응용통계연구 제 6권 2호
1993년 9월 pp. 393-399.

로지스틱 회귀모형에서 최우추정량의 정확도 산정[1)]

이 기 원[2)], 손 건 태[3)], 정 윤 식[4)]

요 약

반응이 두 가지로 나타나는 자료에서 설명변수와 반응변수와의 관계를 연구할 때 많이 사용되는 로지스틱 회귀모형에 대하여 그 모수들을 최우추정법으로 구할 때 추정량의 표준오차는 보통 로그우도함수의 2차도함수에 바탕을 두어 계산하게 된다. 한편 피셔정보량이 로그우도함수의 1차도함수를 재곱한 통계량의 기대값으로도 계산된다는 점에 착안하여 얻어지는 피셔정보량의 추정량도 이와 거의 비슷한 대표본 성질을 갖는 것으로 알려져 있다. 이러한 피셔정보량의 추정량들은 최우추정량을 구할 때의 반복 알고리즘과 깊은 관련을 갖고 있다. 어느 방법이 더 효과적으로 최우추정량을 계산하는 지 평균반복횟수를 비교하고 대표본분산의 추정량으로서 각 방법에서 계산되는 분산의 추정량들을 비교하였다.

1. 서론

어떠한 사건이 일어날 확률과 설명변수와의 관계를 로짓으로 연결짓는 로지스틱 회귀모형은 반응변수가 두 가지 값만을 취하는 자료를 적합시킬 때에 많이 쓰이고 있다. 특히 관찰에 의하여 자료를 수집할 때에는 설명변수 및 반응변수가 서로 독립이고 같은 분포를 갖는다고 볼 수 있기 때문에 다음과 같은 방식으로 모형을 세울 수 있다.

먼저 i번째 관찰치가 취하는 설명변수의 값은 $G(\cdot)$을 분포함수로 갖는 확률변수 X_i 의 관찰값이라고 볼 수 있다. 이 때 관찰된 설명변수의 값 $X_i = x_i$ 와 그 때 반응이 1로 나타날 확률 $p_i = P(Y_i = 1 | X_i = x_i)$을 다음과 같은 로짓변환으로 연결지을 수 있다.

$$\log \frac{p_i}{1-p_i} = x_i^t \beta. \tag{1.1}$$

이 변환은 0에서 1사이로 그 취할 수 있는 값에 제한을 받고 있었던 p_i에 비하여 그 범위에 제한을 두기 어려운 $X_i^t\beta$ 에게 실수전체를 취하도록 하여 β의 추정에서 나타날 수 있는 문제점을 없애 주고 있다. 이 변환을 p_i에 대하여 다시 쓰면

$$p_i = \frac{1}{1 + \exp(-x_i^t\beta)} \tag{1.2}$$

이 되어 로지스틱분포(logistic distribution)의 누적분포함수를 이용하고 있음을 알 수 있다.

이와 같은 역할을 하는 변환들로는 정규분포의 누적분포함수를 이용하는 프로빗(probit)변환

1) 이 연구는 교육부의 연구비지원(BSRI-92-116)으로 이루어졌음.
2) 강원도 춘천시 옥천동 1번지 한림대학교 통계학과.
3) 부산직할시 금정구 장전동 부산대학교 통계학과.
4) 부산직할시 금정구 장전동 부산대학교 통계학과.

과 극단값분포(extreme-value distribution)의 누적분포함수를 이용하는 complementary log-log 변환 등이 있다. 이 변환들의 모양과 비교에 대하여는 McCullagh and Nelder(1990)의 3장이 참고가 된다. 이 중 로짓변환의 특징이라면 일반화선형모형(generalized linear models)의 관점에서 볼 때 자연연결함수(natural link function)를 갖기 때문에 여러 가지 계산이 간단해진다는 점을 들 수 있다.

설명변수를 선형으로 조합시켜주는 β 를 여러 가지 방법으로 추정할 수 있다. 이 중 대부분의 통계팩키지에서 채택하고 있는 최우추정방법에 촛점을 두어 살펴보면, 주어진 n 개의 랜덤표본에서 β 의 최우추정량은 다음의 로그우도함수를 최대로 하는 값이다.

$$l_n(\beta, G) = \sum_{i=1}^{n} \{y_i x_i^t \beta - \log[1 + \exp(x_i^t \beta)] + \text{constants}\}. \qquad (1.3)$$

적절한 조건 하에서 최우추정량 $\hat{\beta}_n$ 은 유효추정량(efficient estimator)이 되어 $n^{1/2}(\hat{\beta}_n - \beta)$ 의 분포가 평균이 0이고 분산-공분산 행렬이 피셔정보행렬의 역행렬로 주어지는 다차원 정규분포로 수렴함을 알 수 있다. 여기서 피셔정보행렬 $I(\beta)$는 다음과 같이 계산된다.

$$I(\beta) = E\{-\nabla^2 l_1\} = E\{XX^t p(1-p)\}, \qquad (1.4)$$

여기서 $\nabla^2 l_1$ 은 (1.3)식에서 표본의 크기가 하나일 때를 β 에 대하여 두 번 미분하여 얻은 2차도함수이다. 따라서 $I^{-1}(\beta)/n$ 을 $\hat{\beta}_n$ 의 대표본근사 분산-공분산 행렬로 간주할 수 있어서 최우추정량의 정확도(accuracy)를 산정하는 바탕이 되며 이의 추정량들에 대하여 관심이 간다.

본 연구에서 고려 대상에 넣고 있는 2가지 추정량 중의 하나는 (1.4)식의 표현을 이용한 것으로서 다음과 같이 주어진다.

$$H_1(\hat{\beta}_n) = 1/n \sum_{i=1}^{n} X_i X_i^t \hat{p}_i(1-\hat{p}_i), \qquad (1.5)$$

단 여기서 $\hat{p}_i$ 은 $1/[1 + \exp(-X_i^t \hat{\beta}_n)]$ 이다.

한편 (1.4)식의 또다른 표현은 $E[(\nabla l)^t \nabla l] = E\{X_i X_i^t (Y_i - p_i)^2\}$으로 주어지므로 이를 이용하면 다음과 같은 추정량을 얻을 수 있다.

$$H_2(\hat{\beta}_n) = 1/n \sum_{i=1}^{n} X_i X_i^t (Y_i - \hat{p}_i)^2. \qquad (1.6)$$

본 연구에서는 이 두 가지 추정방법들을 다음과 같이 비교하고자 한다. 즉, H_1에 바탕을 두어 β 의 추정량을 구할 때 소요되는 반복횟수와 H_2에 바탕을 두어 β 의 추정량을 구할 때 소요되는 반복횟수를 비교하고 또한 $\hat{\beta}_n$ 의 정확도 산정에 바탕이 되는 $I^{-1}(\beta)$ 의 추정량이라는 관점에서 $H_1^{-1}(\hat{\beta}_n)$과 $H_2^{-1}(\hat{\beta}_n)$의 평균제곱오차를 비교하고자 한다. 이와 아울러 각 방법으로 계산되는 분산의 추정량을 $\hat{\beta}_n$ 의 소표본 평균제곱오차와도 비교하여 소표본 변동의 측도로서 어느 방법이 더 바람직한 지 알아 보고자 한다.

2절에서는 추정량의 성질 및 모의 실험의 개요에 대하여 설명하고 3절에서는 모의실험 결과를 분석하여, β에 대한 추정방법으로서 (1.5)를 이용하는 것이 더 반복횟수를 줄일 수 있으며 대표본분산 $I^{-1}(\beta)$의 추정방법으로도 평균제곱오차의 측면에서 역시 (1.5)가 더 낫다는 것을 보인다. 부수적으로 소표본 변동의 측도로서는 (1.6)이 더 낫다는 것을 알 수 있게 된다.

2. 추정량의 성질 및 모의실험 개요

2.1. 추정량의 성질

반응이 두 가지로만 나타나고 그 자료수집방법이 관찰에 의할 경우가 있다. 이때에는 다음과 같은 방법으로 로지스틱 회귀모형을 세울 수 있다. 먼저 관찰치 개개인에게 주어지는 설명변수, 즉 나이라든가 사회적 지위 등의 변수에 대하여 미지의 확률분포함수 $G(x)$를 갖는 확률벡터 X를 설정한다. 이 설명변수가 어떤 값 x를 취할 때 반응이 1로 나올 확률 p를 로지스틱분포의 확률분포함수로 보아 (1.2)와 같은 관계를 세운다. 이 식을 선형관계에 촛점을 두어 다시 쓰면 (1.1)에서와 같은 로짓변환이 된다.

이 때 β 에 대한 로그우도함수를 한 개의 관찰치에 대하여 구해 보면

$$l_1(\beta) = yx^t\beta - \log[1 + \exp(x^t\beta)] + \text{constants}, \tag{2.1}$$

이 되고 합성함수에 대한 미분 공식을 이용하여 1차도함수를 구하면

$$\nabla l_1 = x(y - p) \tag{2.2}$$

이 된다. 합성함수에 대한 미분 공식을 다시 한 번 적용하여 2차도함수를 구하면

$$\nabla^2 l_1 = -xx^t p(1 - p) \tag{2.3}$$

이 되어 l_1 을 최대로 하는 β 값이 존재할 경우에는 단 하나로 주어짐을 알 수 있다. 따라서 n개의 관찰치에 대하여 β의 최우추정량은

$$\nabla l_n = \sum_{i=1}^{n} x_i(y_i - p_i) = 0 \tag{2.4}$$

의 해로 주어진다. 단, 여기서 $p_i = 1/[1 + \exp(-x_i^t\beta)]$ 이다. (2.4)의 해는 보통 다음과 같은 과정의 반복적 방법에 의하여 구해진다.

$$\beta_{r+1} = \beta_r + H^{-1}(\beta_r)\nabla l_n(\beta_r), \tag{2.5}$$

단 여기서 r은 반복 횟수를 나타내고 $H(\cdot)$는 알고리듬에 따라 달라진다. 뉴톤-랩슨 방법에 의하면 로그우도함수의 2차도함수에 음수를 취한 행렬, 즉 $-\nabla^2 l_n(\cdot)$를 사용하고 피셔의 스코어링 방법에 의하면 $E-\nabla^2 l_n(\cdot)$, 즉 n 개의 관찰치로부터 구한 피셔정보행렬을 사용하게 된다. 이 문제에서 피셔의 스코어링 방법은 그 기대값이

$$E-\nabla^2 l_n(\beta) = nE[XX^t p(1 - p)] \tag{2.6}$$

와 같이 주어져 아직 적분형태로 남아 있기 때문에 그에 근사한 값

$$\sum_{i=1}^{n} X_i X_i^t p_i(1 - p_i) = -\nabla^2 l_n \tag{2.7}$$

을 사용하게 되어 결국 뉴톤-랩슨 방법과 동일한 결과를 얻게 된다. 이 방법은 또 피셔정보행렬 (1.4)의 추정량으로 (1.5)를 제안한다고 볼 수 있다.

Berndt(1974) 등은 프로빗 모형의 컨텍스트에서 H로 $\sum_{i=1}^{n}\nabla l_i(\nabla l_i)^t$ 를 사용할 것을 제안하였고 Griffiths등(1987)은 역시 프로빗 모형의 컨텍스트에서 이러한 방법들에 대하여 모의실험으로 비교한 결과를 발표한 바 있다. Griffiths등(1987)의 실험결과를 요약하면 뉴톤-랩슨과 스코어링 방법이 빨리 해를 구하고 거의 동일하게 Berndt등의 방법보다 $I^{-1}(\beta)$에 대한

보다 정확한 추정량을 얻게 되는 데 비하여 Berndt등의 방법은 소표본들에서 평균제곱오차에 가장 가까운 분산의 추정량을 제공해 주고 있다. Berndt등의 방법을 로지스틱 모형에 적용하면 피셔정보행렬 (1.4)의 추정량으로 (1.6)을 사용하는 것을 의미한다.

따라서 로지스틱 회귀모형의 컨텍스트에서 표본의 크기에 따라 어느 방법이 더 효과적으로 최우추정량을 계산하고 또한 피셔정보행렬의 역행렬에 더 가까운 대표본분산의 추정량을 구해 주는 지 밝힐 필요가 있다.

2.2. 모의 실험 개요

모의 실험에서는 계산의 편의를 돕기 위하여 모수의 갯수를 1로 하여 절편이 없는 로지스틱 회귀모형을 택하였다. 또 이 실험에서 설명변수 X의 분포는 구간 $(-2\sqrt{3}, 2\sqrt{3})$ 에서 균일한 분포로 택하여 그 평균이 0 분산이 1이 되도록 하였고 기울기 β는 1로 하였다. 즉 랜덤반응변수 Y는 설명변수 X의 관찰된 값 x에 따라 성공확률 $p = 1/[1 + \exp(-x)]$를 갖는 베르누이분포를 따른다. 균일한 분포를 갖는 랜덤넘버들은 IMSL의 서브루틴 ggubs를 이용하여 추출하였다. 최대 허용 반복횟수는 100회로 하였고 각 반복에서 구한 로그우도의 상대오차가 0.0001보다 작아지면 수렴하는 것으로 판정하였다.

이와 같은 방법으로 구한 최우추정량 $\hat{\beta}_n$과 참값 β와의 평균제곱오차는 충분히 큰 M에 대하여 다음 (2.8)식과 같이 주어진다.

$$MSE(\hat{\beta}_n) = E(\hat{\beta}_n - \beta)^2 \approx 1/M \sum_{j=1}^{M} (\hat{\beta}_{j,n} - \beta)^2. \tag{2.8}$$

단 여기서 $\hat{\beta}_{j,n}$ 은 j 번째 반복실험에서 구한 β의 최우추정량이다. 이 경우에 피셔정보량은

$$\begin{aligned} I(\beta) &= E[X^2 p(1-p)] \\ &= \int x^2 \frac{\exp(-x)}{[1+\exp(-x)]^2} G(dx) \end{aligned} \tag{2.9}$$

으로 주어지므로 결국 이 실험에서의 피셔정보량은 단순히

$$\frac{1}{2\sqrt{3}} \int_0^{2\sqrt{3}} x^2 \frac{\exp(-x)}{[1+\exp(-x)]^2} dx \tag{2.10}$$

이 되고 MATHEMATICA v2.0으로부터 이 값을 계산하면 0.167626이 나온다. 따라서 최우추정량 대표본분산은 주어진 표본의 크기 n에 대하여

$$0.167626^{-1}/n = 5.965662/n \tag{2.11}$$

으로 주어진다.

두 가지 방법 중 어느 쪽이 보다 더 대표본분산에 가까운 분산의 추정량을 제공해 주는지 비교하기 위하여 각 방법으로 부터 구한 분산 추정량과 대표본분산간의 평균제곱오차를 다음과 같은 방법으로 구하였다. $k = 1, 2$ 에 대하여

$$\begin{aligned} MSE(H_k^{-1}) &= E[H_k^{-1}(\hat{\beta}_n) - I^{-1}(\beta)]^2 \\ &\approx 1/M \sum_{j=1}^{M} [H_k^{-1}(\hat{\beta}_{j,n}) - I^{-1}(\beta)]^2. \end{aligned} \tag{2.12}$$

이 실험에서 전체 반복횟수 M은 5,000번으로 하였으며 표본의 크기 n은 각각 50, 100, 200, 400으로 하여 평균제곱오차의 크기가 어떤 속도로 줄어드는지 살피기 쉽도록 하였다.

3. 모의실험결과

표 1은 2.2에서 설명한 모의 실험 결과를 요약한 것이다. 단 여기서 방법 1이란 2.2절에서 밝힌 바와 같이 (1.5)식을 이용하는 방법을 의미하고 방법 2란 (1.6)식을 이용하는 것을 의미한다.

먼저 어느 방법이 더 효과적으로 최우추정량을 구해주는지 알기 위하여 각 방법에 대하여 수렴에 필요한 반복횟수를 비교하여 본 결과 랜덤하게 고른 100번의 경우에 대하여 다음과 같은 요약이 얻어졌다.

	N	MEAN	MEDIAN	STDEV	SEMEAN	MIN	MAX	Q1	Q3
방법 1	100	4.110	4.00	0.530	0.053	3.000	5.000	4.000	4.000
방법 2	100	4.990	5.00	1.389	0.139	3.000	10.000	4.000	6.000

이 요약에서 알 수 있는 바와 같이 방법 1의 경우 5번 이내에 해가 구해진 데 반하여 방법2의 경우에는 평균적으로 1번 정도 더 반복이 소요될 뿐만 아니라 그 변화가 심하여 최고 10회까지의 반복이 소요됨을 관찰할 수 있었다. 전체 실험에 있어서는 30회 이상의 반복을 거쳐 해가 얻어지는 경우도 있어서 최대 반복횟수를 100으로 하였다.

결론적으로 방법1이 최우추정량을 구하는 데 있어서 방법2보다 효과적일 뿐 아니라 대표본분산의 추정량으로서 $I^{-1}(\beta)$를 추정하는 데 있어서도 평균제곱오차를 줄여 주고 있으나 몬테-카를로 방법에 의하여 추정된 최우추정량의 소표본 평균제곱오차에는 방법2가 더 근접함을 알 수 있다.

표. 모의실험결과의 요약

단 여기서 β의 참값은 1, $I^{-1}(\beta)$의 참값은 5.965662이다.

	n	50	100	200	400
최우추정량	추정값	1.0443	1.0251	1.0169	1.0064
	MSE	0.1488	0.0672	0.0311	0.0152
방법1	H^{-1}	6.7070	6.2978	6.1503	6.0436
	$MSE(H^{-1})$	6.9420	2.0724	0.7929	0.3499
	H^{-1}/n	0.1341	0.0630	0.0308	0.0151
방법2	H^{-1}	6.9351	6.3768	6.1789	6.0612
	$MSE(H^{-1})$	10.9252	2.6623	0.9710	0.4191
	H^{-1}/n	0.1387	0.0638	0.0309	0.0152

감사의 글

본 논문의 초고에 대하여 많은 건설적 비평을 하여 주심으로 내용을 개선하는 데 많은 도움이 되어 주신 심사위원께 감사를 드립니다.

참 고 문 헌

[1] Berndt, E.R., Hall, B.H., Hall, R.E., and Hausman, J.A. (1974), Estimation and Inference in Non-Linear Structural Models, Annals of Economic and Social Measurement, 3, 653-665.

[2] Griffiths, W.E., Hill, R.C., and Pope, P.J. (1987). Small Sample Properties of Probit Model Estimators, Journal of Americal Statistical Association, Vol. 82, No. 399, 929-937.

[3] McCullagh, P. and Nelder, J.A.(1990), Generalized Linear Models, 2nd ed. London, Chapman and Hall.

Assessing the Accuracy of the Maximum Likelihood Estimator in Logistic Regression Models

Kee-Won Lee, Keon-Tae Sohn, and Younshik Chung

Abstract

When we compute the maximum likelihood estimators of the parameters for the logistic regression models, which are useful in studying the relationship between the binary response variable and the explanatory variable, the standard error calculations are usually based on the second derivative of log-likelihood function. On the other hand, an estimator of the Fisher information motivated from the fact that the expectaion of the cross-product of the first derivative of the log-likelihood function gives the Fisher information is expected to have similar asymptotic properties. These estimators of Fisher information are closely related with the iterative algorithm to get the maximum likelihood estimator. The average numbers of iterations to achieve the maximum likelihood estimator are compared to find out which method is more efficient, and the estimators of the variance from each method are compared as estimators of the asymptotic variance.