응용통계연구 제 **16**권 *1*호, *2003*년, *pp. 55-69*

# 범주형 자료 분석을 위한 LAD 추정량 *

최현집 [1)]

요 약

일반적인 다차원 분할표 분석을 위해 고려할 수 있는 로그 선형 모형(log-linear model)과 순위 변수(ordered variables)가 고려된 여러 연관성 모형(association models)을 위한 가중값이 부여된 LAD(least absolute deviations) 추정량을 제안하고 추정을 위한 반복 추정법을 제안하였다. 모의실험을 통하여 제안된 LAD 추정량이 최우추정량에 비해 로버스트한 성질을 갖는 다는 것을 밝히고, 이상칸 식별을 위해 많은 선행 연구들에서 인용된 자료들의 경험적 분석을 통해 제안된 추정량과 추정방법이 가질 수 있는 문제점과 특징에 관하여 토론하였다.

주요용어: 다차원 분할표, 최소 절대편차합 추정량, 이상칸, 반복 가중최소제곱방법

## 1. 서 론

자료에 포함된 이상값(outliers)을 식별하거나 혹은 이들에 덜 민감한 추정에 관한 연구는 가장 활발히 연구가 이루어지는 주제 중에 하나이다. 이상값의 식별에 관한 연구는 Barnett과 Lewis(1994) 그리고 로버스트 추정에 관한 연구는 Mayo와 Gray(1997)등에서 여러 제안된 방법들에 관한 참고문헌을 얻을 수 있다.

범주형 자료 분석 혹은 분할표(contingency tables) 분석을 위해 고려되는 모형들은 대체로 최우추정법을 이용하여 모형에 포함된 모수를 추정한다. 따라서 분할표 분석을 위한 모형의 추정에서도 이상값(outliers)에 민감한 최우추정법이 가지고 있는 문제점이 충분히 고려되어야 한다. 분할표 자료에서 이상값은 주어진 모형과 일치된 적합을 나타내지 않은 칸으로 이들 칸을 이상칸(outlier cells)이라고 한다.

이차원 분할표를 위한 독립성 모형(independence models) 하에서 이상칸을 식별하기 위한 연구로는 Fienberg(1969), Brown(1974), Fuchs와 Kenett(1980) 그리고 Korze와 Hawkins (1984)등의 연구가 있다. 그러나 Simonoff(1988)는 제안된 방법들이 이상칸을 이상칸이 아닌 칸으로 판단하는 masking 효과와 이상칸이 아닌 칸을 이상칸으로 판단하는 swamping 효과를 가질 수 있다는 것을 지적하고 삭제된 잔차(deleted residuals)를 이용하여 이들 두 효과에 덜 민감한 이상칸 식별방법을 제안하였다. Simonoff가 제안한 방법은 이차원 분할표가 아닌 일반적인 다차원 분할표와 독립성 모형이 아닌 보다 일반적인 연관성 모형(association models)으로 확장 될 수 있다.

* 이 논문은 2001년도 한국학술진흥재단의 지원에 의하여 연구되었음. (KRF-2001-003-D00032)

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

이러한 잔차에 기반을 둔 이상칸 식별방법 이외에 Yick과 Lee(1998)는 pertubation 진단에 의한 단계적 선택방법을 제안하였다. 그들은 pertubation 진단에 의한 이상칸 식별이 앞에서 언급된 잔차에 기반을 둔 식별방법들에 비해 masking과 swamping 효과에 덜 민감하다는 것을 경험적으로 보였다. 그러나 이들의 연구는 삼차원 이상으로 확장될 수 없다.

이상칸에 보다 덜 민감한 로버스트 추정에 관한 연구로는 관찰칸 값(observed cell counts)에 로그를 취한 이차원 분할표에 median polish를 직접 적용한 Mosteller와 Parunak(1985)의 연구가 있다. 그리고 Rousseuw(1984)가 제안한 LMS(least median of squares)와 LTS(least trimmed squares) 추정 방법을 분할표 분석을 위한 일반적인 모형으로 응용한 Shane과 Simonoff(2001)의 연구가 있다. 그들은 주어진 모형의 로버스트 추정을 위한 기준으로 우도비 검정통계량과 Pearson 통계량 그리고 Gizzle, Starmer와 Koch(1969)가 분할표 모형의 추정을 위해 제안한 가중 최소제곱잔차(weighted least squares residuals) 합을 이용하였다.

Hubert(1997)는 이차원 분할표의 독립성 모형이 갖는 최대 붕괴점(maximum breakdown value)을 유도하였다. 또한 로그 선형 모형(log-linear models)을 로그를 취한 관찰칸 값(observed cell counts)이 종속변수 그리고 분할표의 칸 위치를 나타내는 더미변수(dummy variables)들이 설명변수인 회귀모형으로 나타내고 이러한 모형의 LAD(least absolute deviations) 추정량이 최대 붕괴점 갖는 것을 보였다. 그리고 순위 범주(ordered categories)를 갖는 분할표를 위한 균일 연관 모형(uniform association models)으로도 이러한 사실이 확장될 수 있음을 밝혔다. 그러나 분할표 분석에서 LAD 추정량이 최대 붕괴점을 갖는다는 것만을 보였을 뿐 추정방법에 관한 연구는 수행하지 않았다.

본 연구에서는 범주형 자료 분석을 위해 가중값이 부여된 LAD 추정량의 추정방법에 관하여 연구하였다. 2절에서는 일반적인 다차원 분할표 분석을 위해 고려할 수 있는 로그 선형 모형과 순위 변수가 고려된 여러 연관성 모형의 추정을 위하여 가중값이 부여된 LAD 추정량을 제안하고 Schlossmacher(1973)의 IRWLS(iterative reweighted least squares) 방법을 응용한 반복 추정법을 제안하였다. 3절에서는 모의 실험을 통해 제안된 LAD 추정량이 최우추정량에 비해 로버스트하다는 것을 밝혔으며, 4절에서는 이상칸 식별을 위해 많은 선행 연구들에서 인용한 자료들의 경험적 분석을 통해 제안된 추정량과 반복 추정법이 가질 수 있는 문제점과 특징에 관하여 토론하였다. 마지막으로 5절에서는 본 연구의 결과를 정리하였다.

## 2. 범주형 자료 분석을 위한 LAD 추정량

범주형 자료 분석을 위한 $d$개 칸을 갖는 $D$차원 분할표를 고려하기로 한다. 관찰칸 값으로 구성된 분할표를 $\underline{n} = \{n_i\}$, $i = 1, 2, \ldots, d$, 와 같은 $d \times 1$차 벡터 그리고 $\underline{m} = \{m_i\}$는 $d \times 1$차 기대칸 값(expected cell counts) 벡터로 나타내기로 한다. 각 기대칸 값 $m_i = N\pi_i$로 여기서 $N = \sum_{i=1}^{d} n_i$는 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 로그 선형 모형에 포함된 $p \times 1$차 모수벡터 $\underline{u}$의 추정값 $\hat{\underline{u}}$로부터 얻을

수 있다.

$$\log \underline{m} = \mathbf{X}\underline{u}, \tag{2.1}$$

여기서 $\mathbf{X}$는 모형을 위해 정의된 $d \times p$차 계획 행렬(design matrix)이며, 추정을 위해 행 효과(row effects)와 열 효과(column effects)를 나타내는 모수 $u$들에는 각각 행과 열에 대하여 합하면 '0'이 되는 등의 제약조건이 부여된다.

위 모형에 포함된 $\underline{u}$를 추정하기 위하여 가장 널리 이용되고 있는 최우추정법은 우도비 검정통계량(likelihood ratio goodness-of-fit statistic)

$$G^2 = 2\sum_{i=1}^{d} n_i \log(n_i/\hat{m}_i)^2$$

을 가장 적게 하는 추정 방법으로 생각할 수 있다. 이 이외의 분할표 분석을 위한 모형의 추정 방법은 Agresti(1990)등을 참고할 수 있다.

모형 (2.1)에서 모수의 최우추정값을 얻기 위한 또 다른 방법으로 Grizzle등(1969)이 고려한

$$\sum_{i=1}^{d} \omega_i(\log n_i - \log m_i) \tag{2.2}$$

을 최소로 하는 가중 최소제곱(weighted least squares) 방법을 이용한 반복 추정법을 고려할 수 있다. Grizzle등은 $\log n_i$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.

Shane과 Simonoff(2001)는 식 (2.2)에 의해 Roussew(1984)가 제안한 LMS와 LTS 추정 방법을 적용한 추정량의 추정 방법에 관하여 연구하였다. 그러나 이들 추정량은 그들의 논문에서 지적했듯이 elemental subsets를 이용하기 때문에 이차원 분할표에서도 계산의 양이 많고 계획 행렬이 비정칙 행렬(singular matrix)이 되게 하는 elemental subsets는 고려의 대상에서 제외하여야 하는 등의 문제점들을 갖기 때문에 자료분석을 위해 쉽게 적용하기에는 어려움이 따른다.

이제 같은 맥락에서 식 (2.2)로부터

$$\sum_{i=1}^{d} \omega_i |\log n_i - \log m_i| \tag{2.3}$$

을 최소로하는 LAD 추정량을 고려하기로 한다. 식 (2.3)은 로그를 취한 관찰칸 값과 기대칸 값에 의한 잔차의 절대값에 적절한 가중값 $\omega_i$가 부여된 가중 절대 잔차합(sum of weighted absolute deviations)을 나타내며 일반적으로 널리 알려진 Hubert(1997)가 제안한 절대 잔차합 $\sum_{i=1}^{d} |\log n_i - \log m_i|$과는 다르다.

여기서 식 (2.3)의 추정을 위하여 Schlossmacher(1973)등이 제안한 IRWLS 방법을 확장한 다음과 같은 반복계산법을 제안하기로 한다. 즉, 적절할 가중값 $\omega_i$를 적용하여 반복계산

을 수행하면 $(k+1)$번째 반복을 위한 추정문제는 다음과 같이 표현할 수 있다.

$$\sum_{i=1}^{d} \frac{\omega_i}{|\log n_i - \log \hat{m}_i^{(k)}|} (\log n_i - \log \hat{m}_i^{(k+1)})^2, \tag{2.4}$$

여기서 $\hat{m}_i^{(k)}$는 가중값 $\omega_i/|\log n_i - \log \hat{m}_i^{(k-1)}|$이 부여된 최소제곱법에 의해 $k$번째 반복에서 얻어진 추정칸 값이다. 만일 $|\log \hat{m}_i^{(k+1)} - \log \hat{m}_i^{(k)}| \approx 0$이라면 식 (2.4)는 식 (2.3)에 수렴할 것이며 이때 얻어진 $\underline{\hat{u}} \equiv \underline{\hat{u}}^{(k+1)}$는 가중값 $\omega_i$가 부여된 LAD 추정량이 된다.

제안된 반복계산식 (2.4)를 통해 반복계산을 수행하기 위하여 가중값 $\omega_i$는 Grizzle등(1969)이 제안한 사실을 직접 확장하여 $k$번째 반복에서의 추정칸 값 $\hat{m}_i^{(k)}$를 이용할 수 있다. 단, $k=0$인 반복의 초기값으로는 최우추정법이 가지고 있는 이상칸에 민감한 문제를 피하기 위하여 식 (2.4)에서 $\omega_i \equiv 1$인 가중값이 부여되지 않은 LAD 추정방법에 의해 추정한 추정모수 $\hat{u}^{(0)}$에 의한 $\hat{m}_i^{(0)} = \mathbf{X}\underline{\hat{u}}^{(0)}$를 사용하기로 한다. 또한 Fisher(1961)가 지적한바와 같이 유일한 LAD 추정값은 추정의 대상이 되는 모수의 수 만큼 완전한 적합(perfect fit)이 이루어져야하기 때문에 식 (2.1)에서 고려된 $\underline{u}$의 차수인 $p$개 칸에서 완전한 적합이 이루어져야 한다. 그러나 만일 $|\log n_i - \log \hat{m}_i^{(k)}| \approx 0$이라면 $\omega_i/|\log n_i - \log \hat{m}_i^{(k)}|$는 매우 커지는 문제가 발생한다. 일반적인 최소 절대 잔차합 추정시에 발생할 수 있는 Lange(1999)가 지적한 이러한 문제의 해결은 Merle과 Späth(1974)가 적용한 적절한 $\epsilon > 0$에 대하여 $\max\{|\log n_i - \log \hat{m}_i^{(k)}|, \epsilon\}$값으로 대체하는 경험적인 방법을 적용하기로 한다. 따라서 제안된 식 (2.4)에 의한 추정은 보다 좋은 적합이 이루어지는 칸에 그렇지 않은 칸에 비해 더욱 큰 가중값을 부여하는 성질을 갖는다는 것을 쉽게 알 수 있다.

식 (2.3)을 최소로 하는 추정량의 극한분포(limiting distribution)는 Hubert(1997)와 같이 모형 (2.1)을 칸의 위치를 나타내는 더미변수에 의한 회귀모형으로 간주할 경우에 Knight (1999)에 의해 얻을 수 있다. 특히 Knight는 LAD 추정량의 점근적 성질(asymptotic properties)들에 관한 여러 연구 결과들을 정리하고 식 (2.3)과 같은 형태에 의한 추정량이 가중값을 취하지 않은 LAD 추정량에 비해 더욱 효율적이라는 사실과 이를 이용한 연구 결과를 소개하였다.

이러한 사실로부터 식 (2.4)의 수렴해인 식 (2.3)을 최소로 하는 LAD 추정량을 얻기 위하여 다음과 같은 알고리즘을 제안할 수 있다.

(1 단계) 먼저 식 (2.4)에서 $\omega_i \equiv 1$인 반복계산을 통하여 초기값 $\hat{m}_i^{(0)} = \mathbf{X}\underline{\hat{u}}^{(0)}$을 얻는다.

(2 단계) 앞에서 얻어진 초기값에 의해 $(k+1)$번째 반복을 위한 $\epsilon > 0$인 적절한 값에 대하여 다음과 같은 가중값을 얻는다.

$$\gamma_i^{(k)} = \frac{\hat{m}_i^{(k)}}{\max\{|\log n_i - \log \hat{m}_i^{(k)}|, \epsilon\}}, \tag{2.5}$$

여기서 $\underline{\hat{m}}^{(k)} = \exp(\mathbf{X}\underline{\hat{u}}^{(k)})$이다.

(3 단계) 앞에서 얻은 $\gamma_i^{(k)}$를 대각원소로 갖는 대각행렬을 $\mathbf{\Gamma}^{(k)}$라 하고 $(k+1)$번째

반복에서의 추정값

$$\hat{\underline{u}}^{(k+1)} = (\mathbf{X}'\mathbf{\Gamma}^{(k)}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Gamma}^{(k)}\log(\underline{n})$$

를 얻는다.

(4 단계) $k = k+1$로 한다.

(5 단계) 2~4 단계를

$$\sum_{i=1}^{d} |\log \hat{m}_i^{(k+1)} - \log \hat{m}_i^{(k)}| < \epsilon \sum_{i=1}^{d} |\log \hat{m}_i^{(k+1)}|$$

이 만족될 때까지 반복한다.

위 반복 추정 알고리즘에서 $k$번째 반복에서의 가중값 $\gamma_i^{(k)}$의 분자는 Grizzle등(1969)의 추정방법에서와 유사하게 반복에서 얻어지는 잠정 추정값을 가중값으로 이용하였음을 상기하자. 그러므로 만일 식 (2.5)에서 $\gamma_i^{(k)} \equiv \hat{m}_i^{(k)}$이라면 제안된 방법에 의해 수렴된 추정량은 가중 최소제곱추정량이 된다. 이러한 사실은 제안된 알고리즘이 IMSL , S-Plus 혹은 SAS와 같은 소프트웨어들에서 제공하는 함수를 이용하여 쉽게 프로그램할 수 있는 것을 의미한다. 마지막으로 제안된 추정방법은 분할표를 위한 일반적인 로그 선형 모형인 식 (2.1)의 추정을 위한 추정 방법이므로 다차원 분할표를 위한 로그 선형 모형 그리고 균일 연관 모형과 같은 순위 변수에 의한 분할표 분석을 위한 모형에서도 쉽게 적용될 수 있다.

## 3. 모의실험

이 절에서는 독립성 모형과 균일 연관 모형에 대하여 제안된 추정량의 특성을 설명하기 위하여 모의실험을 수행한 결과를 제시하기로 한다. 모의실험은 $5 \times 5$ 분할표를 대상으로 하였고, 제안된 추정량이 이상값에 반응하는 특성을 설명하기 위하여 여러 이상칸을 추가한 결과 역시 제시하였다.

### 3.1. 독립성 모형

이차원 분할표에 대한 모형 (2.1)의 특수한 경우인 독립성 모형은 다음과 같이 표현할 수 있다.

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

여기서 $i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, c$이며 $r$과 $c$는 각각 행과 열 수를 나타낸다. 이때 모형의 식별을 위해 행 효과들은 $\sum_{i=1}^{r} u_{1(i)} = 0$ 그리고 열 효과들에는 $\sum_{j=1}^{c} u_{2(j)} = 0$과 같은 제약조건을 부여하기로 한다. Hubert(1997)는 이러한 독립성 모형에 대한 LAD 추정량의 최

대 붕괴점이 다음과 같다는 것을 증명하였다.

$$\frac{1}{rc}\left[\frac{\min(r,c)+1}{2}\right] \tag{3.1}$$

이러한 상황하에서 모의실험은 칸 확률이 모두 동일한 균일 확률 구조(uniform probability structure)를 가진 $5 \times 5$ 분할표에 대하여 이루어 졌으며, 제안된 추정량이 이상칸에 반응하는 특성을 설명하기 위하여 다음과 같은 네 가지 경우에 총수 $n = 500$인 분할표를 각각 100번 생성하여 2절에서 제안된 알고리즘에 의해 모형의 추정이 이루어졌다. 이때 식 (2.5)를 위한 값은 Merle과 Späth(1974)와 같이 $\epsilon = 5E - 7$을 이용하였고 이후에 제시되는 분석에서는 모두 이 값을 이용하였다.

i) 균일 확률 구조

ii) (1,1) 칸이 오염된 경우

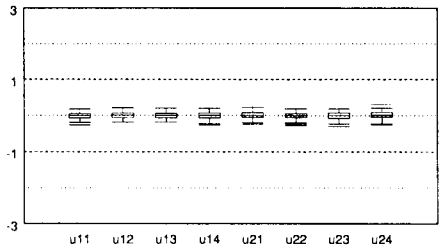
iii) (1,1), (1,2) 칸이 오염된 경우

iv) (1,1), (1,2) (1,3) 칸이 오염된 경우

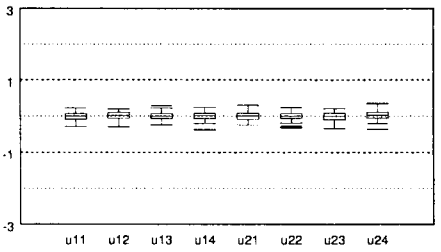
오염된 각 칸에는 칸 도수에 추가로 500을 더하였으며, $5 \times 5$ 분할표이므로 식 (3.1)에 의해 최대 붕괴점은 '3'이다. 따라서 2개 칸 보다 많은 칸이 이상칸인 경우에 LAD 추정량은 붕괴될 수 있다. 참고로 최우추정량의 붕괴점은 '1'이며, Shane과 Simonoff(2001)가 제안한 추정량은 LAD 추정량과 같은 붕괴점을 갖는다.

<그림 1>은 100회 모의실험에 의한 독립성 모형의 추정값들의 분포를 보여주고 있다. 모형에 포함된 행 효과와 열 효과에는 각각 합해서 영이 되는 제약조건이 부여되어 있으므로 각각 네 개의 추정된 행과 열 효과를 나타내었다. u1i는 행 효과 $u_{1(i)}$를 그리고 u2j는 열 효과 $u_{2(j)}$를 나타낸다. <그림 1>의 첫 행은 i)인 경우를 그리고 두 번째에서 네 번째 행은 ii)~iv)인 경우의 최우추정값과 LAD 추정값의 분포를 각각 나타낸다.

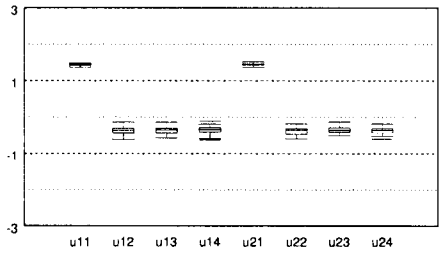
먼저, 첫 번째 행은 칸 확률이 모두 균일한 경우이므로 두 추정값의 분포가 모두 매우 안정적인 분포를 나타내고 있으며, 미세하나마 최우추정값에 비해 LAD 추정값의 산포가 보다 넓게 분포된 것을 식별할 수 있다. 그러나 두 번째 행에서 즉, (1,1)칸이 오염된 경우에 최우추정값은 u11과 u21이 붕괴된 것을 알 수 있고, 이러한 영향이 다른 추정값에도 영향을 주는 것을 알 수 있다. 이에 반해 LAD 추정량은 매우 안정적인 추정이 이루어지는 것을 알 수 있다. 역시 <그림 1>의 세 번째 행의 (1,1)칸과 (1,2)칸이 오염된 경우에 최우추정값은 각 칸에 대응되는 모수들 u11, u21, u22가 상당히 큰 값으로 추정되고 특히 u11이 심각하게 붕괴되는 현상을 알 수 있다. 그러나 이 경우에도 LAD 추정값은 안정적인 추정이 되는 것을 알 수 있다. 단, u11에서는 약간 '0'에서 벗어난 모습을 보여주지만 심각한 정도는 아닌 것을 알 수 있다. 마지막으로 네 번째 경우인 (1,1), (1,2) 그리고 (1,3)칸이 오염된 경우에는 최우추정값과 LAD 추정값 모두 붕괴점을 넘어섰으므로 추정 결과에 문제가 발생할 수 있다는 것을 쉽게 알 수 있다.



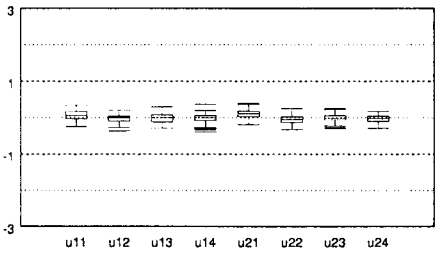

MLE



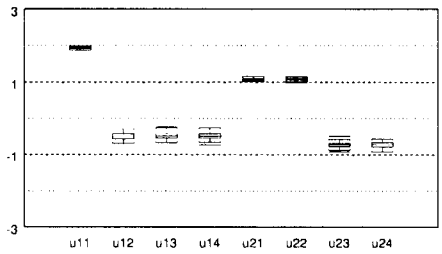

LAD



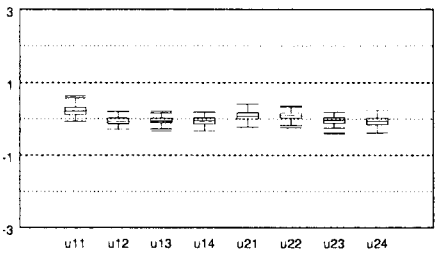

MLE : (1,1)



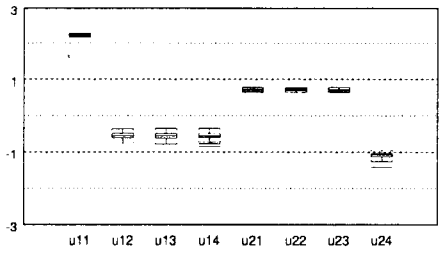

LAD : (1,1)



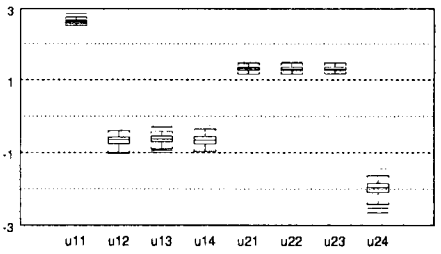

MLE : (1,1), (1,2)




LAD : (1,1), (1,2)




MLE : (1,1), (1,2), (1,3)




LAD : (1,1), (1,2), (1,3)

<그림 1> 모의실험에 의한 독립성 모형의 최우추정값과 LAD 추정값의 분포

### 3.2. 균일 연관 모형

행과 열 점수가 고려된 이차원 분할표에 대한 균일 연관 모형은 다음과 같다.

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \beta\lambda\mu,$$

여기서 $\beta$는 행과 열 점수에 대한 연관 모수(association parameter)를 나타내며 $\lambda$는 행 점수, $\mu$는 열 점수를 나타낸다. 모의 실험에서는 각 행과 열 점수에 (-2, -1, 0 1, 2)와 같은 등간격(unit spaced)인 점수를 부여하고 독립성 모형과 같이 분할표의 총수 $n = 500$에 대하여 100회 반복 추정을 실시하였다. 이때 독립성 모형에서와 유사하게 이상칸에 따른 추정값의 변화를 분석하기 위하여 다음과 같은 네 가지 상황을 고려하였다. 오염된 칸은 역시 해당 칸에 500을 더하였다.

i) 균일 확률 구조

ii) (1,3) 칸이 오염된 경우

iii) (1,3), (2,3) 칸이 오염된 경우

iv) (1,3), (2,3) (3,3) 칸이 오염된 경우

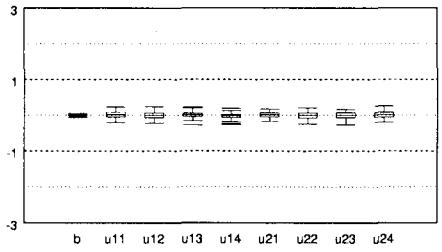
<그림 2>에서 b는 연관모수 $\beta$를 의미하며 각 행은 각각 i)~iv)인 경우의 추정값의 분포를 나타낸다. 균일 연관 모형 역시 독립성 모형과 같이 $r \geq 3, c \geq 4$인 경우에 최대 붕괴점이 식 (3.1)과 같다는 사실이 Hubert(1997)에 의해 알려져 있으므로 균일 연관 모형에서도 독립성 모형과 유사한 결과를 얻을 수 있다는 것을 예상할 수 있다.

먼저 <그림 2>의 첫 행에서는 최우추정값과 LAD 추정값 모두 매우 안정적인 분포를 보여주고 있으며, 최우추정값에 비해 LAD의 산포가 약간 크다는 것을 알 수 있다. 그러나 ii)인 경우를 나타내는 두 번째 행에서 (1,3)칸에 해당되는 u11과 u23의 최우추정값이 붕괴된다는 것을 쉽게 알 수 있다 이에 반해 LAD 추정값의 분포는 매우 안정적인 추정이 이루어짐을 나타내고 있다. 역시 iii)인 경우를 나타내는 세 번째 행에서 (1,2), (2,3)칸에 해당되는 최우추정값은 붕괴되며 특히 u23가 심각하게 큰값으로 추정되는 것을 알 수 있다. 그러나 LAD 추정량은 u23가 약간 큰 값으로 추정되는 모습을 보이나 전반적으로 오염된 칸 즉, 이상칸에 상당히 로버스트한 추정이 된다는 사실을 보여주고 있다. 마지막으로 iv)의 경우인 네 번째 행은 붕괴점 '2' 보다 큰 세 칸에 이상칸이 나타나므로 최우추정값과 LAD 추정량 모두 추정 결과에 상당한 문제가 있을 수 있다는 것을 알 수 있다.

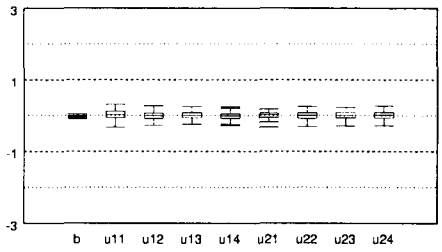
## 4. 실제 자료분석 예

### 4.1. 고고학 유물 자료

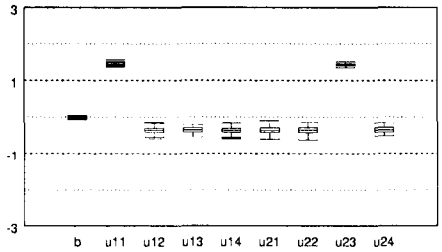
Shane과 Simonoff(2001)에서 발췌한 고고학 유물(archaeological artifact) 자료는 Mosteller와 Parunak(1985)이 관찰칸 값에 로그를 취한 분할표에 median polish 방법을 적용하여 이상칸을 식별하기 위해 인용한 자료중 일부분만으로 구성된 자료이다. 분할표는 송곳(drills),



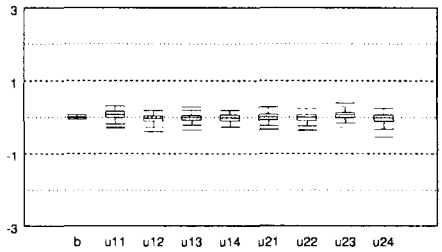

MLE



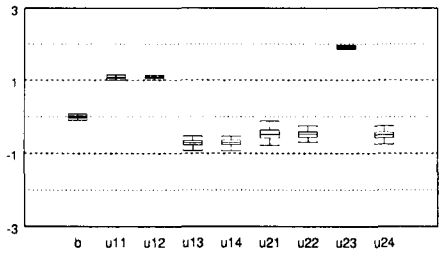

LAD



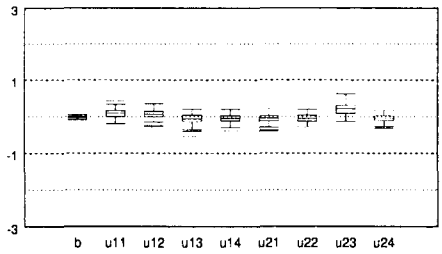

MLE : (1,3)



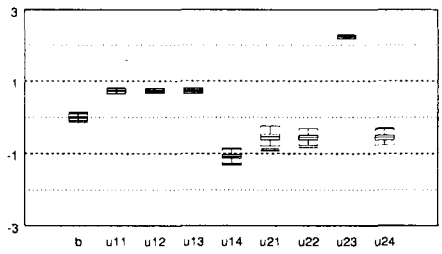

LAD : (1,3)



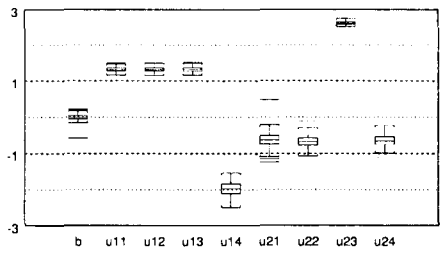

MLE : (1,3), (2,3)




LAD : (1,3), (2,3)




MLE : (1,3), (2,3), (3,3)




LAD : (1,3), (2,3), (3,3)

<그림 2> 모의실험에 의한 균일 연관 모형의 최우추정값과 LAD 추정값의 분포

항아리(pots), 연마석(grinding stones), 칼끝 조각(point fragments)의 네가지 유물이 상존

하는 물(permanent water)에서 떨어진 거리에 의해 교차분류되었다.

Mosteller와 Parunak(1985)등은 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. <표 1>은 고고학 유물 자료의 관찰칸 값과 독립성 모형에 대한 최우추정값 그리고 LAD 추정값을 보여주고 있다. 각 추정값 하단의 괄호안의 값은 표준화된 잔차, $(n_i - \hat{m}_i)/\sqrt{\hat{m}_i}$를 나타낸다.

<표 1> $4 \times 4$ 고고학 유물자료의 관찰칸값과 MLE, LAD 추정값

| | | 바로 인근 | 1/4마일이내 | 1/4~1/2마일 | 1/2~1마일 |
|---|---|---|---|---|---|
| 관찰칸 값 | 송곳 | 2 | 10 | 4 | 2 |
| | 항아리 | 3 | 8 | 4 | 6 |
| | 연마석 | 13 | 5 | 3 | 9 |
| | 칼끝 조각 | 20 | 36 | 19 | 20 |
| MLE | 송곳 | 4.1707 | 6.4756 | 3.2927 | 4.0610 |
| | | (-1.0629) | (1.3850) | (0.3898) | (-1.0227) |
| | 항아리 | 4.8659 | 7.5549 | 3.8415 | 4.7378 |
| | | (-0.8459) | (0.1619) | (0.0809) | (0.5799) |
| | 연마석 | 6.9512 | 10.7927 | 5.4878 | 6.7683 |
| | | (2.2942) | (-1.7633) | (-1.0620) | (0.8578) |
| | 칼끝 조각 | 22.0122 | 34.1768 | 17.3780 | 21.4329 |
| | | (-0.4289) | (0.3119) | (0.3891) | (-0.3095) |
| LAD | 송곳 | 4.2111 | 7.5789 | 4.0000 | 4.2105 |
| | | (-1.0773) | (0.8794) | (0.0000) | (-1.0773) |
| | 항아리 | 4.4444 | 8.0000 | 4.2222 | 4.4444 |
| | | (-0.6851) | (0.0000) | (-0.1081) | (0.7379) |
| | 연마석 | 3.1579 | 5.6842 | 3.0000 | 3.1579 |
| | | (5.5385) | (-0.2870) | (0.0000) | (3.2875) |
| | 칼끝 조각 | 20.0000 | 36.0000 | 19.0000 | 20.0000 |
| | | (0.0000) | (0.0000) | (0.0000) | (0.0000) |

우선 (3,1)에 해당하는 최우추정값에 의한 표준화된 잔차는 2.2942로 비교적 큰 값으로 나타나며 <표 1> 하단의 LAD 추정값에 의한 표준화된 잔차는 5.5385로 매우 큰 값인 것을 알 수 있다. 이러한 사실은 (3,1)칸이 이상칸일 가능성이 있다는 것을 시사하는 것으로 제안된 LAD 추정량이 이상칸에 영향을 받지 않았기 때문에 큰 잔차를 얻은 것으로 해석할 수 있다. 다음으로 LAD 추정에 의해 큰 표준화된 잔차를 갖는 (3,4)칸에 대하여 최우추정에서는 잔차값이 크지 않은 것을 알 수 있다. Shane과 Simonoff(2001)는 이 칸이 (3,1)칸과 함께 이상칸일 가능성이 있다는 것을 자신들의 추정량의 추정 결과에서 지적하였고 제안된 추정량 역시 같은 결과를 얻는다는 사실을 알 수 있다. 마지막으로 이 자료는 $4 \times 4$ 분할표이므로 모형에는 7개 모수가 포함된다. 따라서 유일한 LAD 추정값은 7개 칸에서 완전한

적합이 일어나야 한다. <표 1>에서 제안된 추정값이 7개 칸에서 완전한 적합이 이루어진 것을 확인 할 수 있으므로 제안된 추정방법을 통해 유일한 LAD 추정값이 얻어졌다는 것을 알 수 있다.

<표 2> CMOS 자료의 관찰칸 값과 MLE, LAD 추정값

| | | I | II | III | IV | V |
|---|---|---|---|---|---|---|
| 관찰칸 값 | 낮음 | 47 | 5 | 6 | 2 | 0 |
| | 중간 | 17 | 7 | 10 | 16 | 5 |
| | 높음 | 12 | 4 | 7 | 15 | 9 |
| MLE | 낮음 | 42.7874 | 6.4043 | 5.6581 | 4.2989 | 0.8513 |
| | | (0.6440) | (-0.5549) | (0.1438) | (-1.1088) | (-0.9227) |
| | 중간 | 24.8187 | 6.1510 | 8.9982 | 11.3202 | 3.7119 |
| | | (-1.5694) | (0.3423) | (0.3340) | (1.3909) | (0.6686) |
| | 높음 | 8.3939 | 3.4446 | 8.3476 | 17.3809 | 9.4368 |
| | | (1.2447) | (0.2992) | (-0.4652) | (-0.5711) | (-0.1422) |
| LAD | 낮음 | 47.0000 | 7.3470 | 6.0000 | 5.9668 | 1.6661 |
| | | (0.0000) | (-0.8659) | (0.0000) | (-1.6240) | (-1.2908) |
| | 중간 | 30.5475 | 6.9998 | 8.3797 | 12.2157 | 5.0000 |
| | | (-2.4512) | (0.0001) | (0.5597) | (1.0828) | (0.0000) |
| | 높음 | 11.9084 | 4.0000 | 7.0194 | 15.0000 | 9.0000 |
| | | (0.0266) | (0.0000) | (-0.0073) | (0.0000) | (0.0000) |

### 4.2. 보조 금속 산화 반도체 회로의 접촉창 형성 자료

Shane과 Simonoff(2001)에서 인용한 보조 금속 산화 반도체(CMOS: complementary metal oxide semiconductor) 회로의 접촉창 형성 자료는 세가지 회전속도(낮음, 중간, 높음)와 마이크로 밀리미터 단위로 측정된 다섯 가지 접촉창의 크기( I :생성되지 않거나 혹은 출력됨, II :(0, 2.25), III :[2.25, 2.75), IV:[2.75, 3.25), V :[3.25,$\infty$))에 의해 교차분류된 분할표이다. 이 자료에 관한 자세한 설명은 Simonoff(1988)를 참고할 수 있다. <표 2>는 이 자료의 관찰칸 값과 최우추정과 제안된 LAD 추정에 의한 추정값을 나타내고 있다. 각 추정값 하단의 괄호안의 값은 표준화된 잔차값이다.

이 자료에는 단위 구간 행 점수 (-1, 0, 1)과 열 점수 (-2, -1, 0, 1, 2)가 부여된 균일 연관 모형이 가장 적절한 모형인 것으로 분석되었다. 특히, Shane과 Simonoff(2001)는 (2,1)칸이 적합된 균일 연관 모형과 불일치를 나타내는 이상칸일 가능성이 있다는 것을 지적하였다. <표 2>에서 (2,1)칸의 최우추정에 의한 표준화된 잔차는 그리 큰 값이 아니지만 제안된 LAD 추정에 의한 표준화된 잔차는 -2.4512로 비교적 큰 값으로 나타나기 때문에 그들의 결과와 일치된 추정결과를 얻는다는 사실을 알 수 있다. 참고적으로 최우추정에 의한 연관모수 $\hat{\beta} = 0.5043$이 얻어지며, LAD 추정방법에 의해서는 $\hat{\beta} = 0.3825$를 얻었다.

### 4.3. 임산부 자료

<표 3>의 자료는 Aberdeen에 거주하는 임산부들의 친구와 산부인과 방문횟수에 관한 자료로 다음과 같은 네 변수 A, B, C, D에 의한 사차원 분할표를 구성한다.

<표 3> 임산부 자료의 관찰값 값과 MLE, LAD 추정값

| | | 가까움 | | | | 멂 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 잦지 않음 | | 잦음 | | 잦지 않음 | | 잦음 | |
| | | 첫아이 아님 | 첫아이 | 첫아이 아님 | 첫아이 | 첫아이 아님 | 첫아이 | 첫아이 아님 | 첫아이 |
| 관찰값 | 매일 | 16 | 1 | 14 | 5 | 0 | 0 | 2 | 13 |
| | 매주 | 13 | 6 | 6 | 6 | 10 | 2 | 6 | 6 |
| | 가끔 | 3 | 2 | 2 | 0 | 5 | 0 | 5 | 4 |
| MLE | 매일 | 15.0571 | 2.3784 | 11.2129 | 7.3514 | 4.5421 | 1.7296 | 3.3824 | 5.3459 |
| | | (0.2429) | (-0.8938) | (0.8323) | (-0.8672) | (-2.1312) | (-1.3151) | (-0.7517) | (3.3104) |
| | 매주 | 12.9661 | 2.0480 | 9.6556 | 6.3303 | 7.2674 | 2.7673 | 5.4119 | 8.5535 |
| | | (0.0094) | (2.7615) | (-1.1764) | (-0.1313) | (1.0137) | (-0.4612) | (0.2528) | (-0.8731) |
| | 가끔 | 2.9278 | 0.4625 | 2.1803 | 1.4294 | 4.2393 | 1.6143 | 3.1569 | 4.9895 |
| | | (0.0122) | (2.2606) | (-0.1221) | (-1.1956) | (0.3695) | (-1.2705) | (1.0373) | (-0.4130) |
| LAD | 매일 | 16.0000 | 1.6572 | 10.6667 | 5.0000 | 3.0000 | 0.5969 | 2.0000 | 1.8011 |
| | | (0.0000) | (-0.5105) | (1.0206) | (0.0000) | (-1.73205) | (-0.7726) | (0.0000) | (8.3447) |
| | 매주 | 13.0000 | 1.3465 | 8.6667 | 4.0625 | 9.9941 | 1.9886 | 6.6627 | 6.0000 |
| | | (0.0000) | (4.0104) | (-0.9058) | (0.9613) | (0.0019) | (0.0081) | (-0.2567) | (0.0000) |
| | 가끔 | 3.0000 | 0.3107 | 2.0000 | 0.9375 | 6.6627 | 1.3257 | 4.4418 | 4.0000 |
| | | (0.0000) | (3.0305) | (0.0000) | (-0.9682) | (-0.6442) | (-1.1514) | (0.2649) | (0.0000) |

(변수 A) 출산경력 : 첫아이 아님(not first child), 첫아이(first child)
(변수 B) 병원이용 : 잦지 않음(underutilizer), 잦음(utilizer)
(변수 C) 친구들로 부터의 거리 : 가까움(walking distance), 멂(takes bus)
(변수 D) 방문 : 매일(daily), 매주(weekley), 가끔(less often)

Upton과 Guillen(1995)에서 인용한 이 자료는 [AB][AC][CD] 모형이 가장 적절한 모형인 것으로 분석되었고, 이 모형하에서 (첫아이, 잦음, 멂, 매일)을 나타내는 (2, 2, 2, 1)칸이 이상칸이라는 것이 알려져 있다. <표 3>으로부터 해당칸에 대한 최우추정과 LAD 추정에 의한 표준화된 잔차가 모두 큰 값으로 나타나는 것을 알 수 있다. 그러나 최우추정에 비해 LAD 추정에 의한 결과가 더욱 큰 값인 점에 주목하자.

이 칸 이외에 LAD 추정에 의하면 (2, 1, 1, 2)칸과 (2, 1, 1, 3)칸이 적합된 모형과 큰 불일치를 나타내는 칸인 것을 알 수 있다. 특히 (1, 1, 2, 1)칸은 Upton과 Guillen에 의하면 이상칸인 (2, 2, 2, 1)칸에 의해 swamping 효과가 발생한 칸으로 해석할 수 있는데 <표 3>으로 LAD 추정에서는 최우추정에 비해 작은 표준화된 잔차를 갖기 때문에 이러한 사실을 뒷받침하는 것으로 해석할 수 있다.

## 5. 결론

본 연구에서는 일반적인 다차원 분할표 분석을 위한 가중값이 부여된 LAD 추정량을 제안하고 추정을 위한 반복추정법을 제안하였다. 비록 이차원 분할표에서 독립성 모형과 균일 연관 모형 그리고 사차원 분할표에 대한 조건부 독립성 모형을 적합한 분석 예만을 제시하였지만 제안된 추정량은 모형 (2.1)로 표현할 수 있는 분할표를 위한 어떠한 모형으로도 쉽게 확장되어 적용될 수 있다. 또한 추정을 위해 제안된 알고리즘은 IRWLS 방법을 직접 확장한 알고리즘이므로 실제 자료분석을 위해 쉽게 적용될 수 있다.

여러 모형에 대한 모의실험을 통해 제안된 추정량은 이상칸에 덜 민감한 로버스트 추정량이라는 것을 확인 했으며, 이상칸이 알려진 몇 가지 실제 자료에의 적용 예를 통하여 이러한 사실을 다시 한번 확인할 수 있었다.

## 감사의 글

본 논문을 심사해주신 심사위원들과 편집위원들께 감사드립니다.

## 참고문헌

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

[2] Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data* 3rd, John Wiley: New York.

[3] Brown, M. B. (1974). Identification of the sources of significance in two-way contingency tables, *Applied Statistics*, Vol. 23, 405-413.

[4] Fienberg, S. E. (1969). Preliminary graphical analysis and quasi-independence for two-way contingency tables, *Applied Statistics*, Vol. 18, 153-168.

[5] Fisher, W. D. (1961). A note on curve fitting with minimum deviations by linear programming, *Journal of the American Statistical Association*, Vol. 56, 359-362.

[6] Fuchs, C. and Kennett, R. (1980). A test for detecting outlying cells in the multinomial distribution and two-way contingency tables, *Journal of the American Statistical Association*, Vol. 75, 395-398.

[7] Grizzel, J. E., Stamer, F., and Koch, G. G. (1969). Analysis of categorical data by linear models, *Biometrics*, Vol. 25. 489-504.

[8] Hubert, M. (1997). The breakdown value of the L1 estimator in contingency tables, *Statistics and Probability Letters*, Vol. 33, 419-425.

[9] Knight, K. (1999). Asymptotics for L1-estimators of regression parameters under heteroscedasticity, *The Canadian Journal of Statistics*, Vol. 27, 497-507.

[10] Kotze, T. J. W. and Hawkins, D. M. (1984). The identification of outliers in two-way contingency tables using subtables, *Applied Statistics*, Vol. 33, 215-223.

[11] Lange, K. (1999). *Numerical Analysis for Statisticians*, Springer verlag: New York.

[12] Mayo, M. A. and Gray, J. B. (1997). Elemental subsets: the building blocks of regression, *American Statistician*, Vol. 52, 122-129.

[13] Merle, G. and Späth, H. (1974). Computational experiences with discrete Lp- Approximation, *Computing*, Vol. 12, 315-321.

[14] 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.

[15] Rousseeuw, P. J. (1984). Least median of squares regression, *Journal of the American Statistical Association*, Vol. 79, 871-880.

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

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

[18] Simonoff, J. S. (1988). Detecting outlying cells in two-way contingency tables via backwards stepping, *Technometrics*, Vol. 30, 339-345.

[19] Upton, G, J. G. and Guillen, M. (1995). Perfect cells, direct models and contingency table outliers, *Communications in Statistics-Theory and Methods*, Vol. 24, 1843-1862.

[20] Yick, J. S. and Lee, A. H. (1998). Unmasking outliers in two-way contingency tables, *Comutatonal Statistics and Data Analysis*, Vol. 29, 69-79.

[ 2002년 5월 접수, 2002년 9월 채택 ]

# LAD Estimators for Categorical Data Analysis *

Hyun Jip Choi [1)]

ABSTRACT

In this article, we propose the weighted LAD(least absolute deviations) estimators for multi-dimensional contingency tables and drive an estimation method to estimate the proposed estimators. To illustrate the robustness of the estimators, simulation results are presented for several models including log-linear models and models for ordinal variables in multidimensional contingency tables. Examples were also introduced.

*Keywords:* Multi-dimensional Contingency Tables, LAD Estimators, Outlying Cells, IRWLS methods

* This work was supported by Korea Research Foundation Grant(KRF-2001-003-D00032)

1) Associate Professor, Division of Economics, Kyonggi University, Suwon 442-760, Korea
E-mail: hjchoi@stat.kyonggi.ac.kr