Lecture 13: Survival Analysis & Censored Data

STATS 202: Statistical Learning and Data Science

Linh Tran

Department of Statistics
Stanford University

August 6, 2025

Announcements

- HW4 should be in
- Final predictions due in 4 days (write-up is due in 1 week).
  - **reference your Kaggle leaderboard name on Page 1**
- Final exam is next Saturday
  - Time: Saturday August 16 7:00 PM - 10:00 PM
  - Location: 300-300
  - Practice exam released this Friday (solutions next week)
  - Accommodation requests should already be made
- Course evaluation starts on Aug 11 (on Canvas).

Outline

- Time to event
- Censored data
- Kaplan Meier Curves
- Proportional hazards models
- Time varying covariates

Time to event

Typically used for non-negative random variables $T \geq 0$, e.g.

- Time until person dies
- Time until student graduates
- Number of clicks until customer buys something
- Number of sexual encounters before catching AIDS

Time to event

Requirements for time to event:

1. The intiating event (i.e. time 0)
2. The terminating event (i.e. outcome of interest)
3. A unit of "time"

Time to event

What to do with our random variable T

1. Estimate the probabilty density function (pdf) $f(t)$
2. Estimate the culmulative distribution function (cdf) $F(t)$
3. Estimate the survival function $S(t) = 1 - F(t)$
4. Estimate the hazard function $h(t) = \frac{f(t)}{S(t)}$

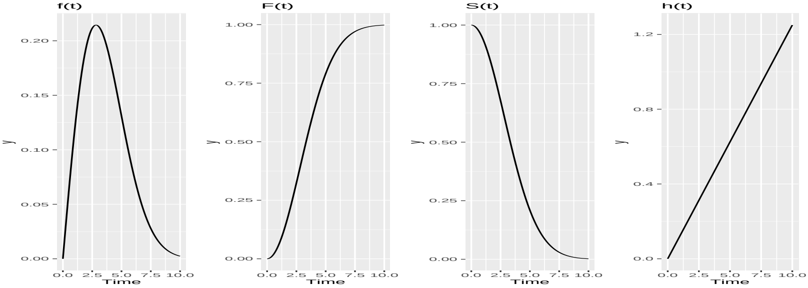
Another way of expressing the hazard function

$$h(t) = \lim_{\Delta_t \to 0} \frac{P(t \leq T \leq t + \Delta_t | T \geq t)}{\Delta_t}$$

n.b. We can also estimate the *cumulative hazard* $\Lambda(t) = -\log S(t)$, or equivalently $S(t) = \exp(-\Lambda(t))$

Example: Applying MLE in a parametric model, e.g. the Weibull distribution.

$$L = \prod_{i=1}^{n} f(t_i) \tag{1}$$



Time to event

Alternative: Estimate a summary statistic, e.g. Mean survival time (aka Life Expectancy)

$$\mathbb{E}[T] = \int_0^\infty S(t)$$

Alternative: Estimate a summary statistic, e.g. Mean survival time (aka Life Expectancy)

$$\mathbb{E}[T] = \int_0^\infty S(t)$$

This can be generalized!

$$\mathbb{E}[T|T \geq t] = \int_t^\infty S(t)$$

n.b. This implies that we can estimate the expectation by first estimating the survival function.

Censored data

Problem: we can't always wait to observe the terminating event (e.g. humans live a long time)

Censored data

Problem: we can't always wait to observe the terminating event (e.g. humans live a long time)

Solution: incorporate an indicator that the terminating event was observed (which assumes right censoring).

Censored data

Problem: we can't always wait to observe the terminating event (e.g. humans live a long time)

Solution: incorporate an indicator that the terminating event was observed (which assumes right censoring).

Formally, we define $C \geq 0$ to be our censoring time (analogous to our event time)

- Our observed time then becomes $Y = \min(T, C)$
- We have an associated indicator $\delta = \mathbb{I}(T \leq C)$

Censored data

Our updated likelihood now has to account for the censoring, i.e. let $q(c)$ and $Q(C)$ be the density and survival functions for C. Then

- If a person is censored, their likelihood is $S(y)q(y)$
- If a person is not censored, their likelihood is $f(y)Q(y)$

Our likelihood is therefore

$$
\begin{aligned}
L &= \prod_{i=1}^{n} [f(y_i)Q(y_i)]^{\delta_i} [S(y_i)q(y_i)]^{1-\delta_i} \\
&= \prod_{i=1}^{n} [f(y_i)^{\delta_i} S(y_i)^{1-\delta_i}][Q(y_i)^{\delta} q(y_i)^{1-\delta_i}] \\
&\propto \prod_{i=1}^{n} f(y_i)^{\delta_i} S(y_i)^{1-\delta_i} = \prod_{i=1}^{n} h(y_i)^{\delta_i} S(y_i)
\end{aligned}
$$

Question: rather than dealing with the survival function, can I just simplify the problem and apply (straight-forward) MLE?

Examples:

- Discarding the censored values
- Treating the censored values as uncensored (i.e set $T = Y$).

Censored data

Question: rather than dealing with the survival function, can I just simplify the problem and apply (straight-forward) MLE?

Examples:

- Discarding the censored values
- Treating the censored values as uncensored (i.e set $T = Y$).

Answer: No! These will result in biased estimates!

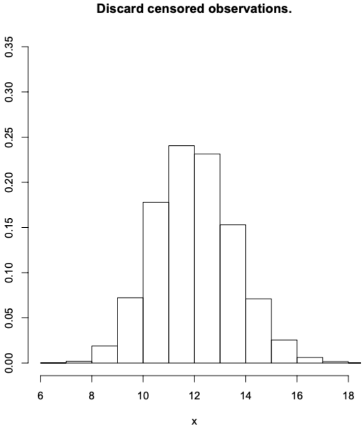
Censored data

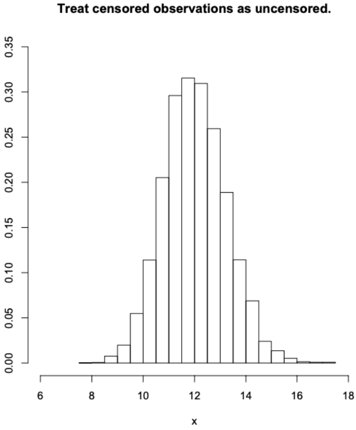
A quick simulation:

- $T_1, ..., T_n \sim Exp(\lambda = 1/20)$
- $C_1, ..., C_n \sim Exp(\lambda = 1/30)$
- Two estimators:
  - $\hat{\mu}_{1n} = \frac{1}{\sum_{i=1}^{n} \delta_i} \sum_{i=1}^{n} Y_i \delta_i$
  - $\hat{\mu}_{2n} = \frac{1}{n} \sum_{i=1}^{n} Y_i$

Censored data

A quick simulation ($\mathbb{E}_0[T] = 1/\lambda = 20$):





Kaplan Meier Estimator

If there is no censoring, estimating the survival function is straight-forward, i.e.

$$\hat{S}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(t_i \geq t) \tag{2}$$

Kaplan Meier Estimator

If there is no censoring, estimating the survival function is straight-forward, i.e.

$$\hat{S}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(t_i \geq t) \tag{2}$$

With censoring, we have pairs of outcomes $(y_1, \delta_1), (y_2, \delta_2), ..., (y_n, \delta_n)$.

- We can form an estimator assuming independent censoring.

Kaplan Meier Estimator

Our setup (for K observed events)

- Order our event times, i.e. $d_1 < d_2 < ... < d_K$

For a given d_k, we have (by the law of total probability)

$$
\begin{aligned}
S(d_k) &= P(T > d_k) \\
&= P(T > d_k | T > d_{k-1}) P(T > d_{k-1}) \\
&\quad + P(T > d_k | T \leq d_{k-1}) P(T \leq d_{k-1}) \\
&= P(T > d_k | T > d_{k-1}) P(T > d_{k-1}) \\
&= P(T > d_k | T > d_{k-1}) S(d_{k-1}) \\
&= P(T > d_k | T > d_{k-1}) \times \cdots \times P(T > d_2 | T > d_1) P(T > d_1)
\end{aligned}
$$

Kaplan Meier Estimator

Our setup (for K observed events)

- Count the number of events at each time, i.e. $q_1 < q_2 < ... < q_K$
- Count the number of "at risk" at each time, i.e. $r_1 < r_2 < ... < r_K$

We can estimate $P(T > d_j | T > d_{j-1})$ using our data, i.e.

$$\hat{P}_n(T > d_j | T > d_{j-1}) = \frac{r_j - q_j}{r_j} \tag{3}$$

n.b. This is the fraction of the risk set that survives past time d_j.

We have

$$\hat{S}_n(d_k) = \prod_{j=1}^{k} \frac{r_j - q_j}{r_j} \tag{4}$$

where r_j is the number at risk and q_j is the number of events (at time j)

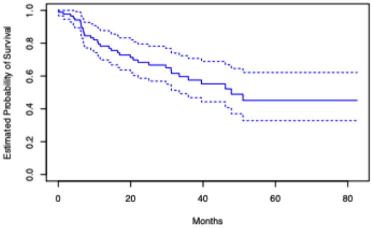


FIGURE 11.2. *For the* BrainCancer *data, we display the Kaplan–Meier survival curve (solid curve), along with standard error bands (dashed curves).*

The log-rank test

Question: What if we have two groups? How do we compare their survival curves?

Recall: For linear models, we can perform a hypothesis test via

$$t = \frac{\hat{\beta}_1 - \mu_0}{\sqrt{\mathrm{var}(\hat{\beta}_1)}} \tag{5}$$

The log-rank test

Question: What if we have two groups? How do we compare their survival curves?

Recall: For linear models, we can perform a hypothesis test via

$$t = \frac{\hat{\beta}_1 - \mu_0}{\sqrt{\text{var}(\hat{\beta}_1)}} \tag{5}$$

We can apply the same concept here, i.e.

$$W = \frac{X - \mathbb{E}[X]}{\sqrt{\text{var}(X)}} \tag{6}$$

e.g. if q_{1k}, r_{1k} are the number of events and at risk for group 1 (at time k), then

$$W_k = \frac{q_{1k} - \hat{\mathbb{E}}[q_{1k}]}{\sqrt{\text{var}(q_{1k})}} : \hat{\mathbb{E}}[q_{1k}] = \frac{r_{1k}}{r_k} q_k \tag{7}$$

The log-rank test

For the log-rank test we apply this across all time points k, i.e. let $X = \sum_{k=1}^{K} q_{1k}$ given us

$$W = \frac{\sum_{k=1}^{K}(q_{1k} - \mathbb{E}[q_{1k}])}{\sqrt{\sum_{k=1}^{K} \text{var}(q_{1k})}} \tag{8}$$

We compare this statistic to a standard normal distribution to calculate the p-value.

Example

Question: Do winners of the Oscar live longer?

An approach:

- Create a data set of actors' lifespans.
- Divide them into whether they've won an oscar.
- Fit KM Curves to each group and test using the log-rank test.

Example

Question: Do winners of the Oscar live longer?

An approach:

- Create a data set of actors' lifespans.
- Divide them into whether they've won an oscar.
- Fit KM Curves to each group and test using the log-rank test.

THIS IS INCORRECT!

Cox-proportional hazards

Many times we'll have more than 1 covariate that we'd like to regress our outcome on.

Our solution is to assume

$$h(t|x_i) = h_0(t)\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right) \tag{9}$$

The Cox-proportional hazards model is described as "semi"-parametric since $h_0(t)$ is unspecified.

Cox-proportional hazards

Assume wlog that we have univariate $x \in \{0, 1\}$. Then

$$\begin{aligned} h(t|x_i = 0) &= h_0(t) \exp(0) \\ h(t|x_i = 1) &= h_0(t) \exp(\beta_j) \end{aligned}$$

So that the hazard ratio is $\frac{h(t|x_i=1)}{h(t|x_i=0)} = \frac{h_0(t)\exp(\beta_j)}{h_0(t)} = \exp(\beta_j)$

n.b. The baseline hazard $h_0(t)$ is for the covariate profile $x = (0, ..., 0)$

Cox-proportional hazards

Question 1: Given that $h_0(t)$ is unspecified, how do we go about estimating the β_j's?

Answer: Apply the same ordering trick that was used in the KM curves, i.e. order the event times and calculate the probabilities

$$\frac{h_0(y_i)\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'}\geq y_i} h_0(y_i)\exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)} \tag{10}$$

Cox-proportional hazards

$$\frac{h_0(y_i)\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'}\geq y_i} h_0(y_i)\exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)} \tag{11}$$

- The probability of an observation failing at each time y_i is ratio of time-specific hazard over total hazard.
- The ratio of hazards cancels out $h_0(t)$, meaning we don't have to worry about it in estimating our β_j's.
- The product of these probabilities over the uncensored observations is called the *partial* likelihood.
- No closed form solution exists for the *partial* likelihood.

Cox-proportional hazards

Question 2: Our partial likelihood only allows us to estimate our β's. What about the survival or hazard function?

Answer: We can estimate the cumulative hazard via

$$\Lambda_0(y) = \sum_{i=1}^{n} \frac{\mathbb{I}(y_i < y)\delta_i}{\sum_{i':y_{i'} \geq y_i} \exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)} \tag{12}$$

The survival curve is then $S(y) = \exp(-\Lambda_0(y))$.

Time-varying covariates

Question 3: What if our features change over time?

Time-varying covariates

Question 3: What if our features change over time?

Solution: We assign the time that corresponds to each of our features for our outcome (along with the indicator of failure).

- The partial likelihood still works out the same!
- Now it's calculated with our covariates specific to the time periods we specify.
- This approach is very similar to "pooled" logistic regression.

References

[1] ISL. Chapter 11