Causal Inference

Chapter 17. Causal Survival Analysis

Yuha Park

August 19, 2021

Seoul National University

Contents

Hazards and risks

Definition

- **Survival analysis** : outcome is time to an event of interest that can occur at any time after the start of follow-up
- **Administrative censoring time** : difference between date of administrative end of follow-up and date at which follow-up begins
- **Censoring in survival analysis** : administrative censoring, loss to follow-up, competing events, and etc.

Measures

Measures that can accommodate administrative censoring and are functions of the survival time T are defined as:

- **survival probability** $P[T > k]$: the proportion of individuals who survived through time k
- **risk** $1 - P[T > k] = P[T \leq k]$: cumulative incidence at time k which is given by one minus the survival probability
- **hazard** $P[T = k | T > k - 1]$: the proportion of individuals at time k who develop the event among those who had not developed it before k

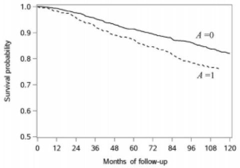

Figure 17.1

Two main ways to arrange analytic dataset

- **Long or wide data format** : each row of the database corresponds to one person (1 row per individual)

- **Person-time data format** : each row of the database corresponds to a person-time (1 row per person-month)
  - D_k: a time-varying indicator (outcome variable) of event for each person at each month k

$$D_k = \begin{cases} 1, & \text{if } T \leq k, \\ 0, & \text{if } T > k. \end{cases}$$

  - $P[D_k = 0]$: survival at k ($= P[T > k]$)
  - $P[D_k = 1]$: risk at k ($= P[T \leq k]$)
  - $P[D_k = 1 | D_{k-1} = 0]$: hazard at k ($= P[T = k | T > k - 1]$)

Person-time data format example

id	k	D_{k+1}	A	L
1	0	$D_1 = 0$	.	.
1	1	$D_2 = 0$	.	.
1	2	$D_3 = 0$	.	.
		$\vdots$		
1	119	$D_{120} = 1$	.	.
2	0	$D_1 = 0$	.	.
2	1	$D_2 = 0$	.	.
		$\vdots$		
2	69	$D_{70} = 1$	.	.
3	0	$D_1 = 0$	.	.
		$\vdots$		

How to estimate measures in person-time data format

The survival probability at k equals the product of one minus the hazard at all previous times.

$$P[D_k = 0] = \prod_{m=1}^{k} P[D_m = 0 | D_{m-1} = 0]$$

- **nonparametric** estimator of hazard at k: Kaplan-Meier or product-limit estimator
- **parametric** estimator of hazard at k: to fit a logistic regression model for $P[D_{k+1} = 1 | D_k = 0]$ at each k

How to estimate the survival curve

What to estimate the survival curve

Our goal is to estimate the survival probabilities $P[D_{k+1} = 0 | A = a]$.

Suppose that individuals start **the follow-up at different dates** but the administrative end of follow-up (AEOF) date is common to all.

- individuals have different administrative censoring times
- C_k: a time-varying indicator for censoring by time k

$$C_k = \begin{cases} 0, & \text{if } AEOF \text{ is greater than } k, \\ 1, & \text{otherwise.} \end{cases}$$

- $P[D_k^{\bar{c}=\bar{0}} = 0 | A = a]$: the survival that would have been observed if the value of the time-varying indicators D_k were known even after censoring where $\bar{c} = (c_1, c_2, \cdots, c_{k_{end}})$

IP weighting of marginal structural models

Suppose we want to compare the counterfactual survivals $P[D_{k+1}^{a=1} = 0]$ and $P[D_{k+1}^{a=0} = 0]$ for $k = 0, 1, \cdots, k_{end} - 1$.

- $D_k^{a,\bar{c}:=\bar{0}} = D_k^a$: a counterfactual time-varying indicator for death at k under treatment level a and no censoring
- Because of confounding, this contrast may not be validly estimated by the contrast of the survivals $P[D_{k+1} = 0|A = 1]$ and $P[D_{k+1} = 0|A = 0]$
- A valid estimation of the quantities $P[D_{k+1}^a = 0]$ for $a = 1$ and $a = 0$ requires adjustment for confounders using **IP weighting**

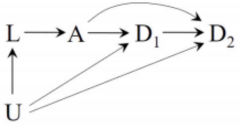

FIGURE 17.5

IP weighting of marginal structural models

The estimation of IP weighted survival curves has two steps under the assumption of exchangeability, positivity, and consistency:

(1) we estimate the stabilized IP weight $SW^A = P(A)/p(A|L)$ for each individual in study population

- The application of the estimated weights SW^A creates a pseudo-population

(2) using the person-time data format, we fit a hazards model like the one described above except that individuals are weighted by their estimated SW_A

- The estimates of $P[D^a_{k+1} = 0 | D^a_k = 0]$ from the IP weighted hazards models can be multiplied over time to obtain an estimate of the survival $P[D^a_{k+1} = 0]$

The parametric g-formula

Under exchangeability, positivity, and consistency,
the survival $P[D^a_{k+1} = 0]$ equals the standardized survival

$$\sum_l P[D_{k+1} = 0 | L = l, A = a] P[L = l].$$

Note that the survival curves estimated via **IP weighting** and **the parametric g-formula** are similar but not identical because they rely on different parametric assumptions.

Accelerated failure time (AFT) model

- T_i^a: the counterfactual time of survival for individual i under treatment level a
- $T_i^{a=1}/T_i^{a=0}$ (survival time ratio): counterfactual survival times under treatment and under no treatment
  - If the ratio > 1, then treatment is beneficial
  - if the ratio < 1, then treatment is harmful
  - if the ratio = 1, then treatment has no effect

Accelerated failure time (AFT) model

A structural nested AFT model:

$$T_i^a / T_i^{a=0} = exp(-\psi_1 a),$$

where ψ_1 measures the expansion of each individual's survival time attributable to treatment.

- If $\psi_1 < 0$, then treatment increases survival time
- If $\psi_1 > 0$, then treatment decreases survival time
- If $\psi_1 = 0$, then treatment does not affect survival time