REVIEW ARTICLE

과대산포가 존재하는 이항형 자료의 회귀분석방법

연세대학교 의과대학 의학통계학과학교실,[1] 의학교육 및 정신과학교실[2]
김동기[1] · 한무영[1] · 전우택[2] · 명성민[1] · 송기준[1]

Regression Methods for Overdispersed Dichotomous Response Data

Dong Kee Kim, PhD[1], Mooyoung Han, MS[1], Wootaek Jeon, MD, PhD[2], Sung Min Myoung, MS[1] and Ki Jun Song, PhD[1]

Department of Biostatistics,[1] Medical Education and Psychiatry,[2] Yonsei University College of Medicine, Seoul, Korea

In neuropsychiatrical research, many problems of statistical inference concern the relationship between the PTSD and traumatic experiences. The logistic model is widely used for modeling a relationship between the covariate and the magnitude of the PTSD. A common complication in the logistic model for dichotomous response data is overdispersion. In this study, two different methods for analyzing dichotomous response data are illustrated and compared. One method is the logistic regression approach, where the numbers of dichotomous responses are predicted by the logistic function of covariates. The other one is the overdispersed logistic regression approach, where the overdispersion is measured by a scale parameter in the variance function of the dichotomous response. In dichotomous response model, when reponses are overdispersed, the overdispersed logistic regression produces more appropriate standard errors of the regression coefficients and the 95% confidence intervals of odds ratios. Therefore, in neuropsychiatrical research, it is recommended to examine the overdispersion problems for their data set before applying the logistic regression model. (J Korean Neuropsychiatr Assoc 2005;44(5):549-552)

KEY WORDS : Logistic regression · Overdispersion · PTSD · Dichotomous response data.

서 론

종속변수가 두 가지 값만을 취하는 이분형(dichotomous, binary) 혹은 질적(qualitative) 변수(질병 발생의 유/무 등)일 경우에 선형회귀모형을 적용하게되는데, 이 경우 선형회귀모형에서 설정한 몇 가지 가정을 심하게 위반하게 된다. 이분형 자료에 대하여 수식적으로 모형화하는 작업은 복잡하기 때문에 구체적인 통계학적 방법이 사용되는데, 주로 회귀분석을 사용하여 독립변수(예를 들면, 약물의 용량이나 과거의 경험 등)와 종속변수(예를 들면, 반응이나 질환의 발생여부)의 정도간의 관계를 규명하게 된다. 이러한 경우에는 분석을 위하여 회귀분석 중에서 로지스틱 회귀분석이 주로 쓰여져 왔다. 로지스틱 회귀분석은 독립변수와 종속변수간의 관계를 로지스틱 함수(logistic function)를 통하여 규명하는 방법이며,[1,2] 로지스틱 회귀분석을 통해 연구자는 이분형 자료에 대한 분석 및 해석을 하는데 있어 만족할 수 있는 결과를 얻을 수 있다.

그러나 로지스틱 회귀모형을 통하여 이분형 분석을 실시할 때에 흔히 나타나는 문제중의 하나는 자료의 변동이 심하여 로지스틱 회귀모형으로 설명되지 못한 변동(variation)이 존재한다는 점이다. 이를 통계학에서는 과대산포(overdispersion)라고 한다. 비연속적 정량분석에서 과대산포는 여러 형태로 나타나는데, 주로 개체간 변동(inter-subject variation) 혹은 임의효과(random effect) 때문에 발생하게 된다.[3] 구체적으로 설명하면 반응변수들이 군집표본추출(cluster sampling)에 의해서 관찰되었을 때, 일반적으로 이항모수(binomial parameter)가 군(cluster)에서 군으로 변화한다는 것을 가정할 수 있다. 또한 약물의 효과와 안정성을 파악하는 분석의 경우는 각 용량별로, 설문 문항을 통해 분석하는 경우 역시 해당 점수별로 숫자의

접수일자 : 2005년 8월 24일 / 심사완료 : 2005년 9월 10일
Address for correspondence
Dong Kee Kim, Ph.D. Department of Biostatistics, Yonsei University College of Medicine, 134 Shinchon-dong, Seodaemun-gu, Seoul 120-752, Korea
Tel : +82.2-2228-2491, Fax : +82.2-364-8037
E-mail : dkkimbios@yumc.yonsei.ac.kr
본 연구는 연세대학교 신진교수연구비 지원으로 이루어졌음.

변동이 심하게 나타날 수 있는데 이러한 것들이 과대산포의 문제를 내포하지 않나 의심스럽게 보인다.[4,5] 이러한 과대산포를 해결하기 위하여, 본 연구에서는 로지스틱 회귀분석을 발전시킨 방법을 제시하고자 한다. 본 연구에서는 이를 과대산포 로지스틱 회귀분석 모형(overdispersed logistic regression model)이라고 부르고자 한다.

정신과 연구에서 외상후 스트레스 장애(post traumatic stress disorder, PTSD) 자료에 대해 로지스틱 회귀모형이 주로 쓰여왔다. 정문용 등[6,7]은 참전재향군인에서 PTSD의 유병상태와 관련 요인을 분석하였으며, 전우택 등[8]은 탈북자의 PTSD 상태와 탈북과정의 외상 경험과의 관계를 로지스틱 회귀모형으로 분석하였다. 본 연구에서는 PTSD 분석에 주로 쓰이는 로지스틱 회귀모형과 과대산포 로지스틱 회귀분석 결과를 제시하여, 그 통계학적 특성을 비교해 보고자 한다.

로지스틱 회귀모형과 과대산포 로지스틱 회귀분석

로지스틱 회귀모형(Logistic regression model)

로지스틱 회귀분석을 이용한 모형은 다음과 같이 정의된다.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

여기에서 β_0는 절편이며, $\beta_1, \cdots, \beta_p$는 p개의 독립변수 $x_1, \cdots, x_p$에 해당하는 회귀계수이다. j번째 변수에 해당하는 회귀계수의 추정치와 표준오차를 $\hat{\beta}_j$와 se_j라고 하면, 교차비(odds ratio)에 대한 95% 신뢰구간은 다음과 같다.

교차비(odds ratio) $= \exp(\hat{\beta}_j)$

교차비의 95% 신뢰구간(95% confidence interval) $= \exp(\hat{\beta}_j \pm 1.96 \times se_j)$

과대산포 로지스틱 회귀모형(Overdispersed logistic regression model)

과대산포 로지스틱 회귀모형은 다음의 분산을 가정한다.

분산 $= \Phi p(1-p)$

여기서 Φ는 과대산포의 정도를 측정하는 모수이다. Φ 값에 따라 모형을 분류하면 Table 1과 같다. Φ가 1이면 로지스틱 회귀모형과 동일하며, Φ가 1보다 크면 과대산포가 존재하는 경우이다. Φ가 1보다 작으면 오히려 과소산포(underdispersion)인 경우인데, 일반적으로 잘 관찰되지 않으므로 특수한 경우에만 적용된다. 과대산포 로지스틱 회귀모형은 최근 통계 패키지에서 추정이 가능하다.

과대산포 로지스틱 회귀모형을 아용하여 회귀계수를 추정하면 통상적으로 로지스틱 회귀모형에 비해 $\hat{\beta}_j$은 동일하나 se_j가 다르게 구해진다. 그러므로 교차비는 동일하나

Table 1. Classification of the overdispersed binomial model

	Model
$\Phi=1$	Logistic regression model
$0<\Phi<1$	Underdispersed logistic regression model
$\Phi>1$	Overdispersed logistic regression model

Table 2. Logistic regression and overdispersed logistic regression analyses for the PTSD using trauma experiences in North Korea and during defection

Trauma factor	Logistic regression		Overdispersed logistic regression	
	Estimate (s.e.†)	Odds ratio (95% C.I.)	Estimate (s.e.)	Odds ratio (95% C.I.)
Intercept	−0.9241 (0.1676)		−0.9241 (0.1717)	
Trauma in North Korea				
Physical trauma	0.0461 (0.3068)	1.074 (0.574, 1.911)	0.0461 (0.3145)	1.047 (0.565, 1.939)
Political · ideological trauma	−0.1761 (0.2244)	0.839 (0.540, 1.302)	−0.1761 (0.2300)	0.839 (0.534, 1.316)
Family related trauma	0.5030 (0.2330)	1.654 (1.047, 2.611)*	0.5030 (0.2388)	1.654 (1.036, 2.640)*
Trauma during defection				
physical trauma	0.2311 (0.2623)	1.260 (0.754, 2.107)	0.2311 (0.2688)	1.260 (0.744, 2.134)
discovery and capture related trauma	−0.2430 (0.2736)	0.784 (0.459, 1.341)	−0.2430 (0.2804)	0.784 (0.453, 1.359)
family related trauma	0.0397 (0.1937)	1.040 (0.712, 1.521)	0.0397 (0.1985)	1.040 (0.705, 1.535)
betrayal related trauma	0.2015 (0.2014)	1.223 (0.824, 1.815)	0.2015 (0.2064)	1.223 (0.816, 1.833)

‡ : Pearson's degree of overdispersion=1.0504, p-value=0.3068

† : s.e. : standard error, C.I : confidence interval, ‡ : degree of overdispersion means the estimate of Φ in the overdispersed logistic regression model, * : p-value<.05

95% 신뢰구간이 통상 로지스틱 회귀모형과 다르게 얻어진다.

두 방법의 결과 비교

본 연구는 전우택 등[8]의 연구에서 발표되었던 탈북자의 PTSD에 대한 연구자료를 이용하고자 한다. 원저는 PTSD에 영향을 미치는 변수에 대하여 각 요인을 추출한 뒤 그 요인을 설명변수로 하는 로지스틱 회귀모형을 제시하였다. 본 사례연구에서는 동일한 자료를 로지스틱 회귀모형과 과대산포 로지스틱 회귀보형을 적용하고 그 결과를 비교하고자 한다.

탈북자에 대한 설문 문항은 북한 내 외상경험척도(25문항), 탈북과정 외상경험척도(19문항)으로 구성되어 있다. 전우택 등[8]은 이 설문지의 척도를 이용해 북한 내 외상사건과 탈북과정 중 외상사건 중에서 어떤 종류의 외상사건들이 외상후 스트레스성 장애(Post-traumatic stress disorder, PTSD)의 발병과 관련이 있는지 알아보기 위해서 요인분석(factor analysis)을 실시하였다. 북한 내 외상사건은 체적 외상(8개 문항), 정치적 · 사상적 외상(4개 문항), 가족과 연관된 외상(5개 문항)의 3개의 요인으로 나눌 수 있으며, 탈북과정 중 외상사건은 육체적 외상(4개 문항), 발각 및 체포와 연관된 외상(7개 문항), 가족과 연관된 외상(1개 문항), 배신과 연관된 외상(2개 문항)인 4개의 요인으로 나눌 수 있다.

Table 2는 이렇게 나누어진 요인과 PTSD 발병 여부의 관계를 파악하기 위해 로지스틱 회귀분석과 과대산포 로지스틱회귀분석을 실시한 결과로 제시된 모형의 기울기 및 절편에 대한 회귀계수(regression coefficients)의 추정치이다. 여기서 우리가 알 수 있는 것은 로지스틱 회귀모형과 과대산포 로지스틱 회귀모형에서 회귀계수에 대한 추정치는 동일하나 표준오차가 다르다는 점이다. 본 연구에서는 북한 내 요인과 탈북과정 중의 요인 7개 중에서 북한 내 요인 중 가족관련 trauma가 로지스틱 회귀분석 결과(p-value=0.0309)와 과대산포 로지스틱 회귀모형 결과(p-value=0.0352)에서 유의한 결과를 보였으며, 이 때의 odds ratio는 두 분석 모두 1.654로 같았으나 95% 신뢰구간이 로지스틱 회귀모형(1.047, 2.611)과 과대산포 로지스틱 회귀모형(1.036, 2.640)에서 서로 다른 결과를 얻었다. 가족관련 trauma 요인을 제외한 다른 여섯 개의 요인은 유의한 결과를 얻을 수는 없었다. 또한, 과대산포에 대한 유의성 검정 결과 p-value가 0.3068로서 유의한 결과를 얻을 수 없어 본 연구에 사용된 자료의 경우 과대산포가 통계학적으로 유의하지 않음을 알 수 있다. Fig. 1은 과대산포 로지스틱회귀모형을 이용하여 북한 내 가족과 연관된 외상의 점수에 따라 PTSD의 확률을 표현한 것이다. 북한 내 가족과 연관된 외상점수가 증가할수록 PTSD 확률이 증가됨을 알 수가 있고, 과대산포 로지스틱회귀모형으로 추정된 확률곡선이 실제 관찰치와 근접한 것을 알 수가 있다.

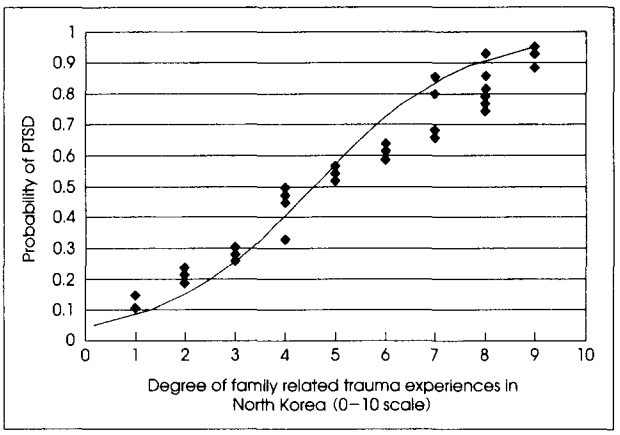

Fig. 1. Observed and estimated probabilities of the PTSD occurrence according to the degree of family related trauma experiences in North Korea. ◆ : observed PTSD percentage, – : estimated probability of PTSD using overdispersed logistic regression.

결 론

본 연구에서는 이분형으로 측정된 자료에 대한 회귀분석 방안을 모색하였다. 이분형적 모형의 경우에는 로지스틱 회귀 모형이 흔히 사용되는데, 로지스틱 회귀모형은 자료가 과대산포 되어있지 않은 경우에는 적절한 방법이라 볼 수 있다. 그러나, 이분형적 모형에서 과대산포가 존재할 때는 로지스틱 회귀분석은 적절하지 못한데, 회귀계수에 대한 추정치는 심각하게 편향되어(biased)있지는 않지만, 표준오차가 정확하지 않게 된다. 그 결과 회귀계수에 대한 신뢰구간 및 p-value가 부적절하게 얻어지게 된다.

탈북자의 PTSD 유무를 알아보는 자료를 가지고 적용하여 보았을 때는 과대산포가 통계학적으로 유의하지 않다는 결과를 얻었다. 그러나 이분형적으로 측정된 많은 자료들의 경우 과대산포의 정도가 심하여 과대산포 로지스틱 회귀모형이 보다 통계학적으로 유의한 경우가 많으므로 로지스틱 회귀모형을 적용하기 전에 과대산포에 대한 검정이 필요하다고 할 수 있다.

본 연구에서는 과대산포를 쉽게 모형화하여 발전시키는 방안을 연구하였다. 이외에도 좀더 복잡하게 모형화하는 방안은 많이 연구되었다. Williams 등은 베타-이항 모형

(beta-binomial model)[3-5]에 기초한 방법을 제시하였고, Kim 등은 transform-both-sides 방법[9,10]을 제시하였다. 그러나 이들 방법은 아직 상용화된 프로그램에 적용하는데 한계가 있다. 본 연구에서는 SPSS와 R을 사용하여 계산하였다. 이 방법은 현재 상용화된 통계 패키지에서 활용 가능하므로 해당 프로그램이 필요한 경우, 저자에게 연락하면 프로그래밍에 대한 정보를 얻을 수 있다.

중심 단어 : 로지스틱 회귀분석 · 과대산포 · 외상후 스트레스성 장애 · 이분형적 자료.

REFERENCES

1) Govindarajulu Z. Statistical techniques in bioassay. New York: Karger; 2001.
2) McCullagh P, Nelder JA. Generalized linear models (2nd Ed.). New York: Chapman and Hall;1989.
3) Williams DA. Extra-binomial variation in logistic linear models. Applied Statistics 1982;31:144-148.
4) Cox DR. Some remarks on overdispersion. Biometrika 1983;70:269-274.
5) Moore DF. Modelling the extraneous variance in the presence of extra-binomial variation. Appl Stat 1987;36:8-14.
6) 정문용, 서 일, 정일진, 김동기, 민경호. 참전재향군인에서 외상후 스트레스 장애의 유병상태와 관련요인 분석-한 병원 입원대상으로. 사회정신의학 2002;7:93-102.
7) 정문용. 외상후 스트레스 장애의 치료. 신경정신의학 2005;44:145-146.
8) Jeon WT, Hong CH, Lee CH, Kim DK, Han M, Min SK. Correlation between traumatic events and posttraumatic stress disorder among North Korean defectors in South Korea. J Traumatic Stress 2005;18:147-154.
9) Kim DK. Regression models for overdispersed jejunal surviving crypts data. In Vitro Cell Dev Biol 2002;38:242-245.
10) Kim DK, Taylor JMG. Transform-both-sides approach for overdispersed binomial data when N is unobserved. J Am Statist Assoc 1994;89:833-845.