

생존분석 시험대비 문제집 (문제 + 해설 포함)

Taenyoung 대비용 종합 세트

October 20, 2025

1 기초 및 연계식

문제 1 (생존함수-위험함수 연계식 증명). 연속형 생존시간 T 에 대해 $S(t) = P(T > t)$, $h(t) = \lim_{\Delta \rightarrow 0} \frac{P(t < T \leq t + \Delta | T > t)}{\Delta}$ 라 하자. 다음을 증명하라:

$$S(t) = \exp\left(-\int_0^t h(u) du\right).$$

해설/정답. $f(t) = -S'(t)$ 이고 $h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -(\ln S(t))'$. 적분하면 $\ln S(t) - \ln S(0) = -\int_0^t h(u) du$. $S(0) = 1$ 으로 $\ln S(0) = 0$, 따라서 $S(t) = \exp\{-\int_0^t h(u) du\}$. \square

문제 2 (Cox HR의 시간불변성). Cox 모형 $h(t|X) = h_0(t) \exp(\beta^\top X)$ 에서 두 개인 X, X^* 의 위험비(HR)가 시간에 의존하지 않음을 보여라.

해설/정답.

$$\frac{h(t|X^*)}{h(t|X)} = \frac{h_0(t) \exp(\beta^\top X^*)}{h_0(t) \exp(\beta^\top X)} = \exp\{\beta^\top (X^* - X)\},$$

$h_0(t)$ 가 소거되어 시간 t 에 무관. \square

2 KM, Greenwood, 로그순위

문제 3 (KM 추정량과 Greenwood 분산(계산)). 사건시점 $t_{(1)} < \dots < t_{(J)}$ 에서 위험집합 크기 n_j , 사건수 d_j 가 주어질 때

$$\hat{S}(t) = \prod_{t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

의 표준오차 $SE\{\hat{S}(t)\}$ 를 **Greenwood**로 구하는 공식을 쓰고, 아래 표에 대해 $\hat{S}(t_{(3)})$ 와 SE , 95% CI(로그-로그 변환)를 계산하라.

j	$t_{(j)}$	n_j	d_j
1	2	100	3
2	5	95	4
3	8	90	6

해설/정답. Greenwood:

$$\widehat{\text{Var}}\{\hat{S}(t)\} = \hat{S}(t)^2 \sum_{t_{(j)} \leq t} \frac{d_j}{n_j(n_j - d_j)}, \quad \text{SE} = \sqrt{\widehat{\text{Var}}}.$$

수치:

$$\hat{S}(t_{(3)}) = \left(1 - \frac{3}{100}\right) \left(1 - \frac{4}{95}\right) \left(1 - \frac{6}{90}\right) = 0.97 \times 0.9579 \times 0.9333 \approx 0.868.$$

분산합: $\frac{3}{100.97} + \frac{4}{95.91} + \frac{6}{90.84} \approx 0.000309 + 0.000462 + 0.000794 \approx 0.001565$. 따라서 $\widehat{\text{Var}} \approx 0.868^2 \times 0.001565 \approx 0.00118$, SE ≈ 0.0343 .

로그-로그 변환 CI: $\eta = \log[-\log \hat{S}]$. $\hat{S} = 0.868 \Rightarrow -\log \hat{S} \approx 0.141$, $\eta \approx \log(0.141) \approx -1.958$. 표준오차는 델타법으로 $\text{SE}_\eta \approx \frac{1}{-\log \hat{S}} \cdot \frac{1}{\hat{S}} \cdot \text{SE}_{\hat{S}} \approx \frac{1}{0.141} \cdot \frac{1}{0.868} \cdot 0.0343 \approx 0.283$. 95% CI for η : $-1.958 \pm 1.96 \times 0.283 \Rightarrow (-2.514, -1.402)$. 역변환: $S_L = \exp\{-\exp(\eta_U)\}$, $S_U = \exp\{-\exp(\eta_L)\}$. $\exp(\eta_U) = \exp(-1.402) \approx 0.246 \Rightarrow S_L \approx e^{-0.246} = 0.782$. $\exp(\eta_L) = \exp(-2.514) \approx 0.081 \Rightarrow S_U \approx e^{-0.081} = 0.922$. 따라서 95% CI $\approx (0.782, 0.922)$. \square

문제 4 (가중/총화 로그순위의 아이디어(서술)). 혼란변수 Z 가 존재할 때 **총화 로그순위** 검정의 통계량 합성 원리를 간단히 서술하고, Fleming-Harrington 계열과 같은 가중 로그순위의 직관을 2-3문장으로 답하라.

해설/정답. 총화 로그순위는 총 $s = 1, \dots, S$ 에서 사건시점 j 마다 관측-기대 차이 $O_{1,sj} - E_{1,sj}$ 를 구해 $U = \sum_s \sum_{j \in \mathcal{J}_s} w_{sj}(O_{1,sj} - E_{1,sj})$ 로 합산하고, $\text{Var}(U) = \sum_s \sum_{j \in \mathcal{J}_s} w_{sj}^2 V_{sj}$ 로 표준화한다. 가중 로그순위는 w_{sj} 를 통해 초기/후기 사건에 민감도를 조정한다. 예컨대 w_j 가 생존함수의 함수(예: $S(t)^p(1 - S(t))^q$)이면 p 가 크면 초기에, q 가 크면 후기에 더 민감해진다. \square

3 Cox: 부분가능도, ties, 추론

문제 5 (부분가능도와 $h_0(t)$ 소거(서술)). Cox 부분가능도의 로그형식

$$\ell(\beta) = \sum_{j=1}^k \left\{ \beta^\top x_{(j)} - \log \sum_{i \in R(t_{(j)})} \exp(\beta^\top x_i) \right\}$$

이 조건부 확률의 곱에서 나오며, $h_0(t)$ 가 소거되는 이유를 3-4문장으로 설명하라.

해설/정답. j 번째 사건시점에서 실패자의 위험비가 위험집합 내에서 가장 먼저 실패할 조건부확률에 비례하며, 분자/분모 모두 $h_0(t_{(j)})$ 를 포함하나 동일 시점에서는 상쇄된다. 사건시점별 조건부확률을 곱하면 전체(부분)가능도가 되고, 로그를 취해 합으로 표현된다. 이 과정에서 $h_0(t)$ 모양은 필요치 않아 β 추정이 가능하다. \square

문제 6 (동일시각 사건 처리(ties) 개념 비교). 동일한 시점 t 에 d 건의 사건이 발생할 때 **Exact, Breslow, Efron** 방법의 차이를 2-3문장씩 요약하라.

해설/정답. *Exact*: d 건의 모든 발생 순열을 고려한 정확 가능성. 계산량 큼, 소표본/동률 빈도 높을 때 정확. *Breslow*: d 건을 동시에 발생한 것으로 간주, 분모를 $(\sum_{i \in R(t)} e^{\beta^\top x_i})^d$ 로 근사. 단순하나 편향 가능. *Efron*: 위험집합에서 사건이 하나씩 제거되는 과정을 선형 보정해 분모를 조정. Exact에 더 근접, 실무 기본값으로 자주 사용. \square

4 PH 가정 점검: 로그-로그, 잔차, cox.zph

문제 7 (로그-로그 평행성의 근거(요약)). $S(t|X) = [S_0(t)]^{\exp(\beta^\top X)}$ 에서

$$\log[-\log S(t|X)] = \beta^\top X + \log[-\log S_0(t)]$$

임을 보이고, 두 집단의 로그-로그 곡선이 시간에 대해 평행해야 함을 설명하라.

해설/정답. 양변 로그 후 재로그로 위 식이 도출된다. 두 집단 X_1, X_2 의 차이는 $(\beta^\top X_1 - \beta^\top X_2)$ 라는 상수이고 t 에 무관하므로 평행성 신호가 된다. \square

문제 8 (Schoenfeld 잔차와 cox.zph 전역검정). Schoenfeld 잔차 $r_k = x_{(k)} - \bar{x}(\hat{\beta}, t_{(k)})$ 의 정의와 “PH가 성립하면 시간과 비상관”인 이유를 설명하고, cox.zph에서 transform="identity", "km", "rank"의 목적을 1-2문장씩 쓰며, 전역(Global) p값 해석을 한 줄로 요약하라.

해설/정답. r_k 는 k 번째 사건에서 관측 공변량과 위험집합 가중평균의 차이다. PH가 성립하면 계수효과가 시간에 일정하여 잔차가 시간에 체계적 패턴을 보이지 않으므로 비상관이어야 한다. identity: 원시(선형) 시간, km: Kaplan-Meier 기반 변환으로 말단에서 변동 안정화, rank: 사건순위로 비모수적 시간척도. 전역 p값은 모든 공변량에 대해 PH 가정이 동시에 성립하는지의 귀무가설 검정 결과다. \square

5 조정 생존곡선(Observed vs Expected)

문제 9 (Breslow $\hat{H}_0(t)$ 와 조정 생존곡선). Breslow 누적기저위험

$$\hat{H}_0(t) = \sum_{t_{(j)} \leq t} \frac{d_j}{\sum_{i \in R(t_{(j)})} \exp(\hat{\beta}^\top x_i)}, \quad \hat{S}_0(t) = \exp\{-\hat{H}_0(t)\}$$

을 쓰고, 임의 공변량 x 에 대한 조정 생존함수 $\hat{S}(t|x) = \hat{S}_0(t)^{\exp(\hat{\beta}^\top x)}$ 를 도출하라. 또한 Observed(KM)과 Expected(조정곡선)을 비교해 PH 진단에 어떻게 쓰는지 2-3문장으로 설명하라.

해설/정답. 정의 그대로이며, Cox 모형에서 $S(t|x) = S_0(t)^{\exp(\beta^\top x)}$ 관계를 추정치로 대체한다. PH가 성립하면 관찰된 KM과 모형기반 조정곡선이 체계적으로 벌어지지 않는다. 일정 구간에서 일관된 편차가 있다면 시간가변 효과 또는 모형 부적합 신호로 본다. \square

6 잔차/영향도 (Martingale/Deviance/dfbeta)

문제 10 (잔차와 영향점 진단(핵심 요약)). Martingale 잔차, Deviance 잔차, dfbeta(dfbetas)의 용도와 해석 포인트를 각 1-2문장으로 요약하라.

해설/정답. Martingale: $M_i = \delta_i - \hat{H}_0(T_i) \exp(\hat{\beta}^\top x_i)$ 로 함수형/선형성 위반 탐지. Deviance: $r_{Di} = \text{sign}(M_i) \sqrt{-2(M_i + \delta_i \log(\delta_i - M_i))}$ 로 극단 관측치 탐지에 유용. dfbeta(dfbetas): 개별 관측 제거 시 $\hat{\beta}$ 변화량(표준화)을 측정, 영향력이 큰 관측을 식별한다. \square

7 R 실습 종합(스캐폴드)

문제 11 (실습: KM/로그순위/총화/가중, Cox/ties, cox.zph, 조정곡선). 다음 스캐폴드를 이용해 데이터프레임 $df(time, status, group, x1, x2, Z)$ 에 대해 (i) KM+로그순위

(표준/총화/가중), (ii) Cox(ties: efron/breslow/exact), (iii) *cox.zph*(변환 3종, 전역 p값), (iv) 조정 생존곡선을 수행하라. 핵심 수치결과(계수/HR/CI/logLik/p값)를 표로 요약하고, PH 진단 결과를 5-7줄로 해석하라.

해설/정답.

보고/해석 템플릿 예시:

- 로그-순위(표준) p값 $\approx p_0$, 총화 후 p_{strat} 로 변화. (혼란 조정에 따라 결론 변화 여부 기술)
- Cox(efron) 추정: $\text{HR}(\text{group B vs A}) = \exp(\hat{\beta}_{\text{group}})$, 95% CI, Wald p값 보고.
- ties 방법별 로그가능도/계수 비교: Efron과 Exact가 유사, Breslow는 (데이터 특성에 따라) 경향 차이.
- cox.zph*: 변수별 p값과 전역 p값. 변환 3종에서 결론 일관성 여부 확인.
- 조정 생존곡선 vs KM: 체계적 벌어짐이 있으면 PH 위반/함수형 문제 가능성 코멘트.

□

8 확장 Cox: 시간의존 효과

문제 12 (시간의존 상호작용 검정). 확장 Cox $h(t|X) = h_0(t) \exp\{\beta X + \delta X g(t)\}$ 에서 $H_0 : \delta = 0$ 의 의미를 쓰고, $g(t) = \log t$ 일 때 R에서 `tt=t`을 이용해 구현하는 스니펫을 제시하라.

해설/정답. $H_0 : \delta = 0$ 은 X 의 효과가 시간불변임(=PH 성립)을 뜻한다. 유의한 $\hat{\delta} \neq 0$ 이면 시간가변 HR을 시사.

```
coxph(Surv(time, status) ~ X + tt(X), data=df,
      tt = function(t, x, ...) { x * log(pmax(t, 1)) })
```

□

추가: 점수 올리는 한 줄 메모

- KM 신뢰구간: 로그-로그 변환을 권장(대칭성 개선).
- 가중 로그-순위: 초기/후기 민감도는 가중식으로 제어.
- ties: 실무 기본은 efron; 동률 다발/소표본이면 exact 검토.
- PH 진단: 변수별 + 전역(Global) p값 둘 다 보고.
- 조정곡선: Observed(KM) vs Expected(조정) 체계적 벌어짐 여부 확인.
- 잔차/영향: Martingale(함수형), Deviance(극단치), dfbetas(영향점) 역할 구분.