



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

이 학 석 사 학 위 논 문

극단값 분포를 이용한 AFT
변량효과 모델링 접근법



2020년 2월

부 경 대 학 교 대 학 원

통 계 학 과

현 성 진

이 학 석 사 학 위 논 문

극단값 분포를 이용한 AFT 변량효과 모델링 접근법

지도교수 하 일 도

이 논문을 이학석사 학위논문으로 제출함.



2020년 2월

부 경 대 학 교 대 학 원

통 계 학 과

현 성 진

현성진의 이학석사 학위논문을 인준함.

2020년 2월 21일



위원장 이학박사 노맹석 (인)

위원 이학박사 하일도 (인)

위원 이학박사 박인호 (인)

목 차

표 차례	ii
그림 차례	iii
1. 서론	1
2. 비례위험모형과 AFT 모형의 소개	3
2.1 Cox의 비례위험모형	3
2.2 AFT 모형	4
3. EV 분포를 이용한 AFT 변량효과 모형의 추정	6
3.1 AFT 변량효과 모형	6
3.2 다단계 가능도의 추정절차	9
4. 실제자료 분석	15
4.1 신장 감염 자료	15
4.2 다기관 방광암 임상시험 자료	18
5. 결론 및 토론	21
참고문헌	22
부록 : R 코드	25

표 차례

표 4.1. 신장 감염 자료 :	
변수 설명	16
표 4.2. 신장 감염 자료 :	
AFT 변량효과 모형을 이용한 분석	16
표 4.3. 방광암 자료 :	
변수 설명	18
표 4.4. 방광암 자료 :	
AFT 변량효과 모형을 이용한 분석	19

그림 차례

그림 4.1. 신장 감염 자료 :

각 환자 별 변량효과의 추정치 및 95% 신뢰구간..... 17

그림 4.2. 방광암 자료 :

각 센터 별 변량효과의 추정치 및 95% 신뢰구간..... 19



AFT random-effect modeling approaches
using extreme value distribution

Seong Jin Hyeon

Department of Statistic, The Graduate School,
Pukyong National University

Abstract

Cox's proportional hazards regression model is a semi-parametric method which has often used in the analysis of survival data. However, if the assumption of proportional hazards is not satisfied, the interpretation of the estimates can be difficult. Thus, accelerated failure time (AFT) regression model which usually assumes the distribution of survival time has been used. The AFT model is a linear model that is easier to interpret than the proportional hazards model. In this thesis, we propose a method of estimating AFT random-effect model with the extreme value distribution for analyzing multivariate (or clustered) survival data. For the inference we use a h-likelihood approach and derive the estimation procedure. The proposed method is illustrated with two real data sets, i.e. kidney infection data and multi-center bladder cancer trial data.

keywords: AFT model, Extreme value distribution, H-likelihood, Random effects, Multivariate survival data

1. 서론

개체 (subject)를 일정시간 추적하여 특정 사건이 발생하기까지의 시간을 측정하고 사건발생에 영향을 미치는 요인을 찾고자 할 때, 생존분석에서는 사건발생시간과 요인간의 관계를 분석, 평가한다.

생존분석(survival analysis)에서 비례위험 (proportional hazards; PH) 모형은 생존시간에 영향을 미치는 공변량 (covariate)을 찾는 방법으로서, 생존시간과 공변량의 관계가 위험함수를 통하여 표현된다. 비례위험 모형은 개별 생존시간의 위험률에 대하여 작용하는 고정효과(예: 회귀계수)에 의하여 모델링 된다. 준모수적 (semi-parametric) 모형인 비례위험 모형은 생존시간의 기저분포 (baseline distribution) 가정을 필요로 하지 않으며 위험함수와 공변량의 관계를 분석할 수 있다. 하지만 위험률이 시간에 관계없이 일정하다는 비례위험에 대한 가정이 요구되며, 이 가정이 만족되지 않으면 회귀 추정값에 대한 해석이 어려울 수 있다.

생존분석에서 모수적 (parametric) 방법인 가속화 실패시간 (accelerated failure time; AFT) 모형은 생존시간이 적절한 모수적 분포를 따른다고 가정하며 비례위험 모형의 대안으로 소개되었다 (Lawless, 1982). AFT 모형에서는 고정효과가 개별 생존시간에 선형적으로 작용하기 때문에 고정효과의 해석이 비례위험 모형보다 용이하다. AFT 모형은 가정한 모수적 분포가 잘못 지정되더라도 회귀계수의 추론이 다소 강건한 (robust) 결과를 줄 수 있다 (Ha 등, 2002, 2017).

본 논문에서는 AFT 변량효과 모형에서 오차항의 분포에 대해 극단값 (extreme value; EV) 분포를 사용하여 생존자료들 간 상관성을 갖는 다변

량 생존자료를 분석할 수 있는 변량효과를 허용하는 AFT 모형에 대한 접근 방법을 제안한다. 모형추론을 위해 다단계 가능도(h-likelihood) 방법을 사용한다. 다단계 가능도는 고정모수(fixed parameters)만을 다루는 기존의 가능도와는 달리 고정모수 뿐만 아니라 관측 되지 않는 변량효과를 포함하는 가능도이다. 특히 다단계 가능도는 주변가능도(marginal likelihood)를 구하기 위해 요구되는 어려운 적분이 필요하지 않으면서, 다양한 변량효과 모형에 대해 효율적인 통계적 추론결과를 제공한다 (Lee 등, 2017; Ha 등 2017).

변량효과를 가지는 AFT 모형에 대한 접근 방법을 제시하기에 앞서 비례 위험 모형과 AFT 모형에 관한 배경지식을 기술한다. 비례위험과 AFT에 관한 모형을 간략하게 기술한다. 또한 극단값 분포를 가지는 AFT 변량효과 모형의 추정절차와 적합 알고리즘을 제안한다. 더 나아가, 잘 알려진 신장 감염 자료 (McGilchrist 와 Aisbett, 1991)와 다기관 임상시험 자료 (multi-center clinical data)를 통해 제안된 방법을 예증한다. 본 논문의 모든 계산은 R 프로그램을 사용하였다.

본 논문의 구성은 다음과 같다. 2장에서는 비례위험 모형과 AFT모형에 대하여 소개한다. 3장에서는 변량효과를 갖는 AFT 모형을 EV 분포와 다단계가능도 접근법을 사용한 추정절차를 유도한다.. 4장에서는 신장 감염 자료 및 다기관 임상시험 자료를 사용하여 실제자료를 분석한다. 5장에서는 연구결과를 토론하고 향후 과제를 제시한다. 마지막으로 부록에서 실제자료 분석에 대해 개발된 R 코드를 제시한다.

2. 비례위험모형과 AFT 모형의 소개

생존분석에서의 회귀 분석에 대한 접근은 생존시간 (survival time)을 종속 변수로 하고 생존시간에 영향을 미치는 요인을 독립변수로 하는 통상적인 선형 다중회귀분석이다. 하지만 생존시간의 분포가 정규분포라는 조건하에서 선형회귀분석이 이루어질 수 있다. 생존시간은 양수의 값을 가지며 생존함수의 분포는 복잡한 형태이기 때문에 정규분포에 가까운 함수를 얻기 어려울 때가 많다. 또한, 일반적인 회귀분석의 방법으로는 생존분석 자료에서 자주 발생하는 중도절단 자료를 처리할 수 없다. 그러나 영국의 통계학자 D. R. Cox에 의하여 제안된 비례위험모형 (Cox, 1972)에 의하여 위험함수에 미치는 요인의 효과를 추정할 수 있게 되었다.

2.1 Cox의 비례위험모형

Cox의 비례위험 모형은 생존시간과 관련된 공변량이 있는 경우 매우 널리 사용되어 왔다.

콕스의 비례위험 모형은 다음과 같다 .

$$\lambda(t;x) = \lambda_0(t) \exp(x^T \beta) \quad (2.1)$$

여기서 $\lambda(t;x)$ 는 공변량 x 를 가지는 개인에 대한 시간 t 에서의 위험함수,

$\lambda_0(t)$ 는 공변량 값이 0인 경우 시간 t 에서 음의 값을 갖지 않는 기저 (baseline) 위험함수, $x = (x_1, x_2, \dots, x_p)^T$ 인 개인의 공변량 벡터, $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$ 는 공변량 x 에 대응하는 회귀모수의 벡터를 말한다.

비례위험 모형은 공변량의 형태가 모수적인데 비하여 기저위험 형태는 비모수적이기 때문에 준모수적 모형으로 부르며, 생존시간의 분포에 대하여 아무런 가정을 하지 않고 주어진 자료에 대한 관측치를 토대로 회귀계수 추론을 하는 방법이다. 비례위험 모형은 위험률이 시간에 관계없이 일정하다는 가정이 만족하지 않으면 추정값에 대한 해석이 어려울 수 있다. 그러나 위험함수 또는 생존함수의 분포에 대하여 구체적인 함수를 적용하면 모수적 모형이 된다. 생존분석 모형이 모수적 모형을 가지면 생존함수에 대한 추정이 용이해진다.

2.2 AFT 모형

생존분석에서 대표적인 모수적 회귀모형인 AFT 모형은 공변량이 생존시간에 승법적 (multiplicative)인 영향을 미치기 때문에 가속화 실패시간 모형이라고 부른다. AFT 모형은 로그실패 시간과 공변량의 직접적인 관계를 가법적으로 모델링하여 비례위험 모형의 한 대안으로서 생존자료에 대한 선형 회귀모형으로 사용되어 왔다.

T_i ($i=1, \dots, n$)를 i 번째 개체의 생존시간이라고 할 때, AFT 모형은 로그생존시간과 공변량과의 선형적 관계로써 다음과 같이 정의된다.

$$\log T = x^T \beta + \epsilon \quad (i = 1, \dots, n) \quad (2.2)$$

여기서 선형예측식 (linear predictor) $x^T \beta$ 는 절편 항을 포함하고, ϵ 은 랜덤 오차항 (random error)이다.

생존시간에 대한 선형 모형인 AFT 모형 (2.2)는 다음과 같이 표현 될 수 있다.

$$T = \exp(x^T \beta) T_0 \quad (2.3)$$

여기서 $T_0 = \exp(\epsilon)$ 이다. 모형 (2.3)으로 부터 공변량 x 는 실패가 일어날 때 까지의 시간 T 를 가속화 (또는 감속화)하는 역할을 의미한다. 따라서 모형 (2.2)를 가속화 실패시간 회귀모형이라 부른다.

생존시간 T 는 양의 값을 가지며 다른 변환도 사용할 수 있다. 하지만 로그 변환이 자주 사용된다.

3. EV 분포를 이용한 AFT 변량효과 모형의 추정

이 장에서는 EV분포를 사용한 AFT 변량효과 모형에 대한 추정절차를 유도한다. 일반적으로 변량효과는 생존 자료에서 개인 간 이질성과 개인 내의 상관관계를 설명하는 데 사용될 수 있다.

3.1 AFT 변량효과 모형

중도절단 하에서 군집화 된 사건 발생 시간 (clustered time-to-event)은 q 번째 개체 (또는 군집)의 n_i 번째 관측치($i=1, \dots, q; j=1, \dots, n_i$)로부터 얻어진다. 여기서 n_i 는 군집크기 (cluster size)를 말한다. T_{ij} 는 i 번째 개체의 j 번째 관측에 대한 생존시간이라 하고 C_{ij} 는 대응하는 중도절단 시간 (censoring time)이다. V_i 는 i 번째 개체의 관측되지 않는 (unobserved) 변량효과라고 하자. 여기서 $n = \sum_{i=1}^q n_i$ 로 총 표본의 크기를 말한다. 예를 들어, 다기관 임상시험에서 n_i 는 i 번째 센터의 환자 수이고 n 은 모든 q 개 센터의 총 환자의 수이다. 특히 모든 i 에 대해, $n_i=1$ 인 경우를 일변량 생존자료 (univariate or independent survival data) 그리고 $n_i=2$ 인 경우를 이변량 생존자료 (bivariate survival data)로 각각 부른다. 본 연구에서는 이와 같은 형태의 자료를 포함하는 다변량 생존자료 (multivariate survival data)를 다룬다.

로그생존시간과 공변량과의 선형적 관계인 AFT 변량효과 모형은 다음과 같이 정의된다.

$$\log T_{ij} = x_{ij}^T \beta + V_i + \epsilon_{ij} \quad (3.1)$$

여기서 $x_{ij} = (1, x_{ij1}, \dots, x_{ij,p-1})^T$ 는 T_{ij} 의 공변량 벡터이고,
 $\beta = (\beta_0, \beta_1, \dots, \beta_{p-1})^T$ 는 $p \times 1$ 회귀계수 벡터이다.

본 논문에서 변량효과인 V_i 의 분포는 $V_i \sim iid N(0, \alpha)$ 인 정규분포를 가정한다. $V_i \sim N(0, \alpha)$ 일 때, V_i 의 확률밀도함수는 다음과 같이 주어진다.

$$f(v_i) = (1/\sqrt{2\pi\alpha}) \exp(-v_i^2/2\alpha) \quad (3.2)$$

그리고 ϵ_{ij} 는 서로 독립인 랜덤 오차항이다.

본 논문에서는 ϵ_{ij} 의 분포로써 위치모수 (location parameter)가 0이고 척도모수 (scale parameter) σ 를 가지는 EV 분포를 가정한다.
 대응하는 확률밀도함수는 다음과 같다.

$$f(\epsilon) = \sigma^{-1} \exp\{(\epsilon/\sigma) - \exp(\epsilon/\sigma)\} \quad (3.3)$$

여기서 ϵ 는 실수범위의 값을 가진다. 그리고 ϵ 의 생존함수 (survival function)는 다음과 같다.

$$S(\epsilon) = \exp\{-\Lambda(\epsilon)\} \quad (3.4)$$

여기서 $\Lambda(\epsilon) = \exp(\epsilon/\sigma)$ 으로 ϵ 의 누적분포함수이다.

즉, $V_i = v_i$ 로 주어진 $\log T_{ij}$ 는 위치모수가 $x_{ij}^T \beta + v_i$ 이고 척도모수가 σ 인 EV 분포가 된다. 그리고 $V_i = v_i$ 로 주어진 T_{ij} 의 조건부 확률분포는 척도모수가 $\exp\{-(x_{ij}^T \beta + v_i)\phi\}$ 이고 형상모수 (shape parameter)가 $\phi = 1/\sigma$ 을 가지는 와이블 분포 (Weibull distribution)를 따른다.

본 논문에서 모형추론을 위해 중도절단시간 $C_{ij} (i=1, \dots, q, j=1, \dots, n_i)$ 와 관련하여 다음과 같이 두 가지 가정을 따른다 (Ha 등., 2017).

가정 1 : $V_i = v_i$ 가 주어졌을 때 순서쌍 $\{(T_{ij}, C_{ij}), j=1, \dots, n_i\}$ 가 조건부 독립이고, T_{ij} 와 C_{ij} 또한 $j=1, \dots, n_i$ 에 대해서 조건부 독립이다.

가정 2 : $V_i = v_i$ 가 주어졌을 때, $\{C_{ij}, j=1, \dots, n_i\}$ 는 T_{ij} 에 대하여 조건부 무정보적 (conditionally non-informative)이다.

여기서 무정보적 (non-informative)은 중도절단 시간의 분포가 실패시간 (failure time) 분포의 관심 모수에 의존하지 않음을 의미한다 (Klein 과 Moeschberger, 2003).

이러한 두 가정에 고려된 AFT 변량효과 모형 (3.1)에 대한 다단계 가능도 추론을 하고자 한다.

3.2 다단계 가능도의 추정절차

이 절에서는 다단계 가능도 함수를 사용하여 추정하는 방법 및 절차를 유도한다. 중도절단성 (censoring)으로 인해 모든 생존시간 T_{ij} 가 관측되지는 않기 때문에, 통상적으로 생존분석에서 관측 가능한 확률 변수 (observable random variables)는 다음과 같이 정의된다.

$$Y_{ij} = \min(\log T_{ij}, \log C_{ij}) \text{ and } \delta_{ij} = I(T_{ij} \leq C_{ij}) \quad (3.5)$$

여기서 $I(\cdot)$ 는 지시함수 (indicator function)이며, δ_{ij} 는 중도절단 지시함수 (censoring indicator function)이다.

가정1과 가정2 하에서 h 로 표현되는 변량효과를 가지는 가속화 실패시간 모형 (3.1)의 다단계 가능도는 다음과 같이 정의된다 (Lee와 Nelder, 1996; Ha 등, 2001).

$$h = h(\beta, \sigma, \alpha) = \sum_{ij} l_{1ij} + \sum_i l_{2i}, \quad (3.6)$$

여기서 l_{ij1} 은 다음과 같다.

$$l_{1ij} = l(y_{ij}, \delta_{ij} | v_i; \beta, \sigma) = \delta_{ij} \{ \log \sigma^{-1} + (y_{ij} - x_{ij}^T \beta - v_i) / \sigma \} - \Lambda_{ij} \quad (3.7)$$

여기서 l_{1ij} 는 $V_i = v_i$ 일 때, (Y_{ij}, δ_{ij}) 에 대한 조건부 로그가능도이다.

그리고 $\Lambda_{ij} = \exp[\{y_{ij} - (x_{ij}^T \beta + v_i)\}/\sigma]$ 이고 y_{ij} 는 Y_{ij} 에 대한 관측값이다.

V_i 에 대한 로그가능도인 l_{2i} 는 다음과 같다.

$$l_{2i} = l_{2i}(v_i; \alpha) = -(1/2)\log(2\pi\alpha) - v_i^2/2\alpha \quad (3.8)$$

AFT 변량효과 모형의 다단계 가능도 (3.6)에서 β_k ($k=0, 1, \dots, p-1$)와 v_i ($i=1, \dots, q$)에 대한 추정 방정식은 다음과 같이 주어진다.

$$\frac{\partial h}{\partial \beta_k} = -\sigma^{-1} \sum_{ij} (\delta_{ij} - \Lambda_{ij}) x_{ijk} \quad (k=0, 1, \dots, p-1)$$

$$\frac{\partial h}{\partial \sigma} = -\sigma^{-1} \sum_{ij} (\delta_{ij} - \Lambda_{ij}) m_{ijk} - \sum_{ij} \sigma^{-1} \delta_{ij}$$

$$\frac{\partial h}{\partial v_i} = -\sigma^{-1} \sum_j (\delta_{ij} - \Lambda_{ij}) - (v_i/\alpha)$$

여기서 $m_{ij} = (y_{ij} - \mu_{ij})/\sigma$ 이다.

그리고 $\mu_{ij} = E(\log T_{ij} | v_i) = x_{ij}^T \beta + v_i = x_{ij}^T \beta + z_{ij}^T v$ 이다.

이에 따라, 음의 헤시안 행렬 (negative Hessian matrix)는 다음과 같다.

$$H = H(\beta, \sigma, v) = \begin{pmatrix} -\frac{\partial^2 h}{\partial \beta \partial \beta^T} & -\frac{\partial^2 h}{\partial \beta \partial \sigma} & -\frac{\partial^2 h}{\partial \beta \partial v^T} \\ -\frac{\partial^2 h}{\partial \sigma \partial \beta^T} & -\frac{\partial^2 h}{\partial \sigma^2} & -\frac{\partial^2 h}{\partial \sigma \partial v^T} \\ -\frac{\partial^2 h}{\partial v \partial \beta^T} & -\frac{\partial^2 h}{\partial v \partial \sigma} & -\frac{\partial^2 h}{\partial v \partial v^T} \end{pmatrix} \quad (3.9)$$

여기서 $-\frac{\partial^2 h}{\partial \beta_k \partial \beta_s} = \sum_{ij} x_{ijs} (\sigma^{-2} \Lambda_{ij}) x_{ijk},$

$$-\frac{\partial h}{\partial \sigma^2} = \sum_{ij} \sigma^{-2} [m_{ij} \Lambda_{ij} (2 + m_{ij}) - \delta_{ij} (2m_{ij} + 1)],$$

$$-\frac{\partial^2 h}{\partial v_i^2} = \sum_j \sigma^{-2} \Lambda_{ij} + \alpha^{-1},$$

$$-\frac{\partial^2 h}{\partial \beta_k \partial v_i} = \sum_j \sigma^{-2} \Lambda_{ij} x_{ijk} z_{ij\bar{i}},$$

$$-\frac{\partial h}{\partial \beta_k \partial \sigma} = \sum_{ij} \sigma^{-2} [\Lambda_{ij} (1 + m_{ij}) - \delta_{ij}] x_{ijk},$$

$$-\frac{\partial h}{\partial \sigma \partial v_i} = \sum_j \sigma^{-2} [\Lambda_{ij} (1 + m_{ij}) - \delta_{ij}] z_{ij\bar{i}}.$$

따라서 행렬표현은 다음과 같다.

$$\frac{\partial^2 h}{\partial \beta \partial \beta^T} = X^T W X, \quad \frac{\partial^2 h}{\partial v \partial v^T} = Z^T W Z + Q, \quad \text{그리고} \quad \frac{\partial^2 h}{\partial \beta \partial v^T} = X^T W Z \quad \text{이다.}$$

여기서 X 와 Z 는 각각 회귀모수 β 와 변량효과 v 에 대한 x_{ij} 와 z_{ij} 의 모형 행렬이다.

그리고 $W = \text{diag}(\sigma^{-2} \Lambda_{ij})$ 는 가중치 $\sigma^{-2} \Lambda_{ij}$ 를 갖는 대각행렬이고, $Q = \alpha^{-1} I_{q \times q}$ 이다.

산포모수 α 의 추론을 위하여 다음과 같은 조정된 단면 다단계 가능도 (adjusted profile h-likelihood)를 사용한다 (Lee와 Nelder, 2001a).

$$p_{\tau}(h) = \left[h - \frac{1}{2} \log \det \left\{ \frac{H}{2\pi} \right\} \right] \Big|_{\tau = \hat{\tau}} \quad (3.10)$$

여기서 $\tau = (\beta^T, v^T, \sigma)^T$ 이며 $H = -\frac{\partial^2 h}{\partial \tau \partial \tau^T}$ 이다.

α 에 대한 추정 방정식은 다음과 같다.

$$\frac{\partial p_{\tau}(h)}{\partial \alpha} = \frac{\partial h}{\partial \alpha} - \frac{1}{2} \text{tr} \left(\hat{H}^{-1} \frac{\partial \hat{H}}{\partial \alpha} \right) \quad (3.11)$$

여기서 $\frac{\partial h}{\partial \alpha} = -\frac{q}{2\alpha} + \frac{\sum_i \hat{v}_i^2}{2\alpha^2}$ 이며,

$$\hat{H} = H_p(\beta, \sigma, v) \Big|_{\beta = \hat{\beta}, \sigma = \hat{\sigma}, v = \hat{v}} = H(\tau) \Big|_{\tau = \hat{\tau}} \quad (3.12)$$

이다.

$$\text{즉, } \frac{\partial p_{\tau}(h)}{\partial \alpha} = -\frac{q}{2\alpha} + \frac{\sum_i \hat{v}_i^2}{2\alpha^2} - \frac{1}{2} \text{tr} \left(\hat{H}^{-1} \frac{\partial \hat{H}}{\partial \alpha} \right) = 0 \text{ 이다.}$$

따라서 $\hat{\alpha}$ 는 다음과 같이 주어진다.

$$\hat{\alpha} = \frac{\sum_i \hat{v}_i^2}{q - \gamma} \quad (3.13)$$

여기서 $\gamma = -\alpha \text{tr} \left(\hat{H}^{-1} \frac{\partial \hat{H}}{\partial \alpha} \right)$ 이다.

다단계 가능도의 추정절차는 다음과 같은 세 단계의 절차로 수행된다.

단계 1: α 가 주어졌을 때, $\tau = (\beta^T, \sigma, v^T)^T$ 를 추정하기 위하여 (3.6)의 h 를 최대화 하기 위한 뉴튼-랩슨 공식 (Newton-Rapshon formula)를 사용한다.

뉴튼-랩슨 공식은 아래와 같이 주어진다.

$$\begin{pmatrix} \hat{\beta} \\ \hat{\sigma} \\ \hat{v} \end{pmatrix}^{(i+1)} = \begin{pmatrix} \hat{\beta} \\ \hat{\sigma} \\ \hat{v} \end{pmatrix}^{(i)} + (H^{-1})^{(i)} \begin{pmatrix} \frac{\partial h}{\partial \beta} \\ \frac{\partial h}{\partial \sigma} \\ \frac{\partial h}{\partial v} \end{pmatrix}^{(i)} \quad (i=0,1,2,\dots) \quad (3.14)$$

단계 2: $\tau = \hat{\tau}$ 가 주어졌을 때, (3.10)의 $p_\tau(h)$ 를 최대화 해주는 산포모수 α 를 추정한다.

산포모수 α 의 추정량은 (3.13)과 같이 주어진다.

단계 3: $\hat{\tau}$ 의 이전 추정치와 현재 추정치의 차가 최대가 되는 값이 10^{-6} 보다 작을 때 반복을 멈춘다.

본 논문에서 제안된 추정절차에 대한 계산은 R 프로그램을 이용하였다 (부록참조).



4. 실제자료 분석

이 장에서는 신장 감염 자료와 다기관 방광암 임상시험 자료에 대하여 설명하고, 3장에서 개발한 다단계 가능도 추론방법을 이용하여 두 개의 실제 자료를 분석한다.

4.1 신장 감염 자료

신장 감염 자료는 휴대용 투석기를 사용하는 38명의 환자에서 두 번의 신장 감염을 관측한 자료로 구성된다 (McGilchrist와 Aisbett, 1991). 따라서, 동일한 환자 내에서 두 번의 반복적인 관측 시간은 상관관계가 있을 수 있다. 신장 감염은 카테테르 (체내에 삽입하여 소변 등을 뽑아내는 도관) 삽입 위치에서 발생할 수 있다. 만약 사건 (즉 감염)이 발생하지 않으면 중도절단으로 처리 된다. 이 자료의 27.3%는 중도절단으로 처리가 되었다. 이 자료에서 생존시간은 카테테르를 삽입한 후 환자가 감염된 시간을 말한다. 공변량은 성별 (Sex) 과 나이 (Age) 그리고 신장질환의 각각 다른 세 가지 유형의 지시변수로서 사구체신염 (glomerulonephritis; GN), 급성 관형신피증 (acute tubular nephropathy; AN), 다낭성신종 (polycystic kidney disease; PKD)으로 총 다섯가지 변수이다. Table 4.1은 신장 감염 자료 분석에 이용된 변수의 설명이다.

표 4.1 신장 감염 자료 : 변수 설명

Variable	Explanation
id	Patient's id number
Time	Time to kidney infection
Status	Status (0 : no event, 1 : infection)
Age	Years
Sex	Sex (0 : male, 1 : female)
Disease	Disease type(GN, AN, PKD, Other)

신장 감염 자료를 AFT 변량효과 모형에 대하여 다단계 가능도로 추정 한 결과는 다음과 같다.

표 4.2 신장 감염 자료 : AFT 변량효과 모형을 이용한 분석

Variable	Estimate	SE
Intercept	1.983	0.728
Age	-0.002	0.011
Sex : 여자	1.648	0.332
GN1	-0.135	0.397
AN1	-0.534	0.389
PKD	1.238	0.584
$\hat{\sigma}$	0.948	0.982
$\hat{\alpha}$	0.240	

Table 4.2에서 유의한 결과를 가지는 변수는 sex와 PKD이다. 분석 결과에 의하면, 여성이 남성보다 $\exp(1.648) = 5.20$ 배 만큼 감염시간을 유의하게 증가시키며, PKD에 걸린 환자가 다른 종류의 질병에 걸린 환자보다

$\exp(1.238) = 3.45$ 배 만큼 감염시간을 유의하게 증가시킴을 보인다.

신장 감염 자료에서 각 환자 별 변량효과의 추정치 및 신뢰구간은 Figure 4.1과 같다.

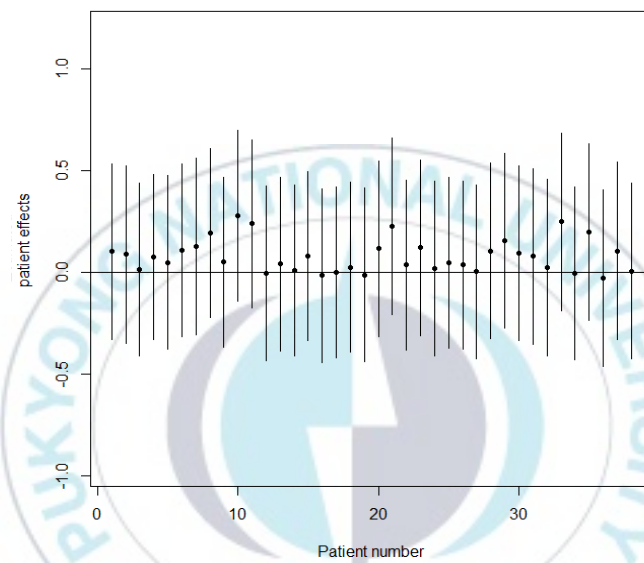


그림 4.1 신장 감염 자료 : 각 환자 별 변량효과의 추정치 및 95% 신뢰구간

변량효과 v_i 의 95% 신뢰구간은 다음과 같이 주어진다 (Ha 등, 2017).

$$\hat{v}_i \pm 1.96 SE(\hat{v}_i - v_i) \quad (i = 1, \dots, 38) \quad (4.1)$$

Figure 4.1은 38명의 환자에 대한 변량효과 v_i 의 추정치 및 변량효과의 95% 신뢰구간을 나타낸다. 위 그림에서 모든 환자의 95% 신뢰구간이 0의 값을 포함하고 있으므로 환자 간 감염시간에 대한 변량효과에 차이가 없음을 보인다.

4.2 다기관 방광암 임상시험 자료

다기관 방광암 임상시험자료는 벨기에 EORTC에서 실시한 임상시험에 참여한 21개 센터에서 모집된 410명의 방광암 환자에게서 수술 후 첫 번째 재발이 일어날 때까지의 시간을 관측한 자료로 구성된다 (Sylvester 등, 2006). 이 자료의 49.8%는 연구 종료 시까지 사건발생이 일어나지 않아서 중도절단으로 처리가 되었다. 공변량은 화학요법 (chemotherapy) 치료지표인 CHEMO와 수술 전 기저 재발률을 나타내는 지표인 TUSTAT으로 총 두가지 변수이다. Table 4.2는 다기관 방광암 임상시험 자료 분석에 이용된 변수의 설명이다.

표 4.3 다기관 방광암 임상시험 자료 : 변수 설명

Variable	Explanation
Center	Institution number of 21 centers
Surtime	Time to the first recurrence from randomization
Status	Status (0 : no event, 1 : recurrence)
CHEMO	Treatment indicator representing chemotherapy (0 : No, 1 : Yes)
TUSTAT	Indicator representing prior recurrent rate (0 ; Primary, 1 : Recurrent)

다기관 방광암 임상시험 자료를 AFT 변량효과 모형으로 추정한 결과는 다음 표와 같다.

표 4.4 방광암 자료 : AFT 변량효과 모형을 이용한 분석

Variable	Estimate	SE
Intercept	6.652	0.193
CHEMO	0.877	0.207
TUSTAT	-0.640	0.175
$\hat{\sigma}$	1.192	0.075
$\hat{\alpha}$	0.272	

Table 4.4에서 유의한 결과를 가지는 변수는 CHEMO와 TUSTAT이다. 분석 결과에 의하면, 화학치료를 받은 환자가 화학치료를 받지 않은 환자보다 $\exp(0.877) = 2.4$ 배 만큼 재발시간을 유의하게 증가시키며, 이전에 재발이 일어난 환자가 재발이 일어나지 않은 환자보다 $\exp(-0.640) = 0.53$ 배 만큼 재발시간이 유의하게 감소됨을 보인다.

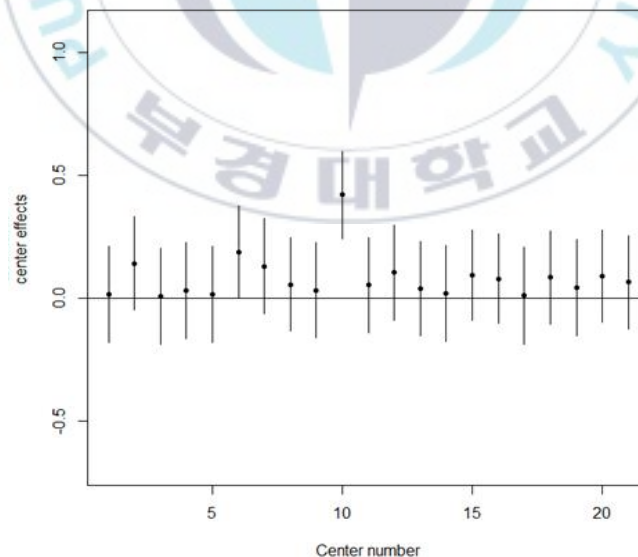


그림 4.2 방광암 자료 : 각 센터 별 변량효과의 추정치 및 95% 신뢰구간

Figure 4.2는 21개 센터에 대한 변량효과 즉, 센터 효과의 추정치 및 변량 효과의 95% 신뢰구간을 나타낸다. 위 그림에서 10번째 센터의 변량효과 95% 신뢰구간이 0의 값을 포함하지 않는다. 이는 10번째 센터가 다른 20개의 센터보다 매우 이질적인 생존시간을 나타내며, 특히 이 센터는 다른 센터보다 재발시간을 길게 해 줌으로 인해 치료가 잘 되는 환자가 상대적으로 많음을 암시한다. 특히 이러한 신뢰구간은 어떤 특별한 센터의 이질성을 파악하는데 매우 유용함을 예증해 줄 수 있다 (Ha 등, 2017).



5. 결론 및 토론

본 논문에서는 오차항의 분포에 대해 극단값 분포를 사용한 AFT 변량효과 모형에서 다단계 가능도 추론법을 제안하였다. 특히 다단계 가능도는 어려운 적분을 요구하지 않고 변량효과에 대한 효율적인 통계적 추론을 제공한다라는 장점이 있다. 그 예증을 위하여 신장 감염 자료와 다기관 방광암 임상시험 자료를 적용하여 분석하였다.

신장 감염 자료의 분석 결과, 유의한 결과를 가지는 변수는 sex와 PKD이다. sex에서 여성의 신장 감염시간이 남성보다 5.2배 증가하고, PKD에 걸린 환자의 신장 감염시간이 다른 종류의 질병에 걸린 환자보다 3.45배 증가됨을 보인다. 그리고 각 환자 신장 감염시간에 대한 변량효과들 간 차이가 없다고 볼 수 있다.

다기관 방광암 임상시험 자료의 분석 결과, 유의한 결과를 가지는 변수는 CHEMO와 TUSTAT이다. CHEMO에서 화학치료를 받은 환자의 방광암 재발시간이 화학치료를 받지 않은 환자보다 2.4배 증가하고, TUSTAT에서 이전에 재발이 일어난 환자의 방광암 재발시간이 이전에 재발이 일어나지 않은 환자보다 0.53배 감소됨을 보인다. 또한, 10번째 센터의 변량효과의 추정치가 다른 병원과 큰 차이를 보이며 재발시간 또한 증가됨을 알 수 있다.

추후 EV 분포를 갖는 AFT 변량효과 모형의 추론에 대한 이론적 연구를 위하여 모의실험 연구가 진행되어야 할 것이며, 모형검토 또한 필요하다고 사료된다.

참고문헌

- [1] Cai, T., Huang, and Tian. L. (2009) Regularized estimation for the accelerated failure time model *Biometrics* 65, 394-404.
- [2] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society-Series B* 34, 187-220.
- [3] Ha, I. D., Jeong, J.H. and Lee, Y. (2017). Statistical modelling of survival data with random effects: h-likelihood approach. Springer.
- [4] Ha, I. D., Lee, Y. and Song, J.-K. (2002). Hierarchical likelihood approach for mixed linear models with censored data. *Lifetime Data Analysis* 8, 163-176.
- [5] Ha, I. D., Noh, M. and Lee, Y. (2012) frailtyHL: a package for fitting frailty models with h-likelihood. *R Journal* 4, 307 - 320.
- [6] Hutton, J. L. and Monaghan, P. F. (2002). Choice of parametric accelerated life and proportional hazard models for survival data: Asymptotic results. *Lifetime Data Analysis* 8, 375-393.
- [7] Klein, J. P. and Moeschberger, M. L. (2003) *Survival analysis : techniques for censored and truncated data*, 2nd edn. Springer.

- [8] Lawless, J. F. (1982). Statistical models and methods for lifetime data. 1st edn. New York: Wiley.
- [9] Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). Journal of the Royal Statistical Society B 58, 619–678.
- [10] Lee, Y., Nelder, J. A. and Pawitan, Y. (2017). Generalised Linear Models with Random Effects: unified analysis via h-likelihood. 2nd edn. London: Chapman and Hall.
- [11] McGilchrist, C. A. and Aisbett, C. W. (1991). Regression with frailty in survival analysis. Biometrics 47, 461–466.
- [12] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. Journal of the Royal Statistical Society A 135, 370–384.
- [13] Park, E. and Ha, I. D. (2018). Penalized variable selection for accelerated failure time models. Communications for Statistical Applications and Methods 25, 591–604.
- [14] Sylvester, R., van der Meijden, A. P. M., Oosterlinck, W., Witjes, J., Bouffieux, C., Denis, L., Newling, D. W. W. and Kurth, K. (2006). Predicting recurrence and progression in individual patients

with stage Ta T1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials. *European Urology* 49, 466–477.

- [15] Zhou, M. (2006). Empirical likelihood analysis of the rank estimator for the censored accelerated failure time model. *Biometrika* 92, 492–498.



부록

R 코드

1) 신장 감염 자료

```
library(survival)
data(kidney)
attach(kidney)

GN1<-ifelse(disease=="GN",1,0)
AN1<-ifelse(disease=="AN",1,0)
PKD1<-ifelse(disease=="PKD",1,0)
sex<-kidney$sex

fit=survreg(Surv(time,status)~age+sex+GN1+AN1+PKD1,
data=kidney, dist="weibull") # AFT model without random effect

summary(fit)

### X matrix
X<-model.matrix(~age+sex+GN1+AN1+PKD1)

### Z matrix
group<-factor(kidney$id)
Z<-model.matrix(~group+0)

### Observable random variables
```

```

n=nrow(X)
p=ncol(X)
q=ncol(Z)

n_i=2
y <- c(time)
del <- c(status)

### Newton Method
# initial value
beta<-c(2.1065,-0.0019,1.6079,-0.0493,-0.5206,1.3425)
sigm<- 0.87
alpha<- 0.05
v<-c(rep(0,q))

iter<-1

repeat{

y_star=log(y)-(X%*%beta+Z%*%v)

mij=y_star/sigm
lamij=(exp(mij))

### 1st derivative
diff1_b=-(t(X)%*%(del-lamij))/sigm
diff1_s=-(sum(del)+sum(mij*(del-lamij)))/sigm
diff1_v=-v/alpha-t(Z)%*%((del-lamij)*mij/sigm)

diff1_theta=rbind(diff1_b,diff1_s,diff1_v)

```

```

### 2nd derivative
W=diag(c(lamij/(sigm^2)))
Q=diag(c(rep(1/alpha,q)))

diff2_b=t(X)%*%W*%X
diff2_s=sum((2*mij*(lamij-del)+(mij^2)*lamij-del)/(sigm^2))
diff2_v=t(Z)%*%W*%Z+Q

diff2_bs=t(X)%*%(lamij+lamij*mij-del)
diff2_bv=t(X)%*%W*%Z
diff2_sv=t(Z)%*%((lamij+lamij*mij-del)*(1/(sigm^2)))

### h-likelihood
H<-rbind(cbind(diff2_b,diff2_bs,diff2_bv ),
          cbind(t(diff2_bs),diff2_s,t(diff2_sv)),
          cbind(t(diff2_bv),diff2_sv,diff2_v ))

theta<-as.matrix(c(beta,sigm,v))

theta_a=theta+solve(H)%*%diff1_theta
theta_a=as.matrix(theta_a)

beta_h<-theta_a[1:p]
sigm_h<-theta_a[p+1]
v_h<-theta_a[(p+2):(p+q+1)]

### Estimation of alpha
PP=matrix(rep(0,p*p),nrow=p)
PS=matrix(rep(0,p*1),nrow=p)
PV=matrix(rep(0,p*q),nrow=p)

```

```

SS=matrix(rep(0,1*1),nrow=1)
SV=matrix(rep(0,1*q),nrow=1)
dQ=diag(c(rep(-1/alpha^2,q)))

dH=rbind(cbind(PP,PS,PV),
          cbind(t(PS),SS,SV),
          cbind(t(PV),t(SV), dQ))

alpha_h=(t(v_h)%*%v_h)/(q+alpha*sum(diag((solve(H)%*%dH))))

  if((max(abs(theta_a-theta))<10^(-8))|(iter>200))(break)
iter<-iter+1
  beta<-beta_h
  sigm<-sigm_h
  v<-v_h
}
iter

beta_h
sigm_h
v_h
alpha_h

#SE

se_beta_hat=sqrt(diag(solve(H)))
SE_beta_h<-se_beta_hat[1:p]
SE_sigm_h<-se_beta_hat[p+1]

result1<-cbind(beta_h,SE_beta_h)
result2<-cbind(sigm_h,SE_sigm_h)

```

```

result<-rbind(result1,result2)
#-----
iter
alpha_h
result
#-----
summary(fit)

### computation of 95% CI for each random effect
iH=solve(H)
var<- diag(iH)[(p+2):(p+q+1)]      # computation of var(v_h-v)
SE<- sqrt(var) # SE(v_h-v)
lb<- v_h -1.96*SE                  # lower bound
ub<- v_h +1.96*SE                  # upper bound
CI<- cbind(lb,ub)                  # computation of CI
id <- 1:q
plot(v_h ~ id, ylim=c(min(lb)-0.5, max(ub)+0.5), ylab="Estimated
      patient effects",xlab="Patient number",pch=20,type="p") # plot for
CI
abline(h=0)
for (i in 1:q){
  x1<- c(i,i)
  y1<- c(lb[i],ub[i])
  lines (y1~x1)
}

```

2) 다기관 방광암 임상시험 자료

```
library(frailtyHL)
data(bladder0)
bla0<-bladder0[bladder0$Surtime!=0,]    # Delete: Survival time with zero
value
bla0$Surtime<- bla0$Surtime
bla0$Status<- bla0$Status
bla0$Center<- bla0$Center
bla0$Chemo<- bla0$Chemo
bla0$Tustat<- bla0$Tustat
attach(bla0)

fit=survreg(Surv(Surtime, Status)~Chemo+Tustat,
            data=bla0, dist="weibull")    # AFT model without random effect
summary(fit)

### X matrix
X<-model.matrix(~Chemo+ Tustat)
#X<-model.matrix(~bla0$Chemo+ bla0$Tustat)

### Z matrix
group<-factor(Center)
Z<-model.matrix(~group+0)

### Observable random variables
n=nrow(X)
p=ncol(X)
q=ncol(Z)

ni <- t(Z)%*%as.matrix(rep(1,n))    #cluster n_i
y <- c(bla0$Surtime)
del <- c(bla0$Status)
```

```

### Newton Method
# initial value
beta<-c(6.8, 0.8, -0.64)
sigm<- 1
alpha<- 0.01
v<-c(rep(0,q))

# 이하 R 코드는 1)의 신장 감염 자료 코드와 동일함.

### computation of 95% CI for each random effects
iH=solve(H)
var<- diag(iH)[(p+2):(p+q+1)] # computation of var(v_h-v)
SE<- sqrt(var) # SE(v_h-v)
lb<- v_h -1.96*SE # lower bound
ub<- v_h +1.96*SE # upper bound
CI<- cbind(lb,ub) # computation of CI
center <- 1:q
plot(v_h ~ center, ylim=c(min(lb)-0.5, max(ub)+0.5), ylab="Estimated
center effects",xlab="Center number",pch=20,type="p") # plot for CI
abline(h=0)
for (i in 1:q){
x1<- c(i,i)
y1<- c(lb[i],ub[i])
lines (y1~x1)
}

```