한국통계학회논문집
2009, 16권, 2호, 229–238

범주형 자료분석을 위한 최대절사우도추정

최현집[1,a]

[a]경기대학교 응용정보통계학과

요약

범주형 자료분석을 위해 고려할 수 있는 모형들은 일반적으로 최우추정에 의하여 적합이 이루어지므로 이상값에 쉽게 영향을 받을 수 있다. 본 연구에서는 분할표 자료에 포함된 이상칸(outlying cell)에 영향을 받지 않는 최대절삭우도 추정값(maximum trimmed likelihood estimates)을 얻기 위한 추정 방법을 제안하였다. 제안된 방법은 우도에 의존하여 분할표에 포함된 칸을 제거해나가며 절사우도의 최대값을 찾기 때문에 완전탐색(complete enumeration)에 비해 계산의 양이 매우 적다. 따라서 일반적인 다차원 분할표 자료분석을 위해 쉽게 적용될 수 있다. 실제 자료분석 예를 통해 제안된 추정방법을 설명하였으며, 모의실험을 통해 문제점과 특징을 토론하였다.

주요용어: 분할표, 이상칸, 최대절사우도 추정량.

1. 서론

여러 범주형 변수들에 의해 생성되는 분할표 분석의 주된 관심사 중에 하나는 분할표를 구성하는 변수들의 연관관계를 식별하는데 있으며, 이를 위해 적용되는 모형들은 대체로 최우추정법을 이용하여 모형에 포함된 모수를 추정한다. 특히 로그선형모형(log-linear models)은 모형에 포함된 모수들에 의해 변수들 사이의 연관관계를 식별하게 할 뿐만 아니라 모수들이 교차적비(cross-product ratios)의 함수로 표현되어 범주들의 연관성을 식별할 수 있기 때문에 가장 널리 적용되고 있다. 따라서 분할표 분석을 위한 로그선형모형의 추정에서도 Shane과 Simonoff (2001) 등이 지적한 바와 같이 이상값(outliers)에 민감한 최우추정법이 가지고 있는 문제점이 충분히 고려되어야 한다. 분할표 자료에서 이상값은 주어진 모형과 일치된 적합을 나타내지 않는 칸으로 이들 칸을 이상칸(outlying cells)이라고 한다.

로그선형모형을 위한 로버스트 추정 방법으로 Rousseeuw (1984)가 제안한 LMS(least median of squares)와 LTS(least trimmed squares) 추정 방법을 응용한 Shane과 Simonoff (2001)가 제안한 방법 그리고 최현집 (2003)이 제안한 절대 편차합을 최소로 하는 LAD(least absolute deviation) 추정방법 등을 고려할 수 있다. 두 연구 모두 Grizzle 등 (1969)이 제안한 가중 최소제곱잔차(weighted least squares residuals) 합을 이용한 반복계산에 의하여 추정값을 얻는다는 공통점을 가지고 있다.

이들 방법과는 달리 추정을 위해 우도함수 값을 절사한 후에 최우추정량을 얻는 Hadi와 Luceño (1997)가 제안한 최대절사우도 추정법을 고려할 수 있다. 최대절사우도 추정량(MTLE: maximum trimmed likelihood estimators)은 오차항의 분포가 정규분포인 경우에 절사모수(trimming parameter) 따라 LMS와 LTS 추정량을 특수한 경우로 나타낼 수 있으며, 따라서 최우추정에 있어서 매우 로버스트하다는 장점을 가진다. 그러나 추정을 위한 목적함수(objective function) 값이 우도값들의 순서에 의해

[1] (443-760) 경기도 수원시 영통구 이의동 산 94-6, 경기대학교 응용정보통계학전공, 부교수.
E-mail: hjchoi@kyonggi.ac.kr

결정되기 때문에 이산최적(discrete optimization) 문제와 동등하게 되어 계산의 양이 매우 많아지게 된다. 이러한 문제를 해결하기 위하여 Neykov와 Muller (2002)는 Rousseeuw (1984)가 제안한 FAST-LTS 추정 방법을 직접 응용한 FAST-TLE 방법을 제안하였으며, 추정하고자 하는 모형에 따라 Čížek (2006), Cheng과 Biswas (2008), Neykov 등 (2007) 등의 추정 방법을 고려할 수 있다.

본 연구에서는 분할표 분석을 위한 로그선형모형의 최대절사우도 추정값을 얻기 위한 추정방법을 제안하고자 한다. 먼저 2절에서는 로그선형모형의 추정을 위한 최대절사우도 추정량을 정의하고, FAST-TLE 등의 추정방법이 직접 적용되기 어렵다는 것을 설명하고자 한다. 또한 이러한 문제를 해결하기 위한 추정 방법을 제안할 것이다. 3절에서는 이상칸 식별을 위해 많은 선행 연구들에서 인용한 자료들의 분석 예를 보일 것이며, 4절에서는 모의 실험을 통해 제안된 추정 방법의 특징을 설명하고자 한다. 마지막으로 5절에서는 본 연구의 결과를 정리하고 제안된 추정방법이 가질 수 있는 문제점에 관하여 토론할 것이다.

2. 범주형 자료 분석을 위한 최대절사우도추정량

먼저 n개 칸을 갖는 다차원 분할표를 고려하기로 한다. 관찰칸 값으로 구성된 분할표를 $\underline{x} = \{x_i\}$, $i = 1, 2, \ldots, n$과 같은 $n \times 1$차 벡터 그리고 $\underline{m} = \{m_i\}$는 $n \times 1$차 기대칸 값(expected cell counts) 벡터로 나타내기로 한다. 각 기대칸 값 $m_i = N\pi_i$로 여기서 $N = \sum_{i=1}^{n} x_i$는 분할표의 총합을 그리고 $\underline{\pi} = \{\pi_i\}$는 칸 확률(cell probabilities) 벡터이다. 이 때 $\underline{\pi}$는 알려져 있지 않기 때문에 추정 기대칸 값 벡터 $\hat{\underline{m}}$은 관찰칸 값 벡터와 분석을 위해 고려된 다음과 같은 로그선형모형에 포함된 $p \times 1$차 모수벡터 $\underline{\theta}$의 추정값 $\hat{\underline{\theta}}$로부터 얻을 수 있다.

$$\log \underline{m} = \mathbf{D}\,\underline{\theta}\,, \tag{2.1}$$

여기서 $\mathbf{D}$는 모형을 위해 정의된 $n \times p$차 계획 행렬(design matrix)이며, 추정을 위해 모수 θ들에는 합이 0이되는 제약조건을 부여하기로 한다.

Hadi와 Luceño (1997)에 의하여 포아송 표본추출모형 하에서 모형 (2.1)을 위한 최대절사우도 추정량은 다음의 목적함수를 최대로 하는 $\underline{\theta}$로 정의할 수 있다.

$$\sum_{i=1}^{h} l_{(i)}(x_i; \underline{\theta})\,, \tag{2.2}$$

여기서 $l_{(i)}(x_i; \underline{\theta})$는 다음과 같이 크기 순으로 정렬된 포아송 분포의 로그우도 $\log f(x_i; \underline{\theta})$ 그리고 h는 절사모수를 나타낸다.

$$l_{(1)}(x_i; \underline{\theta}) \geq l_{(2)}(x_i; \underline{\theta}) \geq \cdots \geq l_{(h)}(x_i; \underline{\theta}) \geq \cdots \geq l_{(n)}(x_i; \underline{\theta}). \tag{2.3}$$

Neykov 등 (2007)이 지적한 바와 같이 목적함수 (2.2)의 절사는 적합된 모형이 사실일 경우에 우도가 작은 $n - h$개 관찰칸을 제거하는 것이 된다. 즉, 절사모수 h에 의한 $\hat{\underline{\theta}}$은 분할표의 h개 칸에 의하여 적합이 일어난 모형의 최우추정량이 된다. 이는 다시 말해 전체 $\binom{n}{h}$개 가능한 h개 칸들에 대한 $\hat{\underline{\theta}}$ 중에서 가장 큰 목적함수 값을 얻게 되는 h개 칸들에 의한 $\hat{\underline{\theta}}$이 최대절사우도 추정량이 된다는 것을 의미한다. Hadi와 Luceño (1997) 그리고 Neykov와 Muller (2002) 역시 지적한 이러한 사실은 다수의 국소 최대값(local maximum values)이 존재하는 목적함수 (2.2)의 특성 때문에 정확한 추정을 위해서는 모든 가능한 $\binom{n}{h}$개 칸들의 집합에 대하여 최우추정을 수행하여야 한다는 것을 의미한다. 그러나 일반적으로 이차원 분할표의 경우에도 $\binom{n}{h}$는 매우 크므로 완전 탐색(complete enumeration)에 의한 추정은 계산

의 양이 매우 많다. 같은 이유로 연속형 자료에 대한 근사적인 추정값을 얻기 위한 다수의 추정 방법이 제안되어 왔으며, 이들은 대체로 Rousseeuw와 Driessen (2006)이 제안한 LTS 추정량을 위한 추정방법에 기반하고 있다. 이들 중 특히 Neykov와 Muller (2002)는 Rousseeuw와 Driessen (2006)이 제안한 FAST-LTS를 직접 응용한 FAST-TLE를 제안하였다.

연속형 자료를 위한 FAST-TLE는 만일 h가 결정되었다면 우선 임의의 h개 초기 관찰값을 선정한 후에 이들로 부터 최우추정량을 얻고, 이를 통해 얻어진 우도들을 식 (2.3)과 같이 재정렬한 후에 다시 큰 순서대로 h개 관찰값을 선택하여 재적합시키는 반복계산을 수행한다. 이 때 이전 h개 관찰값에 의한 목적함수는 새롭게 결정된 h개 관찰값의 목적함수 보다 크거나 같은 단조 증가를 취하게 된다. 그러므로 반복은 이전 h개 관찰값의 최우추정값에 의한 목적함수 값과 새롭게 결정된 목적함수 값의 차이가 적절히 작을 때 멈추게 되고, 마지막으로 선택된 h개 관찰값의 최우추정값이 최대절사우도 추정값이 된다. 이 때 각 초기 관찰값에 대한 반복은 최대 모든 $\binom{n}{h}$개 부분집합에 대한 계산이 수행되므로 반복계산은 항상 수렴하게 된다. 그러나 반복의 시작을 위한 초기 h개 관찰값에 따라 최종 수렴된 추정량은 국소 최대값에 따른 추정값이 될 수 있으므로 다수의 초기 h개 관찰값을 선택하여 각 초기 자료에 대한 반복계산 후 가장 큰 목적함수를 갖는 추정값을 선택한다. 따라서 이 추정 방법에 의하여 얻어진 추정값은 주어진 모형의 정확한 해가 아닌 근사적인 추정값이 된다.

모형 (2.1)에서 모수의 최우추정값을 얻기 위한 방법으로 Grizzle 등 (1969)이 고려한

$$\sum_{i\in I} \omega_i(\log x_i - \log m_i)$$

을 최소로 하는 반복가중최소제곱(iteratively reweighted least squares) 추정법을 고려할 수 있다. 여기서 $I = \{1, 2, \ldots, n\}$은 분할표를 구성하는 n개 칸들의 첨자 집합을 나타낸다. Grizzle 등 (1969)은 $\log x_i$의 근사분산이 $1/m_i$라는 사실로부터 가중값 $\omega_i \equiv \hat{m}_i$로 할 것을 제안하였고, 이러한 가중값을 이용한 추정값이 근사적으로 최우추정값과 동등함을 보였다.

이로부터 로그선형모형의 추정을 위해 FAST-TLE를 직접 적용할 경우에, 선택된 h개 초기 관찰값들의 집합을 $H(H \subset I)$라고 하면, 부분관찰칸 H를 위한 최우추정량은

$$\sum_{i\in H\subset I} \omega_i(\log x_i - \log m_i)$$

를 최소로하는 반복가중최소제곱 추정량이 될 것이다. 그러나 이 경우에 모형 (2.1)의 계획행렬 $\mathbf{D}$의 특징 때문에 임의로 선정된 부분칸 집합 H에 의해 모형의 모수를 추정할 수 없는 경우가 생길 수 있다. 이차원분할표를 예로 들면 H에는 행과 열에 최소한 하나 이상의 칸이 반드시 포함되어야 한다. 다시 말해 선택된 h개 칸들에 의한 집합 H에 대한 계획행렬을 $\mathbf{D}_H$라고 하면 $\text{rank}(\mathbf{D}_H) = p$가 되어야 한다. 그러나 $\binom{n}{h}$개 가능한 칸 집합 중에서 이러한 조건을 만족하는 임의의 칸 집합을 선택하기는 쉽지 않다. 또한 선택된 h개 칸들에 의한 (2.3)과 같은 정렬에 의해 결정되는 h개 칸들 역시 $\text{rank}(\mathbf{D}_H) = p$가 되어야 하나 이러한 조건을 만족시키는 것 역시 어렵다. 결국 로그선형 모형의 최대절사우도 추정을 위해서 FAST-TLE 추정방법을 이용하는데는 제약이 따르게 된다.

분할표 분석을 위해 Shane과 Simonoff (2001)가 제안한 방법 역시 많은 수의 부분 관찰칸들의 집합 H를 임의로 선정하여 이들 중에서 모수의 추정이 가능한 칸 집합에만 최우추정을 수행하기 때문에 FAST-TLE와 동일한 제약을 갖는다. 또한 두 추정 방법 모두 h가 p에 가까울 수로 $\binom{n}{h}$가 커지기 때문에 이러한 제약은 더욱 심해지게 된다.

목적함수 (2.2)의 최대값이 순서에 영향을 받지 않고 유일하게 존재한다고 가정하자. 이러한 경우에 $n - h$개 관찰칸의 제거는 먼저 n개 칸들에 의한 최우추정량을 구하고 이 추정량에 의한 우도함수 값

을 구하여 정렬한 후에 가장 작은 우도함수 값을 제거하여 $n-1$개 칸을 얻는 직관적인 방법을 고려할 수 있다. 이러한 과정을 h 칸들이 남을 때 까지 반복하여 얻어진 부분관찰칸값들의 집합 H_h^*에 의한 최우추정량은 최대절삭우도 추정량이 될 것이다. 그러나 목적함수 (2.2)는 다수의 국소 최대값을 가지므로 이 방법을 일반적인 상황으로 직접 확장할 수는 없다.

그러나 $H_h^* \subset I$이며 $n > h$이므로, 크기가 $n-1$인 I의 부분집합을 H_{n-1}이라고 하면, $\binom{n}{n-1}$개 H_{n-1}들 중에 하나는 반드시 H_h^*를 부분집합으로 가질 것이다. 이를 H_{n-1}^*라고 하면 H_{n-1}^*로 부터 고려할 수 있는 $\binom{n-1}{n-2}$개 H_{n-2}들 중 하나인 H_{n-2}^* 역시 H_h^*를 부분집합으로 갖는다. 그러므로 순차적으로 크기가 줄어든 H_{h+1}^*를 통해 최대절사우도 추정량을 추정해주는 H_h^*를 선택할 수 있을 것이다. 여기서 H_{n-t}^*, $i = 1, 2, \ldots, n-h-1$들을 선택하기 위하여 각각 앞에서 가정한 상황에서와 같이 순차적인 제거에 의해 크기가 h인 칸에 이르렀을 때의 목표함수 값이 가장 큰 집합을 얻는 방법을 고려하기로 한다.

이러한 사실을 바탕으로 분할표 자료의 최대절삭우도 추정량을 추정하기 위한 다음과 같은 추정 방법을 제안하기로 한다.

1 단계: 절삭 모수 h을 결정하고 $t = 1$, 칸 집합 $H_n^* = \{1, 2, \ldots, n\}$ 그리고 목적함수의 초기값으로 $O(H_n) = 0$을 부여한다.

2 단계: $H_{(n-t+1)}^*$로부터 모든 가능한 크기 $n-t$인 $H_{(n-t)}^{(i)}$, $i = 1, 2, \ldots, n-t$를 생성하고 각 $H_{(n-t)}^{(i)}$에 대하여 다음을 반복한다.

1. $j = 0$을 부여한다.
2. $H_{(n-t-j)}^{(i)}$로부터 최우추정값을 얻는다.
3. $H_{(n-t-j)}^{(i)}$에 속한 칸들의 우도들을 정렬하여 가장 작은 우도를 갖는 칸 c를 결정한다.
4. $H_{(n-t-j)}^{(i)}$로부터 칸 c를 제거한 다음 반복을 위한 $H_{(n-t-j-1)}^{(i)}$를 생성한다.
5. $j = j+1$로 증가시키고 j가 $n-h-t$에 이를때까지 위 반복을 수행하여 얻은 칸 집합에 의한 목적함수 값 $O(H_{(n-t)}^{(i)}) = O(H_{(n-t-j)}^{(i)})$를 구한다.

3 단계: 2 단계에서 얻어진 목적함수 값들에서 가장 큰 목적함수 값을 갖는 칸 $H_{(n-t)}^*$을 결정한다. 즉, $O(H_{(n-t)}^*) = \max\{O(H_{(n-t)}^{(i)}),\ i = 1, 2, \ldots, n-t\}$.

4 단계: $t = t+1$로 증가시키고 2 ~ 3 단계를 H_{n-t}^*의 크기가 $h+1$에 이를 때까지 반복하고, $t = h+1$이면 목적함수 값 $O(H_{(n-t)}^*)$을 갖는 최대절사우도 추정량을 얻을 수 있는 H_h^*를 결정한다.

위 추정 알고리즘에서 얻어진 H_h^*을 통해 주어진 h에 대한 최대절사우도 추정값을 얻을 수 있다. 그러나 제안된 추정 알고리즘 역시 완전 탐색을 수행하는 것이 아니라 우도에 의존하여 탐색 공간을 줄여나가며 H_h^*을 결정하기 때문에 목적함수 값이 국소 최대값일 수 있다. 제안된 방법을 통한 추정값이 완전 탐색에 의한 추정값에서 얻을 수 있는 최대 목적함수를 얼마나 정확히 얻게 하는지에 관한 사실은 3절의 모의실험 결과를 통해 설명하기로 한다.

제안된 추정 방법은 우도에 의존하여 초기칸 집합을 선정하고, 이들 초기칸 집합의 크기를 순차적으로 줄여나간다. 그러므로 요구되는 최우추정의 수는 n과 h에 의존하여 다음과 같은 식에 의하여 얻을 수 있다.

$$\sum_{t=h+2}^{n} \sum_{i=h+1}^{t-1} (t \times i). \tag{2.4}$$

표 1: 고고학 유물 자료

	바로 인근	1/4마일이내	1/4~1/2마일	1/2~1마일
송곳	2	10	4	2
항아리	3	8	4	6
연마석	13	5	3	9
칼끝 조각	20	36	19	20

이 식으로부터 제안된 방법에서 고려하는 칸 집합의 수는 완전 탐색에서 요구되는 칸 집합의 수 보다 매우 작다는 것을 예상할 수 있다. 또한 완전 탐색에서는 비교적 n과 h가 작게 변하여도 반복수가 크게 증가하는데 반하여 제안된 방법은 약 $n \times h$ 만큼의 배율로 변하게 된다. 5×5 분할표를 예로 들면 $n = 25$이고 $h = 20$일 때 총 5,285개 칸에서 최우추정이 이루어지는 반면, 완전 탐색에서는 53,130개의 칸에 대한 추정이 이루어져야 한다. 그러나 만일 $h = 17$이라면 제안된 방법은 12,922개 칸 집합에 대하여 추정을 수행하는데 반하여 완전 탐색을 위해서는 1,081,575개 칸 집합에서 추정이 이루어져야 하므로 계산 부담이 상당히 크게 증가하게 된다.

3. 실제 자료분석 예

제안한 추정 방법의 이해를 위하여 Shane과 Simonoff (2001)에서 발췌한 표 1의 고고학 유물(archaeological artifact) 자료를 통해 앞에서 제안한 추정 방법을 설명하기로 한다. 이 자료는 Mosteller와 Parunak (1985)이 관찰칸 값에 로그를 취한 분할표에 median polish 방법을 적용하여 이상칸을 식별하기 위해 인용한 자료중 일부분만으로 구성된 자료이다. 분할표는 송곳(drills), 항아리(pots), 연마석(grinding stones), 칼끝 조각(point fragments)의 네가지 유물이 상존하는 물(permanent water)에서 떨어진 거리에 의해 교차분류되었다.

이 자료의 연관관계를 식별하기 위하여 모형 (2.1)의 특수한 경우인 다음과 같은 독립성 모형(independence models)을 고려하기로 한다.

$$\log m_{ij} = \theta + \theta_{1(i)} + \theta_{2(j)},$$

여기서 $i = 1, 2, \ldots, 4$, $j = 1, 2, \ldots, 4$이며 모형의 식별을 위해 행 효과 모수에는 $\sum_{i=1}^{4} \theta_{1(i)} = 0$ 그리고 열 효과 모수들에는 $\sum_{j=1}^{4} \theta_{2(j)} = 0$과 같은 제약조건을 부여하기로 한다.

위 모형의 최대절사우도 추정값을 구하기 위한 절사 모수는 $p \le h \le n$ 중 한 값을 선택할 수 있다. 이 값들 중에서 Neykov와 Muller (2002) 그리고 Rousseeuw와 Driessen (2006) 등은 $(n + p + 1)/2$을 제안하였고, 이 경우의 최대 붕괴점은 Mili와 Coakley (1996)에 의하여 LMS 추정량과 같이

$$\frac{1}{n}\left[\frac{n - \mathrm{N}(X) + 1}{2}\right]$$

임이 알려져 있다. 여기서 $\mathrm{N}(X)$는 분할표에 부여된 제약조건에 의한 상수이며, $[\cdot]$는 정수 연산자이다. 따라서 표 1의 4×4 분할표는 $n = 16$이고, 독립성모형의 모수의 수 $p = 1 + 3 + 3 = 7$이므로 $h = 12$가 된다. 또한 $\mathrm{N}(X) = n - 4 = 12$이므로 4×4 분할표를 위한 독립성 모형의 붕괴점은 2/16로 2개 이상의 이상칸이 존재할 경우에 최대절사우도 추정량은 이상칸의 영향에 의해 왜곡될 수 있다.

표 1은 4×4 분할표로 $n = 16$개 칸으로 구성되어 있다. 이들 칸에 대하여 행에 우선 번호를 부여하여 칸 집합 $H_{16}^{*} = \{1, 2, \ldots, 16\}$을 구할 수 있다. 이로부터 제안된 방법의 2 단계에서 크기가 $n - 1 = 15$인 16개 부분 관찰칸들의 집합 $H_{15}^{(1)} = \{1, 2, \ldots, 16\}$, $H_{15}^{(2)} = \{1, 3, \ldots, 16\}$, …, $H_{15}^{(16)} = \{1, 2, \ldots, 15\}$를 얻을 수 있다. 이들 16개의 부분 관찰칸 집합 H_{15}에 대하여 각각 2 단계의 하위 반복을 수행하면 3 단

표 2: 최대절삭우도 추정값에 의한 잔차와 표준화 잔차

	바로 인근	1/4마일이내	1/4~1/2마일	1/2~1마일
송곳	−0.3077	1.9054	0.2380	−1.8357
	(−0.2025)	(0.6697)	(0.1277)	(−0.9373)
항아리	0.3077	−1.4437	−0.3890	1.5250
	(0.1875)	(−0.4698)	(−0.1857)	(0.7209)
연마석	11.4429	−0.4617	0.4617	6.4119
	(9.1703)	(−0.1976)	(0.2898)	(3.9857)
칼끝 조각	8.1543	−5.5507	−0.3107	0.3107
	(2.3692)	(−0.8611)	(−0.0707)	(0.0700)

계에서 칸 {13}이 제거된 $H^*_{15} = \{1, 2, \ldots, 12, 14, 15, 16\}$을 선택할 수 있다. 이때 $O(H^*_{15}) = -22.5255$가 된다.

이제 H^*_{15}을 기반으로 2, 3 단계의 반복을 수행하면, H^*_{15}에서 칸 {9}가 제거된 $H^*_{14} = \{1, 2, \ldots, 8, 10, 11, 12, 14, 15, 16\}$을 얻고 앞의 반복과 마찬가지로 $O(H^*_{14}) = -22.5255$를 얻을 수 있다. 마지막으로 같은 반복을 수행하면 $O(H^*_{13}) = -22.5255$인 $H^*_{13} = \{1, 2, \ldots, 8, 10, 11, 14, 15, 16\}$를 얻게 되고, 완전 탐색과 같은 칸 {14}가 제거된 $H^*_{12} = \{1, 2, \ldots, 8, 10, 11, 14, 16\}$을 얻을 수 있다.

한가지 주목할 점은 이 자료에서는 $O(H^*_{15})$와 $O(H^*_{14})$가 같게 얻어졌다. 즉, H^*_{15}를 통해서 선택된 H_{12}가 H^*_{12}와 같다. 이러한 사실은 제안된 방법 역시 FAST-TLE와 마찬가지로 다수의 중복 계산이 발생할 수 있다는 것을 의미한다. 그러나 주어진 자료에 따라서 H^*_{15}에 의해 추정된 최우추정값에 따른 우도의 순위와 H^*_{14}에 의한 우도의 순위는 다를 수 있기 때문에 $O(H^*_{15})$와 $O(H^*_{14})$는 항상 같을 수 없다.

Mosteller와 Parunak (1985)은 자신들이 제안한 방법에 의해 (3, 1) 칸 즉, (연마석, 바로 인근) 칸이 독립성 모형하에서 이상칸 이라는 것을 식별하였고, 이러한 식별이 자료가 가진 특성을 설명하는데 중요한 사실임을 지적하고 있다. H^*_{12}에 의해 추정된 최대절사우도 추정값에 의한 잔차와 표준오차가 정리된 표 2에서 볼 수 있듯이 최대절사우도 추정값에 의한 (3, 1) 칸의 표준화 잔차 역시 9.1703으로 매우 큰 것을 볼 수 있다. 이러한 사실은 Mosteller 등의 결과와 같이 (3, 1) 칸이 이상칸일 가능성이 있다는 것으로 해석할 수 있다. 또한 (3, 4) 칸의 표준화 잔차 역시 3.9857로 비교적 크기 때문에 Shane과 Simonoff (2001)가 지적한 바와 같이 이 칸이 (3, 1) 칸과 함께 이상칸일 가능성이 있는 것으로 분석할 수 있다.

4. 모의 실험

제안된 추정방법의 정확성을 평가하기 위하여 Shane과 Simonoff (2001)와 최현집 (2003)에서 사용한 5×5 분할표를 위한 독립성 모형을 이용하였다. 모의실험을 위한 주변분포(marginal distribution)는 주변분포의 영향에 따른 특징을 평가하기 위해 행과 열 모두 {0.10, 0.15, 0.20, 0.25, 0.30}를 이용하였다. 따라서 생성된 임의 분할표의 왼쪽 상단에 속한 칸들은 작은 칸 값을 갖고 오른쪽 하단에 속한 칸들은 큰 칸 값을 갖게 된다. 이와 연계하여 이상칸의 수에 따른 영향 역시 평가하기 위하여 다음과 같은 네 가지 경우에 $N = 500$인 분할표를 각각 100번 생성하여 2절에서 제안된 추정방법과 완전 탐색에 의한 추정을 수행하였다. 두 방법에 의한 추정값의 차이 여부는 계산오차를 감안하여 두 방법에 의한 목적함수 값이 5E-8보다 작으면 동일한 것으로 평가하였으며, 오염된 각 칸에는 칸 도수에 추가로 100을 더하였다.

1. (1, 2) 칸이 오염된 경우

2. (1, 2), (2, 1) 칸이 오염된 경우

표 3: 한 칸이 오염된 경우에 대한 모의실험 결과

절사모수	오염칸	도수	계산시간		반복횟수		
			완전탐색	제안방법	완전탐색	제안방법	비율
23	(1,2)	100	0.6847	1.4049	300	600	2.0000
	(4,5)	100	0.6728	1.3184			
22	(1,2)	100	4.8354	3.4235	2,300	1,727	0.7509
	(4,5)	100	4.9872	3.6272			
21	(1,2)	99	28.1430	6.8106	12,650	3,311	0.2617
	(4,5)	99	27.3593	6.8817			
20	(1,2)	97	112.9572	10.1436	53,130	5,285	0.0995
	(4,5)	99	112.1662	10.5973			

3. (5,4) 칸이 오염된 경우

4. (5,4), (4,5) 칸이 오염된 경우

모의실험에 사용된 5×5 분할표의 독립성 모형에 포함된 모수의 수 $p = 9$이므로 절사모수는 $9 \le h \le 23$ 중에 한 값을 선택할 수 있다. 이 값들 중 Rousseeuw와 Driessen (2006)에 의해 제안된 절사모수 $h = 17$이고 이때 최대 붕괴점은 3이 된다. 그러나 2절에서 언급한 바와 같이 이 경우 완전 탐색은 총 1,081,575개 칸 집합에 대한 추정이 이루어져야 한다. 최우추정은 일반적인 상황을 고려하여 반복가중 최소제곱 추정방법을 이용하였으므로 모든 칸 집합에 대하여 추정을 수행할 경우 계산의 양은 훨씬 더 증가할 것이다. 따라서 한번 생성된 분할표에 대한 완전 탐색에만도 상당한 시간이 필요하므로 5×5 분할표 임에도 $h = 17$일 때의 정확한 최대절사우도 추정값을 얻기는 매우 어렵게 된다. 따라서 모의실험에서는 계산의 양을 고려하여 $h = 20, 21, 22, 23$의 네 값을 이용하였다.

먼저, 표 3에 한개의 칸이 오염된 경우의 모의실험 결과를 정리하였다. 표 3에서 도수는 완전 탐색에 의한 추정결과와 제안된 방법의 결과가 일치한 도수를 의미한다. 또한 계산시간은 완전 탐색에 소요된 시간과 제안된 추정방법을 수행한 평균 계산시간을 나타낸다. 모의실험은 Intel Core2 Duo CPU가 탑재된 컴퓨터에서 수행되었다. 그리고 반복횟수는 두 방법에서 필요로하는 칸 집합들의 수를 의미한다.

우선 제안된 방법의 추정값은 거의 정확하게 완전 탐색에 의한 추정값과 일치하는 것을 볼 수 있다. 다만 절사모수 $h = 20$인 경우에는 약 97 ~ 99%의 정확성을 보이며, 이러한 사실로부터 h가 20보다 작은 경우에는 즉, 추정하기 위해 필요한 부분칸 집합의 수가 증가할 수록 정확도는 감소할 것이라고 예상할 수 있다. 그러나 제안된 방법에서 필요로하는 계산시간은 h가 작아질수록 급격하게 줄어든다. $h = 20$이고 칸 (4,5)가 오염된 경우에 제안된 방법과 완전 탐색에 소요되는 계산시간의 비율은 0.0945로 제안 방법이 완전 탐색에 비해 약 9.5% 정도의 시간만을 필요로 한다. 다만, 표의 첫 행에서 절사 모수가 붕괴점에 가까운 값을 제거한 크기를 가질 경우에는 제안된 방법의 반복횟수가 오히려 큰 것을 볼 수 있다. 이는 2절에서 지적하였듯이 제안된 방법 역시 FAST-TLE와 마찬가지로 중복계산을 포함한다는 것을 의미한다. 따라서 식 (2.4)에 의하여 주어진 n과 h에 의해서 반복수를 확정적으로 계산할 수 있으므로, 이 값과 완전 탐색에서 요구되는 계산 수를 비교하여 그 크기에 따라 두 방법을 적절히 선택하여 효율성을 얻는 계산 전략이 필요할 것이다. 그러나 h가 커지면 상대적으로 이상칸의 영향을 받을 가능성이 높아지게 되며, 이러한 이유로 Rousseeuw와 Driessen (2006) 등은 $h = 0.75n \simeq 17$을 제안하고 있다.

칸 (1,2)는 모의실험에서 주변분포에 의해 칸 비율이 작은 칸이고, 반면에 칸 (4,5)는 칸 (1,2)에 비하여 칸 비율이 크다. 이는 칸 (1,2)에는 도수 100을 추가하여 오염시켰을 경우에 칸 (4,5)에 비하여 오

표 4: 두 칸이 오염된 경우의 모의실험 결과

절사모수	오염칸	도수	계산시간
23	(1,2), (2,1)	100	2.0960
	(4,5), (5,4)	100	1.9720
22	(1,2), (2,1)	100	0.7508
	(4,5), (5,4)	100	0.7213
21	(1,2), (2,1)	100	0.2401
	(4,5), (5,4)	100	0.2473
20	(1,2), (2,1)	99	0.0902
	(4,5), (5,4)	98	0.0929

염의 정도가 심한 것을 의미한다. 그러나 표 4에서 볼 수 있듯이 이러한 주변분포의 영향은 거의 나타나지 않는 다는 사실을 확인할 수 있다.

표 3은 두 칸이 오염되었을 때의 모의실험 결과를 보여준다. 한 칸이 오염된 경우와 마찬가지로 두 방법에 의한 추정값은 거의 완벽하게 일치하는 것을 확인할 수 있다. 그러나 한개 칸이 오염된 경우과 마차가지로 계산수가 커지면 일치되는 도수가 감소하는 동일한 현상이 발생한다. 절사 모수가 같기 때문에 요구되는 계산횟수는 한개 칸이 오염된 경우와 같다. 따라서 계산시간의 변화 역시 한개 칸이 오염된 경우와 같이 h가 감소함에 따라 급격히 감소하게 된다. 마지막으로 주변분포에 따른 오염된 칸의 위치 역시 한개 칸이 오염된 경우와 마찬가지로 정확도에 큰 영향을 주지 않는다는 것을 알 수 있다.

5. 결론

분할표 분석을 위한 최대절사우도 추정량의 추정을 위하여 본 연구에서 제안된 방법은 반복가중최소제곱 추정법을 이용하고 있다. 따라서 로그선형모형이 아닌 범주에 순위가 부여된 균일연관모형(uniform association models)과 같은 여러 연관성 모형에도 쉽게 적용될 수 있다. 특히, 반복가중최소제곱 추정법은 R, SAS, SPSS 등 일반적인 통계자료분석 환경에서 어렵지 않게 구현될 수 있기 때문에 최대절사우도 추정을 필요로 하는 연구자들에 의해 쉽게 구현되어 적용할 수 있을 것이다. 참고로 자료분석 예와 모의실험은 R로 구현하여 수행하였다.

모의실험 결과에서 살펴본바와 같이 제안된 방법은 계산의 양이 많지 않음에도 상당히 정확한 추정을 수행한다. 다만, 중복계산이 포함되는 영향으로 절사모수가 큰 경우에는 완전 탐색에서 요구되는 계산의 양에 비해 증가하는 문제를 가진다. 그러나 이러한 경우에는, 제안된 방법의 계산횟수는 확정적으로 얻을 수 있기 때문에, Rousseeuw와 Driessen (2006)이 지적한 바와 같이 두 방법의 계산횟수를 비교하여 분석을 수행하는 계산 전략을 이용할 수 있다. 분할표 자료인 경우에는 4절의 모의실험 결과에 의하여 붕괴점 보다 작은 수의 칸이 제거될 경우에는 완전탐색 방법을 이용하고 그렇지 않은 경우에는 제안된 방법을 이용하는 전략을 제안할 수 있다.

또한 제안된 방법은 주변분포 그리고 칸의 위치에 영향을 받지 않는다. 그러므로 우도에 의존하여 순차적인 제거를 통해 최대절사우도 추정값을 얻게하는 제안된 방법은 구현의 용이성과 정확성 등에 의해 일반적인 분할표 분석을 위해 어렵지 않게 적용될 수 있을 것이다.

6. 감사의 글

논문의 심사를 위해 수고하여 주신 편집위원장, 편집위원 그리고 익명의 심사위원들께 감사드립니다.

참고 문헌

최현집 (2003). 범주형 자료 분석을 위한 LAD 추정량, <응용통계연구>, **16**, 55–69.

Cheng, T. and Biswas, A. (2008). Maximum trimmed likelihood estimator for multivariate mixed continuous and categorical data, *Computational Statistics and Data Analysis,* **52**, 2042–2065.

Čížek, P. (2006). Trimmed likelihood-based estimation in binary regression models, *Austrian Journal of Statistics,* **2 & 3**, 223–232.

Grizzle, J. E., Stamer, C. F. and Koch, G. G. (1969). Analysis of categorical data by linear models, *Biometrics,* **25**, 489–504.

Hadi, A. S. and Luceño, A. (1997). Maximum trimmed likelihood estimators: A unified approach, examples, and algorithms, *Computational Statistics and Data Analysis,* **25**, 251–272.

Mili, L. and Coakley, C. W. (1996). Robust estimation in structured linear regression, *The Annals of Statistics,* **24**, 2593–2607.

Mosteller, F. and Parunak, A. (1985). Identifying extreme cells in a sizable contingency table: Probabilistic and exploratory approaches, In *Exploring Datat Tables, Trends and Shapes,* 189–225.

Neykov, N. and Muller, C. H. (2002). Breakdown point and computation of trimmed likelihood estimators in generalized linear models, In *Developments in Robust Statistics,* 277–286.

Neykov, N., Filzmoser, P., Dimova, R. and Neytchev, P. (2007). Robust fitting of mixtures using the trimmed likelihood estimator, *Computational Statistics and Data Analysis,* **52**, 299–308.

Rousseeuw, P. J. (1984). Least median of squares regression, *Journal of the American Statistical Association,* **79**, 871–880.

Rousseeuw, P. J. and Driessen, K. (2006). Computing LTS regression for large data sets, *Data Mining and Knowledge Discovery,* **12**, 29–45.

Shane, K. V. and Simonoff, S. S. (2001). A Robust approach to categorical data analysis, *Journal of Computational and Graphical Analysis,* **10**, 135–157.

2009년 2월 접수; 2009년 2월 채택

Maximum Trimmed Likelihood Estimator for Categorical Data Analysis

Hyunjip Choi[1,a]

[a]Dept. of Applied Information Statistics, Kyonggi Univ.

Abstract

We propose a simple algorithm for obtaining MTL(maximum trimmed likelihood) estimates. The algorithm finds the subset to use to obtain the global maximum in the series of eliminating process which depends on the likelihood of cells in a contingency table. To evaluate the performance of the algorithm for MTL estimators, we conducted simulation studies. The results showed that the algorithm is very competitive in terms of computational burdens required to get the same or the similar results in comparison with the complete enumeration.

Keywords: Contingency table, outlying cell, maximum trimmed likelihood estimator.

[1] Professor, Department of Applied Information Statistics, Kyonggi University, 94-6 Yiui-Dong, Yeongtong-Gu, Suwon, Kyonggi-Do 449-791, Korea. E-mail: hjchoi@kyonggi.ac.kr