The Korean Communications
in Statistics Vol. 6, No. 2, 1999
pp. 357-366

A Simulation Approach for Testing Non-hierarchical Log-linear Models

Hyun Jip Choi[1], Chong Sun Hong[2], Woo Rhee Lee[3]

Abstract

Let us assume that two different log-linear models are selected by various model selection methods. When these are non-hierarchical, it is not easy to choose one of these models. In this paper, the well-known Cox's statistic is applied to compare these non-hierarchical log-linear models. Since it is impossible to obtain the analytic solution about the problem, we proposed a alternative method by extending Pesaran and Pesaran's (1993) simulation approach. We find that the values of the proposed test statistic and the estimates are very much stable with some empirical results.

1. 서 론

여러 범주형 변수들에 의해 구성된 다차원 분할표의 자료구조를 식별하기 위한 로그선형모형은 변수들의 연관을 식별하는데 매우 유용한 도구이다. 그러나 분할표의 차원이 증가할수록 모형의 수가 급격히 증가하기 때문에 주어진 분할표에 가장 잘 적합되는 모형을 식별하기는 쉽지 않다. 이러한 문제를 해결하기 위해 여러 가지 모형선택방법이 제안되고 있으나, 사용된 모형선택방법에 따라 최종적으로 선택되는 모형이 서로 다르다는 문제가 발생할 수 있다. 이러한 경우에 서로 다른 모형선택방법에 의해 최종 선택된 모형들 중에서 분할표에 가장 잘 적합된다고 생각되는 모형을 선택하는 것은 쉽지 않다. 또한 서로 다른 모형선택방법에 의해 선택된 최종모형들이 비계층 로그선형모형(non-hierarchical log-linear model)일 경우에 이들 두 모형을 비교하는 것은 더욱 어렵다.

비계층 모형에 대한 검정문제는 특히 계량경제학 분야에서 활발히 연구되고 있다. 이들 연구는 대체로 주어진 특정 모형의 비교를 위해 Cox(1960, 1961)가 제안한 비계층 가설에 대한 검정 방법을 직접 적용하고 있다. 그러나 Cox의 검정통계량은 대수적으로 검정통계량을 산출하기 어렵기 때문에 Pesaran(1974), Pesaran과 Deaton(1978), 그리고 Davison과 MacKinnon(1981)등의 연구와 같이 선형회귀모형 혹은 단순한 비선형 회귀모형등과 같이 대수적으로 검정통계량을 얻을 수 있는 일부 제한된 영역에서만 적용되고 있다. 그러나 대수적인 해가 존재하지 않는 문제의 한가지 해결방법으로 Pesaran과 Pesaran (1993)이 제시한 모의실험을 이용한 검정통계량 산출방법을 고려할 수 있다.

1) Full-time Lecturer, Department of Applied Information Statistics, Kyonggi University, Suwon, 442-760, Korea.
2) Professor, Department of Statistics, Sung Kyun Kwan University, Seoul, 110-745, Korea.
3) Professor, Department of Applied Information Statistics, Kyonggi University, Suwon, 442-760, Korea.

본 논문에서는 비계층 로그선형모형을 비교하기 위한 검정 문제에 관하여 연구하였다. 제2절에서는 Cox 연구결과를 비계층 로그선형모형에 적용하여, 로그선형모형에서는 Cox의 검정통계량의 대수적 해를 쉽게 얻을 수 없음을 토론하였다. 그리고 제3절에서는 Pesaran과 Pesaran(1993)이 제안한 모의실험을 통해 비계층 로그선형모형을 비교하기 위한 검정방법을 제안하고, 분할표 자료에 대한 모의실험을 통한 검정이 갖는 특징에 관하여 토론하였다. 마지막으로 제4절에서는 제3절에서 고려된 모의실험을 통한 비교검정을 실제 자료를 통해 실시하고, 그 결과를 분석하였다. 제안된 모의실험에 의한 검정으로 경험적인 결과가 나타나는데 상당히 안정적인 검정결과가 발생함을 파악할 수 있었다.

2. 비계층 로그선형모형의 검정

일반적인 다차원 분할표를 관찰칸값(observed cell count)에 의해 $\{x_i\ ;\ i=1,2,\ldots,n\}$로 나타내고, 기대칸값(expected cell count)에 의한 분할표는 $\{m_i\ ;\ i=1,2,\ldots,n\}$로 나타내기로 한다. 여기서 n은 분할표의 총칸수를 나타내고 총표본수는 $\sum_{i=1}^{n} x_i = N$이다. 포아송 추출모형을 가정하면, x_i들은 서로 독립이므로 다음과 같은 결합분포함수를 얻을 수 있다.

$$f(\underline{x};\underline{m}) = \prod_{i=1}^{n} e^{-m_i} m_i^{x_i} / x_i!\,, \tag{1}$$

여기서 $\underline{x}$와 $\underline{m}$은 각각 관찰칸값과 기대칸값으로 구성된 ($n\times 1$)차 벡터를 나타낸다. 이로부터 일반적인 로그선형모형은 다음과 같이 나타낼 수 있다.

$$\log \underline{m} = \boldsymbol{D}\,\underline{u}, \tag{2}$$

여기서 $\underline{u}$는 추정의 대상이 되는 ($p\times 1$)차 모수벡터 그리고 $\boldsymbol{D}$는 적절한 제약조건에 의한 ($n\times p$)차 계획행렬(design matrix)이다. 모형 (2)에 의해 결합분포함수 (1)은 $\underline{m} = \exp(\boldsymbol{D}\underline{u})$ 이므로 $f(\underline{x};\underline{u})$과 같이 표현할 수 있다. 이제 다음과 같은 비계층 가설을 고려해보기로 하자.

$$\begin{aligned} H_f &: \log \underline{m}_f = \boldsymbol{D}_f \underline{u}_f\ , \\ H_g &: \log \underline{m}_g = \boldsymbol{D}_g \underline{u}_g\ , \end{aligned} \tag{3}$$

여기에서 $\boldsymbol{D}_f$와 $\boldsymbol{D}_g$는 각각 모수벡터 $\underline{u}_f$와 $\underline{u}_g$에 부여된 제약조건에 의한 계획행렬이며, $\underline{m}_f$와 $\underline{m}_g$는 각 가설에서의 기대값벡터이다. 또한 모수벡터들은 각각 $\underline{u}_f \in \Omega_f$, $\underline{u}_g \in \Omega_g$로 $\Omega_f \not\subset \Omega_g$과 같은 관계를 가지며, 가설 H_f에서의 결합분포함수를 $f(\underline{x};\underline{u}_f)$ 그리고 가설 H_g에서의 결합분포함수를 $g(\underline{x};\underline{u}_g)$와 같이 나타내기로 한다.

가설 (3)에서 고려된 두 비계층 로그선형모형을 검정하기 위하여 Neyman-Pearson의 우도비 검정통계량을 확장한 통계량을 고려할 수 있다(Cox 1961, 1962).

$$T_f(\hat{\underline{u}}_f, \hat{\underline{u}}_g) = \sum_{i=1}^{n} \hat{l}_{fg,i} - \hat{E}_f\{\sum_{i=1}^{n} \hat{l}_{fg,i}\}, \tag{4}$$

여기서 $\hat{\underline{u}}_f$와 $\hat{\underline{u}}_g$는 각각 가설 H_f와 H_g에서의 최우추정량이며, $\hat{m}_{f,i}$와 $\hat{m}_{g,i}$는 각각 $\hat{\underline{u}}_f$와 $\hat{\underline{u}}_g$에 의한 i번째 기대값의 최우추정량을 나타낸다. 그리고

$$\hat{l}_{fg,i} = x_i \log(\hat{m}_{f,i} / \hat{m}_{g,i})$$

는 각 가설에서의 로그우도함수의 커널(kernel)에 의한 i번째 칸의 로그우도비를 나타낸다.

$\hat{E}_f\{\sum_{i=1}^{n} \hat{l}_{fg,i}\}$는 가설 H_f하에서의 기대 로그우도비의 일치추정량(consistent estimator)을 나타낸다. 만일 가설 (3)의 두 모형이 계층구조를 형성하면 $\hat{E}_f\{\sum_{i=1}^{n} \hat{l}_{fg,i}\} = 0$으로 통계량 (4)는 일반적인 로그우도비 통계량과 같다. 그러나 비계층 구조에서 $\hat{E}_f\{\sum_{i=1}^{n} \hat{l}_{fg,i}\}$는 '0'이 아니며, 따라서 Cox의 검정통계량 (4)는 기대 로그우도비의 추정에 초점이 맞추어져 있다. 우선 가설 H_f에서 $\hat{\underline{u}}_g$의 극한값이 존재한다고 하고(Cox 1961) 이때의 극한값을 $\underline{u}_*$라고 한다면, Pesaran(1987)에 의해 가설 H_f하에서의 기대 로그우도비는 다음과 같이 유도된다.

$$\begin{aligned} E_f\{\sum_{i=1}^{n} l_{fg,i}\} &= \sum_{i=1}^{n} \sum_{x_i=0}^{\infty} l_{f*,i} f(x_i; \underline{u}_f) \\ &\equiv C(\underline{u}_f, \underline{u}_*), \end{aligned} \tag{5}$$

여기서 $l_{f*,i} = x_i \log(m_{f,i}/m_{*,i})$이며, $m_{*,i}$는 $\underline{u}_*$에 의한 i번째 기대값을 나타낸다. 한 가지 주목할 점은 기대 로그우도비 (5)는 Kullback-Leibler(1951)의 정보(information divergence)를 직접 활용한 형태를 취하며 이를 근접측도(measure of closeness), $C(\underline{u}_f, \underline{u}_*)$, 라고 한다.

로그선형모형을 위한 통계량 (4)의 근사분산은 분할표의 각 칸이 서로 동일한 분포를 갖지 않으므로 Cox(1962)에 의해 다음과 같이 얻을 수 있다.

$$\nu_f^2(\underline{u}_f, \underline{u}_*) = \sum_{i=1}^{n} V_f(l_{f*,i}) - \psi'(\underline{u}_f, \underline{u}_*) F(\underline{u}_f)^{-1} \psi(\underline{u}_f, \underline{u}_*), \tag{6}$$

여기서 $V_f(\cdot)$은 가설 H_f에서의 분산을 나타내며,

$$\begin{aligned} \psi(\underline{u}_f, \underline{u}_*) &= \sum_{i=1}^{n} E_f\left[l_{f*,i} \left\{ \frac{\partial \log f(x_i; \underline{u}_f)}{\partial \underline{u}_f} \right\} \right], \\ F(\underline{u}_f) &= -\sum_{i=1}^{n} E_f\left\{ \frac{\partial^2 \log f(x_i; \underline{u}_f)}{\partial \underline{u}_f \partial \underline{u}_f'} \right\} \\ &= \sum_{i=1}^{n} E_f\left[\left\{ \frac{\partial \log f(x_i; \underline{u}_f)}{\partial \underline{u}_f} \right\} \left\{ \frac{\partial \log f(x_i; \underline{u}_f)}{\partial \underline{u}_f'} \right\} \right]. \end{aligned}$$

그리고 모형 (2)에 의해 $\underline{d}_i'$, $i=1,2,\ldots,n$, 을 계획행렬 D의 i번째 행벡터라고 하면,

$$\frac{\partial \log f(x_i;\, \underline{u})}{\partial \underline{u}} = (x_i - m_i)\underline{d}_i',$$
$$\frac{\partial^2 \log f(x_i;\, \underline{u})}{\partial \underline{u}\, \partial \underline{u}'} = -m_i\, \underline{d}_i\, \underline{d}_i'$$

과 같다.

이와 같은 식들에 의해 근접측도 (5)의 일치추정량인 $C(\hat{\underline{u}}_f, \hat{\underline{u}}_*)$으로 추정된 $T_f(\hat{\underline{u}}_f, \hat{\underline{u}}_*)$와 근사분산 (6)의 일치추정량 $\hat{\nu}_f^2(\hat{\underline{u}}_f, \hat{\underline{u}}_*)$을 사용하여, 다음과 같은 표준화된 Cox 통계량은 적절한 가정에 의해 근사적으로 표준정규분포를 따른다.

$$N_f(\hat{\underline{u}}_f, \hat{\underline{u}}_*) = \frac{\sqrt{n}\, T_f(\hat{\underline{u}}_f, \hat{\underline{u}}_*)}{\hat{\nu}_f(\hat{\underline{u}}_f, \hat{\underline{u}}_*)} . \tag{7}$$

그러므로 가설 (3)은 검정통계량 (7)에 의해 검정이 수행되고, 이를 위해서는 $C(\hat{\underline{u}}_f, \hat{\underline{u}}_*)$와 $\hat{\nu}_f^2(\hat{\underline{u}}_f, \hat{\underline{u}}_*)$를 얻어야만 한다. 또한 근접측도와 근사분산의 추정을 위해서는 먼저 가설 H_f에서의 $\hat{\underline{u}}_g$의 추정량인 $\hat{\underline{u}}_*$를 구하는 것이 선행되어야만 한다.

로그선형모형에서 $\underline{u}_*$의 추정에는 크게 두 가지 문제를 갖는다. 첫째로, 로그선형모형에서는 모형 (2)의 모수벡터 $\underline{u}$는 직접해(direct estimates)가 존재하는 경우를 제외하고는 Deming과 Stephan(1940)에 의한 IPF(iterative proportional fitting) 방법 또는 Grizzle, Stamer와 Koch(1969)에 의한 반복 RWLS(reweighted least squares) 방법에 의해 추정된다. 따라서 직접해가 존재하는 경우에 $\underline{u}_*$의 추정량을 유도하기 쉽지 않을 뿐만 아니라, 직접해가 존재하지 않는 경우에는 $\underline{u}_*$의 추정량을 유도할 수 없다. 둘째로, $\underline{u}_*$의 추정식이 유도된다고 할지라도 $\hat{\underline{u}}_*$를 얻기 위해서는 가설 H_f하에서 추정된 기대값을 이용하여 계산하여야 한다. 그러나 이러한 유사최우추정량(pseudo MLE)이 로그선형모형에서 $\underline{u}_*$의 일치추정량이 된다는 연구를 발견할 수 없었다. 그러므로 검정통계량 (7)에 의해 비계층 가설을 검정하기 위해서는 먼저 $\underline{u}_*$의 추정량을 구하는 문제가 해결되어야 한다.

3. 모의실험에 의한 비계층 로그선형모형의 검정

앞에서 언급한대로 $\underline{u}_*$의 추정량을 직접 유도하기 쉽지 않기 때문에 모의실험을 통해 비계층 로그선형모형을 검정하기로 한다. 즉, $\underline{u}_*$의 최우추정량을 가설 H_f하에서 최우추정량 $\hat{\underline{u}}_f$에 의한 분포 $f(\underline{x};\, \hat{\underline{u}}_f)$로부터 생성된 자료에서 추정하고자 한다.

$\underline{x}_j$, $j=1,2,\ldots,R$, 를 가설 H_f에서의 $\hat{\underline{u}}_f$에 의해 생성된 관찰값벡터라고 하자. $\underline{x}_j$는 Emrich와 Piedmonte(1991) 그리고 Lee(1993)등이 제안한 방법에 의해 얻을 수 있으나, 주어진 자료에서 $\hat{\underline{u}}_f$가 추정되어 있으므로 Gange(1995)의 방법을 이용하여 어렵지 않게 구할수 있다. 구한 $\underline{x}_j$에 대하여 가설 H_g하에서의 $\underline{u}_g$의 최우추정량을 $\hat{\underline{u}}_g^j$이라 하면 $\underline{u}_*$의 가능한 추정량으로

다음을 고려할 수 있다.

$$\hat{\underline{u}}_*(R) = \frac{1}{R}\sum_{j=1}^{R}\hat{\underline{u}}_g^j . \tag{8}$$

추정량 (8)은 $\hat{\underline{u}}_f$에 의해 생성된 $\underline{x}_j$의 함수이므로 $\hat{\underline{u}}_f$의 함수가 된다. 또한 $\underline{x}_j$들은 서로 독립적으로 얻어지므로 $\hat{\underline{u}}_g^j$들 역시 서로 독립이다. 따라서 대수의 약법칙(weak law of large numbers)에 의해 R이 충분히 크면, $\hat{\underline{u}}_*(R) \xrightarrow{P} \underline{u}_*$이라는 사실을 어렵지 않게 알 수 있다 (Pesaran과 Pesaran 1993).

이렇게 얻은 모의실험 추정량 $\hat{\underline{u}}_*$을 이용하여 근접측도 (5)의 한가지 가능한 추정량으로 다음과 같은 모의실험 추정량을 고려할 수 있다.

$$C_R(\hat{\underline{u}}_f, \hat{\underline{u}}_*(R)) = \frac{1}{R}\sum_{j=1}^{R}\{\bar{l}_f(\underline{x}_j; \hat{\underline{u}}_f) - \bar{l}_g(\underline{x}_j; \hat{\underline{u}}_*(R))\} , \tag{9}$$

여기서 $\bar{l}_f = n^{-1}\sum_{i=1}^{n}\log f(x_i; \hat{\underline{u}}_f)$ 그리고 $\bar{l}_g = n^{-1}\sum_{i=1}^{n}\log f(x_i; \hat{\underline{u}}_g)$은 각각 최우추정량 $\hat{\underline{u}}_f$와 $\hat{\underline{u}}_*(R)$에 의한 로그우도의 평균을 나타낸다. $\bar{l}_f(\underline{x}_j; \hat{\underline{u}}_f) - \bar{l}_g(\underline{x}_j; \hat{\underline{u}}_*(R))$, $j=1,2,\ldots,R$, 은 서로 독립적으로 얻어지므로 R이 충분히 크면 $C_R(\hat{\underline{u}}_f, \hat{\underline{u}}_*(R))$은 $C(\hat{\underline{u}}_f, \hat{\underline{u}}_*)$에 수렴한다.

모의실험 추정량 (8)과 (9)는 Pesaran과 Pesaran(1993)이 제안한 모의실험 추정량을 로그선형모형으로 확장한 결과이다. 그들은 로지스틱모형(logistic model)과 프로빗모형(probit model)을 비교하는 문제의 경험적 분석에서 (8)과 (9)에 의한 추정이 비교적 적은 반복에서도 안정적이고 효율적인 추정이라고 지적하고 있다. 또한 그들은 비계층 가설을 위해 많은 논문들에서 활용하고 있는 다음과 같은 근사분산의 추정량이 White(1982)가 고려한 가정에 의해 (6)의 일치추정량이라는 것을 제안하였다.

$$\hat{\nu}_f^2(\hat{\underline{u}}_f, \hat{\underline{u}}_g) = n^{-1}\sum_{i=1}^{n}(\hat{l}_{fg,i} - \bar{l}_{fg})^2 - \psi_n'(\hat{\underline{u}}_f, \hat{\underline{u}}_g)A_n(\hat{\underline{u}}_f)^{-1}\psi_n(\hat{\underline{u}}_f, \hat{\underline{u}}_g) , \tag{10}$$

여기서

$$\begin{aligned}
\bar{l}_{fg} &= n^{-1}\sum_{i=1}^{n}\hat{l}_{fg,i} , \\
\psi_n(\underline{u}_f, \underline{u}_g) &= n^{-1}\sum_{i=1}^{n} l_{fg,i}\{\partial\log f(x_i; \underline{u}_f)/\partial\underline{u}_f\} , \\
A_n(\underline{u}_f) &= \left\{-n^{-1}\sum_{i=1}^{n}\{\partial^2\log f(\underline{x}; \underline{u}_f)/\partial\underline{u}_f\partial\underline{u}_f'\}\right\} .
\end{aligned}$$

이들로부터 모의실험에 의한 검정은 통계량 (4)와 유사한 다음과 같은 통계량에 의해 검정을 수행한다.

$$T_f(\hat{\underline{u}}_f, \hat{\underline{u}}_*(R)) = \bar{l}_{fg} - C_R(\hat{\underline{u}}_f, \hat{\underline{u}}_*(R)) .$$

또한 위의 통계량을 근사분산 (10)에 의해 표준화한 다음과 같은 검정통계량은 가설 H_f에서 n이 충분히 크면 근사적으로 표준정규분포를 따른다.

$$N_f(\hat{\underline{u}}_f, \hat{\underline{u}}_*(R)) = \frac{\sqrt{n}\, T_f(\hat{\underline{u}}_f, \hat{\underline{u}}_*(R))}{\hat{\nu}_f(\hat{\underline{u}}_f, \hat{\underline{u}}_g)} \quad . \tag{11}$$

결국 로그선형모형에서 (8)과 (9)는 어렵지 않게 얻을 수 있으므로 가설 (3)은 모의실험에 의한 통계량 (11)에 의해 검정할 수 있다.

4. 실증 분석

이제까지 고려된 모의실험에 의한 비계층 로그선형모형의 검정문제를 Ries와 Smith (1963)에서 발췌한 잘 알려진 세제 상표선호도 자료를 통해 분석해보기로 한다. 자료는 3×2×2×2 분할표이며, 분할표를 구성하는 네 변수는 다음과 같다.

변수 1 : 사용하는 물의 부드러움 정도 (soft, medium, hard)
변수 2 : 상표 M의 이전 사용여부 (yes, no)
변수 3 : 물의 온도 (high, low)
변수 4 : 선호하는 상표 (X, M)

이 자료에 가장 잘 적합하는 모형의 선택을 위해 Bishop, Fienberg와 Holland(1975) 그리고 홍종선과 최현집(1999) 등에서 소개하고 있는 여러 가지 모형선택방법을 고려할 수 있다. 특히 이들이 고려한 계층구조에 의하면 다음 모형을 자료에 가장 잘 적합되는 모형으로 선택할 수 있다.

$$\begin{aligned}\log m_{ijkl} = \; & u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{4(l)} \\ & + u_{13(ik)} + u_{24(jl)} + u_{34(kl)} \; .\end{aligned} \tag{12}$$

위 모형은 $G^2 = 11.89$ ($df = 14$, $p-$값 $= 0.6154$)로 자료에 잘 적합된다. 그리고 자료에 잘 적합되는 또 다른 모형으로 다음과 같은 모형을 선택할 수도 있다.

$$\begin{aligned}\log m_{ijkl} = \; & u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{4(l)} \\ & + u_{12(ij)} + u_{24(jl)} + u_{34(kl)} \; .\end{aligned} \tag{13}$$

위 모형의 적합도 검정통계량 값은 $G^2 = 16.91$ ($df = 14$, $p-$값 $= 0.261$)이다.

모형 (12)와 (13)은 최고차항의 첨자를 이용하여 각각 [13][24][34]와 [12][24][34]와 같이 표현하기로 한다. 이들 두 모형은 모두 직접해가 존재하지 않는다는 사실에 주목하기 바란다. 모형

(12)와 (13)은 계층구조를 형성하지 않으므로 다음과 같은 가설을 모의실험에 의해 검정하기로 한다.

$$\begin{aligned} H_f &: [12][24][34] \\ H_g &: [13][24][34] \end{aligned} \tag{14}$$

위 가설 (14)의 평가를 위해 두 모형에 대한 적합도 검정결과에 의존해보기로 한다. 두 모형은 모두 자료에 잘 적합되고, 모형 [13][24][34]의 검정통계량 값이 모형 [12][24][34]에 비해 적다. 검정통계량 값이 작다는 것은 자료에 잘 적합되는 것을 의미하므로 각 모형의 검정통계량 값에 의존하면 모형 [13][24][34]가 더 잘 적합되는 모형으로 판단할 수 있다. 그러나 이들 두 모형의 검정통계량의 차이는 5.02로 비교적 작다. 또한 두 모형은 서로 다른 계층구조에 속하며, 자유도가 같기 때문에 이 값에 대한 통계적 유의성을 판단하기는 어렵다. 즉, 각 모형의 검정통계량 차이에 의해 모형 [13][24][34]가 더 잘 적합된다는 것은 통계적으로 유의한 결론이 아니다.

두 모형의 적합성을 판단하기 위해 제3절에서 고려한 모의실험에 의한 검정을 수행해보기로 한다. 이를 위해서는 먼저 $\underline{u}_*$에 대한 추정이 이루어져야 하며, <표 1>에 여러 반복수에 의한 $\hat{\underline{u}}_*$에 대한 추정결과가 정리되어 있다. <표 1>에서 모형 (12)와 (13)에 대한 최우추정량은 첫 번째 열과 두 번째 열에 제시되어 있으며, 세 번째 열부터 제시되어 있는 추정값은 각 반복수에 해당하는 모의실험 추정량 $\hat{\underline{u}}_*(R)$이다. 그리고 제일 마지막 행은 각 추정값에 의한 최대 로그우도비를 나타내고 있다. <표 1>에서 우선 최대 로그우도가 500회 반복 이후에 상당히 안정된 형태를 취하는 것에 주목하기 바란다. 그리고 각 반복에 따른 추정값들의 변동폭도 크지 않다는 것을 알 수 있다.

가설 H_f에서의 모형 [12][24][34]은 변수 1과 변수 3이 서로 독립인 것을 의미한다. 이때 모형 [12][24][34]에 대한 충분주변합(sufficient configuration)은 각각 모형표현에 포함된 [12], [24] 그리고 [34]에 의한 주변표이므로 가설 H_f에서의 $u_{13(ik)}$에 대한 추정값은'0'에 가까운 값을 가질 것으로 예상할 수 있다. <표 1>의 추정결과에서 $u_{13(ik)}$에 대한 추정값은 '0'에 가까운 상당히 작은 값으로 추정되어 이러한 사실을 뒷받침 해주고 있다.

<표 1>에서 얻은 $\hat{\underline{u}}_*(R)$에 의해 가설 (3)을 검정하기 위한 모의실험에 의한 검정 결과를 <표 2>와 같이 얻을 수 있다. <표 2>에서 C는 해당되는 반복수에 의한 근접측도의 모의실험 추정량 (9)의 추정값이며, 그리고 N_f는 검정통계량 (11)을 그리고 괄호 안의 값은 해당 통계량에 대한 $p-$값이다.

검정통계량 (11)에서 근사분산 (10)은 반복수에 영향을 받지 않으며, 주어진 자료에서 $\hat{\nu}_f^2(\hat{\underline{u}}_f, \hat{\underline{u}}_g) = 0.05989$로 얻어졌다. 이 추정값과 근접측도의 추정값에 의해 가설 (14)를 검정하기 위한 <표 2>에 정리된 검정통계량 값을 살펴보기로 하자. 어떤 반복수에 대해서도 해당되는 $p-$값이 상당히 적은 것을 볼 수 있다. 또한 주어진 행에서 각 반복에 따른 근접측도의 변화의 폭이 그리 크지 않다. 결국 가설 (14)에 대한 검정통계량 (11)에 의한 검정결과인 <표 2>를 통해 가설 H_g에서의 모형 [13][24][34]가 자료에 더 잘 적합되는 모형인 것으로 판단할 수 있다.

<표 1> 가설 H_f와 H_g에서의 최우추정량과 모의실험에 의한 최우추정량

	$\hat{u}_f$ (std)	$\hat{u}_g$ (std)	$\hat{u}_*(R)$ 반복수(R) 100	200	500	1000
$\hat{u}$	3.68604 (0.0331)	3.68270 (0.0333)	3.68285 (0.0096)	3.68222 (0.0104)	3.68241 (0.0111)	3.68288 (0.0099)
$\hat{u}_{1(1)}$	-0.02999 (0.0449)	-0.05907 (0.0475)	-0.03569 (0.0447)	-0.02719 (0.0468)	-0.02966 (0.0448)	-0.02648 (0.0468)
$\hat{u}_{1(2)}$	0.02528 (0.0443)	0.02779 (0.0461)	0.02721 (0.0425)	0.02328 (0.0433)	0.02459 (0.0457)	0.02293 (0.0456)
$\hat{u}_{2(1)}$	-0.04387 (0.0319)	-0.04343 (0.0319)	-0.04386 (0.0278)	-0.04057 (0.0321)	-0.04222 (0.0307)	-0.04327 (0.0291)
$\hat{u}_{3(1)}$	-0.27526 (0.0328)	-0.27826 (0.0329)	-0.27529 (0.0307)	-0.27765 (0.0336)	-0.27529 (0.0322)	-0.27564 (0.0327)
$\hat{u}_{4(1)}$	-0.01659 (0.0331)	-0.01659 (0.0331)	-0.01936 (0.0298)	-0.01731 (0.0354)	-0.01634 (0.0328)	-0.01481 (0.0339)
$\hat{u}_{12(11)}$	-0.01116 (0.0449)	-	-	-	-	-
$\hat{u}_{12(21)}$	0.04412 (0.0443)	-	-	-	-	-
$\hat{u}_{13(11)}$	-	-0.10159 (0.0475)	0.00609 (0.0438)	0.00214 (0.0453)	-0.00190 (0.0444)	-0.00115 (0.0467)
$\hat{u}_{13(21)}$	-	0.00345 (0.0461)	-0.00248 (0.0494)	-0.00421 (0.0434)	0.00133 (0.0477)	-0.00024 (0.0442)
$\hat{u}_{24(11)}$	-0.14377 (0.0319)	-0.14377 (0.0319)	-0.14324 (0.0303)	-0.14340 (0.0295)	-0.14705 (0.0310)	-0.14295 (0.0304)
$\hat{u}_{34(11)}$	-0.06836 (0.0328)	-0.06836 (0.0328)	-0.07226 (0.0292)	-0.07059 (0.02969)	-0.06795 (0.0324)	-0.06895 (0.0307)
$\log f(x;\hat{u})$	-74.6628	-72.1508	-75.5341	-75.2258	-75.1387	-75.1367

<표 2> 모의실험을 통한 근접측도의 추정값과 Cox 검정통계량의 값

	반복수(R) 100		200		500		1000	
	C	N_f (p-값)	C	N_f (p-값)	C	N_f (p-값)	C	N_f (p-값)
$\hat{u}_*(R)$	0.0363	-2.8221 (0.0074)	0.0235	-2.5649 (0.0148)	0.0198	-2.4923 (0.0177)	0.0198	-2.4907 (0.0179)

참고로 가설 (14)에서 H_f와 H_g에서의 모형의 역할이 바뀐 다음과 같은 가설을 고려해보기로 한다.

$$H_f: [13][24][34]$$
$$H_g: [12][24][34]$$

위 가설을 위해 모수추정과 근접측도 추정을 위해 모두 $R=200$회 반복한 모의실험에 의한 검정결과 $N_f=-0.48002$ (p-값$=0.35553$)을 얻을 수 있었다. 따라서 H_g : [12][24][34]에 비해 모형 H_f : [13][24][34]를 선택하게 되고 이는 앞의 결과와 일치한다.

5. 결 론

분할표 분석을 위한 로그선형모형은 범주형 변수의 수가 증가할수록 고려되어야하는 모형의 수가 급격히 증가한다. 또한 계층구조만을 고려하여도 자료에 잘 적합되는 모형을 선택하기 위해 고려되어야 하는 모형의 수는 매우 많다. 따라서 여러 가지 모형선택방법들이 제안되어 있으나 이들은 모두 주어진 계층구조에서 모형을 선택하기 때문에 서로 다른 계층구조를 고려할 경우에 최종 선택된 모형은 다를 수 있다. 이때 만일 두 계층구조에서 선택된 모형이 계층구조를 형성할 경우에 두 모형의 적합성을 비교하는 것은 잘 알려진 우도비 검정통계량의 분할을 통해 평가할 수 있으나, 만일 자유도가 같을 경우에 두 모형은 비교할 수 없다. 또한 두 최종모형이 비계층 구조를 형성할 경우에 비계층 로그선형모형을 비교하기 위한 연구는 거의 알려져 있지 않다.

본 논문에서는 비계층 로그선형모형의 검정문제에 관하여 연구하였다. 이를 위하여 Cox(1960, 1961)의 통계량을 이용한 검정문제를 비계층 로그선형모형으로 확장하였으며, 로그선형모형에서는 Cox 통계량의 대수적인 해를 얻기 어렵거나 불가능하다는 것을 지적하였다. 이러한 문제를 해결하기 위해 Pesaran과 Pesaran(1993)이 제안한 모의실험을 통한 검정방법을 로그선형모형에 적용하고, 모의실험에 의한 검정결과를 평가하기 위하여 잘 알려진 세제 상표선호도 자료를 이용하여 분석하였다. 자료에 잘 적합된다고 판단되는 두개의 비계층 모형을 통해 분석한 결과, 비교적 적은 반복($R=500$)에서도 $\underline{u}_*$의 모의실험 추정량과 근접측도에 대한 추정값에 대한 변화폭이 크지 않은 것을 알 수 있었으며, 이는 주어진 자료의 총 관찰값 $N=1008$이라는 점을 감안하면 상당히 안정적인 결과라고 판단할 수 있다. 또한 분석에서 고려된 자료의 총칸수 $n=24$이므로, 모의실험에 의한 검정이 총칸수 n이 크지 않은 분할표에서도 본 논문에서 소개한 모의실험에 의한 검정을 충분히 적용할 수 있다는 사실을 발견하였다.

참 고 문 헌

[1] 홍종선, 최현집 (1999). 로그선형모형을 이용한 범주형 자료분석, 자유아카데미.

[2] Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis : Theory and Practice*, MIT Press.

[3] Cox, D. R. (1961). Tests of separate families of hypothesis, *Proceedings of the 4th Berkeley symposium*, 1, 105-123.

[4] Cox, D. R. (1962). Further results on tests of separate families of hypotheses, *Journal of Royal Statistical Society* B, 24, 406-424.

[5] Davison, R. and J. G. MacKinnon (1984). Model specification tests based on artificial linear regressions, *International Economic Review*, 25, 485-502.

[6] Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known, *Annals of Mathematical Statistics*, 11, 427-444.

[7] Emrich, L. J. and Piedmonte, M. R. (1991). A method for generating high- dimensional multivariate binary variate, *The American Statistician*, 45, 302-304.

[8] Gange, S. J. (1995). Generating multiplicate categorical variates using the iterative proportional fitting algorithm, *The American Statistician*, 49, 134-138.

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

[10], Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *Annals of Mathematical Statistics*, 22, 79-86.

[11] Lee, A. J. (1993). Generating random binary deviates having fixed marginal distributions and specified degrees of association, *The American Statistician*, 47, 209, 215.

[12] Pesaran, M. H. (1974). On general problem of model selection, *Review of Economic Studies*, 42, 153-171.

[13] Pesaran, M. H. (1987). Global and partial non-nested hypotheses and asymptotic local power, *Econometric Theory*, 3, 69-97.

[14] Pesaran, M. H. and Deaton, A. S. (1978). Testing non-nested nonlinear regression models, *Econometrica*, 46, 677-694.

[15] Pesaran, M. H. and Pesaran, B. (1993). A simulation approach to the problem of computing Cox's statistic for testing nonnested models, *Journal of Econometrics*, 57, 377-392.

[16] Ries, P. N. and Smith, H. (1963). The Use of Chi-square for Preference Testing in Multidimensional Problems, *Chemical Engineering Progress*, 59, 39-43.