The Korean Communications
in Statistics Vol. 9, No. 1, 2002
pp. 281-289

Bootstrapping Logit Model[1)]

Dae hak Kim[2)], Hyeong Chul Jeong[3)]

Abstract

In this paper, we considered an application of the bootstrap method for logit model. Estimation of type I error probability, the bootstrap p-values and bootstrap confidence intervals of parameter were proposed. Small sample Monte Carlo simulation were conducted in order to compare proposed method with existing normal theory based asymptotic method.

Keywords : Logit Model, Bootstrap method, Confidence interval, p-value, Simulation

1. 서론

의학이나 생물학 등 생명과학연구에서의 자료나 설문을 통한 사회조사에서의 자료들은 흔히 범주형으로 주어진다. 이들 자료들은 범주형 변수에 따라 다차원 분할표를 형성하거나 반응값이 0,1의 이항범주나 다항범주 등을 따르게 된다. 이와 같은 범주형 자료의 구조를 파악하고자 할 때 단순히 통상적인 카이제곱 검정이나 연관성의 측정만으로는 많은 한계가 있으며, 이에 로짓모형(logit model), 로그선형모형(log-linear model), 상관모형 등의 활용이 Goodman(1979, 1986), Agresti(1990) 등에 의해 연구되어 왔다. 그러나 범주형자료에 대한 모형론적 접근에서 모수의 추정과 검정 및 그에 따른 모형선택 등의 통계적 추론에 사용되는 대부분의 방법들이 전통적인 근사이론에 의해 이루어져 왔다.

한편 컴퓨터의 계산능력 발전으로 통계학의 많은 부분에서 컴퓨터 지향적인 방법론들이 소개되고 있다. 컴퓨터를 활용한 방법으로 Efron(1979)에 의해 소개된 붓스트랩 방법이나 Kennedy(1995), Manly(1997) 등의 임의순열검정(permutation test)을 들 수 있다. 붓스트랩 방법이나 임의순열방법은 통계량의 표준오차, 편향 그리고 표본분포를 추정하는 방법으로 주어진 자료에 근거한 시뮬레이션 기법 중의 하나이다.

본 연구에서는 로짓모형에 대하여 붓스트랩 방법을 적용하여 모수의 신뢰구간 추정과 가설검정

1) This research was supported by the Catholic University of Daegu research grants in 2000.
2) Professor, Department of Statistical Information, Catholic University of Daegu, Kyungbuk, 712-702, Korea,
E-mail : dhkim@cataegu.ac.kr
3) Assistant Professor, Department of Information Statistics, Pyongtaek University, Kyunggi, 450-701, Korea,
E-mail : jhc@ptuniv.ac.kr

을 제안하고 제안된 방법의 결과와 전통적인 근사이론에 의한 결과를 비교하고자 한다. 언급한 바와 같이 범주형 자료에 대한 대부분의 통계적 추론은 전통적인 근사이론에 의존하고 있기 때문에, 자료의 크기가 작거나, 구조적 영모형(null model) 등에서는 전통적인 근사이론을 적용함이 적당하지 않을 수 있다. 그러나 붓스트랩 방법은 범주형자료와 같은 격자분포에서도 수리적인 장점을 지니고 있음이 밝혀지고 컴퓨터의 발전과 더불어 범주형 자료에 대한 붓스트랩 방법론의 적용이 요구되어 왔다(Woodroofe and Jhun; 1989). Jeong(1997)은 로그선형모형에서 붓스트랩방법을 적용하였고 Jhun 과 Jeong(2000)은 다항분포의 모수들간의 선형결합의 신뢰구간 구축 시 붓스트랩방법의 장점을 연구한 바 있다.

본 연구에서는 계산 집중적(computer intensive)인 붓스트랩 방법을 로짓모형에 적용하여 붓스트랩 방법을 이용한 구간추정과 가설검정의 결과와 기존의 근사이론과 비교하였다. 2절에서는 로짓모형의 일반적인 통계적 추론을 살펴보고 3절에서는 단순로짓모형에서 붓스트랩 방법을 이용한 제 1종 오류(Type I error)와 검정력 그리고 모수의 신뢰구간을 구축하는 방법을 제안하고 소표본의 경우에 모의실험을 통해 신뢰구간의 포함확률(coverage probability)를 추정하고자 한다. 4절에서는 다중로짓모형에서 붓스트랩 방법의 제안과 모의실험을 통한 제 1종 오류와 신뢰구간을 구하고 근사이론과 비교하였다.

2. 로짓모형의 모수 추정

Z를 이항형 반응 $(Z=0,1)$ 그리고 $X_1,\cdots,X_p$를 p개의 설명변수라 하자. 관측자료

$$(z_1,x_{11},\cdots,x_{ip}),(z_2,x_{21},\cdots,x_{2p}),\cdots,(z_n,x_{n1},\cdots,x_{np})$$

에 대한 로짓모형은

$$l(p)=\log\frac{P(Z=1|x_1,\cdots,x_p)}{P(Z=0|x_1,\cdots,x_p)}=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p \tag{1}$$

이다. 성공 $(Z=1)$의 조건부 확률을 $\pi(\boldsymbol{x})=P(Z=1|x_1,\cdots,x_p)$라고 놓자. Y를 관측도수가 n_i인 부모집단(subpopulation) i에서, 주어진 설명변수 $\boldsymbol{x_i}=(x_{i1},x_{i2},\cdots,x_{ip})$ 에서의 성공횟수라 두자. 그러면 $\{Y_i,i=1,\cdots,I\}$ 는 독립된 I개의 집단에서 이항분포를 따르며 $E(Y_i)=n_i\pi(\boldsymbol{x_i})$이다. 여기서 $n_1+n_2+\cdots+n_I=n$ 이며 $(Y_1,\cdots,Y_I)$의 결합확률함수는 I개의 이항분포의 곱에 비례하게 된다. $n_i=1$인 경우의 자료를 비그룹화 자료(ungrouped data)라고 하고, $n_i>1$ 인 경우의 자료를 그룹화 자료(grouped data)라고 한다. 이제 모수 $\boldsymbol{\beta}=\{\beta_0,\beta_1,\cdots,\beta_p\}$의 최대우도추정은 다음과 같이 주어지는 우도함수의 최대화로 얻어진다.

$$\prod_{i=1}^{I}\pi(\boldsymbol{x_i})^{y_i}[1-\pi(\boldsymbol{x_i})]^{n_i-y_i}=\left\{\prod_{i=1}^{I}[1-\pi(\boldsymbol{x_i})]^{n_i}\right\}\exp\left[\sum y_i\log\left(\frac{\pi(\boldsymbol{x_i})}{1-\pi(\boldsymbol{x_i})}\right)\right] \tag{2}$$

이제, (2)의 로그우도함수로부터 우도방정식은 다음과 같이 주어진다.

$$\sum_i y_ix_{ia}-\sum_i n_i\hat{\pi}_ix_{ia}=0,\quad a=1,\cdots,p \tag{3}$$

여기서, $\hat{\pi}_i=\exp(\sum_j\hat{\beta}_jx_{ij})/[1+\exp(\sum_j\hat{\beta}_jx_{ij})]$이다. $\mathbf{X}$를 $\{x_{ij}\}$의 $I\times(p+1)$ 행렬이라고 두

면 (3)식의 우도방정식은 $\mathbf{X'y}=\mathbf{X'\hat{m}}$ 로 되고 이때 $\hat{m_i}=n_i\hat{\pi_i}$ 이다. Newton-Raphson 방법을 사용하여 모수 $\boldsymbol{\beta}$ 의 추정은 다음의 (4)식을 반복하여 계산할 수 있다.

$$\boldsymbol{\beta}^{(t+1)}=\boldsymbol{\beta}^{(t)}+\{\mathbf{X'Diag}[n_i\pi_i^{(t)}(1-\pi_i^{(t)})]\ \mathbf{X}\}^{-1}\mathbf{X'}(\mathbf{y}-\mathbf{m}^{(t)}) \tag{4}$$

위의 (4)식은 일반화선형모형 $\mathbf{z}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ 에서 $\boldsymbol{\varepsilon}$의 공분산 행렬을 $\mathbf{V}$라 놓을 때, 가중최소제곱법(weighted least squares method)에 의한 모수 $\boldsymbol{\beta}$의 추정량 $(\mathbf{X'V}^{-1}\mathbf{X})^{-1}\mathbf{X'V}^{-1}\mathbf{z}$의 반복에 의한 극한값으로 표현할 수 있다(Agresti, 1990). 또한 모수 $\boldsymbol{\beta}$의 공분산 행렬은 $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})=\{\mathbf{X'Diag}[n_i\hat{\pi}_i(1-\hat{\pi}_i)\mathbf{X}\}^{-1}$가 된다. 로짓모형에서

$$\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\ \rightarrow^{d}\ \mathrm{N}[\mathbf{0},\ n(\mathbf{X'Diag}[n_i\hat{\pi}_i(1-\hat{\pi}_i)\mathbf{X})^{-1}]$$

로 분포수렴하므로 정규분포에 근거한 통계적 추론을 유도할 수 있다(Agresti, 1990).

3. 단순로짓모형

3.1 붓스트랩 유의확률 추정

단순로짓모형에서 X_1의 계수 β_1의 가설검정에 대한 문제를 생각하자. 가설 $H_0: \beta_1=0$ 을 검정하기 위해, Wald 검정, 일반화 우도검정(generalized likelihood ratio test), 피셔의 점수 검정(fisher's score test) 등이 활용된다. Wald 검정법은 ML 추정값 $\hat{\beta}$ 에서 로그우도함수의 모양에 영향을 받으며 $\hat{\beta}^2/Cov(\beta)$ 의 형태로 카이제곱분포를 따른다(Agresti, 1995).

본 연구에서는 이와같은 대표본 이론에 의존하지 않고 $\hat{\beta}$의 분포를 붓스트랩 방법으로 유도하여 가설 $H_0: \beta_1=0$에 대한 검정을 실시하고자 한다. 제안하는 붓스트랩 방법의 알고리즘은 다음과 같다.

[단계1] 관측자료에 $l(p)=\beta_0+\beta_1 x_1$ 의 로짓모형을 적합시켜 β_0, β_1을 계산한다.

[단계2] $\beta_1=0$인 가정하에서, 베르누이분포 $B(1, \hat{P})$ 로부터 n개의 붓스트랩 표본 $\{z_j^*; j=1,\cdots,n\}$ 을 얻는다. 여기서 확률 $\hat{P}$는 $\hat{P}(Z=1)=\sum y_i/n$ 로 $l(p)=\beta_0$ 에서 얻은 값이다.

[단계3] 단계2의 붓스트랩 표본 z_j^*을 순서대로 공변량 X_1에 결합한다. 또한 붓스트랩 표본을 로짓모형에 적합하여 β_0, β_1 의 붓스트랩 MLE $\hat{\beta}_0^*$와 $\hat{\beta}_1^*$ 를 계산한다.

[단계4] 단계2와 단계3을 M번 반복하여 $|\hat{\beta}_b^*| \geq |\hat{\beta}_b|$ 의 상대적 비율로 유의확률을 추정한다.

3.2 붓스트랩 신뢰구간 구축

이제 모수 β_1의 붓스트랩 신뢰구간 추정을 살펴보자. $\{z_j; j=1,\cdots,n\}$ 이 로짓모형 $l(p)=X\beta$

에 의한 반응범주라 할 때, 모수 $\boldsymbol{\beta}$ 의 ML 추정 $\hat{\boldsymbol{\beta}}$ 는 식 (4)와 같이 주어진다. 여기서 식 (4)를 통하여 주어진 반응범주에서 기대확률 $\{\hat{\pi}_i; i=1,\cdots,n\}$을 계산할 수 있다. 만일 반응범주가 한 범주에서 n_i번 관찰된다면 기대확률은 $\{\hat{\pi}_i; i=1,\cdots,I\}$이 될 것이다. 모형에 의한 추정값 $\{\hat{\pi}_i; i=1,\cdots,n\}$을 이용하여 붓스트랩 표본 $\{z_i^*; i=1,\cdots,n\}$를 얻고 붓스트랩 표본 $\{z_i^*; i=1,\cdots,n\}$에 의한 추정값을 $\hat{\boldsymbol{\beta}}^*$라 놓으면, $\hat{\boldsymbol{\beta}}^*$를 모형에 기초한 붓스트랩 추정량(model-based bootstrap estimator)이라 한다. 로짓모형에서 $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ 은 $\mathrm{N}[\mathbf{0}, n(\mathbf{X}'\mathbf{Diag}[n_i\hat{\pi}_i(1-\hat{\pi}_i)\mathbf{X})^{-1}]$로 분포수렴하므로, Khintchine의 약대수법칙(Khintchine's weak law of large numbers)에 의해 다음의 사실을 알 수 있다(Jeong, 1997).

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^*-\hat{\boldsymbol{\beta}}) \rightarrow^d \mathrm{N}[\mathbf{0}, n(\mathbf{X}'\mathbf{Diag}[n_i\hat{\pi}_i(1-\hat{\pi}_i)\mathbf{X})^{-1}]$$

이와같은 일치성(consistency)에 의해 붓스트랩 신뢰구간은 다음과 같이 구축한다.

[단계1] 관측자료에 $l(p)=\beta_0+\beta_1 x_1$ 의 로짓모형을 적합시켜 β_0, β_1을 계산한 후 적합된 모형으로부터 반응범주의 각 수준에서 성공확률 $\{\hat{\pi}_i; i=1,\cdots,n\}$를 추정한다.

[단계2] $\{\hat{\pi}_i; i=1,\cdots,n\}$로부터 붓스트랩 표본 $\{z_i^*; i=1,\cdots,n\}$를 발생하여 붓스트랩 최우추정치 $\hat{\beta}_0^*$와 $\hat{\beta}_1^*$ 를 계산한다.

[단계3] [단계2]를 B 회 반복하여 다음을 계산한다.

$$B^*(b)=\frac{\sqrt{n}(\hat{\beta}_i^*-\hat{\beta}_i)}{ASE(\hat{\beta}_i^*)},\quad b=1,2,\cdots,B$$

[단계4] $100(1-\alpha)\%$ 붓스트랩 신뢰구간은 다음과 같다.

$$L_i^{Boot}=\hat{\beta}_i-B_{(1-\alpha/2)}SE(\hat{\beta}_i),\quad U_i^{Boot}=\hat{\beta}_i-B_{(\alpha/2)}SE(\hat{\beta}_i) \tag{5}$$

여기서, $B_{(1-\alpha)}$는 붓스트랩 분포 $B^*(\cdot)$에서 $100(1-\alpha)\%$ 에 해당되는 분위점이다.

한편 Wald 신뢰구간은 정규분포를 사용한다는 점에서 붓스트랩 신뢰구간과 차이가 있다. Wald의 신뢰구간은 다음과 같다.

$$L_i^{Wald}=\hat{\beta}_i-Z_{(1-\alpha/2)}SE(\hat{\beta}_i),\quad U_i^{Wald}=\hat{\beta}_i-Z_{(\alpha/2)}SE(\hat{\beta}_i) \tag{6}$$

여기서, $Z_{(1-\alpha)}$는 표준정규분포에서 $100(1-\alpha)\%$ 에 해당되는 분위점이다.

3.3 모의실험을 통한 비교

본 절에서는 모의실험을 통하여 Wald 방법과 붓스트랩 방법의 검정력과 신뢰구간을 살펴보겠다. 모의실험을 위하여 다음의 세 가지 모형을 고려하였다.

모형 1: $l(p)=0+0x_1$ **모형 2**: $l(p)=-2.5+0.5x_1$ **모형 3**: $l(p)=-5.0+1.0x_1$

모형1은 각 설명변수의 수준에서 확률의 변화 증가율 $\beta\pi(x_i)(1-\pi(x_i))$ 가 0인 상황이며, 모형 2는 확률의 변화 증가율이 중간효율수준(median effective level)에서 0.125로 비교적 완만 상태이며, 모형 3은 중간효율수준에서 확률의 변화증가율이 0.25로 급격한 상승을 지니는 모형이라고 할 수 있다. 그리고 설명변수 x_1 의 수준은 1부터 10까지 정수의 값을 취하도록 하였다. 이 모형에 의한 각 설명변수의 수준에서 성공의 기대확률은 <표 1>처럼 나타난다.

<표1> 각 모형의 설명변수의 수준에 따른 성공의 기대확률

x_1	1	2	3	4	5	6	7	8	9	10
모형1	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
모형2	0.119	0.182	0.269	0.378	0.500	0.622	0.731	0.818	0.881	0.924
모형3	0.018	0.047	0.119	0.269	0.500	0.731	0.881	0.953	0.982	0.993

각 모형에서 검정력과 신뢰구간을 구하기 위해, 명목수준은 $1-\alpha=0.90,0.95$를 고려하였고 표본의 크기는 $n=10,20,30,40,50$ 로 설정하였다. 이는 주어진 x_i 의 수준에서 각각 1회, 2회, 3회, 4회, 5회의 동일한 횟수의 반복 실험을 한 그룹 데이터를 의미한다. 모의실험 절차를 요약하면 다음과 같다.

[단계1] 주어진 모형 $l(p)$ 로부터 베르누이 랜덤표본을 n개 발생한다.
[단계2] 주어진 랜덤표본로부터 로짓모형을 적합하여 3.1 소절의 알고리즘에 의해 Wald 유의확률(P_w)과 붓스트랩 유의확률(P_b)을 계산한다.
[단계3] 주어진 랜덤표본로부터 3.2 소절의 알고리즘에 의해 주어진 명목수준 $100(10\alpha)\%$에서 Wald 의 신뢰구간(L_i^{Wald}, U_i^{Wald})과 붓스트랩 신뢰구간(L_i^{Boot}, U_i^{Boot})를 구축한다.
[단계4] 위의 [단계1]과 [단계3]를 M 회 반복한다.
[단계5] 주어진 명목 수준 α에 의해 다음의 통계량을 계산한다.

$\hat{\text{P}}\text{ower}_{\text{w}} = \#[P_w \le \alpha]/M$, $\hat{\text{P}}\text{ower}_{\text{B}} = \#[P_b \le \alpha]/M$

$\text{Coverage}^{\text{Wald}} = \#[\text{L}_i^{\text{Wald}} \le \beta_i \le \text{U}_i^{\text{Wald}}]/\text{M}$, $\text{Coverage}^{\text{Boot}} = \#[\text{L}_i^{\text{Boot}} \le \beta_i \le \text{U}_i^{\text{Boot}}]/\text{M}$, $i=0,1$

여기서, $\{\beta_i ; i=0,1\}$는 모형에서 주어지는 모수 값이다. 또한 신뢰구간의 평균 폭과 신뢰구간 폭의 표준오차(SE)를 계산한다.

모의실험 결과는 아래 표에 주어져 있다. <표 2>는 모형 1에서 β_1의 1종 오류 추정결과라 하겠다. 유의수준 0.05에서 살펴보면, Wald의 검정이 다소 보수적으로 보인다. 그러나 표본의 크기가 40이하에서 대체적으로 두 방법 모두 보수적인 입장을 취하고 있음을 볼 수 있다. 이와 같은 현상은 다른 유의수준에서도 계속 유지되고 있다. 여기서 표본의 크기가 작을 때 ($n \le 40$) 1종 오류를 조절하는 능력은 붓스트랩 검정이 Wald 방법보다 좋음을 알 수 있다. 그리고 표본의 크기가 50이 되면서 두 방법 모두 1종 오류 수준에 근사하게 됨을 볼 수 있다. <표 3>은 모형 2와 모형 3의 모의실험결과는 경험적 검정력을 나타내고 있다. 유의수준 0.05에서 검정력의 크기는 두 방법간에

다소 차이가 있음을 보여준다. 표본의 크기가 10일 때 붓스트랩 검정력은 0.340이며 Wald 의 검정력은 0.191로 표본의 크기가 작을 때 현저히 낮은 수준임을 볼 수 있다. 모형 2에서 표본의 크기가 커지면서 두 방법들의 검정력간에는 큰 차이가 없어지나, 붓스트랩 검정력, Wald 검정력의 순이 그대로 유지되고 있음을 볼 수 있다. 모형 3에서도 비슷한 현상이 발생한다. 이상에서 보면 붓스트랩 방법의 검정력이 높으며, 1종 오류를 조절하는 능력이 약간 좋음을 알 수 있다.

<표 2>모형 1의 β_1에 대한 1종 오류(Type I error rates)

방법	α \ n	10	20	30	40	50
wald	0.05	0.024	0.029	0.040	0.038	0.050
	0.1	0.057	0.081	0.086	0.092	0.104
붓스	0.05	0.035	0.036	0.043	0.043	0.051
트랩	0.1	0.076	0.084	0.084	0.089	0.099

<표 3>모형 2와3의 β_1에 대한 경험적검정력

방법	모형	α \ n	10	20	30	40	50
wald	모형2	0.05	0.191	0.700	0.908	0.979	0.993
		0.1	0.352	0.858	0.983	0.991	0.999
	모형3	0.05	0.525	0.895	1.000	1.000	1.000
		0.1	0.881	1.000	1.000	1.000	1.000
붓스 트랩	모형2	0.05	0.340	0.740	0.912	0.978	0.993
		0.1	0.484	0.854	0.985	0.992	0.999
	모형3	0.05	0.777	0.999	1.000	1.000	1.000
		0.1	0.899	1.000	1.000	1.000	1.000

<표 4> 모형 1에서 신뢰구간의 포함확률 (붓스트랩반복 1,000회)

신뢰	n	10		20		30		40		50	
수준	방법	Wald	Boot	Wald	Boot	Wald	Boot	Wald	Boot	Wald	Boot
0.9	포함확률	.943	.885	.919	.910	.910	.901	.908	.904	.896	.891
	폭	1.244	1.129	.566	.547	.442	.433	.377	.372	.334	.331
	SE	.060	.058	.004	.003	.001	.001	.001	.001	.000	.000
0.95	포함확률	.976	.941	.971	.947	.959	.948	.962	.950	.950	.947
	폭	1.482	1.997	.674	.634	.527	.506	.449	.436	.398	.389
	SE	.072	.165	.005	.011	.001	.001	.001	.001	.000	.001

이제, 신뢰구간에 대해 살펴보자. 신뢰구간은 두 방법 모두 근사적으로 정규분포를 따른다고 할 수 있으므로 표본의 크기가 커짐에 따라 포함확률간에는 큰 차이가 발생하지 않음을 예상할 수 있다. 표본의 크기가 10으로 작을 때 Wald 의 신뢰구간이 명목수준 90%와 95%에서 가장 보수적임을 알 수 있다. 또한, 붓스트랩 신뢰구간이나 임의순열 신뢰구간은 표본의 작을 때 양 극단에 영향을 받으므로 신뢰구간의 평균폭이 Wald 의 신뢰구간의 평균폭에 비해 큼을 볼 수 있다. 표본

의 크기가 20이상 커짐에 따라 붓스트랩 신뢰구간은 주어진 명목수준에 상당히 근사하게 되며 신뢰구간의 평균폭도 작아짐을 볼 수 있다. 명목수준 95%에서 살펴보면, Wald 의 신뢰구간은 주어진 명목수준을 초과하는 현상을 보여주고 있으나, 붓스트랩 신뢰구간은 상당히 안정적인 수렴을 하고 있음을 볼 수 있다. 그러나 표본의 크기가 커짐에 따라 두 방법간에는 큰 차이가 나타나지 않는데 이는 두 방법 모두 정규분포로 수렴하는 성질이 있기 때문이다. 모형 2와 모형 3의 신뢰구간에서도 모형 1과 비슷한 현상을 발생하므로 여기서는 제시하지 않았다.

4. 다중로짓모형

4.1 다중로짓모형에서 붓스트랩방법

이항형 반응 ($Z=0,1$)에 대해 p개의 설명변수 $X_1,\cdots,X_p$를 고려하자. 먼저, 로짓모형 (1)에서 가설 $H_0:\beta_p=0$를 검정하기 위한 붓스트랩 검정을 고려해보자. 다음의 과정으로 붓스트랩 검정을 실시할 수 있다.

[단계1] 관측자료에 $l(p)=\beta_0+\beta_1x_1+\cdots+\beta_px_p$ 의 로짓모형을 적합하여 $\hat{\boldsymbol{\beta}}$ 를 계산한다. 또한 $H_0:\beta_p=0$하에서 $l(p)=\beta_0+\beta_1x_1+\cdots+\beta_{p-1}x_{p-1}$ 을 적합한다.

[단계2] $\beta_p=0$ 인 가정하에서, 추정된 $l(p)$로부터 설명변수 $x_1,x_2,\cdots,x_{p-1}$에서 $\hat{\pi}(\boldsymbol{x})$의 확률을 계산한다. 베르누이분포 $B(1,\hat{\pi}(\boldsymbol{x}))$에서 n 개의 모수적 붓스트랩 표본 z_i^*을 발생한다.

[단계3] 단계2의 붓스트랩 표본 z_i^*을 순서대로 설명변수 $x_1,x_2,\cdots,x_{p-1},x_p$ 에 결합한 후 붓스트랩 표본으로부터 다중로짓모형에 적합하여 적합하여 MLE $\hat{\beta}_0^*,\cdots,\hat{\beta}_p^*$ 를 계산한다.

[단계4] 단계2와 단계3을 M번 반복하여 $|\hat{\beta}_p^*|\geq|\hat{\beta}_p|$ 의 상대적 비율로 유의확률을 추정한다.

붓스트랩 방법은 앞의 3.1 소절에서 언급한 방법과 큰 차이가 나지 않는다. 신뢰구간 구축에 있어서는 앞의 3.2 소절에서 소개한 단순로짓모형에서의 신뢰구간 구축방법과 다중로짓모형에서의 신뢰구간 구축과는 별 차이가 없다.

4.2 모의실험을 통한 비교

본 절에서는 다중로짓에 대해서 모의실험을 통해 Wald 방법과 붓스트랩 방법의 제 1종 오류와 신뢰구간을 살펴보겠다. 모의실험을 위해 모형으로 $\beta_0=\beta_1=\beta_2=0$ (모형4)을 고려하였고, 모의실험에서 설정한 설명변수 x_1과 x_2와 각 설명범주의 수준에서 확률은 <표5>와 같이 하였다. 고려된 모형에서 제 1종 오류와 신뢰구간을 구하기 위해, 명목 수준은 $1-\alpha=0.90,0.95$를 또 표본의 크기는 $n=20,30,40,50$으로 설정하였다. 이는 주어진 설명변수의 수준에서 각각 2회, 3회, 4회, 5회의 동일한 횟수의 반복 실험을 한 그룹데이터임을 의미한다. 모의실험 절차는 앞의 3.4절과 동일하다.

<표 5 > 모형4의 설명변수의 수준과 반응범주의 확률값

x_1	1	2	3	4	5	6	7	8	9	10
x_2	1	1	1	2	2	2	2	3	3	3
확률	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

<표 6>모형 4의 β_2에 대한 1종 오류(Type I error rates)

방법	α \ n	20	30	40	50
wald	0.05	0.028	0.028	0.039	0.052
	0.1	0.090	0.091	0.090	0.100
붓스	0.05	0.019	0.027	0.043	0.050
트랩	0.1	0.057	0.083	0.089	0.095

<표 7> 모형 4에서 신뢰구간들의 포함확률 (붓스트랩반복 1,000회)

신뢰	n	20		30		40		50	
수준	방법	Wald	Boot	Wald	Boot	Wald	Boot	Wald	Boot
0.9	포함확률	.910	.916	.908	.898	.904	.902	.901	.899
	폭	6.429	7.100	5.019	.4990	4.275	4.251	3.785	3.772
	SE	.025	.106	.010	.014	.007	.007	.004	.005
0.95	포함확률	.972	.958	.972	.959	.961	.952	.952	.945
	폭	7.661	8.621	5.981	5.810	5.094	4.976	4.510	4.434
	SE	.029	.135	.012	.018	.008	.001	.005	.006

모의실험 결과는 <표 6>과 <표 7>에 주어져 있다. <표 6>에서 1종 오류를 살펴보면 두 방법들의 제 1종 오류가 주어진 유의수준보다 아래서 형성되고 있으므로 두 방법들이 보수적인 성향을 지니고 있음을 보여주고 있다. 본 모의실험에서는 표본의 크기가 커짐에 따라 유의수준 0.01과 0.05에서 제 1종 오류를 조절하는 능력은 붓스트랩과 Wald 방법 간에 큰 차이가 없음을 볼 수 있었다. 주어진 명목수준에 대해 신뢰구간의 포함확률을 계산한 결과는 <표 7>이다. 여기서 우리는 Wald의 신뢰구간은 주어진 명목수준을 모두다 넘어서는 경향이 있음을 발견할 수 있다. 즉 Wald 의 신뢰구간은 과대적합하며, 상대적으로 보수적인 방법임을 의미한다. 이는 표본의 크기가 작을 때 그 경향이 더 나타나는데, 범주형 자료에서 정규근사에 의한 적합도검정이나 통계적 추론에서 가능한 한 귀무가설을 기각하지 않으려는 경향이 있음을 본 모의실험에서도 재확인하여 주는 결과라 하겠다. 표본의 크기가 40이상이 되면서 신뢰구간의 포함확률에서 Wald 방법과 붓스트랩 방법 간에는 큰 차이가 나타나지 않음을 확인할 수 있겠다.

5. 토의와 결론

본 연구에서 다룬 내용은 로짓모형에 대하여 붓스트랩 방법을 적용하여 기존의 전통적인 근사이론과 비교하였다. 로짓모형에서의 붓스트랩 검정방법을 제안하였고 로짓모형에서 붓스트랩분포를 이용하여 신뢰구간을 구축하는 방법을 제시하였다.

모의실험을 통하여 Wald의 검정이 가장 보수적인 성향을 지니고 있으며, 검정력이 낮음을 볼 수 있었다. 그리고 붓스트랩검정은 Wald 의 방법을 보완하는 방법임을 확인할 수 있었다. 붓스트랩 방법은 표본 추출과정에서 미리 가중치가 부여되기에 계산에 있어 다소 간편하다고 하겠다. 신뢰구간 구축 결과는 검정의 결과와 유사하다. 신뢰구간에서도 Wald 의 신뢰구간이 다소 보수적인 성향을 지니고 있음을 재 확인할 수 있었다. 그런데, 두 가지 방법 역시 모두 정규분포에 수렴하므로 표본의 크기가 커짐에 따라 주어진 명목수준에 근사하는 경향에는 큰 차이가 없다고 하겠다. 결론적으로 붓스트랩 방법은 로짓모형에서 일반적으로 사용되는 정규근사의 추론에 대한 비모수적 대안이라고 하겠다.

참고문헌

[1] Agresti A. (1990) *Categorical Data Analysis,* New York : John Wiley & Sons.

[2] Agresti A. (1995) *An Introduction to Categorical Data analysis*, New York : John Wiley & Sons.

[3] Efron B. (1979) Bootstrap methods; another look at the jackknife, *Annals of Statistics*, Vol. 7, 1-26.

[4] Goodman L.A. (1979) Simple models for square contingency tables with ordered categories *Biometrika*, Vol. 66, 413-418.

[5] Goodman L.A. (1986) Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables, *International Statistical Review*, Vol. 54, 243-209.

[6] Jeong H.C (1997) *Bootstrap Methods for Categorical Data Analysis*, Ph. D. dissertation, Korea University.

[7] Jhun M. and Jeong H.C (2000) Applications of Bootstrap Methods for Categorical Data Analysis, *Computational statistics & Data Analysis,* Vol. 35, 83-91.

[8] Kennedy, P.E. (1995) Randomization Tests in Econometrics, *Journal of Business and Economic Statistics*, Vol. 13, 85-94.

[9] Manly, B.F.J (1997) *Randomization, Bootstrap and Monte Carlo Methods in Biology,* Chapman and Hall, London.

[10] Woodroofe M. and Jhun M. (1989) Singh's theorem in the lattice case, *Statistics and Probability Letters*, Vol. 7, 201-205.

[2001년 12월 접수, 2002년 4월 채택]