive multiple imputation method for ignorable nonresponse, *The Korean Communications in Statistics*, **11**, 399-411.
- 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.
- Meeden, G. (2000). A decision theoretic approach to imputation in finite population sampling, *Journal of American Statistical Association*, **95**, 586-595.
- Nusser, S. M., Carriquiry, A. L., Dodd, K. W., and Fuller, W. A. (1996). A semiparametric transformation approach to estimating usual intake distributions, *Journal of American Statistical Association*, **91**, 1440-1449.
- Rao, J. N. K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation, *Biometrika*, **57**, 377-387.
- Paddock, S. M. (2002). Bayesian nonparametric multiple imputation of partially observed data with ignorable nonresponse, *Biometrika*, **89**, 529-538.
- Raftery, A. E. and Lewis, S. M. (1992). How many iterations in the Gibbs sampler? In: Bayesian Statistics IV (J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith, eds), Oxford University Press, UK, 763-773.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, Inc. New York.

[2005년 3월 접수, 2005년 6월 채택]

Imputation for Binary or Ordered Categorical Traits Based on the Bayesian Threshold Model*

Seung-Chun Lee¹⁾

ABSTRACT

The nonresponse in sample survey causes a problem when it comes time to analyze dataset in public-use files where the user has only complete-data methods available and has limited information about the reasons for nonresponse. Recently imputation for nonresponse is becoming a standard approach for handling nonresponse and various imputation methods have been devised. However, most imputation methods concern with continuous traits while many interesting features are measured by binary or ordered categorical scales in sample survey. In this note, an imputation method for ignorable nonresponse in binary or ordered categorical traits is considered.

Keywords: Imputation, Ordered categorical variable, Hierarchical Bayesian threshold model, Logistic regression model

* This Work was Supported by Hanshin University Research Grant in 2005.

1) Professor, Dept. of Statistics, Hanshin University, 411 Yangsan-Dong, Osan, Kyunggi-Do, 447-791

E-mail: seung@hanshin.ac.kr