Survival Analysis: PH Model

Lu Tian and Richard Olshen

Stanford University

Partial Likelihood Function

- The log-PL is

$$\log\{PL(\beta)\} = \sum_{i=1}^{n} \delta_i \left\{ \beta' Z_i - \log \left(\sum_{j=1}^{n} I(U_j \geq U_i) e^{\beta' Z_j} \right) \right\}$$

- The score function

$$S_n(\beta) = \sum_{i=1}^{n} \delta_i \left\{ Z_i - \frac{\sum_{j=1}^{n} I(U_j \geq U_i) Z_j e^{\beta' Z_j}}{\sum_{j=1}^{n} I(U_j \geq U_i) Z_j e^{\beta' Z_j}} \right\}$$

$$= \sum_{i=1}^{n} \int_0^{\infty} \left\{ Z_i - \frac{\sum_{j=1}^{n} I(U_j \geq s) Z_j e^{\beta' Z_j}}{\sum_{j=1}^{n} I(U_j \geq s) Z_j e^{\beta' Z_j}} \right\} dN_i(s)$$

  where $N_i(t) = I(U_i \leq t)\delta_i$.

- $\hat{\beta}$ is the solution to the estimating equation $S_n(\beta) = 0$.

Partial Likelihood Function

- The observed information matrix

$$
\begin{aligned}
I(\beta) &= -\frac{\partial^2 S_n(\beta)}{\partial \beta^2} \\
&= \sum_{i=1}^{n} \delta_i \left[\frac{\sum_{j=1}^{n} I(U_j \geq U_i) Z_j^{\otimes 2} e^{\beta' Z_j}}{\sum_{j=1}^{n} I(U_j \geq U_i) Z_j e^{\beta' Z_j}} - \frac{\left\{ \sum_{j=1}^{n} I(U_j \geq U_i) Z_j^{\otimes 2} e^{\beta' Z_j} \right\}^{\otimes 2}}{\left\{ \sum_{j=1}^{n} I(U_j \geq U_i) Z_j e^{\beta' Z_j} \right\}^{2}} \right] \\
&= \sum_{i=1}^{n} \int_0^{\infty} \frac{S^{(2)}(\beta, s) S^{(0)}(\beta, s) - \{S^{(1)}(\beta, s)\}^{\otimes 2}}{\{S^{(0)}(\beta, s)\}^2} dN_i(s)
\end{aligned}
$$

  where $S^{(k)}(\beta, s) = \sum_{j=1}^{n} I(U_j \geq s) Z_j^{\otimes k} e^{\beta' Z_j}, k = 0, 1, 2.$

- $I(\beta)$ is semi-positive definite, i.e., $\log\{PL(\beta)\}$ is a concave function.

How to solve the equation

- Rewrite the score function as

$$\sum_{i=1}^{n} Z_i \left\{ \delta_i - \sum_{j:U_j \le U_i} \frac{\delta_i e^{\beta' Z_i}}{\sum_{k=1}^{n} I(U_k \ge U_j) e^{\beta' Z_k}} \right\}$$

$$= \sum_{i=1}^{n} Z_i \left\{ \delta_i - \hat{H}(U_i, \beta) e^{\beta' Z_i} \right\}$$

- Fisher's scoring algorithm can be used to solve the equation.

Variable selection in Cox

- Since the log partial likelihood function in concave, the negative partial likelihood function can be viewed as a convex loss function as the squared loss in the linear regression.
- Impose a L_1 penalty to choose informative covariates:

$$\min \left[-\log\{PL(\beta)\} + \lambda|\beta| \right],$$

  where λ is a positive tuning parameter to be chosen from the cross-validation.
- Efficient numerical algorithm exists due to the convexity of the loss function.

Statistical Inference

- As $n \to \infty$, $\sqrt{n}(\hat{\beta} - \beta_0) \to N(0, \Sigma)$ where Σ can be estimated by $n^{-1} I(\hat{\beta})$.
- For large n

$$\hat{\beta} - \beta \sim N\left\{0, I^{-1}(\hat{\beta})\right\}$$

Statistical Inference

- To test $H_0 : \beta = \beta_0$:
  1. Wald test: $T = (\beta - \beta_0)' I(\hat{\beta})(\hat{\beta} - \beta_0) \sim \chi^2_p$ under H_0.
  2. Score test: $T = S(\beta_0)' I^{-1}(\beta_0) S(\beta_0) \sim \chi^2_p$ under H_0.
  3. Partial Likelihood Ratio test: $T = 2[\log\{PL(\hat{\beta})\} - \log\{PL(\beta_0)\}] \sim \chi^2_p$ under H_0.

Score Test vs Logrank Test

- Data: (U_i, δ_i, Z_i) where $Z_i = 1$ or 0.
- The score test statistics

$$
\begin{aligned}
S_n(0) &= \sum_{i=1}^{n} \delta_i \left\{ Z_i - \frac{\sum_{j=1}^{n} I(U_j \geq U_i) Z_j}{\sum_{j=1}^{n} I(U_j \geq U_i)} \right\} \\
&= \sum_{j=1}^{K} \sum_{i:U_i=\tau_j} \delta_i \left\{ Z_i - \frac{\sum_{j=1}^{n} I(U_j \geq U_i) Z_j}{\sum_{j=1}^{n} I(U_j \geq U_i)} \right\} \\
&= \sum_{j=1}^{K} \left\{ \left(\sum_{i:U_i=\tau_j} \delta_i Z_i \right) - \left(\sum_{i:U_i=\tau_j} \delta_i \right) \frac{\sum_{l=1}^{n} I(U_l \geq \tau_j) Z_l}{\sum_{l=1}^{n} I(U_l \geq \tau_j)} \right\} \\
&= \sum_{j=1}^{K} \left\{ d_{j1} - d_j \frac{Y_1(\tau_j)}{Y(\tau_j)} \right\} = \sum_{j=1}^{K} (O_j - E_j)
\end{aligned}
$$

where $0 < \tau_1 < \cdots < \tau_K$ are all the observed failure times.

Score Test vs Logrank Test

- The variance of $U(0)$ under H_0 is

$$
\begin{aligned}
I(0) &= \sum_{i=1}^{n} \delta_i \frac{S^{(2)}(0,U_i)S^{(0)}(0,U_i) - \{S^{(1)}(0,U_i)\}^2}{\{S^{(0)}(0,U_i)\}^2} \\
&= \sum_{j=1}^{K} \left(\sum_{i:U_i=\tau_j} \delta_i \right) \times \left\{ \frac{S^{(2)}(0,\tau_j)S^{(0)}(0,\tau_j) - \{S^{(1)}(0,\tau_j)\}^2}{\{S^{(0)}(0,\tau_j)\}^2} \right\} \\
&= \sum_{j=1}^{K} d_j \left\{ \frac{Y_1(\tau_j)}{Y(\tau_j)} - \frac{Y_1(\tau_j)^2}{Y(\tau_j)^2} \right\} \\
&= \sum_{j=1}^{K} \frac{d_j Y_1(\tau_j) Y_2(\tau_j)}{Y(\tau_j)^2} \approx \sum_{j=1}^{K} \frac{d_j Y_1(\tau_j) Y_2(\tau_j)(Y(\tau_j) - d_j)}{Y(\tau_j)^2 (Y(\tau_j) - 1)}
\end{aligned}
$$

- The score test is approximately equivalent to the logrank test.

Breslow Estimator

- How to estimate the baseline hazard function?
- $0 < \tau_1 < \cdots < \tau_K$ are all the observed failure times and for $\tau_j \leq t < \tau_{j+1}$

$$\hat{H}(t) = \sum_{i=1}^{j} \frac{1}{\sum_{k=1}^{n} I(U_k \geq U_i) e^{\hat{\beta}' Z_k}}.$$

- In general

$$\hat{H}(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i(s)}{S^{(0)}(\hat{\beta}, s)}$$

Breslow Estimator

- One can use the Breslow's estimator to predict survival probability:

$$\hat{S}(t|Z) = \exp\left\{-\hat{H}(t)e^{\hat{\beta}'Z}\right\}.$$

Time dependent covariates

- The covariates could depend on time and the interest is to study the relationship between hazard function and $Z(t)$
- Cox model:

$$h(t|Z(t)) = h_0(t)e^{\beta' Z(t)}$$

- Examples of time-dependent covairates: internal and external (air pollution level, blood pressure, glucose level et al.)

Partial likelihood function

- Data $\{U_i, \delta_i, Z_i(s), s \in [0, U_i]\}$
- Partial likelihood function is

$$\prod_{i=1}^{n} \left(\frac{e^{\beta' Z_i(U_i)}}{\sum_{j=1}^{n} I(U_j \geq U_i) e^{\beta' Z_j(U_i)}} \right)^{\delta_i}$$

Model Checking based on the time-dependent covariates

- To check the PH assumption

$$h(t|Z) = h_0(t)e^{\beta' Z}$$

- To consider a bigger model

$$h(t|Z) = h_0(t)e^{\beta' Z + \gamma'\{B(t)\times Z\}}$$

  If PH model is correct, then $\gamma = 0$.