Journal of Statistical
Theory & Methods
1996, Vol. 7, pp 145~ 159

비례위험모형분석을 위한 한글멀콕스(HMULCOX)[1]

이상복[2] · 박의준[3]

요약 다변량 발병시간자료는 각 개개 환자에게 있어 합병증이 발생되거나 혹은 유사 환자군(집락) 내의 발병시간이 상관되어진 생의학자료에서 흔히 볼 수 있다. HMULCOX는 그런 자료를 분석하기 위한 한글 통계 패키지 가운데 하나이다. 이 프로그램은 관련된 발병시간들이 독립이 아닐때에도 COX 비례 위험 모형의 주변확률분포를 계산해 준다. 주어진 조건으로는 주변확률모형의 기본위험율은 일정한 상수, 혹은 변수라도 관계없다. 또한 치료실패율의 치료변수들(공변량)의 효과에 대해 다양한 통계적 추론이 가능하다. 기본적으로 주변확률분포접근법으로 설계되었지만 HMULCOX는 여러 가지 추론 방법을 선택하는 데 일반적으로 충분하다. 이 프로그램으로 2개의 예를 들어 실행하겠다.

주제어: 비례위험모형, 시간종속성, 다변량발병시간, 주변확률분포, 한글통계패키지

1. 소개

최근 한국형 통계패키지의 설계 개발은 국내에서 활발히 일어나고 있다(이정진외 1994, 1995.) 시간에 따라 변하는 공변량을 가진 비례위험모형에서의 모수β의 추정은 일반적으로 매우 중요하다.

1990년에 MULCOX[5]이라 불리는 컴퓨터 프로그램이 발표 되었다. 각 연구대상이 합병증 발병시간의 공변량 영향에 대한 COX 회귀분석이라는 방법을 Wei, Lin 그리고 Weissfeld[9]가 이론적으로 밝혔다. 생의학 연구에서는 합병증 증상이 흔히 발생함으로 이러한 종류의 프로그램은 매우 필요했다. 다변량 발병시간 자료의 예로써 신체 각 조직의 병이 진전되는 시간, 혹은 각 개인 환자들에 종양의 재발, 감염 경로 혹은 천식환자들의 감염 시간을 들 수 있다.

구체적으로 설명하면 첫번째 현상으로, 시간이 지나면서 변하는 공변량을 사용할 수 있으면 시간에 따라 변하는 공변량의 위험 요인을 연구할 수 있고 또한 비례 위험율을 조절하는 간단한 방법을 얻을 수 있다. 시간 종속적 공변량은 특히 다변량수명시간모형이 적합하다. 그 이유로 감염된 경험이 2차 감염의 위험에 얼마나 영향을 미치는가를 평가하는 문제에 있어서 다변량수명시간분포를 적용시킬 수 있다. 두번째 현상으로는 발병시간이라는 집락 자

[1] 본 연구는 1996년도 효성가톨릭대학교 학술연구비 지원결과의 일부임
[2] 대구효성가톨릭대학교 통계학과
[3] 김천전문대 전산정보처리과

료는 같은 집락내의 종속된 발병시간들에 따라 대상들을 자연적으로 혹은 인공적인 집락으로 묶을 때 나타 난다. 그런 데이터의 예는 왼쪽 눈의 실명과 오른쪽 눈의 실명시간, 대응된 집락 실험에서 종양의 발생시간, 가족 구성원의 유전병이 나타나는 나이를 집락으로 구성하는 것이다.

따라서 다변량발병시간 자료를 분석하기 위한 도구로 1993년에 MULCOX2[6]가 개발되었다. 이 프로그램은 다양한 형태의 시간 종속 공변량을 사용할 수 있다. 이 프로그램은 합병증 자료 뿐만 아니라 집락된 발병 시간 자료를 다룬다. 더우기 한 대상의 임의 시점에서의 발병 위험도 분석한다. 또한 연구치료집단에 늦게 합류하거나, 재발된 환자 자료에 대한 치료 방법 등의 시간에 따라 달라지는 상황에서도 분석 가능하다.

HMULCOX의 이론적 배경은 참고문헌[7]에 자세히 서술되어 있다. 주된 방법은 완전히 명시되지 않은 관련된 발병시간의 종속성을 가지지 않는 비례위험모형을 가진 다변량 발병시간의 주변분포를 계산하는 Lee, Wei와 Amato[4]와 Wei, Lin과 Weissfeld[9]에 의한 주변확률분포접근방법이다. 잘 알고 있는 부분우도비는 종속성을 계산하기 위해 적절히 바꿀 수 있다. HMULCOX는 재발된 병을 분석하기 위해 Andersen과 Gill[3]과 Prentice, William과 Peterson[8]의 것들을 포함해서 여러 가지 택일적으로 사용할 수 있는 접근법이다.

다음절에서 HMULCOX는 사용자들에게 쉽게 접근하여 사용될 수 있도록 통계적 방법과 계산법을 설명했다. 컴퓨터 프로그램 그 자체는 3절에서 기술할 것이고, 프로그램의 주된 구조를 설명하기 위해 4절에서 두 가지 생의학의 예를 들어 설명하였다.

2. 한글멀콕스 통계적 계산 방법

본 장에서는 HMULCOX이해를 돕기위해 본 프로그램의 계산 알고리즘을 간략히 소개한다. n명의 환자가 있고 각 환자는 질병의 형태가 K개 혹은 집락이 K개를 가질 수 있다고 가정하자. 각 환자는 합병증 자료를 대상으로도 하고 집락된 자료를 대상으로 하기도 한다. 일반적으로 전자는 질병의 형태가 명백한 구별이 있는 반면에 후자에서는 질병의 형태가 임의적이다. 만약 집락들(질병 요인)의 수가 일정하기 않다면 가장 큰 집락으로 K를 놓는다.

X_{ik} 는 i번째 사람의 k번째의 질병을 관측시간으로 놓는다.그리고 Δ_{ik} 는 0 혹은 1의 값을 갖는 지수로서 $\Delta_{ik}=0$ 이면 X_{ik} 는 사망 혹은 임의중단관측시간이다. 또한 $Z_{ik}(t)=\{Z_{1ik}(t),\cdots,Z_{pik}(t)\}$ 은 k형의 질병을 갖는 i번째 사람에 관한 시간종속공변량(치료변수)인 $P\times 1$ 벡터이다. X_{ik} 혹은 Z_{ik} 가 결측치이면 $X_{ik}=0$ 그리고 $\Delta_{ik}=0$로 놓는다. 자연적으로 그런 경우는 통계량을 계산할 수 없을 것이다.

비례위험모형을 갖는 질병의 각 형에 대한 주변분포를 계산할 수 있다. K형태의 질병 가운데 기본위험함수가 동일하든가 다른지에 따라서 K형태 질병에 대한 i번째 사람에 대한 위험함수는 아래와 같다.

$$\lambda_k(t,Z_{ik})=\lambda_0(t)e^{\beta' Z_{ik}(t)} \tag{2.1}$$

혹은

$$\lambda_k(t,Z_{ik}) = \lambda_{0k}(t)e^{\beta' Z_{ik}(t)} \tag{2.2}$$

$\lambda_0(t)$ 와 $\lambda_{0k}(t)\,(k=1,\cdots,K)$ 는 특성화되지 않은 기본위험 함수이고 $\beta=(\beta_1,\cdots,\beta_p)'$ 은 미지의 회귀모수로 $P\times 1$ 벡터이다. 합병증 자료의 경우에 다른 질병의 형태 $\lambda_{ok}(t)$ $(k=1,\cdots,K)$ 가 필요하고 반면에 집락된 자료를 위해서는 식(2·1)을 일반적인 기본위험함수로 사용 한다. 식(2·1)과 식(2·2)에서 주변확률모형에서 β 는 똑같다. 이 구조는 항상 적절한 공변량에 대해서 적용 할 수 있다.

$Y_{ik}(t)$는 지수로 시점 t에서 k형태 질병에 대한 i번째 사람이 위험에 놓여 있으면 1 (치료 가능), 위험이 없으면 0(치료 불능)의 값을 갖는다. 대부분 응용에서 $Y_{ik}(t)$는 시간구간 $(0,X_{ik}]$에서 1의 값을 갖고 (치료받는중), X_{ik}후에는 0의 값을 갖는다(치료 안받음). β의 추정량을 구하기 위해 아래와 같은 표현식을 사용하기로 하자.

$$S_k^{(0)}(\beta,t) = \sum_{i=1}^{n} Y_{ik}(t)e^{\beta' Z_{ik}(t)},$$

$$\overline{S}^{(0)}(\beta,t) = \sum_{i=1}^{K} S_k{}^{(0)}(\beta,t),$$

$$S_k^{(1)}(\beta,t) = \sum_{i=1}^{n} Y_{ik}(t)e^{\beta' Z_{ik}(t)} Z_{ik}(t)$$

$$\overline{S}^{(1)}(\beta,t) = \sum_{i=1}^{K} S_k{}^{(1)}(\beta,t),$$

$$S_k^{(2)}(\beta,t) = \sum_{i=1}^{n} Y_{ik}(t)e^{\beta' Z_{ik}(t)} Z_{ik}(t)Z_{ik}(t)',$$

$$\overline{S}^{(2)}(\beta,t) = \sum_{i=1}^{K} S_k{}^{(2)}(\beta,t),$$

추정치 $\hat{\beta}$는 다음식으로부터 유도된다.

$$U(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K} \Delta_{ik}\left\{ Z_{ik}(X_{ik}) - \frac{\overline{S}^{(1)}(\beta,X_{ik})}{\overline{S}^{(0)}(\beta,X_{ik})}\right\}$$

가 식(2.1)의 정규방정식이고

$$U(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K} \Delta_{ik}\left\{ Z_{ik}(X_{ik}) - \frac{\overline{S}_k^{(1)}(\beta,X_{ik})}{S_k^{(0)}(\beta,X_{ik})}\right\}$$

가 식(2.2)의 정규방정식이다. β에 관한 정규방정식 U(β)의 도함수 행렬은

$$A(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K} \Delta_{ik}\left\{ \frac{\overline{S}^{(2)}(\beta,X_{ik})}{\overline{S}^{(0)}(\beta,X_{ik})} - \frac{\overline{S}^{(1)}(\beta,X_{ik})\overline{S}^{(1)}(\beta,X_{ik})'}{\overline{S}^{(0)}(\beta,X_{ik})^2}\right\}$$

$$A(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{K} \Delta_{ik}\left\{ \frac{\overline{S}_k^{(2)}(\beta,X_{ik})}{S_k^{(0)}(\beta,X_{ik})} - \frac{\overline{S}_k^{(1)}(\beta,X_{ik})\overline{S}_k^{(1)}(\beta,X_{ik})'}{S_k^{(0)}(\beta,X_{ik})^2}\right\}$$

각각 (2.1),(2.2)에 해당된다. 따라서 $A(\beta)$ 양정치 행렬이기 때문에, 뉴턴 랩슨 알고리즘으로 $\{U(\beta)=0\}$를 만족하는 $\hat{\beta}$이 구해진다.

통계량 $U(\beta)$는 근사적인 p변량정규분포로서 평균은 0, 분산공분산 행렬은

$$B(\hat{\beta})=\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K}W_{ik}(\hat{\beta})W_{il}(\hat{\beta})'$$

이다. 이때

$$W_{ik}(\beta)=\Delta_{ik}\left\{Z_{ik}(X_{ik})-\frac{\bar{S}^{(1)}(\beta,X_{ik})}{\bar{S}^{(0)}(\beta,X_{ik})}\right\}-\sum_{j=1}^{n}\sum_{i=1}^{K}\frac{\Delta_{jl}Y_{ik}(X_{jl})e^{\beta'Z_{ik}(X_{jl})}}{\bar{S}^{(0)}(\beta,X_{jl})}\left\{Z_{ik}(X_{jl})-\frac{\bar{S}^{(1)}(\beta,X_{jl})}{\bar{S}^{(0)}(\beta,X_{jl})}\right\}$$

이고

$$W_{ik}(\beta)=\Delta_{ik}\left\{Z_{ik}(X_{ik})-\frac{S_k^{(1)}(\beta,X_{ik})}{S_k^{(0)}(\beta,X_{ik})}\right\}-\sum_{j=1}^{n}\frac{\Delta_{jk}Y_{ik}(X_{jk})e^{\beta'Z_{ik}(X_{jk})}}{S_k^{(0)}(\beta,X_{ik})}\left\{Z_{ik}(X_{jl})-\frac{S^{(1)}(\beta,X_{jl})}{S^{(0)}(\beta,X_{jl})}\right\}$$

는 각각 식(2·1)과 식(2·2)이다. 더우기 추정치 $\hat{\beta}$은 근사적인 P변량정규분포로서 평균은 β이고 공분산행렬 $D(\hat{\beta})=A^{-1}(\hat{\beta})B(\hat{\beta})A^{-1}(\hat{\beta})$을 가진다. $\hat{\beta}$를 나이브와 로버스트 공분산 추정치로서 각각 $A^{-1}(\hat{\beta})$과 $D(\hat{\beta})$라 하고 $U(0)A^{-1}(0)U(0)$와 $U(0)B^{-1}(0)U(0)$를 각각 나이브와 로버스트 로그 순위 통계량이라 부른다. 나이브 통계량은 같은 사람에서 발병 시간이 독립이지 않다면 무효이다.

로버스트 로그 순위 통계량 $U(0)B^{-1}(0)U(0)$는 전체 귀무가설 $H_0:\beta=0$를 검정할 수 있다. 각각 공변량 효과들에 관한 추론은 표준정규분포에서 표준화된 모수추정을 구하는데 쓰여진다. β의 여러 가지 구성에 포함된 일반적인 다변량 선형가설은 L이 정수 행렬 $r\times p$인 $H_0:L\beta=0$ 로서 표현된다. H_0 검정을 위한 로버스트 왈드 통계량은 자유도가 r이고 근사적인 χ^2분포를 가지는$(L\hat{\beta})'\{LD(\hat{\beta})L'\}^{-1}(L\hat{\beta})$이다.

일찍이 언급되었듯이 i번째 사람이 질병유형 중의 k번째형에 불완전한 정보를 가진다면, 수식에서 $X_{ik}=\Delta_{ik}=0$ (Z_{ik}가 어떤 임의의 값을 가지더라도) 놓는다. 그러나 데이터 파일에서 이 사람이 처음 질병 $n_i\{\le K\}$에 완전한 정보를 가지고 마지막 질병 $(K-n_i)$ 형에 결측치를 가진다면 i번째 사람에 대해 그러한 쓸데없는 레코오드를 만들 필요가 없다. 또한 데이터 파일로부터 마지막 질병$(K-n_i)$형을 간단히 포함해도 좋다. 그러므로 집락된 자료가 일반적으로 참이고 질병 순서가 임의순일때는 완전하게 입력해야한다.

시간종속 공변량과 위험율이 임의의 형태이기 위해서는 k번째 질병유형에 대해 i번째 사람을 위한 정보를 다음과 같이 나타낸다.

$$\{Y_{ik}(t),X_{ik},\Delta_{ik},Z_{1ik}(t),\cdots,Z_{pik}(t)\}\quad(t>0).$$

데이터 파일에서 연속적인 레코오드는

$$(S_{ik,1},t_{ik,1},\Delta_{ik,1},Z_{1ik,1},\cdots,Z_{pik,1}),\cdots,(S_{ik,n_{ik}},t_{ik,n_{ik}},\Delta_{ik,n_{ik}},Z_{1ik,n_{ik}},\cdots,Z_{pik,n_{ik}})$$

이다.

$l=1,\cdots,n_{ik}$ 동안 i번째 사람이 시간구간 $(S_{ik,l},t_{ik,l}]$에서 질병의 k번째형에 위험이 있고

$(Z_{1ik,l},\cdots,Z_{pik,l})$는 그 구간을 넘은(치료 불가능한) 공변량 벡터이다. $\Delta_{ik,n_{ik}}=\Delta_{ik}$ 이고 $l=1,\cdots,n_{ik}-1$(만약 $n_{ik}>1$이면)에서 어떤 임의의 값 즉, $\Delta_{ik,l}$을 0으로 놓는다. 그 구간 $(S_{ik,l},t_{ik,l}](l=1,\cdots,n_{ik})$는 그 공변량 벡터 $Z_{ik}(t)$는 상수이고 각각에 더 좋은 구간을 나눈다면 $Y_{ik}(t)=1$인 발병시간 모두를 먼저 명시되게 구성한다. $Z_{ik}(t)$가 t(계산 목적상) 시점에서 연속이고 $Z_{ik}(t)$는 두번째 발병 시점으로 평가된 $Z_{ik}(t)$가 되는 구간을 넘는 공변량을 벡터로서 발병된 두 인접한 구별 사이에 상수 값으로 충분하다. 대부분 응용에서 공변량은 섞여져 있고 치료대상들은 치료결과가 나타나거나 혹은 임의 중단될 때까지 t=0에서 위험에 있다. 그러면 단지 한 레코오드 $(0,x_{ik},\Delta_{ik},Z_{1ik},\cdots,Z_{pik})$는 k번째 질병 유형에 i번째 사람을 위해 필요하다. 가능한 많은 레코오드들 때문에 자료 파일에서 각 레코오드에 질병 유형과 사람을 표시하는 것이 필요하다.

3. 컴퓨터 프로그램

3.1 HMULCOX 개요

시작 프로그램은 주메뉴로서 *C*언어로 작성되었으며 그림1과 같고 도움말 자료의 입력및 수정, 실행, 출력, 기타로 이루어졌으며, 입력및 수정은 그림2와 같고 출력은 그림3과 같다. HMULCOX 컴퓨터 프로그램은 배정도 계산 표현으로 표준 FORTRAN-77로 쓰여졌다. 외부 부프로그램 혹은 함수들은 전혀 사용되지 않았다. 그 프로그램은 FORTRAN 컴파일러를 가진 어떤 컴퓨터에서도 실행할 수 있다. HMULCOX에 사용되는 *cpu*시간은 데이터의 크기와 컴퓨터 설치에 달려있다.

이 프로그램은 n, p 와 K의 임의의 값을 가질 수 있다. i와 k 즉 n_{ik}의 각 조합을 위한 시간 구간의 수는 또한 임의의 값을 갖는다. M을 n_{ik}의 최대값으로 놓고 자료와 계산될 결과는 1차원 배열로 저장되어 진다. 즉 A(이 배열은 n,p,K와 M의 값에 따라 나누어져 있다.)이다. A의 차원은 필요하다면 사용자에 의해 변경되어 질 수 있다.

도움말	입력및수정	실행	출력	기타
			화면출력 프린터 출력	도스로 끝

그림1 주메뉴

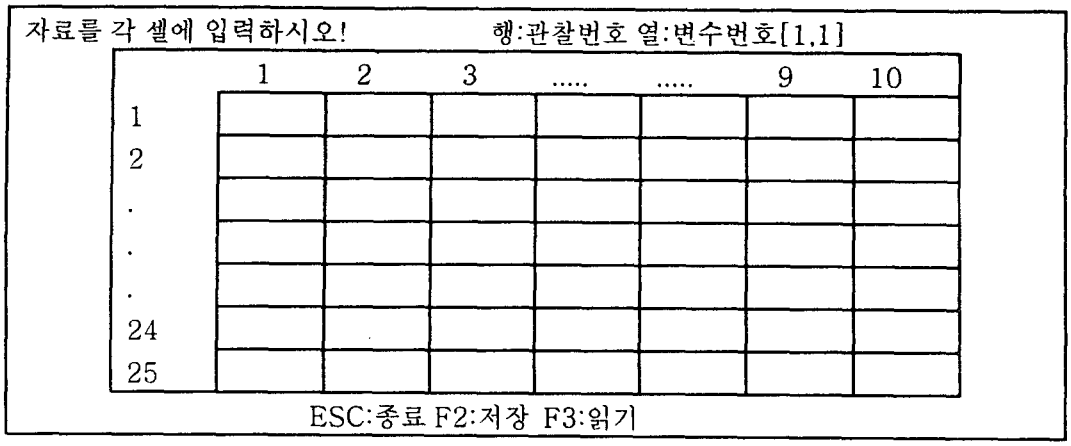

그림2 입력 및 수정

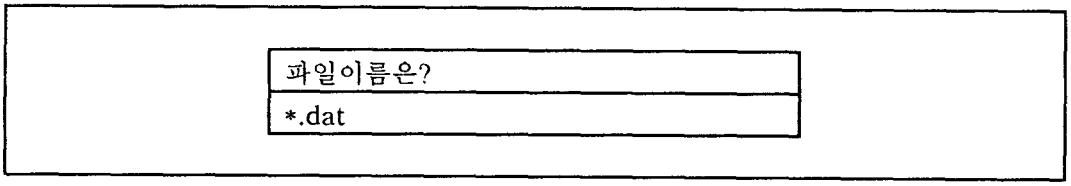

그림3 화면출력, 프린터출력

3.2 입력

HMULCOX는 입력을 자료입력과 제어모수 입력 두개의 분리된 그룹으로 입력한다. 데이터 파일에서 레코드는 아래 순서에 보여지는 변수들에 따라 값을 준다.

$$(ID1_i, ID2_i, S_{ik,l}, t_{ik,l}, \Delta_{ik,l}, Z_{1ik,l}, \cdots, Z_{(p)ik,l})$$

ID1은 사람이고 ID2는 질병 유형인 $(i=1,\cdots,n; k=1,\cdots,n_i; l=1,\cdots,n_{ik})$이다(마지막 $(K-n_i)$질병 유형이 l번째 사람에 대해 결측치인 상황을 적응시키기에 K보다 더 적게 되는 n_i의 가능성을 줄수 있다는 데 주의하라.). 같은 사람의 레코오드는 $ID2_k$의 비감소순으로 연속적으로 되어져야 한다.

제어모수는 HMULCOX의 실행으로 간단하게 키보드로부터 입력할 수 있다. 원시코드는 배치파일로부터 제어변수 입력에서 쉽게 변경할 수 있다.

데이터 파일의 형태가 명기되어 질 때, 제공된 자료 항목이 공백 혹은 콤마로 구별되어 진다면 사용자는 FREE 혹은 free로 입력하면 된다.

3.3 출력

계산된 결과는 사용자에 의해 입력된 출력 파일명으로 출력된다. 그 출력은 2절에서 묘사된 추정치와 검정통계량을 포함 한다.

4. 응용

이 절에서 두 생의학 연구에서부터 취해진 자료로 HMULCOX의 사용을 보여 준다. 그 연

구와 분석은 Lin[7]에 상세하게 묘사되어 졌다. 여기 HMULCOX로 부터 처리 된 결과를 본다.

4.1 정신분열증후군(The Schizophrenia Study) 연구의 예

John Hopkins 대학의 Ann E. Pulver에 의해 시행된 유전에 관한 역학적 연구에 있어서 관련된 487명 중 93명이 여자 정신분열증 환자군으로 등록되었다. 가족 중에서 연관된 변수는 1에서 12의 범위를 가진다. 그러므로 n=93 그리고 K=12 이다. 주된 의문은 관련된 변수 중 질병에 영향을 주는 환자군에 있어서 정신분열증이 처음 발생시의 나이와 관련이 있는가에 대한 것이다. 성별에 관련된 변수도 예측되어 진다는 것이다. 식(2.1)에서

$$Z_{ik} = (Z_{1ik}, Z_{2ik})' \quad (i = 1,\cdots,93; k = 1,\cdots,12)$$

이고

$$Z_{1ik} = \begin{cases} 1 & \text{처음 발생시의 i번째 환자군 나이가} \le 16\text{이면} \\ 0 & \text{그렇지 않으면} \end{cases}$$

$$Z_{2ik} = \begin{cases} 1 & \text{만약 i번째 환자군의 k번째 관련이 남자면} \\ 0 & \text{만약 i번째 환자군의 k번째 관련이 여자면} \end{cases}$$

이다.

데이터 파일은 그림4에 보여진다. 가족의 번호는 같게 하고 가족의 번호, 가족의 크기, 임상 연구 시작시간(0), 관련성의 관측시간, 관련을 위한 발병 지수(병은 1, 임의 절단은 0), 환자군에 있어 발생시의 나이를 위한 지수, 그리고 관련의 성별을 위한 지수로 구성된 487개 레코오드가 있는데 여기서는 PC 용량에 문제가 있어 6명에 대해서만 실행 하여 보았다. 제어모수의 입력과 컴퓨터 실행으로부터 출력은 각각 그림5, 그림6에 보여 준다.

4.2 만성종양(Chronic Granulomatous Disease, CGD) 연구의 예

플라스보로 통제된 실험이 CGD 환자의 감염율을 감소하기 위해 감마 인터페론의 효과를 연구하기 위해 실행되어 졌다. 실험 결과에 의하면 위약 투여 환자 65명 중 30명과 감마 인터페론 투여 환자 63명 중 14명이 하나 혹은 그 이상의 감염이 되어졌다. 이 연구로부터 자료는 그림4에 나타나져 있다. 5개의 열은 환자의 고유번호, 감염정도를 정수로 표현, 관측시간, 감염 지수(1은 감염, 0은 치료), 치료 지수(1은 감마인터페론, 0은 플라스보)이다. 각 레코오드는 감염수를 위한 기록이다. 즉, 마지막 잘 알 수 없는 연구에 대해 각 환자를 위한 마지막 기록은 환자에 의해 방문된 것이다.

```
1  1 0  47  0  0  1   5  2 0  55  0  0  0
1  2 0  42  0  0  0   5  3 0  28  0  0  1
1  3 0  21  0  0  1   5  4 0  26  0  0  1
1  4 0  20  0  0  1   5  5 0  22  0  0  1
1  5 0  19  0  0  1   5  6 0  20  0  0  1
2  1 0  29  1  0  1   5  7 0  29  0  0  0
2  2 0   6  0  0  1   5  8 0  24  0  0  0
2  3 0  47  1  0  0   6  1 0  69  0  0  1
2  4 0  33  0  0  0   6  2 0  18  0  0  1
3  1 0  63  0  0  1   6  3 0  16  0  0  1
3  2 0  57  0  0  0   6  4 0  15  0  0  0
3  3 0  52  0  0  1   6  5 0  52  1  0  0
3  4 0  48  0  0  1
3  5 0  47  0  0  1
3  6 0  45  0  0  1
3  7 0  30  1  0  1
3  8 0  53  0  0  0
3  9 0  50  0  0  0
3 10 0  40  0  0  0
3 11 0  37  0  0  0
4  1 0  55  0  1  1
4  2 0  47  1  1  0
4  3 0  28  0  1  1
4  4 0  27  0  1  0
5  1 0  56  0  0  1
```

그림 4 spn.dat

```
제목을 입력해 주세요
 test
입력된 자료명을 쳐주세요
 spn.dat
출력될 자료명을 쳐주세요
 spn.txt
UNITS(사람수)를 쳐주세요
 6
질병의 최대수를 쳐주세요
 12
최대시간 구간의 수를 쳐주세요
 1
모형의 공변량(치료법)을 쳐주세요
2
1번째 공변량의 이름은
 age
2번째 공변량의 이름은
 gender
자료 파일의 형은
 free
모형1이면 1, 모형 2이면 2
 1
다변량 가설의 수는
 0
잠깐만 기다리세요
```

그림 5 입력

처음에는 3가지 감염에 대한 시도를 해 본다. 그러므로 n=128과 K=3을 가진다. 2절에서 묘사된 주변분포접근을 위해 환자는 전 연구에 관해 3가지 감염에 대해 각각의 위험을 가지는 것으로 고려되어 진다. 특별한 감염시간이 임의 중단되었다면, 똑같은 환자에 관해서 모든 부차적 감염시간은 똑같은 시간에 임의 중단되어진 것(결측치가 아니라)으로서 취급되어 진다. 그러므로 모든 환자에 대해서 3가지 감염에 대해 각각 하나의 레코오드를 가질 것이다. 모든 위험 구간의 왼쪽 끝점은 0이다.

세가지 감염의 치료 효과를 추정하기 위해 i번째 환자가 감마 인터페론을 사용하면 $R_1=1$ 이고 그렇치않으면 $R_1=0\, Z_{i1}=(R_i,0,0)', Z_{i2}=(0,R_i,0)', Z_{i3}=(0,0,R_i)'$ $(i=1,\cdots,128)$으로서 식(2.2)를 사용한다. 이 분석을 위해 원본의 파일(그림 7)로부터 만들어진 그림 8인 자료파일을 만들었다.

읽어보세요!

```
**************************
*                        *
*     다변량 발병자료 시간의      *
*                        *
*       COX 회귀분석          *
*                        *
**************************
제    목 : test
자료 파일명 : spn.dat
출력 파일명 : spn.txt

일반적인 기본 위험 함수의 추정
UNITS(사람수) =  6
질 병 의 수 =  12
중도절단되지 않은 질병의 수
  질병의 형 :  1  2  3  4  5  6  7  8  9 10 11 12

  질병의 수 :  1  1  1  0  1  0  1  0  0  0  0  0

나이브 로그 순위 통계량 =     .69P-값 =  .70566

로버스트 로그 순위 통계량 =    1.03P-값 =  .59744

모수  추정치 나이브 S.E  나이브 N.S.E 로버스트 S.E 로버스트 N.S.E

----  ------ ---------- ----------- --------------------------

age  .73617  1.12692   .65326   .54862    1.34184

gender-.50600  .91967   -.55019   .80847    -.62587

 베타의  나이브 공분산행렬

---------------------------------

 .127E+01 -.838E-01

 -.838E-01  .846E+00

베타의  로버스트 공분산행렬

---------------------------------

 .301E+00 -.145E-01

 -.145E-01  .654E+00
```

up/pgup:앞페이지 down/pgdn:다음페이지 Esc:그만보기

그림 6 출력(spn.txt)

```
4054 1 293 0 1
4077 1 255 0 0
4109 1 213 0 0
4111 1 203 0 0
4001 1 219 1 1
4001 2 373 1 1
4001 3 414 0 1
4002 1   8 1 0
4002 2  26 1 0
4002 3 152 1 0
6025 1 146 1 0
6025 2 316 0 0
6025 1 316 0 0
6027 1 315 0 1
```

그림 7 원본의 파일

```
4054 1 0 293 0 1 0 0
4054 2 0 293 0 0 1 0
4054 3 0 293 0 0 0 1
4077 1 0 255 0 0 0 0
4077 2 0 255 0 0 0 0
4077 3 0 255 0 0 0 0
4109 3 0 213 0 0 0 0
4111 1 0 203 0 0 0 0
4111 2 0 203 0 0 0 0
4111 3 0 203 0 0 0 0
4001 1 0 219 1 1 0 0
4001 2 0 373 1 0 1 0
4001 3 0 414 0 0 0 1
4002 1 0   8 1 0 0 0
```

그림 8 cgd.d0

```
4002 2 0  26 1 0 0 0
4002 3 0 152 1 0 0 0
6025 1 0 146 1 0 0 0
6025 2 0 316 0 0 0 0
6025 3 0 316 0 0 0 0
6025 1 0 316 0 0 0 0
6026 2 0 316 0 0 0 0
6026 3 0 316 0 0 0 0
6027 1 0 315 0 1 0 0
6027 2 0 315 0 0 1 0
6027 3 0 315 0 0 0 1
```

일찍이 언급된 바와 같이 HMULCOX는 Andersen과 Gill[3]과 Prentice et al[8]등에의 방법을 사용하는 것을 의미한다. Andersen-Gill모형하에 한 대상에서는 재발사건의 위험은 일반적인 비례위험모형을 만족하고, 그런 종속이 정한 기간이 공변량으로서 모델에서 명시적으로 포함되지 않는다면 같은 단위에 나타나진 어떤 먼저 나타난 질병도 영향을 줄 수가 없다. 그림9-11은 일변량 공변량으로서 치료지수가 Andersen-Gill 모형에 적합된 입력과 출력을 보여준다. (그림9에 보여진) 이 분석을 위한 자료파일은 위험 구간(k=1이 0이고, k=2는 처음 감염이고, k=3은 두번째 감염인) 그림 7에서 새로운 행을 삽입하면서 (즉 그림 9의 3행)원 파일로 부터 만들어 졌다.

```
4054 1   0 293 0 1
4077 1   0 255 0 0
4109 1   0 213 0 0
4111 1   0 203 0 0
4001 1   0 219 1 1
4001 2 219 373 1 1
4001 3 373 414 0 1
4002 1   0   8 1 0
4002 2   8  26 1 0
4002 3  26 152 1 0
6025 1   0 146 1 0
6025 2 146 316 0 0
6026 1   0 316 0 0
6027 1   0 315 0 1
```

그림 9 CGD.D1

```
제목을 입력해 주세요          1번째 공변량의 이름은
 CGD TEST                   TREAT
입력된 자료명을 쳐주세요      자료 파일의 형은
 CGD.D1                     FREE
출력될 자료명을 쳐주세요      모형 1이면 1, 모형 2이면 2
 CGD.01                      1
UNITS(사람수)를 쳐주세요      다변량 가설의 수는
 9                          0
질병의 최대수를 쳐주세요      잠깐만 기다리세요
 3
최대 시간 구간의 수를 쳐주세요
 1
모형의 공변량(치료법)을 쳐주세요
 1
```

그림 10 입력

앞의 분석에와 같이 각 개인 환자에 대한 기록은 한개 하여야 한다. 그러나 두번째 혹은 세번째 감염을 위한 정보는 사이 시간이 60일 보다 더 된다면 두번째 구간에서 0의 값을

처음구간에서 1 값을 갖는 시간종속공변량 두 개의 기록으로서 나타내질 필요가 있다. 이 분석에서 자료 파일은 그림 12에서 보여진다. 제어모수의 입력과 출력은 그림 13, 그림 14에서 각각 보여진다.

Prentice et al.[8]에서 재발 질병에 대한 2가지 모형을 제의했다. 첫번째는 총시간을 다루는 것이고, 두번째는 사이 시간이다.(Prentice et al.에 더 많은 설명을 위해 LIN[7]를 보시오) 1개 공변량으로서의 치료지수를 갖는 총시간 모형을 분석하기 위해 식(2.1) 대신에 특성화된 식 (2.2)로서 그림 9에서 보여진 자료 파일을 사용한다. 그러면 그림16에서 주어진 결과를 얻을 수 있다. Prentice et al. 사이 시간 모형을 위한 자료 파일은 그림 11에서 보여준다. 이 파일은 총시간 모형에 사용된(그림 9에 보여진) 것과 (두번째와 세번재 감염을 위한) 사이시간은 총시간으로 대체되고 위험구간의 왼쪽 끝 점은 이전 감염시간이라기 보다 0이다. 사이 시간들의 일반적 치료 효과를 위한 추정은 표준오차추정치 1.126를 갖는 −0.8547이다.

```
4054 1   0 293 0 1 0
4077 1   0 255 0 0 0
4109 1   0 213 0 0 0
4111 1   0 203 0 0 0
4001 1   0 219 1 1 0
4001 2 219 279 0 1 1
4001 2 279 373 1 1 0
4001 3 373 414 0 1 1
4002 1   0   8 1 0 0
4002 2   8  26 1 0 1
4002 3  26  86 0 0 1
4002 3  86 152 1 0 0
6025 1   0 146 1 0 0
6025 2 146 206 0 0 1
6025 2 206 316 0 0 0
6026 1   0 316 0 0 0
6027 1   0 315 0 1 0
```

그림12 cgd.d3

```
제목을 입력해 주세요
 cgd test
입력된 자료명을 채주세요.
 cgd.d3
출력될 자료명을 채주세요.
 cgd.o3
UNITS(사람수)를 쳐주세요.
 9
질병의 최대수를 쳐주세요
 3
최대 시간 구간의 수를 쳐주세요
 2
모형의 공변량(치료법)을 쳐주세요
 2
1번째 공변량의 이름은
 treat
2번째 공변량의 이름은
   infhist
모형 1이면 1, 모형2이면 2
   1
다변량 가설의 수는
   0
잠깐만 기다리세요
```

그림13 입력

```
4054 1 0 293 0 1
4077 1 0 255 0 0
4109 1 0 213 0 0
4111 1 0 203 0 0
4001 1 0 219 1 1
4001 2 0 154 1 1
4001 3 0  41 0 1
4002 1 0   8 1 0
4002 2 0  18 1 0
4002 3 0 126 1 0
6025 1 0 146 1 0
6025 2 0 170 0 0
6026 1 0 316 0 0
6027 1 0 315 0 1
```

그림16 cgd study

읽어보세요!

```
 **************************
*                          *
*      다변량 발병자료 시간의      *
*                          *
*        COX 회귀분석          *
*                          *
**************************
제      목 : test
자료 파일명 : cgd.d1
출력 파일명 : cgd.o1

일반적인 기본 위험 함수의 추정
UNITS(사람수) =  9
 질 병 의  수 =  3
중도절단되지 않은 질병의 수
  질병의 형 :  1  2  3

  질병의 수 :  3  2  1

나이브 로그 순위 통계량 =      .60976  P-값 =   .43488

로버스트 로그 순위 통계량 =     .58366  P-값 =   .44488

모수  추정치  나이브 S.E  나이브 N.S.E  로버스트 S.E 로버스트 N.S.E

----  ------  ----------  ------------ ------------ --------------

trt   -.85475  1.12659    -.75870   .97425      -.87733

 베타의  나이브  공분산행렬

---------------------------------

  .127E+01

베타의  로버스트 공분산행렬

---------------------------------

  .949E+00
```

up/pgup:앞페이지 down/pgdn:다음페이지 Esc:그만보기

그림11 출력 (cgd.o1)

읽어보세요!

```
**************************
*                        *
*      다변량 발병자료 시간의      *
*                        *
*         COX 회귀분석          *
*                        *
**************************
제      목 : test
자료 파일명 : cgd.d3
출력 파일명 : cgd.o3

일반적인 기본 위험 함수의 추정
UNITS(사람수) =   9
 질 병 의  수 =   3
중도절단되지 않은 질병의 수
   질병의 형 :   1  2  3
   질병의 수 :   3  2  1

나이브 로그 순위 통계량 =  3.32656   P-값 =   .18952

로버스트 로그 순위 통계량 =  1.24892   P-값 =   .53555

모수   추정치  나이브 S.E   나이브 N.S.E  로버스트 S.E 로버스트 N.S.E

-----  -------  -----------  ------------  ------------  --------------

treat   -.64388  1.16448     -.55293     .96344      -.66831

infhist 1.88339  1.44148     -1.30657    .83114      2.26604

 베타의  나이브 공분산행렬

---------------------------------

  .136E+01   .325E+00

  .325E+00   .208E+01

 베타의  로버스트 공분산행렬

---------------------------------

  .928E+00   .570E+00

  .570E+00   .691E+00
```

up/pgup:앞페이지 down/pgdn:다음페이지 Esc:그만보기

그림 14 출력(cgd.o3)

읽어보세요!

```
**************************
*                        *
*     다변량 발병자료 시간의      *
*                        *
*       COX 회귀분석          *
*                        *
**************************
제     목 : test
자료 파일명 : cgd.d1
출력 파일명 : cgd.o1

일반적인 기본 위험 함수의 추정
UNITS(사람수) =   9
 질 병 의  수 =   3
중도절단되지 않은 질병의 수
   질병의 형 :  1  2  3
   질병의 수 :  3  2  1

나이브 로그 순위 통계량 =    .60976P-값 =   .43488

로버스트 로그 순위 통계량 =    .58366P-값 =   .44488

모수  추정치  나이브 S.E  나이브 N.S.E  로버스트 S.E 로버스트 N.S.E

----  ------  ----------  ------------ ------------ --------------

treat  -.85475  1.12659   -.75870   .97426      -.87733

베타의  나이브 공분산행렬

---------------------------------

  .127E+01

 베타의  로버스트 공분산행렬

---------------------------------

  .949E+00
```

up/pgup:앞페이지 down/pgdn:다음페이지 Esc:그만보기

그림15 출력 cgd.o1

참고문헌

[1] 이정진, 강근석(1994). 한국형 통계패키지 개발연구, 「응용통계연구」, 제7권, 제2호, 279-288

[2] 이정진, 강근석, 이원오, 김지현, 이창수, 김성철(1995). 전문가용 한국형 통계패키지 개발연구 I, 한국통계학회 논문집, 제2권, 2호 pp.434-444.

[3] P.K.Andersen and R.D.Gill, Cox's regression model for counting processes.: a large samples study, Ann.Statist. 10(1982)1100-1120

[4] E.W.Lee, L.J.Wei and D.A.Amato, Cox-type regression analysis for large numbers of small groups of correlated failure time observations, in Survival Analysis:State of the Art, Ed. J.P. Klein and P.K.Goel,pp237-247(Kluwer Academic Publishers, Dordrecht, 1992).

[5] D.Y.Lin, MULCOX:a computer program for the Cox regression analysis of multiple failure time varuables,Comput. Methods Programs Biomed. 32 (1990)125-135.

[6] D.Y.Lin,MULCOX2:a general computer program for the Cox regression analysis of multivariate failure time data, Computer Methods and Programs in Biomedicine,40 (1993) 279-293.

[7] D.Y.Lin,Cox regression analysis of multivariate failure time data:the marginal approach, Stat. Med.(submitted).

[8] R.L.Prentice, B.J.Williams and A.V.Peterson,On the regression analysis of multivariate failure time data, Biometrika 68(1981) 373-379.

[9] L. J. Wei, D. Y. Lin and L. Weissfeld, Regression analysis of multivariate incomplete failure time data by modeling marginal distributions, J. Am. Stat. Assoc. 84 (1989) 1065-1073.