Journal of the Korean Data & Information Science Society
2017, **28**(5), 1021–1026
http://dx.doi.org/10.7465/jkdi.2017.28.5.1021
한국데이터정보과학회지

계수적 반응을 갖는 종양 억제 혼합물 실험에서 모형 비교[†]

김정일[1]

[1]강원대학교 정보통계학과

접수 2017년 8월 21일, 수정 2017년 9월 14일, 게재확정 2017년 9월 19일

요 약

화학, 제약, 식품 등 여러 분야에서 활용되는 혼합물 실험은 반응변수가 설명변수들의 절대량이 아닌 상대적인 혼합비율에 의해 영향을 받고 구조상 공선성이 존재하게 되는 성질이 있으며 양적인 반응변수들에 대한 실험이 많아 대부분 정규분포를 가정하고 선형모형을 적용하여 분석하고 있다. 이 논문에서는 반응변수가 계수형인 혼합물 실험의 사례로 Chen 등(1996)에 소개된 종양 억제 효과에 대한 실험에 나타난 지방, 탄수화물, 섬유질과 같은 식이요법 관련 혼합물 성분들과 종양 발현 여부인 계수형 반응변수를 갖는 자료를 대상으로 셰페의 2차 다항모형과 성분들간의 비선형적 관계를 보완하기 위해 대안으로 제시된 베커의 수정 모형들, 그리고 공선성을 완화하기 위해 제시된 Akay와 Tez (2011)의 성분비 변환 모형을 설명변수들의 선형결합으로 활용하여 설정한 로지스틱회귀모형들을 분류 정확도 기준을 적용하여 비교하고 결과를 설명하였다.

주요용어: 계수형 반응변수, 로지스틱회귀모형, 분류 정확도, 셰페의 2차 다항모형, 혼합물 실험.

1. 서론

화학, 제약, 식품 등 여러 분야에서 활용되는 혼합물 실험 (mixture experiments)은 반응변수가 혼합물의 성분 (components, ingredients)인 설명변수들의 절대량이 아닌 상대적인 혼합비율에 의해 영향을 받는 성질이 있으며 구조상 공선성이 존재한다. 혼합물 실험에서는 휘발유의 성분비에 의해 영향받는 옥탄가와 같이 양적인 반응변수들이 많으며 대부분 정규분포를 가정하고 선형모형을 적용하여 분석하고 있으나 종양 발현 여부와 같은 계수형 반응변수의 경우는 이항분포나 포아송분포 등 이산형분포를 가정하여 일반화선형모형 (generalized linear models)을 적용하여 분석할 수 있다. Chen 등 (1996)은 종양 억제 효과에 대한 혼합물 실험에서 Scheffe의 정준 2차다항모형과 설명변수들간의 비선형 관계를 보완하기 위해 대안으로 제시된 Becker (1968)의 수정모형들을 설명변수들의 선형결합 (linear predictor)으로 적용하여 종양 발현 여부인 반응변수가 이항분포를 한다고 가정하고 로지스틱회귀모형을 설정하여 분석하였고, 같은 실험에 대해 Akay와 Tez (2011)는 공선성을 고려한 모형설정의 관점에서 설명변수에 대한 변환을 적용한 로지스틱회귀모형을 추가적으로 제시하였다.

혼합물 실험의 목적에는 반응변수와 혼합물 성분들간의 적절한 모형을 설정하고 관련성을 파악하여 활용하는 것이 있으며, 이 논문에서는 종양 억제 효과에 대한 혼합물 실험에서 Chen 등 (1996)과 Akay와 Tez (2011)가 제시한 로지스틱회귀모형들을 분류 정확도 (classification accuracy) 기준을 사용하여 비교하고 공선성을 고려하여 로지스틱능형회귀분석도 함께 검토한다. 혼합물 실험의 구조와 Scheffe의 정준 2차다항모형과 대안으로 제시된 Becker (1968)의 수정모형들에 대해 다음 2절에서 소개하고, 계수형 반응변수를 갖는 사례로 Chen 등 (1996)이 소개한 종양 억제 효과에 대한 혼합물 실험 자료를 3절에서 분석하고 결과를 설명한다.

[†] 2015년도 강원대학교 대학회계 학술연구조성비로 연구하였음(관리번호-520150417).
[1] (200-701) 강원도 춘천시 강원대학길1, 강원대학교 정보통계학과, 교수.
E-mail: jikim@kangwon.ac.kr

2. 혼합물 성분과 모형

혼합물 실험에서 혼합물 성분이 q개인 경우 i번째 성분의 비율을 x_i로 나타내면 q개의 성분비율은 다음과 같은 제약조건을 갖는다.

$$x_1 + x_2 + \ + x_q = 1, x_i \geq 0, i = 1, 2, ..., q. \tag{2.1}$$

이러한 혼합물 실험의 실험공간은 $(q-1)$차원 심플렉스(simplex) 공간이며, Cornell (2002)에서 제시된 바와 같이 각 혼합물 성분비율에 제약이 주어지면 실험공간은 심플렉스의 일부가 된다. 이 실험공간에서 심플렉스 중심배열법 등과 같은 실험설계에 의하여 실험한 자료에 적절한 모형을 설정하고 적합시켜 분석한다.

혼합물 실험에서 흔히 사용되는 Scheffe의 정준 2차 다항모형은 반응변수를 Y라 할 때, 제약조건 (2.1)을 고려하여 상수항과 제곱항을 포함하지 않는 다음과 같은 식으로 나타낸다.

$$E(Y) = \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q-1} \sum_{i<j}^{q} \beta_{ij} x_i x_j. \tag{2.2}$$

그러나 위 모형은 하나의 성분비율과 다른 성분비율들의 비선형적 관계를 나타내지 못한다는 점을 보완하여 Becker (1968)는 다음과 같은 3가지 형태의 모형을 제안하였다.

$$H1 : E(Y) = \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q-1} \sum_{i<j}^{q} \beta_{ij} min(x_i, x_j)$$

$$H2 : E(Y) = \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q-1} \sum_{i<j}^{q} \beta_{ij} \frac{x_i x_j}{(x_i + x_j)} \tag{2.3}$$

$$H3 : E(Y) = \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q-1} \sum_{i<j}^{q} \beta_{ij} (x_i x_j)^{1/2}$$

한편 Akay와 Tez (2011)는 공선성을 고려하여 다음과 같이 하나의 성분비율과 다른 성분비율들의 비를 새로운 설명변수로 변환하여 사용하는 방법을 대안으로 제시하였다.

$$z_i = \frac{x_i}{x_q}, i = 1, 2, ..., q-1. \tag{2.4}$$

혼합물 실험에서는 식(2.1)에 주어진 제약으로 인하여 공선성이 존재하게 되어 주어진 모형의 항들간에 상관이 있을 가능성이 많아 모형의 축소나 능형회귀분석을 고려하게 된다.

3. 사례

암컷 쥐에 대해서 세 가지 식이성 성분인 지방, 탄수화물, 섬유질 (fat, carbohydrate, fiber)이 유선 종양 (mammary gland tumors)의 발현에 주는 영향을 연구한 혼합물 실험에 관한 자료가 Chen 등(1996)에 소개되어 있으며 Table 3.1에 나타나 있다. 이 자료는 지방 (x_1), 탄수화물 (x_2), 섬유질

(x_3) 열량의 비율이 서로 다른 9가지 조건 각각에서 쥐 30마리 중 종양이 관측된 개체 수 (tumor)를 나타낸다. 실제 실험에서 혼합물 성분들의 비율은 다음과 같은 조건으로 제한되었다.

$$0.133 \le x_1 \le 0.701, 0.267 \le x_2 \le 0.863, 0.003 \le x_3 \le 0.050, x_1 + x_2 + x_3 = 1.$$

Table 3.1 The number of mammary gland tumors observed in 30 rats

Group	component proportions			
	Fat(x_1)	Carbohydrate(x_2)	Fiber(x_3)	Tumor
1	0.175	0.775	0.050	17
2	0.153	0.820	0.027	15
3	0.133	0.863	0.004	17
4	0.491	0.470	0.039	24
5	0.440	0.538	0.022	21
6	0.390	0.607	0.003	23
7	0.701	0.267	0.032	18
8	0.638	0.343	0.019	23
9	0.576	0.421	0.003	26

이 종양 자료에 대하여 Chen 등 (1996)은 (2.2), (2.3)에 주어진 형태의 Scheffe 2차 다항모형과 Becker (1968)의 H2 모형을 설명변수들의 선형결합으로 설정하고 로지스틱 회귀분석을 행하여 다음과 같은 모형을 가장 적합한 것으로 제시하였다.

$$logit(\hat{\pi}) = -1.240x_1 - 0.968x_2 - 6.435x_3 + 10.447x_1x_2. \tag{3.1}$$

$$logit(\hat{\pi}) = -1.220x_1 - 0.936x_2 - 8.215x_3 + 10.292\frac{x_1x_2}{x_1 + x_2}. \tag{3.2}$$

위 모형들에 대해 Akay와 Tez (2011)는 과산포(under/over-dispersion) 문제가 있음을 지적하고 (2.4)에서와 같이 혼합물 성분수들의 비로 정의된 설명변수에 대해 제곱근 변환을 하여 로지스틱 회귀분석을 행하고 아카이케 정보기준 (AIC) 통계량과 같은 변수선택 판정기준을 적용하여 다음 모형을 가장 적합한 것으로 제시하였다.

$$logit(\hat{\pi}) = 0.290 + 0.181\sqrt{\frac{x_1}{x_3}} - 0.077\sqrt{\frac{x_2}{x_3}}. \tag{3.3}$$

변수선택 판정기준은 절대적이라기보다 참고사항이라 할 수 있으므로 제시된 세 모형을 분류 정확도 기준으로 타당성 (validity)을 비교하고 공선성을 고려한 능형로지스틱회귀모형도 함께 비교하여 최종적인 모형을 선택하는 방법을 수행하고자 한다. 여기서는 (2.2)에 주어진 형태의 Scheffe 정준 2차 다항모형을 설명변수들의 선형결합으로 설정하고 능형모수 (biasing parameter) k가 0.005인 능형로지스틱회귀분석을 하였고 적합된 모형은 다음과 같다.

$$logit(\hat{\pi}) = -0.102x_1 - 0.875x_2 - 1.906x_3 + 8.152x_1x_2 - 23.831x_1x_3 + 2.497x_2x_3. \tag{3.4}$$

혼합물 실험에서 능형계수를 결정하는 문제에 대하여 Jang과 Yoon (1997), Jang과 Cook (2011)의 연구가 있으며 계수형 반응변수를 포함하는 경우에 대한 적용도 고려할 필요가 있겠으나 여기서는 논의하지 않는다.

분류를 위해 혼합물 성분의 서로 다른 9가지 조건에서 각 30마리 쥐에 대한 관측 수인 270개의 자료를 135개의 훈련용 (training) 자료와 나머지 135개는 시험용 (test) 자료로 랜덤하게 분할하였다. 훈련용 자료에 비교 대상 모형들을 적합시키고 시험용 자료에 적용하여 분류의 정확도를 계산한다. 분류의 분리점 (threshold, T_L)은 종양이 발현하지 않은 경우를 발현한 것으로 오분류할 비율이 10%가 넘지 않도록 지정하였고 500번 반복하여 분류 정확도의 평균을 계산하였으며 이 절차를 100회 수행한 결과의 요약이 Table 3.2에 주어져 있으며 비교 대상인 네 모형들 중 식 (3.2)에 주어진 형태인 하나의 성분비율과 다른 성분비율들의 비선형적 관계를 나타내도록 보완한 Becker의 $H2$ 모형이 분류 정확도 기준에서 상대적으로 타당한 것으로 나타났다.

Table 3.2 Logistic regression models with classification accuracy

Models	summary of classification accuracy					
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
(3.1)	0.5093	0.5153	0.5172	0.5171	0.5188	0.5257
(3.2)	0.6060	0.6093	0.6108	0.6110	0.6121	0.6188
(3.3)	0.5103	0.5137	0.5151	0.5151	0.5163	0.5212
(3.4)	0.5049	0.5114	0.5137	0.5139	0.5159	0.5218

4. 결론

혼합물 실험에서는 반응변수가 양적인 경우가 많고 혼합물의 성분인 설명변수들의 절대량이 아닌 상대적인 혼합비율에 의해 영향을 받는 특이한 관계를 나타내기 위하여 Scheffe의 다항모형이나 관련 모형들이 제시되어 있으나 반응변수가 계수형인 경우에 대한 연구는 많지 않으나 일반화선형모형을 적용하여 분석하는 방법이 시도되고 있다. 종양 발현여부에 대해 설명변수들의 선형결합을 Scheffe의 다항모형, Becker H2 모형, 혼합물 성분비 변환 모형들로 나타내고, 로짓 연결함수를 사용한 로지스틱회귀분석에 의해 제시된 여러 모형들을 분류 정확도 기준으로 비교하였고 사례로 주어진 자료에 대해서는 혼합물 성분비간의 비선형관계를 보완한 Becker H2 모형을 활용한 로지스틱회귀모형이 타당한 것으로 나타났으며 분류 정확도 기준을 다른 판정기준과 함께 참고하면 도움이 될 것으로 보인다. 분류 정확도 기준보다 Hong과 Wu (2014)와 Hong과 Yang (2015)에 언급된 좀 더 전반적인 판단기준인 ROC (receiver operating characteristic)곡선의 아래 면적 AUC에 의한 비교도 추가적인 연구 대상으로 생각된다.

변수선택에 의한 모형설정과 더불어 혼합물 실험의 구조에 내재하는 공선성 문제에 대해 능형회귀모형이나 Park과 Key (2013)에 소개된 벌점 로지스틱 회귀모형 등을 고려할 수 있으며 계수형 반응변수를 포함하는 혼합물 실험의 자료에 대해 능형모수를 평가하는 방법에 대한 연구도 필요하다고 판단된다.

References

Akay, K. U. and Tez, M (2011). Alternative modeling techniques for the quantal response data in mixture experiments. *Journal of Applied Statisitics*, **36**, 373-390.

Becker, N. G. (1968). Models for the response of mixture. *Journal of Royal Statistical Society B*, **30**, 349-358.

Chen, J. J., Li, L. A. and Jackson, C. D. (1996). Analysis of quantal response data from mixture experiments. *Environmetrics*, **7**, 503-512.

Cornell J. A. (2002). *Experiments with mixtures*, John Wiley & Sons, Inc., New York.

Hong, C. S. and Wu, Z. Q. (2014). Alternative accuracy for multiple ROC analysis. *Journal of the Korean Data & Information Science Society*, **25**, 1521-1530.

Hong, C. S. and Yang, D. S. (2015). ROC curve and AUC for linear growth models. *Journal of the Korean Data & Information Science Society*, **26**, 1367-1375.

Jang, D. H. and Anderson-Cook, C. M. (2011). Fraction of design space plots for evaluating ridge estimators in mixture experiments. *Quality and Reliability Engineering International* , **27**, 27-34.

Jang, D. H. and Yoon, M. (1997). Graphical methods for evaluating ridge regression estimator in mixture experiments. *Communications in Statistics-Simulations and Computation*, **26**, 1049-1061.

Park, C. and Kye, M. J. (2013). Penalized logistic regression models for determining the discharge of dyspnea patients. *Journal of the Korean Data & Information Science Society*, **24**, 125-133.

Journal of the Korean Data & Information Science Society
2017, **28**(5), 1021–1026

http://dx.doi.org/10.7465/jkdi.2017.28.5.1021
한국데이터정보과학회지

A comparison of models for the quantal response on tumor incidence data in mixture experiments[†]

Jung Il Kim[1]

Department of Information Statistics, Kangwon National University

Received 21 August 2017, revised 14 September 2017, accepted 19 September 2017

Abstract

Mixture experiments are commonly encountered in many fields including food, chemical and pharmaceutical industries. In mixture experiments, measured response depends on the proportions of the components present in the mixture and not on the amount of the mixture. Statistical analysis of the data from mixture experiments has mainly focused on a continuous response variable. In the example of quantal response data in mixture experiments, however, the tumor incidence data have been analyzed in Chen *et al*. (1996) to study the effects of 3 dietary components on the expression of mammary gland tumor. In this paper, we compared the logistic regression models with linear predictors such as second degree Scheffe polynomial model, Becker model and Akay model in terms of classification accuracy.

Keywords: Classification accuracy, logistic regression, mixture experiments, quantal response, second degree Scheffe polynomial model.

[†] This reserach was supported by 2015 Research Grant from Kangwon National University(No. 520150417).

[1] Professor, Department of Information Statistics, Kangwon National University, Gangwon-do 200-701, Korea. E-mail: jikim@kangwon.ac.kr