응용통계연구 제 11권 1호, 1998년, pp. 83-95

무응답을 포함하는 범주형 자료의 분석 *

박태성† 이승연‡

요 약

본 논문에서는 여론조사를 비롯한 표본조사에서 얻어지는 범주형 자료에서 결측치(missing observation)나 무응답(nonresponse)이 발생했을 때 이러한 자료를 적절하게 처리하여 분석할 수 있는 통계모형을 소개하고 실제 사례로서 1948년도에 미국에서 실시한 대통령 선거에 대한 여론조사 자료를 분석하였다. 당시 미국 여론조사 기관에 서는 Dewey후보가 압승을 거둘 것으로 예상을 했었지만 실제 선거에서는 Truman후보가 승리했었다.

1. 서론

최근 들어 사회 전반에 걸쳐 여러 종류의 여론조사가 보편화되어 있다. 이러한 여론 조사는 대개 설문을 통해서나 전화 인터뷰를 통해서 실시되며 응답자가 각 설문에 대하여 주어진 항목 중에서 답을 고르는 형식으로 자료가 수집되기 때문에 범주형 자료가 많이 얻어진다. 범주형 자료는 분할표(contingency table) 혹은 도수표(frequency table)를 사용하여 정리할 수 있다. 범주형 자료에 대한 고전적인 분석은 분할표의 각 칸(cell)의 기대도수나 칸 확률의 로그 변환된(log-transformed) 값이 열변수 효과와 행변수 효과의 선형식으로 표현된다고 가정하는 대수 선형모형(log-linear model)을 사용하여 이루어진다(Agresti, 1990).

대표적인 여론조사 사례는 제 15대 국회의원 당선자를 예측하기 위해 3개 방송사가 공동으로 실시한 '투표자 전화공동조사'이다. 1996년 4월 11일에 치러진 국회의원 선거 직후 개표가 시작되기도 전에 방송3사가 5개의 여론조사기관에 의뢰하여 실시한 전화여론조사 결과를 토대로 당선 예상자들을 발표하였다. 전국 2백53개 선거구 중에서 당선자 예측이 빗나간 곳은 39군데로 결과적으로 '안한 것만 못한 결과를 초래하여 여론조사에 대한 신뢰도에 먹칠을 하였다'는 모 일간지의 평가를 받았다. 이 전화 여론조사의 誤測에 대한 이유로는

(1) 여론조사기관의 정확하지 못한 조사와 분석

(2) 방송사의 섣부른 예측방송

등으로 요약할 수 있다. 특히 이번 전화여론조사에서는 전화조사대상자의 거짓 응답과 무응답(nonresponse)에 대한 적절치 못한 처리로 인해 여당의 득표율을 실제보다 높게 예측하는 결과를 초래하였다. 이같은 결과는 전화조사에서의 무응답자가

*이 논문은 1995년도 한국학술진흥재단의 공모과제 연구비에 의하여 연구되었음.

† (449-791) 경기도 용인시 모현면 왕산리 산 89, 한국외국어대학교 통계학과 부교수

‡ (134-071) 서울시 광진구 군자동 98, 세종대학교 응용통계학과 부교수

실제 투표에 불참했음에도 불구하고 분석과정에서 여당 쪽으로 분류가 되었던지

실제 투표에서 무소속이나 자민련 혹은 국민회의 후보에게 투표했음에도 불구하고 분석과정에서 여당 쪽으로 분류가 된 경우로 나누어 볼 수 있다.

이처럼 실제조사를 통해서 얻어진 범주형 자료는 종종 무응답(nonresponse data)이나 결측치 (missing data)를 포함한다. 이러한 무응답 값이나 결측치들이 반응변수와 아무런 상관이 없는 경우는 '무응답 자료가 발생할 확률이 무응답 값에 아무런 영향을 받지 않는다 (missing at random)'라고 하며 이러한 무응답들을 '무시할 수 있는 무응답(ignorable nonresponse)'이라고 한다. 그러나 무응답들이 실제 관측이 안된 무응답 값과 연관성을 가지고 있는 경우는 '무응답 자료가 발생할 확률이 무응답 값에 영향을 받는다'라고 하며 이러한 무응답들을 '무시할 수 없는 무응답(nonignorable nonresponse)'이라고 부른다. 예를 들어 제15대 총선 전화 여론조사에서 응답자는 상대방에게 자신의 전화번호가 노출되어 있다는 것을 알고 있기 때문에 여당 후보자에게 대한 지지는 명확하게 이야기할 수 있지만 야당후보자에 대한 지지를 공개적으로 나타내는 것을 꺼리는 경향이 있었다. 특히 공무원의 신분인 경우에는 이런 경향이 두드러진다고 할 수 있다. 따라서 야당 후보를 지지하는 경우에는 무응답이 발생할 가능성이 높아진다. 즉 무응답이 발생할 확률이 지지하는 후보자에 따라 영향을 받는다고 할 수 있다.

범주형 자료에서 무응답이 발생한 경우를 다루기 위한 많은 방법들이 제안되었다. 그러나 대부분의 분석방법들은 무응답들이 '무시할 수 있는 무응답'이라는 가정을 전제로 하고 있다. 그러나 이러한 가정을 근거로 한 분석방법은 관측이 안된 무응답 값과 연관성을 가지고 있는 '무시할 수 없는 무응답'이 발생하는 경우에는 잘못된 결론을 유도하게 된다. 무응답이 '무시할 수 있는 무응답'이라는 가정에 근거한 분석 방법으로는 Hocking and Oxspring (1971), Chen and Fienberg (1974, 1976), Fuchs (1982) 등이 있으며 모두 대수선형모형을 정의한 후에 최대우도(maximum likelihood, ML)추정법을 사용하여 모수를 추정하였다.

무응답들이 '무시할 수 없는 무응답'인 경우에는 그 처리가 매우 어려우며 분석이 까다로워진다. 이런 자료를 분석하기 위한 방법으로는 Pregibon (1977), Little (1980, 1982), Albert (1985) 등과 최근의 Clogg et. al (1991)이 제안한 베이지안 방법이 있으나 그 응용범위가 극히 제한적이기 때문에 일반화시킬 수 없는 단점이 있다. Fay (1986), Baker and Laird (1988), Chambers and Welsh (1993)은 대수선형모형을 이용하여 무응답을 유발시키는 메커니즘을 설명하기 위한 모형을 제안하였다. 이 방법은 EM 알고리즘을 이용하여 최대우도추정량 (maximum likelihood estimator)을 구할 수 있으나 종종 칸 도수의 추정치가 모수의 변방(boundary)에서 변방값 (boundary solution)을 갖는 경우가 발생한다. 변방값이 발생하는 경우는 칸 도수의 추정치가 0이든지 관측된 도수의 총 합으로 주어지는 경우로 대수선형모형의 특정 모수가 무한대의 추정값을 갖게 되어 Fisher 정보행렬을 이용한 최대우도추정량의 대표본 분포를 유도하기가 어려워진다.

Park and Brown (1994)은 ML추정법이 갖고 있는 변방값의 문제를 해결하기 위해 무응답 값이 발생한 칸에 Dirichlet 사전확률분포를 가정하고 사후밀도함수를 구한 후 극대점(mode)을 추정값으로 사용할 것을 제안하였다. 또한 시뮬레이션 비교연구를 통해 Park and

Brown (PB)의 추정량이 최대우도추정량 보다 작은 평균제곱오차(mean square error)를 갖고 있음을 보였다.

표 1.1: 1948년도에 실시한 미국대통령 후보에 대한 예비 선거

Economic	Time of	Known preference (Y)			Unknown
class(X_E)	Survey(X_T)	Truman	Dewey	Other	preference
A	July	43	161	10	29
A	Aug.	54	179	27	30
A	Sept.	39	189	8	26
A	Oct.	40	175	9	18
B	July	168	430	39	84
B	Aug.	205	486	69	103
B	Sept.	174	533	51	94
B	Oct.	203	508	45	109
C	July	453	582	77	244
C	Aug.	527	670	150	251
C	Sept.	488	717	108	256
C	Oct.	552	623	83	315
D	July	284	221	42	144
D	Aug.	307	212	65	155
D	Sept.	334	235	60	178
D	Oct.	354	211	38	217

본 연구에서는 '무시할 수 없는 결측치'를 갖는 범주형 자료를 분석하기 위한 통계모형들을 소개하고 이 모형의 추정법으로 ML추정법과 Park and Brown의 추정법을 소개하고, Park and Brown이 사용한 사전확률분포보다 좀더 일반적인 형태의 사전확률 분포를 사용하는 새로운 방법을 고려해보았다. 실제 사례로서 1948년도 미국 여론 조사 기관에서 실시한 여론조사 자료를 분석하였다. 표 1.1에 있는 자료는 Roper 여론조사 자료로 1948년도에 실시한 미국 대통령 후보의 지지도에 관한 예비조사 자료이다 (Barker and Laird, 1988). X_T는 조사기간을 나타내는 변수로 7월, 8월, 9월, 10월의 범주값을 갖고 X_E는 조사참여자의 경제적인 수준을 나타내는 변수로 A, B, C, D의 범주값을 가지고 Y는 후보자에 대한 지지도를 나타내는 반응변수로 Truman, Dewey, Other의 범주값을 갖는다. 후보자에 대한 선택을 못한 경우와 실제 응답을 얻지 못한 경우에는 무응답으로 분류하였고 응답자와 무응답자를 나누기위해 지시변수 R을 사용하였다. 표 1.1에서 볼 수 있는 바와 같이 경제수준과

조사 기간에 따라 무응답자의 비율이 크게 영향을 받고 있음을 알 수 있다. 특히 경제수준이 C와 D인 경우에 10월말의 조사에서 가장 높은 무응답율을 보이고 있다. 이 자료로부터 Truman 후보자에 대한 지지율을 예측하고자 한다. 당시 이 자료를 분석한 여론조사 기관에서는 Dewey후보가 승리할 것으로 예상을 했었지만 실제 선거에서는 Truman후보가 승리했다.

2장에서는 이러한 무응답 자료를 갖는 범주형 자료를 분석할 수 있는 통계모형들을 정의하고 3장에서는 모형을 추정하기 위한 ML추정법과 Park and Brown추정법을 소개하고 새로운 베이지안 추정법을 제시하였다. 4장에서는 표 1.1의 자료를 분석하여 Truman후보자의 지지율을 예측해보았고 마지막으로 5장에서는 결론을 정리하였다.

2. 무응답모형(NONRESPONSE MODEL)

두 변수 X와 Y가 관측이 된다고 가정하고 X는 독립변수(independent variable)를 나타내며 X가 취할 수 있는 범주의 수는 I라고 하자. 또 Y는 J개의 값을 갖는 종속변수(dependent variable)를 나타낸다고 하자. 또 편의상 변수 Y에 대해서만 무응답이 발생한다고 가정하자. 두 변수값이 모두 관측된 자료의 도수는$I \times J$분할표에 정리되고 X값만 관측된 자료는 별도의 $I \times 1$주변합 분할표에 정리된다. 여기서 종속변수 Y의 관측 여부를 나타내는 지시변수 R을 정의하고 (만약 Y가 관측되었으면 $R = 1$, 관측되지 않았으면 $R = 2$) 세 변수 X, Y, R의 값에 의해 만들어지는 확대된 분할표를 고려하자. 아래의 분할표는 $I = J = 2$인 경우에 원래의 분할표와 확대된 분할표를 보여준다. 여기서 $R = 1$인 경우에는 X와 Y가 모두 관측된 경우로써 $X = i$이고 $Y = j$인 경우의 도수는 n_{ij1}으로 표시되어 있고 $R = 2$인 경우에는X만 관측된 경우로 $X = i$ 일 때의 주변합 n_{i+2} 만 관측이 된 경우이다. 그러나 확대된 분할표에서는 n_{ij2} 가 관측이 되지 않았기 때문에 ?로 표시되어 있다.

원래의 분할표

$R = 1$인 경우

n_{111}	n_{121}
n_{211}	n_{221}

$R = 2$인 경우

n_{1+2}
n_{2+2}

확대된 분할표

$R = 1$인 경우

n_{111}	n_{121}
n_{211}	n_{221}

$R = 2$인 경우

?	?
?	?

(합)

n_{1+2}

n_{2+2}

본 논문에서 사용할 대수선형모형은 무응답을 유발시키는 메커니즘을 설명하기 위한 모형으로서 세 변수 X, Y, R에 대하여 정의되는 모형이다. 즉 확대된 분할표에 대한 대수선형모형이 된다. 여기서 만약 'Y가 관측될 확률이 Y의 값에 아무런 영향을 받지 않는다'면 대수선형모형에 YR항을 포함시킬 필요가 없으나 'Y가 관측될 확률이 Y의 값에 영향을 받는다'면 YR항을 포함한 대수선형모형을 사용해야 한다. 예를 들면 대수선형모형 (XR, YR)은 YR항을 포함하고 있으므로 '무시할 수 없는 무응답을 갖는 모형 (nonignorable nonresponse model)'이고 (XY, XR)은 YR항을 포함하고 있지 않으므로 '무시할 수 있는 무응답을 갖는 모형 (ignorable nonresponse model)' 이 된다. 즉 YR항을 포함하는 모형은 '무시할 수 없는

무응답모형'이라고 부르고 YR항을 포함하지 않는 모형을 '무시할 수 있는 모형'이라고 부르자.

확대된 분할표에서 (ijk)칸의 기대도수를 m_{ijk}로 표시하고 m_{ijk}로 이루어진 벡터를 m으로 표시하면 무응답모형은 아래와 같이 대수선형모형으로 정의할 수 있다.

$$\log m = Z\beta$$

여기서 Z는 계획행렬(design matrix)이고 β는 모수벡터이다. 이러한 모형은 독립변수가 두 개 이상인 경우로 쉽게 확장할 수 있다. 즉 $X = (X_1, X_2, \cdots, X_p)$이고 X_i의 범주의 수가 I_i이라고 하면 관측된 자료는 $I_1 \times I_2 \times \cdots \times I_p \times J$ 분할표에 정리된다. 이 자료에 대해서도 YR을 포함하는 대수선형모형을 같은 방법으로 정의할 수 있다.

확대된 분할표에서 $R = 2$인 경우는 각 칸의 도수가 관측이 안되었고 단지 그 주변합(marginal sum) 만이 관측이 된 상태이므로 이 자료에 YR항을 포함하는 대수선형모형을 적합시키면 경우에 따라서는 관측된 자료의 분할표에 있는 칸의 숫자보다 모형의 모수가 많게 되는 과모수화(overparameterization)현상이 발생할 수가 있으므로 모형의 선정에 주의를 해야한다. 또한 모수를 추정하기 위해서는 EM 알고리즘 등을 이용한 반복적인 방법이 많이 사용된다. 모형의 추정에 관한 더 자세한 설명은 Little and Rubin(1987, p237)과 Baker and Laird(1988)를 참조하기 바란다.

3. 모형의 추정

세 변수 X, Y, R에 대하여 $X = i, Y = j, R = k$가 관측될 확률을 π_{ijk}라고 나타내면 $\{n_{ij1}\}$과 $\{n_{i+2}\}$가 관측될 확률로부터 우도함수가 다음과 같이 얻어진다.

$$L = \left(\prod_i \prod_j \pi_{ij1}^{n_{ij1}} \right) \left(\prod_i \pi_{i+2}^{n_{i+2}} \right) \tag{3.1}$$

첨자 i는 독립변수 X의 수준을 나타내고 j는 종속변수 Y의 수준을 나타내고 k는 지시변수 R의 수준을 나타낸다. 첨자 i는 2개 이상의 독립변수에 대해서는 $i = (i_1 i_2 \cdots)$의 형태로 확장될 수 있으나 편의상 i로 표시하자.

3.1절에서는 이 우도함수를 최대로 만드는 Baker and Laird (1988)의 최대우도추정법을 소개하고, 3.2절에서는 Park and Brown (1994)이 제안한 방법인 무응답칸 (n_{ij2})에 대하여 Dirichlet 사전분포를 가정하고 이로부터 사후확률분포을 유도하여 이를 최대로 만드는 방법을 소개하였다. 3.3절에서는 Park and Brown (1994)의 방법을 일반화하여 $R = 1$인 경우와 $R = 2$인 경우 모두 사전도수를 설정하는 새로운 방법을 제안하였다.

3.1. 최대우도추정법(ML 추정법)

상수를 제외한 로그우도함수는 다음의 형태를 갖는다.

$$l = \sum_i \sum_j n_{ij1} \log(\pi_{ij1}) + \sum_i n_{i+2} \log(\pi_{i+2})$$

무응답모형의 모수에 대한 최대우도추정법에서 l을 최대로 만드는 값은 EM 알고리즘을 이용해서 찾을 수 있다. m_{ijk}를 무응답모형의 기대도수라고 하면 E-과정은 관측이 안된 $R = 2$인 경우의 칸 도수를

$$n^*_{ij2} = n_{i+2} \frac{m_{ij2}}{m_{i+2}} \tag{3.2}$$

으로 추정하고 M-단계에서는 관측된 n_{ij1}와 E-단계에서 얻은 n^*_{ij2}을 완전한 자료라고 가정한 후에 대수선형모형을 추정하는 단계이다. M-단계는 일반적인 통계패키지에 있는 대수선형모형의 프로그램을 이용하여 추정할 수 있다. E-단계와 M-단계를 계속 반복해서 얻은 수렴값이 m_{ijk}의 최대우도추정량이 된다. 이 추정량으로부터 무응답모형의 모수를 추정할 수 있고 Fisher의 정보행렬을 이용하여 대표본 분산을 얻을 수 있다. 그러나 '무시할 수 없는 무응답모형'에 대하여 최대우도추정법을 사용하면 종종 칸 도수의 추정치가 모수의 변방 (boundary)에서 변방값 (boundary solution)을 갖는 경우가 발생한다. 즉 칸 도수의 추정치가 0 이든지 관측된 도수의 합 (n_{i+2})으로 주어지게 된다. 또 이런 경우에는 대수선형모형의 YR모수는 무한대의 추정값을 갖게 되므로 Fisher 정보행렬을 이용한 최대우도추정량의 대표본 분포를 유도하기가 어렵게 된다. 더구나 이러한 변방값은 $R = 2$인 칸의 주변합 (marginal sum)에 민감하게 영향을 받는다. 예를 들어 주변합의 조그만 도수의 변화가 칸 $(i, j, 2)$의 기대도수의 추정량을 0에서 n_{i+2}으로 변화시키고 또 n_{i+2}에서 0으로 바꾸기도 한다(Park and Brown, 1994). 이러한 최대우도추정법이 갖고있는 문제점을 극복하기 위해 Park and Brown (1994)은 다음절에서 소개하는 추정법을 대안적으로 제안하였다.

3.2. Park and Brown (1994)의 추정법

Park and Brown (1994)의 방법은 대수선형모형을 이용하여 추정을 하되 기존의 고전적인 방법인 최대우도추정법이 갖고 있는 단점, 즉 추정치가 모수의 변방 (boundary)에서 변방값 (boundary solution)을 갖는 문제점을 극복하기 위해 무응답이 발생한 칸에 Dirichlet 사전확률분포를 가정한 후 이 사전분포와 우도함수로부터 사후분포를 유도하여 이 사후분포를 최대로 만드는 추정량을 구하는 방법이다. Park and Brown은 무응답칸(n_{ij2})에 대하여 Dirichlet 사전분포를 켤레 (conjugate)형태로 $\prod_i \prod_j p_{ij2}^{\delta_{ij2}}$ 가정하고 이로부터 다음과 같이 주어진 로그사후확률분포를 유도하여 이를 최대로 만드는 추정량을 제안하였다. 여기서 δ_{ij2}는 사전도수를 나타낸다.

$$l_p = \sum_i \sum_j n_{ij1} \log(\pi_{ij1}) + \sum_i [n_{i+2} \log(\pi_{i+2}) + \sum_j \delta_{ij} \log(\pi_{ij2})]$$

무응답모형의 모수에 대한 추정법은 l_p을 최대로 만드는 값을 EM 알고리즘을 이용해서 찾는 것이다. m_{ijk}를 무응답모형의 기대도수라고 하면 E-과정은 관측이 안된 $R = 2$인 경

우의 칸 도수를

$$n^*_{ij2} = n_{i+2}\frac{m_{ij2}}{m_{i+2}} + \delta_{ij2} \tag{3.3}$$

으로 추정하고 M-단계에서는 관측된 n_{ij1}와 E-단계에서 얻은 n^*_{ij2}을 완전한 자료라고 가정한 후에 대수선형모형을 추정하는 단계이다. Park and Brown은 $R = 2$인 칸에 대해서만 Dirichlet 사전확률분포를 가정하여 사전도수를 적용하였으므로 전통적인 베이지안 방법으로 간주하기가 곤란하다. 이를 일반화한 방법은 다음 절에서 소개되는 $R = 1$인 경우와 $R = 2$인 경우 모두 사전도수를 설정하는 방법이다.

3.3. 베이지안 추정법

응답칸(n_{ij1})과 무응답칸(n_{ij2})에 대하여 Dirichlet 켤레사전분포인 $\prod_k \prod_i \prod_j \pi_{ijk}^{\delta_{ijk}}$로 가정하고 이로부터 다음의 대수변환된 사후확률분포을 유도하여 이를 최대로 만드는 추정량을 고려해보자.

$$l_p = \sum_i \sum_j [n_{ij1} + \delta_{ij1}] \log(\pi_{ij1}) + \sum_i [n_{i+2} \log(\pi_{i+2}) + \sum_j \delta_{ij2} \log(\pi_{ij2})]$$

l_p를 모형의 모수에 대하여 미분시킨 후에 정리하게 되면

$$n^*_{ijk} = \begin{cases} n_{ij1} + \delta_{ij1}, & k = 1 \\ n_{i+2}\frac{m_{ij2}}{m_{i+2}} + \delta_{ij2}, & k = 2 \end{cases} \tag{3.4}$$

로 주어진다. 식 (3.4)는 모든 i와 j에 대하여 $\delta_{ij1} = \delta_{ij2} = 0$인 경우에는 ML방법의 식 (3.2)가 되고 모든 i와 j에 대하여$\delta_{ij1} = 0$인 경우는 Park and Brown (1994) 방법의 식 (3.3)과 일치하게된다. 따라서 새로 제안된 베이지안 방법은 기존의 ML방법과 Park and Brown (1994)방법을 포함하는 포괄적인 방법이라고 할 수 있다.

여기서 δ_{ijk}를 기존의 경험이나 유사한 형태의 자료로부터 경험적으로 구할 수 있으면 식 (3.4)를 이용하여 반복적인 방법으로 대수선형모형의 모수를 추정할 수 있다. 만약 δ_{ijk}가 기지의 상수값이든지 아니면 어떤 기지의 모수값을 갖는 또 다른 사전분포를 따르는 확률변수라고 가정하면 모수의 추정은 완전한 형태의 베이지안 추정을 따르게 된다. 만약 이러한 사전 정보가 없다면 현재 관찰된 자료로부터 δ_{ijk}값을 결정하기 위해 다음과 같은 경험적인 베이지안(empirical Bayesian) 형태의 방법을 생각해 보자. 먼저 $\Delta_1 = \sum_i \sum_j \delta_{ij1}$이고 $\Delta_2 = \sum_i \sum_j \delta_{ij2}$라고 정의하자. 또한 δ_{ijk}가 응답칸의 도수 n_{ij1}에 비례한다고 가정하면

$$\delta_{ij1} = \Delta_1 \frac{n_{ij1}}{n_{++1}} \text{ 이고 } \delta_{ij2} = \Delta_2 \frac{n_{ij1}}{n_{++2}}$$

이 성립하므로 식 (3.4)는

$$n^*_{ijk} = \begin{cases} n_{ij1} + \Delta_1 \frac{n_{ij1}}{n_{++1}}, & k = 1 \\ n_{i+2}\frac{m_{ij2}}{m_{i+2}} + \Delta_2 \frac{n_{ij1}}{n_{++2}}, & k = 2 \end{cases}$$

가 된다. 또한 이 식에 Park and Brown (1994)이 사용한 제약조건과 유사한 제약 조건인 $n^*_{i+1} = n_{i+1}$과 $n^*_{i+2} = n_{i+2}$을 적용하면

$$n^*_{ijk} = \begin{cases} \frac{n_{i+1}}{n_{i+1}} + \delta_{i+1}\left(n_{ij1} + \Delta_1 \frac{n_{ij1}}{n_{++1}}\right), & k = 1 \\ \frac{n_{i+2}}{n_{i+2}+\delta_{i+2}}\left(n_{i+2}\frac{m_{ij2}}{m_{i+2}} + \Delta_2 \frac{n_{ij1}}{n_{++2}}\right), & k = 2 \end{cases}$$

가 얻어진다. 이 식의 형태는 복잡해 보이지만 모든 i와 j에 대하여 $\delta_{ij1} = \delta_{ij2} = 0$인 경우에는 식 (3.2)의 ML 방법이 되고 모든 i와 j에 대하여 $\delta_{ij1} = 0$인 경우는 Park and Brown (1994)의 식 (3.3)과 일치하게 된다. 마지막 단계로서는 Δ_1과 Δ_2의 값을 결정하는 단계이다. 역시 Park and Brown (1994)의 방법을 따라서 $\Delta_1 + \Delta_2$를 대수선형모형의 모수의 수(p)라고 가정하고 Δ_k를 n_{++k}에 비례하게 설정하면 된다. 다음절에서는 이 절에서 소개된 방법을 이용하여 표 1.1에 있는 자료를 분석해 보았다.

표 3.1: ML 추정결과 : Truman 후보에 대한 지지율

무응답모형		변방값	Truman 후보의 지지율	Goodness of fit	
				G^2	df
무시할 수 있는 모형	1. $(X_E X_T Y, X_T R, X_E R)$	No	41	9.85	9
	2. $(X_E X_T Y, X_T R)$	No	41	184.72	12
	3. $(X_E X_T Y, X_E R)$	No	41	25.34	12
무시할 수 없는 모형	4. $(X_E X_T Y, X_T R, X_E R, Y R)$	Yes	52	3.44	8
	5. $(X_E X_T Y, X_T Y R)$	Yes	52	3.74	8
	6. $(X_E X_T Y, X_T R, Y R)$	Yes	52	4.44	11
	7. $(X_E X_T Y, X_E Y R)$	Yes	52	7.97	8
	8. $(X_E X_T Y, X_E R, Y R)$	Yes	52	7.87	11
	9. $(X_E X_T Y, Y R)$	Yes	52	8.90	14

4. 사례연구

본 절에서는 표 1.1의 자료에 대하여 ML방법과 베이지안 방법을 이용하여 무응답 모형을 적합시켜 보았다. 이 자료의 분석에서 가장 관심이 있는 것은 Truman후보에 대한 지지율을 추정하는 것이다. 이 중에서 특히 조사기간이 10월일 때 Truman후보의 지지율을 추정해보자. Baker and Laird (1988)는 ML방법을 사용하여 '무시할 수 있는 모형'과 '무시할 수 없는 모형'을 적합시킨 결과 지지율이 투표자의 경제적인 수준에 따라 상당한 차이가 있음을 보였고 '무시할 수 있는 모형'보다 '무시할 수 없는 모형' 이 Truman후보에 대하여 더 높은 지지율을 나타냄을 보였다. 표 3.1에 있는 결과는 ML방법을 이용하여 3가지 '무시할

수 있는 모형'과 6가지 '무시할 수 없는 모형'을 추정한 결과로부터 Truman 후보에 대한 전체적인 지지율을 추정하여 정리한 표이다.

이 표의 결과는 Baker and Laird (1988)의 결과와 약간의 차이를 보이고 있는 데 그 이유는 우리가 좀더 강한 수렴 조건을 사용했기 때문으로 사료된다. 이 표에서 알 수 있듯이 '무시할 수 있는 모형'은 변방값 해를 갖고 있지 않으며 Truman후보에 대한 지지율이 41%로 낮게 나오나 '무시할 수 없는 모형'은 Truman후보에 대한 지지율이 52% 정도로 실제 투표에서 나온 58%와 아주 비슷한 값을 갖게 된다. 또한 모형의 적합도를 나타내는 G^2통계량의 값도 '무시할 수 있는 모형'에서는 큰 값을 갖게 되어 모형이 적합하지 않음을 나타내나 '무시할 수 없는 모형'의 경우에는 작은 값을 갖게 되어 모형이 잘 적합되었음을 나타낸다. 그러나 '무시할 수 없는 모형'들은 모두 다 변방값을 갖고 있어 대수선형모형의 모수 β의 추정에 어려움이 있게 된다.

다음으로 이 '무시할 수 없는 모형' 중에서 적합이 잘된 모형 4, 5, 6에 대하여 Park and Brown의 방법과 새로운 베이지안 방법을 사용하여 모형을 적합시켜 보았다. 표 4.1에는 이 세 모형에 대하여 ML 추정결과와 Park and Brown의 방법 결과와 베이지안 방법의 결과가 함께 정리되어있다. Park and Brown의 방법과 베이지안 방법에서는 $\Delta_1 + \Delta_2$의 값은 대수선형모형의 모수의 수인 p를 사용하였다. ML 추정 결과를 보면 무응답의 모든 도수가 Truman으로 분배되고 Dewey와 Other 범주로는 전혀 분배가 되지 않아 변방값이 발생하였으나 Park and Brown의 방법과 베이지안 방법은 무응답의 일부 도수가 Dewey와 Other 범주로 분배되었기 때문에 변방값의 해가 발생하지 않았다. 따라서 변방값을 갖는 칸 도수의 MLE로부터 대수선형모형의 모수의 추정값을 구할 수 없으나 Park and Brown의 방법과 베이지안 방법으로부터 대수선형모형의 모수에 대한 추론이 가능하게 된다. 표 4.1에서 보는 것처럼 Park and Brown(PB) 방법은 ML방법 보다 상당히 작은 값의 지지율을 갖고 있고 베이지안 방법은 Park and Brown의 방법과 ML방법의 중간 정도의 지지율을 보여줌을 알 수 있다.

이 여러 모형의 추정 결과들로부터 어떤 모형을 선택해야할 것인가? 일반적으로 좋은 모형은 적은 수의 모수를 가지고 자료를 잘 적합하는 모형이다. 이 자료에 대한 분석 결과로부터 모형 6이 가장 적은 수의 모수를 가지고 있고 대체적으로 적합이 잘 되었음을 알 수 있다. 모형 6에 대하여 ML 방법을 사용한 경우에는 G^2통계량의 값이 4.44이었고 Park and Brown의 방법의 경우에는 31.37이고 베이지안 방법의 경우에는 9.84이었다. ML 방법은 변방값 때문에 문제가 되고 Park and Brown 방법은 G^2통계량의 값이 너무 크기 때문에 사용하기가 곤란하므로 베이지안 방법의 결과를 사용하는 것이 가장 타당하다고 생각된다.

마지막으로 사전도수의 합 $\Delta_1 + \Delta_2 (= p)$이 변함에 따라서 Truman후보의 전체 지지율이 어떻게 변하는 지를 민감도분석(sensitivity analysis)을 실시해 보았다. 이 지지율은 $\Delta_1 + \Delta_2$의 값이 증가함에 따라 점차적으로 감소하는 추세를 보이다가 대략 20을 넘어서게 되면 일정한 값을 갖는 안정화모양을 보여주었다. 따라서 본 자료에서 사용한$\Delta_1 + \Delta_2$의 값은 적절한 선택이었다고 판단된다.

표 4.1: 모형추정결과 : 경제상태에 따른 Truman 후보에 대한 지지율 (%)

무응답모형	추정법	변방값	Economic Class				전체	Goodness of fit	
			A	B	C	D	지지율	G^2	df
4. $(X_E X_T Y,$	MLE	Yes	24	36	55	69	52	3.44	8
$X_T R, X_E R, Y R)$	PB	No	18	27	44	59	42	9.50	8
	베이지안	No	19	29	47	63	44	7.97	8
5. $(X_E X_T Y,$	MLE	Yes	24	36	55	70	52	3.74	8
$X_T Y R)$	PB	No	20	34	52	67	48	29.48	8
	베이지안	No	22	35	54	69	51	8.61	8
6. $(X_E X_T Y,$	MLE	Yes	24	36	55	70	52	4.44	11
$X_T R, Y R)$	PB	No	20	31	51	66	47	31.37	11
	베이지안	No	22	34	53	68	50	9.84	11

5. 정리

본 논문에서는 결측치와 무응답값을 포함하는 범주형 자료를 분석할 수 있는 대수선형 모형을 살펴보았다. 특히 '무시할 수 없는 무응답'값을 갖는 범주형 자료를 분석하기 위한 모형을 자세히 고찰해 보았다. 또 이 모형을 추정하기 위한 ML추정법과 Park and Brown추정법을 소개하고 Park and Brown이 사용한 사전확률분포보다 좀 더 일반적인 형태의 사전확률을 사용하는 새로운 방법을 제안하였다. 또한 관찰된 자료로부터 사전확률분포의 모수값을 구할 수 있는 간단한 방법을 소개하였다. 따라서 이 방법은 베이지안 방법이라기보다는 경험적 베이지안 방법에 더 가깝다고 할 수 있다. 앞으로 모의실험 연구를 통해서 ML, Park and Brown, 베이지안 방법을 서로 비교해보고자 한다. 이 비교에는 기대도수의 추정값들로부터 편의(bias)와 평균제곱오차 (mean square error, MSE) 등을 사용할 것이다.

무응답 값을 갖는 범주형 자료를 분석하기 위해 몇몇의 베이지안 방법들이 제안되었지만 대부분이 무응답을 무시할 수 있는 경우만을 다루었다. '무시할 수 있는 무응답'이라는 가정 하에서 제안된 베이지안 방법으로는 Basu and Pereira (1982), Albert and Gupta (1983), Dickey, Jiang, and Kadane (1987) 등이 있으나 추정 과정이 복잡하여 모수 추정에 어려움이 있기 때문에 널리 사용되고 있지는 못한 상태이다. 본 논문에서 살펴본 Park and Brown의 방법과 새로운 베이지안 방법은 '무시할 수 있는 무응답'과 '무시할 수 없는 무응답'을 둘 다 처리할 수 있는 장점을 갖고 있고 추정 과정도 대체적으로 간단하기 때문에 쉽게 사용될 수 있는 방법이라고 생각된다.

최근 들어 전화나 설문을 통한 표본조사를 통해서 많은 형태의 범주형 자료가 수집되고 분석되고 있지만 결측치를 포함하고 있는 경우에 이를 잘 처리할 수 있는 보편적인 통계기법은 아직 많이 개발되지 않은 상태이다. 갤럽과 같은 조사기관에서는 여론 조사에서 소위 이야기하는 부동층의 향방을 결정하기 위해 판별분석과 같은 다변량 분석 기법을 사용

하고 있다. 이 방법은 무응답에 영향을 미칠 수 있는 여러 변수들을 기초로 해서 측정된 자료의 집단 성향을 밝혀낸 후에 미지의 사례가 어느 집단에 속할 것인가를 예측하는 방법이다. 본 연구에서 다루었던 대수선형모형을 사용한 방법도 판별분석 방법과 마찬가지로 우리 나라의 여론조사에서도 적절하게 사용할 수 있으리라고 생각한다. 특히 총선의 사전 예비조사 등과 같이 '무시할 수 없는 무응답'이 자주 발생하는 경우에서 유용하게 사용할 수 있으리라고 기대하며 특히 무응답값에 영향을 미칠 수 있는 좋은 변수들(예. 지역)이 관찰된 경우에는 효과적으로 적용할 수 있는 방법이 되리라 생각한다. 그러나 이 모형을 사용한 방법은 설문 자체를 거절하여 독립변수들조차 관찰이 안된 무응답의 경우를 다루기는 적절하지 못한 방법이다.

감사의 글

본 논문을 심사해주신 심사위원님들께 감사를 드립니다.

참고문헌

[1] Agresti, A. (1990). *Categorical Data Analysis.* John Wiley & Sons, New York.

[2] Albert, J. H. (1985). Bayesian estimation methods for incomplete two-way contingency tables using prior beliefs of association. *Bayesian Statistics*, 2, 589-602.

[3] Albert, J. H. and Gupta, A. K. (1983). Bayesian estimation method for 2x2 contingency tables using mixtures of Dirichlet distributions. *Journal of the American Statistical Association*, Vol. 78, 708-717.

[4] Baker, S. G. and Laird, N. M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, Vol. 83, 62-69.

[5] Basu, D. and Pereira, C. A. (1982). On the Bayesian analysis for categorical data: The problem of nonresponse. *Journal of Statistical Planning and Inference*, Vol. 6, 345-362.

[6] Chambers, R. L. and Welsh, A. H. (1993). Log-linear models for survey data with non-ignorable non-response. *Journal of the Royal Statistical Society*, Series B, Vol. 55, 157-170.

[7] Chen, T., and Fienberg, S. E. (1976). The analysis of contingency tables with incompletely classified data. *Biometrics*, Vol. 32, 133-144.

[8] Chen, T., and Fienberg, S. E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, Vol. 30, 629-642.

[9] Clogg, C. C., Rubin, D. B., Schenker, N., and Schultz, B. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association*, Vol. 86, 68-78.

[10] Dickey, J. M., Jiang, J. M., and Kadane, J. B. (1987). Bayesian methods for censored categorical data. *Journal of the American Statistical Association*, Vol. 82, 773-781.

[11] Fay, R. E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association*, Vol. 81, 354-365.

[12] Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of the American Statistical Association*, Vol. 77, 270-278.

[13] Hocking, H. O. and Oxspring, H. H. (1971). Maximum likelihood estimation with incomplete multinomial data. *Journal of the American Statistical Association*, Vol. 66, 65-70.

[14] Little, R. J. A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, Vol. 77, 237-250.

[15] Little, R. J. A. (1980). Superpopulation models for nonresponse. Nonresponse in Sample Surveys: The Theory and Current Practice. Part V , Panel on Incomplete Data National Academy of Sciences, Washington, D. C.

[16] Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* John Wiley & Sons, New York.

[17] Park, T. and Brown, M. B. (1994). Models for categorical data with nonignorable nonresponse. *Journal of the American Statistical Association*, Vol. 89, 44-52.

[18] Pregibon, D. (1977). Typical survey data: Estimation and imputation. *Survey Methodology*, Vol. 2, 70-102.

[1997년 4월 접수, 1997년 8월 최종수정]

Analysis of Categorical Data with Nonresponses *

Taesung Park† Seung-Yeoun Lee‡

ABSTRACT

Statistical models are proposed for analyzing categorical data in the presence of missing observations or nonresponses which might occur in the sampling surveys and polls. As an illustration, we analyzed real polling data of the pre-presidential election in the USA, 1948. It had been predicted that Dewey would win the election. However, Truman won in the actual election.

*This research was supported by Non Directed Research Fund from Korea Research Foundation, 1995.

† Associate Professor, Department of Statistics, Hankuk University of Foreign Studies, Kyungki-Do 449-791, Korea.

‡ Associate Professor, Department of Applied Statistics, Sejong University, Kwangjin-Gu, Seoul 134-071, Korea.