Journal of the Korean Data Analysis Society (June 2015)
Vol. 17, No. 3 (B), pp. 1261-1271

이분형 반복측정 자료와 생존 자료의 결합분석 모형 : 실제 자료 분석 사례를 중심으로*

조혜수[1], 정인경[2]

요 약

결합분석 모형은 두 종류 이상의 결과변수를 동시에 분석할 수 있는 방법으로, 각 결과변수에 대한 모형을 임의효과를 공유하는 형태로 결합하는 것이 가장 일반적이다. 연속형 반복측정 자료와 생존 자료의 결합모형에 관한 연구는 많이 이루어졌으나, 이분형 반복측정 자료를 다룬 연구는 많지 않다. 본 연구에서는 일반화선형혼합모형과 세 가지 다른 형태의 생존모형을 임의효과를 공유하는 결합분석 모형으로 구축하여 미국 샌안토니오 노인코호트 연구(San Antonio longitudinal study of aging; SALSA) 자료를 분석하였다. 연구 대상자들의 신체적 허약 상태를 반복 측정한 결과와 사망까지의 시간의 두 결과변수를 개별적으로 분석했을 경우와 비교하였고, 공유되는 임의효과의 계수의 추정을 통해 두 결과변수 간의 관련성에 대해 살펴보았다. 결합분석 모형을 이용하여 분석했을 경우에는 회귀계수가 기존 분석 결과와 다소 다른 값으로 추정되었고 몇몇 공변량에 대해서는 통계적 유의성이 달라졌다. 다른 형태의 결합분석 모형 간 큰 차이는 없었으나, Cox의 비례위험모형을 이용한 결합분석 모형의 경우에 공유되는 임의효과의 계수가 유의하게 추정되어 두 결과변수 간의 관련성을 설명할 수 있었다.

주요용어 : 결합모형, 임의효과 공유, 일반화선형혼합모형, 비례위험모형.

1. 서론

보건학이나 의학 분야의 종적연구(longitudinal study)에서 동일한 대상자에 대하여 반복측정 자료와 생존시간 자료가 동시에 수집되는 경우가 많다. 서로 다른 두 종류의 결과변수를 각각 다른 모형으로 개별적으로 분석하는 것이 고전적으로 이루어졌던 방법이다. 반복측정 자료는 반복측정 분산분석모형(repeated measures ANOVA)이나(Kwon, Cho, 2007) 혼합효과모형(mixed effects model)을 이용하여(Kim, 2005) 분석할 수 있고, 생존시간 자료는 준모수적 모형인 Cox regression을 활용하거나(Kim, Bae, 2006) 지수분포나 Weibull 분포 등을 이용한 회귀모형(Kim, Ahn, 2011)을 사용할 수 있다. 하지만, 두 결과변수 간에 연관성이 존재할 때, 예를 들어, 암환자들을 대상으로 삶의 질과 사망까지의 시간에 대한 연구를 하는 경우, 삶의 질을 나타내는 척도를 매년 측정하고 사망까지의 시간 자료를 수집한다면, 그 둘 사이에는 연관성이 존재할 것으로 생각할 수 있고, 그러한 경우에는 두 결과변수의 연관성을 고려한 분석 방법을 이용하는 것이 더 적절하다. 두 결과변수를 각각

*연세대학교 의과대학 2013년도 일반교수연구비에 의하여 이루어졌음(6-2013-0139).
*본 논문에서 사용된 자료는 미국 샌안토니오 노인코호트 연구(San Antonio longitudinal study of aging; SALSA)의 연구책임자 Dr. Helen P. Hazuda에게 사용 허락을 받았음.
*이 논문은 제1저자 조혜수의 석사학위논문(Cho, 2014)을 바탕으로 추가연구하여 작성한 것임.
[1]120-752 서울시 서대문구 연세로 50-1, 세브란스병원 임상시험센터 직원. E-mail : chsoo@yuhs.ac
[2](교신저자) 120-752 서울시 서대문구 연세로 50-1, 연세대학교 의과대학 의학정보통계학과 부교수.
E-mail : ijung@yuhs.ac
[접수 2015년 3월 31일; 수정 2015년 5월 19일, 2015년 6월 11일; 게재확정 2015년 6월 14일]

개별적으로 분석하는 경우에 비해 더 효율적인 추정치를 얻을 수 있고, 편향(bias)을 줄일 수 있기 때문이다(Ibrahim, Chu, Chen, 2010). 최근 두 가지 이상의 다른 종류의 결과변수를 하나의 모형 안에서 분석하는 결합 모형(joint model)에 대한 연구가 많이 이루어지고 있는데, 대부분의 연구에서 연속형 형태의 반복측정 자료를 다루고 있다(Tsiatis, DeGruttola, Wulfsohn, 1995; Wulfsohn, Tsiatis, 1997; Wang, Taylor, 2001; Tsiatis, Davidian, 2004). 하지만, 반복측정 자료가 이분형이거나 연속형으로 얻어진 자료일지라도 이분화해서 사용하는 것이 요구되는 경우, 이분형 반복측정 자료와 생존자료를 하나의 모형에서 동시에 분석하는 것이 필요하다.

최근 Choi, Cai, Zeng, Olshan(2015)이 이분형 반복측정 자료와 생존시간 자료의 결합분석(joint analysis)을 위한 모형을 제안하였다. 생존시간 자료는 Cox의 비례위험모형으로, 이분형 반복측정 자료는 일반화선형혼합모형(generalized linear mixed model, GLMM)으로 모형화하면서 두 결과변수의 연관성을 반영하기 위하여 임의효과(random effect)를 공유하는 형태로 결합모형을 구성한다. Choi, Cai, Zeng, Olshan(2015)의 연구에서는 EM 알고리즘(expectation-maximization algorithm)을 통한 최대우도추정(maximum likelihood estimation)방법으로 모형을 추정하는데, 이를 이용하기 위해서는 특별한 프로그래밍이 필요하며 접근하기가 쉽지 않다. Vonesh, Greene, Schluchter(2006)가 제안한 방법도 (일반화)선형혼합모형과 모수적 또는 준모수적 생존모형을 임의효과를 공유하는 형태로 결합모형을 구성하고 최대우도추정방법으로 모형을 추정하는데, 우도함수를 근사하여 계산하고 최대우도추정치를 구하는 과정은 SAS의 NLMIXED procedure를 이용하여 구현될 수 있는 방법을 사용하므로 좀 더 접근하기 쉬운 장점이 있다. 하지만, Vonesh, Greene, Schluchter(2006)의 연구에서도 연속형 반복측정 결과변수의 자료를 사례로 소개하고 있으며 실제로 이분형 자료에 적용된 사례는 찾아보기 어렵다.

본 연구에서는 미국 샌안토니오 노인코호트 연구(The San Antonio longitudinal study of aging: SALSA)의 자료를 이용하여 이분형 반복측정 자료와 생존시간 자료의 결합분석모형을 실제로 구현하여 분석해 보고자 한다. 샌안토니오 노인코호트 연구(SALSA) 자료는 65세 이상의 멕시코계 미국인과 유럽계 미국인을 대상으로 하여 신체적인 허약 상태의 여부를 총 네 번 반복 측정한 결과와 대상자들의 사망여부와 생존시간에 대한 정보를 포함하고 있다. 조사된 노인들의 신체적 허약 상태가 인종 간에 차이가 있는지 알아보기 위해 분석한 내용의 연구결과(Espinoza, Jung, Hazuda, 2010)와 노인들의 사망 확률이 인종에 따라 차이가 있는지 알아보기 위해 실시한 생존 자료에 대한 분석 결과(Espinoza, Jung, Hazuda, 2013)는 이미 발표된 바 있다. 하지만 두 결과변수 간의 관련성이 존재할 것이라고 생각되므로 이 관련성을 반영하는 결합분석모형으로 분석하여 기존의 연구결과와 비교하고 차이를 알아보고자 한다. 또한 구축한 결합분석모형의 추정 결과를 통해서 두 결과변수의 연관성에 대한 유의성을 검토해본다.

2절에서는 반복측정 자료와 생존 자료의 관련성을 임의효과 공유를 통해 반영하고 있는 결합분석 모형을 소개한다. 3절에서는 SALSA 자료에 대하여 결합분석 모형을 여러 형태로 구축하고 그 결과를 비교해본다. 마지막으로 4절에서는 결론 및 고찰을 제시한다.

2. 결합분석 모형

Vonesh, Greene, Schluchter(2006)와 Choi, Cai, Zeng, Olshan(2015)의 연구 내용을 바탕으로 이분형 반복측정 자료와 생존시간 자료의 결합분석 모형을 구축한다. 두 결과변수의 관련성을 반영하기 위하여 연구 대상자가 갖는 임의효과를 모형들 간에 공유시킴으로써 결합모형을 구축할 수 있다. 본 연구에서는 결합분석 모형 안에서 정의되는 반복측정 자료에 대한 분석모형을 부차모형

(Submodel) 1, 생존분석 모형을 부차모형(Submodel) 2로 정의하고, 임의효과 u_i를 두 모형이 공유하며 부차모형 2에서 그에 대한 계수 ψ로 두 변수 간 관련성을 표현한다. 부차모형 1은 일반화선형모형으로, 부차모형 2는 모수적 또는 준모수적 생존모형의 여러 형태로 정의할 수 있다. n명의 대상자에 대하여, i번째 대상자의 j번째 반복 측정된 관측치를 y_{ij}라 하고, 생존시간을 T_i라 하자. 생존시간은 우측 중도절단(right censoring)이 가능하다고 가정하면 실제로 관찰되는 생존시간은 $T_i^* = \min(T_i, C_i)$이고, 표시함수 $\delta_i = I(T_i \le C_i)$로 중도절단 여부를 정의한다. 임의효과 u_i가 주어졌을 때, 반복측정 자료와 생존시간 자료는 독립이라고 가정한다. 각 부차모형은 임의효과가 주어졌을 때의 조건부 모형으로 다음과 같이 나타낼 수 있다.

· Submodel 1: $g(E[y_{ij} \mid u_i]) = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_\pi + u_i$

· Submodel 2: $h(t \mid u_i) = h_0(t)\exp(\alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_\pi + \psi u_i) = h_o(t)\exp(\eta_i)$ (1)

여기서, $x_{1i}, x_{2i}, \cdots, x_\pi$은 i번째 대상자의 공변량을 나타내는데, 두 부차모형에 포함되는 변수는 동일하지 않아도 무방하다. $g()$는 연결함수로 이분형 자료에 대해서는 보통 로짓(logit) 함수를 사용한다. $h()$는 생존시간에 대한 위험함수(hazard function)로 기저위험함수 $h_0()$의 형태에 따라 여러 형태의 생존분석 모형을 구축할 수 있다. 본 연구에서는 Cox의 준모수적 모형, 지수 모형, Weibull 모형 세 가지 형태의 비례위험을 가정하는 모형으로 부차모형 2를 구축한다. 공변량과 생존시간과의 관계를 직접 모형화하는 가속화 고장시간(accelerated failure time; AFT) 모형의 형태로 부차모형 2를 구축하는 것도 가능하다. 지수모형과 Weibull 모형은 AFT 모형의 형태로도 표현할 수 있다.

위와 같이 구축된 모형은 최대우도추정법으로 회귀계수를 추정하여 모형에 대한 추론을 한다. 임의효과 u_i가 주어졌을 때, 반복측정 변수와 생존시간 변수는 독립이라 가정하므로 두 변수의 조건부 결합분포의 확률밀도함수는 각 변수의 조건부 주변분포의 확률밀도함수들의 곱으로 표현된다. 이것에 u_i의 확률밀도함수를 곱하여 u_i에 대해 적분하면 두 변수의 결합분포에 대한 확률밀도함수를 구할 수 있다. 따라서 구하고자 하는 우도함수는 다음과 같이 표현된다.

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}, \psi; \boldsymbol{y_i}, T_i^*) = \prod_{i=1}^{n} \int_{u_i} f_1(\boldsymbol{y_i} \mid u_i) f_2(T_i^* \mid u_i)^{\delta_i} S(T_i^* \mid u_i)^{1-\delta_i} f_3(u_i) du_i$$

여기서, $\boldsymbol{\beta} = (\beta_o, \beta_1, \beta_2, \cdots, \beta_p)'$, $\boldsymbol{\alpha} = (\alpha_o, \alpha_1, \alpha_2, \cdots, \alpha_p)'$, $\boldsymbol{y_i} = (y_{i1}, y_{i2}, \cdots, y_{in_i})'$이고, $f_1(\boldsymbol{y_i} \mid u_i)$는 $\boldsymbol{y_i}$의 조건부 확률밀도함수, $f_2(T_i^* \mid u_i)$는 T_i^*의 조건부 확률밀도함수, $S(T_i^* \mid u_i)$는 조건부 생존함수, $f_3(u_i)$는 u_i의 확률밀도함수이다. u_i에 대해 보통 정규분포를 가정하는데, 위의 우도함수는 직접 계산하는 것이 불가능하고, 수치적분을 통해 근사하는 방법을 사용한다. 본 연구에서는 Cox의 비례위험모형으로 결합모형을 구축하는 경우에는 Choi, Cai, Zeng, Olshan(2015)이 제시한 R 함수를 이용하여 EM 방법으로 추정하였고, 지수모형과 Weibull 모형으로 구축하는 경우에는 SAS NLMIXED procedure를 이용하여 Laplace 근사법을 통하여 모형을 추정하였다.

3. 실제 자료 분석

3.1. 미국 샌안토니오 노인코호트 연구(SALSA) 자료

미국 샌안토니오 노인코호트 연구(The San Antonio longitudinal study of aging: SALSA) 자료는 65세 이상의 멕시코계 미국인과 유럽계 미국인을 대상으로 하여 신체적인 허약 상태의 여부를 관측

하여 1(허약) 또는 0(비허약)으로 기록하였고, 이 이분형 자료는 연구 시작 시점에 관측한 값을 포함하여 총 네 번의 반복측정 결과로 얻어졌다. 각 연구 대상자의 개인적인 특성을 나타내는 성별, 나이, 인종 등을 포함하는 여러 가지 변수 그리고 노인들의 사망여부, 사망시점까지의 시간 즉 생존시간에 대한 정보를 포함하고 있는 자료이다. 분석에 사용되는 변수를 요약하여 자료를 정리하면 Table 1과 같다.

Table 1. SALSA data

ID	Sex	Age	Ethnicity	…	Time	Frailty	FU	Dead
1	1	70	0	…	0	0	3.8	1
1	1	70	0	…	1	.	3.8	1
1	1	70	0	…	2	.	3.8	1
1	1	70	0	…	3	.	3.8	1
2	0	71	1	…	0	0	10.3	0
2	0	71	1	…	1	0	10.3	0
2	0	71	1	…	2	.	10.3	0
2	0	71	1	…	3	.	10.3	0
⋮	⋮	⋮	⋮	…	⋮	⋮	⋮	⋮
736	0	69	0	…	2	0	7.9	0
736	0	69	0	…	3	0	7.9	0

SALSA 자료에서 ID는 환자의 고유번호를 나타내고 한 환자에 대해서 네 번 반복 측정하여 관측치를 얻은 자료이기 때문에 4개의 동일한 ID가 존재하게 된다. 성별을 나타내는 변수(sex)는 남자는 1, 여자는 0으로 정의되어 있고, 인종을 나타내는 변수(ethnicity)는 1은 멕시코계 미국인(Mexican American: MA), 0은 유럽계 미국인(European American: EA)을 의미한다. 노인의 신체적 허약 상태의 여부(frailty)는 시간의 흐름에 따라 연구시작 시점에서 관측된 값을 포함하여 총 네 시점에서 관측된 이분형 자료로 1은 그 노인은 신체적 허약 상태임을 의미하고 0은 비허약 상태라고 여길 수 있다. Frailty가 점으로 표시된 것은 결측을 의미한다. FU은 환자의 사망 또는 연구기간 동안 사망이 관측되지 않았다면 그 개체의 중도절단 시점까지의 시간을 나타낸다. 생존 자료에서의 사건(dead)은 관측 기간 동안 환자가 사망하면 1 그렇지 않은 경우 즉, 중도절단은 0으로 정의되어있다. 그 외에 교육(education)은 노인의 교육 기간을 의미하는 변수로 한 단위 증가한다는 것은 교육받은 기간이 1년 증가함을 뜻한다. 수입(income)은 한 달 동안의 가계 수입을 범주화하여 정의한 변수이고 범주의 수준이 증가한다는 것은 수입 수준이 높음을 의미한다. 당뇨여부(diabetes mellitus: DM)와 두 개 이상의 만성질환을 동시에 앓고 있는지의 여부(comorbidity)는 각각 해당하면 1, 그렇지 않으면 0으로 측정된 변수이다. 또한 정신상태 검사(mini-mental state examination)를 통해 측정된 노인의 인지장애 평가 변수인 MMSE는 값이 클수록 인지수준이 높다는 것을 의미하는 변수로 측정되었고, 노인 우울 척도(geriatric depression scale)로 측정된 우울정도를 나타내는 GDS는 값이 클수록 우울증상이 심하다는 것을 나타내는 변수로 정의되어 있다.

반복측정 자료에서는 결과 변수뿐만 아니라 그 외의 공변량에 대해서도 반복적으로 측정하여 시간이 흐름에 따라 변화하는 값을 갖는 경우가 있지만 본 논문에서 사용된 SALSA 자료는 신체적 허약 여부를 나타내는 반복 측정된 변수 frailty를 제외한 다른 변수는 연구 시작시점에서 측정된 값으로 모든 시점에서 일정한 값을 갖고 있다.

3.2. SALSA 자료 분석

기존의 SALSA 자료에 대한 반복측정자료 분석(Espinoza, Jung, Hazuda, 2010)과 생존자료 분석

(Espinoza, Jung, Hazuda, 2013) 연구의 결과가 논문으로 각각 발표되었다.

Espinoza, Jung, Hazuda(2010)의 논문은 65세 이상의 멕시코계 미국인과 유럽계 미국인을 대상으로 시간의 흐름에 따라 각 인종 간에 신체적 허약이 나타나는 정도의 차이가 있는지 알아보기 위한 연구의 결과이다. 나이, 성별, 사회·경제적 지위, 연구 시점 이전에 허약증세가 있었는지에 대한 전 허약증상(pre-frailty)여부, 당뇨여부, 두 가지 이상의 만성질환을 갖고 있는지의 여부 등을 독립변수로 설정하고, 연구 시작시점을 제외한 세 시점에서 관측된 신체적 허약 여부를 관측한 이분형 결과 자료를 종속변수로 하여 일반화추정방정식(generalized estimating equation)을 통해 각 변수에 대한 영향을 분석하였다. 멕시코계 미국인이 유럽계 미국인에 비해 신체적 허약 상태일 가능성이 유의하게 낮았고, 전 허약증상이 있었던 경우와 교육기간이 짧을수록 또는 수입이 적은 경우일수록 허약 상태로 관측될 확률이 높게 나타난다는 연구결과를 보여주고 있다.

Espinoza, Jung, Hazuda(2013)의 논문에서는 연구 대상자의 사망여부와 사망 시점 또는 중도절단 시점까지의 시간이 측정된 자료를 Cox의 비례위험 모형으로 구축한 연구의 결과를 제시하고 있다. 나이, 성별, 인종, 교육, 수입, 연구시작 시점에서의 허약 여부, 당뇨, 두 가지 이상의 만성질환을 동시에 갖는지의 여부, 인지수준(MMSE), 우울증상의 정도(GDS), 체질량 지수(BMI)를 공변량으로 하여 생존 모형을 구축하였다. 사망발생 위험에 대해서 인종간의 유의한 차이는 보이지 않았으며, 나이가 많은 경우, 성별이 남자인 경우에 사망 위험률이 유의하게 높았다. 또한 당뇨 질환자이거나 두 가지 이상의 만성질환을 갖는 노인들일수록, 우울증상 정도가 높은 수준으로 관측된 노인의 사망 위험률이 높았고 인지수준이 높을수록 위험률은 낮아진다는 결론을 얻었다.

하지만 노인의 허약 상태를 반복적으로 측정하는 과정에서 중도탈락이 발생하게 되고, 이로 인해 발생하는 결측값이 환자의 생존 자료와 관련되어 있을 것으로 예상된다. 이런 자료의 특성을 반영하기 위해 반복측정 자료를 일반화선형혼합모형으로 모형화하면서 각 대상자의 임의효과를 생존 모형과 공유시킴으로써 두 모형을 결합하여 두 결과변수를 동시에 분석해 보고자 한다. 3.2.1)절과 3.2.2)절에서는 결합분석에 앞서 각 결과변수에 대해 일반화선형혼합모형과 세 가지 형태의 생존분석 모형으로 분석하였고, 3.2.3)절에서 결합모형을 이용한 분석을 실시하였다.

1) 반복측정 자료 분석

기존의 세 시점의 반복측정 자료 분석을 통한 모형과는 다르게, 본 논문에서는 샌안토니오 노인 코호트 연구를 시작하는 시점에서 노인의 허약 상태를 측정한 변수를 반복측정 된 결과변수로 포함하여 총 네 번의 시점에서 관측된 결과에 대한 모형을 구축한다.

멕시코계 미국인(MA)과 유럽계 미국인(EA)을 구분 짓는 변수인 인종(ethnicity)과 나이(age), 성별(sex), 교육기간(education), 수입(income), 당뇨여부(DM), 두 가지 이상의 만성질환을 동시에 앓고 있는지의 여부(comorbidity)를 포함하는 노인의 정보들을 독립변수로 설정하였다. 추가적으로 각 시점 사이의 허약 상태의 변화보다 연구시작 시점과 나머지 시점 사이에서 노인들의 허약 상태가 변하는 점을 반영하기 위해 첫 번째 시점은 0, 나머지 시점을 1로 정의한 변수 T_ind를 공변량으로 추가하여 모형을 구축하였다. 절편에 대한 각 대상자의 임의효과 u_i는 측정된 독립변수 외에 측정되지 않은 다른 효과를 포함하며 대상자들 간의 변이가 존재한다는 것을 의미한다. 이들을 고려한 반복측정 자료에 대한 일반화선형혼합모형(GLMM)의 식을 나타내면 다음과 같다.

$$\begin{aligned} g(E(Frailty_{ij}|u_i)) = {} & \beta_o + \beta_1 Ethnicity_i + \beta_2 Age_i + \beta_3 Sex_i + \beta_4 Education_i \\ & + \beta_5 Income_i + \beta_6 DM_i + \beta_7 Comorbidity_i + \beta_8 T\_ind_i + u_i \end{aligned} \quad (2)$$

위 식에서의 연결함수 $g()$는 로짓함수로 정의할 수 있으며 따라서 아래와 같은 식을 통해 허약상태($Frailty$)에 대한 임의효과가 주어졌을 때의 조건부확률을 얻을 수 있고

$$p(Frailty_{ij}=1\mid u_i)=\frac{\exp(g(E(Frailty_{ij}\mid u_i)))}{1+\exp(g(E(Frailty_{ij}\mid u_i)))},$$

전체 자료를 이용하여 모형을 추정하기 위한 주변 우도함수를 표현하면 아래와 같다.

$$L=\prod_{i=1}^{n}\int_{u_i}\left\{\prod_{j=1}^{n_i}p(Frailty_{ij}=1\mid u_i)^{Frailty_{ij}}(1-p(Frailty_{ij}=1\mid u_i))^{1-Frailty_{ij}}\right\}du_i$$

임의효과 u_i에 대해 평균이 0 분산이 σ_u^2인 정규분포를 가정하고, 수치적분을 이용하여 근사하는 방법으로 우도함수를 구하여 모형의 회귀계수를 추정하게 되는데, SAS NLMIXED procedure를 사용하여 SALSA 자료의 반복측정 자료에 대한 모형을 추정하였고 결과는 Table 2와 같다.

Table 2. Parameter estimates from GLMM for frailty of SALSA data

Effect	Estimate	S.E.	Odds Ratio (95% CI)	p-value
Ethnicity (MA vs EA)	-1.133	0.372	0.32 (0.16-0.67)	0.002
Age (1-year increment)	0.095	0.044	1.10 (1.01-1.20)	0.031
Sex (male vs female)	0.702	0.315	2.02 (1.09-3.74)	0.026
Education (1-year increment)	-0.107	0.045	0.90 (0.82-0.98)	0.017
Income (1-category increment)	-0.263	0.062	0.77 (0.68-0.87)	<0.001
DM (diabetes mellitus)	1.242	0.351	3.46 (1.74-6.90)	<0.001
Comorbidity	0.584	0.295	1.79 (1.00-3.20)	0.048
T_ind	1.576	0.251	4.84 (2.95-7.92)	<0.001
σ_u^2	4.707	1.125		<0.001

Table 2의 내용을 보면 멕시코계 미국인(MA)이 유럽계 미국인(EA)에 비교하여 신체적 허약의 결과를 보일 확률이 유의하게 낮다는 것을 알 수 있다. 또한 나이가 많을수록, 남자인 경우일수록, 당뇨를 앓는 노인일수록, 만성질환을 동시에 두 가지 앓고 있는 노인일수록 허약할 확률이 높게 나타났다. 교육수준과 수입수준이 높은 노인일수록 신체적 허약의 가능성이 낮아진다는 결과를 보이고 있다. T_ind 변수에 대해서는 양의 방향으로 유의한 결과를 보이는데 이는 연구 시작 시점보다 그 뒤 시점에서 허약의 상태가 관측될 확률이 높다는 것을 의미한다. σ_u^2이 유의하므로 대상자들 간 변이가 큰 것으로 생각된다.

2) 생존 자료 분석

생존시간 자료에 대해 다양한 형태의 생존분석 모형을 통해 공변량에 대한 효과를 추정하고 각 모형의 결과를 비교해보고자 한다. 식 (1)에서 $h()$는 생존시간에 대한 위험함수로 Cox의 모형에서는 기저위험함수 $h_0()$에 특별한 가정이 필요하지 않고, 지수모형에서는 $h_0()$를 시간에 관계없이 일정한 값(λ)으로 가정하고, Weibull 모형에서는 두 모수 λ와 γ로 표현되는 시간의 함수 형태로 가정한다. η_i는 독립변수들의 선형조합을 나타내는 부분으로 생존 자료만을 분석할 때는 임의효과 u_i는 포함하지 않는다. 본 연구에서는 SALSA 자료에 대한 생존모형의 독립변수로 반복측정 자료 분석 시 사용한 변수 외에 인지수준(MMSE), 우울정도(GDS), 연구시작 시점에서의 허약여부

(frailty0)를 추가하고, T_ind는 포함하지 않았다. 또한 인종과 성별의 교호작용 효과가 유의함을 발견하여 포함하여 분석을 진행하였다. 이 독립변수들로 이루어진 선형 모형식 η_i를 나타내면 아래의 식과 같다.

$$\eta_i = \alpha_1 Ethnicity_i + \alpha_2 Age_i + \alpha_3 Sex_i + \alpha_4 Education_i + \alpha_5 Income_i + \alpha_6 Frailty_{0i} + \alpha_7 DM_i + \alpha_8 Comorbidity_i + \alpha_9 MMSE_i + \alpha_{10} GDS_i + \alpha_{11} Ethnicity_i * Sex_i$$

위 모형들은 SAS나 R 등으로 쉽게 추정할 수 있으며, 분석결과를 Table 3에 정리하였다. Weibull 모형에서의 기저위험함수는 시간에 따라 증가하는 추세로 추정되었고, 대략 6.8년 정도의 시점에서의 값이 지수모형에서의 기저위험 값과 비슷하다. 세 가지 모형에서 모두 교육수준, 연구시작 시점에서의 허약상태의 차이에 따른 사망 위험의 차이가 존재한다고 볼 수 없었다. 세 가지 모형에서 모두 나이가 증가하는 경우와 당뇨가 있는 경우, 두 가지 이상의 만성질환을 갖는 경우, 우울정도가 높은 수준의 노인일수록 사망 위험이 유의하게 높게 나타난다는 결과를 보였고, 그와 반대로 인지수준이 높을수록 사망 위험이 낮았다. 지수모형을 제외한 두 모형에서 수입수준이 높을수록 사망 위험이 낮았다. 인종과 성별의 교호작용 효과가 유의한데, 이것은 성별에 따른 차이가 인종별로 다르다는 것을 의미한다. 세 모형의 결과는 거의 유사한 것을 알 수 있다.

Table 3. Parameter estimates from the three survival analysis models for SALSA data

	Cox PH model		Exponential model		Weibull model	
	Estimate	p-value	Estimate	p-value	Estimate	p-value
Ethnicity	-0.504	0.069	-0.471	0.087	-0.493	0.075
Age	0.082	<0.001	0.079	0.001	0.082	<0.001
Sex	0.609	0.011	0.544	0.023	0.589	0.014
Education	0.038	0.127	0.033	0.182	0.036	0.147
Income	-0.065	0.045	-0.060	0.062	-0.064	0.047
Frailty at baseline	0.370	0.088	0.343	0.117	0.380	0.080
Diabetes	0.490	0.004	0.435	0.011	0.467	0.006
Comorbidity	0.382	0.015	0.358	0.022	0.379	0.016
MMSE	-0.075	0.006	-0.074	0.006	-0.074	0.006
GDS	0.038	0.012	0.033	0.027	0.037	0.014
Ethnicity*Sex	0.727	0.022	0.704	0.027	0.729	0.022
λ			0.00059	<0.001	0.00015	<0.001
γ					1.512	<0.001

3) SALSA 자료의 결합분석

SALSA 자료의 결합분석 모형 안에서 정의되는 반복측정 자료에 대한 모형식 부차모형 1에 대한 구조는 일반화선형혼합모형식 식 (2)와 같은 형태이고 이 모형에서의 임의효과 u_i를 생존분석 모형에서 공유하며 이 효과에 대한 계수 ψ를 추정하게 된다. 세 가지 결합분석 모형(Model 1, 2, 3)을 고려하였고, 각 모형 안에서 정의되는 생존모형 부분 부차모형 2는 Cox의 비례위험모형(Model 1), 지수회귀모형(Model 2), Weibull 모형(Model 3) 세 가지 형태의 모형으로 구축한다. 3.2.1절과 3.2.2절에서 사용한 독립변수를 그대로 사용하였는데, 부차모형 2에서 연구시작 시점에서의 허약여부(frailty0)는 포함하지 않았다. 부차모형 1에서 시작시점에서의 허약여부가 결과변수로 포함되어 있고, 임의효과를 통하여 부차모형 2와 연결되어 있기 때문이다. 부차모형 1, 2를 모형식으로 표현하면 다음과 같다.

· Submodel 1: $g(E(Frailty_{ij}\mid u_i)) = \beta_o + \beta_1 Ethnicity_i + \beta_2 Age_i + \beta_3 Sex_i + \beta_4 Education_i + \beta_5 Income_i + \beta_6 DM_i + \beta_7 Comorbidity_i + \beta_8 T\_ind_i + u_i$

· Submodel 2: $h(t\mid u_i) = h_0(t)\exp(\alpha_1 Ethnicity_i + \alpha_2 Age_i + \alpha_3 Sex_i + \alpha_4 Education_i + \alpha_5 Income_i + \alpha_6 DM_i + \alpha_7 Comorbidity_i + \alpha_8 MMSE_i + \alpha_9 GDS_i + \alpha_{10} Ethnicity_i * Sex_i + \psi u_i)$

이 결합모형은 2절에서 설명한 것처럼 u_i에 대해 정규분포를 가정하고 우도함수를 구성하여 최대우도추정법으로 회귀계수를 추정할 수 있다. 본 연구에서는 Model 1의 경우에는 Choi, Cai, Zeng, Olshan(2015)이 제시한 R 함수를 이용하였고, Model 2와 Model 3의 경우에는 SAS NLMIXED procedure를 이용하였다.

각 결합분석 모형의 추정 결과는 Table 4와 같다. Model 2와 3의 결과는 거의 유사하고 Model 1의 결과는 추정된 계수 값은 다른 모형과 조금 차이가 있지만, 계수의 방향이나 유의성에는 큰 차이가 없었다. 나이와 두 가지 이상의 만성질환을 동시에 앓고 있는지의 여부(comorbidity)의 회귀계수는 Model 2와 3에서는 유의하지 않았으나 Model 1에서는 유의한 결과를 보였고, Sex의 회귀계수는 Model 1에서만 유의하지 않았다. 주목할 만한 것은 임의효과의 계수 ψ의 추정치가 Model 1에서는 매우 유의하였고, 다른 모형에서는 유의수준 0.1 하에서는 유의하다고 볼 수 있다. 따라서 반복 측정된 노인들의 허약 상태를 나타내는 변수와 생존시간 간에 유의한 관련성이 존재한다고 생각된다.

Table 4. Parameter estimates from joint analysis models for SALSA data

		Model 1		Model 2		Model 3	
		Estimate	p-value	Estimate	p-value	Estimate	p-value
Submodel 1	Intercept	-7.759	0.009	-6.687	0.116	-6.786	0.110
	Ethnicity	-1.186	<0.001	-1.309	0.006	-1.308	0.006
	Age	0.122	0.003	0.086	0.140	0.088	0.131
	Sex	0.513	0.071	0.862	0.039	0.872	0.036
	Education	-0.114	0.006	-0.121	0.043	-0.122	0.042
	Income	-0.296	<0.001	-0.307	<0.001	-0.307	<0.001
	Diabetes	1.192	<0.001	1.627	0.001	1.621	0.001
	Comorbidity	1.090	<0.001	0.626	0.110	0.628	0.107
	T_ind	1.782	<0.001	2.020	<0.001	2.042	<0.001
Submodel 2	Ethnicity	-0.483	0.091	-0.384	0.153	-0.396	0.143
	Age	0.092	<0.001	0.092	<0.001	0.097	<0.001
	Sex	0.597	0.016	0.540	0.024	0.585	0.015
	Education	0.039	0.141	0.033	0.180	0.035	0.151
	Income	-0.096	0.005	-0.067	0.034	-0.072	0.023
	Diabetes	0.569	0.001	0.527	0.002	0.565	0.001
	Comorbidity	0.513	0.003	0.348	0.024	0.373	0.016
	MMSE	-0.071	0.009	-0.081	0.002	-0.082	0.002
	GDS	0.032	0.042	0.030	0.037	0.033	0.023
	Ethnicity*Sex	0.719	0.027	0.621	0.045	0.642	0.039
	ψ	0.244	0.006	0.057	0.100	0.065	0.093
	λ			0.00032	<0.001	0.00008	<0.001
	γ					1.466	<0.001

4. 결론 및 고찰

본 연구에서는 일반화선형혼합모형과 여러 형태의 생존모형을 임의효과를 공유하는 결합분석 모형으로 구축하여 미국 샌안토니오 노인코호트 연구(SALSA) 자료를 분석하였다. 연구 대상자들의 신체적 허약 상태를 반복 측정한 결과와 사망까지의 시간의 두 결과변수를 개별적으로 분석했을 경우와 비교하였다. 결합분석 모형의 결과는 두 결과변수의 관련성을 고려하지 않고 개별적으로 분석했던 결과와 약간의 차이를 보였다. 회귀계수가 다소 다른 값으로 추정되었으며 몇몇 공변량에 대해서는 통계적 유의성이 달라졌다. 일반화선형혼합모형으로 노인들의 허약 상태에 대해서만 분석했을 경우에는 고려한 모든 공변량이 유의한 영향을 주는 것으로 나타났으나, Cox의 비례위험모형을 통한 결합분석(Model 1)의 결과에서는 성별의 효과가 유의하지 않았고, 지수모형과 Weibull 모형을 통한 결합분석(Model 2와 3)의 결과에서는 나이와 두 가지 이상의 만성질환보유 여부의 효과가 유의하지 않았다. 생존시간에 대한 분석 결과는 결합분석 모형과 생존분석 모형만의 결과에 큰 차이는 없었다. 다만, 지수모형으로 생존분석을 실시한 경우 유의하지 않았던 수입변수가 결합분석 모형의 결과에서는 유의한 결과로 나타났다. 따라서 세 가지 결합분석 모형에서 모두 노인의 수입수준이 높을수록 사망 위험이 낮아진다는 결과를 얻을 수 있었다. 공유하고 있는 임의효과의 계수 ψ의 추정 결과는 세 가지 모형에서 모두 양의 방향으로 나타났다. 모수적 비례위험모형을 가정한 결합문석 모형(Model 2와 3)에서는 유의수준 0.05에서는 유의하지 않았으나, Cox의 비례위험모형을 이용한 결합분석 모형(Model 1)에서는 매우 유의한 결과를 보였다. 이러한 결과는 신체적 허약 상태로 관측된 노인일수록 사망 위험률이 높다는 것을 의미하는 것으로 해석할 수 있다.

Ibrahim, Chu, Chen(2010)의 연구에서 지적하였듯이 결합분석 모형에서 두 결과변수의 관련성을 나타내는 ψ의 값이 유의하게 추정되는 경우, 두 결과변수를 개별적으로 분석하게 되면 편향된 추정결과를 얻게 될 가능성이 있을 수 있다. 두 분석의 결과가 다른 경우 왜 다른지에 대한 좀 더 자세한 연구가 필요할 것으로 생각된다.

References

Cho, H. S. (2014). *Joint analysis models for longitudinal binary outcome and survival data,* Master's thesis, Graduate School, Yonsei University. (in Korean).

Choi, J., Cai, J., Zeng, D., Olshan, A. F. (2015). Joint analysis of survival time and longitudinal categorical outcomes, *Statistics in BioSciences,* 7, 19-47.

Espinoza, S. E., Jung, I., Hazuda, H. (2010). Lower frailty incidence in older Mexican Americans than in older European Americans: The San Antonio longitudinal study of aging, *The American Geriatrics Society,* 58, 2142-2148.

Espinoza, S. E., Jung, I., Hazuda, H. (2013). The hispanic paradox and predictors of mortality in an aging biethnic cohort of Mexican Americans and European Americans: The San Antonio longitudinal study of aging, *The American Geriatrics Society,* 61, 1522-1529.

Ibrahim, J. G., Chu, H., Chen, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data, *Journal of Clinical Oncology*, 28, 2796-2801.

Kim, H. (2005). A note about SAS for repeated measurements, *Journal of the Korean Data Analysis Society,* 7, 2067-2080. (in Korean).

Kim, S. C., Ahn, S. J. (2005). Fitting of a distribution of a grouped survival data, *Journal of the Korean Data Analysis Society,* 13, 2887-2900. (in Korean).

Kim, S. Y., Bae, J. S. (2006). Clinical trial data and survival analysis, *Journal of the Korean Data Analysis Society,* 8, 533-545. (in Korean).

Kwon, H. K., Cho, J. S. (2007). EEG 3-way repeated ANOVA of prefrontal lobe of left and right brain which influences brain activity by the science learning types, *Journal of the Korean Data Analysis Society*, 9, 1107-1118. (in Korean).

Tsiatis, A. A., Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview, *Statistica Sinica,* 14, 809-834.

Tsiatis, A. A., DeGruttola, V., Wulfsohn, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error: Applications to survival and CD4 counts in patients with AIDS, *Journal of American Statistical Association*, 90, 27-37.

Vonesh, E. F., Greene, T., Schluchter, M. D. (2006). Shared parameter models for the joint analysis of longitudinal data and event times, *Statistics in Medicine*, 25, 143-163.

Wang, Y., Taylor, J. M. G. (2001). Jointly modelling longitudinal and event time data, with applications to AIDS studies, *Journal of American Statistical Association,* 96, 895-905.

Wulfsohn, M. S., Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error, *Biometrics,* 53, 330-339.

Joint Analysis Models for Longitudinal Binary Outcome and Survival Data*

Hyesoo Cho[1], *Inkyung Jung*[2]

Abstract

Joint analysis models, generally in a shared random effects model form, are used to simultaneously model two or more types of outcomes such as longitudinal outcomes and survival time data. Most research on joint analysis models deals with continuous longitudinal outcomes and research dealing with binary outcomes is scarce. In this study, we formulate joint analysis models for longitudinal binary outcomes and survival time data using a generalized linear mixed model and a survival model sharing a random effect for correlation between the two outcomes. We analyzed the San Antonio longitudinal study of aging (SALSA) data using the joint analysis models to simultaneously model the longitudinal frailty status of the elderly over four time points and their survival time. The joint analysis results were somewhat different from those from separate analyses on two outcome variables. Through the coefficient of shared random effect, we explained the association between the two outcome variables.

Keywords : Joint modeling, shared random effects, generalized linear mixed model, proportional hazards model.

*This study was supported by a faculty research grant of Yonsei University College of Medicine for 2013 (6-2013-0139).

*The data used in this study was provided by Dr. Helen P. Hazuda, the principal investigator of San Antonio Longitudinal Study of Aging (SALSA).

[1]Clinical Trials Center, Severance Hospital, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 120-752, Korea.
E-mail : chsoo@yuhs.ac

[2](Corresponding Author) Associate Professor, Department of Biostatistics and Medical Informatics, College of Medicine, Yonsei University, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 120-752, Korea.
E-mail : ijung@yuhs.ac

[Received 31 March 2015; Revised 19 May 2015, 11 June 2015; Accepted 14 June 2015]