The Korean Communications
in Statistics Vol. 10 No. 2, 2003
pp. 291-303

Outlying Cell Identification Method Using Interaction Estimates of Log-linear Models

Chong Sun Hong[1], Min Jung Jung[2]

Abstract

This work is proposed an alternative identification method of outlying cell which is one of important issues in categorical data analysis. One finds that there is a strong relationship between the location of an outlying cell and the corresponding parameter estimates of the well-fitted log-linear model. Among parameters of log-linear model, an outlying cell is affected by interaction terms rather than main effect terms. Hence one could identify an outlying cell by investigating of parameter estimates in an appropriate log-linear model.

Keywords : Contingency table, Identification, Interaction, Log-linear model, Main effect, Outlying cell, Residual.

1. 서론

분할표 자료에서 이상칸(outlying cell)을 식별하기 위한 다양한 기준들이 제시되었는데 한 종류는 다양한 형태의 잔차들이고 다른 종류는 적합도 검정통계량의 차이의 함수로 표현된다. Barnet과 Lewis(1994)는 이러한 다양한 기준들에 대하여 설명하고 이중 몇 가지 기준을 이용하여 모형에 불일치한 칸을 이상칸으로 식별하였다. Fienberg(1969), Kotz와 Hawkins(1984)는 로그 오즈(log odds)로 표현되는 테트라드(tetrads)를 기준으로, 그리고 Mostellar와 Parunak(1985)는 탐색적 접근방법으로 이상칸을 식별할 수 있다고 하였다.

Haberman(1973)은 (1.1)식과 같은 수정된 잔차(adjusted residual)을 제안하였다. 이는 관찰값 x_{ij}와 칸 기대값의 추정량 $\widehat{m}_{ij}$과의 차이를 측정하였으며 정규확률그림에 수정된 잔차를 표시하여 극단적인 칸을 이상칸으로 식별하였다.

$$r_{ij} = \frac{x_{ij} - \widehat{m}_{ij}}{\sqrt{var(\widehat{m}_{ij})}} = \frac{x_{ij} - (x_{i+}\, x_{+j})/N}{\sqrt{x_{i+}\, x_{+j}(N - x_{i+})(N - x_{+j})/N^3}}, \tag{1.1}$$

1) Professor, Department of Statistics, Sungkyunkwan University, Seoul 110-745 KOREA. cshong@skku.ac.kr

2) Researcher, Research & Research, Inc. Yeoksam Dong, Kannam Gu, Seoul, Korea. prograde@korea.com

여기서 $N=\sum_i\sum_j x_{ij}$, $x_{i+}=\sum_j x_{ij}$, $x_{+j}=\sum_i x_{ij}$ 는 각각 총합, i번째 행 주변합, 그리고 j번째 열 주변합을 나타낸다.

Brown(1974)은 $I\times J$인 분할표에서 하나의 칸 (i,j)를 제거한 후 (i,j)칸의 기대값을 (1.2)식과 같이 추정하였다.

$$m_{ij}^{*}=\frac{(x_{i+}-x_{ij})(x_{+j}-x_{ij})}{N-x_{i+}-x_{+j}+x_{ij}}. \tag{1.2}$$

추정량 m_{ij}^{*}를 준독립성(quasi-independence) 모형 하에서 (i,j)칸의 기대값으로 고려한다(Goodman, 1968). 적합도 검정통계량에 영향을 주는 칸은 (1.3)식과 같은 삭제된 잔차(deleted residual)를 이용하여 이상칸으로 식별되었다.

$$r_{ij}^{*}=\frac{x_{ij}-m_{ij}^{*}}{\sqrt{m_{ij}^{*}}}. \tag{1.3}$$

다차원 분할표에서 직접계산이 가능한 최고가능도추정량이 존재하는 분해모형(decomposable model)인 경우, Upton과 Guillen(1995)은 (1.3)식의 삭제된 잔차를 확장하여 하나의 이상칸을 식별하기 위한 일반적인 공식을 제안하였다. 그러나 설정된 모형의 최소충분통계량(minimal sufficient statistics)의 수가 증가하거나 둘 이상의 이상칸이 존재하는 경우에는 제안된 공식을 이용하는 것이 어렵다.

이들 기준에 의해 하나 이상의 이상칸을 식별하기 위한 몇 가지 식별방법들이 제안되었는데 그 중 하나는 Fuchs와 Kenett(1980)이 제안한 전진단계방법(forwards-stepping method)으로 이상칸으로 의심나는 가장 극단적인 칸으로부터 덜 극단적인 칸의 순서로 이상칸을 식별하는 방법이다. 다른 하나는 Simonoff(1988)가 제안한 후진단계방법(backwards-stepping method)으로 덜 극단적인 칸으로부터 가장 극단적인 칸의 순서로 이상칸을 식별하는 방법이다. Simonoff는 하나이상의 이상칸이 존재하는 경우 전진단계방법을 이용하여 이상칸을 식별하면 가장효과(masking effect ; 이상칸을 이상칸으로 식별하지 못하는 효과)를 유발할 수 있고 삭제된 잔차에 의한 후진단계방법을 이용하면 가장효과와 편승효과(swamping effect ; 이상칸이 아닌 칸을 이상칸으로 식별하는 효과)가 제한된다고 하였다. 그럼에도 불구하고 이들 방법은 범주형 변수의 범주 수가 증가할수록 이상칸을 식별하는데 많은 시간이 소요될 뿐만 아니라 다차원 분할표에 대해서는 이상칸을 식별할 수 없다. 이종철과 홍종선(2000)은 후진단계방법을 응용하여 적은 계산으로 동시에 두 개 이상의 다중 이상칸 식별이 가능하며 임의의 다차원 로그선형모형(log-linear model)에 대하여도 적용할 수 있는 다중 이상칸 식별방법을 제안하였다. 또한 이종철과 홍종선(2001)은 이차원 분할표 자료를 대응분석(correspondence analysis) 하여 대응분석 그림분석을 통하여 이상칸을 식별하는 방법도 제안하였다.

본 논문에서는 이상칸에 해당하는 로그선형모형의 모수 추정값과 그 이상칸과의 관계를 이용하여 분할표 자료의 이상칸을 식별하는 방법을 제안한다. 이차원과 삼차원 분할표자료에 이상칸이 존재하는 경우를 고려하여 자료에 적합한 로그선형모형을 선택하고 모형에 포함된 모수의 추정값

의 유의성을 살펴보고자 한다. 이상칸의 기대값에 해당하는 로그선형모형의 모수는 주효과항(main effect term)과 교호작용항(interaction term)으로 구분되는데 주효과항과 교호작용항의 추정량 중에서 밀접한 관계를 갖고 있는 모수가 무엇인지를 탐색적으로 발견하고자 한다. 따라서 교호작용항이 포함된 모형과 교호작용항이 포함되지 않고 오직 주효과항만 포함된 모형으로 구분하여 살펴보기로 한다. 본 논문의 2절에서는 교호작용항을 포함한 로그선형모형들의 모수 추정값과 이상칸과의 관계를 유도하여 모수 중 교호작용항의 유의성을 판단함으로써 이상칸을 식별할 수 있는 방법을 설명하고, 3절에서는 교호작용항이 포함되지 않은 로그선형모형으로는 이상칸을 식별하기 어렵다는 것을 토론한다. 마지막 결론에서는 기존의 문헌에서 소개된 이상칸 식별방법과는 다르게 분할표 자료에 적합한 로그선형모형의 모수 추정량으로 이상칸을 식별할 수 있다는 장점과 향후 연구과제를 언급하였다.

2. 교호작용항과 이상칸

본절에서는 하나의 이상칸이 존재하는 이차원과 삼차원 분할표자료를 고려하면서, 교호작용항을 포함하는 로그선형모형 중에서 이차원일 때는 포화모형(saturated model), 삼차원인 경우에는 조건부 독립모형(conditional independent model)과 부분연관모형(partial association model)의 모수추정량을 살펴보자.

2.1 이차원 포화모형

【표 2.1】 이차원 포화 모형

A \ B	1	2	3	4	5
1	13	14	19	20	19
2	11	10	12	52	12
3	10	11	15	16	15
4	14	16	20	22	21

【표 2.1】 의 자료는 행변수 A의 범주 수준이 4이며 열변수 B의 범주 수준이 5인 4×5 분할표이며, 완전독립모형(completely independent model) 하에서 가능도비 검정통계량값은 G^2=28.06이고 이에 대응하는 p-값=0.0054 로 독립성 모형이라는 가설을 기각할 수 있으므로 【표 2.1】에 적합한 모형은 (2.1)식과 같은 이차원 포화모형(saturated model)이다([AB] 모형).

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \tag{2.1}$$

여기서 u항들의 일반적인 제약식은 서술을 생략한다(자세한 사항은 Bishop et al. (1991)을 참조). 【표 2.1】에서 이상칸 여부를 검정하기 위해 모든 칸에 대하여 (1.1)식의 수정된 잔차 $|r_{ij}|$를 계산한 후 살펴보면, (2,4)칸의 수정된 잔차는 $|r_{24}|$=5.342로 가장 유의한 검정결과를 나타낸다. 또

한 (2,4)칸의 삭제된 잔차가 $|r_{24}^{*}|=10.183$으로 다른 칸의 $|r_{ij}^{*}|$보다 크며 유의한 검정 결과를나타내고 있다. 따라서 (2,4)칸이 이상칸이라고 판단할 수 있다. 【표 2.1】의 자료에 이차원 포화모형을 적합시킨 후 로그선형모형의 모수 중에서 주효과항과 교호작용항 추정량과 표준오차를 구하고 각각의 추정량의 유의함을 검정한 결과는 다음과 같이 【표 2.2】와 【표 2.3】에 나열하였다.

【표 2.2】 로그선형모형의 주효과항 추정량

추정량	추정값	표준오차	X^2	p-값
$\hat{u}_{1(1)}$	0.0606	0.0971	0.3895	0.5326
$\hat{u}_{1(2)}$	-0.0328	0.1044	0.0988	0.7533
$\hat{u}_{1(3)}$	-0.1793	0.1053	2.8968	0.0888
$\hat{u}_{1(4)}$	0.1515	0.0943	2.5779	0.1084
$\hat{u}_{2(1)}$	-0.2810	0.1268	4.9094	0.0267
$\hat{u}_{2(2)}$	-0.2291	0.1247	3.3766	0.0661
$\hat{u}_{2(3)}$	0.0262	0.1133	0.0534	0.8173
$\hat{u}_{2(4)}$	0.4455	0.1000	19.8408	0.0000
$\hat{u}_{2(5)}$	0.0384	0.1129	0.1155	0.7340

【표 2.3】 로그선형모형의 교호작용항 추정량

추정량	추정값	표준오차	X^2	p-값	추정량	추정값	표준오차	X^2	p-값
$\hat{u}_{12(11)}$	0.0282	0.2127	0.0176	0.8944	$\hat{u}_{12(31)}$	0.0057	0.2320	0.0006	0.9803
$\hat{u}_{12(12)}$	0.0504	0.2075	0.0591	0.8080	$\hat{u}_{12(32)}$	0.0491	0.2248	0.0478	0.8270
$\hat{u}_{12(13)}$	0.1005	0.1863	0.2914	0.5893	$\hat{u}_{12(33)}$	0.1040	0.2014	0.2669	0.6055
$\hat{u}_{12(14)}$	0.2675	0.1763	2.3030	0.1291	$\hat{u}_{12(34)}$	0.2508	0.1910	1.7236	0.1892
$\hat{u}_{12(15)}$	0.0883	0.1860	0.2256	0.6348	$\hat{u}_{12(35)}$	0.0919	0.2012	0.2084	0.6480
$\hat{u}_{12(21)}$	0.0454	0.2256	0.0405	0.8404	$\hat{u}_{12(41)}$	0.0115	0.2075	0.0031	0.9560
$\hat{u}_{12(22)}$	0.1926	0.2304	0.6992	0.4031	$\hat{u}_{12(42)}$	0.0931	0.1996	0.2175	0.6409
$\hat{u}_{12(23)}$	0.2656	0.2130	1.5545	0.2125	$\hat{u}_{12(43)}$	0.0610	0.1827	0.1114	0.7386
$\hat{u}_{12(24)}$	0.7814	0.1527	6.1938	0.0000	$\hat{u}_{12(44)}$	0.2631	0.1708	2.3718	0.1235
$\hat{u}_{12(25)}$	0.2778	0.2128	1.7039	0.1918	$\hat{u}_{12(45)}$	0.0976	0.1805	0.2923	0.5888

【표 2.2】를 통하여 주효과항의 추정량 중에서 $\hat{u}_{2(1)}$, $\hat{u}_{2(4)}$이 작은 p-값을 나타내기 때문에 유의함을 알 수 있다. 이상칸이 포함된 행($i=2$)과 열($j=4$)의 주효과항 추정량 $\hat{u}_{1(2)}$와

$\hat{u}_{2(4)}$의 모수추정량이 모두 유의하게 나타나지 않았으며 다른 주효과항 추정량 $\hat{u}_{2(1)}$이 유의하다는 것을 파악하였다. 그리고 【표 2.3】 에서는 교호작용항의 추정량 중 $\hat{u}_{12(24)}$만이 유의하며, 이 상칸이 포함된 행($i=2$)과 열($j=4$)의 교호작용항 추정량과 관계가 있음을 발견하였다. 이 결과는 앞에서 이상칸을 잔차로 식별한 결과와 일치하므로 로그선형모형의 모수 중에서 유의한 교호작용항 추정량 $\hat{u}_{12(24)}$에 의해 $i=2$, $j=4$일 때의 대응하는 (2,4)칸이 이상칸이라는 결론을 내릴 수 있다.

그러므로 이차원 포화모형에서 로그선형모형에 포함된 주효과항의 추정량이 유의하더라도 이상칸을 판단할 수 없으나, 교호작용항 추정량 $\hat{u}_{12(ij)}$이 유의한 경우에는 교호작용항에 대응하는 (i, j)칸이 이차원 분할표에서 이상칸이라고 식별할 수 있음을 발견하였다.

2.2 삼차원 조건부 독립모형

【표 2.4】 삼차원 조건부 독립모형

	C = 1			C = 2			C = 3		
A \ B	1	2	3	1	2	3	1	2	3
1	30	11	29	41	15	40	24	9	23
2	37	13	36	12	18	48	29	11	28
3	69	25	67	93	34	90	54	20	53

【표 2.4】 자료는 행변수 A와 열변수 B, 그리고 층변수 C의 범주 수준이 모두 3인 3×3×3 분할표이며, 이 자료에 모든 로그선형모형을 적합 시켜본 결과 가장 적합한 모형은 (2.2)식과 같은 조건부 독립모형(conditional independent model)이며([AB][AC] 모형), 이 모형 하에서의 가능도비 검정통계량은 $G^2=19.00$이고 이에 대응하는 p-값=0.089이다. 따라서 【표 2.4】 에 적합한 모형은 (2.2)식과 같은 조건부 독립모형이다.

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} \tag{2.2}$$

각 칸의 이상칸 여부를 검증하기 위해 (1.1)식의 수정된 잔차 $|r_{ijk}|$를 계산하면, (2,1,2)칸의 수정된 잔차값의 절대값이 $|r_{212}|=6.4131$으로 다른 칸의 잔차보다 유의한 값을 가지므로 (2,1,2)칸은 이상칸이라 판단할 수 있다.(삼차원 이상의 자료에서는 삭제된 잔차를 구할 수 없기 때문에 언급을 생략함.) 【표 2.4】 에 적합한 조건부 독립모형의 주효과항들은 고려하지 않고 오직 교호작용항들의 추정량을 구한 결과는 【표 2.5】 와 같다. (여기서도 주효과항의 추정량의 유의함과 이상칸과는 연관이 없으므로 이에 대하여는 서술하지 않았다.)

【표 2.5】 로그선형모형의 교호작용항 추정량

추정량	추정값	표준오차	X^2	p-값	추정량	추정값	표준오차	X^2	p-값
$\hat{u}_{12(11)}$	0.0856	0.0742	1.3289	0.2490	$\hat{u}_{13(11)}$	-0.0507	0.0756	0.4502	0.5023
$\hat{u}_{12(12)}$	-0.0385	0.0943	0.1669	0.6829	$\hat{u}_{13(12)}$	0.0929	0.0699	1.7642	0.1841
$\hat{u}_{12(13)}$	-0.0470	0.0736	0.4084	0.5228	$\hat{u}_{13(13)}$	-0.0421	0.0808	0.2721	0.6019
$\hat{u}_{12(21)}$	-0.1722	0.0748	5.2977	**0.0214**	$\hat{u}_{13(21)}$	0.0910	0.0733	1.5418	0.2144
$\hat{u}_{12(22)}$	0.0832	0.0905	0.8447	0.3581	$\hat{u}_{13(22)}$	-0.1789	0.0704	6.4579	**0.0110**
$\hat{u}_{12(23)}$	0.0890	0.0710	1.5730	0.2098	$\hat{u}_{13(23)}$	0.0879	0.0780	1.2706	0.2596
$\hat{u}_{12(31)}$	0.0867	0.0618	1.9676	0.1607	$\hat{u}_{13(31)}$	-0.0403	0.0634	0.4037	0.5252
$\hat{u}_{12(32)}$	-0.0447	0.0780	0.3282	0.5667	$\hat{u}_{13(32)}$	0.0860	0.0573	2.2507	0.1336
$\hat{u}_{12(33)}$	-0.0420	0.0609	0.4758	0.4903	$\hat{u}_{13(33)}$	-0.0457	0.0681	0.4516	0.5016

【표 2.5】 를 통해서 교호작용항의 추정량 중에서 $\hat{u}_{12(21)}$와 $\hat{u}_{13(22)}$만이 유의한 결과를 보이고 있음을 알 수 있다. 2.1절에서 논의한 이차원 포화모형에서와 같이, 유의한 교호작용항 추정량 $\hat{u}_{12(21)}$으로부터 $i=2$, $j=1$인 경우와 $\hat{u}_{13(22)}$으로부터 $i=2$, $k=2$인 경우인 $i=2$, $j=1$, $k=2$ 인 칸에 대응하는 (2,1,2)칸을 이상칸이라 판단할 수 있다. 그러므로 삼차원 조건부 독립모형에서도 로그선형 모형의 교호작용항의 추정량이 유의하므로 추정량들의 i와 j와 k에 대응하는 (i,j,k)칸이 이상칸이라 식별할 수 있다.

2.3 삼차원 부분연관모형

【표 2.6】 삼차원 부분 연관 모형

	C = 1			C = 2			C = 3		
A \ B	1	2	3	1	2	3	1	2	3
1	23	11	12	24	12	12	31	15	16
2	72	35	36	75	37	38	38	47	49
3	53	26	27	56	27	26	72	35	38

(2.3)식과 같은 부분연관모형(partial association model)은 【표 2.6】 의 3×3×3 분할표 자료에 가장 적합한 모형이며, 이 모형 하에서의 가능도비 검정통계량은 $G^2=10.12$이고 이에 대응하는 p-값=0.2499이다. 자료를 가장 잘 설명하는 모형은 부분연관모형([AB][AC][BC] 모형)이다.

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} \tag{2.3}$$

【표 2.6】 자료의 이상칸 여부를 검정하기 위해 모든 칸에 대하여 수정된 잔차를 계산한 결과,

다른 잔차보다 유의한 절대값을 가진 (2,1,3)칸의 수정된 잔차값은 $|r_{213}|=1.93548$ 이며 따라서 (2,1,3)칸은 【표 2.6】 자료에 적합한 부분연관모형에 대한 이상칸임을 파악할 수 있다. 그리고 부분연관모형을 적합시킨 후 로그선형모형의 교호작용항 추정량에 대한 결과는 【표 2.7】과 같다.

【표 2.7】를 살펴보면, 교호작용항 중에서 $\hat{u}_{12(21)}$, $\hat{u}_{13(23)}$ 그리고 $\hat{u}_{23(13)}$만이 유의한 추정량임을 나타내고 있다. 유의한 추정량인 $\hat{u}_{12(21)}$으로부터 $i=2$, $j=1$인 경우와 $\hat{u}_{13(23)}$으로부터 $i=2$, $k=3$인 경우, 그리고 $\hat{u}_{23(13)}$으로부터 $j=1$, $k=3$인 경우에 대응하는 (2,1,3)칸이 이상칸임을 식별할 수 있다. 따라서 삼차원 분할표 자료에 적합하는 부분연관모형에서는 로그선형모형에 포함되고 있는 세 개의 교호작용항 $\hat{u}_{12(ij)}$, $\hat{u}_{13(ik)}$, $\hat{u}_{23(jk)}$에 의해 이상칸을 식별할 수 있음을 발견하였다. 삼차원에서는 조건부 독립모형과 부분연관모형인데 이러한 모형들의 교호작용항이 유의하므로 이에 대응하는 칸이 이상칸임을 발견하였다.

2.1절과 유사하게 삼차원 포화모형에서도 일차 교호작용항과 이차 교호작용항의 유의함을 검정하면서 이상칸을 식별할 수 있는데 모형의 적합도 검정의 일반적인 측면에서 자료에 적합한 모형으로 포화모형을 사용하지 않으므로 본 연구에서는 언급하지 않기로 한다. 그러나 2.1절에서 논의한 이차원 범주형 자료에 대한 분석에서는 교호작용항을 포함하고 있는 로그선형모형은 포화모형뿐이기 때문에 포화모형에 대한 연구를 이상칸의 식별방법에 관한 문제 제기차원에서 다룬 것이다. 그러므로 우리는 자료에 적합한 로그선형모형의 모수 중에서 주효과항을 제외한 교호작용항의 유의성을 기준으로 유의한 교호작용항에 대응하는 칸을 이상칸이라고 식별할 수 있음을 탐색적으로발견하였다.

【표 2.7】 로그선형모형의 교호작용항 추정량

추정량	추정값	표준오차	X^2	p-값	추정량	추정값	표준오차	X^2	p-값
$\hat{u}_{12(11)}$	0.0573	0.0780	0.5395	0.4627	$\hat{u}_{13(23)}$	-0.1676	0.0636	6.9539	**0.0084**
$\hat{u}_{12(12)}$	-0.0347	0.0917	0.1427	0.7056	$\hat{u}_{13(31)}$	-0.0398	0.0686	0.3372	0.5614
$\hat{u}_{12(13)}$	-0.0226	0.0905	0.0626	0.8024	$\hat{u}_{13(32)}$	-0.0514	0.0680	0.5705	0.4501
$\hat{u}_{12(21)}$	**-0.1219**	**0.0608**	4.0154	**0.0451**	$\hat{u}_{13(33)}$	0.0912	0.0643	2.0085	0.1564
$\hat{u}_{12(22)}$	0.0641	0.0697	0.8443	0.3582	$\hat{u}_{23(11)}$	0.0763	0.0626	1.4868	0.2227
$\hat{u}_{12(23)}$	0.0578	0.0690	0.7030	0.4018	$\hat{u}_{23(12)}$	0.0847	0.0620	1.8683	0.1717
$\hat{u}_{12(31)}$	0.0646	0.0627	1.0605	0.3031	$\hat{u}_{23(13)}$	**-0.1611**	**0.0608**	**7.0053**	**0.0081**
$\hat{u}_{12(32)}$	-0.0294	0.0735	0.1599	0.6893	$\hat{u}_{23(21)}$	-0.0418	0.0731	0.3266	0.5677
$\hat{u}_{12(33)}$	-0.0352	0.0728	0.2337	0.6288	$\hat{u}_{23(22)}$	-0.0255	0.0722	0.1251	0.7236
$\hat{u}_{13(11)}$	-0.0398	0.0855	0.2168	0.6415	$\hat{u}_{23(23)}$	0.0673	0.0685	0.9661	0.3257
$\hat{u}_{13(12)}$	-0.0367	0.0845	0.1879	0.6646	$\hat{u}_{23(31)}$	-0.0346	0.0724	0.2281	0.6330
$\hat{u}_{13(13)}$	0.0764	0.0800	0.9136	0.3392	$\hat{u}_{23(32)}$	-0.0592	0.0719	0.6764	0.4108
$\hat{u}_{13(21)}$	0.0796	0.0654	1.4839	0.2232	$\hat{u}_{23(33)}$	0.0937	0.0677	1.9172	0.1662
$\hat{u}_{13(22)}$	0.0880	0.0647	1.8520	0.1736					

3. 주효과항과 이상칸

2절에서는 교호작용항을 포함한 로그선형모형이 적합한 자료에 이상칸이 존재하는 경우를 살펴보았지만, 본절에서는 교호작용항을 포함하지 않고 주효과항을 포함하고 있는 로그선형모형이 잘 설명하는 자료에 이상칸이 존재하는 경우를 고려하자.

3.1 이차원 완전독립모형

【표 3.8】 자료에 대하여 두 변수의 독립성을 가정한 (3.1)식과 같은 로그선형모형을 적합시킨 결과 최대우도비 검정통계량값은 G^2=13.3이고 이에 대응하는 p-값=0.3474이므로 【표 3.8】 자료에 적합한 모형은 완전독립모형(completely independent model)이다([A][B] 모형).

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} \tag{3.1}$$

분할표 자료 【표 3.8】의 모든 칸에 대하여 수정된 잔차 $|r_{ij}|$를 계산하면, $|r_{33}|$=3.8275으로 다른 수정된 잔차보다 큰 절대값을 갖으며 유의하다. 따라서 (3,3)칸이 이상칸이라고 판단할 수 있다. 완전독립모형을 적합시킨 후 로그선형모형의 주효과항에 대한 결과는 【표 3.9】와 같다.

【표 3.8】 이차원 완전독립모형

A \ B	1	2	3	4	5
1	29	35	27	14	56
2	21	24	19	10	39
3	14	16	33	7	27
4	27	32	25	13	52

【표 3.9】 로그선형모형의 주효과항 추정량

추정량	추정값	표준오차	X^2	p-값
$\hat{u}_{1(1)}$	0.2345	0.07148	10.763	0.0010
$\hat{u}_{1(2)}$	-0.1195	0.08018	2.219	0.1362
$\hat{u}_{1(3)}$	-0.2722	0.08461	10.344	0.0013
$\hat{u}_{1(4)}$	0.1570	0.07321	4.602	0.0319
$\hat{u}_{2(1)}$	-0.0434	0.0945	0.2109	0.6460
$\hat{u}_{2(2)}$	0.1185	0.0891	1.7694	0.1834
$\hat{u}_{2(3)}$	0.0901	0.0900	1.0020	0.3168
$\hat{u}_{2(4)}$	-0.7700	0.1263	37.1253	0.0000
$\hat{u}_{2(5)}$	0.6047	0.0760	63.2247	0.0000

【표 3.0】에서 로그선형모형의 주효과항인 $u_{1(i)}$과 $u_{2(j)}$에 관하여 살펴보면, $\hat{u}_{1(1)}$, $\hat{u}_{1(3)}$, $\hat{u}_{1(4)}$, $\hat{u}_{2(4)}$, $\hat{u}_{2(5)}$의 추정량이 유의하다. 여기서 이상칸이 포함된 행과 열의 주효과항 추정량인 $\hat{u}_{1(3)}$와 $\hat{u}_{2(3)}$ 만이 유의하게 나타나는 것이 아니라 다른 추정량이 유의한 것을 알 수 있다. 【표 3.8】 자료에서 수정된 잔차를 통하여 식별한 이상칸은 (3,3)칸이지만, 이차원 완전독립모형에서의 주효과항 추정량으로는 이상칸을 식별하기는 어렵다는 것을 발견하였다.

3.2 삼차원 완전독립모형

【표 3.10】의 3×3×3 분할표 자료는 완전독립모형 하에서 가능도비 검정통계량값은 $G^2 = 5.72$이고 이에 대응하는 p-값=0.9992 으로 이 자료에 적합한 모형은 (3.2)식과 같은 삼차원 완전독립모형이다([A][B][C] 모형).

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} \tag{3.2}$$

【표 3.10】에서 모든 칸에 대한 수정된 잔차를 계산하면, (1,1,1)칸에 대응하는 수정된 잔차의 절대값이 $|r_{111}| = 3.6396$으로 다른 칸의 잔차보다 큰 절대값을 가지며 유의하다. 그러므로 (1,1,1)칸이 이상칸임을 판단할 수 있다. 【표 3.10】에 완전독립모형을 적합시킨 후 로그선형모형의 모수 추정량을 구한 결과는 다음과 같다.

【표 3.10】 삼차원 완전독립모형

	C = 1			C = 2			C = 3		
A \ B	1	2	3	1	2	3	1	2	3
1	10	17	22	35	27	35	35	26	35
2	18	14	19	30	22	30	29	22	29
3	12	9	12	19	14	19	18	14	18

【표 3.11】 로그선형모형의 주효과항 추정량

추정량	추정값	표준오차	X^2	p-값
$\hat{u}_{1(1)}$	0.2371	0.0564	17.6716	0.0000
$\hat{u}_{1(2)}$	0.1095	0.0580	3.5565	0.0593
$\hat{u}_{1(3)}$	-0.3466	0.0654	28.1079	0.0000
$\hat{u}_{2(1)}$	0.0536	0.0578	0.8599	0.3538
$\hat{u}_{2(2)}$	-0.1684	0.0612	7.5775	0.0059
$\hat{u}_{2(3)}$	0.1148	0.0569	4.0630	0.0438
$\hat{u}_{3(1)}$	-0.3608	0.0657	38.9849	0.0000
$\hat{u}_{3(2)}$	0.1913	0.0570	9.7858	0.0018
$\hat{u}_{3(3)}$	0.1694	0.0573	8.8545	0.0029

【표 3.11】에서 주효과항들의 추정량을 살펴보면 이상칸이 포함된 첫 번째 행과 첫 번째 열 그리고 첫 번째 층에 대응하는 주효과항 추정량이 유의하게 나타나는 것이 아니라 이상칸의 위치와 관계없는 $\hat{u}_{1(1)}$, $\hat{u}_{1(3)}$, $\hat{u}_{2(2)}$, $\hat{u}_{3(1)}$, $\hat{u}_{3(2)}$, $\hat{u}_{3(3)}$이 유의한 것을 알 수 있다. 따라서 2.1절과 3.1절과 동일하게 삼차원 완전독립모형 하에서도 로그선형모형의 주효과항으로는 이상칸을 식별하기는 불가능하다고 판단할 수 있다.

3.3 삼차원 한 변수 독립모형

【표 3.12】 삼차원 한 변수 독립모형

	C = 1			C = 2			C = 3		
A \ B	1	2	3	1	2	3	1	2	3
1	19	35	29	19	36	30	11	22	18
2	41	78	65	42	80	67	26	48	41
3	25	47	39	61	48	40	15	29	24

【표 3.0】의 분할표 자료에 (3.2)식과 같은 한 변수 독립모형(model with one factor independent of the other two)을 적합시킨 결과, 가능도비 검정통계량은 $G^2=17.38$이고 대응하는 p-값=0.316이므로 자료에 적합한 모형은 한 변수 독립모형이다([AB][C] 모형).

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} \tag{3.3}$$

【표 3.0】의 모든 칸에 대한 수정된 잔차를 계산한 결과, (3,1,2)칸에 수정된 잔차는 $|r_{312}|=4.2022$으로 유의하므로 (3,1,2)칸이 이상칸임을 인지할 수 있다. 또한 【표 3.0】 자료에 한 변수 독

립모형을 적합시킨 후 로그선형모형의 주효과항과 교호작용항에 대한 결과는 【표 3.13】에 나열하였다.

【표 3.13】 로그선형모형의 모수 추정량

추정량	추정값	표준오차	X^2	p-값	추정량	추정값	표준오차	X^2	p-값
$\hat{u}_{1(1)}$	-0.4118	0.0526	61.3005	0.0000	$\hat{u}_{12(11)}$	-0.0958	0.0806	1.4130	0.2346
$\hat{u}_{1(2)}$	0.3896	0.0432	81.4042	0.0000	$\hat{u}_{12(12)}$	0.0508	0.0696	0.5337	0.4651
$\hat{u}_{1(3)}$	0.0222	0.0464	0.22792	0.6331	$\hat{u}_{12(13)}$	0.0450	0.0726	0.3846	0.5351
$\hat{u}_{2(1)}$	-0.2684	0.0509	27.7849	0.0000	$\hat{u}_{12(21)}$	-0.0977	0.0659	2.1966	0.1383
$\hat{u}_{2(2)}$	0.2257	0.0448	25.3716	0.0000	$\hat{u}_{12(22)}$	0.0447	0.0573	0.6079	0.4356
$\hat{u}_{2(3)}$	0.0427	0.0468	0.83543	0.3607	$\hat{u}_{12(23)}$	0.0531	0.0597	0.7902	0.3740
$\hat{u}_{3(1)}$	0.1224	0.0437	7.8259	0.0052	$\hat{u}_{12(31)}$	0.1935	0.0687	7.9367	0.0048
$\hat{u}_{3(2)}$	0.2348	0.0427	30.3116	0.0000	$\hat{u}_{12(32)}$	-0.0955	0.0627	2.3200	0.1277
$\hat{u}_{3(3)}$	-0.3572	0.0496	51.9544	0.0000	$\hat{u}_{12(33)}$	-0.0980	0.0655	2.2411	0.1344

【표 3.13】의 결과를 살펴보면, 주효과항들은 이상칸의 위치인 $i=3$, $j=1$, $k=2$와 무관하게 주효과항 { $\hat{u}_{1(i)}$: $i=1,2$, $\hat{u}_{2(j)}$: $j=1,2$, $\hat{u}_{3(k)}$: $k=1,2,3$}이 모두 유의하므로 주효과항 추정량으로는 이상칸을 식별할 수 없다. 그리고 교호작용항을 살펴보면, $\hat{u}_{12(31)}$ 항만이 유의한 결과를 나타내고 있다. 그러므로 한 변수 독립모형에 적합한 자료에 존재하고 있는 이상칸은 모형에 포함하고 있는 유일한 교호작용항인 $\hat{u}_{12(ij)}$과 밀접한 관계가 있음을 발견할 수 있고, 따라서 로그선형모형의 교호작용항 추정량으로 식별한 이상칸은 $i=3$, $j=1$일 때 대응하는 {$(3,1,k)$; $k=1,2,3$}칸이라 예측할 수 있다. 이러한 결과는 한 변수 독립모형에서 로그선형모형의 추정량으로 이상칸을 식별할 때 층변수 C에 대해서는 식별되지 않는다는 것을 의미한다. 그러므로 삼차원에서 한 변수 독립모형([AB][C] 모형)은 독립적인 관계를 갖고 있는 한 변수를 차원 축소(collapsible)가 가능한 모형이므로 축소시켜 이차원 포화모형([AB] 모형)으로 자료분석을 하면서 이상칸 식별하는 방법을 추천한다.

4. 결론 및 향후연구과제

본 논문에서는 이상칸이 존재하는 이차원과 삼차원 분할표에서 이 자료를 잘 설명해주는 로그선형모형을 선택하고 설정된 로그선형모형의 모수의 유의성을 기준으로 이상칸을 식별하는 방법을 제안하였다. 자료에 적합한 로그선형모형의 모수 추정량과 이상칸의 관계를 탐색적인 연구를 통해, 이상칸에 해당하는 로그선형모형의 주효과항에는 영향을 미치지 않고 교호작용항에 유의한 영향을 미친다는 것을 발견하였다.

이상칸이 존재하는 이차원 범주형 자료에 적합한 모형 중 주효과항만이 포함된 이차원 완전독

립모형([A][B] 모형)에서는 이상칸에 대응하는 로그선형모형의 주효과항이 유의성을 나타내지 않으므로 이상칸을 식별할 수 없었으며, 포화모형([AB] 모형)에서는 이상칸에 해당하는 로그선형모형의 교호작용항 추정량이 유의성을 나타냈기 때문에 유의한 교호작용항에 대응하는 칸을 이상칸으로 식별할 수 있었다. 즉 교호작용항 $\hat{u}_{12(ij)}$이 유의하면, $\hat{u}_{12(ij)}$의 i와 j에 대응하는 (i,j)칸이 이상칸임을 식별할 수 있다.

이상칸이 존재하는 삼차원 범주형 자료에 적합한 모형 중 완전독립모형([A][B][C] 모형)에서는 로그선형모형의 주효과항만으로 이상칸을 식별할 수 없었으며, 한 변수 독립모형(예를 들어 ([AB][C] 모형))에서는 유의한 교호작용항(예를 들어 $\hat{u}_{12(ij)}$)으로 이상칸 위치의 일부(예를 들어 행과 열)에서만 식별이 가능하다. 이 모형은 축소 가능한 모형이므로 이차원 포화모형으로 축소시켜 교호작용항의 유의함을 살펴보면, 한 변수 독립모형의 경우에는 이상칸을 식별할 수 있겠다. 그리고 삼차원 조건부 독립모형(예를 들어 [AB][AC] 모형)에서는 로그선형모형의 교호작용항(예를 들어 $\hat{u}_{12(ij)}$와 $\hat{u}_{13(ik)}$이 유의한 경우에는 추정량들의 i, j 그리고 k에 대응하는 (i,j,k)칸이 이상칸임을 식별할 수 있으며, 부분연관모형([AB][AC][BC] 모형)에서도 로그선형모형의 교호작용항 $\hat{u}_{12(ij)}$, $\hat{u}_{13(ik)}$, $\hat{u}_{23(jk)}$들이 유의한 경우에 추정량들의 i, j 그리고 k에 대응하는 (i,j,k)칸이 이상칸임을 식별할 수 있음을 발견하였다.

이와 같은 연구를 통하여 하나의 이상칸이 포함된 분할표에서 로그선형모형의 주효과항 추정량으로는 이상칸을 식별할 수 없으며 교호작용항으로 이상칸을 식별할 수 있다는 탐색적인 결과를 얻었으며 이로부터 대안적인 이상칸의 식별 방법으로 자료에 적합한 로그선형모형의 모수들 중 교호작용항의 추정량이 유의한 경우에는 그 추정량에 대응하는 칸을 이상칸으로 판단하는 방법을 본 논문에서 제안한다.

두 개 이상의 다중 이상칸이 존재하는 이차원 또는 삼차원 범주형 자료의 경우에는 여기에서 제안한 방법으로는 식별하는데 한계가 있다. 따라서 다중 이상칸이 존재하는 범주형 자료에서는 교호작용항이 포함된 로그선형모형의 추정량으로 이상칸을 식별하는 방법의 사용을 제한하여야 하는 문제점이 있다. 그리고 본 논문에서 연구한 로그선형모형의 교호작용항 추정량을 이용한 이상칸 식별방법은 사차원 이상의 다차원 분할표 자료에서 계속 연구되어야 하며 이 과정에서 이차 이상의 고차 교호작용항의 영향에 관한 연구를 향후 연구과제로 남겨둔다.

참고 문헌

[1] Barnet, V. and Lewis, T. (1994). *Outliers in Statistical Data*, 3rd, John Wiley & Sons.

[2] Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1991). *Discrete Multivariate Analysis: Theory and Practice*, The MIT Press.

[3] Brown, M. B. (1974). Identification of the Sources of Significance in Two-Way Contingency Tables, *Journal of the American Statistical Association*, Vol. 61, 65-975.

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

[5] Fuchs, C. and Kenett, 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.

[6] 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. 29, 205-220.

[7] Haberman, S. J. (1973). The analysis of residuals in cross-classification tables, *Biometrics*, Vol. 29, 205-220.

[8] Lee, Jong Cheol and Hong, Chong Sun (2000). Identification of Multiple Outlying Cells in Multy-way Tables, *The Korean Communications in Statistics*, Vol 7, No. 3, 687-698.

[9] Lee, Jong Cheol and Hong, Chong Sun (2001). An Identification of Outlying Cells in Contingency Tables via Correspondence Analysis Map, *The Korean Communications in Statistics*, Vol 8, No. 1, 39-49.

[10] Kotze, T. J. W. and Hawkins, D. M. (1984). The identification of outliers in two-way contingency tables using 2×2 subtables, *Applied Statistics*, Vol. 33, 215-223.

[11] Mosteller, F. and Parunak, A. (1985). Identifying extreme cells in a sizable contingency table : probabilistic and exploratory approaches, *In Exploring Data Tables, Trends and Shapes*, John Wiley & Sons, pp 189-224.

[12] Simonoff, J. S. (1988). Detecting Outlying Cells in Two-Way Contingency Tables Via Backwards-Stepping, *Technometrics*, Vol. 30, 339-345.

[13] Upton, G. and Guillen, M. (1995). Perfect Cells, Direct Models and Contingency Table Outliers, *Communications in Statistics, Part A - Theory and Methods*, Vol. 24, 1843-1862.

[2003년 1월 접수, 2003년 5월 채택]