Journal of the Korean
Data & Information Science Society
2000, Vol. 11, No. 2, pp. 295 ~ 302

와이블 모형의 모수 추정에서 분할법의 효율성

정대현 [1] · 김주성[2] · 원동유[3]

요 약

생존분석에서 널리 이용되고 있는 모형 중 하나인 와이블 모형에 대해 효율적인 모수 추정에 대해 연구하였다. 공변량을 포함하고 있는 와이블 모형의 모수를 추정하기 위하여 전 치료기간을 여러 단계로 나누어 최대우도법을 적용하는 분할법을 소개하였다. 실제의 자료를 적용하여 분할법과 비분할법으로 모수를 추정하여 비교한 결과 분할법의 효율성을 입증하였다.

주제어: 분할, 와이블 분포, 최대 우도, 뉴톤-랩슨 방법

1. 서론

와이블 분포는 여러 상황에 적용 가능한 위험함수를 제공함으로 생존분석에서 발생하는 다양한 자료를 분석하기 위하여 자주 이용되고 있다. 와이블 분포의 확률밀도함수는

$$f(t;\beta,\delta) = \left(\frac{\delta}{\beta}\right)\left(\frac{t}{\beta}\right)^{\delta-1} exp\left[-\left(\frac{t}{\beta}\right)^{\delta}\right] \ , \ t \geq 0$$

이며 β는 일반적으로 공변량의 함수로 보는 것이 타당하고 δ는 분포의 형태를 나타내는 양의 상수로 가정한다 (Pike(1966), Peto and Lee(1973) and Nelson(1972)). 상수 δ의 값에 따라 와이블 모형은 증가, 감소 또는 일정한 위험함수를 갖는다.

환자 n명의 생존시간은 $T_1, \cdots, T_n$이라고 하고 중도절단된 시간을 $L_1, \cdots, L_n$이라고 하자. T_i가 L_i보다 작은 경우에만 실제 생존시간을 관측할 수 있으므로 ξ_i를 다음과 같이 정의하면

$$\xi_i = \begin{cases} 1, & T_i \leq L_i \\ 0, & T_i > L_i \quad i = 1, \cdots, n \end{cases}$$

[1]충북 청주시 흥덕구 개신동 산 48 충북대학교 자연과학대학 통계학과 부교수 ,
[2]충북 청주시 흥덕구 개신동 산 48 충북대학교 자연과학대학 통계학과 교수 ,
[3]충북 청주시 흥덕구 개신동 산 48 충북대학교 대학원 전자계산학과 박사과정.

는 i번째 환자의 중도절단 여부를 나타내는 변수이다. 실제 관측할 수 있는 자료는 중도절단된 시간 또는 생존시간이므로 실제 관측시간 t_i를 $t_i = \min(T_i, L_i)$ 로 볼 수 있다. 일반적인 와이블 모형에서 위험함수는

$$\lambda(t;x) = \delta t^{\delta-1}\exp(-x'\beta\delta), \quad t \geq 0$$

이고, 생존함수는

$$S(t;x) = \exp\{\exp(-x'\beta\delta)t^{\delta}\}$$

이다. $x = (x_1, \cdots, x_p)$는 공변수이며 $\beta = (\beta_1, \cdots, \beta_p)$는 공변수 벡터 x의 계수이며 δ는 양의 상수이다.

n명의 환자를 관측한 후 자료 (t_i, ξ_i, x_i)를 얻었을 때 우도함수와 로그-우도함수는 각각 다음과 같다. 단, $x_i = (x_{i1}, \cdots, x_{ip})$.

$$\mathcal{L}(\delta,\beta) = \prod_{i=1}^{n} \exp\left\{e^{-x_i'\beta\delta}t_i^{\delta}\right\}\left(\delta t_i^{\delta-1}e^{-x_i'\beta\delta}\right)^{\xi_i},$$

$$l(\delta,\beta) = \sum_{i=1}^{n} e^{-x_i'\beta\delta}t_i^{\delta} + \sum_{i=1}^{n}\xi_i(log\delta + (\delta-1)logt_i - x_i'\beta\delta).$$

(δ,β)에 대한 최대우도추정치를 얻기 위해 로그 우도함수를 δ와 β_k에 대하여 일차, 이차 편미분을 하여 다음 식들을 얻었다.

$$\frac{\partial l(\delta,\beta)}{\partial\delta} = \sum_{i=1}^{n}\xi_i\left(\frac{1}{\delta} + logt_i - x_i'\beta\right) + \sum_{i=1}^{n} e^{-x_i'\beta\delta}t_i^{\delta}\left[-x_i'\beta + logt_i\right], \tag{1}$$

$$\frac{\partial l(\delta,\beta)}{\partial\beta_k} = -\sum_{i=1}^{n}\delta x_{ki}(\xi_i + e^{-x_i'\beta\delta}t_i^{\delta}), \tag{2}$$

$$\frac{\partial^2 l(\delta,\beta)}{\partial\beta_k\partial\beta_{k'}} = \sum_{i=1}^{n} e^{-x_i'\beta\delta}\delta^2 x_{ki}x_{k'i}t_i^{\delta}, \qquad k,k' = 1,\cdots,p$$

$$\frac{\partial^2 l(\delta,\beta)}{\partial\delta^2} = -\sum_{i=1}^{n}\frac{\xi_i}{\delta^2} + \sum_{i=1}^{n} e^{-x_i'\beta\delta}\{(-x_i'\beta)^2 t_i^{\delta} - 2x_i'\beta t_i^{\delta}logt_i + t_i^{\delta}(logt_i)^2\},$$

$$\frac{\partial l(\delta,\beta)}{\partial\delta\partial\beta_k} = -\sum_{i=1}^{n}\xi_i x_{ki} + \sum_{i=1}^{n} e^{-x_i'\beta\delta}\left\{-x_{ki}\delta\left[-x_i'\beta t_i{}^{\delta} + t_i^{\delta}logt_i\right] - x_{ki}t_i^{\delta}\right\}.$$

식(1)과 식(2)를 뉴톤-랩슨 방법(Lawless(1982))과 SAS를 이용하여 (δ,β)의 최우추정치를 얻을 수 있다. 와이블 모형의 경우 SAS를 이용하여 모수 추정을 할 수 있다. 전 치료기간을 여러 구간으로 나누어 구간마다 서로 다른 모수를 가정하여 추정하는 분할법을 Shuster(1990)는 지수모형 그리고 Chung과 Won(1999)은 Gompertz 모형에 적용하였다. 2장에서는 와이블 모형의 모수 추정 방법으로 분할법을 소개하였고, 3장에서는 실제 자료를 이용해 비분할된

와이블 모형과 분할된 와이블 모형의 모수를 추정하여 서로 비교하고 4장에서는 앞에서 나온 결과들을 요약하여 결론을 내렸다.

2. 분할법을 이용한 와이블 모형의 모수 추정

어느 질병의 자료가 여러 단계로 이루어져 있을 때 환자들의 생존시간은 치료가 진행되면서 치료의 효과에 따라 변한다고 볼 수 있다. 어느 단계에서 치료의 효과가 탁월하다면 다음 치료 단계까지의 생존할 확률이 커지게 되므로 각 단계에 치료의 효과를 알아보기 위하여 조건부 생존확률을 이용하는 것이 바람직하다. 어느 질병의 치료 기간이 $(I-1)$ 단계로 나뉘어져 있다면 치료 기간을 I개의 구간으로 나누어진다. 환자의 혈압, 나이 등의 특성을 고려한 $(p-1)$개의 공변량$(x_1,\cdots,x_{p-1})$ 을 가정하자. 치료 그룹 g 번째에 있는 환자에 대한 생존기간의 위험함수와 생존함수는 각각 다음과 같다.

$$\begin{aligned}\lambda(t;x,g) &= \delta_g t^{\delta_g-1}\exp(-x'\beta\delta_g),\\ S(t;x) &= \exp\left\{\exp(-x'\beta\delta_g)\left[\tau_{i-1}^{\delta_g}-t^{\delta_g}\right]\right\},\\ &\quad t\in[\tau_{i-1},\tau_i),\\ &\qquad i=1,\cdots,I,\quad g=1,2\quad l=1,\cdots,n_{ig}.\end{aligned}\tag{3}$$

단, $\tau_0=0$와 $\tau_I=\infty$ 이라고 한다. 여기서, δ_g는 치료 그룹 g와 관련된 상수이며 β는 공변량들과 관련된 회귀계수이다.

R_{ig} 는 그룹 g에서 시점 τ_{i-1} 에 생존한 환자들의 집합이고 D_{ig}는 그룹 g에서 구간 $[\tau_{i-1},\tau_i)$에서 사망한 환자들의 집합이라고 하자. t_l을 l번째 환자가 관측된 생존시간이라고 하면, t_{l_i} 는 l번째 환자가 생존한 것으로 관측되었을 때 구간 i에서의 부분이다. 즉,

$$t_{l_i} = \begin{cases} t_l-\tau_{i-1}, & \text{구간 } i\text{에서 } l\text{번째 환자가 사망하거나 중도절단된 경우,}\\ \tau_i-\tau_{i-1}, & \text{구간 } i\text{를 지나 } l\text{번째 환자가 계속 생존하는 경우}\end{cases}$$
$$i=1,\cdots,I.$$

우도함수와 로그-우도함수는 각각 다음과 같다.

$$\mathcal{L}(\delta,\beta)=\prod_{i=1}^{I}\prod_{g=1}^{2}\left[\prod_{l\in R_{ig}}\exp\{e^{-x'\beta\delta_g}(\tau_{i-1}^{\delta_g}-(t_{li}+\tau_{i-1})^{\delta_g})\}\prod_{l\in D_{ig}}\delta_g t_l^{\delta_g-1}e^{-x'\beta\delta_g}\right],$$

$$l(\delta,\beta)=\sum_{i=1}^{I}\sum_{g=1}^{2}\Big\{\textstyle\sum_{l\in R_{ig}}e^{-x'\beta\delta_g}(\tau_{i-1}^{\delta_g}-(t_{li}+\tau_{i-1})^{\delta_g})$$
$$+\textstyle\sum_{l\in D_{ig}}(log\delta_g+(\delta_g-1)logt_l-x'\beta\delta_g)\Big\}.$$

모수 (δ,β)에 대한 최대우도추정치를 얻기 위해 다음 식들을 얻었다.

$$\frac{\partial l(\delta,\beta)}{\partial \delta_g} = \sum_{i=1}^{I}\sum_{l\in D_{ig}}(1/\delta_g + logt_l - x'\beta)$$
$$+\sum_{i=1}^{I}\sum_{l\in R_{ig}} e^{-x'\beta\delta_g}\left[-x'\beta(\tau_{i-1}^{\delta_g} - (t_{li}+\tau_{i-1})^{\delta_g}) + \tau_{i-1}^{\delta_g} log\tau_{i-1}\right.$$
$$\left.-(t_{li}+\tau_{i-1})^{\delta_g} log(t_{li}+\tau_{i-1})\right], \quad (4)$$

$$\frac{\partial l(\delta,\beta)}{\partial \beta_k} = \sum_{i=1}^{I}\sum_{g=1}^{2}\{-\sum_{l\in D_{ig}}\delta_g x_{kl} - \sum_{l\in R_{ig}} e^{-x'\beta\delta_g}\delta_g x_{kl}(\tau_{i-1}^{\delta_g} - (t_{li}+\tau_{i-1})^{\delta_g})\}, \quad (5)$$

$$\frac{\partial^2 l(\delta,\beta)}{\partial \beta_k^2} = \sum_{i=1}^{I}\sum_{g=1}^{2}\sum_{l\in R_{ig}} e^{-x'\beta\delta_g}(\delta_g x_{kl})^2(\tau_{i-1}^{\delta_g} - (t_{li}+\tau_{i-1})^{\delta_g}), \quad (6)$$

$$\frac{\partial^2 l(\delta,\beta)}{\partial \beta_k \partial \beta_{k'}} = \sum_{i=1}^{I}\sum_{g=1}^{2}\sum_{l\in R_{ig}} e^{-x'\beta\delta_g}\delta_g^2 x_{kl}x_{k'l}(\tau_{i-1}^{\delta_g} - (t_{li}+\tau_{i-1})^{\delta_g}), \quad (7)$$

$$\frac{\partial^2 l(\delta,\beta)}{\partial \delta_g \partial \delta_{g'}} = 0 \qquad g\neq g' \quad ,$$

$$\frac{\partial^2 l(\delta,\beta)}{\partial \delta_g^2} = -\sum_{i=1}^{I}\sum_{l\in D_{ig}}\frac{1}{\delta_g^2} + \sum_{i=1}^{I}\sum_{l\in R_{ig}} e^{-x'\beta\delta_g}\{(-x'\beta)^2(\tau_{i-1}^{\delta_g} - (t_{li}+\tau_{i-1})^{\delta_g})$$
$$-2x'\beta\left[\tau_{i-1}^{\delta_g} log\tau_{i-1} - (t_{li}+\tau_{i-1})^{\delta_g} log(t_{li}+\tau_{i-1})\right] + \tau_{i-1}^{\delta_g}(log\tau_{i-1})^2$$
$$-(t_{li}+\tau_{i-1})^{\delta_g}(log(t_{li}+\tau_{i-1}))^2\}, \quad (8)$$

$$\frac{\partial l(\delta,\beta)}{\partial \delta_g \partial \beta_k} = -\sum_{i=1}^{I}\sum_{l\in D_{ig}} x_{kl} + \sum_{i=1}^{I}\sum_{l\in R_{ig}} e^{-x'\beta\delta_g}\left\{(-x_{kl}\delta_g)\left[-x'\beta(\tau_{i-1}^{\delta_g} - (t_{li}+\tau_{i-1})^{\delta_g})\right.\right.$$
$$\left.+\tau_{i-1}^{\delta_g} log\tau_{i-1} - (t_{li}+\tau_{i-1})^{\delta_g} log(t_{li}+\tau_{i-1})\right]$$
$$\left.-x_{kl}(\tau_{i-1}^{\delta_g} - (t_{li}+\tau_{i-1})^{\delta_g})\right\}. \quad (9)$$

식(4)와 식(5)를 각각 0으로 놓고 뉴톤-랩슨 방법을 이용해 분할된 와이블 모형의 모수 (δ,β)의 최우추정치$(\hat{\delta},\hat{\beta})$를 구할 수 있다. 식 (7)이 항상 0보다 작으므로 주어진 δ_g에 대하여 방정식 $\frac{\partial l(\delta,\beta)}{\partial \beta_k} = 0$을 만족하는 β를 구할 수 있다.

추정치에 대한 근사적 분산과 공분산을 얻기 위해서, $(\hat{\delta},\hat{\beta})$를 대입한 관측된 정보행렬(information matrix)을 계산하면 다음과 같다.

$$I(\hat{\delta},\hat{\beta}) = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix},$$

여기서 I_{11}은 $g\times g$ 대각 행렬이고, 대각 원소는 식(8)에 (-1)을 곱한 것이다. I_{21}은 $g\times p$ 행렬

이고 I_{12} 는 I_{21}의 전치 행렬로 식(9)에 (-1)을 곱한 것이다. I_{22}는 $p \times p$ 행렬로 식(6)과 식(7)에 각각 (-1)을 곱한 것으로 이루어진다. $(\hat{\delta}, \hat{\beta})$에 대한 근사적 분산과 공분산은 $I(\hat{\delta}, \hat{\beta})^{-1}$이다.

3. 실제 자료의 적용

앞절에서 설명된 분할법이 와이블 분포의 모수 추정시 비분할법보다 더 효율적임을 보이기 위하여, Pediatric Oncology Group(POG)가 1981년에 1125명의 소아들을 대상으로 조사한 급성 백혈병에 관한 자료를 이용하였다.

와이블 모형에서 공변량으로 나이, 성별과 처리법을 다음과 같이 고려하였다.

$$x_1 = \begin{cases} 0 & \text{, 환자 나이가 10세 미만인 경우} \\ 1 & \text{, 환자 나이가 10세 이상인 경우} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{, 남자} \\ 2 & \text{, 여자} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{, SAM(standard plus high dose methotrexate pulses)} \\ 2 & \text{, S(standard)} \end{cases}$$

모수 추정 결과는 표 1에서 알 수 있듯이 치료 방법은 치료 기간에 유의하다고 할 수 없다. 표 2에 나타난 것처럼 치료 방법을 제외한 축소모형을 적합시켜 본 결과 나이와 성별은 소아암 환자의 치료 기간에 유의한 영향을 준다고 볼 수 있다.

치료 방법에 따라 두 그룹으로 나누어 분할법을 이용하여 와이블 모형의 모수를 추론한 결과 표 3과 같다.

표 1. 와이블 모형의 비분할법

Parameter*	Estimate	Std Error	p-value
β_0	2.1391	0.3161	< 0.0001
β_1	-1.1744	0.1887	< 0.0001
β_2	0.7986	0.1586	< 0.0001
β_3	0.0261	0.1490	0.8607
δ	0.6734	0.0303	< 0.0001

β_1은 나이와 관련된 회귀계수,

β_2는 성별과 관련된 회귀계수,

β_3는 처리와 관련된 회귀계수,

δ는 척도 모수.

표 2. 축소된 와이블 모형의 비분할법

Parameter*	Estimate	Std Error	p-value
β_0	2.1777	0.2278	< 0.0001
β_1	-1.1749	0.1887	< 0.0001
β_2	0.7989	0.1585	< 0.0001
δ	0.6734	0.0303	< 0.0001

β_1은 나이와 관련된 회귀계수,
β_2는 성별과 관련된 회귀계수,
δ는 척도 모수.

표 3. 와이블 모형의 분할법

Parameter*	Estimate	Std Error	p-value
β_0	2.1324	0.14625	< 0.0001
β_1	-1.0936	0.11887	< 0.0001
β_2	0.7430	0.10067	< 0.0001
δ_1	0.73301	0.03756	< 0.0001
δ_2	0.7414	0.03849	< 0.0001

β_1은 나이와 관련된 회귀계수,
β_2는 성별과 관련된 회귀계수,
δ_1은 SAM 치료법에서의 척도 모수,
δ_2은 S 치료법에서의 척도 모수.

표 2와 표 3을 보면 알 수 있듯이 분할법을 이용한 결과 분할된 와이블 모형에서 나이와 성별에 관련된 모수의 최우추정치가 작은 표준오차를 가지므로 분할법이 더 효율적임을 알 수 있었다.

4. 결론

본 연구에서는 생존분석에서 자주 이용되는 모형 중 하나인 와이블 모형의 효율적인 모수를 추정하기 위하여 분할법을 소개하고 그 효율성에 대하여 알아보았다. POG의 실제 자료를 이용하여 공변량들을 포함하는 와이블 모형에서 모수를 추정하기 위하여 비분할법보다 분할법이 더 효율적임을 알아보았다.

분할법의 장점은 여러 단계로 이루어진 어떤 질병의 치료 과정 중 각 치료 단계 사이에 서로 다른 위험률을 가정하는 것이다. 그럼으로써 전 치료 기간동안 일정한 위험률을

가정하는 것보다 더 다양하고 융통성있는 모형으로 자료를 분석할 수 있다. 그 결과 각 구간내에서 전 단계의 치료 방법의 효과에 따라 변하는 생존함수를 용이하게 얻을 수 있다. 계산적으로는 추정하고자하는 모수의 MLE의 존재성을 보장 받을 수 있다는 것도 장점이다(Geiser. et al(1998)). 장기간의 치료 기간을 필요로하는 질병일 경우에는 모형 (3)을 이용하여 마지막 단계의 치료를 받는 시점을 참고로하여 완치율을 쉽게 계산할 수 있다.

분할법의 단점은 전 구간을 여러 구간으로 나누고 각 구간마다 서로 다른 형태 모수를 가정하기 때문에 다수의 모수를 포함하는 모형을 다루어야 하는 것이다. 분할법의 효율성을 극대화하기 위하여 전 구간을 여러 개의 구간으로 분할하는 것을 정형화 할 수 없다는 것이 하나의 문제점이라 할 수 있다.

요즈음 연구가 활발해지고 있는 계수과정 이론을 본 연구에 이용하여 보는 것도 하나의 새로운 과제가 될 것이다.

참 고 문 헌

1. D. Chung and D. Won(1999). Assessing cure rates via piecewise Gompertz model with covariates, *J. of the Korea Data & Information Science Soc.*, 10, 445-455.

2. Geiser, P. W., Chang, M. , P. V. Rao, Shuster, J. J. and Pullen, J.(1998). Modelling cure rates using the Gompertz model with covariate information, *Statistics in Medicine*, 17, 831-839

3. Lawless, J. F.(1982). Statistical Models and Methods for Lifetime Data, *New York, Wiley.*

4. Nelson, W. B.(1972). Graphical analysis of accelerated life test data with the inverse power law model, *IEEE Trans. Reliab.*, R21, 2-11.

5. Peto, R., and P. Lee (1973). Weibull distributions for continuous carcinogensis experiments, *Biometrics*, 29, 457-470.

6. Pike, M.C.(1966). A method of analysis of a certain class of experiments in carcinogensis, *Biometrics*, 22, 142-161.

7. Shuster, J. J.(1990). Handbook of Sample Size Guideines for Clinical Trials, *Boca Raton, CRC Press Inc.*

Piecewise Weibull Model with Covariates

Daehyun Chung [4] · Ju-Sung Kim [5] · Dong-Yu Won [6]

Abstract

We study the efficient method to estimate the parameters for the Weibull model with covariates which occupies an important position in survival analysis. A treatment period may be divided by the stages of treatments under the different treatment arams. The piecewise method is considered to obtain the estimators of the parameters by maximum likelihood method. We explore the real data to show that the piecewise is more efficient than the nonpiecewise to estimate the parameters.

Key Words and Phrases: Piecewise, Weibull distribution, Maximum likelihood, Newton-Raphson method

[4]Associate Professor, Dept. of Statistics, Chungbuk National University, Chugbuk,
[5]Professor, Dept. of Statistics, Chungbuk National University, Chugbuk,
[6]Graduate Student, Dept. of Computer Science, Chungbuk National University, Chugbuk.