7장 Cox 비례위험모형

2020년 가을학기

전북대학교 통계학과

Outline

Cox 비례위험모형

공변량이 한 개인 경우 비례위험모형

동점처리

회귀계수에 대한 검정

층화된 비례위험모형

Cox 비례위험모형

Cox 비례위험모형

- 위험함수가 공변량에 의해 영향을 받는다면 모형에 포함 $\rightarrow$ 회귀모형
- 생존시간의 분포에 대한 정보가 있다면 모수회귀모형
- 분포에 대한 가정이 적절하지 못할 경우 모수회귀모형의 사용은 추정된 회귀계수의 정확성에 대한 신뢰성을 잃음

비례위험모형 (proportional hazard model: PH model)

- 위험함수에 공변량에 대한 회귀식을 포함하는 모형- 모수모형의 가정에 영향을 받지 않은 방법
- 반응변수인 생존시간과 공변량의 관계를 위험함수를 통해 표현
- 기저분포에 대한 가정 필요
- 시간 가변 공변량 (time-varying covariate)을 회귀모형에 포함시킬 수 있음
- 추정된 회귀계수는 공변량과 위험함수의 관계를 나타냄

Cox 비례위험모형

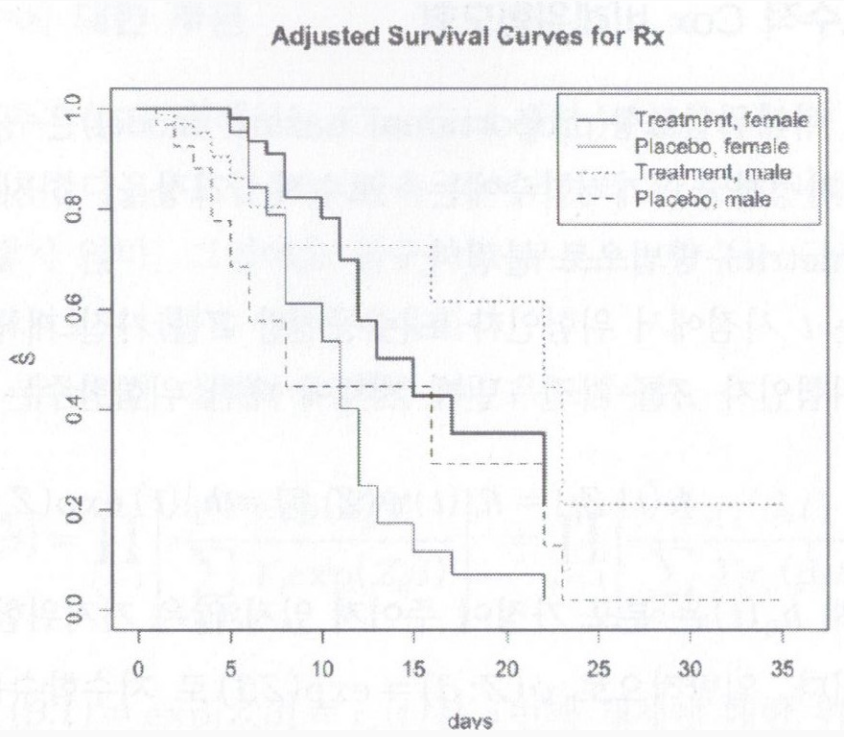

Figure 1: 처리/위약그룹 공변량과 성별 공변량이 포함된 비례위험 모형의 생존함수 그래프

Example
공변량의 예

- 인구특성: 나이, 성별, 사회경제적 상태, 교육정도
- 행동특성: 식습관, 흡연여부, 운동정도 , 음주정도
- 생리특성: 혈압, 혈중 글루코오즈 수준, 헤모글로빈 수치, 심장박동
- 위험인자 (risk factor) : 치매의 위험을 증가시키는 APP (Amyloid beta-protein precursor) 유전인자 여부, 호르몬 수치, 가족 병력

예제 7.1 : 모수회귀모형과 비례위험모형의 차이점
한 개의 공변량 Z가 있는 경우 고려

(a) 모수회귀모형: $\log T_i = Z_i\beta + \epsilon_i$

- 공변량 Z_i가 반응변수 $\log T_i$와 선형관계로 표현
- 회귀계수 β는 공변량의 변화에 따른 log(생존시간)의 관계

(b) 비례위험모형 : $h(t|Z_i) = h_0(t)\exp(Z_i\beta)$

- 회귀계수 β는 공변량의 변화에 따른 위험률의 관계

공변량이 한 개인 경우 비례위험모형

데이터 구조 - 공변량이 한 개인 경우

- T_i : i번째 개체가 연구에 있는 시간
- δ_i: i번째 개체에 대한 사건발생/중도절단 지시변수

$$\delta_i = \begin{cases} 1 & \text{관측된 (observed) 경우} \\ 0 & \end{cases}$$

- $Z_i(t)$: t시점에서 i번째 개체에 대한 위험인자 또는 공변량
- $t_{(1)} < t_{(2)} < \cdots < t_{(n)}$을 편의상 $t_1 < \cdots < t_n$

예제 7.2

(a) 대장암환자 수술 후 생존시간 데이터 분석 - 성별변수를 공변량으로 고려

(b) 폐암환자 수술 후 생존시간 데이터 분석 - 흡연여부를 공변량으로 고려

준모수적 Cox 비례위험모형

Cox 비례위험모형 - 준모수적 (semiparametric) 방법

- 공변량에는 모수적 함수형태 가정
- 기저함수에는 모수적 가정하지 않음

$h(t|Z)$: t시점에서 위험인자 또는 공변량 Z를 가진 개체에 대한 위험률

$$h(t|Z) = h_0(t)\psi(Z;\beta) = h_0(t)\exp(Z\beta)$$

- $h_0(t)$:분포가정이 주어져 있지 않은 기저위험함수 (baseline hazard function)
- $\psi(Z;\beta)$: 공변량의 함수모형
  - $\exp(Z\beta)$인 경우 (가장 일반적)
    - β : 공변량 효과를 추정하는 회귀계수
    - $\psi(Z;\beta) = \exp(Z\beta)$일 때 공변량 Z가 한 단위 증가할 때 위험률 $exp(\beta)$만큼 증가
  - 선형함수 $1 + Z\beta$, 로지스틱 함수 $\log(1 + e^{Z\beta})$ 등도 사용됨

준모수적 Cox 비례위험모형

예제 7.3 :한 개의 공변량을 고려하는 경우, 두 개체간 사건 발생 위험비 비교
개체 i와 개체 j의 사건발생 위험비

$$\frac{h(t|Z_i)}{h(t|Z_j)} = \frac{h_0(t)\exp(Z_i\beta)}{h_0(t)\exp(Z_j\beta)} = \exp[(Z_i - Z_j)\beta]$$

시간에 의존하지 않고 회귀계수 β와 공변량 값의 차 $(Z_i - Z_j)$에 의존

예제 7.4 :공변량이 치료법 $Z = 1$ 또는 0인 이항변수인 경우
개체 i와 개체 j의 사건발생 위험비 (서로 다른 치료법을 처치받은 환자에 대한 사건 발생 위험비)

$$\frac{h(t|Z_i)}{h(t|Z_j)} = \frac{h_0(t)\exp(Z_i\beta)}{h_0(t)\exp(Z_j\beta)} = \exp[(Z_i - Z_j)\beta] = \exp(\beta)$$

$\exp(\beta)$: 시간이 변해도 불변 $\Rightarrow$ Cox "비례위험 (proportional hazard)"

회귀계수에 대한 추론

회귀계수 β 추정 - 부분우도함수 (partial likelihood function) 이용 (Cox 제안)

동점이 없는 생존시간 데이터에 대한 부분우도함수

$$PL(\beta) = \prod_{i=1}^{n} \left[\frac{Y_i \exp(Z_i\beta)}{\sum_{l \in R_i} Y_i \exp(Z_l\beta)} \right]_i^{\delta} = \prod_{i=1}^{n} \left[\frac{Y_i r_i(\beta, t)}{\sum_{l \in R_i} Y_l r_l(\beta, t)} \right]^{\delta_i}$$

- $r_i(\beta, t) = \exp[z_i\beta] \equiv r_i(t)$: i번째 개체에 대한 위험함수 (risk score)
- $Y_i = Y_i(t)$: i번째r가 t시점에서 위험집합에 속하면 1, 아니면 0

로그 부분우도함수

$$l(\beta) = \log PL(\beta) = \sum_{i=1}^{n} \delta_i \left[\log[Y_i \exp(Z_i\beta)] - \log \sum_{l \in R_i} Y_i \exp(Z_l\beta) \right]$$

회귀계수에 대한 추론

점수함수 (score function) $U(\beta)$
로그 부분우도함수 $l(\beta)$를 β에 대해 미분하면

$$U(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} \delta_i \left[\frac{Z_i Y_i \exp(Z_i \beta)}{Y_i \exp(Z_i \beta)} - \frac{\sum_{l \in R_i} Z_l Y_l \exp(Z_l \beta)}{\sum_{l \in R_i} Y_l \exp(Z_l \beta)} \right]$$

최대부분우도추정량 $\hat{\beta}$: $U(\hat{\beta}) = 0$

회귀계수 추정량 $\hat{\beta}$의 통계적 성질

1. 일치성 (consistency)
2. 근사적 정규성 (asymptotic normality) : 마팅게일 중심극한정리

$$\hat{\beta} - \beta \dot{\sim} N(0, I^{-1}(\beta)), \quad I(\beta) = -\frac{\partial^2 l(\beta)}{\partial \beta^2}$$

   $100(1-\alpha)$%신뢰구간 : $\hat{\beta} \pm z_{\alpha/2}[\hat{I}^{-1}(\hat{\beta})]^{1/2}$

3. 효율성 (efficiency) : β에 대한 추정량들 중 MPLE (maximum partial likelihood estimator)가 최소분산을 가짐

회귀계수에 대한 추론

예제 7.7

	생존시간	중도절단여부	공변량
ID	t	δ	Z
1	2	1	2
2	2	0	2
3	3	1	1
4	4	1	3

부분우도함수 $PL(\beta)$

동점처리

동점처리

- 비례위험모형은 모든 시점들이 서로 동일하지 않다는 가정하에서 부분우도함수 유도
- 실제 데이터에서 생존시간이 동일한 경우 종종 발생

(1) 정확방법

(2) Breslow 방법

(3) Efron 방법

- 동점 개수가 크지 않은 경우 Breslow 방법과 Efron 방법의 우도값은 유사
- R에서 Efron 방법이 디폴트

회귀계수에 대한 검정

회귀계수에 대한 검정

$H_0 : \beta_j = \beta_j^{(0)}$

$\hat{\beta}$의 근사적 정규성 이용한 Z검정 :$Z = \dfrac{\hat{\beta}_j}{se(\hat{\beta}_j)}$

$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}$
$\boldsymbol{\beta}$: p개의 공변량의 회귀계수벡터

(1) 우도비 검정 : $L = 2[l(\hat{\boldsymbol{\beta}}) - l(\boldsymbol{\beta}^{(0)})] \overset{H_0}{\sim} \chi_p^2$

(2) Wald 검정 : $W = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)})' \hat{I}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)})$

(3) 점수검정 : $L = U'(\boldsymbol{\beta}^{(0)}) \hat{I}(\boldsymbol{\beta}^{(0)}) U(\boldsymbol{\beta}^{(0)})$

층화된 비례위험모형

층화된 비례위험모형

각 개체가 서로 다른 그룹 또는 층에 속한 경우 층에 따라 기저위험함수가 서로 다르지만 회귀계수는 동일하다고 가정하는 경우

Example
여러 개의 병원에서 진행된 임상시험 (multicenter clinical trials)

$\Rightarrow$ 환자 모집단의 다양성, 환경적, 지리적 차이 등에 의한 병원 간 이질성 존재

병원 개개의 효과에는 관심 없음

$\Rightarrow$ 병원을 공변량에 포함하는 것보다 병원 별로 기저함수를 다르게 가정

- k번째 층 i번째 개체의 위험함수 : $h_{ki}(t|Z) = h_k(t)\exp(Z'_{ki}\beta)$
- K층을 가진 비례위험모형의 부분우도함수 : $L(\beta) = \prod_{k=1}^{K} L_k(\beta)$
- 로그부분우도함수 : $l(\beta) = \sum_{k=1}^{K} l_k(\beta)$

NOTE

- 층화모형의 장점 - 교락변수 (confounding variable)에 대한 효과 보정
- 단점 - 층의 개수가 크면 추정된 회귀계수의 정확도 감소

층화된 비례위험모형

예제 7.11
미국에서 베테랑인 폐암환자 137명에 대해 8개변수 조사