수정된 FS방법을 이용한 일반화된 지수생존모형의 추정

하일도* 조건호**

* 경산대학교 정보과학부 조교수
** 경산대학교 정보과학부 조교수

초 록

일반화된 지수생존모형(generalized exponential survival model)을 고려하여 이 모형의 모수를 추정하는 수정된 FS(modified Fisher scoring)방법을 제안한다. 이를 위해 우도방정식(likelihood equations)을 유도하고 초기추정치(initial estimate)를 포함한 추정알고리즘(estimating algorithm)을 개발한다.

I. 서론

임의중도절단된 생존자료(randomly censored survival data)의 분석은 최근에 의·약학 및 생리학분야에서 뿐만 아니라 공학이나 사회과학 등의 분야에서도 널리 응용되고 있다. 하지만 이러한 자료를 정확하게 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 생존자료의 분석을 위해, 1970년대 이후 많은 학자들(예: Prentice(1973), Greenberg등(1973), Krall등(1975))은 지수생존모형(exponential survival model)을 고려하였다. 그들은 모형모수를 추정하기 위해 관측된 정보행렬(observed information matrix)에 근거한 뉴톤-랩슨(Newton-Raphson; N-R)방법을 사용하였다. 이러한 자료분석 방법에는 다음과 같은 두 가지 문제점이 수반된다. 첫째, 보다 다양한 임의중도절단된 생존자료를 분석할 수 있는 일반화된 지수생존모형이 요구된다. 둘째, 관측된 정보행렬은 일반적으로 모형모수에 대한 초기추정치에 민감하여 역행렬이 존재하지 않는 경우가 많기 때문에 N-R방법을 사용하는데 있어서 수렴성에 한계가 있다.

이 논문에서는 이러한 문제점을 극복하기 위해 Nelder와 Wedderburn(1972)이 개발한 일반화된 선형모형
(generalized linear model; GLM)에서 연결함수(link function)를 허용한 일반화된 지수생존모형을 고려하여, 이 모형의 모수를 추정하는 수정된 FS 방법을 제안한다. 이 방법은 관측된 정보행렬 대신 피셔 정보행렬(Fisher information matrix)의 합리적인 한 추정량을 사용한다.

이 논문의 주요구성은 아래와 같다. 2절에서는 고려한 모형을 정의하고 3절에서는 우도방정식을 유도한다. 4절에서는 수정된 FS방법을 제안한다. 또한, 초기추정치를 포함한 추정 알고리즘이 제시된다. 마지막으로 5절에서는 수렴측면에서 우리의 방법을 예증한다.

II. 자료구조와 모형

i번째 ($i=1,2,\cdots,n$)환자에 대해, T_i를 연구중인 생존시간 그리고 C_i를 T_i에 대응되는 임의중도절단시간이라 하자. 그러면 관측 가능한 확률변수들은 다음과 같다.

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

여기서 $I(\cdot)$는 지표함수(indicator function)이다. 또한 T_i와 C_i는 서로 독립이라 가정하자.

이 논문에서 우리는 다음과 같은 일반화된 지수생존모형을 고려한다.

$$\eta_i = g(\mu_i) = x_i^t\beta \quad (i=1,2,\cdots,n), \tag{2.1}$$

여기서, η_i와 $g(\cdot)$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_i의 모평균(population mean)이며 $x_i^t = (x_{i1}, x_{i2}, \cdots, x_{ip})$는 i번째 환자에 대응되는 기지의 공변량들(known covariates)의 $1\times p$벡터이고 $\beta = (\beta_1, \beta_2, \cdots, \beta_p)^t$는 미지의 모형모수 (또는 회귀모수(regression parameter))들의 $p\times 1$벡터이다.

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

III. 우도방정식

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

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

여기서,

$$\ell_i = \ell(\beta; y_i, \delta_i, x_i^t) = -\delta_i \log(\mu_i) - \frac{y_i}{\mu_i}.$$

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

$$\ell_j'(\beta) = \sum_i^n \frac{(y_i - \delta_i\mu_i)x_{ij}}{\mu_i^2}\left(\frac{1}{g'(\mu_i)}\right) = 0,$$
$$(j=1,2,\cdots,p) \tag{3.1}$$

여기서 $\ell_j(\beta) = \sum_{i=1}^n \partial\ell_i/\partial\beta_j$ 이고 $g'(\mu_i) = \partial g(\mu_i)/\partial\mu_i$이다. (3.1)로부터 p-차원의 점수벡터(score vector)는 다음과 같이 표현할 수 있다.

$$\ell'(\beta) = X^t W u, \tag{3.2}$$

여기서 $\ell'(\beta)$는 j번째원소가 $\ell_j'(\beta)$인 $p\times 1$벡터, X는 i번째 행벡터(row vector)가 x_i^t인 $n\times p$ 모형행렬(model matrix), W는 i번째 원소가 $w_i = \{\mu_i^2[g'(\mu_i)]^2\}^{-1}$인 $n\times n$ 대각선 가중행렬(diagonal weight matrix)이고, u는 i번째 원소가 $u_i = (y_i - \delta_i\mu_i)g'(\mu_i)$인 $n\times 1$벡터이다.

우도방정식 (3.1) 또는 (3.2)에 있는 $X^tWu = 0$는 평균함수 μ_i들에 의존하는 β_j들의 비선형함수(nonlinear function)이므로, β의 해인 최대우도추정량 (maximum likelihood estimator; MLE) $\hat{\beta}$을 얻기 위해서는 N-R 또는 FS(Fisher scoring)와 같은 반복적 방법(iterative method)을 사용하여 이

우도방정식을 β에 대해 풀어야 한다.

Ⅳ. 수정된 FS방법

N-R방법을 통해 우도방정식 (3.1)을 풀기 위해서 요구되는 $\ell(\beta)$의 음 2차미분(negative second derivatives)은 다음과 같이 주어진다.

$$-\ell''(\beta) = X^t W^* X, \qquad (4.1)$$

여기서 $-\ell''(\beta)$는 (j,k)번째 원소가

$$-\ell_{jk}''(\beta) = \sum_{i=1}^{n} x_{ij} w_i^* x_{ik} \quad , (j,k=1,2,\cdots,p)$$

인 $p\times p$행렬이고 W^*는 i번째 원소가

$$w_i^* = -w_i\left\{(\delta_i - \frac{2y_i}{\mu_i}) - (y_i - \delta_i\mu_i)\frac{g''(\mu_i)}{g'(\mu_i)}\right\} \qquad (4.2)$$

인 $n\times n$대각선 가중행렬이다. 모형행렬 X가 완전 열계수(full column rank)이고 W^*의 대각선상의 원소 w_i^*가 모든 i에 대해 양수이면 (4.1)은 양정치 행렬(positive definite matrix)이 된다. 하지만 X가 완전 열계수행렬이라 할 지라도 w_i^*가 모든 i에 대해 양수를 만족하기는 쉽지 않다. 그 이유는 식 (4.1)이 자료, 모수, 연결함수에 매우 의존되어 있기 때문이다. 나아가 (4.2)에 있는 Y_i의 기대값이 중도절단시간 C_i에 의존되기 때문에 β의 피셔 정보행렬을 구하기가 어렵다. 따라서, MLE를 얻기 위해 N-R과 FS방법을 바로 적용할 수 없다. 한 대안으로 우리는 피셔 정보행렬의 한 추정량을 사용하는 FS방법-수정된 FS방법-을 제안한다.

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

$$I(\beta) = E(X_D^t W_D X_D), \qquad (4.3)$$

여기서 D는 중도절단이 없는(uncensored)환자들에 대한 관측치들의 집합을 나타내며, X_D는 i번째 행벡터 $x_i^t\,(i\in D)$를 갖는 $r\times p$ 모형행렬, W_D는 i번째 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.

$$E(\partial\ell_i/\partial\beta_j) = 0, \quad j=1,2,\cdots,p. \qquad (4.4)$$

그러면, (3.1)과 (4.4)에 의해

$$E(Y_i) = \mu_i E(\delta_i) \quad (i=1,2,\cdots,n) \qquad (4.5)$$

이고 (4.2)와 (4.5)에 의해

$$E(w_i^*) = w_i E(\delta_i) \quad (i=1,2,\cdots,n) \qquad (4.6)$$

이다. 따라서 (4.1)의 기대값과 (4.6)을 이용하면 (4.3)의 증명이 완료된다.

정리 1의 (4.3)으로부터, $I(\beta)$의 합리적인 한 추정량을 다음과 같이 제안한다:

$$\hat{I}(\hat{\beta}) = X_D^t\,\widehat{W_D}X_D, \qquad (4.7)$$

여기서 $\widehat{W_D}$는 i번째 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_D가 완전열계수이고 W_D의 대각선상의 값이 양수이면 추정량 (4.7)은 양정치행렬이 된다. 사실, W_D의 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: β의 초기추정치 $\hat{\beta}^{(0)}$를 얻는다.

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

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

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 (3.2)와 피셔 정보행렬의 추정량 (4.7)이다.

단계 3: 아래의 반복식을 이용하여 β의 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 초기추정치를 얻기 위해 우리는 임의중도절단자료를 완전자료 (complete data 또는 uncensored data)로 간주한 후 GLM에서 이용되는 초기추정치를 사용한다. 즉,

$$\hat{\beta}^{(0)} = (X^t W^{(0)} X)^{-1} X^t W^{(0)} z^{(0)}, \quad (4.8)$$

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 종속벡터(adjusted dependent vector)이다. 보다 자세한 것은 Nelder와 Wedderburn(1972), McCullagh와 Nelder(1989, page 41)를 보라.

나아가, 수렴기준으로 β에 대한 이전의 추정치(previous estimates)와 현재의 추정치(cur-rent estimates)간의 절대차 (absolute difference)의 최대값 (이때 상한값=0.001)을 사용한다.

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 (Cox와 Hinkley, 1974, page 281)하에서의 $\hat{\beta}$의 점근정규성(asymptotic normality)과 (4.7)로부터, $\hat{\beta}$의 추정된 $p\times p$ 점근공분산행렬 (asymptotic covariance matrix)은

$$\widehat{\mathrm{Cov}}(\hat{\beta}) = [\hat{I}(\hat{\beta})]^{-1} = [X_D^t \hat{W}_D X_D]^{-1}. \quad (4.9)$$

이므로 $\hat{\beta}_j$ $(j=1,2,\cdots,p)$의 표준오차(standard errors; SEs)는 (4.9)의 (j,j)번째 대각선원소의 제곱근(square root)으로 주어진다.

V. 예제

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

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

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

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

Note: w_i^*와 w_i^D는 각각 W^*와 W_D의 대각선상의 원소를 나타내며, 이것은 각각 (4.1)과 (4.7)로부터 얻어진다.

이러한 제약을 갖는 모형 (2.1)에서 β를 추정함에 있어 N-R방법은 <표 1>에서 보는 바와 같이 W^*의 대각선상의 원소 w_i^*들이 양수가 되지 않을 가능성이 많다. 즉, 모든 i에 대해 $w_i^* > 0$이기 위해서는 다음 식

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

이 성립해야 하는 어려움이 존재한다. 하지만 제안된 방법은 모든 i에 대해 $\mu_i = 0$인 특별한 경우를 제외하고는 W_D의 대각선상의 원소 w_i^D는 항상 양수임을 알 수 있다. 즉, <표 1>로부터 N-R방법은 초기추정치에 민감하여 수렴하는데 에는 어려움이 수반되는 반면, 우리의 방법은 수렴측면에서 우월함을 알 수 있다.

참 고 문 헌

Cox, D. R. and Hinkley, D.V., Theoretical Statistics. London: Chapman and Hall, 1974.

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.

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

McCullagh, P. and Nelder, J. A., Generaliz-ed linear models, 2nd ed., London: Chap-man and Hall, 1989.

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

Prentice, R. L., "Exponential survival with censoring and explanatory variables." Bio-metrika, Vol. 60, 1973, 279-288.