

## 공변량을 포함한 사망시간과 치료횟수의 결합모형의 개발

박희창<sup>1</sup> · 박진표<sup>2</sup>

### 요약

본 논문에서는 환자의 수명과 치료횟수의 모형화를 위해 공변량을 포함한 결합모형을 개발하고자 한다. 개발된 모형을 Autologous Blood and Marrow Transplant Registry에 등록된 환자의 자료에 적용하여 모형에 포함된 모수의 추론과정을 논의하고자 한다.

### 1. 서론

일반적으로 생존분석분야에서 자주 발생하는 문제는 설명요인의 효과를 임의의 시간구간에서 일어나는 사건의 발생횟수에 대해 모형화하는 것이다. 이러한 문제가 일어나는 경우를 구체적으로 살펴 보면, 치명적인 질환을 가진 환자의 입원횟수나 응급실에서 심장발작을 일으킨 환자의 심장약 복용횟수를 모형화한다던가, 또는 골수이식환자의 회복과정에서 그들에게 취해진 수혈횟수에 대한 모형화 등을 그 예로 들 수 있다. 이러한 반복적으로 발생하는 사건에 대한 분석방법은 오랫동안 연구되어 왔는데, Cox와 Lewis(1996)는 반복적으로 발생하는 사건의 분석 방법의 여러 가지 양상을 논의 하였고, Lawless(1987)와 Thall(1988), 그리고 Lawless와 Nadeau(1995)는 여러 가지 시간간격 에서 사건발생의 기대값에 근거한 방법을 독자적으로 발표하였다. Wei등(1989)은 특정한 사건이 발생할 때까지의 시간에 대한 분석방법을 토의하였으며, Aalen과 Husebye(1991)는 사건 발생간격에 대해 언급하였다. 또한 Andersen 등(1993)은 counting process적인 접근으로 반복적으로 발생하는 사건의 분석방법을 논의하였다.

이 논문에서는 환자의 사망시간과 치료 횟수에 관하여 모형화하는데 있어서 관측가능한 공변량을 포함한 결합모형을 개발하고자 한다. 2절에서는 이 논문에서 사용되는 기호와 기본 가정, 그리고 모형의 특성을 기술하고, 3절에서는 모형에 포함된 모수의 추론과정을 논의하며, 4절에서는 개발된 모형을 Autologous Blood and Marrow Transplant Registry(ABMTR)에 등록된 환자들 중 사망위험성이 상당히 높은 유방암환자 701명의 자료에 적용하고자 한다.

<sup>1</sup>(641-773) 경남 창원시 사림동 9, 창원대학교 통계학과 부교수

<sup>2</sup>(631-701) 경남 마산시 합포구 월영동 449 번지, 경남대학교 컴퓨터공학과 부교수

## 2. 모형의 개발

이 절에서는 사망시간  $X$ 와 환자의 관찰시간  $T$ 까지 일어난 사건발생횟수  $N(t)$ 의 결합 분포에 대한 모형화를 고려하고자 한다. 중도 절단여부를 표현하기 위해 만약  $T = X$ 이면  $\delta = 1$ 로,  $T > X$ 이면  $\delta = 0$ 으로 표기한다. 조사대상환자에 대해 두 집단의 공변량을 고려하는데, 이들은 사망시간과 사건발생횟수 각각에 대한 잠재적인 설명요인이거나 두 변량 모두에 대한 잠재적인 설명요인이다. 사망시간에 대해서는  $P_d$ 개의 공변량을 고려하고, 사건발생횟수에 대해서는 공변량의 수를  $P_c$ 개로 한다. 따라서  $Z_d$ 를 절편항을 나타내기 위한 첫 번째 원소값 1과  $P_d$ 개의 공변량으로 구성된 사망시간에 대한 공변량 벡터  $(P_d + 1)$ 로 정의하고,  $Z_c$ 를 이와 유사하게 첫 번째 원소의 값 1과 사건발생횟수에 대한  $P_c$ 개의 공변량으로 구성된 공변량벡터  $(P_c + 1)$ 로 정의한다.  $Z_d$ 와  $Z_c$ 에는 동일한 요인이 포함될 수도 있다.

공변량  $Z_d$ 가 주어졌을 때 환자의 사망시간은 다음과 같은 위험률을 갖는 와이블분포를 따르는 것으로 가정한다.

$$h(t|Z_d) = \alpha t^{\alpha-1} \exp(\beta Z_d), \quad t \geq 0, \alpha \geq 0 \quad (1)$$

식 (1)은 표준적인 와이블 회귀모형이다. 사건발생횟수에 대해서는 공변량  $Z_c$ 가 주어졌을 때 치료횟수  $N(t)$ 는 다음과 같은 비율을 갖는 Poisson 과정을 따르는 것으로 가정한다.

$$\lambda(t|Z_c) = t^\phi \exp(\gamma Z_c), \quad t \geq 0, \phi \geq 0 \quad (2)$$

이러한 가정하에서 개발된 모형의 성질을 규명하기 위해  $X$ 와  $N(\cdot)$ 의 결합확률을 구하여야 하는데,  $N(t)$ 는 관찰기간  $t$ 의 값이 환자의 사망시간보다 짧은 경우에만 관측가능하며, 만약  $X > t$ 인 경우에는  $N(t) = N(X)$ 가 된다는 사실하에서 다음과 같이 얻어진다.

$$P[X = t, N(s) = k] = \alpha t^{\alpha-1} \exp\{\beta Z_d\} \exp[-t^\alpha \exp\{\beta Z_d\}] \times \frac{\{(min(s, t))^\phi \exp[\gamma Z_c]\}^k \exp[-(min(s, t))^\phi \exp[\gamma Z_c]]}{k!} \quad (3)$$

또한  $t$ 이상의 값을 취하는  $X$ 와  $N(\cdot)$ 의 결합확률을 구할 수 있는데, 먼저  $t \geq s$ 인 경우에는 다음과 같다.

$$P[X \geq t, N(s) = k] = \exp[-t^\alpha \exp\{\beta Z_d\}] \frac{\{s^\phi \exp[\gamma Z_c]\}^k \exp[-s^\phi \exp[\gamma Z_c]]}{k!} \quad (4)$$

또한  $s > t$ 인 경우에는 다음과 같이 얻어진다.

$$\begin{aligned}
 P[X \geq t, N(s) = k] &= \frac{\alpha \exp[\beta Z_d + k\gamma Z_c]}{k!} \\
 &\times \int_t^s u^{k\phi + \alpha - 1} \exp[-u^\alpha \exp(\beta Z_d) - u^\phi \exp(\gamma Z_c)] du \\
 &+ \exp[-s^\alpha \exp(\beta Z_d)] \frac{[s^\phi \exp(\gamma Z_c)]^k \exp[-s^\phi \exp(\gamma Z_c)]}{k!}
 \end{aligned} \tag{5}$$

이러한 모형에 대해  $X$ 의 주변분포를 구할 수 있는데, 이는 다음과 같이 얻어진다.

$$S(t) = \exp[-t^\alpha \exp(\beta Z_d)] \tag{6}$$

$N(s)$ 의 주변분포를 얻기 위해서  $P[X \geq 0, N(s) = k]$ 을 구하면 되는데, 이는 식(5)로부터 다음과 같이 구해진다.

$$\begin{aligned}
 P[N(s) = k] &= \frac{\alpha \exp[\beta Z_d + k\gamma Z_c]}{k!} \\
 &\times \int_0^s u^{k\phi + \alpha - 1} \exp[-u^\alpha \exp(\beta Z_d) - u^\phi \exp(\gamma Z_c)] du \\
 &+ \exp[-s^\alpha \exp(\beta Z_d)] \frac{[s^\phi \exp(\gamma Z_c)]^k \exp[-s^\phi \exp(\gamma Z_c)]}{k!}
 \end{aligned} \tag{7}$$

### 3. 모수의 추정

모형에 포함된 모수의 추정은 관찰기간동안 환자들에게 행해진 수혈의 총횟수에 근거하고 있다.  $T_i$ 를  $i$ 번째 환자의 관찰기간이라고 하고,  $\delta_i$ 를  $i$ 번째 환자의 사망여부를 나타내는 지시변수라 하자. 즉, 환자가 사망했을 경우에는  $\delta_i$ 의 값이 1이 되며, 환자가 중도 절단된 경우에는  $\delta_i$ 의 값이 0이 된다.  $N_i = N_i(T_i)$ 를  $i$ 번째 환자가 받는 총수혈횟수라 하자. 본 논문에서는 Lawless(1987)의 연구와는 달리 각 개인이 경험한 사건의 정확한 발생시간은 알려져 있지 않고 단지 전체 사건발생횟수만 알려진 자료에 대하여 분석하는 것이다.  $Z_{di}$ 와  $Z_{ci}$ 를  $i$ 번째 환자의 공변량벡터라고 하자. 관찰기간동안 식 (6)으로부터 얻어지는 사망한 환자들에 대한 확률값과 식 (7)로부터 얻어지는 중도 절단된 환자들에 대한 확률값을 이용하여 우도함수를 구할 수 있는데, 표본의 크기  $n$ 을 근거로 하여 대수우도함수를 구하면 대수우도함수  $LL$ 의 값이 다음의  $L_1(\alpha, \beta)$ 와  $L_2(\phi, \gamma)$ 의 합으로 얻어지는데, 전자는 일반적인

Weibull 회귀모형 으로부터 얻어지는 우도함수이고, 후자는 Poisson 우도함수이다.

$$L_1(\alpha, \beta) = \sum_{i=1}^n \left[ \delta_i \{ (\alpha - 1) \ln[T_i] + \ln[\alpha] + \beta Z_{di} \} - T_i^\alpha \exp[\beta Z_{di}] \right]$$

$$L_2(\phi, \gamma) = \sum_{i=1}^n \left[ N_i \{ \phi \ln[T_i] + \gamma Z_{ci} \} - T_i^\phi \exp[\gamma Z_{ci}] \right] \quad (8)$$

모수  $\alpha, \phi, \beta$ , 그리고  $\gamma$ 에 대한 추정값은 대수우도함수  $LL$ 을 수치 해석적으로 최대화함으로써 얻을 수 있다.  $LL$ 의 각 모수에 대한 1차 편미분값인 score 통계량과 관측된 정보행렬(observed information matrix)을 부록에 제시하였는데, 이는 각 모수의 2차 편도함수의 음의 값으로 구성된다. 모수의 관측된 정보행렬의 역행렬이 최우추정량의 공분산 행렬이 된다. 또한 모형에 포함된 모수에 대한 검정은 Wald검정법을 이용하여 수행할 수 있다.

#### 4. 예 제

이 절에서는 ABMTR에 등록된 환자들 중 사망위험성이 상당히 높은 유방암환자 701명의 자료를 이용하여 위에서 논의한 모형에 대한 추론문제를 토의하고자 한다. 이들 환자들은 1990년부터 1994년 사이에 이식수술을 받았으며, 1996년까지 추적조사가 진행되었다. 평균추적기간은 16.5개월이고 그 범위는 0.1에서 66개월이다. 환자들 중에서 361(51.1%)명이 관찰기간동안 사망하였고, 환자들은 동일한 골수 기증자로부터 평균 5회의 수혈을 받은 것으로 나타났다. 환자들 중 144명은 한 번도 수혈을 받지 않았으며, 가장 많이 수혈을 받은 환자는 그 수가 172회나 된다.

본 논문에서는 5종류의 잠재적인 위험요인을 고려하였는데 이들 중 두 요인은 연속적인 값을 취하는 공변량으로서 이식 당시의 환자의 연령(평균:44세, 범위:24.5 ~ 64.4세)과 진단에서부터 이식까지의 대기시간(평균:626일, 범위:92 ~ 4811일)이다. 다른 하나의 요인은 이식년도로서 이진(binary) 변량(90 ~ 92:44%, 93 ~ 94:56%)으로 고려하였으며, 나머지 두 요인은 범주형 변량으로서 이들 중 하나는 이식 당시 환자의 질병단계를 나타내는 것으로 primary stage 2-3 disease(PS;32.3%), metastatic disease in complete remission(CR; 18.7%), partial remission(PR; 28.0%), 그리고 resistant disease(RD; 21.1%)로 구분하였으며, 다른 하나는 이식조직원(graft source)으로, 골수(bone marrow(BM), 49.4%), 말초혈관줄기세포(peripheral blood stem cell (PBSC), 39.5%), 그리고 골수와 말초혈관줄기세포(Both)로 구분하였다. 이에 대한 보다 구체적인 내용은 Antman 등(1997)의 연구에 잘 나타나 있다.

모형에 포함된 모수의 추론을 위해 본 논문에서는 식(8)의 우도함수를 수치적으로 최대화하기 위해 Marquardt(1963)의 방법을 사용하였는데, 이는 steepest decent법과 Newton-Raphson기법을 절충한 수치 해석적 방법이다. 이 방법을 사용한 이유는 개발된 모형에 포

함된 모수의 초기추정값을 얻기가 어려울 뿐만 아니라 추정해야 할 모수의 수가 상당히 많기 때문이다. 또한 모수의 추정과정을 위해 FORTRAN으로 프로그램을 작성하였다.

최적 모형을 구하기 위해 먼저 사망시간과 수혈횟수 양쪽 모두에 5종류의 요인 전부를 개발된 모형에 고려하였다. < 표 1 > 에 모수의 추정값과 표준오차, 그리고 모수의 값이 0이라는 가설에 대한 Wald 검정통계량과 유의 확률값을 제시하였다. 그런 후 유의수준 5%하에서 유의한 모든 요인을 최종 모형에 포함시키기 위해 후진소거과정을 통하여 한 번에 하나씩 요인을 제거하였다. 그 결과 최종 모형은 < 표 2 >와 같이 주어졌다. 이 표로부터 알 수 있는 바와 같이 사망시간에 대한 위험요인은 질병의 단계, 이식조직원, 그리고 대기시간 등이다. 수혈횟수에 대해서는 위에서 고려한 5종류의 요인 전부 즉, 질병의 단계와 이식조직원, 이식년도, 연령, 그리고 대기시간이 유의한 것으로 나타났다.

본 논문에서는 공변량을 포함한 환자의 수명과 치료횟수의 결합모형을 개발하였다. 또한 개발된 모형을 ABMTR에 등록된 환자의 자료에 적용하여 모형에 포함된 모수의 추론과정을 논의하였다. 모형진단에 관한 연구나 모의실험을 통한 효율성을 비교 검토하는 문제는 추후 연구과제로 남긴다.

## 참고 문헌

1. Aalen, O. O. and Husebye, E. (1991). Statistical Analysis of Repeated Events Forming Renewal Processes, *Statistics in Medicine*, 10, 1227-1240
2. Anderson, P.K., Borgan, O., Gill, R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.
3. Bandyopadhyay, D. and Basu, A.P. (1990). On a Generalization of Model by Lindly and Singpurwalla, *Annals of Applied Probability*, 22, 498-500.
4. Cox, D.R. and Lewis, P.A.W. (1966). *The Statistical Analysis of Series of Events*, Methuen, London.
5. Gupta, P.L. and Gupta, R.D. (1990). A Bivariate Environmental Stress Model, *Advanced in Associations in Applied Probability*, 22, 501-503
6. Klein, J. P., Moeschberger, M. L. (1997). *Survival Analysis: Techniques For Censored And Truncated Data*, Springer-Verlag, New York.
7. Klein, J. P., Moeschberger, M. L., Li, Y. H. and Wang S. T. (1992). *Estimating random effects in the Framingham heart study*, in J. P. Klein and P. Goel, Editors, *Survival Analysis: State of The Art*, 99-120. Kluwer Academic Publishers, Boston.
8. Lawless, J.F. (1987). Regression methods for Poisson process data, *Journal of the American Statistical Association*, 82, 808-815.
9. Lawless, J.F. and Nadeau, C. (1995). Some Simple Robust Methods for the Analysis of Recurrent Events, *Technometrics*, 35, 158-168.

10. Lee, S.(1968). Inference for a bivariate survival function induced through the environment, Ph.D. Dissertation, The Ohio State University.
11. Lee, S. and Klein, J.P.(1988). Bivariate Models with a Random Enviromental Factor, *Indian Journal of Productivity, Reliability and Quality Control*, 13, 1-18.
12. Lee, S. and Klein, J.P.(1989). *Statistical Methods for Combining Laboratory and Field Data Based On a Random Environmental Stress Model*, Recent Developments in Statistics and Their Applications, Klein and Lee, Eds., pp. 87-116.
13. Lindley, D. V. and Singpurwalla, N. A.(1986). Multivariate Distributions for the Reliability of a System of Components Sharing a Common Environment, *Journal of Applied Probability*, 23, 418-431.
14. Marquardt, D. (1963). An algorithm for least-squares estimation of non-linear parameters, *SIAM Journal of Applied Mathematics*, 11, 431-441.
15. Nayak, T.K.(1987). Multivariate Lomax Distribution: Properties and Usefulness in Reliability Theory, *Journal of Applied Probability*, 24, 170-177.
16. Thall, P.F.(1988). Mixed Poisson Regression Models for Longitudinal Interval Count Data, *Biometrics*, 44, 197-209.
17. Vaupel, J. W., Manton, K. G. and Stallard, E. (1979). The impact of Heterogeneity in individual Frailty and the dynamics of mortality, *Demography*, 16, 439-454.
18. Wei, L.J., Lin, D.Y. and Weissfeld, L.(1989). Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distribution, *Journal of the American Statistical Association*, 84, 1065-1073.
19. Whitmore, G. A. and Lee, M. T.(1991). A Multivariate Survival Distribution Generated by an Inverse Gaussian Mixture of Exponentials, *Technometrics*, 33, 9-50.

< 표 1 > 초기모형에 포함된 모수의 최우추정값

구분	요인	추정값	표준오차	자유도	Wald통계량	p값
사망시간	절편	-4.77	0.389			
	질병의 단계			3	118.32	< 0.0001
	CR	0.67	0.193	1	11.92	0.0006
	PR	1.53	0.173	1	78.99	< 0.0001
	RD	1.74	0.178	1	95.45	< 0.0001
	이식조직원			2	7.61	0.0223
	PBSC	-0.17	0.122	1	1.95	0.1631
	Both	0.31	0.175	1	3.05	0.0806
	이식년도(90-92)	0.03	0.120	1	0.08	0.7832
	연령	$5.0 \times 10^{-3}$	0.007	1	0.45	0.5001
	대기시간	$-1.9 \times 10^{-4}$	$6.0 \times 10^{-5}$	1	9.87	0.0017
	$\alpha$	1.03	0.047			
	Weibull 대수우도값	-1599.07				
수혈횟수	절편	2.39	0.079			
	질병의 단계			3	801.59	< 0.0001
	CR	-0.04	0.037	1	1.42	0.2338
	PR	-0.27	0.037	1	55.86	< 0.0001
	RD	0.58	0.033	1	304.15	< 0.0001
	이식조직원			2	236.31	< 0.0001
	PBSC	-0.37	0.027	1	181.01	< 0.0001
	Both	0.05	0.035	1	1.70	0.1924
	이식년도(90-92)	-0.12	0.025	1	23.74	< 0.0001
	연령	0.01	0.002	1	83.61	< 0.0001
	대기시간	$-3.2 \times 10^{-5}$	$1.2 \times 10^{-5}$	1	7.20	0.0073
	$\phi$	-0.18	0.010			
	Poisson 대수우도값	-6450.54				
	전체 대수우도값	-8049.61				

< 표 2 > 최종모형에 포함된 모수의 최우추정값

구분	요인	추정값	표준오차	자유도	Wald통계량	p값
사망시간	절편	-4.54	0.205			
	질병의 단계			3	122.22	< 0.0001
	CR	0.67	0.192	1	12.05	0.0005
	PR	1.53	0.172	1	79.73	< 0.0001
	RD	1.76	0.176	1	100.00	< 0.0001
	이식조직원			2	7.67	0.0216
	PBSC	-0.17	0.118	1	2.17	0.1409
	Both	0.30	0.173	1	3.03	0.0820
	대기시간	$-1.8 \times 10^{-4}$	$5.9 \times 10^{-5}$	1	9.61	0.0019
	$\alpha$		1.04	0.046		
Weibull 대수우도값		-1599.32				
수혈횟수	절편	2.39	0.079			
	질병의 단계			3	801.59	< 0.0001
	CR	-0.04	0.037	1	1.42	2338
	PR	-0.27	0.037	1	55.86	< 0.0001
	RD	0.58	0.033	1	304.15	< 0.0001
	이식조직원			2	236.31	< 0.0001
	PBSC	-0.37	0.027	1	181.01	< 0.0001
	Both	0.05	0.035	1	1.70	0.1924
	이식년도(90-92)	-0.12	0.025	1	23.74	< 0.0001
	연령	0.01	0.002	1	83.61	< 0.0001
대기시간	$-3.2 \times 10^{-5}$	$1.2 \times 10^{-5}$	1	7.20	0.0073	
$\phi$		-0.18	0.010			
Poisson 대수우도값		-6450.54				
전체 대수우도값		-8049.86				

부 록

식(8)의 대수우도함수의 1차 편도함수인 score통계량은 다음과 같다.

$$\frac{\partial LL}{\partial \alpha} = D/\alpha + S_{DLT} - \sum_{i=1}^n \ln[T_i]Y_{1i}; \tag{9}$$

$$\frac{\partial LL}{\partial \phi} = S_{NLT} - \sum_{i=1}^n \ln[T_i]Y_{2i}; \tag{10}$$



$$\frac{\partial LL}{\partial \beta_j} = \sum_{i=1}^n \delta_i Z_{di} - \sum_{i=1}^n Z_{di} Y_{1i}, j = 0, 1, \dots, P_d; \quad (11)$$

$$\frac{\partial LL}{\partial \gamma_j} = \sum_{i=1}^n N_i Z_{ci} - \sum_{i=1}^n Z_{ci} Y_{2i}, j = 0, 1, \dots, P_c; \quad (12)$$

여기서  $D$ 는 사망자수이며,  $S_{DLT} = \sum_{i=1}^n \delta_i \ln[T_i]$ ,  $S_{NLT} = \sum_{i=1}^n N_i \ln[T_i]$ ,  $Y_{1i} = T_i^\alpha \exp[\beta Z_{di}]$ , 그리고  $Y_{2i} = T_i^\phi \exp[\gamma Z_{ci}]$  이다.

대수우도함수의 2차 편도함수는 다음과 같다.

$$\frac{\partial^2 LL}{\partial \alpha^2} = -D/\alpha^2 - \sum_{i=1}^n (\ln[T_i])^2 Y_{1i}; \quad (13)$$

$$\frac{\partial^2 LL}{\partial \alpha \partial \beta_j} = - \sum_{i=1}^n Z_{di} \ln[T_i] Y_{1i}, j = 0, 1, \dots, P_d; \quad (14)$$

$$\frac{\partial^2 LL}{\partial \phi^2} = - \sum_{i=1}^n (\ln[T_i])^2 Y_{2i}; \quad (15)$$

$$\frac{\partial^2 LL}{\partial \phi \partial \gamma_j} = - \sum_{i=1}^n Z_{ci} (\ln[T_i]) Y_{2i}, j = 0, 1, \dots, P_c; \quad (16)$$

$$\frac{\partial^2 LL}{\partial \beta_j \partial \beta_k} = - \sum_{i=1}^n Z_{di} Z_{dik} Y_{1i}; \quad (17)$$

$$\frac{\partial^2 LL}{\partial \gamma_j \partial \gamma_k} = - \sum_{i=1}^n Z_{ci} Z_{cik} Y_{2i}; \quad (18)$$

$$\frac{\partial^2 LL}{\partial \alpha \partial \phi} = 0; \quad (19)$$

$$\frac{\partial^2 LL}{\partial \alpha \partial \gamma_j} = 0; \quad (20)$$

$$\frac{\partial^2 LL}{\partial \phi \partial \beta_j} = 0; \quad (21)$$

$$\frac{\partial^2 LL}{\partial \beta_j \partial \gamma_k} = 0; \quad (22)$$

## Joint Modeling of Death Times and Counts with Covariate

Hee-Chang Park <sup>3</sup> · Jin-Pyo Park <sup>4</sup>

### Abstract

In this paper we suggest the joint model of death times and counts with covariates. We assume that the death times follow a Weibull distribution with rate that depends on covariates. For the counts, a Poisson process is assumed with the intensity depending on time and the covariates. We obtain the maximum likelihood estimators of model parameters. This model is applied to data set of patients with breast cancer who received a bone marrow transplant.

---

<sup>3</sup>Associate Professor, (641-773) Department of Statistics, Changwon National University

<sup>4</sup>Associate Professor, (631-701) Department of Computer Engineering, Kyungnam University