응용통계연구 제 13권 1호, 2000년, pp. 197-206

반복비율적합에 의한 다차원 분할표의 결측칸값 추정

최현집[1] 신상준[2]

요 약

반복비율적합 방법을 확장하여 준독립성모형하에서 불완전한 다차원 분할표에 포함된 결측칸의 최우추정값을 얻기위한 추정방법을 제안하였다. 제안된 방법은 주변합이 영이 아닌 모든 불완전한 분할표에 적용할 수 있으며 주어진 준로그선형모형의 구조를 해치지 않는다. 또한 결측칸의 위치와 수에 영향을 받지 않고 항상 수렴한다는 것을 확인하였다.

주요용어: 결측칸, 불완전한 분할표, 준로그선형모형, 반복비율적합.

1. 서 론

분할표의 빈칸(empty cell)은 크게 표본추출 영값(sampling zero)을 갖는 칸과 구조적인 영값(structural zero)을 갖는 칸으로 나눌 수 있다. 표본추출 영값은 전체 표본수를 증가 시키면 칸값이 나타날 수 있는 경우로 칸 확률이 존재하는 경우를 말한다. 그러나 구조적인 영값은 자료구조상 분할표를 구성하기 이전에 칸 확률이 존재하지 않는 다는 것을 알거나 혹은 어떤 이유에서든 자료가 수집되지 않은 경우를 말하며, 구조적인 영값을 포함한 분할표를 불완전한 분할표(incomplete contingency table)라고 한다. 이러한 구조적인 영값을 갖는 칸은 관찰되지 않았거나 관찰할 수 없는 칸을 의미하므로 결측칸(missing cell)이라고 한다.

Goodman(1968)은 불완전한 분할표 분석을 위해 준독립성모형(quasi independence model)을 제안하고 결측칸이 나타날 수 있는 여러 상황과 결측칸을 제외한 나머지 칸들의 최우추정값을 얻기 위한 반복계산법을 제안하였다. 특히, Fienberg(1970)와 Savage(1973)는 준독립성모형을 적합시켰을 때 결측칸을 제외한 나머지 칸에 대한 유일한 최우추정량이 존재한다는 것을 밝혔으며, 여러 형태의 불완전한 이차원 분할표의 최우추정량을 얻기 위한 방법은 Bishop, Fienberg와 Holland(1975) 그리고 홍종선과 최현집(1999)에 잘 정리되어 있다. 이러한 준독립성모형을 다차원 분할표로 확장한 준로그선형모형(quasi log-linear model)에 관해서는 Fienberg(1972)와 Harberman(1974)에서 자세히 다루고 있다.

불완전한 분할표는 이차원일 경우에도 결측칸의 위치와 수에 따라 준독립성모형을 위한 최우추정값을 얻기 위한 식(closed form expression)이 존재하지 않을 수 있다. 이러한 경

1) (442-760) 경기도 수원시 팔달구 이의동 산 94-6, 경기대학교 경제학부 응용정보통계전공 조교수
E-mail: hjchoi@stat.kyonggi.ac.kr
2) (442-760) 경기도 수원시 팔달구 이의동 산 94-6, 경기대학교 경제학부 응용정보통계전공

우에 결측칸을 포함하고 있지 않은 완전한 분할표(complete contingency table)의 최우추정을 위한 반복비율적합(IPF : iterative proportional fitting) 방법을 결측칸을 제외한 칸들에 대하여 직접 적용하여 최우추정값을 얻는다. 즉, 반복비율적합의 매 주기에서 결측칸은 완전한 적합이 이루어진 것으로 간주하고 반복계산의 각 단계에서 제외시킨다.

Wagner(1970)는 이차원 정방형 분할표(two dimensional square contingency table)의 대각칸이 결측칸인 경우에 준독립성모형 하에서 행과 열의 주변비율의 최우추정값을 얻기 위한 반복추정방법을 제안하였다. Wagner의 방법은 주변비율의 최우추정값을 얻기 때문에 결측칸의 추정이 필요한 경우에 추정된 주변비율에 의해 준독립성모형하에서의 결측칸의 추정값을 얻을 수 있다. 그러나 Wagner의 방법은 유일한 해가 존재하지 않을 수 있고, 대각칸이 아닌 비대각 결측칸을 추정하기 위한 방법으로 확장할 수 없다. 이차원 분할표의 임의 결측칸의 추정방법에 관한 연구로는 Graf, Alf Jr., 와 Williams(1997)등이 있다. Graf등이 제안한 방법은 이차원 분할표에 나타날 수 있는 결측칸의 위치와 칸의 수등에 제약이 없이 반복추정에 의해 최우추정값을 얻을 수 있다. 또 그들이 제안한 방법에 의해 추정된 결측칸의 최우추정값은 결측칸을 제외한 나머지 칸들의 준독립성 관계에 영향을 미치지 않는다. 이러한 사실로 부터 그들은 불완전한 분할표에 대하여 준독립성모형하에서 결측칸을 추정한 후에 이들을 관찰값인 것으로 간주하고 독립성모형을 적합하는 방법을 제안하였다. 따라서 Goodman과는 달리 관찰할 수 없는 결측칸을 가진 불완전한 분할표 분석을 위해 먼저 결측칸을 추정한 후에 독립성모형을 적합시키는 분석방법을 고려할 수 있다.

그러나 이들 연구는 모두 이차원 분할표의 결측칸값을 추정하기 위한 방법으로 삼차원 이상의 일반적인 다차원 분할표의 결측칸값을 추정하기 위하여 확장하는데는 어려움이 있다. 즉, Wagner의 방법은 이차원 정방형 분할표의 대각칸이 결측칸인 경우에 국한되며 Graf등이 제안한 방법은 이차원 분할표에서 직접해(direct solution)가 존재하는 독립성모형인 경우에 주어진 추정식을 이용한다. 따라서 삼차원 분할표에서 직접추정이 가능한 완전독립성모형(complete independence model)에서 조차 반복을 위한 식을 유도하기가 어렵기 때문에 이를 다차원 분할표로 확장하는 것은 더욱 어렵다. 또한 그들은 예를 통해 제안한 방법의 수렴성을 제시하고 있을 뿐 제안된 방법의 수렴성을 밝히지는 못했다.

본 논문에서는 일반적인 다차원 분할표의 결측칸값을 추정하기 위한 추정방법에 관하여 연구하였다. 먼저, 결측칸에 적절한 초기값을 부여한 반복비율적합 방법을 확장한 추정방법을 제안하고, 제안된 방법에 의해 얻은 추정값은 준로그선형모형의 구조를 해치지 않는 최우추정값임과 항상 수렴한다는 것을 확인하였다. 제2절에서는 일반적인 다차원 분할표의 결측칸을 추정하기 위한 반복추정방법을 제안하고, 최우추정값이 존재하기 위한 조건, 초기값 선정 문제 그리고 수렴성 문제를 다루었다. 제3절에서는 제2절에서 제안한 방법을 적용한 실제자료의 분석을 통한 예를 소개하였다. 마지막으로 제4절에서는 제안된 방법과 이전 연구를 비교하였으며, 추정된 결측칸의 추정분산을 쉽게 얻을 수 있음을 토론하였다.

2. 일반적인 다차원분할표의 결측칸값 추정

일반적인 다차원 분할표의 칸의 전체 집합을 T 그리고 결측칸을 제외한 칸의 집합을 S라고 하자. 이로부터 T에서 S를 제외한 즉, $T\backslash S$를 결측칸의 집합으로 나타내기로 한다. 또한 분할표의 칸을 나타내는 첨자의 집합을 θ라 하고, $\theta \in S$인 경우에 x_θ는 관찰칸값, m_θ를 관찰칸값에 대응하는 기대칸값(expected cell count) 그리고 $\theta \in T\backslash S$인 결측칸값을 M_θ로 나타내기로 한다. 결측칸 M_θ는 $0 < M_\theta < \infty$이라고 가정하자. 즉, 결측칸의 추정값이 존재한다고 가정한다. 이제 θ_q, $q = 1, 2, \ldots, Q$, 를 주어진 불완전한 분할표에 가장 잘 적합되는 준로그선형모형의 최소충분통계량(minimal sufficient statistics)의 첨자의 집합이라고 하면 x_{θ_q}는 최소충분통계량을 그리고 $\hat{m}_{\theta_q}$는 x_{θ_q}에 의한 최우추정값의 주변합을 나타낸다.

이러한 상황에서 M_θ를 추정하기 위한 다음과 같은 반복비율적합방법을 응용한 반복계산법을 제안하기로 한다.

- (1 단계) 결측칸 M_θ에 영보다 큰 적절한 초기값 $\hat{M}_\theta^{(0)}$을 부여한다.
- (2 단계) 부여된 초기값이 관찰값인 것으로 간주하고 다음과 같은 반복비율적합을 수행한다.

  1. $\theta \in T$에 대하여 적절한 초기값 $\hat{m}_\theta^{(0)}$을 부여한다.
  2. $\hat{M}_\theta^{(0)}$와 $\hat{m}_\theta^{(0)}$ 그리고 최소충분통계량 x_{θ_q}, $q = 1, 2, \ldots, Q$, 을 이용하여 다음의 반복계산을 수행한다. 만일 $\theta_q \notin T\backslash S$이면

  $$\hat{m}_\theta^{rq} = \hat{m}_\theta^{(r-1)q} \frac{x_{\theta_q}}{\hat{m}_{\theta_q}^{(r-1)q}}, \tag{2.1}$$

  혹은 $\theta_q \in T\backslash S$이면

  $$\hat{m}_\theta^{rq} = \hat{m}_\theta^{(r-1)q} \frac{x_{\theta_q}^{*(t-1)}}{\hat{m}_{\theta_q}^{(r-1)q}}, \tag{2.2}$$

  여기서 $x_{\theta_q}^{*(t-1)} = x_{\theta_q} + \hat{M}_{\theta_q}^{(t-1)}$이며, $\hat{M}_{\theta_q}^{(t-1)}$은 $\hat{M}_\theta^{(t-1)}, \theta \in T\backslash S$, 만의 주변합을 나타낸다.

  3. $r = 1, 2, \ldots$에 대하여 $\hat{m}_\theta^{rq} \approx \hat{m}_\theta^{(r-1)q}$인 적절한 정도에 이를때까지 반복을 수행한다.

- (3 단계) 이전 단계에서 얻은 추정값을 이용하여 결측칸 추정을 위한 t번째 반복에서의 잠정추정값 $\hat{M}_\theta^{(t)} = \hat{m}_\theta^{rq}, \theta \in T\backslash S$, 를 얻는다.

$t = 1, 2, \ldots$에 대하여 $\hat{M}_\theta^{(t)} \approx \hat{M}_\theta^{(t-1)}, \theta \in T\backslash S$, 인 적절한 정도(accuracy)에 이를때까지 반복을 수행하며, 결측칸의 추정값 $\hat{M}_\theta = \hat{M}_\theta^{(t)}$를 얻는다.

준로그선형모형은 완전한 분할표의 로그선형모형에서와 같이 최소충분통계량에 의해 추정이 이루어지므로 최소충분통계량이 영인 분할표에서는 준로그선형모형이 정의되지 않

는다. 반복비율적합 방법을 확장한 위 방법은 준로그선형모형의 최소충분통계량에 의해 반복계산이 수행되므로 최소충분통계량이 영인 분할표의 결측칸은 추정할 수 없다. 또한 위 추정방법은 $(t-1)$번째 반복의 잠정추정값이 관찰값인 것으로 간주하고 t번째 반복의 추정값을 얻는다. 따라서 포아송추출모형(Poisson sampling model)을 가정하면 준로그선형모형의 기대칸값과 결측칸값을 추정하기 위한 $(t-1)$번째 반복에서의 로그우도함수의 커넬을 다음과 표현할 수 있다.

$$\sum_{\theta \in S} x_\theta \log m_\theta + \sum_{\theta \in T \setminus S} \hat{M}_\theta^{(t-1)} \log M_\theta \,. \tag{2.3}$$

위 로그우도함수의 커넬로부터 준로그선형모형의 최우추정값과 결측칸의 추정을 위한 다음과 같은 우도방정식들을 얻을 수 있다.

$$x_{\theta_q} + \hat{M}_{\theta_q}^{(t-1)} \;=\; \hat{m}_{\theta_q} + \hat{M}_{\theta_q}^{(t)}, \; q = 1, 2, \ldots, Q. \tag{2.4}$$

즉, 제안된 추정방법에 의한 $\hat{M}_\theta^{(t)}$는 우도방정식 (2.4)의 해이며, $\hat{M}_\theta^{(t)} \approx \hat{M}_\theta^{(t-1)}$이라면 $\hat{M}_{\theta_q}^{(t)} \approx \hat{M}_{\theta_q}^{(t-1)}$일 것이므로 (2.4)는 준로그선형모형의 최우추정을 위한 정규방정식과 같다. 그리고 제안된 추정방법은 영보다 큰 초기값에 의해 반복계산이 수행되므로 $\hat{M}_\theta^{(t-1)}$은 항상 영보다 크고, $\hat{M}_\theta^{(t)}$ 역시 항상 영보다 크다. 또한 (2.3)은 $x_\theta = \hat{m}_\theta$와 $\hat{M}_\theta^{(t-1)} = \hat{M}_\theta^{(t)}$에서 최대가 되므로 Bishop등(1975, P.85)에 의해 $\hat{M}_\theta^{(t)}$는 항상 수렴한다. 그러므로 제안된 방법에 의한 수렴된 추정값 $\hat{M}_\theta = \hat{M}_\theta^{(t)}$는 준로그선형모형의 최우추정값에 영향을 미치지 않으며 (2.3)을 최대로하는 최우추정값이 된다.

준로그선형모형을 위한 반복비율적합 방법은 초기값에 민감하지 않다는 사실이 알려져 있다. 특히, Biship등(1975)은 주어진 모형의 구조를 해치지 않는 값이면 어떠한 초기값이라 할지라도 반복에 영향을 주지 않는 다는 것을 밝혔으며, 계층모형(hierarchical model)인 경우에 $\hat{m}_\theta^{(0)} = 1$과 같은 초기값을 제안하고 있다. 따라서 제안된 방법에 의한 결측칸의 추정값은 준로그선형모형의 최소충분통계량의 함수로 얻어지므로 일반적인 반복비율적합에서 사용되는 초기값, 즉 준로그선형모형의 구조를 해치지 않는 적절한 값으로 선택할 수 있고 $\hat{M}_\theta^{(0)} = 1$과 같은 초기값을 고려할 수 있다. 그리고 제안된 추정방법은 반복의 각 단계에서 주어진 초기값을 관찰값인 것으로 간주하고 반복비율적합을 수행한다. 반복비율적합은 Dempster, Laird와 Rubin(1977)의 EM 알고리즘의 특수한 경우라는 것은 널리 알려져 있으며, 특히 제안된 방법의 (2 단계)는 주어진 초기값에서 최대값을 얻기 위한 조건부 최대화 단계(conditional maximization step) 그리고 추정단계(estimation step)인 (3 단계)를 갖는 Meng과 Rubin(1993)의 ECM 알고리즘의 특수한 경우로 생각할 수도 있다.

반복비율적합은 IMSL 혹은 SAS와 같은 소프트웨어들에서 제공하는 함수에 의해 쉽게 프로그램할 수 있으므로 제안된 방법을 실제 자료에 적용하는 것은 그리 어려운 문제가 아니다. 또한 제안된 방법은 준로그선형모형의 최소충분통계량이 영이 아닌 모든 분할표에 적용할 수 있으므로, 결측칸의 수와 위치에 영향을 받지 않는다. 그리고 제안된 방법을 통해 결측칸과 결측칸이 아닌 칸들에 대한 준로그선형모형 하에서의 추정값을 동시에 얻을 수 있다. 따라서 주어진 준로그선형모형의 적합성 검정을 수행하기에 용이하다.

3. 예

표 3.1의 자료는 Wagner(1970)와 Graf등(1997)에서 인용한 세 종류의 다람쥐 원숭이(squirrel monkey) R, S 그리고 U의 생식표현(genital display)에 관한 자료이다.

표 3.1: 세 종류의 다람쥐 원숭이

표현 \ 응답	R	S	U	
R	-	1	8	9
	-	(2.2155)	(6.7845)	
S	29	-	46	75
	(27.7845)	-	(47.2155)	
U	2	3	-	5
	(3.2155)	(1.7845)	-	
	31	4	54	89

표 3.1에서 대각칸은 관찰할 수 없는 결측칸을 나타내고, 칸값 아래 괄호안의 값은 준독립성모형에 의한 추정칸값이다. 따라서 $T\backslash S = \{(1,1),(2,2),(3,3)\}$이고, 준독립성모형의 최소충분통계량은 $x_{\theta_1} = x_{i+}$와 $x_{\theta_2} = x_{+j}$이다. 이제 결측칸을 추정하기 위해 초기값 $\hat{M}_{ij}^{(0)} = 1$ 그리고 반복비율적합을 위해 초기값 $\hat{m}_{ij}^{(0)} = 1$를 부여하면 제안된 방법의 (2 단계)에서의 칸 $(1,1)$의 첫번째 잠정추정값은 $\theta_1 \in T\backslash S$이고, $r = 1$ 그리고 $q = 1$이므로 (2.2)에 의해 다음과 같이 계산된다.

$$\hat{m}_{11}^{1\cdot 1} = \hat{m}_{11}^{(1-1)1} \frac{x_{1+}^{*(1-1)}}{\hat{m}_{1+}^{(1-1)1}}$$
$$= 1 \cdot \frac{9+1}{3} = 10/3.$$

유사한 방법에 의해 칸 $(2,2)$에 대해서 $r = 1$ 그리고 $q = 2$이므로

$$\hat{m}_{22}^{1\cdot 1} = \hat{m}_{22}^{(1-1)2} \frac{x_{2+}^{*(1-1)}}{\hat{m}_{2+}^{(1-1)2}} = 76/3,$$

그리고 칸 $(3,3)$에 대해서 $\hat{m}_{33}^{1\cdot 1} = 6/3$을 얻을 수 있다. 이제 이들값과 $r = 1$에서의 S에 대한 잠정추정값을 이용하여 다음 반복을 수행하며, (3 단계)에서 다음과 같은 결측칸에 대한 잠정추정값을 얻을 수 있다.

$$\hat{M}_{11}^{(1)} = 3.4782,$$
$$\hat{M}_{22}^{(1)} = 4.1404,$$
$$\hat{M}_{33}^{(1)} = 3.5869.$$

이러한 반복을 $|\hat{M}_{ij}^{(t)} - \hat{M}_{ij}^{(t-1)}| < 0.0001$인 정도에 이를때까지 반복하여 표 3.2와 같은 결측칸의 최우추정값과 결측칸을 제외한 칸들의 최우추정값을 얻을 수 있다.

표 3.2: 준독립성모형하에서의 최우추정값

	R	S	U
R	3.9924	2.2155	6.7845
S	27.7845	15.4183	47.2155
U	3.2155	1.7845	5.4644

표 3.2에서 대각칸을 제외한 나머지 칸들의 최우추정값이 표 3.1의 준독립성모형 하에서의 최우추정값과 같다는 점에 주목하기 바란다. 또한 제안된 방법에 의한 대각칸의 추정값들은 Wagner 그리고 Graf등에서 얻은 최우추정값과 일치한다. 다만 제안된 추정방법은 Graf등의 방법에 비해 매 t번째 반복에서 반복비율적합을 실시한다는 점에서 계산의 양이 많다. 그러나 Graf의 방법은 삼차원이상 다차원으로 확장하기 어렵다는 점에 주목하기 바란다(제4절 참고) .

다음으로 제안된 방법의 다차원 분할표의 적용의 예로 Bishop등(1975)에서 사용된 다음과 같은 $2 \times 2 \times 2$분할표를 고려해보기로 한다.

표 3.3: 솜꼬리 토끼 자료

변수 1		1		2	
변수 2		1	2	1	2
변수 3	1	8	26	32	189
	2	8	41	41	-

표 3.3의 자료는 Michigan주의 한 자연생태공원의 솜꼬리 토기(cottontail rabbit)의 총수를 추정하기 위한 자료로 다음과 같은 세 변수로 구성되어 있다.

- 변수 1 : 조사기간의 전기에 포획된 토끼의 표시유무(1:표시, 2:표시않음)
- 변수 2 : 조사기간의 후기에 포획된 토끼의 표시유무(1:표시, 2:표시않음)
- 변수 3 : 조사기간 종료후 사냥꾼에 의해 포획된 토끼의 표시유무(1:표시, 2:표시않음)

칸 $(2,2,2)$는 포획되지 않은 토끼의 수를 나타내므로 관찰되지 않은 결측칸이 된다. Bishop등은 표 3.2의 자료를 가장 잘 적합되는 준로그선형모형으로 세 변수가 완전한 독립인 모형을 선택하였다. 따라서 최소충분통계량 $x_{\theta_1} = x_{i++}$, $x_{\theta_2} = x_{+j+}$, $x_{\theta_3} = x_{++k}$를 얻을 수 있다. 이로부터 초기값 $\hat{M}_{ijk}^{(0)} = 1$ 그리고 $\hat{m}_{ijk}^{(0)} = 1$를 부여하고 제안된 반복추정방법을

이용하여 $\hat{M}_{222} = 240.7448$을 얻을 수 있다. 이 추정값은 Bishop등이 제안한 포획-재포획모형(capture-recapture model)에서 총표본수를 추정하기 위한 칸 $(2,2,2)$의 최우추정값과 일치한다.

4. 토 의

분할표에 포함된 결측칸의 추정값은 주어진 준로그선형모형의 구조를 해치지 않아야 한다. 즉, (i,j)칸만이 결측인 이차원 분할표에서 완전독립성모형인 경우에 결측칸 M_{ij}는 다음을 만족하여야 한다.

$$M_{ij} = \frac{(x_{i+} + M_{ij})(x_{+j} + M_{ij})}{(N + M_{ij})} , \tag{4.1}$$

여기서 $N = \sum_i \sum_j x_{ij}$로 결측칸을 제외한 칸의 총합을 나타낸다. 이제 (4.1)을 M_{ij}에 대하여 정리하면 주어진 완전독립성모형의 최소충분통계량 x_{i+}와 x_{+j}의 함수인 다음과 같은 결측칸에 대한 추정량을 얻을 수 있다.

$$\hat{M}_{ij} = \frac{x_{i+}x_{+j}}{(N - x_{i+} - x_{+j})} . \tag{4.2}$$

Graf등(1997)은 이러한 사실로부터 둘 이상의 결측칸을 추정하기 위하여 (4.2)를 이용한 반복계산을 위한 식을 유도하고 있다. 따라서 이를 삼차원으로 확장하면 직접해(direct solution)가 존재하는 완전독립성모형인 경우에 (i,j,k)칸만이 결측이라도 M_{ijk}의 추정을 위해서는 다음을 만족하는 식을 유도하여야 한다.

$$M_{ijk} = \frac{(x_{i++} + M_{ijk})(x_{+j+} + M_{ijk})(x_{++k} + M_{ijk})}{(N + M_{ijk})^2} . \tag{4.3}$$

그러나 (4.3)은 M_{ijk}의 다항함수이므로 M_{ijk}의 해조차 찾기가 쉽지 않고 결국 반복계산을 위한 식을 유도한다는 것은 더욱 어려운 일이다. 더욱이 주어진 준로그선형모형이 직접해가 존재하지 않는 모형인 경우에는 반복계산을 위한 식을 찾는 것은 더더욱 어렵다. 이와는 달리 제안된 추정방법은 반복을 위한 매 단계에서 주어진 모형의 최소충분통계량의 함수인 (2.2)를 통해 반복비율적합을 수행한다. 따라서 일반적인 다차원분할표에서 최소충분통계량이 영이 아닌 어떠한 준로그선형모형에 대해서도 결측칸의 추정이 가능하다. 또한 주어진 준로그선형모형의 최소충분통계량의 함수에 의해 반복이 수행되기 때문에 제안된 방법에 의한 추정값은 항상 (4.2)와 (4.3)과 같은 조건을 만족한다. 반복비율적합은 주어진 로그선형모형에 의한 기대칸값의 최우추정을 위한 반복계산방법으로 다음과 같은 추정공분산행렬을 갖는다는 것이 널리 알려져 있다(Agresti(1990), Christensen(1997)).

$$\hat{\mathrm{V}}ar(\underline{\hat{m}}) = \mathbf{D}(\underline{\hat{m}})\mathbf{X}(\mathbf{X}'\mathbf{D}(\underline{\hat{m}})\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}(\underline{\hat{m}}) - \underline{\hat{m}}\underline{\hat{m}}'/N , \tag{4.4}$$

여기서 $\underline{\hat{m}}$은 추정기대칸값 벡터, $\mathbf{D}(\underline{\hat{m}})$는 추정기대칸값을 대각원소로 갖는 대각행렬 그리고 $\mathbf{X}$는 주어진 로그선형모형의 모수인 각 연관항(association term)에 부여된 적절한 제

약조건에 의한 계획행렬(design matrix)이다. 제안된 방법은 결측칸 추정을 위해 반복의 매 단계에서 잠정추정량을 관찰값인 것으로 간주하고 추정을 수행한다. 또한 수렴된 추정값은 (4.2)와 (4.3)과 같은 조건을 만족하기 때문에 준로그선형모형의 모수 추정값에 영향을 주지 않는다. 따라서 결측칸의 추정값에 대한 추정분산은 (4.4)를 직접 적용하여 쉽게 구할 수 있다. 참고로 제3절의 예에서 삼차원분할표인 솜꼬리 토끼 자료의 결측칸 M_{222}의 표준오차는 (4.4)에 의해 14.4672임을 쉽게 구할 수 있다.

감사의 글

본 논문을 심사해주신 심사위원들과 편집위원들께 감사드립니다.

참고문헌

[1] 홍종선, 최현집 (1999). 〈로그선형모형을 이용한 범주형 자료분석〉. 자유아카데미.

[2] Agresti, A. (1990). *Categorical Data Analysis*, Wiley.

[3] Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975). *Discrete Multivariate Analysis : Theory and Practice*, The MIT Press.

[4] Christensen, R. (1997). *Log-Linear Models and Logistic Regression*, Springer-Verlag.

[5] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of Royal Statistical Society* Series B, Vol. 39, 1-38.

[6] Fienberg, S.E. (1970). Quasi-independence and maximum likelihood estimation in incomplete contingency tables, *Journal of the American Statistical Association*, Vol. 65, 1610-1616.

[7] Fienberg, S.E. (1972). The analysis of incomplete multi-way contingency tables, *Biometrics*, Vol. 28, 177-202.

[8] Goodman, L.A. (1968). The analysis of cross-classified data : independence, quasi-independence, and interaction in contingency tables with or without missing cells, *Journal of the American Statistical Association*, Vol. 63, 1091-1131.

[9] Graf, R.G., Alf Jr., E.F. and Williams, S. (1997). A BASIC program for estimating missing cell frequencies in Chi square tests for association, *InterStat*, http://interstat.stat.vt.edu/InterStat/.

[10] Harberman, S.J. (1974). *The Analysis of Frequency Data*, University of Chicago Press.

[11] Meng, X.L. and Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, Vol. 80, 267-278.

[12] Savage, I.R. (1973). Incomplete contingency tables : condition for the existence of unique MLE, In *Mathematics and Statistics. Essays in Honor of Harold Bergström*, Sweden, Chalmers Institute of Technology, 87-99.

[13] Wagner, S.S. (1970). The maximum-likelihood estimate for contingency tables with zero diagonal. *Journal of the American Statistical Association*, Vol. 65, 1362-1383.

[1999년 6월 접수, 2000년 2월 채택]

Estimating Missing Cells in Contingency Table with IPF

Hyun Jip Choi [1] Sang Jun Shin [2]

ABSTRACT

For estimating missing cells in contingency table, we suggest an iterative method which extends IPF (Iterative Proportional Fitting) method. The suggested method is not restricted by the number and the location of missing cells, and does not distort the given quasi-independency.

Keywords: Missing cells; Incomplete tables; Quasi-Log Linear models; Iterative proportional fitting.

1) Assistant Professor, Division of Economics, Kyonggi University. E-mail: hjchoi@stat.kyonggi.ac.kr
2) Graduate Student, Department of Applied Information Statistics, Kyonggi University.