The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.2.309
28(2), 309–324

Survival Analysis using SRC-Stat Statistical Package

Il Do Ha[a,1] · Maengseok Noh[a] · Youngjo Lee[b] · Johan Lim[b] · Jaeyong Lee[b] · Heeseok Oh[b] · Dongwan Shin[b] · Sanggoo Lee[b] · Jinuk Seo[b] · Yonhtae Park[b] · Sungzoon Cho[b] · Jonghun Park[b] · Youkyung Kim[b] · Kyungsang You[b]

[a]Department of Statistics, Pukyong National University
[b]Data Science for Knowledge Creation Research Center, Seoul National University

(Received March 23, 2015; Revised March 31, 2015; Accepted March 31, 2015)

Abstract

In this paper we introduce how to analyze survival data via a SRC-Stat statistical package. This provides classical survival analysis (e.g. Cox's proportional hazards models for univariate survival data) as well as advanced survival analysis such as shared and nested frailty models for multivariate survival data. We illustrate the use of our package with practical data sets.

Keywords: Cox's proportional hazards models, frailty models, H-likelihood, multivariate survival data, random effects.

1. 서론

SRC-Stat 통계패키지는 서울대학교「데이터과학과 지식창출 연구센터」에서 미래창조과학부와 한국연구재단이 추진하는 선도연구센터 지원사업 및 에스이(랩)과 서울대학교 빅데이터 센터 등의 지원으로 개발한 통계패키지(가제: SRC-Stat)이다. 특히 다양한 분야(의학, 자연과학, 금융, 사회과학 등)의 데이터를 메뉴 형식으로 접근하여 손쉽게 분석할 수 있는 고급형 국산 통계패키지이다. 현재 국내 교육기관에서 무상(다운로드 http://srcdsc.snu.ac.kr/srcstat/)으로 사용할 수 있도록 2013년 9월 베타 버전으로 보급한 이후, 현재까지 다양한 통계적 방법 및 소프트웨어 기술이 계속 업데이트되고 있다.

본 논문에서는 SRC-Stat를 이용하여 생존자료를 분석하는 방법을 소개한다. 생존분석은 의·약학, 유전학, 자연과학, 공학, 사회과학 등 다양한 분야에서 수집되는 생존시간(survival time 또는 time-to-event)자료의 분석에 적용된다. 생존시간은 어떤 정해진 연구시점으로 부터 특정 사건(예: 병의 재발, 기계고장, 퇴사, 고객이탈 등)의 발생시점까지의 시간, 또는 일정한 기저(baseline)시점(예: 수술시점)부터 관심대상의 끝점(endpoint; 예를 들어, 재발이나 사망)까지의 시간으로 정의된다. 생존분석의 정의를 보다 일반화하여 한 마디로 표현하면 "생존시간에 관한 자료를 분석하는 통계적 방법"이라 할 수 있다.

생존자료의 큰 특징 중 하나는 중도절단(censoring)이다. 중도절단의 형태 및 종류는 다양하지만, 여기서는 생존분석에서 가장 많이 사용되는 임의중도절단(random censoring) 되는 경우를 다루기로 한

This research was supported by an NRF grant funded by Korea government (MSIP) (No. 2011-0030810).
[1]Corresponding author: Department of Statistics, Pukyong National University, 45 Yongso-ro, Nam-gu, Busan 608-737, Korea. E-mail: idha1353@pknu.ac.kr

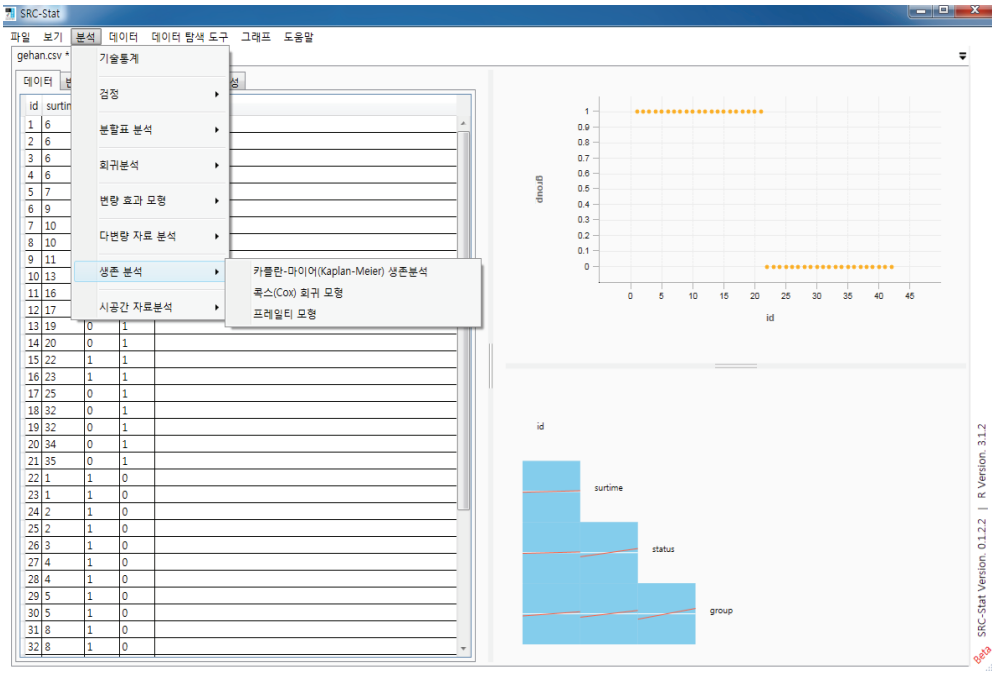

Figure 1.1. Composition of survival analysis in SRC-Stat

다. 따라서 생존자료는 두 변수, 즉 정량적인(continuous) 생존시간 변수와 사건발생에 관한 이분형인(binary) 중도절단여부 변수로 구성되는 불완전 자료(incomplete data)이다. 중도절단성을 갖는 생존자료의 특성 때문에, 정량적인 완전자료(complete data)에 대한 표준적인 방법(예: t-검정, 윌콕슨(Wilcoxon)검정)을 생존분석에 바로 적용하는 것은 모순된 결론에 도달할 수 있다. 생존분석에서 개발된 대부분의 추론 방법들은 이러한 표준적인 방법을 중도절단 자료로 확장시킨 형태이다.

일반적으로 생존분석에 대한 공통적인 관심 추론분야는 다음과 같이 3개 정도로 요약된다:

A) 생존율의 추정: 특정 암 환자가 수술 후 생존할 확률 또는 환자들의 50%가 생존할 시점(즉 median survival time)의 추정 등.

B) 두 집단의 생존분포 비교: 두 처리군(시험군/대조군)의 생존율의 비교 (즉 $H_0 : S_1(t) = S_2(t)$ for all t) 등.

C) 공변량(covariates)의 영향파악: 생존시간이나 위험률에 유의한 영향을 미치는 설명변수 (또는 위험인자)들의 조사 등.

위의 세 분야에 대해 가장 표준적으로 사용하는 비모수적 분석법으로, A)에 대해서는 카플란-마이어 (Kaplan과 Meier, 1958)의 생존율 추정법을, B)에 대해서는 로그순위(log-rank) 검정법 (Mantel-Haenszel, 1959) 그리고 C)에 대해서는 콕스 (Cox, 1972)의 비례위험(proportional hazards; PH)모형 추론법이다. 사실 이러한 저자들은 20세기 가장 괄목할만한 생존분석법을 제시하였으며, 특히 임상시험에서 가장 지대한 영향을 끼친 방법이라 할 수 있다 (Fleming와 Lin, 2000). 나아가, 노르웨이의 저명한 통계학자 Aalen (1975)은 counting-process 마팅게일(martingale) 이론을 개발하여 생존분석에서

소표본 및 대표본에 대한 통합적 이론을 제시하였다. 본 논문에서는 SRC-Stat를 통해 A)–C)에 대해 분석하는 방법을 다음 절에서 소개하고자 한다.

통상적으로 생존자료 분석은 한 개인에게 사건이 단 한 번 발생(즉 single event)하는 경우의 독립인 생존자료에 대한 분석법을 주로 다룬다. 하지만 최근에는 한 환자에 대한 여러 번의 사건 발생으로 얻어지는 상관된(correlated) 생존자료(예: 재발시간) 분석에도 많은 관심이 모아지고 있다. 또한 사람의 기관(organ; 눈, 신장, 폐 등) 및 임상시험 참여기관(center)이나 쌍둥이 및 가족군과 같이 하나의 군(cluster)으로 부터 얻어지는 생존시간의 자료는 통상적으로 상관되어져 있다고 할 수 있다. 이러한 상관된 형태의 자료(즉, multiple or clustered event-time data)를 다변량 생존자료(multivariate survival data 또는 correlated survival data)라 부른다. 하지만, 자료 분석 상 제약(예: 상관성, 이질성, 중도절단성)으로 인해 그 추론법이 다소 복잡하며 최근에 많은 연구가 진행되고 있다. 이에 대한 대표적인 분석모형으로는 콕스모형에 변량효과(random effect)의 한 형태인 프레일티(frailty; Vaupel 등, 1979)를 허락하여 그 상관성 및 이질성을 동시에 모형화 하는 프레일티 모형이 최근에 매우 폭넓게 연구 되어져 오고 있다 (Nielsen 등, 1992; Hougaard, 2000; Duchateau와 Janssen, 2008). 특히, 다변량 생존자료에 대해 콕스 모형의 적합은 회귀모수에 대해 심각한 과소추정을 이끌 수 있다 (Andersen 등, 1997; Ha 등, 2001). 프레일티 모형의 추론을 위해 프레일티에 대해 적분을 통해 제거함으로써 얻어지는 주변우도(marginal likelihood)를 사용할 수 있지만, 이 방법은 종종 어려운 적분문제에 직면하게 된다 (Vaida와 Xu, 2000). 이를 극복하기 위해 Lee와 Nelder (1996)는 적분이 요구되지 않는 다단계우도(hierarchical likelihood; h-likelihood)를 다단계 일반화선형모형(hierarchical generalized linear models; HGLMs)의 추론에서 처음으로 소개하였으며, 다양한 변량효과모형에서 통합된 효율적인 추론절차를 제공하는 장점이 있다 (Lee 등, 2006). Ha 등 (2001)은 이러한 다단계우도를 프레일티 모형으로 확장하였다. 따라서 본 패키지에서는 프레일티 모형의 분석을 위해 다단계우도 방법을 사용한다.

본 패키지에서 제공되는 생존분석법은 Figure 1.1에서 제시한 바와 같이 크게 세 분야로서 카플란-마이어, 콕스회귀모형 그리고 프레일티 모형이다. 이를 위한 본 논문의 구성 체계는 다음과 같다. 2절에서는 기초적 생존분석법으로 카플란-마이어의 생존율 추론 및 콕스의 비례위험모형을 소개하고, 3절에서는 다변량 생존자료의 분석에 폭 넓게 사용되고 있는 프레일티 모형 분석법을 다룬다. 마지막으로 4절에서는 토론과 추후 연구과제에 대하여 논의하고자 한다.

2. 기초적 생존분석: 단변량 생존자료

본 패키지는 단변량 생존자료(univariate survival data) 분석을 위한 대표적인 두 가지 비모수적 생존분석법(즉, 카플란-마이어 생존율 분석법과 콕스 비례위험모형 분석)을 제공한다. 여기서 생존시간들은 서로 독립이고 중도절단시간은 생존시간과 독립이라고 가정한다.

2.1. 카플란-마이어(Kaplan-Meier) 생존분석

SRC-Stat 통계패키지를 이용한 카플란-마이어 생존분석의 절차 및 사용방법의 설명은 다음과 같다.

1) 기본개념

카플란-마이어 방법은 각 개체(subject)의 생존시간에 대한 생존함수(즉 생존율)를 추론하는 비모수적인 방법이다. 생존율의 카플란-마이어 추정량(또는 누적한계 추정량; product limit estimator)은 생존구간을 랜덤구간으로 취급하여, 연구대상자 중 사건이 관측되는 시점마다 생존율이 점프(jump)를 갖으며 중도절단 시 그 생존율이 불변인 단계(step)함수이다. 여기서 중도절단이 전혀 없는 경우 이 추정

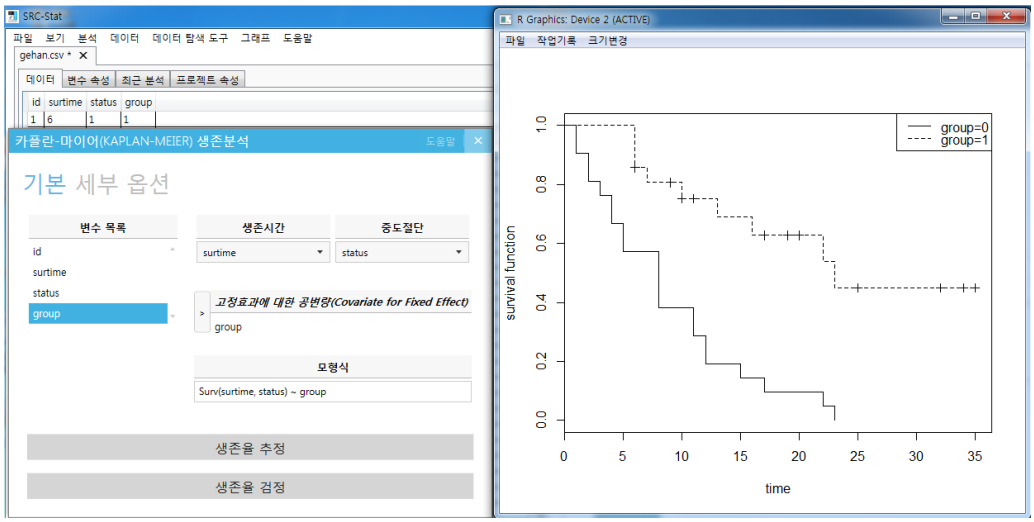

Figure 2.1. Kaplan and Meier survival analysis for comparing two groups for the Gehan data

량은 경험적(empirical) 생존함수가 된다. 특히 두 군간 생존율 비교가 임상시험 등 의학연구에서 자주 사용되며, 여기서 두 군의 카플란-마이어 생존율의 그림과 두 군의 생존율 검정(보통 로그-순위 검정)이 함께 널리 사용된다. 이 방법은 표본수가 적어도 사용 가능하며, 카플란-마이어 추정량의 일치성(consistency)과 점근정규성(asymptotic normality)과 같은 대표본 성질을 만족 (Breslow와 Crowley, 1974)하기 때문에 생존분석에서 기본적으로 가장 많이 사용되는 방법 중 하나이다.

2) 카플란-마이어 추정량의 형태

생존율 $S(t)$에 대한 카플란-마이어 추정량 $\widehat{S}(t)$은 다음과 같이 정의된다:

$$\widehat{S}(t) = \prod_{k:y_{(k)} \le t} \left\{ 1 - \frac{d_{(k)}}{n_{(k)}} \right\},$$

여기서 $y_{(k)}$는 k $(k = 1, \ldots, D)$번째로 사건이 관측되는 생존시간, $n_{(k)}$는 $y_{(k)}$에서의 위험대상자의 수, 그리고 $d_{(k)}$는 $y_{(k)}$에서의 동점(ties)의 수이다.

3) 프로그램 사용법

카플란-마이어 생존분석을 위한 프로그램의 단계별 사용법 절차는 다음과 같다.

단계 1: Figure 2.1의 메뉴의 **분석**에서 "생존분석" 클릭 후 "카플란-마이어(Kaplan-Meier) 생존분석"을 선택한다.

단계 2: Figure 2.1의 **기본**에서 "생존시간"과 "중도절단"에 대응하는 변수를 각각 선택한 후, "변수 목록"에서 그룹변수를 선택하여 "고정효과에 대한 공변량"으로 옮긴다.

단계 3: Figure 2.1의 좌측하단에서 "생존율 추정"을 클릭하여 카플란-마이어 생존율 추정 및 생존곡선 그래프의 결과를 얻는다. 그 다음으로 "생존율검정"을 클릭하여 log-rank test(디폴트)의 결과를 얻는다. 만약 다른 검정이 필요할 경우 "세부옵션"에서 "가중치 지정법"에 있는 다른 검정법(generalized Wilcoxon test 또는 weighted log-rank test)을 선택할 수 있다.

4) 기본사항

Figure 2.1의 좌측에 있는 기본사항의 설명은 다음과 같다.

- 생존시간: 분석대상이 되는 생존시간 (예: surtime 또는 time)을 지정한다.
- 중도 절단: 생존시간에 대응하는 중도절단 여부를 나타내는 중도절단 지시함수(censoring indicator; 보통 status, censor 또는 event)를 지정하며, 통상적으로 중도절단을 "0"으로 표시한다.
- 고정효과에 대한 공변량: 생존시간에 대한 설명변수로서 그룹변수를 지정한다(보통 x로 표시).
- 모형식: 하나의 survival object로서 이 함수는 기본적으로 `Surv(time, status)~x`로 사용 할 수 있다. 여기서 time은 생존시간을, status는 중도절단 여부를, 그리고 x는 그룹을 각각 나타내는 변수이다.
- 생존율 추정: 생존함수의 카플란-마이어 추정값, 표준오차 및 95% 신뢰하한/상한값을 제공한다. 각 군별 카플란-마이어 생존곡선을 그래프로 또한 제시한다.
- 생존율 검정: 두 군간(세 군이상 도 가능) 생존함수의 동등성 여부를 검정하며, 디폴트는 로그-순위 검정이다. 통상적으로 두 군간 생존곡선이 비례할 경우 로그-순위 검정를 사용하며, 그렇지 않은 경우 일반화된 윌콕션검정(generalized Wilcoxon test 또는 Gehan test)를 사용할 수 있다.

5) 선택사항

Figure 2.1의 좌측의 **세부옵션**에 있는 군간 생존율 비교를 위한 가중치 지정법은 다음과 같다.

가중치 지정법: 군간 생존율 차이의 검정에서 사전 가중치를 지정한다. 그 형태는 $\widehat{S}(t)^{\rho}$ (Harrington과 Fleming, 1982)로서, $\rho = 0$이면 로그-순위검정(또는 Mantel-Haenszel 검정), $\rho = 1$이면 Gehan검정 (Gehan, 1965), 그리고 $\rho = 0.5$이면 가중 로그-순위검정(또는 Tarone-Ware 검정) (Tarone과 Ware, 1977)를 각각 의미한다.

6) 결과(Output)

Figure 2.1의 최측 하단에 있는 "생존율 추정" 및 "생존율 검정"을 클릭함으로써 나타나는 결과의 설명은 다음과 같다.

- time: 사건이 관측된 생존시간 (즉 $y_{(k)}$)
- n.risk: 각 시점에서의 위험대상자 수 (즉 $n_{(k)}$)
- n.event: 각 시점에서의 동점(ties)의 수 (즉 $d_{(k)}$)
- surv: 카플란-마이어 생존율의 추정치
- std.err: 카플란-마이어 추정량의 표준오차
- lower (upper): 카플란-마이어 추정량의 95% 신뢰구간의 하한값(상한값)
- p: 근사적인 카이제곱 검정(디폴트는 로그순위 검정)의 p-값을 보여준다.

예제 2.1: 다음은 Gehan (1965)에 의해 제시된 임상시험 생존자료로서, R 패키지(library(MASS); data(gehan))에서 자료를 다운로드 할 수 있다. 이 시험은 급성백혈병 환자에 대해 6-메르캅토퓨린(6-mercaptopurine; 6-MP)이라는 신약의 효과를 알아보는 것이 목적이다. 이를 위해 한 군(21명)은 6-MP를 다른 한 군은 가짜약(placebo; 21명)을 무작위배정(randomization)한 후, 환자의 호전(remission)기간(단위: 주)을 아래와 같이 관측하였다. 여기서 "+"는 우측 중도절단(right censoring)을 나타낸다.

- 처리군(6-MP): 6 6 6 6^+ 7 9^+ 10 10^+ 11^+ 13 16 17^+ 19^+ 20^+ 22 23 25^+ 32^+ 32^+ 34^+ 35^+
- 대조군(placebo): 1 1 2 2 3 4 4 5 5 8 8 8 8 11 11 12 12 15 17 22 23

하나의 관심사항은 두 군간 생존율 (즉 호전율)에 차이가 있는지를 조사하는 것이다. Figure 2.1의 메뉴에서 "데이터"와 "데이터 가져오기"를 클릭하여 데이터 오픈을 먼저 한 후, Figure 2.1의 좌측에서 생존시간은 "surtime", 중도절단은 "status", 고정효과에 대한 공변량은 "group"으로 지정하면 된다. 다만 대조군은 "group = 0" 처리군은 "group = 1"로 코딩되어 있다. Figure 2.1의 우측은 두 군의 추정된 카플란-마이어 생존율을 나타내는 그림으로서, 두 군의 생존율은 평행함을 보이며 특히 처리군이 대조군보다 훨씬 높은 생존율이 보임을 알 수 있다. 비록 그림으로 제시 되어 있지만 두 군 생존율 비교에 대한 검정으로서 디폴트인 로그순위 검정의 p-값은 거의 0에 가까우며, 세부옵션에 있는 다른 두 방법(gehan 및 Tarone-Ware 검정법)도 이와 유사한 p-값을 준다. 따라서 두 군은 유의한 차이가 있음을 알 수 있으며, 처리군이 대조군보다 훨씬 긴 호전기간을 준다는 사실을 파악할 수 있다.

2.2. 콕스의 비례위험모형

2.1절에서는 서로 다른 군간 생존시간의 차이를 검정했지만, 생존시간에 영향을 미치는 공변량의 효과를 파악하는 것 또한 생존분석에서 매우 중요하다. 이를 위해 생존시간의 위험률(hazard rates)과 공변량간의 관계를 묘사하는 콕스의 비례위험 회귀모형이 자주 사용된다. 본 통계패키지를 이용한 콕스회귀모형의 분석절차 및 사용방법은 아래와 같다.

1) 기본개념

콕스모형은 각 개인의 위험률과 공변량간의 관계를 설명하기 위한 대표적인 생존분석 회귀모형으로서, 위험률에 영향을 미치는 공변량의 효과(예: 처리효과 또는 예후요인효과)를 추론하는 것이 주요한 사용목적 중 하나이다. 특히 이 모형은 비례위험성의 가정과 준모수적(semi-parametric)모형이라는 두 가지 큰 특성을 갖고 있다.

Cox (1972)는 기저위험함수에 대한 아무런 정보 없이 회귀모수를 추정하기 위한 편우도(partial likelihood)를 소개하였다. Andersen과 Gill (1982)은 counting process이론을 사용하여, 편우도를 최대로 하는 회귀모수의 추정량은 일치성과 점근정규성과 같은 대표본 성질을 만족한다는 사실을 증명하였다. 나아가, Johansen (1983)은 콕스의 편우도는 단면우도(profile likelihood)가 된다는 사실도 보였다.

2) 모형의 형태

콕스회귀모형의 형태는 각 개인의 생존시간에 대한 위험률과 공변량간의 관계가 아래와 같이 표현된다:

$$\lambda(t;x) = \lambda_0(t)\exp\left(x^T\beta\right), \tag{2.1}$$

여기서 $\lambda_0(t)$는 미지의 기저(baseline)위험함수로서, p개의 공변량들의 벡터 $x = (x_1, \ldots, x_p)^T$의 모든 값들이 0일 때의 위험함수이다. x에 대응하는 회귀모수들의 벡터 $\beta = (\beta_1, \ldots, \beta_p)^T$는 주요한 관심추론의 대상이다. 이 모형의 주요한 특징은 다음과 같다:

i) 준 모수적 모형(Semi-parametric model): $\lambda_0(t)$의 형태는 모르지만(non-parametric), 공변량의 함수 형태는 $\exp(x^T\beta)$로 안다(parametric).

ii) 비례위험성(PH): 두 군간 위험률의 비, 즉 상대위험(relative risk; RR)이 시간에 관계없이 일정한 상수에 비례(proportional)한다.

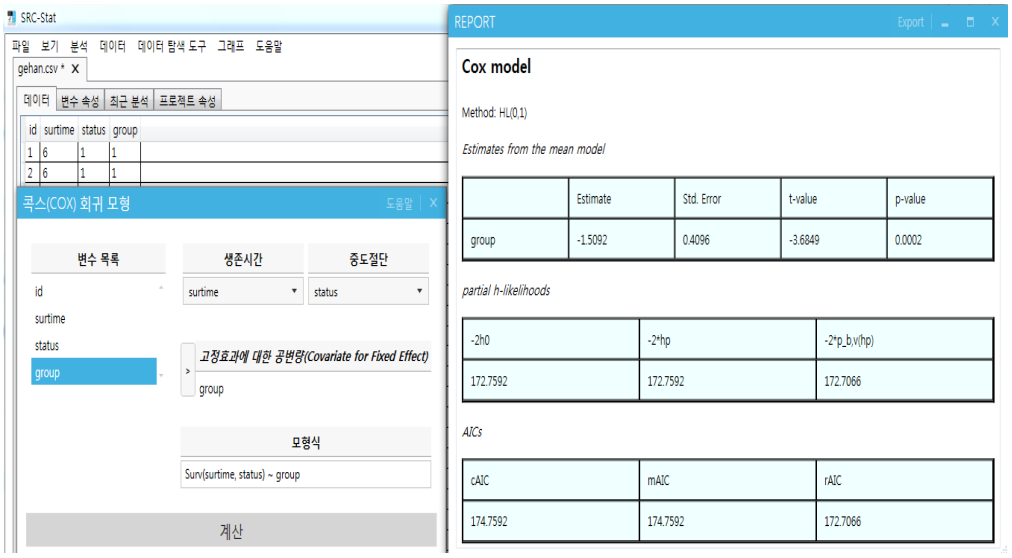

Figure 2.2. Results of fitting Cox's PH models for the Gehan data

iii) 무 절편항(No intercept term): $\lambda_0(t)$가 절편항 역할을 하기 때문에, $x^T\beta$는 동일성(identifiability) 문제로 인해 절편항을 포함하지 않는다.

특히 콕스모형은 비례위험이라는 가정으로 인해 관심모수 의 해석이 쉽다. 하나의 예로서, 편의상 공변량이 하나인 군(즉 시험군은 $x = 2$, 대조군은 $x = 1$로 표현)을 고려해 보자. 한 명은 시험군에 다른 한 명은 대조군에 배정되어 있는 두 개인에 대한 위험률의 비(RR)는 $h_0(t)$가 서로 상쇄되어 없어지기 때문에 다음과 같이 시간에 대해서 상수가 됨을 알 수 있다:

$$\mathrm{RR}(t;x) = \frac{\lambda_2(t)}{\lambda_1(t)} = \frac{\exp(2\beta)}{\exp(\beta)} = \exp(\beta), \tag{2.2}$$

여기서 $\lambda_j(t) = \lambda(t; x = j)$이다. 따라서 "시험군은 대조군에 비해 $\exp(\beta)$만큼 위험이 증가 (if $\beta > 0$) 또는 감소 (if $\beta < 0$)한다"라고 해석할 수 있다. 특히, 공변량이 이분형(binary)이고 생존자료에 동점(ties)이 없는 경우 회귀모수의 추론결과는 PH가정을 가지는 로그-순위 검정결과와 동일하다.

3) 프로그램 사용법

콕스 회귀분석을 위한 프로그램의 단계별 사용법 절차는 다음과 같다.

단계 1: Figure 2.2의 메뉴의 **분석**에서 "생존분석" 클릭 후 "콕스(Cox) 회귀 모형"을 선택한다.

단계 2: Figure 2.2의 좌측에서 "생존시간"과 "중도절단"에 대응하는 변수를 각각 선택한 후, "변수 목록"에서 공변량들을 선택하여 "고정효과에 대한 공변량"으로 옮긴다.

단계 3: Figure 2.2의 좌측 하단에서 "계산"을 클릭하여 적합된 추론결과를 얻는다.

4) 기본사항

Figure 2.2의 좌측에 있는 기본사항에서 생존시간과 중도절단은 2.1절과 같다.

- 고정효과에 대한 공변량: 생존시간에 대한 설명변수로서 그룹변수를 포함한 다양한 독립변수들을 지정한다(보통 x로 표시).
- 모형식: 하나의 survival object로서 이 함수는 기본적으로 `Surv(time, status)~x`로 사용 할 수 있다.

5) 결과(Output)

Figure 2.2의 우측에 있는 결과의 설명은 다음과 같다. 콕스모형에서는 프레일티 항이 없으므로, 아래에 제시한 콕스모형의 다단계우도(h-likelihood)는 Breslow의 편우도(partial likelihood)가 된다. 다단계우도와 관련된 자세한 설명은 3.1절에서 다루기로 한다.

- Estimate: 적합 된 콕스모형의 회귀모수 추정치
- Std. Error: 대응하는 표준오차
- t-value: t-통계량(= Estimate/Std. Error)
- p-value: t-통계량의 유의확률
- partial h-likelihoods: partial h-likelihoods의 결과를 제시한다.
- $-2h0$: $-2 *$ (partial h-likelihood의 첫 번째 component의 값), 콕스모형에서 partial h-likelihood의 첫 번째 component는 Breslow의 partial likelihood와 동일하다.
- $-2hp$: $-2 *$ (partial h-likelihood의 값)이지만, 콕스모형에서는 프레일티 항이 없으므로 $-2hp$는 $-2h0$와 같다
- $-2p_{b,v}(hp)$: $-2 *$ (adjusted partial h-likelihood의 값), 콕스모형에서는 로그프레일티(log-frailty) v가 0이므로 $-2p_{b,v}(hp)$는 $-2p_b(hp)$와 동일하다.
- AICs: 적합된 모형에 대해 세 가지 종류의 Akaike 정보기준(AIC; Akaike information criteria)를 제시한다. 모형선택(model selection)을 할 경우 고려된 모형들 중 가장 작은 AIC값을 주는 모형을 선택한다.
- $\text{cAIC} = -2h0 + 2\text{dfc}$: 자유도 dfc를 갖는 conditional AIC, 콕스모형에서는 dfc = 회귀모수의개수.
- $\text{mAIC} = -2hp + 2\text{dfm}$: 자유도 dfm을 갖는 marginal AIC, 콕스모형에서는 cAIC와 동일하다.
- $\text{rAIC} = -2p_{b,v}(hp) + 2\text{dfr}$: 자유도 dfr를 갖는 restricted AIC, 콕스모형에서는 dfr = 0.

예제 2.2: 편의상 예제 1의 생존자료에 대해 콕스모형을 적합한 결과는 Figure 2.2의 우측에 표로 정리되어 있다. 여기서 β에 관한 해석은 다음과 같다:

i) 군의 효과인 β는 매우 유의하다 ($p = 0.0002$). 다시 말하면 치료군($x = 1$)은 대조군($x = 0$)에 비해 위험률이 $\exp(-1.5092) = 0.221$배, 즉 78% 위험이 유의하게 감소한다.

ii) 식 (2.2)의 상대위험도의 95% 신뢰구간은 $\exp(-1.5092 \pm 1.96 * 0.4096) = (0.099, 0.493)$이다.

3. 프레일티 모형: 다변량 생존자료

이 절에서는 콕스모형을 다변량 생존자료분석으로 확장한 프레일티 모형(frailty models)의 사용법 및 결과의 해석 등을 소개한다. 현재 본 패키지는 두 종류의 프레일티 모형으로, 공통(shared)프레일티 모형과 지분(nested or multi-level)프레일티 모형에 대해 적합이 가능하다. 여기서는 프레일티가 주어질 때, 생존시간들은 서로 독립이고, 중도절단시간은 생존시간과 독립이며 비정보적(non-informative) 이라고 가정한다.

3.1. 공통 프레일티 모형

1) 개념

프레일티 모형은 각 개체(subject)나 군집(cluster)의 위험률에 대한 콕스의 준모수적 비례위험모형 (2.1)에 프레일티(frailty)를 허락한 하나의 확장된 생존분석 회귀모형으로서, 주로 다변량생존자료(multivariate or correlated survival-time data)의 회귀분석에 사용되고 있다. 여기서 프레일티는 각 개체의 위험률에 승법적으로 영향을 미치는 관측 안 되는 변량효과(unobserved random effect)를 의미한다. 프레일티의 분포는 통상적으로 로그정규(log-normal) 또는 감마(gamma)분포를 지정한다. 특히 콕스모형은 프레일티 모형에서 모든 로그프레일티(log-frailty)의 값이 0(즉 로그프레일티의 분산이 0)인 경우에 해당되므로 프레일티 모형을 이용해서 바로 적합할 수 도 있다.

2) 모형의 형태

각 개인의 공통된 프레일티 u가 주어질 때, 생존시간의 조건부 위험률은 다음과 같이 표현된다:

$$\lambda(t|u;x) = \lambda_0(t)\exp\left(x^T\beta\right)u, \tag{3.1}$$

여기서 $\lambda_0(t)$는 미지의 기저(baseline)위험함수로서, p개의 공변량들의 벡터 $x = (x_1, \ldots, x_p)^T$의 모든 값들이 0일 때의 위험함수이다. 프레일티 u는 감마분포 또는 로그정규분포를 지정할 수 있다. 만약 모든 개인들의 $u = 1$이면, 프레일티 모형은 콕스의 비례위험 모형이 된다. 특히 감마 프레일티모형과 로그정규 프레일티모형은 프레일티 분포에 대해 각각 감마분포와 로그정규분포를 가정하기 때문에, 프레일티 u에 대해서는 평균이 1이고 분산이 α인 감마분포를, 그리고 로그프레일티 $v\,(= \log u)$에 대해서는 정규분포, 즉 $v \sim N(0, \alpha)$를 각각 지정한다.

3) 다단계 우도

프레일티 모형 (3.1)에 대한 다단계 우도 (Lee와 Nelder, 1996; Ha 등, 2001)의 정의는 다음과 같다:

$$h = h(\beta, \lambda_0, \alpha) = \ell_0 + \ell_1,$$

여기서 $\ell_0 = \log f(y, \delta|u; \beta, \lambda)$는 로그 프레일티 u가 주어질 때 관측되는 확률변수 (y, δ)의 조건부 로그우도(conditional log-likelihood)이고, $\ell_1 = \log f(v; \alpha)$는 로그 프레일티 v의 로그우도 이다. 다만 y는 관측되는 생존시간이며 δ는 중도절단 여부를 나타내는 지시함수이다. 하지만 $\lambda_0(t)$의 함수 형태를 전혀 모르기 때문에 관심 모수 β의 추론을 위해 λ_0를 제거한 단면 다단계우도(profile h-likelihood) h^*를 사용한다. 특히 h^*는 벌점화 편우도(penalized partial likelihood; Therneau와 Grambsch, 2000) h_p와 동치관계이다 (Ha 등, 2001). 따라서 본 논문에서는 편의상 hp를 편 다단계우도(partial h-likelihood)라 부른다. 나아가 프레일티 모수(즉 산포모수) α의 추론을 위해 조정된 단면 다단계우도(즉 조정된 편 다단계우도) $p_{b,v}(hp)$를 사용한다 (Ha와 Lee, 2003, 2005).

4) 프로그램 사용법

프레일티 모형분석을 위한 프로그램의 단계별 사용법 절차는 다음과 같다.

단계 1: Figure 3.1의 메뉴의 **분석**에서 "생존분석" 클릭 후 "프레일티 모형"을 선택한다.

단계 2: Figure 3.1의 **기본**에서 "생존시간", "중도절단"에 대응하는 변수를 각각 선택한 후, "변수 목록"에서 공변량들을 선택하여 "고정효과에 대한 공변량"과 "변량효과에 대한 공변량"으로 각각 옮긴다.

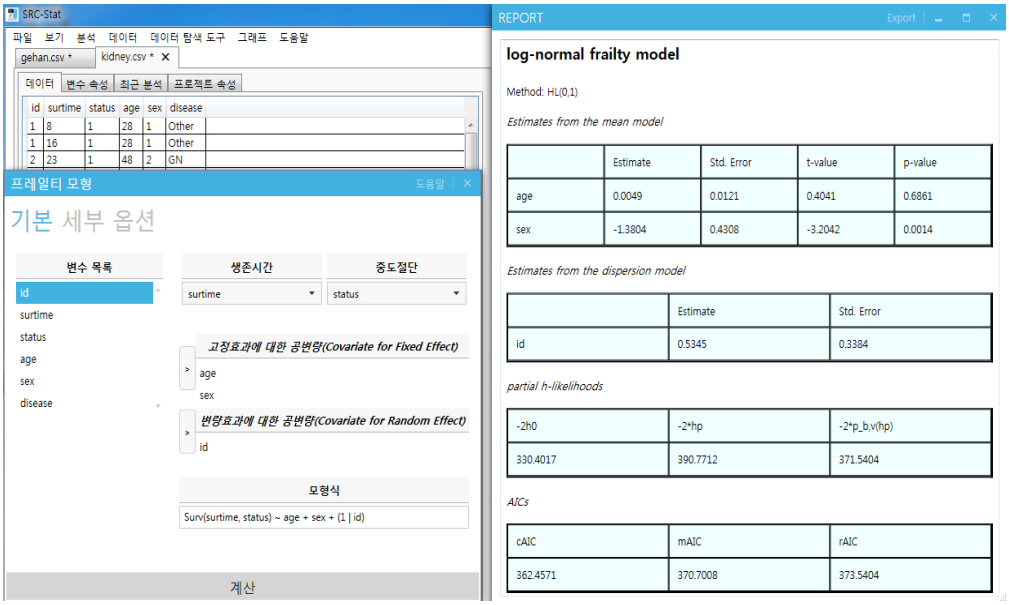

Figure 3.1. Results of fitting lognormal shared frailty models for the kidney infection data

단계 3: Figure 3.1의 좌측 하단에서 "계산"을 클릭하여 적합된 추론결과를 얻는다. 디폴트는 로그정규 프레일티모형이며 그 추정방법은 HL(0, 1)이다.

프레일티 분포를 감마로 지정할 경우 Figure 3.1의 **세부옵션**의 선택사항 "RandDist"에서 "Gamma"를 클릭해야 하며, 모수 추정방법을 변경할 경우 선택사항의 "평균에 대한 라플라스 차수"에서 0 또는 1를, "산포에 대한 라플라스 차수"에서 1 또는 2를 지정할 수 있다. 이에 대한 이론적 설명은 Ha 등 (2012)을 보길 바란다.

5) 기본사항

Figure 3.1의 좌측에 있는 기본사항에서 생존시간, 중도절단 및 고정효과에 대한 공변량은 2.2절의 콕스 모형과 같으며, 단지 모형식 지정에 있어 프레일티 항에 대한 추가사항이 포함된다.

- 모형식: 하나의 survival object로서 이 함수는 기본적으로 `Surv(time, status)~x` +(1|id)로 사용할 수 있다. 여기서 time은 생존시간을, status는 censoring 여부를, x는 공변량이며 id는 각 개체의 식별자(identifier)이다.

6) 선택사항

Figure 3.1의 **세부옵션**에 있는 선택사항의 설명은 다음과 같다.

- RandDist: 프레일티 분포를 지정한다. 정규분포(디폴트)와 감마분포를 지정할 수 있다.
- 평균에 대한 라플라스 차수: 평균모수(mean parameters)를 적합 시키기 위한 리플라스 근사의 차수(order)로서 0 또는 1을 지정한다(디폴트는 0)
- 산포에 대한 라플라스 차수: 산포(dispersion) 모수(보통 프레일티 분산)를 적합 시키기 위한 리플라스 근사의 차수로서 1 또는 2를 지정한다(디폴트는 1)

- 특히, 로그정규 프레일티모형인 경우 HL$(0,1)$(즉 mord = 0, dord = 1을 이용한 h-likelihood 방법), 감마프레일티 경우 HL$(0,2)$(즉 mord = 0, dord = 2를 이용한 h-likelihood 방법)를 주로 사용한다. 하지만 heavy censoring이거나 프레일티의 분산이 큰 경우 로그정규 프레일티에서는 HL$(1,1)$을, 감마프레일티에서는 HL$(1,2)$의 사용을 추천한다.
- 변량효과 분산 고정: "변량효과 분산 고정"의 왼쪽에 있는 조그마한 박스를 클릭하면 프레일티의 분산이 하나의 값으로 고정된다. 이를 클릭하지 않으면 프레일티의 분산이 추정된다.
- 변량효과 분산 초기치: 프레일티의 분산의 초기치를 지정한다(디폴트 = 0.1). 만약 프레일티 모형에서 콕스모형을 적합할 경우, "변량효과 분산 고정"의 박스를 클릭하고 "변량효과 분산 초기치"를 0으로 지정하면 된다.

7) 결과(Output)

Figure 3.1의 우측에 있는 결과의 설명은 다음과 같다.

- Estimates from the mean model: hazard part에서 적합된 회귀모수 추정치와 그 결과를 제공한다.
- Estimate: 적합된 프레일티모형의 회귀모수 추정치
- Std. Error: 대응하는 표준오차
- t-value: t-통계량(= Estimate/Std. Error)
- p-value: t-통계량의 유의확률
- Estimates from the dispersion model: 산포part에서 프레일티 모수 추정치와 그 결과를 제공한다(Estimate: 적합된 프레일티 모수의 추정치; Std. Error: 대응하는 표준오차).
- partial h-likelihoods: partial h-likelihoods의 결과를 제시한다.
- $-2h0$: $-2*$(partial h-likelihood의 첫 번째 component의 값)
- $-2hp$: $-2*$(partial h-likelihood의 값)
- $-2p_{b,v}(hp)$: $-2*$(adjusted partial h-likelihood의 값)
- $-2s_{b,v}(hp)$: $-2*$(adjusted partial h-likelihood의 값 + the second-order term)
- AICs: 적합된 모형의 세 가지 종류의 Akaike Information Criteria를 제시한다
- cAIC $= -2h0 + 2$dfc: 자유도 dfc를 갖는 conditional AIC
- mAIC $= -2p_v(hp) + 2$dfm: 자유도 dfm을 갖는 marginal AIC
- rAIC $= -2p_{b,v}(hp) + 2$dfr: 자유도 dfr를 갖는 restricted AIC

여기서 dfc는 회귀모수와 변량효과를 추정하기 위한 조정된 자유도, dfm은 모수(즉 회귀모수와 산포모수)의 개수, 그리고 dfr은 산포모수의 개수이다. HL$(0,2)$ 또는 HL$(1,2)$를 이용하는 감마프레일티 모형의 경우 mAIC와 rAIC의 계산에서는 2차 라플라스 근사법을 이용한다. 프레일티 모형에 대한 위의 세 가지 AIC들의 자세한 설명은 Ha 등 (2007, 2012)을 보길 바란다.

예제 3.1: 공통 프레일티모형의 예증으로 McGilchrist와 Aisbett (1991)에 의해 제시된 신장 감염시간(kidney infection time) 자료를 고려하자. 이 자료는 R 패키지(library(survival); data(kidney))에서 자료를 다운로드 할 수 있다. 38명 각 환자에 대해 1차 및 2차 감염시간(surtime)을 측정하였으며 약

24% 중도절단(status))을 가진다. 관심 있는 공변량은 나이(age), 성별(sex; 남자 '1', 여자 '2'로 코딩) 및 병의 종류(disease; GN, AN, PKD, other로 4개 범주)이다. 동일한 환자에게 두 개의 감염시간을 가지므로 환자내 자료간 상관성이 있으므로, 이를 모형화 하기 위해 공통된(shared) 환자의 효과를 묘사하는 프레일티를 사용할 수 있다. 나이와 성별을 가지는 로그정규 프레일티 모형을 적합한 결과는 Figure 3.1의 우측에 표로 정리되어 있다. 이에 대한 해석은 다음과 같다:

(i) 나이의 효과는 유의하지 않지만(p-값 $= 0.6861$), 성별효과는 매우 유의하다(p-값 $= 0.0014$). 특히 여자는 남자에 비해 감염 위험률이 $\exp(-1.3804) = 0.251$배 만큼 유의하게 감소한다.

(ii) 로그프레일티(즉 변량효과)의 분산 추정치는 $\hat{\alpha} = 0.5345$(SE $= 0.3384$)이지만, 이에 대한 유의성 검정(즉 $H_0 : \alpha = \text{var}(v) = 0$)은 0에서의 변량효과 분산에 대한 경계공간(boundary space) 문제로 인해, 우도에 기초한 혼합 카이제곱 검정법(mixture chi-squares test)을 사용한다 (Ha 등, 2011, 2012). 이러한 검정을 위해서는 로그프레일티의 분산이 0일 때의 콕스모형을 이 자료에 대해 적합함으로써 얻어지는 우도들의 값이 필요하다. 따라서 두 모형(공통프레일티 모형과 콕스모형)의 데비언스(deviance) $-2p_{b,v}(hp)$의 차이는 $377.4404 - 371.5404 = 5.9 > 2.71$이므로 유의수준 5%에서 귀무가설이 기각되므로 프레일티 효과는 유의함을 알 수 있다. 여기서 377.4404는 콕스모형에서의 데비언스이다.

(iii) Figure 3.1에 있는 세 가지 AIC들 모두 콕스모형의 AIC들(not shown) 보다 작으므로 공통 프레일티 모형이 선택됨을 알 수 있다.

3.2. 지분 프레일티 모형

병원급 센터에서 환자를 모집하는 다기관 임상시험(multi-center clinical trials)에서 각 환자별 재발사건들이 관측되는 지분(nested)디자인 구조인 경우, 두 개의 프레일티 항을 고려할 수 있다. 이에 대한 로그정규 지분(nested 또는 multi-level) 프레일티 모형 (Yau, 2001; Ha 등, 2007)은 다음과 같이 정의된다.

$$\lambda(t|v_1, v_2; x) = \lambda_0(t) \exp\left(x^T\beta + v_1 + v_2\right), \tag{3.2}$$

여기서 $v_1 \sim N(0, \alpha_1)$로서 센터의 로그프레일티이고, $v_2 \sim N(0, \alpha_2)$로서 센터에서 지분된 환자의 로그프레일티이며, v_1과 v_2는 서로 독립이다. 이 모형에 대한 사용방법은 공통 프레일티 모형과 유사하며, 하나의 큰 차이점은 기본사항의 모형식 함수는 `Surv(time, status)~x` $+(1|\text{center}) + (1|\text{id})$으로 지정된다. 여기서 "center"는 각 센터의 식별자이고 "id"는 각 환자의 식별자를 나타낸다. 지분 프레일티모형 (3.2)는 공통 프레일티 모형 (3.1)에 변량효과가 하나 더 추가되는 형태이기 때문에, 이에 대한 다단계우도의 추론은 쉽게 확장이 가능하다 (Ha 등, 2007).

예제 3.2: 지분 프레일티 모형의 예증을 위해 Fleming와 Harrington (1991)에 의해 제시된 만성육아종병(chronic granulomatous disease; CGD)에 대한 재발감염시간 자료를 고려한다. 이 자료는 다기관 임상시험을 통해 13개의 센터에서 모집된 총 128명의 CGD환자에 대해 재발 감염시간을 측정한 것으로서, CGD환자에 대해 심각한 감염율을 줄이는데 있어서 감마인트페론(gamma interferon; rIFN-g)이 효과가 있는지를 조사하는 것이 목적이다. 여기서 생존시간은 재발감염시간(TIME)이며, 연구종료에 의해 약 63%의 중도절단(DEL)이 있다. 자료의 구조를 먼저 살펴보면 동일한 환자의 재발감염시간들은 서로 의존될 수 있다. 하지만 각 환자는 13개의 센터중 하나에 소속되어 있기 때문에 자료의 상관성은 센터효과(center effect)에서도 비롯될 수 있다. 따라서 이 자료의 모형화를 위해 두 개의 프레일티

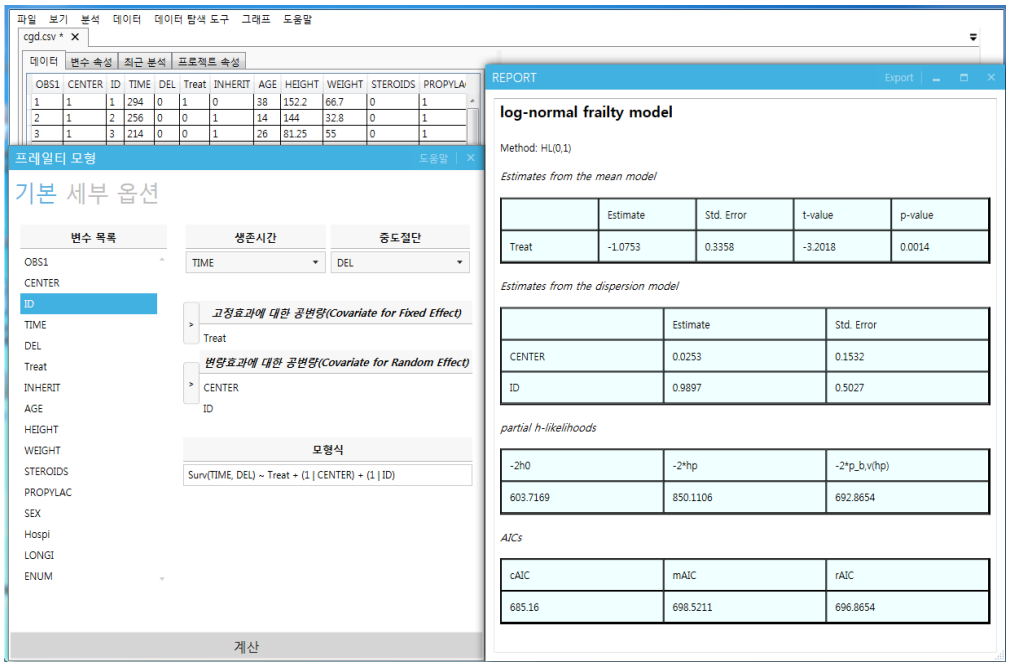

Figure 3.2. Results of fitting nested frailty models for the CGD data

Table 3.1. Model selection among four models for CGD data

Model	cAIC	mAIC	rAIC
M1: 콕스	708.6	708.6	707.4
M2: 센터 프레일티효과	702.9	706.8	705.6
M3: 환자 프레일티효과	685.1	696.6	694.9
M4: 센터 및 환자 프레일티효과	685.2	698.5	696.9

요인, 즉 랜덤센터효과(CENTER)와 센터에서 지분된 랜덤환자효과(ID), 그리고 하나의 공변량(Treat; 감마인트페론은 '1', placebo는 '0'으로 코딩)을 고려 할 수 있다. 대응하는 지분 프레일티 모형을 적합한 결과는 Figure 3.2의 우측에 표로 요약되어 있으며, 그 해석은 다음과 같다:

(i) Treat의 효과는 매우 유의하다(p-값 $= 0.0014$). 따라서 감마인트페론은 플래시보에 비해 감염 위험률이 $\exp(-1.0753) = 0.341$배 만큼 유의하게 감소한다. 다시 말하면 감마인트페론은 CGD환자에 대해 심각한 감염율을 유의하게 감소시킨다고 할 수 있다.

(ii) 환자프레일티의 분산 추정치 $\hat{\alpha}_1 = 0.989\,(\text{SE} = 0.5027)$이 센터프레일티의 분산 추정치 $\hat{\alpha}_2 = 0.0253\,(\text{SE} = 0.1532)$보다 매우 크기 때문에 랜덤환자효과는 보다 이질적임을 알 수 있다.

(iii) 이 자료에 대한 적절한 모형선택을 위해 3개의 축소모형을 포함한 총 4개의 모형을 고려할 수 있다. 따라서 M1(콕스모형), M2(랜덤센터효과만을 고려한 모형), M3(랜덤환자효과만을 고려한 모형), M4(랜덤센터 및 환자효과를 고려한 모형)중에서 적절한 모형을 선택한 결과, Table 3.1에서 제시하는 바와 같이 세 가지 기준 모두 가장 작은 AIC를 주는 모형은 M3이므로 CGD자료의 경우 M3모형을 적절한 모형으로 선택할 수 있다.

4. 토론 및 추후과제

본 논문에서는 SRC-Stat 통계패키지를 이용하여 단변량 및 다변량 생존자료 분석을 수행하는 방법을 소개하였다. 특히 본 패키지는 메뉴 형식이기 때문에 생존분석을 위해 통계 전문가뿐만 아니라 비 통계인도 접근하기가 쉬운 장점이 있다. 2절에서 다룬 카플란-마이어 방법과 콕스모형 추론법은 생존분석에서 가장 기본적이면서 표준적인 분석법이다. 콕스모형은 기저 위험함수에 대해 어떠한 형태의 가정 없이도 관심모수에 대한 추론 및 간편한 해석이 가능하기 때문에 생존자료의 회귀분석 시 가장 많이 사용된다. 하지만 콕스모형은 PH라는 강한 가정을 전제하고 있기 때문에 추론 시 주의가 필요하다 (Lawless, 2003). 3절에서는 최근에 많은 연구가 진행되고 있는 다변량 생존자료 분석에 매우 유용한 프레일티 모형 사용법을 소개하였다. 사실 프레일티 모형은 단변량 생존자료 분석에도 사용이 가능하며 이 경우는 개인간 이질성을 모형화 해 준다 (Hougaard, 2000). 또한 프레일티 모형은 비비례위험모형(non-PH)이 되기 때문에 콕스의 PH모형에 대한 하나의 대안이 될 수도 있다 (Ha와 MacKenzie, 2010).

프레일티 모형은 보다 다양한 생존분석 분야로 적용 및 확장이 가능하기 때문에 우리는 다음 네 가지 사항들을 본 패키지에 추가하는 것을 현재 연구하고 있다:

(i) 상관된 프레일티를 갖는 모형(Frailty models with correlated frailties; Ha 등, 2011).

(ii) 프레일티의 구간 추정(Interval estimation for frailties; Ha 등, 2013).

(iii) 프레일티 모형에서 변수선택(Variable selection in frailty models; Ha 등, 2014).

(iv) 경쟁위험 프레일티 모형(Competing risks frailty models; Ha 등, 2015).

위와 같은 확장된 프레일티 모형의 패키지를 개발함으로써 기존의 통계패키지(SAS, R, SPSS, Stata 등)와의 차별성으로 인해 효율적이면서도 경쟁력 있는 고급 생존분석의 극대화를 도출할 수 있을 것으로 사료된다.

References

Aalen, O. O. (1975). *Statistical Inference for a Family of Counting Process*, Ph.D. dissertation, University of California, Berkeley.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study, *The Annals of Statistics*, **10**, 1100–1120.

Andersen, P. K., Klein, J. P., Knudsen, K. and Palacios, R. T. (1997). Estimation of variance in Cox's regression model with shared gamma frailties, *Biometrics*, **53**, 1475–1484.

Breslow, N. E. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship, *The Annals of Statistics*, **2**, 237-453.

Cox, D. R. (1972). Regression models and life tables(with discussion), *Journal of the Royal Statistical Society B*, **34**, 187–220.

Duchateau, L. and Janssen, P. (2008). *The Frailty Models*, Springer, New York.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*, Wiley, New York.

Fleming, T. R. and Lin, D. Y. (2000). Survival analysis in clinical trials: Past developments and future directions, *Biometrics*, **56**, 971–983.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily single-censored samples, *Biometrika*, **52**, 203–223.

Ha, I. D., Christian, N. J., Jeong, J.-H., Park, J. and Lee, Y. (2015). Analysis of clustered competing risks data using subdistribution hazard models with multivariate frailties, *Statistical Methods in Medical Research*, In press.

Ha, I. D. and Lee, Y. (2003). Estimating frailty models via Poisson hierarchical generalized linear models, *Journal of Computational and Graphical Statistics*, **12**, 663–681.

Ha, I. D. and Lee, Y. (2005). Comparison of hierarchical likelihood versus orthodox best linear unbiased predictor approaches for frailty models, *Biometrika*, **92**, 717–723.

Ha, I. D., Lee, Y. and MacKenzie, G. (2007). Model selection for multi-component frailty models, *Statistics in Medicine*, **22**, 4790–4807.

Ha, I. D., Lee, Y. and Song, J.-K. (2001). Hierarchical likelihood approach for frailty models, *Biometrika*, **88**, 233–243.

Ha, I. D. and MacKenzie, G. (2010). Robust frailty modelling using non-proportional hazards models, *Statistical Modelling*, **10**, 315–332.

Ha, I. D., Noh, M. and Lee, Y. (2012). frailtyHL: A package for fitting frailty models with h-likelihood, *The R Journal*, **4**, 307–320

Ha, I. D., Pan, J., Oh, S. and Lee, Y. (2014). Variable selection in general frailty models using penalized h-likelihood, *Journal of Computational and Graphical Statistics*, **23**, 1044–1060.

Ha, I. D., Sylvester, R., Legrand, C. and MacKenzie, G. (2011). Frailty modelling for survival data from multi-centre clinical trials, *Statistics in Medicine*, **30**, 28–37.

Ha, I. D., Vaida, F. and Lee, Y. (2013). Interval estimation of random effects in proportional hazards models with frailties, *Statistical Methods in Medical Research*, Published online: 29/January/2013.

Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data, *Biometrika*, **69**, 553–566.

Hougaard, P. (1999). Fundamentals of survival data, *Biometrics*, **55**, 13–22.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, Springer, New York.

Johansen, S. (1983). An extension of Cox's regression model, *International Statistical Review*, **51**, 165–174.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimator from incomplete observations, *Journal of the American Statistical Association*, **53**, 457–481.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*, 2nd edn, Wiley, New York.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion), *Journal of the Royal Statistical Society B*, **58**, 619–678.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via h-Likelihood*, Chapman and Hall, London.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of National Cancer Institute*, **22**, 719–748.

McGilchrist, C. A. and Aisbett, C. W. (1991). Regression with frailty in survival analysis, *Biometrics*, **47**, 461–466

Nielsen, G. G., Gill, R. D., Andersen, P. K. and Rensen, S. (1992). A counting process approach to maximum likelihood estimation in frailty models, *Scandinavian Journal of Statistics*, **19**, 25–44.

Tarone, R. E. and Ware, J. (1977). On distribution-free test for equality of survival distributions, *Biometrika*, **64**, 156–160.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*, Springer, New York.

Vaida, F. and Xu, R. (2000). Proportional hazards models with random effects, *Statistics in Medicine*, **19**, 3309–3324.

Vaupel, J. W., Manton, K. G. and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography*, **16**, 439–454.

Yau, K. K. W. (2001). Multilevel models for survival analysis with random effects, *Biometrics*, **57**, 96–102.

SRC-Stat 통계패키지를 이용한 생존분석

하일도[a,1] · 노맹석[a] · 이영조[b] · 임요한[b] · 이재용[b] · 오희석[b] · 신동완[b] · 이상구[b] · 서진욱[b] · 박용태[b] · 조성준[b] · 박종헌[b] · 김유경[b] · 유경상[b]

[a]부경대학교 통계학과, [b]서울대학교 데이터과학과 지식창출 연구센터

(2015년 3월 23일 접수, 2015년 3월 31일 수정, 2015년 3월 31일 채택)

요 약

본 논문에서는 SRC-Stat 통계패키지를 이용하여 생존자료를 분석하는 방법을 소개한다. 본 패키지는 단변량 생존자료 분석을 위한 콕스의 비례위험모형 뿐만아니라, 다변량 생존자료분석을 위한 공통 및 지분 프레일티 모형과 같은 고급 생존분석법을 제공한다. 잘 알려져 있는 실제자료의 사용을 통해 본 패키지의 유용성을 예증한다.

주요용어: 콕스비례위험모형, 프레일티 모형, 다단계우도, 다변량생존자료, 변량효과.

이 논문은 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. 2011-0030811).

[1]교신저자: (608-737) 부산광역시 남구 용소로 45, 부경대학교 통계학과. E-mail: idha1353@pknu.ac.kr