BSTT536: Survival Data Analysis

Instructor: Hua Yun Chen, PhD

Division of Epidemiology and Biostatistics
School of Public Health
University of Illinois at Chicago

Table of Content

Two-group comparison

1. The survival functions for two groups

$$S_1(t) = \exp\{-H_1(t)\}, \text{ and } S_2(t) = \exp\{-H_2(t)\}.$$

2. In the two group comparison

$$S_2(t) = \exp\{-H_2(t)\} = \exp\left\{-H_1(t)\frac{H_2(t)}{H_1(t)}\right\}.$$

3. If we assume that

$$\frac{H_2(t)}{H_1(t)} = C,$$

then $H_2(t) = CH_1(t)$.

Two-group comparison in regression form

1. Under the proportional hazard assumption, $H_2(t) = e^{\beta} H_1(t)$ where $\beta = \log C$.
2. Equivalently, in terms of hazard rates,

$$h_2(t) = e^{\beta} h_1(t).$$

3. Define $Z = 0$ or 1 for a subject in first or second group. Then

$$h(t \mid Z) = h_1(t) \exp(\beta Z)$$

Multi-group comparison in regression form

1. For k group comparison,

$$S_j(t) = \exp\{-H_j(t)\} = \exp\left\{-H_1(t)\frac{H_k(t)}{H_1(t)}\right\},$$

where $H_j(t)$ is the cumulative hazard function for group j with $j = 1, \cdots, k$.

2. Let

$$\frac{H_j(t)}{H_1(t)} = C_j = \exp(\gamma_j)$$

for $j = 2, \cdots, k$. Let $Z_j = 1$ or 0 according to subject j in group j or not for $j = 2, \cdots, k$. The regression form is

$$h(t \mid Z_2, \cdots, Z_k) = h_1(t)\exp(\gamma_2 Z_2 + \cdots + \gamma_k Z_k).$$

Multi-group comparison in regression form (continuing)

1. If the k categories can be treated as ordinal and modeled linear, i.e.,
$$\gamma_j = (j-1)\beta,$$
then let Z be the ordinal variable taking values $0, \cdots, k-1$ corresponding to categories $1, \cdots, k$, the model can be reduced to
$$h(t \mid Z) = h_1(t)\exp(\beta Z).$$
2. In general, proportional hazard regression can be specified as
$$h(t \mid Z) = h_0(t)\exp(\beta_1 Z_1 + \cdots + \beta_p Z_p),$$
where $h_0(t)$ is called baseline hazard and $Z = (Z_1, \cdots, Z_p)^t$ is the vector of covariates.

Proportional hazards regression

1. Let $\beta = (\beta_1, \cdots, \beta_p)$, the conditional survival function

$$S(t|Z) = \exp\{-H_0(t)\exp(\beta Z)\}.$$

2. The proportion hazard model can be interpreted through

$$\log \frac{h(t|Z)}{h(t|Z_0)} = \beta(Z - Z_0).$$

   That is, the logarithm of the ratio of two hazards is proportional to a linear function of the covariates and is independent of the time.

3. The parameters in the proportional hazards model includes relative hazards ratio parameter β and the baseline hazard $H_0(t)$.

Parameter estimation in the proportional hazards regression

1. The observed data (X_i, δ_i, Z_i), $i = 1, \cdots, n$. The likelihood

$$\prod_{i=1}^{n} \{h_0(X_i) \exp(\beta Z_i)\}^{\delta_i} \exp\left\{-H_0(X_i) e^{\beta Z_i}\right\}.$$

2. *Maximizing the likelihood over H_0 leads to a profile likelihood to be maximized for estimating β as

$$PL(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\beta Z_i)}{\sum_{\{j|X_j \geq X_i\}} \exp(\beta Z_j)}\right]^{\delta_i}.$$

   This likelihood is called the partial likelihood. Cox (1975)'s Biometrika paper gave a justification for this name. $\{j|X_j \geq t\}$ is usually called at risk set.

3. We can treat the partial likelihood just as the ordinary likelihood in estimating and making inference of the unknown parameter β.

Estimation and inference for β

1. The log partial likelihood score equation for β is

$$\frac{\partial \log PL(\beta)}{\partial \beta} = \sum_{i=1}^{n} \delta_i \left\{ Z_i - \frac{\sum_{\{j|X_j \geq X_i\}} Z_j \exp(\beta Z_j)}{\sum_{\{j|X_j \geq X_i\}} \exp(\beta Z_j)} \right\} = 0.$$

   The estimator of β, denoted by $\hat{\beta}$, solves the score equation.

2. The information matrix for estimating β from the partial likelihood is

$$-\frac{\partial^2 \log PL(\beta)}{\partial \beta^2} = \sum_{i=1}^{n} \delta_i \left[\frac{\sum_{\{j|X_j \geq X_i\}} Z_j^2 \exp(\beta Z_j)}{\sum_{\{j|X_j \geq X_i\}} \exp(\beta Z_j)} \right.$$

$$\left. - \left\{ \frac{\sum_{\{j|X_j \geq X_i\}} Z_j \exp(\beta Z_j)}{\sum_{\{j|X_j \geq X_i\}} \exp(\beta Z_j)} \right\}^2 \right].$$

Estimation and inference for β

1. The variance of the partial likelihood estimator of β can be estimated consistently by

$$\hat{V}(\hat{\beta}) = \left\{ -\frac{\partial^2 \log PL(\beta)}{\partial \beta^2} \right\}^{-1}.$$

2. Inference on β_0 can be carried out based on

$$\frac{\hat{\beta} - \beta_0}{\sqrt{\hat{V}(\hat{\beta})}} \sim N(0, 1).$$

*Breslow estimator for $H_0(t)$

1. Estimate the jump at the uncensored survival time X_k by

$$\hat{h}_k = \frac{d_k}{\sum_{\{j|X_j \geq X_k\}} \exp(\beta Z_j)},$$

   where d_k is the total number of uncensored survival times at X_k.

2. Estimate $H_0(t)$ by

$$\hat{H}_0(t) = \sum_{\{k|\delta_k=1, X_k \leq t\}} \hat{h}_k.$$

3. The variance of $\hat{H}_0(t)$ is more involved.

Two-sample comparison by Cox regression model

1. The model for two sample comparison

$$h_1(t) = h_2(t)\exp(\beta).$$

2. The partial likelihood

$$\prod_{i=1}^{n}\left[\frac{\exp(\beta Z_i)}{\sum_{\{j|X_j\geq X_i\}}\exp(\beta Z_j)}\right]^{\delta_i}.$$

where $Z_i = 1$ if subject i is in group 1, and $Z_i = 0$ if subject i is in group 2.

3. The score at $\beta = 0$ is

$$\sum_{i=1}^{n}\delta_i\left\{Z_i - \frac{\sum_{\{j|X_j\geq X_i\}}Z_j}{\sum_{\{j|X_j\geq X_i\}}1}\right\} = \sum_{k=1}^{K}\left\{d_{k1} - d_k\frac{n_{k1}}{n_k}\right\}$$

which is the same as the log-rank test score.

Two-sample comparison by Cox regression model (Continuing)

1. The information for estimating β becomes

$$\begin{aligned}
& \sum_{i=1}^{n} \delta_i \left[\frac{\sum_{\{j|X_j \geq X_i\}} Z_j^2}{\sum_{\{j|X_j \geq X_i\}} 1} - \left\{ \frac{\sum_{\{j|X_j \geq X_i\}} Z_j}{\sum_{\{j|X_j \geq X_i\}} 1} \right\}^2 \right] \\
= & \sum_{k=1}^{K} d_k \left\{ \frac{n_{k1}}{n_k} - \left(\frac{n_{k1}}{n_k} \right)^2 \right\} \\
= & \sum_{k=1}^{n} \frac{n_{k1} n_{k2} d_k}{n_k^2} \approx \sum_{k=1}^{n} \frac{n_{k1} n_{k2} d_k (n_k - d_k)}{n_k^2 (n_k - 1)}
\end{aligned}$$

which is the variance used in the log-rank test.

Multiple group comparison by Cox regression model

1. Suppose that G groups are to be compared. Assume that

$$h_g(t) = h_G(t)\exp(\beta_g), g = 1, \cdots, G-1.$$

By defining dummy variables to indicate each group as

$$Z_{ig} = \begin{cases} 1 & \text{if subject } i \text{ is in group } g, \\ 0 & \text{otherwise,} \end{cases}$$

where $g = 1, \cdots, G-1$, the model can be rewritten as

$$h(t|Z_{i2}, \cdots, Z_{iG}) = h_{G-1}(t)\exp(\beta_1 Z_{i1} + \cdots + \beta_{G-1} Z_{i(G-1)}).$$

2. The scores for $(\beta_1, \cdots, \beta_{G-1})$ under the null hypothesis of no difference are

$$\begin{aligned} \sum_{i=1}^{n} \delta_i \left\{ Z_{i1} - \frac{\sum_{\{j|X_j \geq X_i\}} Z_{j1}}{\sum_{\{j|X_j \geq X_i\}} 1} \right\} &= \sum_{k=1}^{K} \left\{ d_{k1} - d_k \frac{n_{k1}}{n_k} \right\}, \\ \cdots & \quad \cdots \\ \sum_{i=1}^{n} \delta_i \left\{ Z_{i(G-1)} - \frac{\sum_{\{j|X_j \geq X_i\}} Z_{j(G-1)}}{\sum_{\{j|X_j \geq X_i\}} 1} \right\} &= \sum_{k=1}^{K} \left\{ d_{k(G-1)} - d_k \frac{n_{k(G-1)}}{n_k} \right\}. \end{aligned}$$

Multiple group comparison by Cox regression model (continuing)

1. The variance-covariance matrix is the inverse of the information matrix with diagonal elements

$$\sum_{i=1}^{n} \delta_i \left[\frac{\sum_{\{j|X_j \geq X_i\}} Z_{jg}^2}{\sum_{\{j|X_j \geq X_i\}} 1} - \left\{ \frac{\sum_{\{j|X_j \geq X_i\}} Z_{jg}}{\sum_{\{j|X_j \geq X_i\}} 1} \right\}^2 \right] = \sum_{k=1}^{K} d_k \left\{ \frac{n_{kg}}{n_k} - \left(\frac{n_{kg}}{n_k} \right)^2 \right\}$$

$$= \sum_{k=1}^{n} \frac{n_{kg}(n_k - n_{kg}) d_k}{n_k^2},$$

and off-diagonal elements

$$\sum_{i=1}^{n} \delta_i \left[\frac{\sum_{\{j|X_j \geq X_i\}} Z_{jg} Z_{jg'}}{\sum_{\{j|X_j \geq X_i\}} 1} - \left\{ \frac{\sum_{\{j|X_j \geq X_i\}} Z_{jg}}{\sum_{\{j|X_j \geq X_i\}} 1} \right\} \left\{ \frac{\sum_{\{j|X_j \geq X_i\}} Z_{jg'}}{\sum_{\{j|X_j \geq X_i\}} 1} \right\} \right]$$

$$= -\sum_{k=1}^{K} d_k \frac{n_{kg} n_{kg'}}{n_k^2}$$

2. The score test is approximately equivalent to log-rank test.

Trend test for multiple groups

1. Suppose that G groups are ordinal and are assigned score W as $W = w_g$, if the subject is in group g, where $g = 1, \cdots, G$. That is, $W = \sum_{g=1}^{G} w_g 1_{\{group=g\}}$. The Cox model is

$$h(t|g) = h_0(t) \exp(\beta w_g).$$

2. The score at $\beta = 0$ is

$$\sum_{i=1}^{n} \delta_i \left\{ W_i - \frac{\sum_{\{j|X_j \geq X_i\}} W_i}{\sum_{\{j|X_j \geq X_i\}} 1} \right\} = \sum_{g=1}^{G} w_g \sum_{k=1}^{K} \left\{ d_{kg} - d_k \frac{n_{kg}}{n_k} \right\},$$

which is the same as the log-rank test score.

Trend test for multiple groups (Continuing)

1. The information at $\beta = 0$ becomes

$$\begin{aligned}
&\sum_{i=1}^{n} \delta_i \left[\frac{\sum_{\{j|x_j \geq x_i\}} W_j^2}{\sum_{\{j|x_j \geq x_i\}} 1} - \left\{ \frac{\sum_{\{j|x_j \geq x_i\}} W_j}{\sum_{\{j|x_j \geq x_i\}} 1} \right\}^2 \right] \\
&= \sum_{k=1}^{K} d_k \sum_{g=1}^{G} w_g^2 \frac{n_{kg}}{n_k} - \sum_{k=1}^{K} d_k \left(\sum_{g=1}^{G} w_g \frac{n_{kg}}{n_k} \right)^2 \\
&= \sum_{k=1}^{K} \frac{d_k}{n_k} \sum_{g=1}^{G} n_{kg} \left(w_g - \frac{\sum_{g=1}^{G} w_g n_{kg}}{\sum_{g=1}^{G} n_{kg}} \right)^2 .
\end{aligned}$$

2. The score test for $H_0 : \beta = 0$ can be obtained.

Estimation of baseline survival function

1. The baseline hazard function can be estimated by Breslow estimator.

$$\hat{H}_0(t) = \sum_{\{k|\, T_k \leq t\}} h_k,$$

where

$$h_k = \frac{d_k}{\sum_{\{j|X_j \geq T_k\}} \exp(\hat{\beta} Z_j)},$$

$T_k,\ k = 1, \cdots, K$ are distinctive uncensored survival times, and d_k is the number of failures at T_k.

Estimation of baseline survival function (Continuing)

1. The baseline survival function can be estimated by

$$\prod_{\{k|T_k \leq t\}} (1 - h_k).$$

2. The survival function with covariate value $Z = Z_0$ can be estimated by

$$\hat{S}(t|Z_0) = \exp\{-H_0(t)\exp(\hat{\beta}Z_0)\}$$

or

$$\prod_{\{k|T_k \leq t\}} \left\{1 - h_k \exp(\hat{\beta}Z_0)\right\}.$$

SAS Proc phreg

```
PROC PHREG data=Myeloma;
        MODEL Time*censor(status)=covariates;
        BASELINE covariates=datasetname out=Pred1 survival=S
                lower=S_lower upper=S_upper;
        OUTPUT out=Outp xbeta=xb resmart=mart resdev=dev;
RUN;
```

1. covariates=datasetname: specify a dataset name containing covariate values used for computing the survival function.
2. xbeta: covariates times regression coefficients.
3. resmart: martingale residuals;
4. resdev: deviance residuals;
5. **See examples in SAS help file under PROC PHREG for more information**.

*Different ways of handling ties

1. Breslow approach has partial likelihood

$$\prod_{k=1}^{K} \frac{\exp\left\{\beta \sum_{\{j|X_j=T_k,\delta_j=1\}} Z_j\right\}}{\left\{\sum_{\{j|X_j\geq T_k\}} \exp(\beta Z_j)\right\}^{d_k}}.$$

   This approach is the same as the profile likelihood approach.

2. Efron's approach has partial likelihood

$$\prod_{k=1}^{K} \frac{\exp\left\{\beta \sum_{\{j|X_j=T_k,\delta_j=1\}} Z_j\right\}}{\prod_{l=1}^{d_k}\left\{\sum_{\{j|X_j\geq T_k\}} \exp(\beta Z_j) - \frac{l-1}{d_k}\sum_{\{j|X_j=T_k,\delta_j=1\}} \exp(\beta Z_j)\right\}}.$$

3. Cox approach (or discrete time approach)

$$\prod_{k=1}^{K} \frac{\exp\left\{\beta \sum_{\{j|X_j=T_k,\delta_j=1\}} Z_j\right\}}{\sum_{\sigma} \exp\left\{\beta \sum_{\{j|X_{\sigma(j)}=T_k,\delta_{\sigma(j)}=1\}} Z_{\sigma(j)}\right\}},$$

   where σ is a permutation of $\{j|X_j \geq T_k\}$.

Example 1: Analysis of the myeloma data

1. Fit a Cox regression model to the data.
2. Estimate baseline hazard function.
3. Perform variable selection in Cox regression.