STAT347: Generalized Linear Models
Lecture 14

Today's topics: Survival analysis

- Examples of survival analysis datasets
- Basic concepts in survival analysis: survival function, hazard rate, censoring
- Kaplan-Meier estimator of the survival function

Three references for the survival analysis:

- Lecture notes (`http://www.math.ucsd.edu/~rxu/math284/`) on survival analysis from Ronghui Xu at UCSD. We focus the main ideas from her Lecture 1-5 (slect1.pdf - slect5.pdf).
- Textbook *Computer Age Statistical Inference* (`https://web.stanford.edu/~hastie/CASI/`) Chapters 9.1-9.4
- Brad Efron's GLM notes Part 3 3.6 - 3.7 (`http://statweb.stanford.edu/~ckirby/brad/STATS305B_Part-3_corrected-2.pdf`)

1 Examples of survival analysis

1.1 NCOG study

A randomized clinical trial conducted by the Northern California Oncology Group (NCOG) compared two treatments for head and neck cancer: chemotherapy (Arm A of the trial, $n = 51$ patients) and chemotherapy plus radiation (Arm B, $n = 45$ patients). The data records the survival time in number of days past treatment for each patient. The numbers followed by + patients still alive on their final day of observation. For example, the sixth patient in Arm A was alive on day 74 after his treatment, and then "lost to follow-up"; we only know that his survival time exceeded 74 days.

Arm A:

7	34	42	63	64	74+	83	84	91	108	112
129	133	133	139	140	140	146	149	154	157	160
160	165	173	176	185+	218	225	241	248	273	277
279+	297	319+	405	417	420	440	523	523+	583	594
1101	1116+	1146	1226+	1349+	1412+	1417				

Arm B:

37	84	92	94	110	112	119	127	130	133	140
146	155	159	169+	173	179	194	195	209	249	281
319	339	432	469	519	528+	547+	613+	633	725	759+
817	1092+	1245+	1331+	1557	1642+	1771+	1776	1897+	2023+	2146+
2297+										

- The main question: is the Arm B is more effective treatment than Art A?

- Instead of just compare the mean survival time, we would like to know more information about the survival time distribution (the survival curve)
- How to deal with "lost to follow-up" (censoring) ?

1.2 Duration of nursing home stay

The National Center for Health Services Research studied 36 for-profit nursing homes to assess the effects of different financial incentives on length of stay. "Treated" nursing homes received higher per diems for Medicaid patients, and bonuses for improving a patient's health and sending them home. Study included 1601 patients admitted between May 1, 1981 and April 30, 1982.

Variables include:

- LOS - Length of stay of a resident (in days)
- AGE - Age of a resident
- RX - Nursing home assignment (1:bonuses, 0:no bonuses)
- GENDER - Gender (1:male, 0:female)
- MARRIED - (1: married, 0:not married)
- HEALTH - health status (2:second best - 5:worst)
- CENSOR - Censoring indicator (1:censored, 0:discharged)

Question: How do we find the treatment effect on stay length after adjusting for other covariates and censoring?

2 Basic concepts

- Survival time: T is a random non-negative variable, the duration from the start of treatment to death.
  - Continuous: T has a density function $f(t)$
  - Discrete: $T \in \{0, 1, 2, 3, \cdots\}$, $f_i = P(T = i)$
- Survival function/curve: $S(t) = P(T \geq t)$
  - Continuous: $S(t) = \int_t^\infty f(t')dt'$
  - Discrete: $S_i = \sum_{j \geq i} f_j$
- Hazard rate/function: $h(t) = f(t)/S(t)$ (or $h_i = f_i/S_i$ for discrete T)
- Accumulative hazard function: $H(t) = \int_0^t h(t)$ (or $H_i = \sum_{j \leq i} h_j$ for discrete T)

The survive function and hazard rate provide more information than $E(T)$.

An important fact is that knowing one of the three functions of $H(t)$, $h(t)$ and $S(t)$ will enable inferring the other two functions.

Continuous case (homework):

$$S(t) = e^{H(t)}$$

Discrete case:

$$S_i = \prod_{j=0}^{i-1} P[T \geq j+1 \mid T \geq j] = \prod_{j=0}^{i-1}(1-h_j)$$

2.1 Censoring

For n samples, denote their survival time as $T_1, T_2, \cdots, T_n$. However, we may not be able to observe every T_i. Censoring can occur when

- When the study ends, some individual have not had the event yet (still alive)
- Some individuals dropout or get lost in the middle of the study.

Typically, individuals do not enter the study at the same time, but it is usually not a concern as T_i is the length of the observation time (can treat the starting time as a covariate to adjust for its possible effect).

A graphical representation of the data with censoring (in class)

Denote each sample's censoring time as $C_1, C_2, \cdots, C_n$. Then what we can actually observe for each sample are $Y_i = \min(T_i, C_i)$ and an indicator of whether censoring occurs:

$$\delta_i = \begin{cases} 1 & \text{if } T_i \leq C_i \text{ (observed death)} \\ 0 & \text{Otherwise} \end{cases}$$

When each sample also has its covariate, what we observe can be denoted as (Y_i, X_i, δ_i) for $i = 1, 2, \cdots, n$.

Throughout the class, we only consider **non-informative censoring**, which is basically requiring

$$T_i \perp C_i \mid X_i$$

which means that the censoring time is not associated with the survival time, at least conditioning on other known covariates X_i.

3 Estimating the survival function

In this section we consider the scenario when there is no observed covariates X_i and **the survival time T_i are i.i.d.**

3.1 Non-parametric approach

When there is no censoring, then the survival function $S(t)$ is a transformation of the cdf, thus we can estimate it by the empirical cdf function.

$$\widehat{S}_n(t) = \frac{1}{n}\sum_i 1_{T_i \geq t}$$

Example: For 20 samples, the survival times are 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

Properties of $\widehat{S}_n(t)$: as $1_{T_i \geq t} \sim \text{Bernoulli}(S(t))$, so that

- $\widehat{S}_n(t)$ converges in probability to $S(t)$ (consistency);
- $\sqrt{n}\left(\widehat{S}_n(t) - S(t)\right) \to N(0, S(t)[1 - S(t)])$ in distribution.

However, when there is censoring this method does not work Example: the survival times are 1, 1, 2, 2+, 3+, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

We don't know how to estimate $S(3)$ from the empirical cdf approach in this example. There is a clever way to do this.

3.1.1 Kaplan-Meier estimator

For the discrete survival time, or we can discretize the survival time into bins. For each bin i or discrete survival time i, assume we observe r_i samples that are still alive at the beginning of this time bin, d_i death during this time bin and c_i drop-outs at the end of this time bin (we assume no drop-outs during this time bin).

At the presence of non-informative censoring, the r_i samples are i.i.d. at this time point and we have

$$d_i \sim \text{Bernoulli}(r_i, h_i)$$

thus an unbiased estimate of h_i is

$$\widehat{h}_i = \frac{d_i}{r_i}.$$

The estimate of S_i will be

$$\widehat{S}_i = \prod_{j \le (i-1)} (1 - \widehat{h}_j)$$

For continuous survival time, the bin can be smaller and smaller, and we get the Kaplan-Meier estimator as

$$\widehat{S}(t) = \prod_{j:\tau_j \le t} \frac{r_j - d_j}{r_j}$$

where $\{\tau_1, \tau_2, \cdots \tau_K\}$ is the set of K distinct uncensored failure times observed in the sample, d_j is the number of death at τ_j and r_j is the total number of people who are at risk right before τ_j.

Each $r_{i+1} = r_i - d_i - c_i$, so $\hat{h}_1, \cdots, \hat{h}_n$ are not independent. How do we estimate $\text{Var}(\widehat{S}(t))$?

The Greenwood formula for estimating the uncertainty in $\widehat{S}(t)$:

$$\log \widehat{S}(t) = \sum_{j:\tau_j \le t} \log(1 - \hat{h}_j)$$

Using the Delta method

$$\begin{aligned}\log \widehat{S}(t) &\approx \sum_{j:\tau_j \le t} \left[\log(1 - h_j) - \frac{1}{1 - h_j}(\hat{h}_j - h_j)\right] \\ &= \text{Const} - \sum_{j:\tau_j \le t} \frac{1}{1 - h_j}(\hat{h}_j - h_j)\end{aligned}$$

Though $\hat{h}_1, \cdots, \hat{h}_n$ are not independent, $\mathbb{E}(\hat{h}_j - h_j \mid h_1, \cdots, h_{j-1}) = 0$ (the partial sums form a martingale). Thus, when calculating the variance, we can "treat" $\hat{h}_j$ as independent.

$$\begin{aligned}\widehat{\mathrm{Var}}\left(\log \widehat{S}(t)\right) &\approx \sum_{j:\tau_j \le t} \left(\frac{1}{1-\widehat{h}_j}\right)^2 \widehat{\mathrm{Var}}(\widehat{h}_j) \\ &= \sum_{j:\tau_j \le t} \frac{\widehat{h}_j}{(1-\widehat{h}_j) r_j} = \sum_{j:\tau_j \le t} \frac{d_j}{(r_j - d_j) r_j}\end{aligned}$$

Using Delta method on $\widehat{S}(t) = e^{\log \widehat{S}(t)}$, we get

$$\begin{aligned}\widehat{\mathrm{Var}}\left(\widehat{S}(t)\right) &= [\widehat{S}(t)]^2 \widehat{\mathrm{Var}}\left(\log(\widehat{S}(t))\right) \\ &= [\widehat{S}(t)]^2 \sum_{j:\tau_j \le t} \frac{d_j}{(r_j - d_j) r_j}\end{aligned}$$