응용통계연구 (2010)
23(6), 1235–1243
DOI: 10.5351/KJAS.2010.23.6.1235

분할표 분석을 위한 절사 LAD 추정량과 최적 절사율 결정

최현집[1]

[1]경기대학교 응용정보통계학과

(2010년 10월 접수, 2010년 11월 채택)

요 약

다차원 분할표를 구성하는 범주형 변수들의 연관관계를 식별하기 위하여 널리 이용되는 로그선형모형을 위한 절사 LAD(least absolute deviations) 추정방법을 제안하였다. 제안된 방법은 가중 LAD 추정을 반복하여 계산이 수행되므로 분할표 분석을 위해 적용할 수 있는 여러 연관성 모형(association models)에 직접 적용할 수 있다. 또한 붓스트랩을 이용한 최적절사율을 결정하는 방법이 갖는 공분산행렬을 과소추정하는 문제를 해결하기위한 절사율 결정 방법을 제안하였다. 모의실험을 통해 제안된 방법이 붓스트랩 방법에 비하여 항상 우수한 절사율을 보인다는 것을 설명하였으며, 제안된 방법들의 실제 자료분석 결과를 제시하였다.

주요용어: 다차원 분할표, 로그선형모형, 절사 LAD 추정량, 최적절사율.

1. 서론

다음과 같은 d개 칸을 갖는 D차원 분할표를 구성하는 범주형 변수들의 연관관계를 식별하기 위한 로그선형모형(log-linear models)을 고려하기로 한다.

$$\log \underline{m} = \boldsymbol{X}\underline{u}, \tag{1.1}$$

여기서 $\underline{m} = \{m_i\}$, $i = 1, 2, \ldots, d$는 $d \times 1$차 기대칸 값(expected cell counts) 벡터, $\boldsymbol{X}$는 모형을 위해 정의된 $d \times p$차 계획 행렬(design matrix)이며, 추정을 위해 행 효과(row effects)와 열 효과(column effects) 등을 나타내는 모수 u들에는 각각 행과 열에 대해 합하면 '0'이 되는 등의 제약조건이 부여된다.

위 모형에 포함된 $\underline{u}$를 추정하기 위하여 가장 널리 이용되고 있는 방법으로 Grizzle 등 (1969)이 고려한

$$\sum_{i=1}^{d} \omega_i (\log n_i - \log m_i)^2 \tag{1.2}$$

을 최소로 하는 가중 최소제곱(weighted least squares) 방법을 이용한 반복 추정법을 고려할 수 있다. Grizzle 등 (1969)은 $\log n_i$의 근사분산이 $1/m_i$라는 사실로부터 $\omega_i \equiv \hat{m}_i$로 할 것을 제안하였고, 이러한 가중값을 이용한 추정값이 근사적으로 최우추정값과 동등함을 보였다.

그러나 최우추정방법은 주어진 모형과 일치된 적합을 나타내지 않는 이상칸(outlying cells)의 영향을 받기 때문에 이상칸의 영향에 덜 민감한 여러 로버스트 추정방법들이 제안되고 있다. 특히 최현집 (2003)은 식 (1.2)로부터

$$\sum_{i=1}^{d} \omega_i \left|\log n_i - \log m_i\right| \tag{1.3}$$

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

을 최소로하는 가중 LAD(least absolute deviation) 추정량을 제안하였으며, 이는 로그를 취한 관찰칸 값, $\underline{n} = \{n_i\}$과 기대칸 값에 의한 잔차의 절대값에 적절한 가중값 ω_i가 부여된 가중 절대 잔차합(sum of weighted absolute deviations)을 최소로 하는 추정방법이 된다. 이 이외의 LAD 추정방법은 최현집 (2003)을 참고할 수 있다.

이들 방법과는 달리 로그선형모형의 추정을 위한 로버스트 방법으로 우도함수 값을 절사한 후에 최우추정량을 얻는 Hadi와 Luceño (1997)가 제안한 최대절사우도 추정방법을 고려할 수 있다. 같은 맥락에서 Godaliza (1991)가 제안한 절사중앙값을 확장한 식 (1.3)을 절사한 다음의 목적함수를 최소로하는 추정량을 고려할 수 있다.

$$\sum_{i=1}^{h} \omega_{(i)} \left| \log n_{(i)} - \log m_{(i)} \right|, \tag{1.4}$$

여기서 $h = [n(1-\alpha)]$로 $[a]$은 a보다 큰 가장 작은 정수로서 α는 절사율, $\omega_{(1)}|\log n_{(1)} - \log m_{(1)}| \leq \omega_{(2)}|\log n_{(2)} - \log m_{(2)}| \leq \cdots \leq \omega_{(h)}|\log n_{(h)} - \log m_{(h)}|$는 가중값이 부여된 편차의 절대값을 크기순으로 정렬한 값을 나타내다. 식 (1.4)에 의한 추정량은 Vandev (1995)가 지적하였듯이 절사중앙값이 갖는 여러가지 장점, 즉 다수의 이상칸들이 포함된 경우에 LAD 추정량에 비하여 더욱 로버스트하다는 장점을 갖을 수 있다는 것을 예상할 수 있다. 그러나 목적함수 자체가 다수의 국소최소값(local minimum)을 갖기 때문에 추정에 어려움이 있을 수 있다. 또한 절사율에 따른 추정에 포함되는 칸의 수 h는 $p \leq h \leq d$ 사이의 값을 사용할 수 있기 때문에 적절한 h를 결정하는 문제 역시 고려되어야 한다.

다변량 절사중앙값을 위한 h를 결정하기 위해 다수의 붓스트랩 표본에서 추정된 값들에 의한 추정 공분산 행렬을 이용하는 방법을 고려할 수 있다. 그러나 붓스트랩 추정 공분산 행렬은 과소추정되는 문제를 가지게 되어, 즉 적은 h값이 결정되는 단점을 가지고 있다. 이동희와 정병철 (2010)은 이러한 문제를 보완하기 위하여 이중 붓스트랩에 의한 최적절사율 결정방법을 제안하였으나 계산의 양이 매우 많다는 단점을 가지게 된다. 특히 범주형 자료 분석을 위해 고려할 수 있는 목적함수 (1.4)를 최소로하는 추정량은 반복계산을 통해 계산이 이루어지기 때문에 붓스트랩을 이용할 경우에 계산의 양은 지수적으로 증가하여 일반 연구에 직접 적용하는데는 많은 어려움이 따른다는 것을 예상할 수 있다.

본 연구에서는 식 (1.4)를 최소로하는 절사 LAD 추정량의 계산방법에 관하여 연구하고자 한다. 이를 위해 먼저 Vandev (1995)가 제한한 계산방법을 확장한 추정방법을 제안할 것이며, 여러 절사율에 의한 추정값들의 특징에 관하여 실제 자료분석 결과를 통해 설명하고자 한다. 이를 바탕으로 이상칸의 영향에 가장 덜 민감한 최적절사율을 결정하기 위한 방법을 제안하고자 한다. 제안된 방법은 공분산 행렬을 이용하는 방법이 갖는 과소추정 문제를 해결하기 위하여 절사를 통해 얻는 불완전한 분할표(incomplete tables)의 적합성(goodness of fits)을 이용한 측도를 사용하게 된다. 제안된 방법의 특징을 설명하기 위하여 모의실험을 수행하였으며, 공분산 행렬을 이용한 방법에 비하여 보다 적절한 절사율을 갖고 또한 이상칸의 영향에 덜 민감하다는 특징을 갖는 다는 것을 설명하였다.

이 논문의 구성은 다음과 같다. 2절에서 절사 LAD 추정량을 위한 반복계산법을 제안하였으며, 실제 자료분석 예를 통해 추정결과가 갖는 여러가지 특징에 관하여 설명하였다. 3절에서는 최적절사율을 결정하기 위한 방법을 제안하고, 모의실험을 통하여 기존에 제안된 방법과 비교하여 제안된 방법이 갖는 특징에 관하여 설명하였다. 마지막으로 4절에서는 본 연구의 결과를 정리하고 연구결과에 대하여 토론하였다.

2. 절사 LAD 추정을 위한 반복계산법

일반적으로 LAD 추정을 위하여 가장 널리 사용되는 방법은 Schlossmacher (1973) 등이 제안한 IR-

WLS(iterative reweighted least squares) 방법과 Barrodale과 Roberts (1973) 등이 제안한 LP(linear programming) 방법을 확장한 방법 등이 제안되고 있으며, 두 방법 모두 경사법(gradient method)에 기반을 둔 방법이라고 설명될 수 있다. 이제 이들 방법을 이용하여 식 (1.4)를 최소로하는 절사 LAD 추정량을 계산하기 위한 계산방법을 고려해보기로 한다.

먼저, h가 주어져 있다면 최현집 (2003)에서 제안된 방법을 응용하여 절사 LAD 추정값을 얻을 수 있다. 다시말해 모든 칸에 가중값 $\omega_i \equiv 1$을 부여된 LAD 추정값을 얻고, 이때 얻어진 모수 추정값에 의한 기대칸값 $\omega_i = \hat{m}_i$을 가중값의 초기값으로 얻는다. 다음으로 이 값들에 의한 가중 절대편차값 $\omega_i|\log n_i - \log m_i|$을 구하여 정렬하고 주어진 h에 따라 절사한 후에 절사되고 남은 칸들에 대하여 가중 LAD 추정값을 얻는다. 이때 절사된 칸은 구조적 영값(structural zero)으로 취급할 수 있기 때문에 추정된 가중 LAD 추정값은 불완전한 분할표를 위한 추정값이 된다. 이렇게 얻어진 추정값을 이용하여 모든 칸에 대한 기대칸값을 얻고 이들 값에 의하여 가중 절대편차값 $\omega_{(i)}|\log n_{(i)} - \log m_{(i)}|$을 구하고 h에 따라 절사한 후에 추정하는 과정을 반복 수행하여 그 수렴값인 절사 LAD 추정값을 얻을 수 있다. 그러나 최현집 (2003) 등에서 지적하였듯이 이러한 계산 방법은 초기값에 대단히 민감하기 때문에 다수의 이상칸이 포함되어 초기값이 $\omega_i \equiv 1$인 추정량이 붕괴되면 절사 LAD 추정량 역시 붕괴될 수 있다는 단점을 가지게 된다.

추정의 대상이 되는 로그선형모형에 포함된 모수 중에서 일반선형모형의 절편에 해당되는 첫번째 모수인 u-항은 분할표의 총수와 관련된 모수가 된다. 다시말해 u-항의 최우추정값은 로그를 취한 관찰칸값의 평균의 함수로 설명할 수 있다. 따라서 절편만이 포함된 로그선형모형의 LAD 추정값은 로그를 취한 관찰칸값의 중앙값이 되는 것을 알 수 있다. 이 경우에는 단순한 중앙값과 같은 붕괴점을 갖기 때문에 다수의 모수가 포함된 로그선형모형의 가중 LAD 추정량에 비해 상당히 로버스트하다고 예상할 수 있다. 따라서 이런 사실을 기반으로 다음과 같은 절사 LAD 추정량을 위한 계산방법을 제안하기로 한다.

Step 1. 절삭 모수 h을 결정하고 $\sum_{i=1}^{d}|\log n_{(i)} - u|$를 최소로 하는 $\hat{u}$를 얻어 $k = 0$인 초기 가중값 $\omega_i^{(0)} = \hat{m}_i^{(0)} = e^{\hat{u}}$, $i = 1, 2, \ldots, d$를 얻는다.

Step 2. 얻어진 가중값에 의하여 $(k+1)$번째 반복을 위하여 $\omega_i^{(k)}|\log n_i - \log \hat{m}_i^{(k)}|$를 구하고 정렬한 후에 주어진 h에 의하여

$$\sum_{i=1}^{h} \omega_{(i)}^{(k)}|\log n_{(i)} - \log m_{(i)}|, \tag{2.1}$$

를 최소로하는 모형 (1.1)의 추정값 $\underline{\hat{u}}^{(k+1)}$을 얻는다. 이렇게 얻어진 추정값에 의하여 모든 칸에 대한 기대칸 값인 $\underline{\hat{m}}^{(k+1)} = \exp(\boldsymbol{X}\underline{\hat{u}}^{(k+1)})$을 구하고 가중값 $\omega_i^{(k+1)} = \hat{m}_i^{(k+1)}$, $i = 1, 2, \ldots, d$를 얻는다.

Step 3. $k = k + 1$로 한다.

Step 4. 2~3 단계를

$$\sum_{i=1}^{h}\left|\hat{m}_{(i)}^{(k+1)} - \hat{m}_{(i)}^{(k)}\right| < \epsilon \sum_{i=1}^{h}\left|\hat{m}_{(i)}^{(k+1)}\right|$$

이 만족될 때까지 반복한다.

제안된 계산방법은 Vandev (1995)의 계산방법을 직접 확장한 형태로 만일 식 (2.1)의 추정을 위하여 IRWLS 방법이 이용되었다면, Vandev가 사용한 Newtop-Raphson 알고리즘의 해를 가중값으로 재 부

표 **2.1.** 4×4 고고학 유물자료의 관찰칸값과 표준화 잔차

	바로 인근	1/4마일이내	1/4∼1/2마일	1/2∼1마일
송곳	2 (−1.0773)	10 (0.8794)	4 (0.0000)	2 (−1.0773)
항아리	3 (−0.6852)	8 (0.0000)	4 (−0.1081)	6 (0.7379)
연마석	13 (6.1333)	5 (0.0000)	3 (0.2223)	6 (3.7333)
칼끝 조각	20 (0.0000)	36 (0.0000)	19 (0.0000)	20 (0.0000)

표 **2.2.** 여러 절사율에 따른 절사칸들과 공분산행렬의 행렬식 값

h	절사된 칸들	행렬식 값
16	-	6.08261e-08
15	(3, 1)	1.91172e-09
14	(3, 1), (3, 4)	6.58969e-10
13	(1, 1), (2, 2), (3, 4)	2.73856e-10
12	(1, 1), (2, 2), (2, 4), (4, 4)	1.47584e-07

여한 형태가 된다. 특히 4번째 반복에서의 수렴기준 역시 경사도를 이용한 반복계산의 수렴성 평가와 유사하며, 이 값에 대한 설명은 최현집 (2003)을 참고할 수 있다.

예제 2.1: Shane과 Simonoff (2001)에서 발췌한 고고학 유물(archaeological artifact) 자료는 분할표를 위한 로버스트 추정방법의 적용 예를 위하여 가장 많이 이용되는 자료이다. 특히 Mosteller와 Parunak (1985) 등은 자신들이 제안한 방법에 의해 $(3,1)$ 칸이 독립성 모형하에서의 이상칸이라는 것을 식별하였고, Shane과 Simonoff (2001) 그리고 최현집 (2003) 역시 같은 결과를 제시하고 있다. 또한 이들은 $(3,1)$ 칸에 더하여 $(3,4)$ 칸 역시 이상칸일 가능성이 있다는 것을 지적하였다.

표 2.1은 고고학 유물자료의 관찰칸값을 나타내고 있으며 관찰칸값 아래 괄호안의 값은 표준회된 잔차 $(n_i - \hat{m}_i)/\sqrt{\hat{m}_i}$를 나타낸다. 표 2.1에서 $(3,1)$ 칸의 표준화된 잔차는 6.1333으로 매우 크고, 또한 $(3,4)$ 칸의 표준화된 잔차 역시 3.7333으로 크기 때문에 앞의 연구결과들에서 얻은 결과들과 같은 설명이 가능한 것을 알 수 있다. 표준화된 잔차는 $h = 14$를 이용하여 얻었다.

이 자료에 대한 최적절사율을 결정하기 위하여 여러 h값에 따른 추정값 $\hat{\underline{u}}$의 붓스트랩 방법에 의한 추정 공분산 행렬의 행렬식 값이 표 2.2에 정리되어 있다. 표 2.2의 내용에 따르면 $h = 13$일 경우의 추정 공분산 행렬의 행렬식 값이 2.73856-e10으로 가장 작게 나타난다. 이렇듯 매우 작은 행렬식 값이 얻어지는 것은 표 2.3에서 볼 수 있듯이 모수의 추정 분산과 공분산이 매우 작은 것에 기인한다. 로그선형모형은 모형의 특성상 연관항으로 구성된 모수벡터의 추정값은 로그를 취한 관찰칸값들의 함수로 나타나고, 또한 이들의 분산은 기대칸값의 역수에 로그를 취한 형태로 나타나기 때문에 매우 작은 값을 갖을 수 밖에 없다. 따라서 붓스트랩 표본에 의한 추정값들의 변동폭은 매우 작을 수 밖에 없다는 것을 알 수 있다. 여기에 더하여 표 2.3에서 볼 수 있듯이 $h = 14$일때와 $h = 13$일때의 행렬식 값의 차이는 매우 작다는 것을 알 수 있다. 특히 $h = 13$인 경우에는 $(1,1)$, $(2,2)$, $(3,4)$ 칸이 절사되어 이상칸으로 의심되는 $(3,1)$ 칸이 모형의 추정을 위해 포함되었음을 알 수 있다. 반면에 $h = 14$인 경우에 절사된 칸은 $(3,1)$, $(3,4)$ 칸이 절사되어 이상칸으로 의심되는 두 칸이 모두 절사되었으므로 $h = 13$에 비하여 이상칸의 영향을 받지 않는 매우 안정적인 추정값을 얻을 수 있음을 의미하는 것으로 판단할 수 있다.

표 **2.3.** $h = 13$인 절사 LAD 추정량의 공분산행렬의 추정값

	u	$u_{1(1)}$	$u_{1(2)}$	$u_{1(3)}$	$u_{2(1)}$	$u_{2(2)}$	$u_{2(3)}$
u	0.0259	−0.0067	−0.0089	0.0287	0.0035	−0.0090	−0.0082
$u_{1(1)}$	−0.0067	0.0404	−0.0028	−0.0299	0.0061	−0.0095	−0.0117
$u_{1(2)}$	−0.0089	−0.0028	0.0390	−0.0342	−0.0069	0.0005	0.0061
$u_{1(3)}$	0.0287	−0.0299	−0.0342	0.1080	0.0046	−0.0009	−0.0047
$u_{2(1)}$	0.0035	0.0061	−0.0069	0.0046	0.0513	−0.0147	−0.0100
$u_{2(2)}$	−0.0090	−0.0095	0.0005	−0.0009	−0.0147	0.0355	0.0082
$u_{2(3)}$	−0.0082	−0.0117	0.0061	−0.0047	−0.0100	0.0082	0.0345

표 **3.1.** 여러 절사율에 따른 절사칸들과 p-값

h	절사된 칸들	행렬식 값	p-값
16	-	6.08261e-08	6.89e-07
15	$(3,1)$	1.91172e-09	0.0590
14	$(3,1)$, $(3,4)$	6.58969e-10	0.7601
13	$(1,1)$, $(2,2)$, $(3,4)$	2.73856e-10	0.0571
12	$(1,1)$, $(2,2)$, $(2,4)$, $(4,4)$	1.47584e-07	7.68e-06

표 2.2의 여러 h에 의한 추정값들은 Barrodale과 Roberts (1973)가 제안한 방법을 확장한 Koenker와 d'Orey (1987)의 방법을 사용하였으며, 이 방법은 R 언어의 quantreg 패키지에 포함된 함수 rq()로 구현되어 있다. 따라서 자료의 분석은 제안된 알고리즘의 목적함수 (2.1)을 위한 추정은 함수 rq()을 이용하였다. 특히 함수 rq()에서는 붓스트랩 방법을 이용한 추정모수의 추정 공분산행렬 역시 계산해주며, 표 2.3은 함수 rq()로부터 얻은 값을 계산한 결과이다.

3. 최적 절사율 추정

2절에서 지적한 바와 같이 로그선형모형을 위한 절사 LAD 추정량의 추정 공분산행렬은 매우 작은 값으로 구성되기 때문에 행렬식 역시 매우 적은 값을 가질 수 밖에 없다. 따라서 여러 h값에 따른 이들 행렬식의 차이는 더욱 작은 값을 갖게 되고 결국 이 값에 의존한 절사율의 결정은 2절의 예제에서 살펴본 봐와 같이 상당히 이상칸에 의하여 왜곡된 절사 LAD 추정값을 얻을 가능성을 포함하고 있다.

여기서 다시한번 2절에서 제안된 추정방법을 살펴보기로 하자. h가 주어졌다면 절사된 칸들은 구조적인 영값으로 취급되기 때문에 반복은 절사되지 않은 칸들에 의한 불완전한 분할표에 대한 적합이 이루어진다. 제안된 계산방법은 이들 칸을 대상으로 가장 적합도가 우수한 절사 LAD 추정값을 계산해주게 된다. 다시 말해 모든 가능한 h에 대하여 절사 LAD 값을 얻고 이들 중에서 가장 적합이 우수한 추정값을 얻게 해주는 h에 의하여 주어진 분할표의 최적 절사율을 얻을 수 있는 방법을 고려할 수 있다.

일반적으로 로그선형모형의 적합성을 측정하기 위해서 우도비검정통계량(likelihodd ratio statistics)이 가장 널리 이용되고 있으며, 최우추정방법에 의한 적합이 아닌 경우에 널리 이용되는 통계량으로 Pearson의 X^2 통계량을 고려할 수 있다. 그러나 주어진 h에 따라 절사된 분할표는 표본의 크기가 변하기 때문에 이 값들을 직접 비교하는 데에는 문제가 있다. 그러나 이러한 문제는 불완전한 분할표의 적합성 평가를 위하여 자유도를 조정하여 해결할 수 있다. 설명을 위하여 다음의 예제를 살펴보기로 한다.

예제 3.1:

표 3.1의 세 번째 열은 여러 h에 의하여 절사된 분할표에 대한 절사 LAD 추정값에 의한 Pearson의 X^2

통계량의 유의확률(p-value)을 보여주고 있다. 예를 들어 절사가 이루어지지 않은 가중 LAD 추정값의 Pearson의 X^2 통계량값은 6.89e-07로 매우 작아 적합이 잘 이루어지지 않은 것으로 판단할 수 있다. 이때의 자유도는 9가 된다. 그러나 절사율 $\alpha = 1/16$인 $h = 15$인 경우에는 $(3,1)$칸이 절사되고 이때 자유도 8인 X^2 통계량값의 유의확률은 0.0590으로 매우 개선되는 것을 알 수 있다. 여기에 한 칸이 더 절사된 절사율 $\alpha = 2/16$인 $h = 14$에서는 자유도 7일때 유의확률은 0.7601로 적합성이 매우 많이 개선되는 것을 알 수 있다. 그러나 반면에 절사율이 더욱 높아진 $\alpha = 3/16$인 $h = 13$에서는 적합성이 오히려 줄어들고 이러한 현상은 이상칸으로 의심되는 칸이 아닌 다른 칸들이 적합에 이용되었기 때문으로 해석할 수 있다.

이렇듯 앞의 예제에서 얻은 사실을 바탕으로 다음과 같은 최적 절사율을 결정하기 위한 방법을 제안하기로 한다.

Step 1. $d = h$인 칸에 의한 절사 LAD 추정량을 구한다.

Step 2. 앞에서 구한 절사 LAD 추정값을 통하여 구한 Pearson의 X^2 통계량을 X_h^2라하고 자유도 $h - p$에서의 유의확률을 p_h로 나타내기로 한다.

Step 3. $h = h - 1$로 하고 1~2 단계를 $h = h^M$에 이를때 까지 반복한다.

Step 4. $\max\{p_h - p_{h-1},\, h = n, n-1, \ldots, h^M\}$인 p_h^*를 갖는 h에 의한 최적절사율 α를 얻는다.

제안된 방법의 h^M은 로그선형모형의 경우에 $p \le h^M \le d$ 중 한 값이 될 수 있다. 이 값들 중에서 가장 이상적인 값으로 Rousseeuw와 Driessen (2006) 등은 $h^M = [d + p + 1]/2$를 제안하였다. 그러나 분할표 자료의 특성상 이 값이 작아지는 경우에는 절사되고 남은 칸들에 의한 모형의 계획행렬이 비정칙행렬(singular matrix)이 될 위험이 커지게 되고, 이들 계획행렬이 비정칙이 되는 경우를 제외할 경우에는 주어진 절사율에 따른 추정값을 얻을 수 없는 치명적인 문제가 생길 수 있다. 따라서 위 계산방법을 위한 h^M은 계산의 편의를 위해 Rousseeuw와 Driessen (2006)가 제안한 $h^M \approx 0.75d$를 이용하기로 한다.

4. 모의 실험

이 절에서는 제안된 최적절사율 결정을 위한 알고리즘과 붓스트랩 방법에 의한 알고리즘을 비교하기 위하여 모의실험을 수행한 결과를 제시하기로 한다. 모의실험은 5×5 분할표를 대상으로 하였고, 제안된 절사 LAD 추정방법이 갖는 이상칸에 반응하는 특성을 설명하기 위하여 여러 이상칸을 추가한 결과를 제시하고자 한다.

모의실험을 위한 분할표의 확률구조는 행과 열의 주변분포(marginal distribution) 모두 $p_i = i/15$, $i = 1, 2, \ldots, 5$를 이용하였다. 이로부터 독립성모형의 칸확률은 $p_{ij} = p_i p_j$가 된다. 여기에 이상칸에 따른 특징을 설명하기 위하여 다음과 같은 다섯가지 경우에 총 표본수 $n = 500$인 분할표를 각각 500번 생성하여 2절에서 제안된 알고리즘에 의하여 여러 h값에 대한 절사 LAD 추정값을 구하였다. 이때 반복의 수렴은 $\epsilon = 5e - 8$을 이용하였다. 다음과 같은 다섯 가지 상황의 오염된 칸들에는 추가로 100을 더하였으며, 주변분포의 특성상 p_i가 작은 행과 열에 상대적으로 심각한 오염이 이루어지고 p_i가 큰 행과 열에서의 p_{ij}는 상대적으로 클 것이므로 보다 덜 심각한 오염이 이루어지는 것으로 판단할 수 있다.

i) $(1,2)$ 칸이 오염된 경우

ii) $(5,4)$ 칸이 오염된 경우

iii) $(1,2)$, $(5,4)$ 칸이 오염된 경우

표 **4.1.** 여러 절사율에 따른 이상칸 포함 도수

h	이상칸									
	$(1,2)$		$(5,4)$		$(1,2)$, $(5,4)$		$(1,1)$,$(1,2)$, $(5,4)$		$(1,2)$, $(5,4)$, $(5,5)$	
	행렬식	유의확률	행렬식	유의확률	행렬식	유의확률	행렬식	유의확률	행렬식	유의확률
25	0	0	0	0	0	0	0	0	0	0
24	28	141	41	165	0	0	0	0	0	0
23	59	110	79	82	32	170	0	0	0	0
22	80	71	93	63	93	120	21	70	41	169
21	93	65	90	71	112	68	85	131	82	133
20	122	54	95	74	128	76	142	147	166	100
19	114	59	101	45	133	65	193	112	209	98

표 **4.2.** 여러 절사율에 따른 평균절사칸수와 평균포함확률

	이상칸									
	$(1,2)$		$(5,4)$		$(1,2)$, $(5,4)$		$(1,1)$, $(1,2)$, $(5,4)$		$(1,2)$, $(5,4)$, $(5,5)$	
	행렬식	p-값	행렬식	p-값	행렬식	p-값	행렬식	p-값	행렬식	p-값
평균절사칸수	4.14	2.916	3.85	2.884	4.468	3.490	4.964	4.364	5.078	4.254
평균포함확률	0.992	1.000	0.998	1.000	0.996	0.998	0.882	0.920	0.996	1.000

iv) $(1,1)$, $(1,2)$, $(5,4)$ 칸이 오염된 경우

v) $(1,2)$, $(5,4)$, $(5,5)$ 칸이 오염된 경우

표 4.1는 이러한 다섯 가지 상황에서 수행된 모의실험의 결과를 보여주고 있다. 여러 절사율은 위한 $h^M = 0.75d \approx 19$를 이용하였고, 2절에서 제안된 절사 LAD 추정방법을 위해서는 예제에서와 마찬가지로 함수 rq()를 이용하였다.

먼저 표의 각 칸 값은 주어진 h와 이상칸에 대한 절사 LAD를 구하고 이때에 절사된 칸이 모의실험에서 오염시킨 칸을 모두 포함하고 있는 경우의 칸도수를 나타낸다. 예를들어 $(1,2)$칸이 오염되었고 $h = 24$인 경우에 추정 공분산행렬의 행렬식이 $h = 24$에서 최소가 되는 경우에 28번 $(1,2)$ 칸을 절사시켰다는 것을 의미한다. 반면에 제안된 유의확률에 의한 결정방법에 의하면 141번이 최적 절사율 $\alpha = 1/25$로 결정되었고 이에 의한 $h = 24$에서 141번 오염된 $(1,2)$ 칸을 절사시켰다는 것을 의미한다. 그러므로 h가 d에 가까운 값일때 이들 칸의 도수가 높다는 것은 같은 절사율에서 보다 높은 이상칸 식별 가능성을 갖는 다는 것으로 해석할 수 있다. 이러한 현상은 다섯 가지 경우 모두에서 나타나는 공통되는 현상으로 유의확률을 크게 개선시키는 절사율을 결정하는 제안된 방법이 행렬식에 의한 결정 방법에 비해 우수하다는 판단의 경험적 증거가 된다.

표 4.2의 평균 절사칸수는 모든 h에 대한 분할표의 총 칸수 중에서 절사된 칸수들의 평균을 의미하며, 이 값은 절사율과 일대일이 된다. 따라서 이 값이 적다는 것은 절사율이 적다는 것을 의미하여 다섯가지 경우 모두 제안된 유의확률에 의한 절사율 결정이 최적절사율을 모두 고르게 적은 값으로 결정되었음을 의미한다. 평균포함확률은 500번의 모의실험중에서 결정된 절사율에 따른 절사 LAD에 사용된 절사칸들이 실제 오염된 칸을 얼마나 포함하고 있는지를 나타내는 값이다. 앞에서 설명한 바와 같이 $(1,1)$, $(1,2)$ 칸은 작은 칸 확률을 갖기 때문에 여기에 100이 더해진것은 $(5,4)$, $(5,5)$ 칸에 비하여 매우 오염정도가 심하게 된다. 특히 $(1,1)$, $(1,2)$, $(5,4)$ 칸에 오염이 이루어진 경우에 평균포함확률이 낮은 현상이 나타나는 것을 볼 수 있으나 제안된 방법에 의한 최적절사율 결정에서 심각한 오염이 이루어진 경우에도 약 92%로 비교적 높은 평균포함확률을 갖는 것을 알 수 있다. 이러한 현상은 다섯가지 경우에 모두 고

르게 나타나는 현상으로 결국 제안된 방법에 의한 최적절사율 결정이 수행한 모의실험에서는 낮은 절사율과 높은 포함확률을 갖는 것으로 판단할 수 있다.

5. 결론 및 토의

분할표 분석을 위한 절사 LAD 추정량의 추정방법을 제안하였다. 제안된 방법은 LAD 추정방법을 직접 응용하고 있기 때문에 예제와 모의실험에서 사용된 독립성 모형 뿐만아니라 순위변수(ordered variables)가 고려된 여러 연관성 모형(association models)에도 적용될 수 있다. 또한 이차원이 아닌 삼차원 이상의 다차원 분할표에도 직접 적용될 수 있다.

본 논문에서 제안된 방법의 반복계산을 위한 가중 LAD 추정에는 계산의 정확성을 담보하기 위하여 Koenker와 d'Orey (1987)가 제안한 방법을 구현한 R 언어의 quantreg 패키지에 포함된 함수 rq()을 사용하였다. 그러나 이 방법 역시 h^M이 모형에 포함된 모수의 수 p에 가까워 질 수로 계획행렬이 비정칙행렬이 될 가능성이 높아져서 추정이 불가능하다는 단점을 갖고, 또한 경사법에 기반을 둔 방법이기 때문에 Birkes와 Dodge (1993)가 지적한 바와 같이 수렴이 이루어지지 않을 수 있다는 문제 역시 생길 수 있다. 이러한 문제들은 모든 가능한 완벽한 칸 집합을 고려하는 완전탐색(complete enumeration) 방법을 통해 해결할 수 있으나 계산의 양이 매우 커져 적용하기 어렵다는 단점이 있다. 계산의 양에서 발생하는 문제는 선형계획법을 이용한 계산의 양보다는 크지만 완전탐색 보다는 계산의 양이 적은 Choi (2008)가 제안한 최적 부분칸(best subsets)를 이용한 방법을 응용할 수 있다.

참고문헌

이동희, 정병철 (2010). 붓스트랩을 활용한 최적 절사공간중위수 추정량, <응용통계연구>, **23**, 375–382.

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

Barrodale, I. and Roberts, F. D. K. (1973). An improved algorithm for discrete l_1 linear approximation, *SIAM Journal of Numerical Analysis*, **10**, 839–848.

Birkes, D. and Dodge, Y. (1993). *Alternative Methods of Regression*, John Wiley & Sons, Inc.

Choi, H. J. (2008). Estimating LAD regression coefficients with best subset points, *Communications in Statistics, Simulation and Computation*, **37**, 1799–1809.

Godalize, A. (1991). Best approximations to random variables based on trimming procedures, *Journal of Approximation Theory*, **64**, 162–180.

Grizzle, J. E., Stamer, 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, example, and algorithms, *Computational Statistics and Data Analysis*, **25**, 251–272.

Koenker, R.W. and d'Orey (1987). Computing regression quantiles, *Journal of the Royal Statistical Society, Series C(Applied Statistics)*, **36**, 383–393.

Mosteller, F. and Parunak, A. (1985). Identifying extreme cells in a sizable contingency table: Probabilistic and exploratory approaches, In *Exploring Data Tables, Trends and Shape*, edited by Hoaglin, D. C., Mosteller, F. and Tukey, J. W., John Wiley & Sons, New York, 189–224.

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

Schlossmacher, E. J. (1973). An iterative technique for absolute deviations curve fitting, *Journal of the American Statistical Association*, **68**, 857–859.

Shane, K. V. and Simonoff, J. S. (2001). A robust approach to categorical data analysis, *Jounal of Computational and Graphical Statistics*, **10**, 135–157.

Vandev, D. L. (1995). Computing of trimmed L_1 median, In *Multidimensional Analysis in Behavioral Sciences, Philosopic to Technical*, 152–157.

Trimmed LAD Estimators for Multidimensional Contingency Tables

Hyun Jip Choi[1]

[1]Department of Information Statistics, Kyonggi University

(Received October 2010; accepted November 2010)

Abstract

This study proposes a trimmed LAD(least absolute deviation) estimators for multi-dimensional contingency tables and suggests an algorithm to estimate it. In addition, a method to determine the trimming quantity of the estimators is suggested. A Monte Carlo study shows that the propose method yields a better trimming rate and coverage rate than the previously suggest method based on the determinant of the covariance matrix.

Keywords: Contingency tables, log-linear models, weighted LAD estimator, trimmed LAD estimator.

[1]Professor, Department of Applied Information Statistics, Kyonggi University, San94-6 Yiui-dong, Yeongtong-Gu, Suwon, Kyonggi-Do 443-760, Korea. E-mail: hjchoi@kgu.ac.kr