STAT347: Generalized Linear Models
Lecture 15

Winter, 2024
Jingshu Wang

Today’s topics:

- Proportional hazard regression model
  - Model setup
  - Partial likelihood
  - Estimation and inference with partial likelihood

Dealing with covariates in survival analysis

- Evaluate how covariates are associated with the survival time
  - Observed data: (Y_s, X_s, δ_s) for $s = 1,2, \ldots, n$ observations

$$Y_s = \min(T_s, C_s), \delta_s = 1_{C_s < T_s}$$

- Generalize from the 2X2 table in log-rank test where X_s is just a group indicator
  - For each discrete survival time i,
  - We observe r_{i1} and r_{i2} samples that are still alive at the beginning of this time bin for each group respectively
  - Observe d_{i1} and d_{i2} death during this time bin for two groups respectively.

	death	alive	total at risk
Group 1	d_{i1}	$r_{i1} - d_{i1}$	r_{i1}
Group 2	d_{i2}	$r_{i2} - d_{i2}$	r_{i2}
Total	d_i	$r_i - d_i$	r_i

$$d_{ik} \sim \text{Binomial}(r_{ik}, h_{ik})$$

$$h_{ik} = P(T_s = i | S_s \geq i, X_s = k)$$

Proportional hazard model

- Define hazard rate $h_s(t) = f_s(t)/S_s(t)$ for an observation s
- We assume that the hazard

$$h_s(t) = e^{X_s^T \beta} h_0(t)$$

- The model is proposed by David Cox (1972, 1975)
- This is a semi-parametric model as we have no assumption on the baseline hazard function $h_0(t)$
- X does not include the intercept for identifiablity
- proportional hazard:

$$\log \left\{ \frac{h_s(t)}{h_0(t)} \right\} = X_s^T \beta$$

Proportional hazard model

- Survival function need to be less than 1, while the hazard rate does not have that constraint.
- The benefit of having a proportional model is that there is no constraint on the range of β to have the hazard rate positive.

$$h_s(t) = e^{X_s^T \beta} h_0(t)$$

- No parametric assumption on the baseline hazard function $h_0(t)$

- Question: how do we estimate the coefficients β without estimating $h_0(t)$

Partial likelihood

- For simplicity, assume no ties: exactly one person die at a time (if there are ties, idea is similar but needs some adjustments)

- Denote the risk set $\mathcal{R}(t) = \{s : y_s \geq t\}$: individuals that are still alive at time t
- At time Y_s where $\delta_s = 0$, conditional on the fact that there are exactly 1 person die, the probability of choosing individual s is

$$L_s = \frac{h_s(y_s)}{\sum_{l \in \mathcal{R}(y_s)} h_l(y_s)} = \frac{e^{X_s^T \beta}}{\sum_{l \in \mathcal{R}(y_s)} e^{X_l^T \beta}}$$

- Partial likelihood:

$$L = \prod_s L_s^{1-\delta_s}$$

- It is “partial" because it ignores all the non-events, times when nothing happened or there were losses to follow-up

Partial likelihood

- Constructing the full likelihood: for each sample s, assume we observe (y_s, δ_s). We build a likelihood for each sample conditional on C_s (treat C_s as fixed):
  - If $\delta_s = 0$, then we observe $T_s = y_s$, the likelihood is $L_s = f(y_s) = S(y_s)h(y_s)$
  - If $\delta_s = 1$, then we only observe $T_s \geq y_s$, the likelihood is $L_s = S(y_s)$

  Thus the full likelihood is

$$L = \prod_s L(s) = \prod_{s=1}^{n} S(y_s)h(y_s)^{\delta_s}$$

- Rewrite the full likelihood as

$$L = \prod_{s=1}^{n} S_s(y_s)h_s(y_s)^{\delta_s} = \prod_{s=1}^{n} \left(\frac{h_s(y_s)}{\sum_{l \in \mathcal{R}(y_s)} h_l(y_s)} \right)^{\delta_s} \left(\sum_{l \in \mathcal{R}(y_s)} h_l(y_s) \right)^{\delta_s} S_s(y_s)$$

Cox (1972) argued that the first term in this product contained almost all the information about β, while the last two terms contained the information about $h_0(t)$, the baseline hazard.

Estimation and inference

The log-likelihood:

$$l(\beta) = \log L = \sum_{s=1}^{n} (1 - \delta_s) \left[X_s^T \beta - \log \left\{ \sum_{t \in \mathcal{R}(y_s)} e^{X_s^T \beta} \right\} \right]$$

- Estimate β: solve the score equation $\dot{l}(\beta) = 0$
- Statistical inference:
  researchers has taken a lot of e ort to show that it has asymptotic distribution
  (not a trivial result)

$$\widehat{\beta} \dot{\sim} N(\beta, \ddot{l}(\widehat{\beta})^{-1})$$

Data example

- Continue Example10 R notebook