시험 대비 생존 분석 심화 개념 정리 (증명 포함)

1 생존 분석의 기초

생존 분석은 어떤 사건(event)이 발생하기까지 걸리는 시간 변수와 이에 영향을 미치는 요인들을 통계적으로 분석하는 방법론입니다.

1.1 주요 함수와 그 관계

- **생존 함수 (Survivor Function, S(t)):** 특정 시점 t를 지나 생존할 확률입니다.

$$S(t) = P(T > t), \quad t > 0$$

- **위험 함수 (Hazard Function, h(t)):** 시점 t까지 생존했을 때, 바로 그 순간(t와 $t + \Delta t$ 사이)에 사건이 발생할 조건부 확률을 나타내는 순간 위험률입니다.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t}$$

- **S(t)와 h(t)의 관계 증명:** 위험 함수의 정의에 따라, $h(t)$는 다음과 같이 표현될 수 있습니다.

$$h(t) = \frac{f(t)}{S(t)}$$

여기서 $f(t)$는 시점 t에서의 사건 발생 확률 밀도 함수(PDF)이며, $f(t) = -S'(t) = -\frac{dS(t)}{dt}$ 입니다. 따라서,

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} \ln S(t)$$

이 식을 0부터 t까지 적분하면,

$$\int_0^t h(u)du = -\int_0^t \frac{d}{du} \ln S(u)du = -[\ln S(t) - \ln S(0)]$$

$S(0) = P(T > 0) = 1$이므로 $\ln S(0) = 0$ 입니다. 따라서,

$$\int_0^t h(u)du = -\ln S(t)$$

이를 $S(t)$에 대해 정리하면 다음과 같은 핵심 관계식이 유도됩니다.

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

2 카플란-마이어(KM) 곡선과 로그-순위 검정

2.1 카플란-마이어 (Kaplan-Meier) 추정량

KM 추정량은 중도절단이 있는 데이터에서 생존 함수를 추정하는 비모수적 방법입니다.

- **KM 공식 (Product-Limit Formula):** 시점 $t_{(j)}$에서의 생존율은 그 이전 시점까지의 생존율에 해당 시점에서 생존할 조건부 확률을 곱하여 계산됩니다.

$$\hat{S}(t) = \prod_{t_{(i)} \leq t} \left(\frac{n_i - m_i}{n_i}\right)$$

여기서 $t_{(i)}$는 i번째 사건 발생 시점, n_i는 해당 시점의 위험 집단(risk set) 크기, m_i는 해당 시점의 사건 수입니다.

- **KM 공식의 유도:** 생존 함수의 정의에 따라, 시점 $t_{(j)}$를 지나 생존하는 것은 시점 $t_{(1)}, t_{(2)}, \ldots, t_{(j)}$에서 연속적으로 사건을 겪지 않는 것과 같습니다. 이는 조건부 확률의 곱으로 표현할 수 있습니다.

$$\hat{S}(t_{(j)}) = \prod_{i=1}^{j} P(T > t_{(i)} | T \geq t_{(i)})$$

  i번째 사건 시점에서 생존할 조건부 확률은 $(n_i - m_i)/n_i$로 추정되므로, 위 공식이 유도됩니다. 또한, 이는 재귀적으로도 표현 가능합니다:

$$\hat{S}(t_{(j)}) = \hat{S}(t_{(j-1)}) \times P(T > t_{(j)} | T \geq t_{(j)})$$

2.2 로그-순위 검정 (Log-Rank Test)

로그-순위 검정은 여러 그룹의 생존 곡선이 통계적으로 동일한지를 검정하는 방법입니다.

- **귀무가설 (H_0):** 모든 그룹의 생존 곡선은 같다 ($S_1(t) = S_2(t) = \cdots = S_G(t)$).
- **원리:** 각 사건 발생 시점 j마다 그룹 1의 **기대 사건 수(e_{1j})**를 계산합니다. 이는 해당 시점의 전체 사건 수($m_j = m_{1j} + m_{2j}$)를 위험 집단의 비율에 따라 배분한 값입니다.

$$e_{1j} = (m_{1j} + m_{2j}) \times \left(\frac{n_{1j}}{n_{1j} + n_{2j}} \right)$$

- **검정 통계량:** 관측된 총 사건 수(O_i)와 기대된 총 사건 수(E_i)의 차이를 이용합니다. 두 그룹 비교 시, 검정 통계량은 다음과 같으며 자유도가 1인 카이제곱 분포를 따릅니다.

$$\text{Log-Rank Statistic} = \frac{(O_2 - E_2)^2}{\text{Var}(O_2 - E_2)} \sim \chi^2(1)$$

  G개 그룹 비교 시에는 자유도가 $G-1$인 카이제곱 분포를 따릅니다.

(보강) 층화/가중 로그-순위 검정 공변량(예: 검사치 범주)을 층화변수 $s = 1, \ldots, S$로 나누고, 각 사건시점의 관측/기대 사건수 차이를 층별로 계산한 후 합산한다.

$$U = \sum_{s=1}^{S} \sum_{j \in \mathcal{J}_s} w_{sj} \left[O_{1,sj} - E_{1,sj} \right], \quad Var(U) = \sum_{s=1}^{S} \sum_{j \in \mathcal{J}_s} w_{sj}^2 \, V_{sj},$$

여기서 w_{sj}는 가중치(기본 로그-순위는 $w_{sj} = 1$), $\mathcal{J}_s$는 층 s의 사건시점 집합이다. 검정통계량 $Z = U/\sqrt{Var(U)}$ 또는 $\chi^2 = Z^2$로 유의성을 판단한다. 실무에서는 `survdiff`의 `rho`로 w_{sj}를 달리할 수 있다 (예: Fleming–Harrington 계열; $\rho = 0$가 표준 로그-순위). 또한 공변량에 따라 층화하려면 `survdiff(time ~ group + strata(z))`처럼 `strata()`를 사용한다.

3 콕스 비례위험(Cox Proportional Hazards) 모델

3.1 콕스 PH 모델의 공식과 가정

콕스 PH 모델은 공변량이 생존 시간에 미치는 영향을 분석하는 준모수적 회귀 모델입니다.

- **모델 공식:**

$$h(t, X) = h_0(t) \cdot \exp\left(\sum_{i=1}^{p} \beta_i X_i \right)$$

  여기서 $h_0(t)$는 **기저 위험 함수(baseline hazard function)**로, 형태를 특정하지 않아 비모수적(non-parametric) 부분이며, $\exp(\ldots)$ 부분은 모수적(parametric) 부분입니다.

- **핵심 가정 (비례 위험 가정):** 두 공변량 벡터 X^*와 X를 가진 개인들의 **위험비(Hazard Ratio)**는 시간에 따라 변하지 않고 일정하다는 것입니다.

$$\frac{h(t, X^*)}{h(t, X)} = \text{constant for all } t$$

3.2 위험비 (Hazard Ratio, HR)의 유도

위험비는 두 그룹의 위험 함수 간의 비율로, 콕스 모델에서 다음과 같이 유도됩니다.

$$\text{HR} = \frac{h(t, X^*)}{h(t, X)} = \frac{h_0(t)\exp(\sum \beta_i X_i^*)}{h_0(t)\exp(\sum \beta_i X_i)} = \exp\left[\sum \beta_i (X_i^* - X_i)\right]$$

여기서 기저 위험 함수 $h_0(t)$가 소거되므로, 위험비는 시간에 의존하지 않는 상수가 됩니다.

3.3 부분 가능도 (Partial Likelihood)

콕스 모델의 회귀 계수 β는 **부분 가능도(Partial Likelihood)**를 최대화하여 추정됩니다. 이는 각 사건 발생 시점에서, 실제로 사건을 겪은 개인이 위험 집단 내 다른 이들보다 먼저 사건을 겪을 조건부 확률들의 곱으로 구성됩니다. j번째 사건이 발생했을 때의 가능도 L_j는 다음과 같습니다.

$$L_j(\beta) = \frac{\text{Hazard for the individual who fails at } t_{(j)}}{\sum_{k \in R(t_{(j)})} \text{Hazards for all individuals in risk set } R(t_{(j)})}$$

$$L_j(\beta) = \frac{h_0(t_{(j)})\exp(\sum \beta_i X_{ij})}{\sum_{k \in R(t_{(j)})} h_0(t_{(j)})\exp(\sum \beta_i X_{ik})} = \frac{\exp(\sum \beta_i X_{ij})}{\sum_{k \in R(t_{(j)})} \exp(\sum \beta_i X_{ik})}$$

전체 부분 가능도는 모든 사건에 대한 L_j의 곱입니다: $L(\beta) = \prod_{j=1}^{k} L_j(\beta)$. 여기서도 $h_0(t)$가 소거되므로, 기저 위험 함수를 몰라도 β를 추정할 수 있습니다.

(보강) 동일시각 사건(ties)의 처리 동일한 시점 t에 사건이 d건 발생하면 부분가능도의 분모/분자 정의가 달라진다. 대표적 방법은 다음과 같다.

- **정확법(Exact/Marginal)**: d건의 사건 발생 순열을 모두 고려한 정확 가능도를 사용. 계산량이 크나 가장 정확.
- **Breslow 근사**: 사건 d건을 한꺼번에 발생한 것으로 보고 분모를 $\left(\sum_{i \in R(t)} e^{\beta^\top x_i}\right)^d$로 근사.
- **Efron 근사**: d건이 위험집합에서 점진적으로 제거된다고 보고 분모에 보정항을 도입(정확법에 더 근접).

실무에선 표본크기/동률 빈도에 따라 `coxph(..., ties="efron")` (기본), `"breslow"`, `"exact"`를 선택한다.

4 비례 위험(PH) 가정 평가

4.1 로그-로그 플롯(Log-log Plots)의 원리 증명

로그-로그 플롯이 평행해야 하는 이유는 콕스 모델의 생존 함수로부터 수학적으로 유도됩니다.

1. 콕스 모델의 생존 함수는 다음과 같습니다:

   $$S(t, X) = [S_0(t)]^{\exp(\sum \beta_i X_i)}$$

   여기서 $S_0(t)$는 기저 생존 함수입니다.

2. 양변에 자연로그를 취합니다 (Log #1):

   $$\ln S(t, X) = \exp\left(\sum \beta_i X_i\right) \cdot \ln S_0(t)$$

3. 양변에 음수를 곱하고 다시 자연로그를 취합니다 (Log #2):

   $$\begin{aligned}\ln[-\ln S(t, X)] &= \ln\left[-\exp\left(\sum \beta_i X_i\right) \cdot \ln S_0(t)\right] \\ &= \ln\left(\exp\left(\sum \beta_i X_i\right)\right) + \ln(-\ln S_0(t)) \\ &= \sum \beta_i X_i + \ln(-\ln S_0(t))\end{aligned}$$

4. 두 개인(공변량 X_1, X_2)에 대한 로그-로그 생존 함수의 차이를 계산하면 다음과 같습니다:

$$\ln[-\ln S(t, X_1)] - \ln[-\ln S(t, X_2)] = \Big(\sum \beta_i X_{1i} + \ln(-\ln S_0(t))\Big) - \Big(\sum \beta_i X_{2i} + \ln(-\ln S_0(t))\Big)$$
$$= \sum \beta_i (X_{1i} - X_{2i})$$

결과적으로, 두 로그-로그 플롯 간의 수직 거리는 시간에 의존하지 않는 상수($\sum \beta_i (X_{1i} - X_{2i})$)가 됩니다. 따라서 두 그래프는 평행해야 합니다.

4.2 시간 의존 변수를 이용한 검정

PH 가정을 통계적으로 검정하는 가장 엄격한 방법은 시간 의존 변수를 포함하는 **확장된 콕스 모델(extended Cox model)**을 사용하는 것입니다.

$$h(t, X) = h_0(t) \exp[\beta X + \delta(X \times g(t))]$$

여기서 $g(t)$는 시간의 함수(예: t 또는 $\ln(t)$)입니다.

- **귀무가설 (H_0):** $\delta = 0$.
- 만약 귀무가설이 기각되면(δ가 0과 유의하게 다르면), X의 효과가 시간에 따라 변한다는 의미이므로, X는 PH 가정을 위배한 것입니다.

4.3 PH 가정의 그래프적/잔차 기반 점검과 실무적 대처

(1) Observed vs Expected 비교(조정 생존곡선) 콕스모형 적합 후 각 사건시점 $t_{(j)}$에서 집단 g의 *기대 사건수*는

$$e_{gj} \;=\; d_j \,\frac{\sum_{i \in R(t_{(j)}) \cap g} \exp(\hat{\beta}^\top x_i)}{\sum_{i \in R(t_{(j)})} \exp(\hat{\beta}^\top x_i)},$$

여기서 d_j는 $t_{(j)}$에서의 총 사건 수, $R(t_{(j)})$는 위험집합입니다. 누적 관측–기대 차이

$$C_g(t) \;=\; \sum_{t_{(j)} \le t} \{O_{gj} - e_{gj}\}$$

를 시점에 따라 그리거나, *조정 생존곡선*

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

을 생성하여 집단별 대표 공변량(또는 평균/중앙 공변량)으로 비교합니다. PH가 성립하면 관찰(KM)과 모형기반(조정) 곡선이 체계적으로 벌어지지 않습니다.

(2) Schoenfeld 잔차 기반 PH 가정 검정(Grambsch–Therneau) k번째 사건시점에서 실패한 개인의 공변량을 $x_{(k)}$라 하고, 위험집합 가중 평균을

$$\bar{x}(\beta, t_{(k)}) \;=\; \frac{\sum_{i \in R(t_{(k)})} x_i \exp(\beta^\top x_i)}{\sum_{i \in R(t_{(k)})} \exp(\beta^\top x_i)}$$

로 두면, *Schoenfeld 잔차*는

$$r_k \;=\; x_{(k)} - \bar{x}(\hat{\beta}, t_{(k)}).$$

PH가 성립하면 r_k는 시간에 의존적 패턴을 갖지 않습니다. 실무에서는 *scaled* 잔차 $\tilde{r}_k$(정보행렬로 정규화)를 구성하여, 각 공변량별로 $\tilde{r}_k$를 사건시간의 함수(예: $\log t$ 또는 순위)에 회귀시켜 *기울기=0*을 검정합니다. 변수별 p값과 전역(global) p값을 함께 보고, 유의하면 해당 공변량의 PH 위반을 시사합니다.

(보강) `cox.zph`의 변환 선택과 전역(Global) 검정 Schoenfeld(및 scaled) 잔차에 대해 `cox.zph`는 시간 축 변환으로 `"identity"`, `"km"`, `"rank"` 등을 제공한다. 변환은 잔차–시간 관계의 선형화를 돕는 목적이며, 변수별 p값과 함께 **전역(Global) p값**으로 모형 전반의 PH 가정을 평가한다. 여러 변환에서 유사한 결론이 나오는지 함께 확인하면 해석이 안정적이다.

(3) 시간의존 상호작용으로 엄밀 검정(확장 Cox) 이미 기술한 확장 콕스

$$h(t \mid X) = h_0(t) \exp\{\beta X + \delta\, X\, g(t)\}$$

에서 $H_0 : \delta = 0$을 Wald/Score/LRT로 검정합니다. $g(t)$는 $\log t$, t, 구간함수 등으로 두며, 유의한 δ는 X 효과가 시간에 따라 변함을 의미(=PH 위반)합니다.

(4) PH 위반 시 대처 전략

- **층화 Cox**: 문제가 되는 범주형 공변량으로 층화하여 $h_0(t)$를 층별로 분리(해당 변수의 HR은 추정하지 않음).
- **구간별 HR(piecewise)**: t를 구간으로 나눠 $X \times I(t \in$ 구간)로 효과를 시점별로 다르게 허용.
- **시간의존 효과 모델링**: $X \times g(t)$의 함수형을 모형 안에 명시(유도한 δ 포함).
- **모형 변경 고려**: PH 가정이 본질적으로 맞지 않는다면 AFT 등 대안 모형 검토.