Chapter 7

How to estimated $S(t)$ and $H(t)$ based on $h(t|z) = h_0(t)\exp\{\beta z\}$?

Preliminary of Bleslow's estimator:

Suppose there are two groups, group 1 and group 2. The hazard rates within these two groups at time t are $h_0(t)$ and $h_1(t)$ respectively. Suppose $h_1(t) = h(t|z=1) = h(t|z=0)\exp\{\beta\} = h_0(t)\exp\{\beta\}$. That is, according to the individuals who are still alive at time $t-\varepsilon$, the chance to die at time t of the individuals from group 1 ($z=1$) is $\exp\{\beta\}$ times as big than that of the individuals from group 0 ($z=0$). On the other hand, if there are n_0 individuals in group 0 who are still alive at time $t-\varepsilon$ and d_0 of them die at time t, a natural estimate of $h_0(t)$ is $\hat{h}_0(t) = d_0/n_0$. Similarly, we have $\hat{h}_1(t) = d_1/n_1$ where n_1 is the number of individuals in group 1 who are still alive at time $t-\varepsilon$ and d_1 is the number of death at time t among that n_1 individuals. Since $h_1(t) = h_0(t)\exp\{\beta\}$, an estimate of $h_0(t)$ based on n_1 and d_1 can be $d_1/(n_1\exp\{\beta\})$. Intuitively, an estimate of $h_0(t)$ based on n_0, n_1, d_0 and d_1 can be $(d_0+d_1)/(n_0+n_1\exp\{\beta\})$.

Bleslow's estimator:

Without loss of generality, assume there are n individuals with r ordered distinct death time $t_{(1)} < \cdots < t_{(k)}$ and d_i of them death at time $t_{(i)}$ ($i = 1, \cdots, k$). Suppose that $\hat{\beta}$ is the MLE from the data based on the Cox's model $h(t|z) = h_0(t)\exp\{\beta z\}$ by using partial likelihood technique. Breslow's estimator of $h_0(t_{(i)})$ is defined as the following

$$\hat{h}_0(t_{(i)}) = \frac{d_i}{\sum_{j\in R(t_{(i)})}\exp\{\hat{\beta}z_j\}}$$

where $R(t) = \{$ patients who are still alive at time $t-\varepsilon\}$. Moreover

$$\begin{aligned}\hat{H}_0(t) &= \sum_{t_{(i)}\le t}\hat{h}_0(t)\\ \tilde{S}_0(t) &= \exp\{-\hat{H}_0(t)\}\end{aligned}$$

or

$$\hat{S}_0(t) = \prod_{t_{(i)}<t}[1-\hat{h}_0(t_{(i)})] = \prod_{t_{(i)}<t}[1-\frac{d_i}{\sum_{j\in R(t_{(i)})}\exp\{\hat{\beta}z_j\}}].$$

Based on fully likelihood function.

Recall that the full likelihood function based on Cox's model is

$$L = \prod_{i=1}^{n}[S(y_i|z_i) - S(y_i+\varepsilon|z_i)]^{\delta_i}[S(y_i+\varepsilon|z_i)]^{1-\delta_i}$$

$$= \prod_{i=1}^{n} [S_0(y_i)^{\exp\{\beta z_i\}} - S_0(y_i+\varepsilon)^{\exp\{\beta z_i\}}]^{\delta_i} [S_0(y_i+\varepsilon)^{\exp\{\beta z_i\}}]^{1-\delta_i}.$$

As with the Kaplan-Meier estimate, it is clear that L is maximize by taking $S_0(t) = S_0(t_i)$ for $t_i < t \leq t_{i+1}$ and allowing probability mass to fall only at the observed distinct failure times $\dot{t}_1 < t_2 < \cdots < t_k$. Define t_{k+1} be a finite value greater than t_k, then the full likelihood function can be rewritten as

$$\begin{aligned} L &= \prod_{i=1}^{k} \prod_{j \in D_i} [S_0(t_i)^{\exp\{\beta z_j\}} - S_0(t_i+\varepsilon)^{\exp\{\beta z_j\}}] \prod_{j \in C_i} [S_0(t_i+\varepsilon)^{\exp\{\beta z_j\}}] \\ &= \prod_{i=1}^{k} \prod_{j \in D_i} [S_0(t_i)^{\exp\{\beta z_j\}} - S_0(t_{i+1})^{\exp\{\beta z_j\}}] \prod_{j \in C_i} [S_0(t_{i+1})^{\exp\{\beta z_j\}}] \end{aligned}$$

where D_i is the set of individuals who are died at time t_i and C_i is the set of individual who are censored in the interval (t_i, t_{i+1}). As with the Kaplan-Meier estimate again, the survival function lead to the consideration of the product of a series conditional survival function. That is

$$S(t_i) = P(T \geq t_i | T \geq t_{i-1}) P(T \geq t_{i-1} | T \geq t_{i-2}) \cdots P(T \geq t_1).$$

Let $\alpha_i = P(T \geq t_{i+1} | T \geq t_i)$ $(i = 1, \cdots, k)$ and $\alpha_0 = P(T \geq t_1) = 1$. Therefore, we have

$$S_0(t_i) = \prod_{t_l < t_i} \alpha_l$$

and the likelihood function L becomes

$$\begin{aligned} L(\alpha, \beta) &= \prod_{i=1}^{k} \prod_{j \in D_i} [(\prod_{t_l < t_i} \alpha_l)^{\exp\{\beta z_j\}} - (\prod_{t_l \leq t_i} \alpha_l)^{\exp\{\beta z_j\}}] \prod_{j \in C_i} [(\prod_{t_l \leq t_i} \alpha_l)^{\exp\{\beta z_j\}}] \\ &= \prod_{i=1}^{k} [\prod_{j \in D_i} (1 - \alpha_i^{\exp\{\beta z_i\}}) \prod_{l \in R_i - D_i} \alpha_i^{\exp\{\beta z_l\}}] \end{aligned}$$

where R_i is the set of individual who are at risk at time t_i. The estimation of the survival function can be carried out by joint estimation of the α's and β. More simply, however, we can take $\beta = \hat{\beta}$ as estimated from the partial likelihood function and then maximize $L(\alpha, \hat{\beta})$ with respect to $\alpha_1, \cdots, \alpha_k$. If only a single failure occurs at t_i, we have

$$\hat{\alpha}_i = \{1 - \frac{\exp\{\hat{\beta} z_i\}}{\sum_{j \in R_i} \exp\{\hat{\beta} z_j\}}\}^{\exp\{-\hat{\beta} z_i\}}.$$

Otherwise, an iterative solution is required; a suitable initial value for the iteration is the Breslow's estimator. Specifically,

$$1 - \alpha_i^{(0)} = d_i / [\sum_{j \in R_i} \exp\{\hat{\beta} z_j\}].$$

The maximum likelihood estimate of the baseline survivor function is

$$\hat{S}_0(t) = \prod_{t_i<t} \hat{\alpha}_i,$$

which, like the Kaplan-Meier estimate, is a step function with discontinuities at each observed failure time t_i. The corresponding estimate of the cumulative hazard function is

$$\hat{H}_0(t) = \sum_{t_i \le t} (1 - \hat{\alpha}_i).$$

Inference for time dependent covariate

If the value of the covariate change over time, the variable is called time dependent covariate in Cox's model. For an example, "blood pressure", "CD4 count" etc.

Extension of the Cox proportional hazards model for time-dependent variables

$$h(t|z\{t\}) = h_0(t) \exp\{\beta_1 z_1(t) + \cdots + \beta_p z_p(t)\}$$

where $h_0(t)$ is the hazard function for an individual for whom all the variables are equal to zero at time origin and remain at this same value through time.

Please notice that

(1) $h(t|z(t))/h_0(t) = \exp\{\beta_1 z_1(t) + \cdots + \beta_p z_p(t)\}$ does depend on time. (i.e. not a proportional hazard function)

(2) $H(t|z(t)) = \int_0^t h_0(y) \exp\{\beta_1 z_1(y) + \cdots + \beta_p z_p(y)\} dy \neq H_0(t) \exp\{\beta_1 z_1(t) + \cdots + \beta_p z_p(t)\}$. Also $S(t|z(t)) \neq S_0(t)^{\exp\{\beta_1 z_1(t) + \cdots + \beta_p z_p(t)\}}$.

Data: $(Y_i, \delta_i, Z_i(\cdot))$, $i = 1, \cdots, n$. The partial likelihood becomes

$$L(\beta) = \prod_{i=1}^{n} [\frac{\exp\{\beta z_i(y_i)\}}{\sum_{j \in R_i} \exp\{\beta z_j(y_i)\}}]^{\delta_i}.$$

All the technique about partial likelihood held the same, a slightly different is that the covariates $z(\cdot)$ is now a function of t.

Some application of time-dependent Cox's model

(1) To examine the assumption of PH in the Cox's model, a time dependent variable can be added to the model. Specifically, assume the original model is $h(t|z) = h_0(t) \exp\{\beta z\}$. Consider an extended model with time dependent covariate as follows

$$h(t|z) = h_0(t) \exp\{\beta z + \alpha z(t)\}.$$

We may test $\alpha = 0$ to see whether the data fit to the PH assumption or not. There are several choice for $z(t)$. Usually, we use $z(t) = zt$, $z(t) = z\sqrt{t}$, $z(t) = zt^2$ or $z(t) = z \ln t$.

(2) In many circumstances, the waiting time from the occurrence of some catastrophic event until a patient receive treatment may be strongly associated with the patient's survival time. For example, in the heart transplant data analysis, a patient should wait until he/she receives a suitable heart which making him/her eligible for a transplant. If the survival time of a patient is shorter than his/her waiting time, the patient will never get a transplant. If we compare the survival time of the patients with or without heart transplant directly, we will pretty sure to conclude that the survival time for heart transplant patients have longer survival time.

Let w_i be the waiting time for the ith individual. The model $h(t|z_i, w_i) = h_0(t)\exp\{\beta z_i + \alpha w_i\}$ is not the time-dependent Cox model that we want. The time-dependent Cox model for this case should be defined through indicator function. Define $z_i^*(t) = \begin{cases} 0, & \text{if } t \le w_i \\ 1, & \text{if } t > w_i \end{cases}$. Consider the model

$$h(t|z_i, w_i) = h_0(t)\exp\{\beta z_i + \alpha z_i^*(t)\}.....(7.1)$$

That is, the ith patient has hazard $\begin{cases} h_0(t)\exp\{\beta z_i\} & \text{, before heart transplant. ie } t \le w_i \\ h_0(t)\exp\{\beta z_i + \alpha\} & \text{, after heart transplant. ie } t > w_i \end{cases}$.

Suppose $z_i = \begin{cases} 0, & \text{if male} \\ 1, & \text{if female} \end{cases}$, then model (7.1) represents both male and female groups have the same effect of "heart transplant". However, the hazards within male and female groups may different. Specifically,

(1) a male before transplant, the hazard is $h_0(t)$

(2) a female before transplant, the hazard is $h_0(t)\exp\{\beta\}$

(3) a male after transplant, the hazard is $h_0(t)\exp\{\alpha\}$

(4) a female after transplant, the hazard is $h_0(t)\exp\{\beta + \alpha\}$

If we believe that the effect of transplant of male and female groups are different, we need to have an interaction term in the model. Specifically, the model becomes

$$h(t|z_i, w_i) = h_0(t)\exp\{\beta z_i + \alpha_1 z_i^*(t) + \alpha_2 z_i z_i^*(t)\}.$$

Note: For the heart transplant data, Cox suggests $h(t|w_i) = h_0(t)\exp\{\alpha_1 + \alpha_2\exp\{-\alpha_3(t - w_i)\}\}$.

Residuals:

Suppose (T_i, Z_i) $(i = 1, \cdots, n)$ are independently generated from the Cox PH model $h(t|z_i) = h_0(t)\exp\{\beta z_i\}$. Define $H_i(t) = H_0(t)\exp\{\beta z_i\}$ where $H_0(t) = \int_0^t h_0(u)du$.

Then, as we know, $H_i(T_i)$ follows $Exp(1)$. Later, suppose C_i are independently generated from some distribution F. Assume given Z_i, C_i and T_i are independent to each other. Now, if we can only observe (Y_i, Z_i, δ_i) where $Y_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$, we still have that $(H_i(Y_i), \delta_i)$ will merge to the censored data from $Exp(1)$.

The Cox-Snell residuals:

Define $\hat{H}_i(t) = \hat{H}_0(t)\exp(\hat{\beta}z_i)$ where $\hat{H}_0(t)$ and $\hat{\beta}$ are the estimates of $H_0(t)$ and β based on the Cox model. Then transform the data (Y_i, Z_i, δ_i) to $(\hat{H}_i(Y_i), \delta_i)$ $(i = 1, \cdots, n)$. Find the K-M estimate of the survival function based on $(\hat{H}_i(Y_i), \delta_i)$. Plots $\ln \hat{S}(t)$ against t and check by eyes to see whether it gives a straight line. Or equivalently, plots $\ln - \ln \hat{S}(t)$ against $\ln t$ to see whether it gives a straight line with unit slop and zero intercept by eyes.

Note: That's always show a straight line even the model is not correct.

Modified Cox-Snell residuals:

Since the exponential distribution has lack of memory properties. Moreover, the expectation of $Exp(1)$ is equal to 1. Thus, define $r_i = \hat{H}_i(Y_i) + (1 - \delta_i)$ and find the K-M estimate of the survival function based on r_i. Again, plots $\ln \hat{S}(t)$ vs t and check by eyes to see whether it gives a straight line. Later, Croulay and Hu suggest to define $r_i = \hat{H}_i(Y_i) + (1 - \delta_i)\ln 2$. The idea to use $\ln 2$ since the median of $Exp(1)$ is $\ln 2$.

Martingale residuals:

Define $rM_i = -(r_i - 1) = \delta_i - \hat{H}_i(Y_i)$. Some properties of rM_i are stated as follows

(1) $E(rM_i) = 0$;

(2) rM_i is not symmetric about 0;

(3) rM_i take value between $-\infty$ and 1;

(4) rM_i is negative for censored data.

Deviance residuals:

Define $rD_i = sign(rM_i)\{-2[rM_i + \delta_i \ln(\delta_i - rM_i)]\}^{1/2}$

Note: The sum of these residuals is not equal to zero.

Score residuals:

The ith score residual for covariate z_j (the jth covariate in the model), is defined as

$$rU_{ji} = \delta_i(z_{ji} - \bar{z}_{ji})$$

where $\bar{z}_{ji} = \sum_{k \in R(y_i)} z_{ji} \exp\{\hat{\beta} z_k\} / \sum_{k \in R(y_i)} \exp\{\hat{\beta} z_k\}$. It is trivial that $\sum_{i=1}^{n} rU_{ji} = 0$ for all j.

Note: When plot rU_{ji}, separating censored and uncensored data since $rU_{ji} = 0$ when $\delta_i = 0$.

Non-residual plots:

(1) Plot K-M for strata factor. For an example, plot 2 K-M curves for the treatment group and for the placebo group respectively. Then, check by eyes to see whether the PH assumption between these two groups is suitable or not.

(2) Let $z_i^* = \hat{\beta}_1 z_{i1} + \cdots \hat{\beta}_p z_{ip}$ (Xbeta in SAS and SPSS) where $\hat{\beta}_i$ $(i = 1, \cdots, p)$ are the estimates based on the Cox model

$$h(t|z_1, \cdots, z_p) = h_0(t) \exp\{\beta_1 z_1 + \cdots, \beta_p z_p\}.$$

Let q_1, q_2, q_3 be the first, second and third quartiles of $z_1^*, \cdots, z_n^*$. Plot 4 $\ln - \ln \hat{S}_{K-M}$ against t curves for the groups $z_i^* < q_1, q_1 \leq z_i^* < q_2, q_2 \leq z_i^* < q_3$ and $z_i^* \geq q_3$ respectively to see whether the these four curves are parallel or not.