Notes: Computer Age Statistical Inference – Ch 9 Survival Analysis

Yingbo Li

06/13/2020

Table of Contents

Survival Analysis

- Life Table and Kaplan-Meier Estimate
- Cox's Proportional Hazards Model

Life table

- An insurance company's life table shows information of clients by their age. For each age i, it contains
  - n_i: number of clients
  - y_i: number of death
  - $\hat{h}_i = y_i / n_i$: hazard rate
  - $\hat{S}_i$: survival probability estimate
- An example life table

Age	n	y	$\hat{h}$	$\hat{S}$
34	120	0	0.000	1.000
35	71	1	0.014	0.986
36	125	0	0.000	0.986
...	...	...	...	...

Discrete survival analysis: notations

- A client's lifetime (time until event): random variable X
  - Also called failure time, survival time, or event time
- Probability of dying at age i

$$f_i = P(X = i)$$

- Probability of surviving past age i

$$S_i = \sum_{j \geq i+1} f_j = P(X > i)$$

- Hazard rate at age i: conditional probability

$$h_i = \frac{f_i}{S_{i-1}} = P(X = i \mid X \geq i)$$

Life table estimations

- Hazard rate estimation: binomial proportions

$$\hat{h}_i = \frac{y_i}{n_i}$$

  - Typical frequentist inference: probabilistic results h_i is estimated by the plug-in principle

- Probability of surviving past age j given survival past age i:

$$P(X > j \mid X > i) = \prod_{k=i+1}^{j} P(X > k \mid X \geq k) = \prod_{k=i+1}^{j} (1 - h_k)$$

- Probability of survival estimation

$$\hat{S}_j = \prod_{k=i_0}^{j} \left(1 - \hat{h}_k\right)$$

  where i_0 is the starting age

Continuous survival analysis: notations

- Time until event T: a continuous positive random variable, with pdf $f(t)$ and cdf $F(t)$
- Survival function (i.e., reverse cdf)

$$S(t) = \int_t^\infty f(x)dx = P(T > t) = 1 - F(t)$$

- Hazard rate, also called hazard function

$$h(t) = \frac{f(t)}{S(t)} = \lim_{\Delta t \to 0} \frac{P(t < T \le t + \Delta t \mid T > t)}{\Delta t}$$

  - In some other books, hazard rate is denoted as $\lambda(t)$

Hazard rate and cumulative hazard function

- Connection between hazard rate $h(t)$ and survival function $S(t)$

$$h(t) = -\frac{\partial \log S(t)}{\partial t} \quad \Longleftrightarrow \quad S(t) = \exp\left\{-\int_0^t h(x)dx\right\}$$

- Cumulative hazard function

$$\Lambda(t) = \int_0^t h(x)dx = -\log S(t)$$

- Knowing any of $S(t)$, $h(t)$, $\Lambda(t)$ allows one to derive the other two
- Example: exponential distributed T

$$f(t) = \lambda e^{-\lambda t} \quad \Longrightarrow \quad S(t) = e^{-\lambda t}, \quad h(t) = \lambda$$

  – Constant hazard rate: menoryless

Censored data

- Censored data: survival times known only to exceed the reported value
  - E.g., lost to followup, experiment ended with some patients still alive
  - Usually denoted as “number+”
- Observation z_i for censored data:

$$z = (t_i, d_i),$$

where t_i is the survival time, and d_i is the indicator

$$d_i = \begin{cases} 1 & \text{if death observed} \\ 0 & \text{if death not observed} \end{cases}$$

Kaplan-Meier estimate

- Among the censored data $z_1, \ldots, z_n$, we denote the ordered survival times as
$$t_{(1)} < t_{(2)} < \ldots < t_{(n)},$$
assuming no ties.

- The Kaplan-Meier estimate for survival probability $S_{(j)} = P(X > t_{(j)})$ is the life table estimate
$$\hat{S}_{(j)} = \prod_{k \leq j} \left(\frac{n-k}{n-k+1} \right)^{d_{(k)}}$$

- Life table curves are nonparametric: no relationship is assumed between the hazard rates h_i

A parametric approach

- Death counts y_k are independent Binomials

$$y_k \overset{ind}{\sim} \mathsf{B}(n_k, h_k)$$

- Logistic regression

$$log\left(\frac{h_k}{1-h_k}\right) = \boldsymbol{\alpha}\mathbf{x}_k$$

  – E.g., cubic regression:

$$x_k = (1, k, k^2, k^3)'$$

  – E.g., cubic-linear spline:

$$x_k = (1, k, (k-k_0)_-^2, (k-k_0)_-^3)'$$

  where $x_- = x \cdot \mathbf{1}_{x \leq 0}$

Cox's proportional hazards model

- Proportional hazards model assumes

$$h_i(t) = h_0(t) \cdot e^{\mathbf{x}_i' \boldsymbol{\beta}},$$

  where $h_0(t)$ is a baseline hazard, which we don't need to specify

- Denote $\theta_i = e^{\mathbf{x}_i' \boldsymbol{\beta}}$, then

$$S_i(t) = S_0(t)^{\theta_i},$$

  where $S_0(t)$ is the baseline survival function

  – Larger value of θ_i indicates more quickly declining (i.e., worse) survival curves
  – Positive value of the coefficient β_j indicates increase of the corresponding covariate x_j associating with worse survival curves

Proportional hazards model: key results

- Let J be the number of observed deaths, occurring at times

$$T_{(1)} < T_{(2)} < \ldots < T_{(J)}$$

  assuming no ties

- Just before time $T_{(j)}$ there is a risk set of individuals still under observation

$$R_j = \{i, t_i \geq T_{(j)}\}$$

- Key results of the proportional hazards model: given one person dies at time $T_{(j)}$, the probablity it is person i, among the set of people at risk, is

$$P(i_j = i \mid R_j) = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{\sum_{k \in R_j} e^{\mathbf{x}_j'\boldsymbol{\beta}}} = \frac{\theta_i}{\sum_{k \in R_j} \theta_j}$$

Parameter estimation: based on the partial likelihood

- Estimaiton of β is to maximize the partial likelihood

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{J} \frac{e^{\mathbf{x}'_{i_j}\boldsymbol{\beta}}}{\sum_{k \in R_j} e^{\mathbf{x}'_j \boldsymbol{\beta}}}$$

where individual i_j dies at time $T_{(j)}$

- Semi-parametric: we do not need to specify the baseline $h_0(t)$, since it is not contained in the objective function

References

- Efron, Bradley and Hastie, Trevor (2016), *Computer Age Statistical Inference*. Cambridge University Press