PH 716 Applied Survival Analysis

Part I: Basic quantities of survial models

Zhiyang Zhou (zhou67@uwm.edu, zhiyanggeezhou.github.io)

2024/Jan/24 18:21:36

"All models are wrong, but some are useful."

— G. E. P. Box. (1976). *Journal of the American Statistical Association*, 71:791–799.

What is a statistical model?

- The (joint) distribution of RV(s) of interest
  - Reformulate linear regression and logit regression models
  - Parametric vs non-parametric vs semi-parametric

Statistical modelling

Random variable (RV) of interest → data collection and cleaning → model specification → model fitting and inference → interpretation

Motivating examples for survial models

- ([DM] Example 1.2) This is a Phase II clinical trial of Xeloda and oxaliplatin (XELOX) chemotherapy given before surgery to 48 advanced gastric cancer patients with paraaortic lymph node metastasis (Wang et al., 2014). An important survival outcome of interest is progression-free survival, which is the time from entry into the clinical trial until progression or death, whichever comes first.

```
head(asaur::gastricXelox)
```

```
##   timeWeeks delta
## 1         4     1
## 2         8     1
## 3         8     1
## 4         8     1
## 5         9     1
## 6        11     1
```

- ([DM] Example 1.5) The purpose of Steinberg et al. (2009) was to evaluate extended duration of a triple-medication combination versus therapy with the nicotine patch alone in smokers with medical illnesses.

```
head(asaur::pharmacoSmoking[, 2:8])
```

```
##   ttr relapse         grp age gender     race employment
## 1 182       0   patchOnly  36   Male    white         ft
## 2  14       1   patchOnly  41   Male    white      other
## 3   5       1 combination  25 Female    white      other
```

```
## 4  16       1 combination  54   Male    white         ft
## 5   0       1 combination  45   Male    white      other
## 6 182       0 combination  43   Male hispanic         ft
```

- The event of interest may be, e.g.,
  - death, cancer incidence, disease remission
  - HIV infection, AIDS onset
  - organ rejection, transplant failure
  - bankruptcy, economic recovery

Basic quantities of survival models

Recall random variables

- An RV is a real-valued function.
- The cumulative distribution function (cdf) of X, say F_X, is defined as

$$F_X(x) = \Pr(X \leq x), \quad x \in \mathbb{R}.$$

  - F_X satisfies following three properties:
    - (right continuous) $\lim_{x \to x_0^+} F_X(x) = F_X(x_0)$;
    - (non-decreasing) $F_X(x_1) \leq F_X(x_2)$ for $x_1 \leq x_2$;
    - (ranging from 0 to 1) $F_X(-\infty) = 0$ and $F_X(\infty) = 1$.
  - Reversely, any function satisfying the three properties must be a cdf for certain RV.
- Discrete RV
  - RV X merely outputs countable different values
  - Probability mass function (pmf): $p_X(x) = \Pr(X = x)$
- Continuous RV
  - RV X is continuous iff its cdf F_X is (absolutely) continuous, i.e., there exists f_X such that

$$F_X(x) = \int_{-\infty}^{x} f_X(u)\mathrm{d}u, \quad \forall x \in \mathbb{R}.$$

  - Probability density function (pdf): $f_X(x) = \mathrm{d}F_X(x)/\mathrm{d}x = \lim_{\delta \to 0^+} \frac{\Pr(x < X \leq x+\delta)}{\delta}$.

Survival function

- T (≥ 0): survival time or, more specifically, the time elapsed from time 0 to the event of interest (abbr. time to event); the RV of interest in survival analysis
  - The time origin, 0, often represents
    - the time point when a disease was first detected,
    - the initiation time of therapy/procedure, or
    - certain meaningful time point
  - E.g.,
    - In a clinical trial: time 0 might be the time of randomization;
    - Among heart transplant patients: time 0 is the end of transplantation
- $S_T(t) = \Pr(T > t) = 1 - F_T(t)$, $t \geq 0$:
  - The probability that the event of interest has not yet occurred by time t
  - (right continuous) $\lim_{t \to t_0^+} S_T(t) = S_T(t_0)$;
  - (non-increasing) $S_T(t_1) \geq S_T(t_2)$ for $t_1 \leq t_2$;
  - (ranging from 0 to 1) $S_T(0) = 1$ and $S_T(\infty) = 0$.

Hazard function (aka. hazard rate)

- $\lambda_T(t) \geq 0$, reflecting how dangerous at time t for a person who survives up to time t
- If T is a continuous RV (i.e., the event may occur at any time point in $[0, \infty)$)
  - $\lambda_T(t) = \lim_{\delta \to 0^+} \frac{\Pr(t \leq T < t+\delta | T \geq t)}{\delta} \overset{Why?}{=} f_T(t)/S_T(t)$
  - Conditional failure rate, not a probability;
- If T is a discrete RV (i.e., the event may only occur at $0 = t_0 < t_1 < t_2 < \cdots$),
  - $\lambda_T(t_j) = \Pr(T = t_j \mid T \geq t_j) = p_T(t_j)/S_T(t_{j-1})$ and zero elsewhere, $j = 1, 2, \ldots$.
    - p_T is the pmf of T

- Ex. 1.1 ([JM] Ex. 2.3): Suppose that T has pmf $p_T(t_j) = 1/3$, $j = 1, 2, 3$. Find the survival and hazard functions.
- Ans. to Ex. 1.1: The survival function is

$$S_T(t) = \Pr(T > t) = \sum_{j: t_j > t} p_T(t_j) = \begin{cases} 1 & \text{if } 0 \leq t < t_1 \\ 2/3 & \text{if } t_1 \leq t < t_2 \\ 1/3 & \text{if } t_2 \leq t < t_3 \\ 0 & \text{if } t \geq t_3. \end{cases}$$

  It is a right-continuous descending step function. The hazard function is then

$$\lambda(t_j) = \frac{p_T(t_j)}{S_T(t_{j-1})} = \begin{cases} 1/3 & j = 1 \\ 1/2 & j = 2 \\ 1 & j = 3 \\ 0 & \text{elsewhere.} \end{cases}$$

Cumulative hazard function

- $\Lambda_T(t) = \int_0^t \lambda_T(s) \mathrm{d}s$ if T is continuous
- $\Lambda_T(t) = \sum_{t_j \leq t} \lambda_T(t)$ if T is discrete

---

- Ex. 1.2 (exponential distribution): Suppose that T has pdf $f_T(t) = \lambda \exp(-\lambda t)$, $\lambda > 0$. Find the survival, hazard and cumulative hazard functions.
  - Memorylessness: constant hazard rate $\Rightarrow \Pr(T > t + t_0 \mid T > t_0) = \Pr(T > t)$ i.e., the survial probability would not be influenced by the history.
  - This is the only distribution with constant hazard rate.
  - Not often used to model human subjects unless followed for relatively short duration; more often used in engineering or for illustrative purposes.
- Ans. to Ex. 1.2: $F_T(t) = 1 - \exp(-\lambda t), t > 0 \Rightarrow S_T(t) = \exp(-\lambda t)$ if $t > 0$ and 1 otherwise. So, $\lambda_T(t) = \lambda$ and $\Lambda_T(t) = \lambda t$.

- Ex. 1.3 (Weibull distribution): Find survival, hazard and cumulative hazard functions for the following three equivalent parameterizations of Weibull pdf: for $t \geq 0$,
  - $f_T(t) = (k/\lambda)(t/\lambda)^{k-1} e^{-(t/\lambda)^k}$ (used in R with k as "shape" and λ as "scale"),
  - $f_T(t) = bkt^{k-1} e^{-bt^k}$ (i.e., $b = \lambda^{-k}$),
  - $f_T(t) = \beta k(\beta t)^{k-1} e^{-(\beta t)^k}$ (i.e., $\beta = \lambda^{-1}$).
- For the Weibull distribution with any of the three parameterizations
  - $k > 1$: monotone increasing hazard rate
  - $k < 1$: monotone decreasing hazard rate

  - $k = 1$: reducing to the exponential distribution
- Ans. to Ex. 1.3:
  - $F_T(t) = 1 - \exp\{-(t/\lambda)^k\}$ for $t \geq 0$ and zero elsewhere
  - $F_T(t) = 1 - \exp(-bt^k)$ for $t \geq 0$ and zero elsewhere
  - $F_T(t) = 1 - \exp\{-(\beta t)^k\}$ for $t \geq 0$ and zero elsewhere

More identities

- When T is continuous
  - $\lambda_T(t) = -\mathrm{d}\{\ln S_T(t)\}/\mathrm{d}t$ (Why?)
  - $\Lambda_T(t) = -\ln S_T(t)$
  - $S_T(t) = \exp\{-\Lambda_T(t)\}$
  - $f_T(t) = S_T(t)\lambda_T(t)$
- Proof of $\lambda_T(t) = -\mathrm{d}\{\ln S_T(t)\}/\mathrm{d}t$ for continuous T: $\lambda_T(t) = f_T(t)/S_T(t) = \{\mathrm{d}F_T(t)/\mathrm{d}t\}/S_T(t) = [\mathrm{d}\{1 - S_T(t)\}/\mathrm{d}t]/S_T(t) = -\mathrm{d}\{\ln S_T(t)\}/\mathrm{d}t$
  - The last equality holds due to the chain rule
- When T is discrete
  - $\lambda_T(t_j) = 1 - S(t_j)/S_T(t_{j-1})$ (Why?)
  - If $t_J \leq t < t_{J+1}$ then $S_T(t) = \prod_{j:t_j \leq t}\{1 - \lambda_T(t_j)\}$ (Why?),
    - Noting that $1 - \lambda_T(t_j) = S(t_j)/S_T(t_{j-1})$
- So, knowing one of f_T, S_T, λ_T and Λ_T is sufficient to confirm the distribution of T

Other quantities

- Mean lifetime (or life expectancy)
  - When T is continuous, $\mathrm{E}(T) = \int_0^\infty t f_T(t)\mathrm{d}t = \int_0^\infty S_T(t)\mathrm{d}t$ (Why?)
    - The area under the survival curve $S_T(t)$
    - Proof: $\mathrm{E}(T) = -\int_0^\infty tS'(t)\mathrm{d}t = -\int_0^\infty tS'(t)dt = -tS(t)\Big|_0^\infty + \int_0^\infty S(t)dt$ (integration by parts)
  - When T is discrete, $\mathrm{E}(T) = \sum_j t_j p_T(t_j)$
- Median lifetime: the smallest t such that $S_T(t) \leq .5$
  - The earliest time at which half the cohort will have died
- Mean residual life (for continuous T): $r(t) = \mathrm{E}(T - t \mid T \geq t) = \int_t^\infty (s - t) f_T(s)\mathrm{d}s/S_T(t)$
  - The average additional lifetime for an individual alive at time t
  - $r(0) = \mathrm{E}(T)$

Censoring

- T is potentially unobserved in full for all the subjects
  - Subjects whose survival times are unobserved during follow-up are said to be censored.
- Cause of censoring
  - Administrative censoring
    - Study ends before the event of interest has occurred
    - Usually independent of survival time
  - Withdrawal from study
    - E.g.,
      - a patient drops out of clinical trial because he/she is too sick to participate
      - a subject discontinues participation in trial because her symptoms have subsided
    - Depending on the censoring
- Types of censoring
  - Right-censoring (by default here):
    - For a censored subject, T is not known exactly, but $\geq$ the censoring time;
    - i.e., there is a lower bound for the event time.
  - Left-censoring

        * For a censored subject there is a upper bound for the event time
        * E.g.,
            · ([DM] Example 3.4) "In early childhood learning centers, interest often focuses upon testing children to determine when a child learns to accomplish certain specified tasks. The age at which a child learns the task would be considered the time-to-event. Often, some children can already perform the task when they start in the study. Such event times are considered left censored."
    - Interval-censoring
        * For a censored subject there are both lower and upper bounds for the event time
        * E.g.,
            · ([DM] Example 3.5) "In the Framingham Heart Study, the ages of first occurrence of the subcategory angina pectoris may be known only to be between two clinical examinations, approximately two years apart (Odell et al., 1992). Such observations would be interval-censored."
- Noninformative censoring (by default here)
    - The censoring time is independent of survival time