응용통계연구 제 **13**권 2호, *2000*년, *pp. 237-245*

2단계 사례-대조자료를 위한 로지스틱 회귀모형의 추론

신미영[1] 신은순[2]

요 약

이 논문에서는 2단계 계획 하에서의 사례-대조 자료를 로지스틱 회귀 모형에 적합시키고 WESML 방법으로 모수를 추정하며 추정량의 점근분포를 찾는다. 또한 WESML 방법과 CML 방법으로 얻은 모수의 추정량과 표준오차를 실제 자료를 이용하여 비교한다.

주요용어: 로지스틱, 2단계 사례-대조자료, WESML.

1. 서론

이산형 로지스틱 회귀분석은 반응변수와 설명변수들 간의 관계를 모형화하여 분석하는 통계기법이다. 로지스틱 회귀모형으로 분석하는 자료는 크게 전향성(Prospective) 연구 자료와 사례-대조(Case-Control) 연구 자료로 나눌 수 있다. 관심있는 반응변수에 영향을 미치는 설명변수 E에 노출된 집단과 그렇지 않은 집단을 대상으로 일정 시간이 지난 후 주어진 E값에 대응되는 반응변수 Y의 값을 측정하여 얻은 자료를 전향성 연구자료라 하며 반응변수의 결과에 따라 사례군($Y=1$)과 대조군($Y=0$)으로 분류한 후 각 집단으로부터 표본을 추출하여 대응되는 설명변수를 관측하여 얻은 자료를 사례-대조 연구 자료라 한다. 전향성 연구에서는 무엇보다도 연구 기간이 장기화될 수 있다는 단점 때문에 사례-대조 연구 방법에 많은 관심을 갖게된다. 사례-대조 로지스틱 회귀모형에서 모수의 추정과 가설검정에 관한 연구는 신미영(1994), Prentice and Pyke (1979)와 Scott and Wild(1989,1991)를 참고하기 바란다.

의학 관련 연구 분야에서는 전향성/사례-대조로 표현되었던 자료구조가, 경제, 경영학 등의 연구 분야에서는 외생(Exogenous)/선택-근거(Choice-Based) 연구라는 표현으로 자료 수집 방법을 명시하기도 한다(Manski and Lerman(1977)과 Manski and McFadden(1981) 참조). 자료 수집의 형태로 보아 외생은 전향성에, 선택-근거는 사례-대조에 대응되는 자료 수집 방법이다.

어느 연구에서나 자료의 크기가 충분히 크다면, 그렇지 않은 경우에 비하여 좋은 연구 결과를 기대할 수 있겠으나, 연구비 또는 연구 인력 부족 등의 이유로 언제나 원하는 정보를 갖춘 충분한 자료를 수집하기가 쉽지 않다. 이에 대한 보완책으로 Breslow and Cain(1988), White(1982)와 Wild(1991)등은 2단계 계획(Two-Stage Design)으로 자료를 추출하여 모수

1) (420-743) 경기도 부천시 원미구, 가톨릭대학교 자연과학부, 부교수
E-mail: myshin@www.cuk.ac.kr
2) (100-715) 서울시 중구 필동, 동국대학교 통계학과, 박사과정
E-mail: shin_eunsoon@hotmail.com

를 추정하는 연구를 하였다. 2단계 계획 연구란 1차 단계에서는 반응변수와 반응변수에 영향을 줄 것으로 생각되는 주 설명변수에 관한 충분히 큰 자료를 수집하고, 1차 단계에서 수집된 자료로부터 2차 단계의 표본을 수집하여 보다 세부적인 공변수의 정보를 얻어 연구하는 방법이다. Breslow와 Cain(1988)은 2단계 계획 하에서 사례-대조 자료를 이용하여 로지스틱 모형의 모수를 조건부 최우추정(CML) 방법으로 추정한 후 추정량의 점근 분포를 유도하였다.

이 논문에서는 2단계 사례-대조 연구자료 분석 방법으로 로지스틱 모형을 사용하고 모수를 가중외생추출 최우추정(WESML) 방법으로 추정하며 예제를 통하여 Breslow와 Cain (1988)이 제안한 CML 방법으로 추정한 모수와 그 결과를 비교해 보기로 한다.

2. 가정 및 기호

이 장에서는 2단계 계획 하에서 추출된 사례-대조 자료를 위한 로지스틱 회귀모형에서의 추정량을 구하기 위해 필요한 가정과 기호들을 설명한다. 1차 단계에서는 사례군($Y=1$)과 대조군($Y=0$)으로 분류된 모집단으로부터 각각 N_1, N_0개의 표본을 추출한 후 설명변수 $E=k$로 분류된 자료의 크기를 $N_{ik}(i=0,1,k=1,\cdots,K)$라 하자. 2차 단계에서는 1차 단계에서 $Y=i$, $E=k$로 분류된 자료 N_{ik}개 중에서 표본을 추출하여 이산형 공변수 Z에 관한 세부 정보를 얻는다. $n_{ij}(i=0,1,j=1,\cdots,J)$는 $Y=i$이며 주변수 수준과 공변수 수준의 조합이 j번째인 자료의 크기라 한다. n_i와 n은 2차단계에서 각각 $Y=i$인 자료와 총 자료의 크기를 나타낸다.

모집단에서 $Y=1$일 확률은 다음과 같은 로지스틱 회귀모형으로 설명된다고 가정한다.

$$pr(Y=1|x_j)=\frac{exp(\beta' x_j)}{1+exp(\beta' x_j)}$$

여기서 $\beta'=(\beta_0\ \beta_1\ \beta_2)$이며, X_j는 상수항 정보와 주변수 수준과 공변수 수준의 조합이 j번째인 정보를 갖는 벡터이다.

이 논문에서 필요한 기호와 가정을 다음과 같이 정의한다.

기호

$$P_{ij}=pr(Y=i|X_j=x_j)$$

$$p_j=pr(X_j=x_j),\ p=(p_1,\cdots,p_J)$$

$$Q_{ji}=pr(X_j=x_j|Y=i)$$

$$q_i=pr(Y=i),\ q=(q_0,q_1)$$

가정 2.1 $\lim_{n\to\infty} n_i/n=\mu_i$ 인 μ_i가 존재한다.

가정 2.2 $\lim_{n\to\infty} n/N=r$ 인 r이 존재한다.

3. 2단계 사례-대조 연구에서 WESML 추정

이 장에서는 Manski와 Lerman(1977)에 의해 제안된 WESML 추정 방법을 2장에서 언급한 2단계 계획 하에서 추출된 사례-대조 자료에 적용하여 로지스틱 회귀모형에서의 추정량을 구하고 추정량의 점근 분포를 유도한다.

전향성연구 자료로 부터 로지스틱 모형의 모수는 다음 로그우도함수를 최대화하여 얻을 수 있다.

$$logL(\beta) \propto \frac{1}{n}\sum_i\sum_j n_{ij}\, logP_{ij} \tag{3.1}$$

사례-대조 연구에서는 위의 로그우도함수 (3.1)식에 가중치 w_i를 준 가중외생추출 우도함수 (3.2)를 최대화하여 추정량을 구하는 방법을 WESML 방법이라 한다.

$$log_w L(\beta) = \frac{1}{n}\sum_i w_i \sum_j n_{ij} logP_{ij} \tag{3.2}$$

Manski와 Lerman(1977)은 q_i가 알려진 경우 가중치로 q_i/μ_i를 사용하였으며 Cosslett (1981)은 가중치 $\frac{q_i}{n_i/n}$을 사용하였을 때 더 유효한 추정량을 구한다는 것을 보였다.

본 논문에서는 1차단계 자료로부터 q_i를 추정하며 2차단계 자료로부터 얻은 정보를 이용하여 가중치 $w_i = \frac{q_i}{n_i/n}$ 를 사용한 가중외생추출 우도함수 (3.3)식을 최대화하는 WESML 추정량을 구하고자 한다.

$$\begin{aligned} W(\beta, q) &= \frac{1}{n}\sum_i \frac{q_i}{n_i/n}\sum_j n_{ij} logP_{ij} \\ &= \sum_i\sum_j q_i \frac{n_{ij}}{n_i} logP_{ij} \end{aligned} \tag{3.3}$$

(3.3)을 최대화하는 β의 WESML 추정량을 $\hat{\beta}$ 이라 하면,

$$\frac{\partial W(\hat{\beta}, q)}{\partial \beta} = W_\beta(\hat{\beta}, q) = 0$$

이 만족된다.

$W_\beta(\hat{\beta}, q)$을 참값 β에 대하여 1차 테일러 급수 전개하면,

$$0 = \sqrt{n} W_\beta(\hat{\beta}, q) \simeq \sqrt{n} W_\beta(\beta, q) + W_{\beta\beta}(\bar{\beta}, q)\sqrt{n}(\hat{\beta} - \beta) \tag{3.4}$$

이다. 여기서 $W_\beta(\beta, q)$은 $W(\beta, q)$의 일차도함수를 나타내며, $W_{\beta\beta}(\bar{\beta}, q)$은 참값 β와 $\hat{\beta}$의 사이값 $\bar{\beta}$에서 평가된 $W(\beta, q)$의 이차도함수를 나타낸다.

보조정리 3.1 $W_{\beta\beta}(\bar{\beta}, q) \xrightarrow{a.s.} -\sum_i \sum_j p_j\, P_{ij}^2(1 - P_{ij}) X_j' X_j$

$$\equiv -J_w$$

증명: 대수의 법칙에 의해 n_{ij}/n_i는 Q_{ji}로 거의 확실하게(almost surely) 수렴한다. 또한 $\hat{\beta}$은 β의 일치추정량이라는 사실과 P_{ij}의 연속성 성질에 의해 β와 $\hat{\beta}$의 사이값 $\bar{\beta}$에서 평가된 P_{ij}의 값 $\bar{P}_{ij}$는 확률적으로 P_{ij}로 수렴한다. 이 사실들과 $W_{\beta\beta}(\beta, q)$의 다음 식에 의해 정리는 증명된다.

$$W_{\beta\beta}(\beta, q) = -\sum_i \sum_j q_i \frac{n_{ij}}{n_i} P_{ij}(1 - P_{ij}) X_j' X_j$$

□

보조정리 3.2 $\sqrt{n} W_\beta(\beta, q) \xrightarrow{D} N(\mathbf{0}, G_w)$,

$$G_w = \sum_i \frac{q_i^2}{\mu_i} \sum_j \sum_k Q_{ji}(\delta_{jk} - Q_{ki})(1 - P_{ij})(1 - P_{ik}) X_j' X_j,$$

여기서 $j = k$ 이면 $\delta_{jk} = 1$, $j \neq k$ 이면 $\delta_{jk} = 1$이다.

증명: $W_\beta(\beta, q)$의 전개식은 다음과 같다.

$$\begin{aligned} W_\beta(\beta, q) &= \frac{\partial W(\beta, q)}{\partial \beta} \\ &= \sum_i \sum_j q_i \frac{n_{ij}}{n_i} u_i (1 - P_{ij}) X_j', \end{aligned}$$

여기서 $i = 0$ 이면 $u_i = -1$, $i = 1$ 이면 $u_i = 1$이다.
사례-대조 연구 자료에서 $\mathrm{E}[n_{ij}] = n_i Q_{ji}$, $\mathrm{Cov}(n_{ij}, n_{ik}) = n_i Q_{ji}(\delta_{jk} - Q_{ki})$ 이므로 $\sqrt{n} W_\beta(\beta, q)$의 기대값과 분산-공분산 행렬은 다음과 같다.

$$\begin{aligned} \mathrm{E}[W_\beta(\beta, q)] &= \mathrm{E}\Big[\sum_i \sum_j q_i \frac{n_{ij}}{n_i} u_i (1 - P_{ij}) X_j' \Big] \\ &= \sum_j p_j \sum_i u_i P_{ij}(1 - P_{ij}) X_j' \\ &= \mathbf{0} \\ \mathrm{Var}[\sqrt{n} W_\beta(\beta, q)] &= \mathrm{Var}\Big[\sum_i \sum_j \sqrt{n} q_i \frac{n_{ij}}{n_i} u_i (1 - P_{ij}) X_j' \Big] \\ &= \sum_i q_i^2 \frac{n}{n_i} \sum_j \sum_k Q_{ji}(\delta_{jk} - Q_{ki})(1 - P_{ij})(1 - P_{ik}) X_j' X_j. \end{aligned}$$

가정 2.1에 의해 $\lim_{n \to \infty} \mathrm{Var}[\sqrt{n} W_\beta(\beta, q)] = G_w$ 이며, 중심극한 정리에 의해 정규성은 증명된다. □

이제 $\hat{\beta}$의 점근분포는 정규분포를 따른다는 대표본성질을 알아보기 위해 (3.4)을 다음과 같이 표현한다.

$$\sqrt{n}(\hat{\beta} - \beta) \simeq -W_{\beta\beta}^{-1}(\bar{\beta}, q) \cdot \sqrt{n} W_\beta(\beta, \hat{q}) \tag{3.5}$$

정리 3.1 $\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{D} N(\mathbf{0}, H), \quad H = J_w^{-1} G_w J_w^{-1}.$

증명: 보조정리 3.2에서 $\sqrt{n}W_\beta(\beta, q)$가 점근 정규분포를 따른다는 사실을 보였다. (3.5)식과 정규분포의 선형함수는 정규분포를 따른다는 성질로부터 정규성은 증명된다. 또한 보조정리 3.1과 3.2에 의하여 $\sqrt{n}(\hat{\beta}-\beta)$의 평균과 공분산 행렬은 다음과 같이 구해진다.

$$\begin{aligned}
\mathrm{E}[\sqrt{n}(\hat{\beta}-\beta)] &= \mathrm{E}\,[-W_{\beta\beta}^{-1}(\bar{\beta}, q)\cdot\sqrt{n}W_\beta(\beta, q)] \\
&= J_w^{-1}\sqrt{n}\ \mathrm{E}[W_\beta(\beta, q)] \\
&= \mathbf{0} \\
\mathrm{Var}[\sqrt{n}(\hat{\beta}-\beta)] &= \mathrm{Var}\,[-W_{\beta\beta}^{-1}(\bar{\beta}, q)\cdot\sqrt{n}W_\beta(\beta, q)] \\
&= J_w^{-1} Var[\sqrt{n}W_\beta(\beta, q)] J_w^{-1} \\
&= J_w^{-1} G_w J_w^{-1}.
\end{aligned}$$

□

4. 예제와 결론

의료인이 법정전염병을 발견 즉시 보건 기구에 신고하는 것은 전염성 질병 예방 차원에서 중요한 문제이나 그렇지 않은 경우도 있으리라 생각되어 법정전염병 신고율에 관심을 갖기로 한다. 이 예제에서 다루게 될 자료는 가톨릭대학교 예방의학교실에서 조사 연구된 자료의 일부로써 법정전염병을 발견 즉시 신고한 경우 $(Y=1)$와 미신고한 경우 $(Y=0)$를 반응변수로 하는 이항자료이다. 법정전염병 확진 여부를 주변수로 하며 신고 불이행시 벌칙 조항을 알고 있는지를 공변수로 측정한다.

반응변수 : 법정 전염병 신고$(Y=1)$, 미신고$(Y=0)$
주변수 : 법정전염병 확진$(E=0)$, 의심$(E=1)$
공변수 : 신고 불이행시 벌칙 조항을 안다$(Z=0)$, 모른다$(Z=1)$

법정전염병 환자를 신고하지 않은 경우와 신고한 경우로부터 법정전염병을 확진$(E=1)$한 경우와 의심$(E=0)$한 경우의 정보를 수집한 1차 단계 자료를 표 4.1에 정리하였다.

2차 단계에서는 1차 단계 자료에서 상대위험도를 고려하여 $Y=i$, $E=j$에 대하여 $N_{ij}/2$개 자료를 무작위 추출하여 법정전염병 신고 불이행시 벌칙 조항을 알고 있는지에 대한 정보를 수집한 결과를 표 4.2에 요약하였다.

사례-대조 연구 자료의 특성을 무시하고 주어진 자료를 전향성 연구 자료라 가정한 후 분석한 결과를 표 4.3에 요약하였으며 Breslow와 Cain(1988)이 2단계 사례-대조 자료를 위한 로지스틱 회귀모형의 모수 추정방법으로 제안한 CML 추정량의 결과는 표 4.4에 요약하였다.

이 논문에서 제시한 2단계 사례-대조 연구 자료 분석 방법은 다음과 같다. 1차 단계 정보를 이용하여 q_i와 Q_{ji}의 추정량 N_i/N와 N_{ij}/N_i를 구하고, 2차 단계에서의 정보를 이용하여 μ_i와 p_j의 추정량 n_i/n와 $(n_{0j}+n_{1j})/n$를 구한다. β의 WESML 추정량 $\hat{\beta}$은 SAS 프로그

표 4.1: 1차 단계 자료 구조

	$Y=0$	$Y=1$
$E=0$	27	292
$E=1$	238	66
소계	265	358

표 4.2: 2차 단계 자료 구조

	(주변수, 공변수)	$Y=0$	$Y=i$
X_1	$E=0, Z=0$	5	59
X_2	$E=0, Z=1$	8	86
X_3	$E=1, Z=0$	42	21
X_4	$E=1, Z=1$	74	11
	소계	129	177

표 4.3: 전향성 연구 로지스틱 분석

변 수	추정량	±	표준오차
절 편	-3.0056	±	0.3931
주변수	3.8439	±	0.3704
공변수	0.8752	±	0.3494

표 4.4: CML 추정방법

변 수	추정량	±	표준오차
절 편	-2.9669	±	0.3234
주변수	3.8386	±	0.2708
공변수	0.8752	±	0.3494

표 4.5: WESML 추정 방법

변 수	추정량	±	표준오차
절 편	-2.9878	±	1.0255
주변수	3.8427	±	0.1368
공변수	0.8725	±	0.1066

램 PROC LOGISTIC에서 가중치 $(N/N_i)\cdot(n_i/n)$를 사용하여 구하며 정리 3.1에서 정의된 $\hat{\beta}$의 분산 H를 추정한다. 그 결과를 표 4.5에 정리하였다.

표 4.4와 표 4.5로 부터 CML, WESML 방법으로 얻은 모수 추정량의 차가 매우 근소함을 확인 할 수 있다. CML 방법으로 얻은 절편 추정량의 표준오차는 WESML 추정 방법으로 얻은 절편 추정량의 표준오차에 비하여 작아 CML 추정 방법이 WESML 추정 방법에 비하여 절편 모수 추정에 있어 더욱 좋은 결과를 얻었다고 할 수 있다.

그러나 주변수 추정량의 경우 WESML 추정 방법으로 얻는 표준오차는 0.1368이였으며, CML 추정 방법으로 얻은 표준오차 0.2708에 비하여 그 크기가 매우 작음을 알 수 있다. 또한 CML 추정 방법으로 얻은 표준오차도 전향성 연구 추정 방법으로 얻은 표준오차 0.3704보다 작았다. 즉, 주변수 추정량의 표준오차의 크기는 WESML, CML, 전향성 연구 로지스틱 추정 방법 순으로 작음을 알 수 있었다. 이는 실제적인 분석에서 가장 관심 있는 주변수의 추정에 있어서는 CML 추정 방법이나 WESML 추정 방법이 전향성 연구 로지스틱 분석에 비하여 더 좋은 결과를 가지며, 그 중에서도 WESML 추정 방법이 가장 좋은 결과를 갖는 다는 사실을 보여준다. 공변수 추정량의 표준오차 크기는 WESML 추정 방법으로 얻은 표준오차가 CML,전향성 연구 로지스틱 추정 방법으로 얻은 표준오차에 비하여 매우 작음을 알 수 있다.

제시된 예제 분석에서는 2단계 사례-대조 자료분석시 CML과 WESML 추정 방법이 자료의 구조를 고려하지 않고 전향성 연구자료로 추정하는 방법보다 대체로 좋은 결과를 얻게 되었으며, 주변수와 공변수의 모수 추정량에 있어서는 WESML 추정방법이 CML 추정 방법에 비하여 더욱 좋은 결과를 얻음을 알 수 있었다.

참고문헌

[1] 신미영 (1994). Logistic regression for retrospective Studies, <품질경영학회지>, 22권 4호, 111-119

[2] Breslow, N.E. and Cain, K.C. (1988). Logistic regression for two-stage case-control data, *Biometrika*, vol. 75, 11-21.

[3] Cosslett, S.R. (1981). Efficient estimation of discrete choice models, *In structural analysis of disease data with economical statistics* MA:MIT Press, 51-111.

[4] Manski, C. and Lerman, S. (1977). Estimation of choice probabilities from choice-based samples, *Econometrica*, vol. 45, 1977-1988.

[5] Manski, C. and McFadden, D. (1981). Alternative estimator and sample designs for discrete choice analysis, *In structural analysis of disease data with economical statistics*, MA:MIT Press, 2-50.

[6] Prentice, R.L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies, *Biometrika*, vol. 66, 403-411.

[7] Scott, A.J. and Wild, C.J. (1989). Hypothesis testing in case-control studies, *Biometrika*, vol. 76, pp 806-808.

[8] Scott, A.J. and Wild, C.J. (1991). Fitting logistic regression models in stratified case-control studies, *Biometrics*, vol. 47, 497-510.

[9] White, J.E. (1982). A two stage design for the study or the relationship between a rare exposure and a rare disease, *American Journal of Epidemiology*, vol. 115, 119-128.

[10] Wild, C.J. (1991). Fitting prospective regression models to case-control data, *Biometrika*, vol. 78, 705-717.

[1999년 12월 접수, 2000년 5월 채택]

Estimation of Logistic Regression for Two-Stage Case-Control Data

Mi-Young Shin[1] Eun-Soon Shin[2]

ABSTRACT

In this paper we consider a logistic regression model based on two-stage case-control sampling and study the Weighted Exogeneous Sampling Maximum Likelihood(WESML) method to get an asymptotically normal estimates of the parameters in a logistic regression model. A numerical example is carried out to demonstrate the differences between the Conditional Maximum Likelihood(CML) estimates and the WESML estimates for two-stage case-control data.

Keywords: Logistic; Two-stage Case-control; WESML.

1) Associate Professor, Division of Natural Science, The Catholic University of Korea.
E-mail: myshin@www.cuk.ac.kr
2) Graduate Student, Dept. of Statistics, Dongguk University.
E-mail: shin_eunsoon@hotmail.com