

임의중도절단자료를 갖는 일반화된 지수회귀모형[†] (Generalized Exponential Regression Model with Randomly Censored Data)

하 일 도*
(li-Do Ha)

요 약 임의중도절단자료(randomly censored data)를 갖는 일반화된 지수회귀모형을 고려하여 이 모형에서 모수를 추정하는 수정된 피셔 점수화(modified Fisher scoring)방법을 제안한다. 이를 위해 우도방정식(likelihood equations)이 유도되고 추정알고리즘(estimating algorithm)이 개발된다. 실제의 자료를 통해 제안된 방법을 예증한다.

Abstract We consider generalized exponential regression model with randomly censored data and propose a modified Fisher scoring method which estimates the model parameters. For this, the likelihood equations are derived and then the estimating algorithm is developed. We illustrate the proposed method using a real data.

1. 서 론

임의중도절단자료란 어떤 정해진 시작점으로부터 특정한 사건의 발생시점까지의 기간(예: 수술 또는 치료 시작으로부터 사망이란 사건이 발생한 시점까지의 기간)인 생존시간(survival time)으로 구성된 생존자료(survival data)중 일부가 중도에 관측되지 못한 경우를 말한다. 현실적으로 이러한 생존자료가 임의중도절단되는 경우가 빈번하다. 예를 들어, 환자가 각기 다른 임의의 시점에 연구에 참여하게 되는 임상실험(clinical trials)의 경우, 환자가 다른 병원으로 옮기거나 치료를 포기하거나 환자가 사망하기 전에 연구가 종결되는 경우 등으로 인하여 어떤 시점 이상의 자료가 수집되지 못한다.

최근에 임의중도절단된 자료분석은 의학 및 생리학분야에서 뿐만 아니라 공학이나 사회과학 등의 분야에서도 널리 응용되고 있다. 하지만 임의중도절단된 자료를 정확하

게 분석하기 위해서는 그 자료에 적절한 통계적모형을 설정 또는 개발한 후, 그 모형모수(model parameter)를 추정하는 방법을 연구하는 것이 선행되어야 한다.

만성병(예: 암)으로 고생하는 환자들에 관한 임의중도절단된 자료의 분석을 위해, 1970년대 이후 많은 학자들(예: [2],[3],[6])은 지수회귀모형(exponential regression model)을 고려하였다. 그들은 모형모수를 추정하기 위해 관측된 정보행렬(observed information matrix)에 근거한 뉴턴-랩슨(Newton-Raphson; N-R)방법을 사용하였다.

위의 자료분석 방법에는 다음과 같은 두 가지 문제점이 수반된다. 첫째, 보다 다양한 임의중도절단된 자료를 분석할 수 있는 일반화된 지수회귀모형 (generalized exponential regression model)이 요구된다. 둘째, 관측된 정보행렬은 일반적으로 모형모수에 대한 초기추정치(initial estimate)에 민감하여 역행렬이 존재하지 않는 경우가 많기 때문에 N-R방법을 사용하는데 있어서 수렴성에 한계가 있다.

이러한 문제점을 극복하기 위해, 이 논문에서는 Nelder와 Wedderburn[5]이 개발한 일반화된 선형모형(generalized linear model; GLM)에서 연결함수(link

[†] 이 논문은 1997년도 경산대학교 기린 연구비 지원에 의한 것임.

* 경산대학교 정보과학부 조교수

function)를 허용한 일반화된 지수회귀모형을 고려한 후, 이 모형의 모수를 추정하는 방법으로 수정된 피셔 점수화(Fisher scoring; FS)방법을 제안한다. 이 방법은 관측된 정보행렬 대신 피셔 정보행렬(Fisher information matrix)의 합리적인 한 추정량(estimator)을 사용한다. 이 논문의 주요구성은 아래와 같다. 2절에서는 고려한 모형을 정의하고 우도방정식을 유도한다. 3절에서는 수정된 FS방법을 제안한다. 또한, 개발된 초기추정치들 갖는 추정 알고리즘이 제시된다. 마지막으로 4절에서는 실제의 자료를 통해서 우리의 방법을 예증한다.

2. 모형과 우도방정식

i 번째 ($i=1, 2, \dots, n$)환자에 대해, T_i 를 연구중인 생존시간 그리고 C_i 를 T_i 에 대응되는 임의중도절단시간이라 하자. 그러면 관측 가능한 확률변수들은 다음과 같다.

$$Y_i = \min(T_i, C_i), \quad \delta_i = I(T_i \leq C_i),$$

여기서 $I(\cdot)$ 는 지표함수(indicator function)이다. 또한 T_i 와 C_i 는 서로 독립이라 가정하자.

이 논문에서 우리는 다음과 같은 일반화된 지수회귀모형을 고려한다.

$$\eta_i = g(\mu_i) = x_i' \beta \quad (i=1, 2, \dots, n), \quad (2.1)$$

여기서, η_i 와 $g(\cdot)$ 는 각각 GLM에서의 선형예측식(linear predictor)과 연결함수(link function)이다. 또한 $\mu_i = E(T_i) (> 0)$

는 서로독립인 지수분포(exponential distribution)를 따르는 생존시간 T_i 의 모평균(population mean)이며

$$x_i' = (x_{i1}, x_{i2}, \dots, x_{ip})$$

는 i 번째 환자에 대응되는 기지의 공변량들(known covariates)의 $1 \times p$ 벡터이고 β 는 미지의 모형모수(또는 회귀모수(regression parameter))들의 $p \times 1$ 벡터이다.

일반적으로, 지수회귀모형인 경우 연결함수로서 로그 연결($\eta_i = \log(\mu_i)$), 항등(identity)연결($\eta_i = \mu_i$), 역(reciprocal)연결($\eta_i = 1/\mu_i$)등을 주로 사용한다. 많은 학자들은 선택된 하나의 연결함수를 갖는 지수회귀모형을 고려한 후 N-R방법을 사용하여 자료를 분석하여 왔다. 하지만 고려된 모형(2.1)은 이러한 개개의 연결함수를 통합한(unified) 일반화된 지수회귀모형이다.

관측값 (y_i, δ_i, x_i') ($i=1, 2, \dots, n$)에 근거한 모형(2.1)의 β 에 대한 로그-우도함수(log-likelihood function)는 다음과 같이 주어진다.

$$\ell(\beta) = \sum_i \ell_i,$$

여기서,

$$\begin{aligned} \ell_i &= \ell(\beta; y_i, \delta_i, x_i') \\ &= -\delta_i \log(\mu_i) - \frac{y_i}{\mu_i}. \end{aligned}$$

그러면 β 에 대한 우도방정식은 연쇄법칙(chain rule)에 의해 다음과 같이 유도된다.

$$\begin{aligned} \ell_j(\beta) &= \sum_i \frac{(y_i - \delta_i \mu_i) x_{ij}}{\mu_i^2} \left(\frac{1}{g(\mu_i)} \right) \\ &= 0, \quad (j=1, 2, \dots, p) \end{aligned} \quad (2.2)$$

여기서 $\ell_j(\beta) = \sum_i \partial \ell_i / \partial \beta_j$ 이고

$g'(\mu_i) = \partial g(\mu_i) / \partial \mu_i$ 이다. (2.2)로부터 p -차원(p -dimensional)의 점수벡터(score vector)는 다음과 같이 표현할 수 있다.

$$\ell'(\beta) = X' W u, \quad (2.3)$$

여기서 $\ell'(\beta)$ 는 j 번째원소가 $\ell_j(\beta)$ 인 $p \times 1$ 벡터, X 는 i 번째 행벡터(row vector)가 x_i' 인 $n \times p$ 모형행렬(model matrix),

W 는 i 번째원소가 $w_i = \{\mu_i^2 [g'(\mu_i)]^2\}^{-1}$ 인 $n \times n$ 대각선 가중행렬(diagonal weight matrix)이고, u 는 i 번째 원소가

$$u_i = (y_i - \delta_i \mu_i) g'(\mu_i)$$

인 $n \times 1$ 벡터이다.

위의 우도방정식(2.2) 또는(2.3)에 있는 $X' W u = 0$ 는 평균함수 μ_i 들에 의존하는 β_j 들의 비선형함수(nonlinear function)이므로, β 의 해인 최대우도추정량(maximum likelihood estimator; MLE) $\hat{\beta}$ 을 얻기 위해서는 N-R 또는 FS와 같은 반복적 방법(iterative method)을 사용하여 이 우도방정식을 β 에 대해 풀어야 한다.

3. 수정된 FS방법

N-R방법을 통해 우도방정식 (2.2)을 풀기 위해서 요구되는 $\ell(\beta)$ 의 음2차미분(negative second derivatives)은 다음과 같이 주어진다.

$$-\ell''(\beta) = X^t W^* X, \quad (3.1)$$

여기서 $-\ell''(\beta)$ 는 (j, k) 번째 원소가

$$-\ell''_{jk}(\beta) = \sum_{i=1}^n x_{ij} w_i^* x_{ik},$$

$(j, k = 1, 2, \dots, p)$ 인 $p \times p$ 행렬이고 W^* 는 i 번째 원소가

$$w_i^* = -w_i \left\{ \left(\delta_i - \frac{2y_i}{\mu_i} \right) - (y_i - \delta_i \mu_i) \frac{g''(\mu_i)}{g'(\mu_i)} \right\} \quad (3.2)$$

인 $n \times n$ 대각선 가중행렬이다. 모형행렬 X 가 완전 열계수(full column rank)이고 W^* 의 대각선상의 원소 w_i^* 가 모든 i 에 대해 양수이면 (3.1)은 양정치행렬(positive definite matrix)이 된다. 하지만 X 가 완전 열계수행렬이라 할 지라도 w_i^* 가 모든 i 에 대해 양수를 만족하기는 쉽지 않다. 그 이유는 식(3.1)이 자료, 모수, 연결함수에 매우 의존되어 있기 때문이다. 나아가 (3.2)에 있는 Y_i 의 기대값이 중도절단시간 C_i 들에 의존되기 때문에 β 의 피서 정보행렬을 구하기가 어렵다. 따라서, MLE를 얻기 위해 N-R과 FS방법을 바로 적용할 수 없다. 한 대안으로 우리는 피서 정보행렬의 한 추정량을 사용하는 FS방법(수정된 FS방법)을 제안한다.

정리 1. 모형 (2.1)하에서, $p \times p$ 피서 정보행렬 $I(\beta)$ 은 다음과 같다.

$$I(\beta) = E(X_D^t W_D X_D), \quad (3.3)$$

여기서 D 는 중도절단이 없는(uncensored) 환자들에 대한 관측치들의 집합을 나타내며, X_D 는 i 번째 행벡터 $x_i^t (i \in D)$ 를 갖는 $r \times p$ 모형행렬, W_D 는 i 번째 대각선원소가 $w_i (i \in D)$ 인 $r \times r$ 행렬이며, $r = \sum_{i=1}^n \delta_i$ 는 임의중도절단이 없는 개수를 나타낸다.

(증명) 모형 (2.1)하에서 우리는 다음을 얻는다.

$$E(\partial \ell_i / \partial \beta_j) = 0, \quad j = 1, 2, \dots, p. \quad (3.4)$$

그러면, (2.2)와 (3.4)에 의해

$$E(Y_i) = \mu_i E(\delta_i) \quad (i = 1, 2, \dots, n) \quad (3.5)$$

이고 (3.2)와 (3.5)에 의해

$$E(w_i^*) = w_i E(\delta_i) \quad (i = 1, 2, \dots, n) \quad (3.6)$$

이다. 따라서 (3.1)의 기대값과 (3.6)을 이용하면 (3.3)의 증명이 완료된다.

정리 1의 (3.3)으로부터, $I(\beta)$ 의 합리적인 한 추정량을 다음과 같이 제안한다:

$$\hat{I}(\hat{\beta}) = X_D^t \widehat{W}_D X_D, \quad (3.7)$$

여기서 \widehat{W}_D 는 i 번째 원소가 $\widehat{w}_i (i \in D)$ 인 $r \times r$ 대각선 가중행렬이며 $\widehat{w}_i (= w_i(\hat{\beta}))$ 는 $\beta = \hat{\beta}$ 에서 계산된 $w_i = w_i(\beta) (i \in D)$ 이다. X_D 가 완전열계수이고 W_D 의 대각선상의 값이 양수이면 추정량 (3.7)은 양정치행렬이 된다. 사실, W_D 의 대각선상의 모든 원소는 특별한 경우(예: $\beta = 0$)를 제외하고는 항상 양수이다.

제안된 FS방법을 통하여 MLE $\hat{\beta}$ 을 얻는 추정알고리즘을 우리는 다음과 같이 제시한다:

단계 0: β 의 초기추정치 $\hat{\beta}^{(0)}$ 를 얻는다.

단계 1: 선택된 연결함수 $g(\cdot)$ 에 대해 다음을 계산한다.

$$\widehat{\mu}_i^{(0)} = g^{-1}(x_i^t \hat{\beta}^{(0)}) \quad (i = 1, 2, \dots, n).$$

단계 2: $\ell'(\hat{\beta}^{(0)})$ 와 $\hat{I}(\hat{\beta}^{(0)})$ 를 계산한다.

여기서 $\ell'(\hat{\beta}^{(0)})$ 와 $\hat{I}(\hat{\beta}^{(0)})$ 는 각각 $\beta = \hat{\beta}^{(0)}$ 에서 계산되는 점수벡터 (2.3)과 피서 정보행렬의 추정량 (3.7)이다.

단계 3: 아래의 반복식을 이용하여 β 의 다음 추정치 $\hat{\beta}^{(1)}$ 를 계산한다.

$$\hat{\beta}^{(1)} = \hat{\beta}^{(0)} + [I(\hat{\beta}^{(0)})]^{-1} \ell'(\hat{\beta}^{(0)}).$$

단계 4: 필요한 수렴기준(convergence criterion)이 만족될 때까지 단계 1-3을 계속 반복한다.

수정된 FS방법의 알고리즘에서 초기추정치들 얻기 위해 우리는 임의중도절단자료를 완전자료(complete data 또는 uncensored data)로 간주한 후 GLM에서 이용되는 초기추정치들 사용한다. 즉,

$$\hat{\beta}^{(0)} = (X^t W^{(0)} X)^{-1} X^t W^{(0)} z^{(0)}, \quad (3.8)$$

여기서 $W^{(0)}$ 와 $z^{(0)}$ 는 각각 $\mu_i = y_i$ 에서 계산되는 행렬 W 와 $z_i^{(0)} (= g(\mu_i))$ 를 갖는 $n \times 1$ 수정된 중

속벡터(adjusted dependent vector)이다. 보다 자세한 것은 참고문헌 [4], [5]를 보라.

한편, N-R방법은 단계 3에서 $\hat{I}(\hat{\beta}^{(0)})$ 대신 $\beta = \hat{\beta}^{(0)}$ 에서 계산되는 관측된 정보행렬 $-\ell''(\hat{\beta}^{(0)})$ 을 사용한다.

적절한 조건[1] 하에서의 $\hat{\beta}$ 의 점근정규성(asymptotic normality)과 (3.7)로부터, $\hat{\beta}$ 의 추정된 $p \times p$ 점근 공분산 행렬 (asymptotic covariance matrix)은

$$\widehat{\text{Cov}}(\hat{\beta}) = [\hat{I}(\hat{\beta})]^{-1} = [X_D' \widehat{W}_D X_D]^{-1}. \quad (3.9)$$

이므로 $\hat{\beta}_j$ ($j=1, 2, \dots, p$)의 표준오차(standard errors; SEs)는 (3.9)의 (j, j)번째 대각선원소의 제곱근(square root)으로 주어진다.

4. 예제

이 절에서는 제안된 FS방법과 N-R방법을 수렴측면에서 비교하려고 한다. 이를 위해 우리는 편의상 항등 연결합 수 즉, $\eta_i = \mu_i$ 를 갖는 모형 (2.1)을 고려한다. 그런데 모형 (2.1)은 모든 i 에 대해 $\mu_i > 0$ 을 만족해야 하므로 항등연결을 사용하는 경우 β 에 대한 허용 가능한 공간(admissible space) B 는 다음과 같은 제약을 갖는다.

$$B = \{ \beta | x_i' \beta > 0, \text{ for all } i \}$$

이러한 제약을 갖는 모형 (2.1)에 있는 β 를 추정하는데 있어서 N-R방법은 <표 1>에서 보는 바와 같이 W^* 의 대각선상의 원소 w_i^* 들이 양수가 되지 않을 가능성이 많다. 즉, $w_i^* > 0$ 이기 위해서는 다음 식

$$2\mu_i^{-1} y_i > \delta_i \quad (\text{단, 모든 } i \text{에 대해 } \mu_i \neq 0).$$

이 성립해야 하는 어려움이 존재한다. 하지만 제안된 방법은 모든 i 에 대해 $\mu_i = 0$ 인 특별한 경우를 제외하고는

W_D 의 대각선상의 원소 w_i^D 는 항상 양수임을 알 수 있다.

<표 1> 항등연결합수를 갖는 모형(2.1)에 대한 N-R방법과 수정된 FS방법의 비교

$-\ell''(\hat{\beta}) = X' \widehat{W}^* X$ (N-R 방법)	$\hat{I}(\hat{\beta}) = X_D' \widehat{W}_D X_D$ (수정된 FS방법)
$w_i^* = -\mu_i^{-2} \delta_i + 2\mu_i^{-3} y_i$ for $i=1, 2, \dots, n$	$w_i^D = \mu_i^{-2}$ for $i \in D$

Note: w_i^* 와 w_i^D 는 각각 W^* 와 W_D 의 대각선상의 원소이며, 이것은 각각 (3.1)과 (3.7)로부터 얻어진다.

따라서 <표 1>의 결과에 대해 실제의 자료를 이용하여 두 방법을 수렴측면에서 비교해보자.

예제 1: Prentice[6]에 의해 소개된 폐암에 관한 생존자료로서, 폐암환자 137명중에서 임의중도절단된 경우는 9명이다. 이 자료에서 생존시간 T 에 대한 여러 공변량 중 편의상 하나의 공변량 x 를 택하였다. 여기서 x 는 초진후 연구시작점까지의 경과시간(단위: 월)이다.

이 자료를 이용하여 우리는 다음과 같은 항등연결을 갖는 지수회귀모형을 고려하자.

$$\mu_i = \beta_0 + \beta_1 x_i \quad (i=1, 2, \dots, 137), \quad (4.1)$$

여기서 β_0 와 β_1 은 추정해야 될 미지의 회귀모수이다. 두 방법을 통해 모형 (4.1)을 적합(fitting) 하기 위해 우리는 다음과 같은 방법을 사용한다.

1) 두 방법 모두 똑같은 초기추정치 (3.8)를 이용한다. 모형 (4.1)의 경우 초기추정치는 $W^{(0)}$ 의 대각선 원소 $w_i^{(0)}$ 와 z 의 i 번째 원소 z_i 로 구성된다. 즉,
 $w_i^{(0)} = y_i^{-2}$, $z_i^{(0)} = y_i$
이다.

2) 수렴기준으로, $\beta = (\beta_0, \beta_1)'$ 에 대한 이전의 추정치(previous estimates)와 현재의 추정치(current estimates) 간의 절대적 상대변화율(absolute relative changes)의 최댓값(이때 상한 값=0.001)을 사용한다.

3) 3절에서 개발된 알고리즘의 프로그램을 위해 SAS/IML(version 6.12)을 사용한다.

이러한 방법을 거쳐서 적합된 결과는 <표 2>와 같다. 이 결과로부터 우리는 재미있는 형태를 발견하였다: 제안된 방법은 8번째 반복(iteration)후 수렴했지만 N-R방법은 발산하였으며, 제안된 방법의 첫 번째 반복의 추정치

를 N-R방법의 초기추정치로 대체한 결과 N-R방법은 다소 느린 22번째 반복후 수렴하였으며 마지막 추정치는 $(\hat{\beta}_0, \hat{\beta}_1) = (140.202, -1.112)$ 로 두 방법 모두 동일한 결

<표 2> Prentice (1973)의 자료를 이용한 N-R방법과 수정된 FS방법의 비교

Iteration	N-R		수정된 FS		N-R	
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
initial	3.477 ^{a)}	-0.063	3.477 ^{a)}	-0.063		
1	5.197	-0.095	152.739	-2.446	152.739 ^{b)}	-2.446
2	7.753	-0.142	149.739	-2.320	143.450	-1.946
3	11.535	-0.211	147.763	-2.148	149.439	-2.180
4	17.091	-0.313	143.463	-1.713	152.141	-2.444
5	25.163	-0.463	138.442	-1.013	145.846	-2.102
6	36.695	-0.678	140.217	-1.113	151.492	-2.390
7	52.733	-0.983	140.201	-1.111	170.470	-3.716
8	74.101	-1.401	140.202	-1.112	214.427	-6.916
9	100.674	-1.959	수렴		122.419	-3.287
⋮						
21	-401671.9	14348.892			140.215	-1.113
22	-803152.8	28688.483			140.202	-1.112
⋮					수렴	
⋮						
34	-3.2892×10^9	117470043				
35	-6.5769×10^9	234888188				
36	발산					
표준오차			13.54	0.51	13.49	0.50

a): N-R방법과 수정된 FS방법은 똑같은 초기 추정치가 사용되었다.

b): N-R방법의 수렴을 위해 수정된 FS방법의 첫 번째 반복의 추정치가 사용되었다.

과를 제공해 줌을 알 수 있다. 덧붙여, 수렴후 계산된 $\hat{\beta}$

의 표준오차는 기대되는 바와 같이 두 방법 모두 비슷한 값을 제공해 준다.

따라서, 적합된 모형(fitted model)은 다음과 같다:

(단, ()은 회귀추정치에 대응하는 표준오차를 나타낸다.)

(1) N-R방법의 경우:

$$\hat{\mu} = 140.202 - 1.112x$$

(13.54) (0.51)

(2) 수정된 FS방법의 경우:

$$\hat{\mu} = 140.202 - 1.112x$$

(13.49) (0.50)

결국, <표 1>과 <표 2>로부터 N-R방법은 초기추정치에 민감하여 수렴하는데에는 어려움이 수반되는 반면, 우리의 방법은 수렴측면에서 우월함을 알 수 있다.

참고 문헌

[1] Cox, D. R. and Hinkley, D.V., Theoretical Statistics. p.281, London: Chapman and Hall, 1974.
 [2] Greenberg, R., Bayard, S., and Byar, D.,

"Selecting concomitant variables using a likelihood ratio step-down procedure and a method of testing goodness of fit of an exponential survival model." Biometrics, Vol. 30, 1974, 601-608.

[3] Krall, J., Uthoff, V., and Harley, J., "A step-up procedure for selecting variables associated with survival." Biometrics, Vol. 31, 1975, 49-57.

[4] McCullagh, P. and Nelder, J. A., Generalized linear models, 2nd ed., p.41, London: Chapman and Hall, 1989.

[5] Nelder, J. A. and Wedderburn, R. W. M., "Generalized linear models." Journal of the Royal Statistical Society, A, Vol. 135, 1972, 370-384.

[6] Prentice, R. L., "Exponential survival with censoring and explanatory variables." Biometrika, Vol. 60, 1973, 279-288.