The Korean Communications
in Statistics Vol. 10 No. 2, 2003
pp. 277-289

Improved Exact Inference in Logistic Regression Model

Donguk Kim[1], Sooyeon Kim[2]

Abstract

We propose modified exact inferential methods in logistic regression model. Exact conditional distribution in logistic regression model is often highly discrete, and ordinary exact inference in logistic regression is conservative, because of the discreteness of the distribution. For the exact inference in logistic regression model we utilize the modified P-value. The modified P-value can not exceed the ordinary P-value, so the test of size α based on the modified P-value is less conservative. The modified exact confidence interval maintains at least a fixed confidence level but tends to be much narrower. The approach inverts results of a test with a modified P-value utilizing the test statistic and table probabilities in logistic regression model.

Keywords: Confidence interval; Conservative; Coverage probability; Exact inference; Logistic regression; Modified exact P-value; $2\times J\times K$ table.

1. 서론

검정통계량 T의 정확분포(exact distribution)에 근거한 정확검정(exact test)에서 검정 통계량의 분포가 이산형 분포인 경우 $H_0: \theta=\theta_0$ 검정에서 1종 오류 확률이 유의수준 α와 같거나 작게 된다. 그리고 유의수준 α에서 모수 θ에 대한 정확 신뢰구간은 모든 θ에 대하여 그 포함확률이 최소한 $1-\alpha$이다. 그러므로 그러한 신뢰구간들은 보수적(conservative)이다. 이 보수성을 줄이기 위한 연구가 많이 이루어졌다. Lancaster(1961)는 관측값의 확률의 절반과 더 극단적인 값의 확률을 더한 것으로 정의되는 'mid-P'를 제안했다. 비록 mid-P를 이용한 검정에서 실제 1종 오류율이 유의수준보다 커질 가능성이 있지만 이 방법의 장점은 보수성을 줄이며 랜덤화에 의존하지 않는다는 것이다. Cohen과 Yang(1994)는 포아송 분포의 평균에 대한 신뢰구간에서 mid-P가 보수성을 줄임을 보였다. Kim(1998)은 수정된 mid-P의 활용을 제안하였으며 수정된 mid-P를 사용한 신뢰구간이 일반적인 mid-P를 사용한 경우보다 모수를 포함하는 포함확률(coverage probability)이 명목수준에 더 근접함을 보였다. 그리고 통계적 추론에서 양측 mid-P의 장점을 보였다.

---

1) Associate Professor, Department of Statistics, Sungkyunkwan University.
E-mail: dkim@skku.ac.kr
2) Department of Statistics, Sungkyunkwan University.

Cohen과 Sackrowitz(1992)는 수정된(modified) P-값이 일반적인 P-값보다 보수성을 줄일 수 있다는 것을 보여준다. 그리고 Kim과 Agresti(1995)는 수정된 P-값을 활용한 수정된 정확추론이 이산성에 기인한 보수성을 줄이며 수정된 정확 신뢰구간이 일반적인 정확 신뢰구간보다 모수를 포함하는 포함확률이 명목수준에 더 근접함을 보였다. 또한 통계적 추론에 양측(two-sided) P-값의 우수성을 보였다.

사례-대조 연구는 일반적으로 이진 반응변수와 설명변수들 사이의 관계를 로지스틱(logistic) 회귀모형을 사용하여 나타낸다(Day et al., 1979; Hirji et al., 1987, 1988; Mehta, 1995). 대표본 근사 방법은 희박자료(sparse data)이거나 소표본인 경우 또는 불균형(unbalanced)자료나 층(stratum)이 많은 경우 그 정확성이 의심스럽다. 우리는 로지스틱 회귀모형에서 회귀모수에 관한 추론을 위하여 우도함수 대신 충분통계량의 정확순열(exact permutation) 분포를 이용한다. 즉, 관심없는 장애모수(nuisance parameter)를 관측된 값으로 고정시켜 관심있는 모수의 충분통계량의 정확순열 분포를 이용한 정확추론을 사용한다(Mehta et al., 1995; Hirji et al., 1994). 그러나 로지스틱 회귀모형에서 이산형 조건부 분포를 사용하는 정확검정은 보수적이 되며 가설검정에서 보수성은 신뢰구간 구축에도 영향을 준다. 로지스틱 회귀모형에서 보수성을 줄이는 문제에 대하여 mid P-값을 이용한 방법이 연구되었으나 이는 정확검정을 제공하지 않는다(Hirji, 1991). 우리는 로지스틱 회귀모형에서 검정통계량의 이산성에 기인한 보수성을 줄일 수 있도록 수정된 P-값을 사용하여 향상된 정확검정 및 정확 신뢰구간을 구축하며 향상된 정확추론을 보인다.

본 논문에서 2절은 로지스틱 회귀모형에서 정확추론을 살펴보며 3절에서 정확검정의 단점인 보수성을 해결하기 위해 로지스틱 회귀모형에서 수정된 P-값을 정의하고 수정된 P-값을 사용한 유의성 검정이 보수성을 줄임을 보인다. 4절에서는 수정된 P-값을 로지스틱 회귀모형에 적용하여 수정된 신뢰구간을 구축하며 향상된 정확추론을 보인다. 그리고 5절은 실제자료에 수정된 정확추론을 적용한다.

2. 로지스틱 회귀모형

이진(binary) 반응변수 Y와 독립변수 X의 로지스틱 회귀모형 자료를 $2\times J\times K$ 분할표로 나타내어 $\{Y_{11k}, Y_{12k}, \ldots, Y_{1Jk}\}$와 $\{n_{+1k}, n_{+2k}, \ldots, n_{+Jk}\}$가 각각 k번째 층에서 $Y=1$인 경우의 j번째 열의 반응변수의 개수와 열의 주변합이라 하자. k번째 층에서 j번째 범주의 설명변수 x_j와 관련된 반응확률을 p_{jk}라 하면 로지스틱 회귀모형은 다음과 같다.

$$\log\left(\frac{p_{jk}}{1-p_{jk}}\right)=\alpha_k+\beta x_j,\ \ j=1,\ldots,J,\ \ k=1,\ldots,K \tag{2.1}$$

여기서 α_k는 층에 관한 절편모수이고, β는 연속한 두 열의 수준들 사이의 공통 로그 오즈비(common log odds ratio)이다.

2.1 조건부 정확 분포

k번째 층에서 관찰된 $\boldsymbol{Y}_{1k}=\{Y_{1jk}, j=1,\dots,J\}$는 열합계가 $\{n_{+jk}, j=1,\dots,J\}$인 J개의 독립적인 이항 확률변수로 나타낼 수 있으며 결합분포는 식(2.2)와 같다.

$$\Pr(Y_{11k}=y_{11k},\dots,Y_{1Jk}=y_{1Jk})=\frac{c_k(\boldsymbol{y}_{1k})\exp\{\sum_{j=1}^{J}y_{1jk}(\alpha_k+\beta x_j)\}}{\prod_{j=1}^{J}[1+\exp(\alpha_k+\beta x_j)]^{n_{+jk}}} \tag{2.2}$$

여기서 $c_k(\boldsymbol{y}_{1k})=\prod_{j=1}^{J}\binom{n_{+jk}}{y_{1jk}}$ 이다.

k번째 층의 비조건부 우도 함수 식(2.2)에서 장애모수 α_k의 충분통계량 $S_k=\sum_j Y_{1jk}$의 값을 고정시켜 장애모수를 제거하면 β에 관한 조건부 우도 함수는 식(2.3)과 같다.

$$\Pr(Y_{11k}=y_{11k},\dots,Y_{1Jk}=y_{1Jk}|S_k=s_k)=\frac{c_k(s_k,\boldsymbol{y}_{1k})\exp(\sum_{j=1}^{J}y_{1jk}x_j\beta)}{\sum_{\boldsymbol{u}\in\Omega_{1k}}c_k(s_k,\boldsymbol{u})\exp(\sum_{j=1}^{J}u_{1jk}x_j\beta)} \tag{2.3}$$

여기서 $c_k(s_k,\boldsymbol{y}_{1k})=\prod_{j=1}^{J}\binom{n_{+jk}}{y_{1jk}}$이고 Ω_{1k}는 k번째 층에서 $\sum_j Y_{1jk}=s_k$을 만족하는 $2\times J$ 분할표의 집합이다.

또한 k번째 층에서 β의 충분통계량은 $T_k=\sum_j Y_{1jk}x_j$이고 k번째 층에서 $S_k=s_k$가 주어졌을 때 T_k의 조건부 분포는 식(2.4)와 같다(Hirji와 Vollset, 1994).

$$\Pr(T_k=t|S_k=s_k)=\frac{c_k(s_k,t)\exp(t\beta)}{\sum_{u\in\Omega_k}c_k(s_k,u)\exp(u\beta)} \tag{2.4}$$

여기서 $c_k(s_k,t)=\sum c_k(s_k,\boldsymbol{y}_{1k})=\sum\prod_{j=1}^{J}\binom{n_{+jk}}{y_{1jk}}$이며 $c_k(s_k,t)$에서 합은 $0\le Y_{1jk}\le n_{+jk}$, $\sum_j Y_{1jk}=s_k$, 그리고 $\sum_j Y_{1jk}x_j=t$를 만족하는 모든 정수 수열$\{Y_{1jk}, j=1,\dots,J\}$에 대해 이루어진다. 즉, $\sum_{j=1}^{J}Y_{1jk}=s_k$를 만족하는 $2\times J$ 분할표 중에서 $\sum_{j=1}^{J}Y_{1jk}x_j$의 값이 같은 t를 만족하는 $2\times J$ 분할표의 $c_k(s_k,\boldsymbol{Y}_{1k})$의 합이다. 그리고 Ω_k는 k번째 층에서 $S_k=s_k$가 주어졌을 때 T_k의 모든 가능한 값의 집합이다.

β에 대한 충분통계량은 $T=\sum_k T_k$이므로 모든 층을 통합하여 T의 조건부 분포는 식(2.5)와 같다.

$$\Pr(T=t \mid S_k=s_k,\ k=1,\dots,K)=\frac{c(t)\exp(t\beta)}{\sum_{u\in\Omega} c(u)\exp(u\beta)} \tag{2.5}$$

여기서 $c(t)=\sum_{\sum t_k=t}\prod_{k=1}^{K} c_k(s_k,t_k)$ 이고, Ω는 $\{S_k=s_k,\ k=1,\dots,K\}$가 주어졌을 때 T의 모든 가능한 값의 집합이다. 즉, $c(t)$는 각 층에서 t_k를 만족하는 모든 경우의 $c_k(s_k,t_k)$에 대해 이들 값들을 모든 층에 대해 곱하면 $T=\sum_k T_k$의 총 경우의 수가 된다. 이들을 T의 값에 대해 묶은 것이 $c(t)$이다. 그리고 β에 대한 정확 추론은 β에 대한 충분 통계량을 사용한다.

3. 수정된 P-값

로지스틱 회귀모형에서 관측 자료를 $2\times J\times K$ 분할표로 나타내어 $Y=\{y_{ijk},\ i=1,2,\ j=1,\dots,J,\ k=1,\dots,K\}$는 관찰된 칸 도수라 하며 이산형 분포를 갖는 검정통계량 T에 대하여 t_0을 T의 관찰값이라 하자. T의 분포가 이산형인 경우 유의수준 α인 검정에서 1종 오류 확률은 명목수준 α보다 작다. 특히 조건부 분포를 사용할 경우 이산성의 정도가 더 심해지므로 귀무가설하에서 1종 오류 확률은 α보다 더 작아지게 된다. 따라서 이산형의 정확분포를 사용한 정확검정은 보수적이 된다. 가설검정에서 보수성은 신뢰구간 추정에도 영향을 미치므로 $100(1-\alpha)\%$ 정확 신뢰구간인 경우 모수를 포함하는 포함확률은 최소한 $100(1-\alpha)\%$이며 이산성이 강할수록 포함확률은 $100(1-\alpha)\%$보다 더 커지게 된다. 일반적으로 T의 값이 클 때 귀무가설을 기각할 경우 단측(one-sided)검정의 일반적인(ordinary) 정확 P-값은 다음과 같이 정의된다.

$$P=P_{H_0}(T\geq t_0) \tag{3.1}$$

검정통계량의 분포가 이산적이기 때문에 만들어진 보수성을 해결하기 위하여 기각역의 임계점 $T=t_0$에서 랜덤화(randomization)를 하던가 또는 $T=t_0$의 확률의 절반을 P-값에 포함시키는 mid-P가 있으나 이들 방법은 정확검정이 아니다. 랜덤화에 의존하지 않고 보수성을 줄이기 위하여 이산성의 정도를 줄인 분포를 갖는 수정된 정확 P-값(modified exact P-value)을 활용할 수 있다(Cohen 과 Sackrowitz, 1992; Kim과 Agresti, 1995). Kim과 Agresti(1995)는 임계점인 $T=t_0$에서 $T=t_0$을 만족하는 분할표의 모든 확률을 P-값에 포함시키는 대신 $T=t_0$을 만족하는 랜덤 분할표 중에서 랜덤 분할표의 표확률이 관측된 표확률보다 작거나 같은 랜덤 분할표의 확률만을 P-값에 포함시킨다. 이 수정된 P-값을 사용한 검정은 정확검정이며 신뢰구간은 정확

신뢰구간이다.

로지스틱 회귀모형에서 조건부 검정에 의해 장애모수의 충분통계량을 관측된 값으로 고정시켜 이를 만족하는 모든 분할표의 집합을 Γ라 하면 다음과 같다.

$$\Gamma=\{\{z_{ijk}\};\, z_{+jk}=n_{+jk},\, z_{1+k}=n_{1+k},\, i=1,2,\, j=1,\ldots,J,\, k=1,\ldots,K\} \tag{3.2}$$

이런 분할표의 집합 Γ가 준거집합(reference set)이라 할 때 다음의 집합 B를 고려하자.

$$B=\{\, Z:\, Z\in\Gamma,\ T=t_0,\ P(\, Z)\le P(\, Y)\} \tag{3.3}$$

여기서 집합 B는 식(3.2)를 만족하는 준거집합에 속하는 랜덤 분할표 $Z=\{z_{ijk},\, i=1,2,\, j=1,\ldots,J,\, k=1,\ldots,K\}$중에서 $T=t_0$을 만족하며 그 발생확률이 관측자료 Y의 발생확률보다 작거나 같은 랜덤 분할표의 모임이며 확률은 귀무가설하에서 계산된다. 그러면 수정된 정확 P-값은 다음과 같이 정의한다.

$$P^*=P_{H_0}(\, T>t_0)+P_{H_0}(B) \tag{3.4}$$

따라서 수정된 P-값은 β의 충분통계량 T의 분포가 T의 관측값 t_0보다 클 때의 확률과 $T=t_0$을 만족하는 랜덤 분할표 중에서 관측된 분할표보다 더 귀무가설을 기각하게 하는 랜덤 분할표의 표확률을 합한 것으로 정의된다. 수정된 P-값은 검정통계량 T의 분포보다 표본공간을 더 세분한다. 즉, 일차(primary) 검정통계량 T의 고정된 값에서 이차(secondary) 통계량인 표확률을 이용하여 T의 분포를 좀더 세밀하게 나눈다. T의 주어진 값에 대해 귀무가설 분포에서 Y의 발생 확률인 표확률이 작을수록 귀무가설을 기각하는 정도가 더 강해진다. 따라서 수정된 P-값은 일반적인 검정통계량 T의 분포만 고려하는 것이 아니라 $T=t_0$에서 우리가 관찰한 자료가 귀무가설을 기각하는 정도까지 고려하게 된다.

$0<\alpha<1$인 임의의 α에 대하여 $P_{H_0}(P^*\le\alpha)\le\alpha$를 만족하기 때문에 이산형 분포를 갖는 임의의 검정통계량에 대해 P^*를 계산할 수 있다. 수정된 P-값은 일반적인 P-값보다 클 수 없으므로 수정된 P-값을 기초로 하는 검정은 덜 보수적이다.

$2\times J\times K$ 분할표에서 $\{Y_{1jk},\, j=1,\ldots,J\}$는 시행횟수가 열주변합인 J개의 독립적인 이항 확률변수로 생각할 수 있다. 식(2.2)의 비조건부 확률분포에서 Y_{1k}의 행주변합($\{\, n_{i+k}\}$)을 조건부로하는 조건부 우도 함수 식(2.3)을 유도하였고 이 Y의 발생 확률인 표확률(table probability)은 식(3.5)와 같이 계산된다.

$$\Pr(\ Y=\ \boldsymbol{y}\,|\,n_{1+k}, k=1,\ldots,K) = \prod_{k=1}^{K} \frac{\prod_{j=1}^{J}\binom{n_{+jk}}{y_{1jk}} \exp(\sum_j y_{1jk}x_j\beta)}{\sum_{r\in R_k}\prod_{j=1}^{J}\binom{n_{+jk}}{r_{1jk}} \exp(\sum_j r_{1jk}x_j\beta)} \tag{3.5}$$

여기서 R_k은 k번째 분할표에서 생성된 준거집합이다.

로지스틱 회귀모형에서 조건부 정확 분포를 이용한 정확검정을 다음의 예제를 이용하여 실시해 보고자 한다. [표 3.1]과 [표 3.2]는 층변수 K, 이진 반응변수 Y와 독립변수 X에 대한 로지스틱 회귀모형의 자료를 분할표로 나타낸 것이다.

층이 있는 자료에서 β에 대한 충분통계량의 조건부 분포를 구하기 위해서는 우선 각 $2\times J$ 분할표에서 열의 주변합과 α_k의 충분통계량 $s_k = \sum_{j=1}^{J} y_{1jk}$이 관측된 분할표에서 얻어진 값을 만족하는 모든 가능한 랜덤 분할표를 생성한다. 그리고 각 층에서 생성한 랜덤 분할표에서 $c_k(s_k, t)$를 구하여 T_k의 분포를 구한 후 $T = \sum_k T_k$를 구한다. 다음으로 각 층에서 얻어진 $c_k(s_k, t)$를 사용하여 T의 조건부 분포 식(2.5)를 구한다.

귀무가설 $H_0: \beta = 0$을 검정하기 위해 조건부 정확 분포를 이용하여 단측검정을 할 경우 일반적인 P-값과 수정된 P-값인 P^*을 비교하면 [표 3.1]에서 $P=0.815$, $P^*=0.509$ 그리고 [표 3.2]에서 $P=0.129$, $P^*=0.052$로 수정된 P^*-값이 정확 검정의 보수성을 많이 줄일 수 있음을 알 수 있다. [표 3.2]에서 $\alpha=0.05$인 일반적인 정확검정은 귀무가설을 채택하게 하나 수정된 정확검정은 귀무가설을 기각할 수 있을 정도로 $\boldsymbol{Y}$의 발생확률은 매우 작다.

수정된 P^*-값이 보수성을 얼마나 줄이는지를 알아보기 위하여 일반적인 P-값과 수정된 P^*-값의 기대값을 각각 구하여 연속형인 경우의 P-값의 기대값인 1/2과 비교하였다. [표 3.1]의 조건부 분포를 사용한 경우 $E_{H_0}P=0.625$, $E_{H_0}P^*=0.562$ 이며 마찬가지로 [표 3.2]인 경우 $E_{H_0}P=0.604$, $E_{H_0}P^*=0.531$ 을 얻었다. 수정된 P^*-값이 일반적인 P-값보다 검정 통계량이 연속형인 경우의 P-값의 기대값인 0.5에 가까운 값을 가지므로 보수성의 정도가 많이 줄어들었음을 알 수 있다.

다음으로 발생 가능한 모든 P-값의 개수를 알아보면 [표 3.1]에서 일반적인 P는 8개이며 수정된 P^*는 20개이다. 또한 [표 3.2]인 경우 일반적인 P는 7개이며 수정된 P^*는 32개다. 이와 같

[표 3.1] 예제1 : 2×2×2 분할표

K	Y	$X=0$	$X=1$
1	0	5	1
	1	0	3
2	0	1	5
	1	3	2

[표 3.2] 예제2 : 2×3×2 분할표

K	Y	$X=0$	$X=1$	$X=2$
1	0	0	1	1
	1	1	2	0
2	0	1	0	2
	1	1	3	0

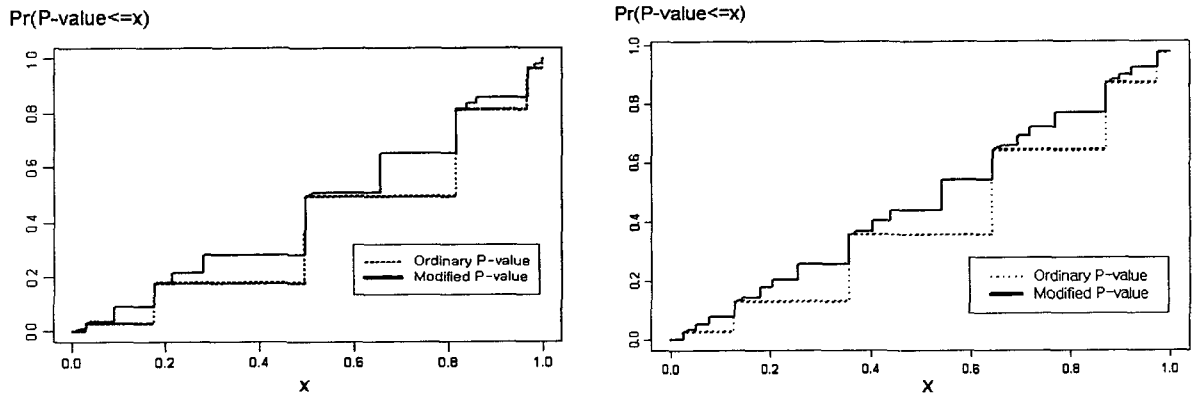
[표 3.3] 표 3.1에서 $T=5$인 경우 P와 P^*의 비교

$T=t_0$	표의개수	일반적인 P-값	수정된 P-값(P^*)	표확률
5	1	0.815	0.500	0.005
	2		0.653	0.144
	3		0.815	0.162
	4		0.509*	0.009*

이 수정된 P^*를 이용한 방법은 P의 분포를 더 세분한다.

[표 3.3]은 [표 3.1]에서 $T=5$인 경우 P와 P^*를 비교한 것이다. 관찰된 자료에서 $P=0.815$이나 수정된 P^*-값은 $T=5$를 만족하는 $\boldsymbol{Y}$의 정수 수열이 4가지 발생하며 $\boldsymbol{Y}$의 발생확률에 따라 0.500, 0.509, 0.653, 0.815 네 개의 값을 갖는다. 그리고 $T=5$를 만족하는 4가지 $\boldsymbol{Y}$의 정수 수열 중 가장 큰 발생확률 0.162를 갖는 $\boldsymbol{Y}$인 경우 P와 P^*는 같은 값을 갖는다.

임의의 검정통계량에 대하여 $0<\alpha<1$일 때 $P_{H_0}(P\text{-값} \le \alpha) \le \alpha$을 만족한다. 다음은 이산형 분포를 갖는 일반적인 P와 수정된 P^*의 누적분포함수를 각각 구하여 이들을 연속형 균일분포의 누적분포함수와 비교한다. [그림 3.1]은 [표 3.1]과 [표 3.2]의 고정된 행합계와 열합계를 만족하는 랜덤 분할표에 대하여 귀무가설하에서 일반적인 P와 수정된 P^*의 누적분포함수를 각각 나타낸다. 수정된 P^*-값은 T의 분포뿐만 아니라 $T=t_0$을 만족하는 $\boldsymbol{Y}$의 정수 수열에 대해 $\boldsymbol{Y}$의 발생확률을 고려하여 계산한 것이므로 수정된 P^*의 누적분포함수가 일반적인 P의 누적분포함수보다 연속형인 경우의 균일분포의 분포함수에 더 가까운 것을 알 수 있다. 독립변수가 더 많은 값을 갖는 [표 3.2]인 경우 P^*의 누적분포함수는 더 세분되어 연속형 균일분포의 분포함수에 더 가까워지는 것을 볼 수 있다.



(a) [표 3.1] (b) [표 3.2]

[그림 3.1] 귀무가설하에서 정확 P-값들의 누적분포함수

4. 수정된 정확 신뢰구간

신뢰구간과 유의성 검정의 관련성에 의하여 T의 정확분포를 사용한 정확검정은 1종 오류 확률이 유의수준 α보다 작으므로 $100(1-\alpha)\%$ 정확 신뢰구간에서 모수를 포함하는 실제 포함확률은 명목수준 $100(1-\alpha)\%$보다 커진다. 수정된 P^*-값이 유의성 검정의 보수성을 줄인 것처럼 수정된 P^*-값을 사용하면 실제 포함확률이 명목수준에 더 가까운 신뢰구간을 구축할 수 있다. 두 개의 단측검정 방법을 역으로 이용하는 $100(1-\alpha)\%$ 정확 신뢰구간은 양쪽의 단측검정에서 꼬리부분의 확률이 $\alpha/2$를 갖도록 결정된다. 회귀모수 β에 대한 검정통계량 T의 조건부 확률밀도함수 식(2.5)를 $P(t;\beta)$이라 하자.

로지스틱 회귀모형에서 단측 P-값을 이용한 $100(1-\alpha)\%$의 일반적인 정확 신뢰구간 (β_L, β_U)은 다음과 같이 정의된다. $t_{\min} \le t_0 \le t_{\max}$에 대하여 신뢰구간의 하한과 상한은

$$\begin{aligned} \beta_L &: \sum_{t \ge t_0} P(t;\beta_L) = \alpha/2 \\ \beta_U &: \sum_{t \le t_0} P(t;\beta_U) = \alpha/2 \end{aligned} \tag{4.1}$$

을 만족하며, $t_0 = t_{\min}$이면 $\beta_L = -\infty$ 그리고 $t_0 = t_{\max}$이면 $\beta_U = +\infty$이다.

수정된 P^*-값을 이용하면 일반적인 P-값을 이용한 정확 신뢰구간보다 신뢰구간의 길이가 작으면서 명목수준을 만족하는 덜 보수적인 신뢰구간을 얻을 수 있다. 수정된 정확 신뢰구간은 다음의 함수를 이용한다.

$$\begin{aligned} P_1^*(\beta) &= \sum_{t > t_0} P(t;\beta) + P[B(\beta);\beta] \\ P_2^*(\beta) &= \sum_{t < t_0} P(t;\beta) + P[B(\beta);\beta] \end{aligned} \tag{4.2}$$

여기서 $B(\beta) = \{Z : Z \in \Gamma, T = t_0, P(Z;\beta) \le P(Y;\beta)\}$이다. β의 하한 β_L^*는 $P_1^*(\beta) \ge \frac{\alpha}{2}$을 만족하는 β의 최소값이며, β의 상한 β_U^*는 $P_2^*(\beta) \ge \frac{\alpha}{2}$를 만족하는 β의 최대값이다. 수정된 정확 신뢰구간은 일반적인 정확 신뢰구간과 같거나 정확 신뢰구간에 포함된다.

[표 4.1]은 [표 3.1]에서 $T=5$인 경우 P와 P^*를 사용한 정확 신뢰구간을 비교한 것이다. 수정된 정확 신뢰구간은 일반적인 정확 신뢰구간과 같거나 포함됨을 알 수 있다. 또한 [표 3.3]에서 T의 관측값인 $T=5$를 만족하는 랜덤 분할표 4개 중 가장 큰 표확률 0.162를 갖는 랜덤 분할표는 $P = P^*$가 된 것과 마찬가지로 이 경우는 수정된 정확 신뢰구간과 일반적인 정확 신뢰구간이 일치한다.

[표 4.1] 표 3.1에서 $T=5$인 경우 P와 P^*를 사용한 β의 95% 정확 신뢰구간

$T=t_0$	β의 하한값	β의 상한값
5	-1.734	2.619
신뢰구간의 길이		4.353

(a) 일반적인 P

$T=t_0$	표의개수	β의 하한값	β의 상한값
5	1	-0.948	1.740
	2	-1.485	2.338
	3	-1.734	2.619
	4	-0.995*	1.791*
신뢰구간의 평균길이		3.413	

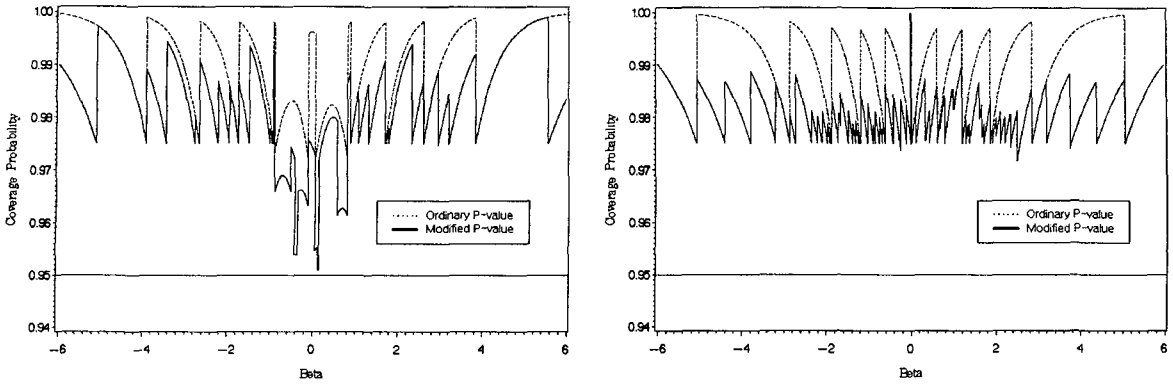
(b) 수정된 P^*

[표 3.1]과 [표 3.2]의 고정된 행합계와 열합계를 만족하는 랜덤 분할표에 대하여 β의 95% 신뢰구간의 평균길이(expected length)를 구해 보았다. [표 3.1]에서 일반적인 P를 이용한 모든 신뢰구간의 평균길이는 5.207이며 수정된 P^*를 이용한 모든 신뢰구간의 평균길이는 4.099이다. 마찬가지로 [표 3.2]인 경우 이들 두 P를 이용한 신뢰구간의 평균길이는 각각 4.636과 3.476이다. 일반적인 P를 이용한 신뢰구간보다 수정된 P^*를 이용한 신뢰구간의 평균길이가 더 작아지는 것을 알 수 있다. [표 4.2]는 [표 3.1]과 [표 3.2]에 대하여 β에 대한 95% 신뢰구간을 나타낸다.

[표 3.1]과 [표 3.2]의 행합계와 열합계가 고정된 $\{Z_{ijk}\}$ 조건부 분포에서 [그림 4.1]은 Γ에 속하는 모든 랜덤 분할표에 대하여 일반적인 P 또는 수정된 P^*를 사용하여 β의 95% 신뢰구간을 각각 만든 후 임의의 β를 실제로 포함하는 포함 확률을 나타낸다. [표 3.1]과 [표 3.2]의 행합계와 열합계가 주어졌을 때 정확분포에 근거한 β에 대한 95% 신뢰구간에서 β가 포함될 실제 확률은 β의 전 구간에서 0.95이상이어야 한다. 일반적인 P를 이용한 경우 β의 전 구간에서 포함확률이 0.975이상이며 1까지 근접하므로 매우 보수적이다. 그러나 수정된 P^*를 이용한 포함확률은 일반적인 P-값을 이용할 때보다 포함확률이 0.975에 가까워지며 β의 절대값이 작은 값에서 0.95에 근접할 수 있다. 따라서 수정된 P^*-값을 사용하면 검정이나 신뢰구간의 보수성을 줄일 수 있다.

[표 4.2] β에 대한 95% 정확 신뢰구간

정확 신뢰구간	표 3.1	표 3.2
일반적인 (ordinary)	(−1.734, 2.619)	(−5.064, 0.611)
수정된 (modified)	(−0.995, 1.791)	(−3.786, 0.424)



(a) [표 3.1] (b) [표 3.2]

[그림 4.1] 표 3.1과 표 3.2의 조건부 분포에서 β의 95% 신뢰구간에 대한 포함확률

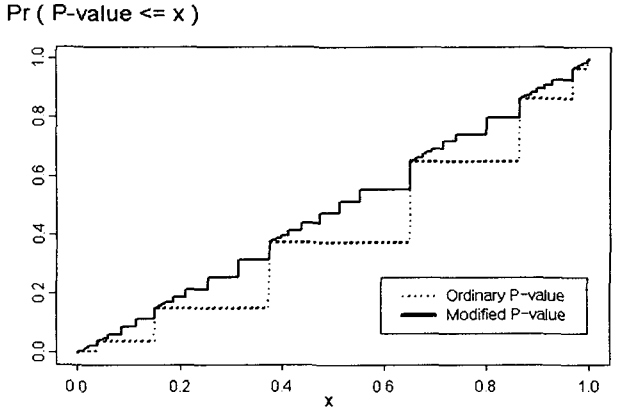
5. 실제예제

[표 5.1]은 Bishop et al.(1975)에서 인용한 자료로 세 그룹의 에스키모인에 대해 41세에서 50세까지 에스키모인의 하악골 전방 부위의 이상돌출(dental disorder torus mandibularis) 발생수를 조사한 결과이다.

이 자료를 이용하여 로지스틱 회귀모형에서 향상된 정확추론을 보이려고 한다. $\{\alpha_k,\ k=1,2,3\}$의 충분통계량 $s_1=16$, $s_2=8$과 $s_3=6$을 만족하는 모든 가능한 분할표를 이용하여 T의 조건부 분포를 구한다. 관찰된 [표 5.1]에서 β의 충분통계량 $T=18$이며 $P=0.349$, $P^*=0.147$이다. 일반적인 P-값과 수정된 P-값의 기대값은 $E_{H_0}P=0.599$, $E_{H_0}P^*=0.520$으로 수정된 P^*-값의 기대값이 0.5에 근접한다. [표 5.1]에 대한 P와 P^*의 개수는 각각 11개와 75개로 많은 차이를 보이며 누적분포함수인 [그림 5.1]을 보면 P^*를 사용할 경우 연속형 균일분포의 분포함수에 거의 근접한다. 일반적으로 층의 개수나 열의 개수가 많아질수록 P^*의 개수가 더 많아지므로 보수성이 줄어든다.

[표 5.1] 2×2×3 분할표

K	Y	$X=0$	$X=1$
1	0	4	0
	1	6	10
2	0	0	2
	1	4	4
3	0	2	5
	1	2	4



[그림 5.1] 표 5.1에서 귀무가설하에서 정확 P-값들의 누적분포함수

[표 5.2] 표 5.1에서 β에 대한 95% 신뢰구간

정확 신뢰구간		점근적 신뢰구간	
일반적인 정확	수정된 정확	비조건부 점근	조건부 점근
(−1.092, 2.246)	(−0.641, 1.715)	(−0.890, 2.017)	(−0.876, 1.922)

[표 5.2]는 β에 대한 점근신뢰구간과 정확신뢰구간을 비교한다. $T_0 = 18$에서 일반적인 P를 이용한 신뢰구간의 길이는 3.338이고 수정된 P^*를 이용한 12개 신뢰구간의 평균길이는 2.610이다. 로지스틱 회귀모형에서 비조건부 최우추정량을 이용한 점근적 신뢰구간과 조건부 최우추정량을 이용한 점근적 신뢰구간은 서로 비슷하나 조건부 점근 신뢰구간이 비조건부 점근 신뢰구간에 포함된다. 이 자료에서는 일반적인 정확신뢰구간은 점근적 신뢰구간을 포함하며 점근적 신뢰구간은 수정된 정확신뢰구간을 포함한다.

6. 결론

로지스틱 회귀모형에서 0이 많이 발생하는 희박자료이거나 또는 소표본인 경우 대표본 근사방법은 바람직하지 않다. 따라서 이 경우 검정통계량의 조건부 정확 분포를 사용하여 조건부 준거집합을 기초로 한 정확검정이 선호된다. 그러나 조건부 정확 검정은 1종 오류 확률이 유의수준보다 작거나 같다는 것을 보장한다 하더라도 검정통계량의 분포가 이산형이므로 이 검정방법은 매우 보수적이다. 검정통계량의 이산성에 따른 보수성을 해결하기 위하여 로지스틱 회귀모형에서 두 개의 검정통계량을 동시에 고려하는 수정된 P^*를 이용한 방법을 제안하였다. 이것은 검정통계량의 정확분포를 준거집합의 표확률을 이용하여 세분화하는 방법이며 정확추론을 제공한다.

수정된 P^*를 이용한 방법은 임의의 랜덤화 없이 주어진 1종 오류 확률을 보장해 주면서 일반적인 P를 기초로 한 방법보다 보수성을 줄일 수 있는 장점이 있다. 보수성을 줄이기 위해 mid-P의 사용은 정확추론을 제공하지 못한다.

로지스틱 회귀모형에서 수정된 P^*를 이용한 검정방법이 보수성을 줄이는 것을 보이기 위해 로지스틱 회귀모형의 자료를 분할표로 나타낸 $2\times2\times2$와 $2\times3\times2$ 그리고 $2\times2\times3$ 분할표에 대하여 수정된 P^*를 사용한 정확추론을 실시해 보았다. [표 3.1]과 [표 3.2]에 대하여 귀무가설하에서 수정된 P^*-값의 기대값 (0.562 , 0.531)이 일반적인 P-값의 기대값 (0.625 , 0.604)보다 검정통계량의 분포가 연속형인 경우의 기대값 0.5에 근사하며, 누적분포함수 그래프 [그림 3.1]을 보면 수정된 P^*-값의 누적 분포 함수가 일반적인 P-값의 누적 분포 함수보다 연속형 균일분포의 분포함수에 더 가깝게 나타난 것을 알 수 있었다.

또한 [표 3.1]과 [표 3.2]를 이용하여 β의 95% 평균 신뢰구간의 길이를 비교해 보면 수정된 P^*를 사용한 경우 신뢰구간의 평균길이 (4.099 , 3.476)가 일반적인 P를 이용했을 때의 평균길이 (5.207 , 4.636)보다 더 작게 나타났다. 회귀모수 β에 대한 신뢰구간에서 포함확률 [그림 4.1]은 수정된 P^*를 이용한 신뢰구간의 포함확률이 일반적인 P를 이용한 신뢰구간보다 명목수준에 더 가깝게 나타남을 알 수 있었다.

검정통계량의 분포를 구성하는 준거집합의 $\boldsymbol{Y}$의 정수 수열의 개수가 많을수록 수정된 P^*를 이용한 통계적 추론은 로지스틱 회귀모형에서 이산성에 기인한 검정이나 신뢰구간의 보수성을 많이 줄일 수 있다. 따라서 $2\times J\times K$ 분할표에서 소표본이며 K가 클 때 수정된 P^*를 사용한 정확추론은 일반적인 추론보다 덜 보수적이 되어 더욱 더 향상된 추론을 제공하게 된다. 우리는 제2의 검정통계량으로 로지스틱 회귀모형에서의 표확률을 이용하였다. 제2의 검정통계량으로 점수(score)통계량을 사용하면 점수통계량은 β의 충분통계량에 의존하므로 $T=\sum T_k$와 같은 표본공간의 분할을 만들어 제2의 검정통계량으로 점수통계량은 적절하지 않다. 또한 T와 관련된 성격으로 $2\times J\times K$ 분할표에서 순서(ordering)와 관련된 동질적인 선형 대 선형모형(homogeneous linear-by-linear association model)을 고려할 때 관심있는 모수의 충분통계량도 T와 유사한 형태이다.

본 논문에서의 계산은 포트란(Fortran) 77을 이용하여 작성하였으며 분할표에서 T의 분포에 대한 이진 반응변수의 순열 분포를 이용한 확률의 계산은 Hirji와 Vollset(1994)이 작성한 프로그램을 이용하였다. 로지스틱 회귀모형에서 양측(two-sided) 정확 P-값과 수정된 양측 정확 P^*-값을 이용한 정확추론은 앞으로의 연구과제이다.

REFERENCE

[1] Agresti, A. and Min, Y. (2001). On small-sample confidence intervals for parameters in discrete distributions, *Biometrics*, **57**, 963-971.

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

[3] Cohen, A. and Sackrowitz, H.B. (1992). An evaluation of some tests of trend in contingency tables, *Journal of the American Statistical Association*, **87**, 470-475.

[4] Cohen, G. and Yang, S. (1994). Mid-P confidence intervals for the Poisson expectation, *Statistics in Medicine,* **13**, 2189-2203.

[5] Day, N.E. and Byar, D.P. (1979). Testing hypotheses in case-control studies-Equivalence of Mantel-Haenszel statistics and logit score tests, *Biometrics*, **35**, 623-630.

[6] Hirji, K.F. (1991). A comparison of exact, mid-P, and score tests for matched case-control studies, *Biometrics,* **47**, 487-496.

[7] Hirji, K.F., Mehta, C.R. and Patel, N.R. (1987). Computing distributions for exact logistic regression, *Journal of the American Statistical Association*, **82**, 1110-1117.

[8] Hirji, K.F., Mehta, C.R. and Patel, N.R. (1998). Exact inference for matched case-control studies, *Biometrics*, **44**, 803-814.

[9] Hirji, K.F. and Vollset, S.E. (1994). Computing exact distributions for several ordered $2 \times K$ tables, *Applied Statistics*, **43**, 541-548.

[10] Kim, D. (1998). Improved mid P-value method for statistical inference in three-way contingency tables, *The Korean Communications in Statistics*, **5**, 905-926.

[11] Kim, D. and Agresti, A. (1995). Improved exact inference about conditional association in three-way contingency tables, *Journal of the American Statistical Association*, **90**, 632-639.

[12] Lancaster, H.O. (1961). Significance tests in discrete distributions, *Journal of the American Statistical Association*, **56**, 223-234.

[13] Mehta, C.R. and Patel, N.R. (1995). Exact logistic regression : theory and examples, *Statistics in Medicine*, **14**, 2143-2160.

[2003년 5월 접수, 2003년 8월 채택]