8장 잔차와 모형 적합도 진단

2020년 가을학기

전북대학교 통계학과

Outline

Cox-Snell 잔차

마팅게일 잔차

편차 잔차

Schoenfeld 잔차

비례위험률 가정에 대한 점검

모형 적합 후 모형 적절성 평가

Cox 비례위험모형에 대한 진단

(i) 공변량들의 적절한 함수 형태

(ii) 비례위험모형 가정의 적절성 검토 - 그래프 이용

(iii) 이상점(outlier) 파악

(iv) 영향점(influence point), 지렛점(leverage point) 파악

$\Rightarrow$ 모형의 적합도 평가를 위해 잔차 사용

잔차에 대한 정의

- Cox-Snell 잔차
- 마팅게일 잔차
- 편차 잔차
- Schoenfeld 잔차

Cox-Snell 잔차

Cox-Snell 잔차

모형이 적절하다면 각 개체에 대한 위험함수

$$H(t|\boldsymbol{Z}_i) = H_0(t)\exp(\boldsymbol{Z}_i\boldsymbol{\beta}) \sim Exp(1) \quad [\text{참고 } 8.1]$$

Cox-Snell 잔차

$$r_i^{CS} = \hat{H}_0(t_i)\exp(\boldsymbol{Z}_i\hat{\boldsymbol{\beta}})$$

$\hat{H}_0(t_i) = \sum_{t_i \le t} \frac{d_i}{\sum_{j\in R(t_i)} \exp(\boldsymbol{Z}_j\hat{\boldsymbol{\beta}})}$ 누적기저위험함수 추정량 (Breslow 추정량)

Note

- $H(t) = -\log S(t) = -\log(1 - F(t))$
- 모형이 타당하다면 $\hat{S}(t_i)$는 $S(t_i)$와 가깝고
- $-\log \hat{S}(t_i) \sim Exp(1)$

Cox-Snell 잔차

- 일반적인 선형회귀모형의 잔차와 다른 형태
- 0을 중심으로 대칭이 아니고 음의 값을 갖지 않음
- 모형이 타당할 경우 평균과 분산이 각가 1인 치우친 분포 형태
- 모형이 타당하다면 누적위험함수 $H_E(t) = t$, 즉 (t_i, r_i^{CS})는 원점을 지나는 직선 형태

마팅게일 잔차

마팅게일 잔차

주어진 공변량에 대해 나머지 다른 공변량의 효과를 적합한 후,
생존시간에 미치는 영향을 Cox 비례위험모형을 통해 잘 설명하는 적절한 함수 형태를 결정하고자 하는 경우

공변량에 대한 변환함수의 예
$Z, \log Z, Z^2, Z \log Z$ 또는 Z의 범위에 따라 다른 함수 형태 고려

마팅게일 잔차 M_i

- Cox-Snell 잔차의 변형
- 관측값 여부 (0 또는 1)에서 누적위험함수값을 빼서 계산

$$M_i = \delta_i - \hat{H}(t_i|\boldsymbol{Z}_i), \quad i = 1, 2, \ldots, n$$

- $\delta_i = 1$ (관측된 경우), 0(중도절단된 경우)
- $\hat{H}(t_i|\boldsymbol{Z}_i)$는 t_i시점에서 누적위험함수 추정값

마팅게일 잔차

예 8.1

(a) A는 2개월에 중도절단되었고 2개월까지의 누적위험률은 0.2라고 한다. A의 마팅게일 잔차는 $M_A =$

(b) B는 13개월에 사망했으며 13개월까지의 누적위험률은 0.7이라고 한다. $M_B =$

마팅게일 잔차는 $O - E$형태, 즉 관측값 O와 기댓값 E의 차이

$$\hat{M}_i = \delta_i - \hat{H}_0(t_i)\exp(\boldsymbol{Z}_i'\hat{\boldsymbol{\beta}}) = \delta_i - r_i^{CS}$$

마팅게일 잔차의 특성

(1) $E(M_i) = 0$, 회귀계수벡터 $\boldsymbol{\beta}$값이 참일 경우 기댓값은 0

(2) $\sum_{i=1}^{n} \hat{M}_i = 0$ 마팅게일 잔차의 합은 0

(3) $Cov(M_i, M_j) = 0$ 회귀계수벡터 $\boldsymbol{\beta}$값이 참일 경우

(4) $Cov(\hat{M}_i, \hat{M}_j) < 0$ 실제 추정된 마팅게일 잔차는 음의 상관성을 가짐

마팅게일 잔차

- 선형회귀모형에서 잔차는 모형의 적합도에 활용
  - 잔차제곱합
  - 반응변수의 추정값과 잔차의 산점도
- **마팅게일 잔차의 역할** - 공변량의 함수형태를 찾는 데 유용

Example
p개의 공변량들 중 한 개 공변량 Z_1에 대한 적절한 함수 형태 $g(Z_1)$을 결정하고자 할 경우

(1) 공변량 Z_1를 제외한 공변량 벡터 $\boldsymbol{Z}^* = (Z_2, \ldots, Z_p)'$에 대해 비례위험모형

$$H(t|\boldsymbol{Z}^*) = H_0(t)\exp(\boldsymbol{Z}'^*\boldsymbol{\beta}^*) \quad \text{적합 후}$$

(2) 마팅게일 잔차 $\hat{M}_i$에 대하여 $(Z_{i1}, \hat{M}_i)$산점도를 그려서

(3) 공변량 Z_1에 대한 적절한 함수 $g(Z_1)$을 추측

(4) 잔차그림에서 직선관계가 보이면 $g(Z_1) = Z_1\beta$, 이차관계이면 $g(Z_1) = Z_1^2\beta$ 등을 고려

편차 잔차

편차 잔차 (deviance residual)

- 마팅게일 잔차의 범위 $-\infty$에서 1 사이로 대칭이 아님
- 마팅게일 잔차에 대해 표준화 변환한 잔차로 0 중심으로 분산 1을 가지며 대칭
- 편차 잔차가 큰 개체일수록 모형을 통한 추정이 좋지 않음을 의미
- 편차 잔차는 선형회귀모형 잔차와 유사
- 모형이 적합하다면 (편차 잔차, 공변량)에 대한 산점도를 그렸을 때 특별한 패턴이 보이지 않음
- 모형적합 후 이상점 점검 -편차 잔차를 통해 이상값 식별

$$D_i = sign(\hat{M}_i) \times \sqrt{-2[\hat{M}_i + \delta_i \log(\delta_i - \hat{M}_i)]}$$

- $\hat{M}_i = 0 \Rightarrow D_i = 0$
- 중도절단이 많지 않은 경우 D는 서로 독립이고 정규분포 형태
- 절댓값이 큰 편차 잔차를 갖는 관측값은 이상값으로 의심

Schoenfeld 잔차

Schoenfeld 잔차

각 개체에 대해 각 공변량에 대한 잔차 $\Rightarrow$ 부분 잔차 (Schoenfeld's partial residual)

Schoenfeld 잔차
j번째 개체의 k번째 공변량에 대한 Schoenfeld 부분 잔차

$$S_{jk} = \delta_j |Z_{jk} - \overline{Z}_k(\beta, t_j)|$$

$$\overline{Z}_k(\beta, t_j) = \frac{\sum_i Y_i r_i(s) Z_{ik}(s)}{\sum_i Y_i r_i(s)} = \frac{\sum_i Y_i \exp(Z_{ik}\beta) Z_{ik}(s)}{\sum_i Y_i \exp(Z_{ik}\beta)}$$

- 각 개체에 대해 사건발생 시 공변량 값에서 해당 기댓값을 뺀 값
  기댓값=위험집합에서 공변량들을 이용하여 계산
- Schoenfeld 잔차는 p개의 열을 가진 행렬로 구성되고 각 사건은 각 행을 구성
- Schoenfeld 잔차는 사건발생 시간과 독립
- 모형이 만족한다면 (시간, Schoenfeld 잔차) 산점도에 특별한 패턴이 없음
- Schoenfeld 잔차는 중도절단 데이터에 대해서는 정의 안됨
- **영향치 식별**에 사용

비례위험률 가정에 대한 점검

비례위험률 가정에 대한 점검

- 그래프적 방법 - Anderson plot, Arjas plot, log cumulative baseline hazard plot, Schoenfeld 잔차(점수 잔차)에 의한 방법

누적기저위험함수 그래프

시간-고정인 변수이며 몇 개의 수준이 있는 공변량에 대해

$$S(t) = \exp(-H_0(t)Z_i\beta)$$

$$\log[-\log S(t)] = \log[H_0(t)] - Z_i\beta$$

- Cox 비례위험모형이 맞다면 각 수준에서의 Kaplan-Meier 추정선은 $\log - \log$그림을 그렸을 때 서로 평행
- 수준별 $\log - \log$추정선이 서로 교차하면 비례성 관계 의심

비례위험률 가정에 대한 점검

누적 Schoenfeld 잔차 (또는 점수 잔차) 그림

- 비례위험모형이 맞다면 Schoenfeld 잔차 합은 Brownian bridge 과정
- 회귀계수추정량 $\hat{\beta}$의 표준화 점수 잔차 (standaridized score residual)은 비례성을 만족하면 ± 1.3581을 벗어날 확률이 0.05이내
- 이 범위를 벗어나면 유의수준 5%에서 비례성 벗어난다고 판단.
- 표준화 점수 잔차 r_i^S 에 대해 (t_i, r_i^S)그림을 그려 비례성 판단