Introduction to Survival Analysis

Ming-Yueh Huang

Institute of Statistical Science, Academia Sinica

July 21, 2025

Analysis of Time Duration

- Time-to-event data
- Major variable of interest: Time duration
    - Initial event
    - Failure event
- Compare: Calendar time

Probability Representation

- A random variable $T \geq 0$
- Population vs. sample (data)
- Statistical parameter
  - Cumulative distribution function $F(t) = \mathrm{pr}(T \leq t)$
  - Survival function $S(t) = \mathrm{pr}(T > t) = 1 - F(t)$
  - Hazard function $\mathrm{d}\Lambda(t) = \mathrm{pr}\{T \in (t - \mathrm{d}t, t] \,|\, T \geq t\} = \mathrm{d}F(t)/S(t^-)$

Sampling

- Golden standard: I.I.D. random sample $\{T_i : i = 1, \ldots, n\}$
- Incident sampling
- Prevalent/cross-sectional sampling
- How if we ignore the sampling bias?

Statistical Inference

- Parametric models
  - e.g., $T \sim Exp(\lambda)$, where λ is an unknown parameter.
  - e.g., $\log T \sim Normal(\mu, \sigma^2)$, where (μ, σ) are unknown parameters.
  - Method of moments or maximum likelihood estimation can be directly applied.
- Nonparametric estimation
  - Empirical distribution $\widehat{F}(t) = n^{-1} \sum_{i=1}^{n} 1(T_i \leq t)$.

⋆ Coding Exercise

Right-Censoring

- Lost of follow-up
- C: censoring time
  - $Y = \min(T, C)$
  - $D = 1(T \leq C)$
- Right-censored survival data: $\{(Y_i, D_i) : i = 1, \ldots, n\}$
- How if we ignore the right-censoring?

  ⋆ Coding Exercise

Parametric Models

- $T \sim F(t;\beta)$
- $C \sim F_C(t;\theta)$ and $C \perp\!\!\!\perp T$
- Likelihood function

$$L(\beta,\theta) = \prod_{i=1}^{n} \{f(Y_i;\beta) S_C(Y_i^-;\theta)\}^{D_i} \{f_C(Y_i;\theta) S(Y_i;\beta)\}^{1-D_i}$$
$$= \prod_{i=1}^{n} f(Y_i;\beta)^{D_i} S(Y_i;\beta)^{1-D_i}$$
$$\times \prod_{i=1}^{n} f_C(Y_i;\theta)^{1-D_i} S_C(Y_i^-;\theta)^{D_i}$$

- $\widehat{\theta}_{\mathrm{ML}} = \mathrm{argmax}_\theta \prod_{i=1}^{n} f(Y_i;\beta)^{D_i} S(Y_i;\beta)^{1-D_i}$
  $\star$ Coding Exercise

Semiparametric Models

- The MLE stays the same even if we allow nonparametric $F_C(t)$.
- The parameters become β and $F_C(t)$.
- Hereafter, we assume the independent and non-informative censoring.

Nonparametric Estimation

- Suppose that $T \in \{0 < t_1 < \cdots < t_K < \infty\}$.
- Parameter of interest
  - Probability mass function: $p(t) = \mathrm{pr}(T = t) = \sum_{k=1}^{K} p_k 1(t = t_k)$, where $p_k = \mathrm{pr}(T = t_k)$.
  - $F(t) = \sum_{t_k \le t} p_k$.
  - $S(t) = \sum_{t_k > t} p_k$.
  - Hazard function $\lambda(t) = \mathrm{pr}(T = t \,|\, T \ge t)$ $= \sum_{k=1}^{K} p_k 1(t = t_k) / \sum_{t_k \ge t} p_k \triangleq \sum_{k=1}^{K} \lambda_k 1(t = t_k)$.
  - That is, $\lambda_k = p_k / \sum_{j=k}^{K} p_j = \lambda(t_k)$.

Kaplan-Meier Estimator

- Key idea

$$p_k = \frac{\text{pr}(T \geq t_2)}{\text{pr}(T \geq t_1)} \cdot \frac{\text{pr}(T \geq t_3)}{\text{pr}(T \geq t_2)} \cdots \frac{\text{pr}(T \geq t_k)}{\text{pr}(T \geq t_{k-1})} \cdot \frac{\text{pr}(T = t_k)}{\text{pr}(T \geq t_k)}$$
$$= \prod_{j=1}^{k-1} (1 - \lambda_j)\lambda_k.$$

- Similarly,

$$S(t_k) = \frac{\text{pr}(T \geq t_2)}{\text{pr}(T \geq t_1)} \cdot \frac{\text{pr}(T \geq t_3)}{\text{pr}(T \geq t_2)} \cdots \frac{\text{pr}(T \geq t_k)}{\text{pr}(T \geq t_{k-1})} \cdot \frac{\text{pr}(T > t_k)}{\text{pr}(T \geq t_k)}$$
$$= \prod_{j=1}^{k} (1 - \lambda_j).$$

Kaplan-Meier Estimator

- Consistent plug-in:

$$\lambda_k = \frac{\mathrm{pr}(T = t_k)}{\mathrm{pr}(T \geq t_k)} = \frac{\mathrm{pr}(T = t_k)\mathrm{pr}(C \geq t_k)}{\mathrm{pr}(T \geq t_k)\mathrm{pr}(C \geq t_k)}$$
$$= \frac{\mathrm{pr}(T = t_k, C \geq t_k)}{\mathrm{pr}(T \geq t_k, C \geq t_k)} = \frac{\mathrm{pr}(Y = t_k, D = 1)}{\mathrm{pr}(Y \geq t_k)}.$$

- Thus, we can estimate λ_k by

$$\widehat{\lambda}_k = \frac{n^{-1}\sum_{i=1}^{n} 1(Y_i = t_k, D_i = 1)}{n^{-1}\sum_{i=1}^{n} 1(Y_i \geq t_k)} \triangleq \frac{\widehat{d}_k}{\widehat{r}_k}.$$

- Accordingly, we have

$$\widehat{S}(t_k) = \prod_{j=1}^{k}(1 - \widehat{\lambda}_j) = \prod_{j=1}^{k}\left(1 - \frac{\widehat{d}_j}{\widehat{r}_j}\right).$$

Nonparametric MLE

- The likelihood function is

$$
\begin{aligned}
\text{Likelihood} &= \prod_{i=1}^{n} \{p(Y_i) S_C(Y_i^-)\}^{D_i} \{f_C(Y_i) S(Y_i)\}^{1-D_i} \\
&\propto \prod_{i=1}^{n} p(Y_i)^{D_i} S(Y_i)^{1-D_i} \\
&= \prod_{i=1}^{n} \left\{ \frac{p(Y_i)}{S(Y_i^-)} \right\}^{D_i} \left\{ \frac{S(Y_i^-)}{S(Y_i)} \right\}^{D_i} S(Y_i) \\
&= \prod_{i=1}^{n} \lambda(Y_i)^{D_i} \left\{ \frac{1}{1-\lambda(Y_i)} \right\}^{D_i} \prod_{0 \le t \le Y_i} \{1 - \lambda(t)\} \triangleq \widehat{L}.
\end{aligned}
$$

Nonparametric MLE

- $\partial \log \widehat{L} / \partial \lambda_k = 0$ gives that

$$\sum_{i=1}^{n} \left\{ \frac{D_i 1(t_k = Y_i)}{\lambda_k} + \frac{D_i 1(t_k = Y_i)}{1 - \lambda_k} + \frac{1(t_k \leq Y_i)}{1 - \lambda_k} \right\} = 0.$$

- Thus,

$$\widehat{\lambda}_k = \frac{\sum_{i=1}^{n} D_i 1(Y_i = t_k)}{\sum_{i=1}^{n} 1(Y_i \geq t_k)} = \frac{\widehat{d}_k}{\widehat{r}_k}.$$

- The Kaplan-Meier estimator is still valid even when T is continuous.

⋆ Coding Exercise

Censored Survival Regression

- Let $\mathbf{Z}$ be a vector of covariates.
- Parameter of interest
  - Conditional distribution and survival functions:
    $F(t \,|\, \mathbf{z}) = \mathrm{pr}(T \le t \,|\, \mathbf{Z} = \mathbf{z})$, $S(t \,|\, \mathbf{z}) = 1 - F(t \,|\, \mathbf{z})$.
  - Conditional hazard function:
    $$\Lambda(t \,|\, \mathbf{z}) = \int_0^t \mathrm{d}_u \Lambda(u \,|\, \mathbf{z}) = \int_0^t \frac{\mathrm{d}_u F(u \,|\, \mathbf{z})}{S(u^- \,|\, \mathbf{z})}.$$
  - When $F(t \,|\, \mathbf{z}) = \int_0^t f(u \,|\, \mathbf{z}) \mathrm{d}u$ for a conditional density $f(t \,|\, \mathbf{z})$, then we define $\lambda(t \,|\, \mathbf{z}) = f(t \,|\, \mathbf{z}) / S(t \,|\, \mathbf{z})$.

Parametric Models

- $T \mid \mathbf{Z} = \mathbf{z} \sim F(t \mid \mathbf{z}; \beta)$
- $C \mid \mathbf{Z} = \mathbf{z} \sim F_{C|\mathbf{Z}}(t \mid \mathbf{z}; \theta)$ and $C \perp\!\!\!\perp T \mid \mathbf{Z}$
- Likelihood function

$$
\begin{aligned}
L(\beta, \theta) &= \prod_{i=1}^{n} \{ f(Y_i \mid \mathbf{Z}_i; \beta) S_C(Y_i^- \mid \mathbf{Z}_i; \theta) \}^{D_i} \\
&\qquad \times \{ f_C(Y_i \mid \mathbf{Z}_i; \theta) S(Y_i \mid \mathbf{Z}_i; \beta) \}^{1-D_i} \\
&= \prod_{i=1}^{n} f(Y_i \mid \mathbf{Z}_i; \beta)^{D_i} S(Y_i \mid \mathbf{Z}_i; \beta)^{1-D_i} \\
&\qquad \times \prod_{i=1}^{n} f_C(Y_i \mid \mathbf{Z}_i; \theta)^{1-D_i} S_C(Y_i^- \mid \mathbf{Z}_i; \theta)^{D_i}
\end{aligned}
$$

- $\widehat{\theta}_{\mathrm{ML}} = \mathrm{argmax}_\theta \prod_{i=1}^{n} f(Y_i \mid \mathbf{Z}_i; \beta)^{D_i} S(Y_i \mid \mathbf{Z}_i; \beta)^{1-D_i}$

Commonly-Used Models

- $T \mid \mathbf{Z} = \mathbf{z} \sim Exp\{\lambda \exp(\mathbf{z}^\intercal \beta)\}$.
  - $\mathrm{E}(T \mid \mathbf{Z} = \mathbf{z}) = \lambda^{-1} \exp(-\mathbf{z}^\intercal \beta)$
  - $\log \mathrm{E}(T \mid \mathbf{Z} = \mathbf{z}) = -\log \lambda - \mathbf{z}^\intercal \beta = \alpha^* + \mathbf{z}^\intercal \beta^*$.
- $T \mid \mathbf{Z} = \mathbf{z} \sim Weibull\{\lambda \exp(\mathbf{z}^\intercal \beta), \gamma\}$.
  - $\mathrm{E}(T \mid \mathbf{Z} = \mathbf{z}) = \lambda^{-1} \exp(-\mathbf{z}^\intercal \beta)$
  - $\log \mathrm{E}(T \mid \mathbf{Z} = \mathbf{z}) = -\log \lambda - \mathbf{z}^\intercal \beta = \alpha^* + \mathbf{z}^\intercal \beta^*$.
  - γ appears in the conditional variance.

$\star$ Coding Exercise