신뢰성응용연구
제7권, 제3호, pp.225-237, 2006

연구논문

QoL에 의한 정보형 중도탈락의 모형화

이 기 훈

전주대학교 경영학부

Modelling the Informative Dropouts with QoL

Ki Hoon Lee

School of Business, Jeonju University

Abstract

This paper proposes a method of modelling the informative dropouts with QoL(quality of life) in survival analysis. QoL is the index to measure the health related quality of life of a patient who got some treatments for a disease. Dropouts are prevalent occurrences on longitudinal study. They are commonly dependent to the QoL of patients, that is, severe disease or death and called informative dropouts. Modelling the mechanism of dropouts could achieve the more accurate inference for survival analysis. A likelihood method is proposed to estimate the survival parameter and test the patterns of dropouts.

1. 서론

삶의 질(Quality of life; QoL)은 의료분야에서 질병치료를 위해 어떤 처치를 받고 있는 환자들의 건강상태를 측정하는 지표(index)이다. 특히 만성병(예: 당뇨병, 관절염, AIDS)의 임상시험(clinical trials)에서 중요한 평가지표로 인식되어 왔고, 생존(survival)이 주요 측정변수(endpoint)인 질병에서도 생명의 양과 더불어 생명의 질을 중요시하는 경향에 따라 QoL 측정의 중요성은 점차 증가되고 있다. 삶의 질이란 개인의 삶에 영향을 주는 여러 가지 요인들을 망라하는 개념이지만 의학적인 차원에서 보면 건강에 관련된 삶의 질(Health related Quality of life; HRQoL)로 축소된다. 건강관련 요인들에 대한 보편적인 변수에 대한 동의는 존재하지 않지만 WHO가 건강을 '단순히 병이나 쇠약함이 없는 것을 의미하는 것이 아니라 신체적, 정신적 그리고 사회적으로 완전한 웰빙 상태를 의미한다' 라고 정의하고 있기 때문에 HRQoL도 이러한 문맥에서 고려되고 있다. 예를 들어 Schumacher et al(1991)[1]은 삶의 질을 병의 증세 정도와 치료의 부작용(예: 구역질, 통증, 불면, 소화불량 등), 신체적·기능적 상태(예: 동작가능성, 자립도, 피로감 등), 정서적인 상태(예: 불안감, 우울증, 만족감 등) 그리고 사회적 기능상태(예: 가족간 교류, 일/여가, 교우관계 등) 등의 요인(factor)으로 나누어 측정하고 있다. QoL을 측정하기 위해 여러 종류의 도구가 사용되어지는데 구체적으로 Rotterdam 증상 체크리스트, Nottingham 건강 프로파일, Sickness Impact 프로파일, Hospital 불안감 및 우울 척도, Short Form(SF) 36 등이 있다.[2]

한 개체에서 일정한 시점에 따라 반복적으로 자료를 얻는 경시적 연구(longitudinal study)에서는 환자들의 이탈로 인한 중도탈락 또는 결측이 자주 발생한다. 어떤 시점 t에서 결측이 생겼을 때 Rubin(1976)[3]과 Little과 Rubin(1987)[4]은 결측을 다음과 같이 세가지 종류로 분류하였다.

1) 완전무작위결측(MACR : missing completely at random): t시점에서 결측은 전에 관측된 값들과 비관측된 t시점의 값과 무관하다.

2) 무작위결측 (MAR : missing at random) : t시점에서의 결측은 전에 관측된 값들과는 관계가 있으나 비관측된 t시점의 값과는 무관하다.

3) 무시할 수 없는 비 응답(MNI : non-ignorable non-response) : t시점에서의 결측은 전의 관측값들과 관계가 있고 t시점에서의 비관측값과 관련될 가능성이 있다.

Diggle과 Kenward(1994)[5]는 이를 중도탈락에 적용하여 중도탈락을 각각 완전무작위 중도탈락(CRD: completely random drop-out), 비관측값과 독립이지만 기존의 값들에는 종

1) Schumacher, M, Olschewski, M, Schulgen, G. (1991). Assessment of Quality of Life in Clinical Trials. *Stat Med*, Vol. 10, 1915-1930.
2) Bowling, A. (1991) *Measuring Health: A review of Quality of Life Measurement Scales.* Buckingham: Open University Press.
3) Rubin, D. B. (1976). Inference and Missing Data, *Biometrika*, Vol. 63, 581-592.
4) Little, R. J. A. , Rubin, D. B. (1987) *Statistical Analysis with Missing Data.* New York, John Wiley.
5) Diggle, P. and Kenward M. G. (1994). Informative Drop-out in Longitudinal Data Analysis (with discussion). *Applied Statistics*, Vol. 43, 49-94.

속된 무작위 중도탈락(RD: random drop-out), 기존의 값들과 비관측값 모두에 종속된 정보형 중도탈락(ID: informative drop-out) 등의 세가지로 분류하였다.

중도탈락(drop-out)은 환자가 완전히 실험에서 이탈하는 것으로 여러 이유로 발생하는데, 주로 사망, 질병, 치료중단, 치료효과미비, 추적실패, 연구종료 등이다. 생존분석의 측정변수가 되는 질병이나 사망이 중도탈락의 주요 원인이라면 이러한 중도탈락은 무작위가 아닌 정보형이라 할 수 있다. 또한 질병이나 사망은 환자의 QoL과 밀접한 관계를 갖고 있기 때문에 중도탈락이 비관측된 값(질병의 발생, 악화, 사망)과 연관이 있는 정보형 중도탈락에서는 QoL 정보를 적극 이용하면 보다 합리적인 생존분석이 가능하다.

이러한 결측 패턴을 모형에 포함하지 않고 단순한 대체(imputation)에 의한 분석을 할 경우 추론에 편의(bias)가 존재하고 검정력에 손실이 있음은 Zwinderman(1992)[6], Roy와 Lin(2005)[7] 등에 의해서 언급되었다.

결측값이 완전 무작위일 때는 일반적인 통계방법을 사용하여도 무방하고, 무작위(MAR)일 때도 Murray와 Findlay(1988)[8]에 의하면 우도함수(likelihood function)에 기초하지 않은 방법은 편의가 존재하지만 우도함수에 기초한 분석법을 사용한다면 결합우도함수가 완전관측자료와 결측과정 등의 두 부분의 곱으로 인수분해할 수 있기 때문에 결측과정이 무시될 수 있다고 한다. 본 논문에서는 완전관측자료와 결측과정이 독립적이지 않은 ID인 경우 우도함수를 이용한 분석방법을 고찰하고자 한다.

2장에서는 기존의 QoL을 고려한 생존분석법을 소개하고 3장에서는 우도함수를 이용하여 중도탈락을 모형화한 추론방법을 제안한다. 4장에서는 결론과 향후연구과제에 관하여 언급하겠다.

2. QoL을 고려한 기존의 분석방법

정보형 중도탈락이 존재하는 자료분석시 편의를 줄이기 위해 생존분석에서 생존(survival)과 QoL이 동시에 고려되는 측정변수가 되어야 하는데 이러한 분석방법은 다음과 같은 3가지 유형으로 나누어 볼 수 있다.

1. 정보형 중도탈락을 가정한 QoL 생존분석
2. QoL이 적용된 생존분석
3. QoL 분석과 생존분석의 동시수행

그러나 1번을 제외한 2, 3번의 방법은 정보형 중도탈락을 모형에 넣어주는 직접적인 방법

---

6) Zwinderman, A. H. (1992). Statistical Analysis of Longitudinal Quality of Life Data with Missing Measurements. Qual Life Res, Vol. 1, 219-224.

7) Roy, J. and Lin, X. (2005). Missing Covariates in Longitudinal Data with Informative Dropouts: Bias Analysis and Inference. *Biometrics*, Vol. 61, 837-846.

8) Murray, G. D., Findlay, J. G. (1988) Correcting for the Bias Caused by Drop-outs in Hypertension Trials. *Statistics in Medicine*. Vol. 7, 941-946.

이 아닌 QoL 정보와 생존을 결합하여 편의를 줄이는 간접적인 방법으로, 정보형 중도탈락을 무시하고 생존분석하는 것보다는 정확성이 높다는 이점뿐이었다. 그러나 1번의 방법은 모형화가 복잡하여 실제 사용에 어렵고, 기본적으로 중도탈락의 형태를 검정하는 절차가 우선 필요하다는 문제점을 갖고 있다. 본 장에서는 2, 3번의 기존방법을 소개하고 3장에서는 1번 방법에 의한 모형화와 중도탈락 검정이 어떻게 동시에 가능한지를 설명하겠다.

2.1. QoL이 적용된 생존분석

생존분석에서 환자관련 공변량(covariate)을 포함하여 생존함수를 모형화할 때, QoL이 한 변량값으로 모형에 포함될 수 있다. 공변량을 포함한 위험함수는 Cox[9]에 의해 제안되었고 비례위험모형(PHM: Proportional Hazard Model)이라 알려져 있다. 공변량은 연속형/순서형/이항형 등이 가능한데 연구기간 동안 고정(예: 성별)되는 것을 고정형 또는 시간비의존형이라 하고 시간에 따라 변화하는 공변량을 시간의존형이라한다. 두 가지 모두 Cox의 비례위험모형에 의해 분석할 수 있으며 고정형인 경우는 층화생존분석을 이용할 수 있다.

고정형 공변량에 대한 Cox의 PHM 위험함수는 다음과 같다.

$$h(t) = h_0(t)e^{\boldsymbol{\beta}'\boldsymbol{x}} \tag{2.1}$$

여기서 $h_0(t)$는 기저위험함수(base hazard function), $\boldsymbol{x} = (x_1, x_2, \cdots, x_m)'$는 고정형 공변량, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)'$는 회귀계수이며 편우도함수에 의해 추정된다.

이 모형은 생존시간의 확률분포에 대한 어떤 특정한 분포를 가정하지 않는다. 그러므로 기저위험률을 임의로 가정할 수 있다. 그러나 위험률의 비례성을 가정하기 때문에 준모수방법(semi-parametric)으로 불리고 있다. 비례성의 가정이란 위험비율(hazard ratio)이 다음과 같이 시간에 관계없이 상수가 된다는 것이다.

$$\frac{h(t)}{h_0(t)} = e^{\boldsymbol{\beta}'\boldsymbol{x}},$$

여기서 위험비율값이 1이면 효과가 존재하지 않는 것이고 1 미만이면 생존확률이 늘어나고, 1 초과이면 생존확률이 줄어드는 것을 의미한다.

비례위험성의 가정도 생존함수의 모형검정 때와 같이 대수누적위험도표로 검정할 수 있다. 각 그룹별로 Kaplan-Meier 방법에 의해 $S(t)$를 추정하고 $\log[-\log S(t)]$와 $\log t$를 도표를 그렸을 때 비례성 가정이 만족되면 이들 도표가 평행이 될 것이다.

공변량 변수의 값이 시간에 따라 변할 때, 시간의존형 공변량 Cox 모형(Cox model with time-dependent covariates)은 다음과 같은 모형을 채택한다.

$$h(t) = h_0(t)e^{\boldsymbol{\beta}'\boldsymbol{x} + \boldsymbol{\delta}'\boldsymbol{z}(t)},$$

여기서 $\boldsymbol{x}$와 $\boldsymbol{\beta}$는 고정된 공변량과 그 회귀계수이고 $\boldsymbol{z}(t)$와 $\boldsymbol{\delta}$는 시간의존형 공변량과 그 회

9) Cox, D. R. (1972), Regression Models and Life Tables, *Journal of the Royal Statistical Society* B, Vol. 34, 187-220.

귀계수이다.

QoL은 시간에 따라 달라지는 값이므로 이를 PHM 모형에 넣고자 할 때 시간의존형 Cox 모형을 사용하여야 한다. 이는 몇 가지 가정이 필요하고 PHM의 장점을 더 이상 보유하지 않지만, 실제 자료분석은 S-PLUS 통계패키지를 이용할 수 있기 때문에 사용상 큰 문제는 없다. 그러나 이러한 분석법에서는 중도탈락한 개체의 생존시간자료를 사용할 수 없으므로 QoL이 단지 공변량으로 사용된 것 이외에 정보형 중도탈락의 정보를 이용할 수는 없다.

2.2 QoL과 생존분석의 동시수행

임상실험에서 생존시간은 늘이지만 부작용 등으로 그 것의 가치가 평가절하될 때나 생존기간에는 별 차이가 없지만 증세의 완화라든가 중독성의 감소 등으로 삶의 질을 높였을 때 이에 대한 평가가 새롭게 이루어져야 한다. 그러나 삶의 질과 생존시간이 각각의 끝점으로 분리되어 분석된다면 이들 간의 조화로 최적의 치료법을 찾기가 어려워진다. 그래서 두 변수를 동시에 고려하는 방법이 제안되어져 왔고 그 종류는 다음과 같다. 첫째 삶의 질과 양을 하나의 측정변수를 하여 질에 의해 조정된 생존분석을 하는 QAL(Quality Adjusted Life) 방법, 둘째, 환자가 삶의 질을 고려한 여러 가지 건강상태(사망 포함)로 이동하는 형태를 모형화한 다단계 모형(multistate model)을 이용해 처리간 차이를 보는 방법, 마지막으로 두변수를 동시에 이루어지는 과정으로 보아 연계모형으로 고려하는 동시모형(Simultaneous Modelling) 방법 등이다.

생존이 주된 끝점인 경우에도 치료에 부작용이 많은 경우 통증, 만족도 등과 같은 부차적인 끝점(endpoint)에 관심을 갖게 된다. 이때 QAL, 즉 QoL로 조절된 생존분석이 하나의 대안이 될 수 있다. 생존에 관하여 처리간에 차이를 분석할 때 생존에 영향을 주는 환자관련 요인을 조절하는 단계가 필요하다. 이런 요인은 주로 나이, 성별, 백혈구수와 같은 생리학적 변수, 종양의 크기와 병 관련 직접적인 변수, 연구에 포함되었을 때 병세의 진척정도 등이 될 수 있다. 이러한 변수는 QoL의 공변량으로 간주될 수 있고 이들은 기저측도(baseline measure)가 될 수 있고, 시간에 따라 변하는 변량이 될 수도 있다.

QALY(Quality Adjusted Life Years)는 Fanshel과 Bush[10]가 제안한 방법으로 생존기간을 QoL에 따라 절감(down weighting)하는 방법이다. 이는 경제학의 비용-효용함수 개념을 차용한 것으로, 삶의 질을 고려한 생존기간 QALY은 다음과 같이 여러 건강상태에서 머무른 시간의 가중평균값이다. 일반적으로 건강상태는 $H_1, H_2, \ldots, H_m$로 표시하며 H_m은 사망상태를 의미한다.

$$\text{QALY} = \sum_{i=1}^{m} u_i t_i \ ,$$

여기서 $u_i (i = 1, \ldots, m)$는 $H_i (i = 1, \ldots, m)$상태에서의 효용이고 H_i상태에서 머무른 시간을 t_i라 한다. 그러나 여기서 QoL의 효용을 양적으로 측정할 수 있는가에 관한 문제와 설사 가능하

10) Fanshel, S., Bush, J. W.(1970) A Health-status Index and Its Applications to Health Services Outcomes. *Operat Res,* Vol. 18, 1021-1066.

더라도 질로 생존기간을 조절하는 방법이 다양하여 최적의 방법을 찾기 어렵다는 문제점을 안고 있다.

이를 개선하기 위해 제안된 Q-TWiST(Quality-adjusted Time Without Symptom of disease or Toxity of treatment)는 유방암치료효과 분석을 위하여 Goldhirsch et al[11] 이 제안한 방법으로 Glasziou et al[12]이 이론적으로 정립하였다.

Q-TWiST 분석을 위한 첫단계는 비교되는 처리간에 차이를 밝힐 수 있는 의학적으로 의미있는 건강상태를 다음과 같이 정의하는 것이다.

- TOX : 중독성 부작용(Toxic side-effects)을 갖는 시간
- TWiST : 증상이나 부작용이 없는 시간
- REL (또는 PROG) : 병의 재발이후의 시간(즉 다시 증상/부작용을 갖는 시간)

각 처리그룹마다 Q-TWiST는 다음과 같이 계산된다.

$$Q-TWiST = u_t TOX + TWiST + u_r,$$

여기서 u_t, u_r 는 각각 그 대응하는 기간의 효용을 의미하는데, 식에서 보는 바와 같이 TWiST는 효용 1을 갖고 TOX와 REL은 0과 1사이의 효용값을 갖는다.

그러나 Q-TWiST분석도 중도탈락이 완전무작위가 아닐 때, Kaplan-Meier 통계량이 편의되어 있다. 이에 대한 해법으로 분할 생존분석이 제안되는데, KM 도표에서 각 건강상태에서 머무른 평균시간을 구하고 이를 각 그룹마다 Q-TWiST모형에 따라 가중평균을 구하게 된다. 각 건강상태에 머무른 시간에 따라 그룹수준에 따라 개별수준이 아니기 때문에 절단된 생존시간에 가중값을 줄 필요가 없으므로 정보형 절단에 따른 문제점을 극복할 수 있다.

다단계(Multistate) 생존분석은 Fix와 Neyman(1951)[13]이 의료연구에서 제안하였는데 상태간의 전이확률(transition probability)과 각 상태에서 머문 시간을 마아코프(Markov) 모형을 통해 분석한 것이다. 이론적인 많은 부분은 계수과정(counting process) 틀에서 고려되고, 상태간 이동은 조건부확률 또는 전이확률로 묘사되며 개체가 어떤 QoL상태로 이동하는 것을 확률과정으로 모형화 하였다. 다단계 생존분석은 어떤 상태로든 이동이 가능하기 때문에 단일방향의 건강상태전이(T→W→R)만을 고려하는 Q-TWiST 방법의 제한점을 극복할 수 있다. 이러한 모형은 특히 당뇨병, 간이식, 심장이식, 유방암, 전립선암 등과 같이 상태가 호전될 수 있는 질병에서 유용하게 사용된다.

건강 전이 확률은 개체가 일정기간동안 한 상태에서 다른 상태로 전이할 가능성을 의미하므로 이는 Markov 확률과정으로 표현할 수 있다. 만약 전이 확률이 어떤 상태에 머무른 체류시간(sojourn time)에 영향을 받으면 semi-Markov 과정이라 한다.

---

11) Goldhirsch, A,, Gelber, R. D, Simes, R. J,, Glasziou, P,, Coates, A. S. (1989). Costs and Benefits of Adjuvant Therapy in Breast Cancer: a Quality-Adjusted Survival Analysis. *Journal of Clinical Oncology*, Vol. 7, 36-44.

12) Glasziou, P. P,, Simes, R. J,, Gelber, R. D. (1990). Quality Adjusted Survival analysis. *Statistics in Medicine*, Vol. 9, 1259-1276.

13) Fix, E,, Neyman, J. A. (1951) Simple Stochastic Model of Recovery, Relapse and Loss of Patients. *Hum Biol*, Vol. 23, 205-241.

표준 생존분석에서 생명에서 사망으로의 전이확률은 Cox의 회귀모형으로 표시될 수 있는데 이 개념을 다단계 생존분석에서 차용할 수 있다. 즉, t 시점에 상태 i 에서 상태 j 로 전이할 확률, $\lambda_{ij}(t)$ 를 Cox 회귀모형을 이용해 다음과 같이 표현할 수 있다.

$$\lambda_{ij}(t) = \lambda_{0ij}(t) e^{\boldsymbol{\beta}_{ij}' \boldsymbol{x}_{ij}} \; ; \; i, j = 1,...,m \, ,$$

여기서 $\lambda_{0ij}(t)$ 는 기저 전이확률이고 $\boldsymbol{\beta}_{ij}$, $\boldsymbol{x}_{ij}$ 는 공변량과 그에 대응하는 회귀계수 벡터이다. 이때 공변량을 시간종속형 $\boldsymbol{x}_{ij}(t)$ 으로 일반화할 수 있다.

실제로 한 개체가 어떤 상태에서 다른 상태로 전이할 때까지의 정확한 시간을 측정하기는 힘들다. 단지 관찰시점에 상태가 변화함을 인지할 수 있을 뿐이다. 이런 상황에서는 전이시간이 절단된 경우이고 위에서 설명한 표준적인 방법으로 모형화하기는 어렵다. 그래서 가장 많이 사용되는 방법이 관찰시점으로 근사하거나 관찰시점의 중간시점으로 근사화 시키는 것이다.

삶의 질과 생존기간이 모두 중요한 끝점일 때 환자에게 두 사건, 즉 삶의 질 상태의 변화와 사망시간은 동시에 일어나는 과정으로 간주된다. 동시모형화는 삶의 질을 시간에 따른 함수로 추정하고 이를 생존함수에 적용하는 방법으로, 생존자료가 QoL을 추정하는데 적용되기 때문에 사망에 의한 중도탈락 형태의 정보형 중도탈락을 조절할 수 있다. 또한 시간종속형 공변량을 가진 PHM 분석의 성능을 개선할 수 있는데, 왜냐하면 모든 공변량 값이 시간에 종속된 QoL 모형으로부터 추정되기 때문이다.

이를 처음 제안한 Faucett과 Thomas(1996)[14]는 Gibbs sampling을 사용하여 공변량 값과 생존시간들을 모형화하였다. 그들은 공변량들은 다음과 같이 변량효과모형을 가정하고 사망시간은 비례위험모형을 가정하였다. 시간 t_{ij} 에서 $i\,(i=1,...,n)$ 개체, $j\,(j=1,...,c)$ 번째 변수의 연속적인 시간종속 공변량이 z_{ij} 이라 하면 다음과 같이 표현할 수 있다.

$$z_{ij} = \boldsymbol{x}_i(t_{ij}) + \varepsilon_{ij} \, ,$$

여기서 $\boldsymbol{x}_i(t_{ij})$ 는 t_{ij} 시점에서 관측되지 않은 실제값, 즉 $E(z_{ij})$ 이며, ε_{ij} 는 독립이며 정규성을 가정한다. $\boldsymbol{x}_i(t_{ij})$ 를 단변량, 단순회귀모형으로 가정하면

$$x_i(t) = \alpha_i + \beta_i t \, ,$$

여기서 변량효과 α_i 와 β_i 는 이변량 정규분포를 가정한다. 동시에 위험률은 PHM에서와 마찬가지로 다음과 같이 모형화 한다.

$$\lambda_i(t) = \lambda_0(t) \, e^{\gamma \boldsymbol{x}_i(t)} \, ,$$

여기서 $\lambda_0(t)$ 는 기저위험함수이며, γ 는 회귀계수이다.

이렇듯 QoL과 생존 두 변수를 동시에 고려하는 분석법에서도 QoL의 정보만 이용할 뿐이지 중도탈락한 개체의 생존정보를 얻어내는 데는 한계가 있다.

---

14) Faucett, C. L., Thomas, D. C. (1996) Simultaneously Modelling Censored Survival Data and Repeatedly Measured Covariates; a Gibbs Sampling Approach, *Statistics in Medicine*, Vol. 15, 1633-1685.

3. QoL과 중도탈락을 모형화한 분석방법

중도탈락이 발생하면 생존 자료들을 모두 얻을 수 있지 않으므로 관측가능한 실제 자료값과 관측은 안되는 원자료값을 다음과 같이 분리하여 정의한다. 임상시험에서 정기적으로 QoL을 측정하고 최종적으로 생존시간을 얻었다고 가정할 때 원래 변수값은 다음과 같이 생존시간과 QoL 측정값의 결합 벡터 형태가 된다.

$$\begin{aligned}\boldsymbol{y}_1 &= (t_1,\ Q_1(1),\ldots,Q_1(t_1))'\\ \boldsymbol{y}_2 &= (t_2,\ Q_2(1),\ldots,Q_2(t_2))'\\ &\vdots\\ \boldsymbol{y}_n &= (t_n,\ Q_n(1),\ldots,Q_n(t_n))',\end{aligned}$$

여기서 사망시점의 QoL은 0이 된다. 즉, $Q_i(t_i)=0\,(i=1,\ldots,n)$

만약 이들 중 k번째 개체가 중도탈락으로 인해 생존시간 측정이 안되고 단지 d_{k-1}시점까지 얻은 QoL만 측정가능하다면 그때 관측값은 다음과 같이 정의한다.

$$\boldsymbol{x}_k = \boldsymbol{y}_k{}^* = (NA,\ Q_1(1),\ldots,Q_1(t_{k-1}),NA,\ldots,NA)'$$

본 논문에서는 편의상 n개체 중에 처음 l개는 중도탈락이 없고 나머지는 중도탈락이 있다고 가정하면 관측값을 다음과 같이 표현할 수 있다.

$$\boldsymbol{x}_i = \begin{cases}\boldsymbol{y}_i, & i=1,\ldots,l\\ \boldsymbol{y}_i{}^*, & i=l+1,\ldots,n.\end{cases}$$

관측값들의 우도함수(likelihood function)를 구하기 위해 우선 중도탈락이 없는 개체들의 우도함수를 다음과 같이 표시한다.

$$f(\boldsymbol{x}_k) = f(\boldsymbol{y}_k)\prod_{i=1}^{t_k}(1-p_i);\ k=1,\ldots,l,$$

여기서 $p_k(1\leq k<n)$는 k시점에 개체들이 중도탈락할 확률이다. 중도탈락이 발생하는 일반적인 경우에 우도함수는 다음과 같다.

$$f(\boldsymbol{x}_k) = f(\boldsymbol{y}_k{}^*)\prod_{i=1}^{d_{k-1}}(1-p_i)\,P^*(D=d_k);\ k=l+1,\ldots,n,$$

여기서 $P^*(D=d_k)$는 d시점에 처음으로 중도탈락이 발생할 확률로서 p_k와는 다르게 $(Q_k(1),\ Q_k(2),\ldots,\ Q_k(d-1),\ NA)$에 의해 계산되어지는데 중도탈락이 MCAR, MAR인 경우에는 p_k와 같다

그러므로 관측값들의 우도함수는 다음과 같다.

$$f(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) = \prod_{i=1}^{l} f(\boldsymbol{y}_i)\prod_{j=1}^{t_i}(1-p_j)\times\prod_{i=l+1}^{n} f(\boldsymbol{y}_i{}^*)\prod_{j=1}^{d_i-1}(1-p_j)P^*(D=,d_i)$$

이때 로그우도함수는 다음과 같다.

$$\begin{aligned}\log f(\boldsymbol{x_1},\ldots,\boldsymbol{x_n}) &= \sum_{i=1}^{l}\log f(\boldsymbol{y_i}) + \sum_{i=1}^{l}\sum_{j=1}^{t_i}\log(1-p_j) \\ &\quad + \sum_{i=l+1}^{n}\log f(\boldsymbol{y_i}^*) + \sum_{i=l+1}^{n}\sum_{j=1}^{d_i-1}\log(1-p_j) + \sum_{i=l+1}^{n}P^*(D=d_i) \qquad (3.1)\\ &= (1)+(2)+(3)+(4)+(5).\end{aligned}$$

(1)을 정리하기 위해 위험함수를 (2.1)에서와 같이 가정한다.

$$h(t)=h_0(t)e^{\boldsymbol{\beta}'\boldsymbol{Q}(t)},$$

여기서 $h_0(t)$는 기저위험함수, $\boldsymbol{Q}(t)=(Q(1),Q(2),\ldots,Q(t))'$는 QoL 측도값으로 고정형 공변량, $\boldsymbol{\beta}=(\beta_1,\beta_2,\ldots,\beta_t)$는 회귀계수이며 편우도함수에 의해 추정된다. 그러므로 중도탈락이 없는 경우의 확률함수는 다음과 같다.

$$f(t)=\lambda e^{\boldsymbol{\beta}'\boldsymbol{Q}(t)}e^{-\int_0^t \lambda e^{\boldsymbol{\beta}\boldsymbol{Q}(u)}du}.$$

그래서

$$(1)=\sum_{i=1}^{l}\log f(\boldsymbol{y}_i)=l\log\lambda+\sum_{i=1}^{l}\boldsymbol{\beta}'\boldsymbol{Q}_i(t_i)-\sum_{i=1}^{l}\int_0^{t_i}\lambda e^{\boldsymbol{\beta}'\boldsymbol{Q}_i(u_i)}du_i$$

(2)식을 정리하기 위하여 중도탈락에 관한 로짓(logit) 확률을 다음과 같이 표현할 수 있는데, 이는 중도탈락의 확률이 QoL의 함수로 표시할 수 있다는 가정아래 중도탈락의 형태를 검정할 수 있는 근거를 마련해준다.

$$\begin{aligned}\mathrm{logit}(p_k) &= b_{k0}+b_1Q(t_k)+b_2Q(t_k-1)+\cdots+b_{t_k}Q(1)\\ &= b_{k0}+\sum_{m=1}^{t_k}b_m\,Q(t_k+1-m),\end{aligned}$$

여기서 $(b_1,b_2,\ldots,b_t)$는 로짓확률의 계수이다. 그러므로

$$(2)=-\sum_{i=1}^{l}\sum_{j=1}^{t_i}\log[1+\exp(b_{j0}+\sum_{m=1}^{t_j}b_m\,Q_j(t_j+1-m)]$$

(3)에는 최종사망시간이 관측되지 못하고 관측직전의 QoL 측도값만 존재하므로 다음과 같이 표현한다.

$$(3)=\sum_{i=l+1}^{n}\log f(\boldsymbol{y}_i^*)=\sum_{i=l+1}^{n}\boldsymbol{\beta}'\boldsymbol{Q}_i(t_i)$$

또한 (4)는 중도탈락이 발생한 개체가 $d_i(i=l+1,\ldots,n)$ 전까지 관측이 되는 확률이다.

$$(4)=-\sum_{i=l+1}^{n}\sum_{j=1}^{d_i-1}\log[1+\exp(b_{j0}+\sum_{m=1}^{d_i-1}b_m\,Q_j(d_i-m)]$$

그리고 (5)는 다음과 같이 정리된다.

$$(5) = \sum_{i=l+1}^{n} \log P^*(D=d_i)$$
$$= \sum_{i=l+1}^{n} (b_{d_i 0} + \sum_{m=1}^{d_i} b_m Q_i(d_i+1-m) - \log[1+\exp(b_{d_i 0} + \sum_{m=1}^{d_i} b_m Q_i(d_i+1-m)]).$$

(5)에서는 중도탈락 시점 d에서의 $Q(d)$가 포함되어 있는데 우리는 이를 추정하여 대치함으로써 CRD, RD, ID에 대한 검정이 가능하도록 하였다. 이를 위하여 QoL벡터가 정규분포를 따른다고 가정한다.

$$\boldsymbol{Q}(t) = (Q(1), Q(2), \ldots, Q(t))' \sim N(\boldsymbol{\mu}(\lambda), \Sigma),$$

여기서 $\boldsymbol{\mu}(\lambda) = (\mu_1, \mu_2, \ldots, \mu_t)'$, Σ는 $(t \times t)$ 공분산 행렬이다.

$\boldsymbol{x}^{(d)} = (x_1, x_2, \ldots, x_d)'$가 평균 $\boldsymbol{\mu}^{(d)} = (\mu_1, \mu_2, \ldots, \mu_d)'$와 공분산 $\Sigma^{(d)} = ((\sigma_{ij}))_{d \times d}$을 갖는 d변량 정규분포를 따를 때 $f(x_d | (x_1, x_2, \ldots, x_{d-1}))$는 다음과 같은 평균과 분산을 갖는 정규분포 확률밀도함수이다.

$$m_d = \mu_d + (\sigma_{1d}, \sigma_{2d}, \cdots, \sigma_{d-1,d})(\Sigma^{(d-1)})^{-1}(\boldsymbol{x}^{(d-1)} - \boldsymbol{\mu}^{(d-1)}),$$

$$v_d = \sigma_{dd} - (\sigma_{1d}, \sigma_{2d}, \cdots, \sigma_{d-1,d})(\Sigma^{(d-1)})^{-1}(\sigma_{1d}, \sigma_{2d}, \cdots, \sigma_{d-1,d})'.$$

위의 사실을 이용하여 $Q(d)$를 $\hat{Q}(d) = m_d$로 대체한다. 또한 생존분석을 위해 (3.1)의 우도함수를 최대화하는 λ와 $\boldsymbol{\beta}$, $(b_1, b_2, \ldots, b_t)$ 계수벡터를 추정하여야 한다. 그런데 이를 위한 계산은 매우 복잡하기 때문에 $\boldsymbol{\beta}$와 $(b_1, b_2, \ldots, b_t)$가 모두 QoL의 계수라는 점, 관측된 QoL값 들이 서로 상관관계가 높다는 점에 착안하여 수식을 좀 더 간단히 바꾸고자 한다. 우선 중도탈락시점의 QoL 값과 직전 QoL 값만을 사용하여 (3), (4), (5)식을 다음과 같이 표현할 수 있다.

$$(3) = \sum_{i=l+1}^{n} \boldsymbol{\beta}' \boldsymbol{Q}_i(t_i) = \sum_{i=l+1}^{n} (b_1 \hat{Q}_i(d_i) + b_2 Q_i(d_{i-1}))$$

$$(4) = -\sum_{i=l+1}^{n} \sum_{j=1}^{d_i - 1} \log[1 + \exp(b_{j0} + b_1 Q_i(d_i - j) + b_2 Q_i(d_i - j - 1))]$$

$$(5) = \sum_{i=l+1}^{n} (b_{d_i 0} + b_1 \hat{Q}_i(d_i) + b_2 Q_i(d_i - 1) - \log[1 + \exp(b_{d_i 0} + b_1 \hat{Q}_i(d_i) + b_2 Q_i(d_i - 1)]).$$

그리고 결측이 없는 개체의 우도한수인 (1), (2)는 다음과 같이 표시된다.

$$(1) = l \log \lambda + \sum_{i=1}^{l} (b_1 Q_i(t_i - 1) + b_2 Q_i(t_i - 2)) - \sum_{i=1}^{l} \int_0^{t_i} \lambda e^{b_1 Q_i(u_i - 1) + b_2 Q_i(u_i - 2)} du_i$$

$$(2) = -\sum_{i=1}^{l} \sum_{j=1}^{t_i} \log[1 + \exp(b_{j0} + b_1 Q_j(t_j - j) + b_2 Q_j(t_j - j - 1,))]$$

여기서 b_1에 마지막 QoL 값인 $Q(t)$를 곱하지 않고 $Q(t-1)$을 곱하는 이유는 $Q(t)$값이 0이고 사망이 $(t-1<x<t)$구간에서 발생하였으므로 $Q(t-1)$을 마지막 QoL 측정값으로 간주할 수 있기 때문이다.

위에 식에 근거하여 로그우도함수의 구체적인 형태를 유도하는 간단한 예를 들어보자. 처리간에 구별 없이 다음 표와 같이 3개의 개체가 최대 5번 QoL이 반복측정된 것을 가정하고 이들의 중도탈락은 3번째 개체에서만 발생했다고 하자.

관측시점		1	2	3	4	5	생존시간	비고
개체	1	$Q_1(1)$	$Q_1(2)$	$Q_1(3)$	$Q_1(4)$	$Q_1(5)=0$	5	
	2	$Q_2(1)$	$Q_2(2)$	$Q_2(3)$	$Q_2(4)=0$	-	4	
	3	$Q_3(1)$	$Q_3(2)$	$Q_3(3)$	NA	NA	0	중도탈락

첫 번째 개체의 로그우도함수는 다음과 같다.

$$\log\lambda + b_1 Q_1(4) + b_2 Q_1(3) - \lambda(e^{b_1 Q_1(2)+b_2 Q_1(1)} + e^{b_1 Q_1(3)+b_2 Q_1(2)} + e^{b_1 Q_1(4)+b_2 Q_1(3)}) + \log[1+\exp(b_{20}+b_1 Q_1(2)+b_2 Q_1(1))] + \log[1+\exp(b_{30}+b_1 Q_1(3)+b_2 Q_1(2))] + \log[1+\exp(b_{40}+b_1 Q_1(4)+b_2 Q_1(3))]$$

두 번째 개체의 로그우도함수는 다음과 같다.

$$\log\lambda + b_1 Q_2(3) + b_2 Q_2(2) - \lambda(e^{b_1 Q_2(2)+b_2 Q_2(1)} + e^{b_1 Q_2(3)+b_2 Q_2(2)}) + \log[1+\exp(b_{20}+b_1 Q_2(2)+b_2 Q_2(1))] + \log[1+\exp(b_{30}+b_1 Q_2(3)+b_2 Q_2(2))]$$

마지막으로 세 번째 개체의 로그우도함수는 다음과 같이 표시할 수 있다.

$$b_1\widehat{Q}_3(4) + b_2 Q_1(3) + \log[1+\exp(b_{20}+b_1 Q_3(2)+b_2 Q_3(1))] + \log[1+\exp(b_{30}+b_1 Q_3(3)+b_2 Q_3(2))] + \log[1+\exp(b_{40}+b_1 \widehat{Q}_3(4)+b_2 Q_3(3))] (b_{40}+b_1\widehat{Q}_3(4)+b_2 Q_3(3) - \log[1+\exp(b_{40}+b_1\widehat{Q}_3(4)+b_2 Q_3(3)]).$$

앞에서 구한 로그우도함수 (3.1)을 최대화하는 모수를 추정하고 그때 최적값을 구하면 모수의 가정에 따라 우리가 원하는 가설을 검정할 수 있다. 예를 들어 (1) '$b_1=b_2=0$' 은 CRD를 의미하고 (2) '$b_1=0, b_2\neq 0$' 은 RD, (3) '$b_1\neq 0, b_2\neq 0$' 은 ID를 의미한다. 그러므로 (1)의 가정하에 (3.1)의 최대값 L_1을 구하고 (2)의 가정하에 (3.1)의 최대값 L_2를 구하여 그 차이 L_2-L_1를 자유도 1인 카이제곱 분포와 비교하여 검정하면 이는 중도탈락이 CRD인지 RD인지를 검정하는 검정법이 된다. 또한 처리 그룹간에 생존기간의 차이를 검정할 때도 '$\lambda_1=\lambda_2=\lambda$' 는 두 처리의 동일성을 '$\lambda_1\neq\lambda_2$' 은 처리효과의 존재를 의미하므로 마찬가지 방법으로 검정할 수 있다. 로그우도함수를 최대화하는 방법으로는 주로 Nelder와 Mead, Broyden-Fletcher-Goldfarb-Shanno(BFGS) 방법 등이 주로 사용된다.

4. 결론

정보적 중도탈락이 발생하였을 때, 중도탈락을 고려하지 않은 표준적인 생존분석은 추론의 편의(bias) 발생 등의 많은 문제점을 갖고 있다. 본 논문에서는 생존이 주요 끝점인 경시적 자료분석에서 삶의 질(QoL)이 같이 측정되었을 때 이를 이용하여 생존함수를 추론하는 우도함수 분석방법을 제안하였다. 기존의 방법들이 중도탈락의 패턴을 직접적으로 추론모형에 포함시키는 방법이 아닌 간접적으로 QoL을 반영하여 가능한 편의를 줄이는 방법이 주로 사용되어 왔던데 반해 우리는 중도탈락과 QoL을 모형에 동시에 포함시켜 모형화 하였다. 이러한 추론법은 중도탈락의 무작위성(randomness)에 대한 검정도 가능하고 생존함수의 동일성에 대한 검정도 가능하다. 그러나 간단한 예제의 우도함수에서 알 수 있듯이 함수의 형태가 무척 복잡하고 현재로서는 통계프로그램인 R을 이용하여 최적의 모수를 추정하고 검정하는 것이 유일한 방법이다. R 프로그램에 능통하지 않은 사용자도 추론이 가능하도록 제안한 우도함수 추론방법을 패키지화하는 것이 앞으로의 과제라 할 수 있다. 또한 중도탈락 시점에서 비관측된 QoL의 적절한 추정방법과 이에 따른 검정법의 특성 연구도 계속되어야 할 연구라 하겠다.

참고문헌

[1] Bowling, A. (1991) Measuring Health: A Review of Quality of Life Measurement Scales. Buckingham: Open University Press.

[2] Cox, D. R. (1972). Regression Models and Life Tables, Journal of the Royal Statistical Society B, Vol. 34, 187-220.

[3] Diggle, P., Kenward M. G. (1994). Informative Drop-out in Longitudinal Data Analysis (with discussion). Applied Statistics, Vol. 43, 49-94.

[4] Fanshel, S., Bush, J. W. (1970). A Health-status Index and Its Applications to Health Services Outcomes. Operat Res, Vol. 18, 1021-1066.

[5] Faucett, C. L., Thomas, D. C. (1996). Simultaneously Modelling Censored Survival Data and Repeatedly Measured Covariates; a Gibbs Sampling Approach, Statistics in Medicine, Vol. 15, 1633-1685.

[6] Fix, E., Neyman, J. A. (1951). Simple Stochastic Model of Recovery, Relapse and Loss of Patients. Hum Biol, Vol. 23, 205-241.

[7] Glasziou, P. P., Simes, R. J., Gelber, R. D. (1990). Quality Adjusted Survival analysis. Statistics in Medicine, Vol. 9, 1259-1276.

[8] Goldhirsch, A., Gelber, R. D., Simes, R. J., Glasziou, P., Coates, A. S. (1989). Costs and Benefits of Adjuvant Therapy in Breast Cancer: a Quality-Adjusted Survival Analysis. Journal of Clinical Oncology, Vol. 7, 36-44.

[9] Little, R. J. A., Rubin, D. B. (1987). Statistical Analysis with Missing Data. New York, John Wiley.

[10] Murray, G. D., Findlay, J. G. (1988) Correcting for the Bias Caused by Drop-outs in Hypertension Trials. Statistics in Medicine. Vol. 7, 941-946.

[11] Roy, J., Lin, X. (2005). Missing Covariates in Longitudinal Data with Informative Dropouts: Bias Analysis and Inference. Biometrics, Vol. 61, 837-846.

[12] Rubin, D. B. (1976). Inference and Missing Data, Biometrika, Vol. 63, 581-592.

[13] Schumacher, M., Olschewski, M., Schulgen, G. (1991). Assessment of Quality of Life in Clinical Trials. Statistics in Medicine, Vol. 10, 1915-1930.

[14] Zwinderman, A. H. (1992). Statistical Analysis of Longitudinal Quality of Life Data with Missing Measurements. Qual Life Res, Vol. 1, 219-224.