Survival Analysis and Censored Data

Survival analysis

Statistical method used to analyse and model the time until an event of interest occurs

- Provides valuable insights into the time-to-event data
- Helps to understand the factors that influence the occurrence of events
- Enables the estimation of hazard rates and comparison of survival curves between different groups
- Assists in making informed decisions
- Primary goal is to estimate the survival function
- Takes censoring into account

Censoring refers to observations where the event of interest has not occurred for some subjects by the end of the study period

Survival and Censoring times

Survival time : Time at which the event of interest occurs

- Ex: Time at which the patient dies in a medical study
- Indicated by 'T'

Censoring time : Time at which censoring occurs

- Ex: Time at which the patient drops out of the study or the study ends
- Indicated by 'C'

Observed time= Y = min(T,C)

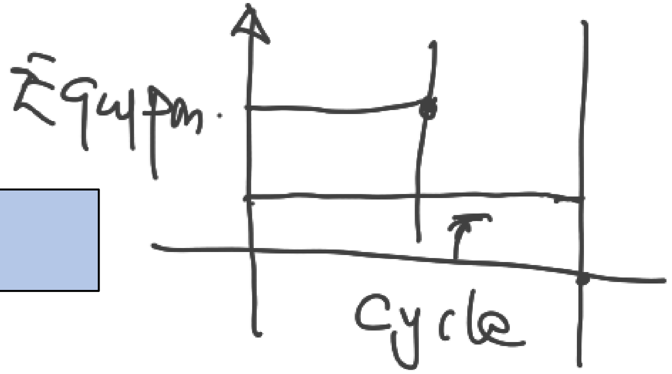
- True survival time T is observed if T < C
- Censoring time C is observed if T > C

$$\textit{Status indicator} = \delta = \begin{cases} 1 & if\ T \leq C \\ 0 & if\ T > c \end{cases}$$

Types of censoring

Right censoring

- Happens when an individual is still under observation at the end of the study
- The event of interest has not occurred for that individual
- The actual event time is unknown
- Most common type of censoring in survival analysis



Occurs when $T \geq Y$

Left censoring

- Happens when the event of interest has occurred before the start of the observation period
- Only the information that the event occurred before the study began is available
- The exact event time is unknown

Occurs when $T \leq Y$

Kaplan-Meier survival curve

Non-parametric statistic estimator

$$\textit{Survival curve/Survival function} = S(t) = Pr(T > t)$$

- Estimating survival function is complicated by the presence of censoring
- This is an approach to overcome this challenge

$$Pr(T > d_k) = Pr(T > d_k | T > d_{k-1}) Pr(T > d_{k-1}) + Pr(T > d_k | T \leq d_{k-1}) Pr(T \leq d_{k-1})$$

0

Where :

- $d_1 < d_2 < . . . < d_K$ denote the K unique event times among the non-censored subjects
- q_k denote the number of subjects for whom **event has occurred at time** d_k
- r_k denotes the number of subjects for whom the **event has not occurred** and are in the study just before d_k , **called the risk set**

$$P(A) = P(B) \cdot P(A/B) + P(B^c) \cdot P(A/B^c)$$

Kaplan-Meier survival curve

- It is impossible for the event to occur to the subject past time d_k if the event has not happened until an earlier time d_{k-1}
- Therefore

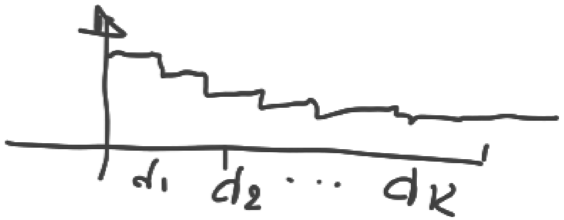
$$\boldsymbol{S(d_k) = Pr(T > d_k) = Pr(T > d_k | T > d_{k-1}) Pr(T > d_{k-1})}$$

- Plugging estimates of each of terms on the right side

$$\boldsymbol{\widehat{Pr}(T > d_j | T > d_{j-1}) = (r_j - q_j) / r_j}$$

$$\boldsymbol{Kaplan - Meier\ estimator\ \hat{S}(d_k) = \prod_{j=1}^{k} \frac{r_j - q_j}{r_j}}$$

- Kaplan-Meier survival curve has a step-like shape since $\hat{S}(t) = \hat{S}(d_k)$ for times between d_k and d_{k+1}



Log-Rank test

- Used to compare the survival curves of two or more groups or treatment arms
- EX: In case of cancer study, it is used to compare the survival of males to that of females to a treatment
- The idea of log-rank test statistic is
  - H_0: E(X) = μ for some random variable X
  - Test statistic is of the form ('1' denotes the group)
  - When the sample size is large, W has approximately a standard normal distribution

$$\boldsymbol{W = \frac{X - \mu}{\sqrt{Var(X)}}, \qquad X = \sum_{k=1}^{K} q_{1k}}$$

$$\boldsymbol{Expected\ value\ of\ X = \mu = \sum_{k=1}^{K} \frac{r_1 k}{r_k} q_k}$$

Log-Rank test

- Variance of q_{1k} is

$$\boldsymbol{Var(q_{1k}) = \frac{q_{k(}\, r_{1k}/r_k)(1 - r_{1k}/r_k)(r_k - q_k)}{r_k - 1}}$$

$$\boldsymbol{Var\left(\sum_{k=1}^{K} q_{1k}\right) \approx \sum_{k=1}^{K} Var(q_{1k}) = \sum_{k=1}^{K} \frac{q_{k(}\, r_{1k}/r_k)(1 - r_{1k}/r_k)(r_k - q_k)}{r_k - 1}}$$

- Log-rank test statistic is

$$\boldsymbol{W = \frac{\sum_{k=1}^{K} q_{1k} - \frac{r_1 k}{r_k} q_k}{\sqrt{\sum_{k=1}^{K} \frac{q_{k(}\, r_{1k}/r_k)(1 - r_{1k}/r_k)(r_k - q_k)}{r_k - 1}}}}$$

Regression models with a survival response

Hazard function

- Also called as hazard rate/ force of mortality
- Defined as

$$\boxed{\boldsymbol{h(t)} = \lim_{\Delta t \to 0} \frac{\boldsymbol{Pr(t < T \leq t + \Delta t | T > t)}}{\boldsymbol{\Delta t}}}$$

- T = Unobserved survival time
- Δt is a small number

$$\boxed{\boldsymbol{h(t)} \approx \frac{\boldsymbol{Pr(t < T \leq t + \Delta t | T > t)}}{\boldsymbol{\Delta t}}}$$

- From
  - $\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$
  - $S(t) = \Pr(T > t)$

$$\boxed{\boldsymbol{h(t)} = \lim_{\Delta t \to 0} \frac{\boldsymbol{Pr((t < T \leq t + \Delta t) \cap (T > t))/\Delta t}}{\boldsymbol{Pr(T > t)}}}$$

Regression models with a survival response

Hazard function

$$\boldsymbol{h(t) = \lim_{\Delta t \to 0} \frac{Pr(t < T \leq t + \Delta t)/\Delta t}{Pr(T > t)} = \frac{f(t)}{S(t)}}$$

- f(t) = Probability density function *i.e.*, Instantaneous rate of death at time t
- The likelihood associated with the *i*th observation is

$$\boldsymbol{L_i = \begin{cases} f(y_i) & \textit{if the } i\text{th} \textit{ observationis not censored} \\ S(y_i) & \textit{if the } i\text{th} \textit{ observationis censored} \end{cases}}$$

$$\boldsymbol{L_i = f(y_i)^{\delta_i} S(y_i)^{1-\delta_i}}$$

- Assuming that the n observations are independent

$$\boldsymbol{L = \prod_{i=1}^{n} f(y_i)^{\delta_i} S(y_i)^{1-\delta_i} = \prod_{i=1}^{n} h(y_i)^{\delta_i} S(y_i)}$$

Regression models with a survival response

Proportional Hazards

$T \not\perp x$ Risk function

- Assumption (exponential survival)

$$\boxed{\boldsymbol{h(t|x_i) = h_0(t) exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}} = h_0(t) \cdot e^{\alpha(x_i, \beta)}$$

- $h_0(t) \geq 0$ is called the baseline hazard (unspecified function)
- x_i is the feature vector
- $\boldsymbol{exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}$ is the relative risk for the feature vector $\boldsymbol{x_i = (x_{i1}, \ldots\ldots x_{ip})^T}$
- The hazard function is flexible as the probability density function is allowed to take any form
- One unit increase in x_{ij} corresponds to an increase in $h(t|x_i)$ by a factor of $\exp(\beta_j)$

Regression models with a survival response

Cox's proportional hazards model

- Makes it possible to estimate β without having to specify the form of $h_0(t)$
- Assumptions
  - Each event occurs at a distinct time
  - δ_i = 1, *i*th observation is uncensored
  - y_i is its future time

(y_i, δ_i)

- Hazard function for the *i*th observation at time y_i

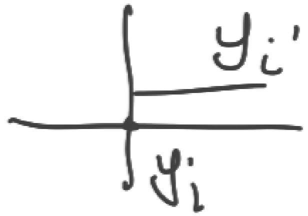
$$\boxed{h(y_i|x_i) = h_0(y_i) exp\left(\sum_{j=1}^{p} x_{ij}\,\beta_j\right)}$$

$$= h_0(y_i)\, e^{\alpha(x_i, \beta)}$$

Regression models with a survival response

Cox’s proportional hazards model

- Total hazard at time y_i for the at risk observations is



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

- The probability that the *i*th observation is the one to fail at time y_i is given by

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

$$= \frac{e^{\alpha(x_i, \beta)}}{\sum_{i' : y_{i'} \geq y_i} e^{\alpha(x_{i'}, \beta)}}$$

Regression models with a survival response

Cox's proportional hazards model

Partial likelihood

- Valid regardless of the true value of $h_0(t)$, making the model very flexible and robust
- Does not correspond exactly to the probability of the data under assumption. However, it is a very good approximation

$$\prod_{i:\delta_i=1} p_i(\beta) \longleftarrow \boxed{PL(\boldsymbol{\beta}) = \prod_{i:\,\delta_i=1} \frac{\exp\left(\sum_{j=1}^{p} x_{ij}\beta_i\right)}{\sum_{i':y_{i'}\geq y_i} \exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)}} \longrightarrow$$

Numerical Optimization

- β is estimated my maximizing the partial likelihood with respect to β

Regression models with a survival response

Cox’s proportional hazards model or log-rank test?

Case: For a single predictor case ($p = 1$), which is assumed to be binary ($x_i \in \{0,1\}$)

In the case of a single binary covariate, the score test for $H_0 : \beta = 0$ in Cox’s proportional hazards model is equal to the log-rank test

- Thus, it does not matter which approach is being used

$$y = \beta_0 + \beta_1 x$$

$$H_0 : \beta_0 = 0 \ , \ \beta_1 = 0$$

$$H_1 : \beta_0 \neq 0 \quad \beta_1 \neq 0$$

Shrinkage for Cox model

- Similar to ‘loss+penalty’ formulation

$$\underset{\beta_0,\beta_1,\ldots\ldots,\beta_p}{\text{minimize}}\left\{\boldsymbol{L}(\boldsymbol{X},\boldsymbol{y},\boldsymbol{\beta})+\boldsymbol{\lambda P}(\boldsymbol{\beta})\right\}$$

L(X,y,β) – Loss function, P(β) – Penalty function, λ – Tuning parameter

- Consider minimizing a penalized version of the negative log partial likelihood

$$-\boldsymbol{log}\left(\prod_{i:\,\delta_i=1}\frac{\boldsymbol{exp}\left(\sum_{j=1}^{p}x_{ij}\beta_i\right)}{\sum_{i':y_{i'}\geq y_i}\boldsymbol{exp}\left(\sum_{j=1}^{p}x_{i'j}\beta_j\right)}\right)+\boxed{\boldsymbol{\lambda P}(\boldsymbol{\beta})}$$

Where (i) $\boldsymbol{P}(\boldsymbol{\beta}) = \sum_{j=1}^{p}\boldsymbol{\beta}_j^2$ for ridge penalty

(ii) $\boldsymbol{P}(\boldsymbol{\beta}) = \sum_{j=1}^{p}|\boldsymbol{\beta}_j|$ for lasso penalty

- When λ = 0, minimization is equivalent to maximizing the usual Cox partial likelihood
- When λ > 0, minimizing yields a shrunken version of the coefficient estimates

Shrinkage for Cox model

- When λ is large, then using a ridge penalty will give small coefficients that are not equal to zero
- When λ is sufficiently large, using a lasso penalty will give some coefficients exactly equal to zero

Partial Likelihood $f^{\underline{n}}$.

$$PL = \prod_{i,\,\delta_i=1} \frac{e^{\alpha(x_i,\beta)}}{\sum_{i',\,y_{i'}\geq y_i} e^{\alpha(x_{i'},\beta)}}$$

$$-\text{Log}\,PL = \frac{1}{N_{\delta_i=1}} \sum_{i,\,\delta_i=1} \left(\underline{\alpha(x_i,\beta)} - \log\left(\sum_{i',\,y_{i'}\geq y_i} e^{\alpha(x_{i'},\beta)}\right)\right)$$

Decision tree $X_i = [x_{1i} \ldots x_{pi}]^T$ and

Deep NN.

ML

$$\boxed{\alpha(\underline{X}_i, \beta)}$$

Chap. 11: An Introduction to statistical learning, Second edition

Hastie / Tibshirani.

AUC for survival analysis

- Area under the curve is a way to quantify the performance of a two-class classifier
- Generalizing the notion to survival analysis:
  - Estimated risk score is calculated using the Cox model coefficients

$$\boldsymbol{Estimated\ risk\ score = \hat{\eta}_i = \hat{\beta}_1 x_{i1} + \ldots\ldots + \hat{\beta}_p x_{ip} \quad for\ i = 1, \ldots\ldots, m}$$

  - If $\hat{\eta}_{i'} > \hat{\eta}_i$, the model predicts that i'th observation has a larger hazard than the ith observation
  - Thus survival time $t_i > t_{i'}$
  - Harrell's concordance index (C-index) computes the proportion of observation pairs for which $\hat{\eta}_{i'} > \hat{\eta}_i$ and $y_i > y_{i'}$

$$\boldsymbol{C} = \frac{\sum_{\boldsymbol{i,i'}:\boldsymbol{y_i > y_{i'}}} \boldsymbol{I}(\hat{\eta}_{i'} > \hat{\eta}_i)\boldsymbol{\delta_{i'}}}{\sum_{\boldsymbol{i,i'}:\boldsymbol{y_i > y_{i'}}} \boldsymbol{\delta_{i'}}}$$

$\mathrm{I}(\hat{\eta}_{i'} > \hat{\eta}_i) = 1$ if $\hat{\eta}_{i'} > \hat{\eta}_i$
$\mathrm{I}(\hat{\eta}_{i'} > \hat{\eta}_i) = 0$ otherwise

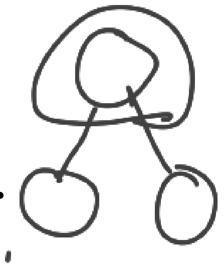
Time – dependent covariates

- Time –dependent covariates: Predictors whose value may change over time
- Ex: Patient's blood pressure $x_i \rightarrow x_i(t)$
- Proportional hazards model has the ability to handle time-dependent covariates
- For the example: the blood pressure, x_{ij} and $x_{i'j}$ is replaced with $x_{ij}(y_i)$ and $x_{i'j}(y_i)$ respectively

$x_{ij}\beta_j \rightarrow x_{ij}(t_i)$

Survival Trees

- Survival trees are a modification of classification and regression tress that use a split criterion
- It maximizes the difference between the survival curves in the resulting daughter nodes.
- Survival trees can then be used to create random survival forests



Tutorial for: RUL, Survival Analysis.
Q-statistics & Discriminate analysis

THANK YOU!