Cox PH Model

Estimation of survival function

Cox PH model

- Recall the Cox's proportional hazard model
$$h(t \mid \boldsymbol{Z}) = h_0(t)\exp(\boldsymbol{Z}^T\boldsymbol{\beta}).$$
- The coefficients $\boldsymbol{\beta}$ quantifies the effects of covariates on relative risk.
- The coefficients $\boldsymbol{\beta}$ can be estimated by maximizing the partial likelihood function.
- How to estimate the baseline?

Baseline survival

- In practice, we are also interested in estimating the survival probability for a new patient with a given set of covariates $S(t \mid \boldsymbol{Z})$.
- Note that in the Cox's model
$$h(t \mid \boldsymbol{Z}) = h_0(t)\exp(\boldsymbol{Z}^T\boldsymbol{\beta})$$
$$\Rightarrow H(t \mid \boldsymbol{Z}) = H_0(t)\exp(\boldsymbol{Z}^T\boldsymbol{\beta})$$
$$\Rightarrow \exp[-H(t \mid \boldsymbol{Z})] = \exp[-H_0(t)\exp(\boldsymbol{Z}^T\boldsymbol{\beta})]$$
$$\Rightarrow S(t \mid \boldsymbol{Z}) = [S_0(t)]^{\exp(\boldsymbol{Z}^T\boldsymbol{\beta})}$$
- After $\boldsymbol{\beta}$ being estimated, we need to estimate the baseline survival $S_0(t)$.

Recall Kaplan-Meier estimator

- Recall the relationships between $h(t)$, $H(t)$, and $S(t)$ for discrete r.v.s
  - $h(t_i) = P(\, T = t_i \mid T \geq t_i\,)$
  - $H(t) = \sum_{t_i \leq t} h(t_i)$
  - $S(t) = \prod_{t_i \leq t} \big(1 - h(t_i)\big)$

Recall Kaplan-Meier estimator

- In the univariate case, individuals all have the same hazard rate $h(t_i)$ at t_i.
- For a subject alive just prior to t_i,

  $P(\text{ Failure at } t_i \mid \text{alive prior to } t_i\,) {=} h(t_i)$
- Therefore

  $$\frac{d_i}{Y_i} \approx h(t_i) \quad or \quad d_i \approx E(\, d_i \mid R_i\,) = Y_i h(t_i)$$
  - $\hat{h}(t_i) = \frac{d_i}{Y_i}$
  - $\hat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{Y_i}$
  - $\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{Y_i}\right)$

Cox PH model

- Different hazard for different individuals
- Let d_i be the total number of events at t_i

$$d_i \approx E(\, d_i \mid R_i\,) = \sum_{j \in R(t_i)} E(\, I(\text{subject } j \text{ has event at } t_i) \mid R_i\,)$$

$$= \sum_{j \in R(t_i)} P(\text{ subject } j \text{ has event at } t_i \mid R_i\,) = \sum_{j \in R(t_i)} h_0(t_i) \exp\big(Z_j^T \beta\big)$$

$$= h_0(t_i) \sum_{j \in R(t_i)} \exp\big(Z_j^T \beta\big)$$

- Therefore estimate $h_0(t_i)$ by

$$\hat{h}_0(t_i) = \frac{d_i}{\sum_{j \in R(t_i)} \exp\big(Z_j^T \beta\big)}$$

Baseline estimation

- Therefore

$$\widehat{H}_0(t) = \sum_{t_i \le t} \hat{h}_0(t_i) = \sum_{t_i \le t} \frac{d_i}{\sum_{j \in R(t_i)} \exp(Z_j^T \beta)}$$

- The baseline survival function can be estimated by

$$\hat{S}_0(t) = \prod_{t_i \le t} \left(1 - \hat{h}_0(t_i)\right) = \prod_{t_i \le t} \left(1 - \frac{d_i}{\sum_{j \in R(t_i)} \exp(Z_j^T \beta)}\right)$$

- This is referred as Product-Limit method in SAS.

Survival estimation

- Therefore the survival function for patients with covariates $\boldsymbol{Z}$ can be estimated by

$$\hat{S}(t \mid \boldsymbol{Z}) = \left[\hat{S}_0(t)\right]^{\exp(Z^T \hat{\beta})}.$$

  - $\hat{S}_0(t)$ is the estimated baseline survival for individuals with $\boldsymbol{Z} = 0$.
- Sometimes the baseline is meaningless that no individual can have $\boldsymbol{Z} = 0$.
- For example, if $Z = age$, all patients are over 30 years old in the study.
- But can construct covariates so that $\boldsymbol{Z} = 0$ is representative by centering the covariates.
- Let $\boldsymbol{Z}^* = \boldsymbol{Z} - \overline{\boldsymbol{Z}}$, then $\boldsymbol{Z}^* = 0$ corresponding to the baseline being "average" individuals.
- A good reason to center the covariates in the analysis.

SAS Example

```
data covin;
input stage age id;
cards;
2 50 1
3 40 2
;
run;
```

- Create a data set to estimate survival functions for two potential patients.
- We will estimate the survival functions

$$S_1(t \mid Stage = 2, age = 50)$$
$$S_2(t \mid Stage = 3, age = 40)$$

SAS Example

- BASELINE Statement in PROC PHREG
  - OUT=SAS-data-set
  - COVARIATES=SAS-data-set
  - TIMELIST=list (eg. timelist=5,20 to 50 by 10)
  - keyword=name (specifies the statistics to be included in the OUT= data set)
  - METHOD(NELSON, PL)

SAS Example

```
proc phreg data=larynx plots(overlay)=survival;
class stage;
model stime*censor(0)=stage age/risklimits
;/*computes confidence intervals for hazard ratios*/
baseline out=out covariates=covin survival=_all_
/method=PL rowid=Id;
hazardratio ‘H1’age / units=10 cl=both;
hazardratio ‘H2’stage / cl=both;
contrast "age" age 10/estimate=exp;
contrast "Stage 1 vs Stage 2" stage 1 -1 0
/estimate=exp;
contrast "Stage 2 vs Stage 3" stage 0 1 -1
/estimate=exp;
run;
```