Collapsibility Criteria using Raindrop Plots

홍종선[1] 김범준[2]

요약

범주형 자료분석에서 차원축소(collapsibility)는 오즈비로 설명되었다. 실제의 $2\times2\times K$ 분할표 자료를 이 이론에 적용시켰을 때 오즈비의 값으로 차원축소가 가능한지의 여부를 판단하기는 어렵다. 오즈비를 시각적으로 표현하는 방법 중에서 Doi, Nakamura와 Yamamoto(2001)가 제안한 Contour plot을 통해서 분할표 자료를 설명하는 것은 가능하지만 차원축소의 가능성을 결정하기에는 한계가 있다. 본 연구에서는 오즈비의 신뢰구간을 시각적으로 표현할 수 있는 방법으로 Barrowman과 Myers(2003)가 제안한 Raindrop plot을 이용하여 $2\times2\times K$ 분할표 자료를 설명할 수 있으며 동시에 차원축소의 가능성을 판단할 수 있는 방법을 제안하고자 한다.

주요 용어 : 오즈비, 교차적비, 차원축소, 로그선형모형.

1. Raindrop plot

삼차원 분할표에서 기본이 되는 $2\times2\times K$ 분할표에 대한 기본적인 개념과 수식을 정의하자. 칸의 빈도 x_{ijk}와 이에 대응하는 칸 확률 p_{ijk}를 가지는 $2\times2\times K$ 분할표와 세 번째 변수에 대하여 차원축소된 2×2 분할표를 고려하자.
k번째 2×2 분할표의 교차적비와 차원축소된 분할표의 교차적비는 다음과 같이 표현된다.

$$\theta_k=\frac{p_{11k}p_{22k}}{p_{12k}p_{21k}},\ \theta_c=\frac{p_{11+}p_{22+}}{p_{12+}p_{21+}} \tag{1.2}$$

2×2 분할표의 주변합을 고정시켰을 때 로그오즈비 $\theta_k^l=\log(\theta_k)$의 조건부 가능도함수는 다음과 같이 얻을 수 있다(Agresti 1990).

$$L(\theta_k^l)=\binom{x_{12k}+x_{21k}}{x_{11k}}\binom{x_{21k}+x_{22k}}{x_{21k}}e^{\theta_k^l}/S(\theta_k^l)$$

여기서 $S(\theta_k^l)$는 분할표의 주변합을 고정시켰을 때 가능한 x_{11k} 값에 대한 가능도함수의 분자의 총합이다.

로그오즈비 θ_k^l의 최대가능도추정량(MLE)을 θ_k^{MLE}라 표시하고, $l(\theta_k^l)$를 로그조건부 가능도함수로 정의하면, 가능도비의 근사분포에 기초하여 θ_k^l에 대한 근사적인 $100\times(1-\gamma)\%$ 신뢰구간은 다음과 같이 정의된다(Cox와 Hinkley(1974) 참조).

$$\left\{\theta_k^l:\ 2[l(\theta_k^{MLE})-l(\theta_k^l)\le\chi^2_{1(1-\gamma)}]\right\},$$

여기서 $\chi^2_{1(p)}$는 자유도 1을 갖는 카이제곱분포의 $100\times p$번째 백분위수이다. 편의상 $l(\theta_k^{MLE})=0$

1) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 경제학부 통계학전공, 교수
2) (110-745) 서울특별시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 박사과정

이라 설정하면, $\chi^2_{1(1-\gamma)}/2$이기 때문에 $100\times(1-\gamma)\%$ 신뢰구간은 다음과 같다.

$$\left\{\theta_k^l : l(\theta_k^l) \geq -\chi^2_{1(1-\gamma)}/2\right\},$$

$\gamma=0.05$일때 $\chi^2_{1(0.95)}/2=1.92$이므로, Raindrop plot은 로그가능도 함수가 -1.92보다 큰 부분을 반사시켜 95%와 99% 신뢰구간을 동시에 나타낼 수 있다.

2. 차원축소와 Raindrop plot

삼차원 범주형 자료에 대한 로그선형모형들을 Christensen(1990)이 사용한 표기 방법을 따라 분류하면 다음과 같다: 포화모형(saturated model)은 [123], 부분연관모형(partial association model)은 [12][13][23], 조건부독립모형(conditionally independent model)은 [12][13], [12][23], [13][23], 한 변수의 독립모형(model with one factor independent of the other two)은 [12][3], [13][2], [1][23], 그리고 완전독립모형(completely independent model)은 [1][2][3]으로 표기한다. 세 번째 변수를 교락(confounder)변수로 간주하여 첫 번째 변수와 두 번째 변수의 교호작용에 대하여 세 번째 변수가 차원축소 가능한 모형은 [12][3], [12][13], [12][23]이며 그 이외 다른 모형은 차원축소 불가능한 모형이다(자세한 이론은 Bishop, Fienberg와 Holland(1975, pp. 39, 47), Agresti(1984, pp. 146), Christensen(1990, pp. 114) 그리고 Ducharme과 Lepage(1986)을 참조할 것).

$2\times2\times K$ 분할표 자료에 대하여 위에서 연구한 오즈비에 관하여 정리하면 다음과 같다.
첫 번째 변수와 두 번째 변수의 교호작용에 대하여 세 번째 변수가 차원축소 가능한 모형인 [12][3], [12][13], [12][23] 모형(Ducharme과 Lepage(1986)에서는 strong collapsibility라고 정의함)에서는 다음과 같은 관계식을 얻는다.

$$\theta_1 = \cdots = \theta_K = \theta_c \neq 1 \tag{2.1}$$

첫 번째 변수와 두 번째 변수의 교호작용에 대하여 세 번째 변수가 차원축소 불가능한 모형 중 [1][2][3], [1][23], [13][2] 모형(Ducharme과 Lepage(1986)는 위의 차원축소 가능한 모형과 더불어 이 모형들을 strict collapsibility 라고 정의함)에서는

$$\theta_1 = \cdots = \theta_K = \theta_c = 1 \tag{2.2}$$

이며, [13][23] 모형에서는

$$\theta_1 = \cdots = \theta_K = 1 \tag{2.3}$$

의 관계를 유도하였다. 그리고 [12][13][23] 모형에서는 이차교호작용항이 존재하지 않기 때문에 오직 다음과 같은 관계식만을 유도할 수 있다.

$$\theta_1 = \cdots = \theta_K \tag{2.4}$$

삼차원 분할표 자료에 대한 여러종류의 로그선형모형에 중 부분연관모형([12][13][23]) 이외의 모형은 직접해(direct solution)를 구할 수 있는 모형들이며, 각 모형에 적합한 칸 확률 p_{ijk}는 주어진 충분합 형태(sufficient configuration)의 함수로 구하고 대응하는 칸 빈도 x_{ijk}는 표본크기 $N=1{,}000$과 모비율 $\{p_{ijk}\}$로 이루어진 다항분포를 따르는 난수를 생성하는 모의실험을 통하여 구한다. 예를 들어, 조건부 독립모형 중 [13][23] 모형의 칸 확률 p_{ijk}는 주어진 충분합 형태 $\{p_{i+k}\}$, $\{p_{+jk}\}$의 주변확률표를 이용하여 $p_{ijk}=p_{i+k}p_{+jk}/p_{++k}$의 관계식을 이용하여 구한다. 직접해가 존재하지 않은 부분연관모형인 경우에는 정동빈, 홍종선과 윤상호(2003)의 연구에서 사용한 방법을 사용하여, 적절한 구간의 균일분포를 따르는 난수를 생성하고 로그선형모형의 여러

모수의 추정값을 얻은 후 이에 대응하는 칸 빈도 x_{ijk}를 생성하였다.

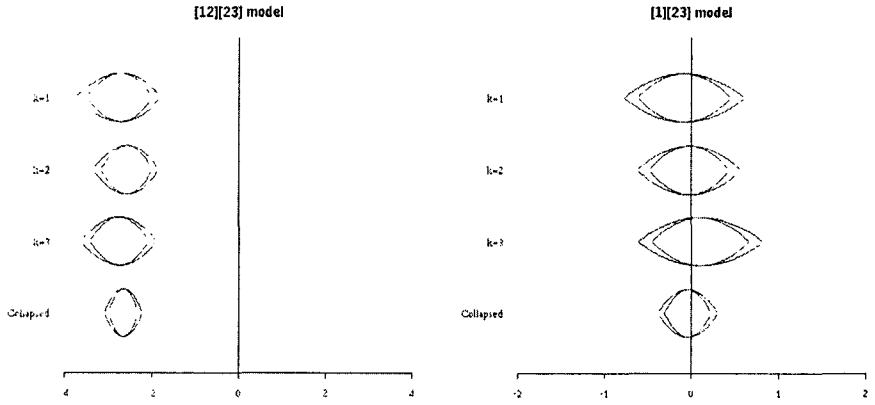
2×2×3 분할표 자료에 대하여 Raindrop plot을 작성하여 수록하였다. Raindrop plot에서는 4개의 Raindrop이 존재하는데 마지막 하단부의 Raindrop이 차원축소된 분할표에 대한 것이다.

여러 종류의 로그선형모형에 적합한 자료를 여러 번 생성하고 대응하는 Raindrop plot을 작성하여 살펴본 결과 다음과 같은 결론을 유도할 수 있었다. 첫 번째 변수와 두 번째 변수의 교호작용에 대하여 세 번째 변수가 차원축소 가능한 모형인 [12][3], [12][13], [12][23] 모형(strong and strict collapsibility)에서는 (2.1) 식에 정리된 이론과 같이 모든 오즈비 θ_k, θ_c값은 유사하며 모두 1에서 멀리 떨어진 값을 갖고 Raindrop plot에서 로그오즈비의 95%, 99% 신뢰구간에 0 값을 포함하지 않는다는 것을 발견하였다. 따라서 이 모형들의 모든 오즈비 θ_k, θ_c 값이 1이 아님을 Raindrop plot을 통해서 통계적으로 확신할 수 있다.

첫 번째 변수와 두 번째 변수의 교호작용에 대하여 세 번째 변수가 차원축소 불가능한 모형 중 [1][2][3], [1][23], [13][2] 모형(strict collapsibility이지만 strong collapsibility는 아님)에 대하여는 (2.2) 식에 정리된 이론을 Raindrop plot으로 확인할 수 있다. 즉 Raindrop plot에서 로그오즈비의 신뢰구간에 0값이 포함되어 있다.

그러나 첫 번째 변수와 두 번째 변수의 교호작용에 대하여 세 번째 변수가 차원축소 불가능한 모형 중 [13][23] 모형에서는 (2.3) 식에서 논의된 바와 같이 차원축소된 분할표에서의 θ_c값이 1에서 멀리 떨어진 경우가 대부분이었다. Raindrop plot에서 로그오즈비 θ_c^l의 신뢰구간에 0 값이 포함되지 않기 때문에 오즈비 θ_c는 모두 1에 가까운 값을 갖는다고 판단할 수 없다. 또한 차원축소 불가능한 모형 중 이차교호작용항이 존재하지 않는 [12][13][23] 모형에서는 오즈비에 대하여 (2.4) 식 이외에 어떠한 이론을 유도할 수 없는데 Raindrop plot을 살펴보면 상이한 형태임을 발견할 수 있었다.

그림 1. 차원축소가능모형과 차원축소불가능모형의 Raindrop Plot 예



3. 결론

2절에서 토론한 여러 로그선형모형에 대하여 적합한 자료를 바탕으로 작성한 Raindrop plot을 살펴보면, θ_k, θ_c에 대한 명확한 이론이 설정된 [12][3], [12][13], [12][23] 모형과 [1][2][3],

[1][23], [13][2] 모형에 대한 결과와 일치하는 것을 살펴볼 수 있다. 즉 [12][3], [12][13], [12][23] 모형에서는 모든 오즈비의 값이 동일하고 로그오즈비의 신뢰구간에 0값이 포함되지 않았으며 [1][2][3], [1][23], [13][2] 모형에서는 모든 오즈비의 값이 1에 가까운 것을 Raindrop plot을 이용해서 통계적으로 확신할 수 있었다.

본 연구에서는 Ducharme과 Lepage(1986)에서는 strong과 strict collapsibility 라는 정의와 Christensen(1990, pp. 113), Agresti(1984, pp. 146)가 내린 일반적인 차원축소에 대한 정의를 따라서 우리는 다음과 같은 결론을 유도할 수 있으며, 이와 같은 판단 기준은 Raindrop plot을 사용하여 보다 통계적으로 설정할 수 있다.

• 첫 번째 변수와 두 번째 변수의 교호작용에 대하여 세 번째 변수가 차원축소 가능한 모형인 [12][3], [12][13], [12][23] 모형은 strong and strict collapsibility 모형으로 정의하는데 이 모형에 대한 Raindrop plot에서 로그우도비의 신뢰구간에 0값을 포함하지 않는다. 따라서 세 번째 변수에 차원축소 가능한 모형의 모든 오즈비 θ_k와 θ_c의 값은 유사하며 그 값들이 1이 아니다라는 (2.1) 식의 이론과 일치함을 발견하였다.

• 세 번째 변수가 차원축소 불가능한 모형인 [1][2][3], [1][23], [13][2] 모형은 strict collapsibility이지만 strong collapsibility는 아닌 모형으로 정의하는데 이 모형에 대한 Raindrop plot에서 로그오즈비의 신뢰구간은 0값을 포함한다. 따라서 이 모형의 모든 오즈비 θ_k와 θ_c의 값은 모두 1에 근접한 유사한 값을 갖고 있음을 발견하였으며 이것을 요약한 관계이론은 (2.2) 식과 같음을 발견하였다.

참고문헌

Agresti, A. (1984). *Analysis of Ordinary Categorical Data*, John Wiley and Sons.

Agresti, A. (1990). *Categorical Data Analysis*, John Wiley and Sons.

Barrowman, N. J. and Myers, R. A. (2003). Raindrop Plots: A New Way to Display Collections of Likelyhoods and Distributions, *The American Statistician*, Vol. 57, 268-274.

Bishop, Yvonne M. M., Fienberg, Steve E., and Holland, Paul W. (1975). *Discrete Multivariate Analysis*, Cambridge, Massachusetts: MIT Press.

Christensen, Ronaldo. (1990). *Log-Linear Models*, New York: Springer-Verlag.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*, Chapman & Hall.

Ducharme, G. R. and Lepage, Y. (1986). Testing Collapsibility in Contingency Tables, *Journal of the Royal Statistical Society*, B, Vol. 48, No. 2, 197-205.

Jeong, D. B., Hong, C. S., and Yoon, S. H. (2003). Empirical Comparisons of Disparity Measures for Partial Association Models in Three Dimensional Contingency Tables, *The Korean Communications in Statistics*, Vol. 10, No. 1, 135-144.

Yamamoto, E. and Doi, M. (2001). Noncollapsibility of Common Odds Ratios without/with Confounding, *Bulletin of The* 53^{th} *Session of the International Statistical Institute*, Book 3, 39-40.