Reliability Data and Survival Analysis
Homework 6

Chenghua Liu
liuch18@mails.tsinghua.edu.cn
Department of Computer Science
Tsinghua University

1

We showed in class that the score test for comparing two treatments and two sample log rank test are equivalent when there are no ties in the censored survival times. This equivalence is also true for the situation where there are more than two treatments. In this problem, you are asked to show part of this when there are three treatments. Namely, suppose we have the following proportional hazards model

$$\lambda(t \mid \cdot)=\lambda_0(t) \mathrm{e}^{Z_1 \beta_1+Z_2 \beta_2}$$

where Z_1 and Z_2 are two dummy variables created for 3 treatments:

$$Z_1=\begin{cases}1 & \text{if treatment } =1 \\ 0 & \text{otherwise}\end{cases}$$
$$Z_2=\begin{cases}1 & \text{if treatment } =2 \\ 0 & \text{otherwise}\end{cases}$$

Given the data $(x_i, \delta_i, z_{1i}, z_{2i})$ for $i=1,2,\ldots,n$, where there are NO ties in the censored survival times, show that the score vector$\left(\frac{\partial \ell}{\partial \beta_1}, \frac{\partial \ell}{\partial \beta_2}\right)^T$ for testing

$$H_0: \beta_1=\beta_2=0$$

is identical to the vector of the 3 -sample log rank test.

solution:

Let's directly prove the promotion situation where totally $p+1$ groups.Recalling Score test, Under $H_0: \beta_1=\beta_2 \cdots=\beta_p=0$, the score $U(0)$ (evaluated under H_0) has the distribution

$$U(0)^{\prime} J(0)^{-1} U(0) \stackrel{a}{\sim} \chi_p^2$$

Where the 0 is a $p \times 1$ vector and $J(0)$ is a $p \times p$ matrix.

Similar to the previous derivation in class, the element of score vector $U(\beta)$ has the expression

$$\frac{\partial \ell(\beta)}{\partial \beta_j} = \sum_u dN(u) \left[z_{I(u)j} - \bar{z}_j(u, \beta)\right]$$

where $z_{I(u)j}$ denotes the j th element of the covariate vector for the individual $I(u)$ who died at time u, and

$$\bar{z}_j(u, \beta) = \frac{\sum_{l=1}^n z_{lj} \exp\left(z_l^T \beta\right) Y_l(u)}{\sum_{l=1}^n \exp\left(z_l^T \beta\right) Y_l(u)} = \sum_{l=1}^n z_{lj} w_l, \quad w_l = \frac{\exp\left(z_l^T \beta\right) Y_l(u)}{\sum_{l=1}^n \exp\left(z_l^T \beta\right) Y_l(u)}$$

is the weighted average of the j th element of the covariate vector for the individuals at risk at time u.

If we denote

$$Z_{I(u)}^{p\times 1} = \begin{pmatrix} z_{I(u)1} \\ \vdots \\ z_{I(u)p} \end{pmatrix}, \quad \bar{Z}^{p\times 1}(u, \beta) = \begin{pmatrix} \bar{z}_1(u, \beta) \\ \vdots \\ \bar{z}_p(u, \beta) \end{pmatrix}$$

then the partial likelihood equation can be expressed as

$$U(\beta) = \sum_u dN(u) \left[Z_{I(u)}^{p\times 1} - \bar{Z}^{p\times 1}(u, \beta)\right]$$

Under $H_0 : \beta = 0, \bar{z}_j(u, 0)$ is simplified to be

$$\bar{z}_j(u, 0) = \frac{\sum_{l=1}^n z_{lj} Y_l(u)}{\sum_{l=1}^n Y_l(u)}$$

which is the proportion of individuals in group j among those at risk at time u. Since we only assume one death at time u, this proportion is the expected number of death for treatment j among those at risk at time u, under the null hypothesis of no treatment difference.

Therefore, the j-th element of $U(0)$ is the sum over the death times of the observed number of deaths from treatment j minus the expected number of deaths under the null hypothesis. This was the numerator of the multi-sample log rank test:$\sum_j \left(\mathbf{O}_j - \mathbf{E}_j\right)$

Now, let's prove the cov are equal.

$$J(\beta) = -\frac{\partial^2 \ell(\beta)}{\partial \beta^T \partial \beta} = -\left[\frac{\partial^2 \ell(\beta)}{\partial \beta_i \partial \beta_j}\right]_{p\times p}$$

is positive definite.And the (i, j) th element of $J(\beta)$ is

$$\begin{aligned} J_{i,j} &= \sum_u dN(u) \left[\frac{\sum_{l=1}^n z_{li} z_{lj} \exp\left(z_l^T \beta\right) Y_l(u)}{\sum_{l=1}^n \exp\left(z_l^T \beta\right) Y_l(u)} - \bar{z}_i(u, \beta) \bar{z}_j(u, \beta)\right] \\ &= \sum_u dN(u) \left[\frac{\sum_{l=1}^n \left(z_{li} - \bar{z}_i(u, \beta)\right)\left(z_{lj} - \bar{z}_j(u, \beta)\right) \exp\left(z_l^T \beta\right) Y_l(u)}{\sum_{l=1}^n \exp\left(z_l^T \beta\right) Y_l(u)}\right] \\ &= \sum_u dN(u) V_{i,j}(u, \beta) \end{aligned}$$

where $V_{i,j}(u,\beta)$ is the weighted sample covariance between the i th and j th element of the covariate vector among individuals at risk at time u with the weight $w_l = \frac{\exp\left(z_l^T \beta\right) Y_l(u)}{\sum_{l=1}^n \exp\left(z_l^T \beta\right) Y_l(u)}$

Therefore, the information matrix is

$$J^{p\times p}(\beta) = \sum_u dN(u) V(u,\beta)$$

Let β be zero vector,

$$\begin{aligned} V^{p\times p}(u,0) &= \left[\sum_{l=1}^n w_l \left(z_l - \bar{z}(u,0)\right)\left(z_l - \bar{z}(u,0)\right)^T\right] \\ &= \left[\frac{\sum_{l=1}^n Y_l(u)\left(z_l - \bar{z}(u,0)\right)\left(z_l - \bar{z}(u,0)\right)^T}{Y(u)}\right] \end{aligned}$$

$$\Rightarrow dN(u_j) V_{kl}^{(u_j)} = \begin{cases} -\frac{d_j Y_k(u_j) Y_l(u_j)(Y(u_j)-d_j)}{Y(u_j)^2 (Y(u_j)-1)} & \text{if } k \neq l \\ \frac{d_j Y_k(u_j)(Y(u_j)-d_j)(Y(u_j)-Y_k(u_j))}{Y(u_j)^2(Y(u_j)-1)} & \text{if } k = l \end{cases}$$

Which is as same as $\mathbf{V}_j$ in the multi-sample log rank test.

$$\left(\sum_j (\mathbf{O}_j - \mathbf{E}_j)\right)' \left(\sum_j \mathbf{V}_j\right)^{-1} \left(\sum_j (\mathbf{O}_j - \mathbf{E}_j)\right) \sim \chi_p^2 \quad \text{under } H_0 (\text{Total } p+1 \text{ groups})$$

2

The following data set contains the survival information from 5 patients and their smoking status z(1), z(3), z(4) and z(5) at each observed death time, 1, 3, 4 and 5 :

x (year)	δ	$z(1)$	$z(3)$	$z(4)$	$z(5)$
3	1	1	0	.	.
2	0	0	.	.	.
1	1	1	.	.	.
4	1	0	1	1	.
5	1	0	0	0	0

where x = time to failure or censoring (you NEED sort the data by x); δ = failure indicator: 1 = failure, 0 = censored ; z = 1 for smoking and z = 0 for nonsmoking. Assume a proportional hazards model with time-dependent covariate z(t)

$$\lambda(t \mid z(t)) = \lambda_0(t) e^{\beta z(t)}$$

(a) Construct the partial likelihood of β using this data set.

(b) Plot the log partial likelihood of β in the range of $[-4, 4]$.

(c) Find $\hat{\beta}$ that maximizes the log partial likelihood function and hence calculate the standard error of your estimate.

(d) Repeat part (c) using R.

solution:

(a)

The partial likelihood function of β for this model is given by

$$PL(\beta) = \prod_u \left[\frac{\exp\left(\beta^T Z_{I(u)}(u)\right)}{\sum_{l=1}^n \exp\left(\beta^T Z_l(u)\right) Y_l(u)} \right]^{dN(u)}$$

where $I(u)$ is the indicator variable that identifies the individual label $\in \{1, 2, \ldots, n\}$ for the individual who dies at time u.We sort the data by x.

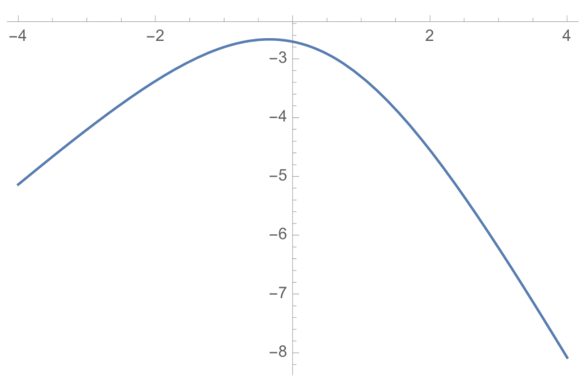
x (year)	δ	$z(1)$	$z(3)$	$z(4)$	$z(5)$
1	1	1	.	.	.
2	0	0	.	.	.
3	1	1	0	.	.
4	1	0	1	1	.
5	1	0	0	0	0

Then the partial likelihood function of β using the above data is

$$L(\beta; x, \delta, z(t)) = \frac{e^\beta}{3+2e^\beta} \times \frac{e^0}{2+e^\beta} \times \frac{e^\beta}{1+e^\beta} \times \frac{e^0}{e^0} = \frac{e^{2\beta}}{(3+2e^\beta)(2+e^\beta)(1+e^\beta)}$$

(b)

We plot the log partial likelihood of β in the range of $[-4, 4]$ as follow.



(c)

We get the value $\beta = 1.07307$ when PL is maximum -2.99552 by mathematica.And the standard error $\sqrt{1/\left[-\ell(\hat{\beta})''\right]} = \sqrt{1/0.654712} = 1.235$

(d)

The R code:

```r
library(numDeriv)
likelihood <- function(beta) {
          psi<-exp(beta)
          L = 2*beta-log(3+2*psi)-log(2+psi)-log(1+psi)
          return(L)
}
result<-optim(par=0,fn=likelihood,method="L-BFGS-B",
          control = list(fnscale=-1),
          lower=-4,upper = 4)
result$par
sqrt(1/(-hessian(func = likelihood,x=result$par)))
```

3

The following small data set contains the survival and covariate information from 4 patients

x	δ	z
2	0	1
5	0	0
2	1	1
3	0	0
1	1	0
4	1	1

where x = time to failure or censoring (you may sort the data by x); δ = failure indicator; 1 = failure, 0 = censored; z = observed value of covariate. Assume a proportional hazards model

$$\lambda(t \mid z) = \lambda_0(t)\exp(z\beta)$$

(a) Based on this data set, compute the maximum partial likelihood estimator $\hat{\beta}$ using the griding method, and then compute the Breslow estimator $\widehat{\Lambda}_0(t)$ of the cumulative baseline hazard function $\Lambda_0(t)$. Plot $\widehat{\Lambda}_0(t)$ as a function of time.

(b) Assume the proportional hazards model is correct, plot the estimated survival curve for S(t | z = 1) and S(t | z = 0) as a function of time t on the same graph. Again, please do it by hand.

solution:

(a)

From the table we know that the data (with ties) :

$$\text{Group 0: } 1, 3^+, 5^+ \Longrightarrow Z_i = 0$$
$$\text{Group 1: } 2, 2^+, 4 \Longrightarrow Z_i = 1$$

j	Ordered failure time X_i	Number at risk Group 0	Number at risk Group 1	Likelihood Contribution $e^{\beta S_j} / \sum_{\ell \in s(j,d_j)} e^{\beta S_{j\ell}}$
1	1	3	3	$e^0 / \left[3+3e^{\beta}\right]$
2	2	2	3	$e^{\beta} / \left[2+3e^{\beta}\right]$
3	4	1	1	$e^{\beta} / \left[1+1e^{\beta}\right]$

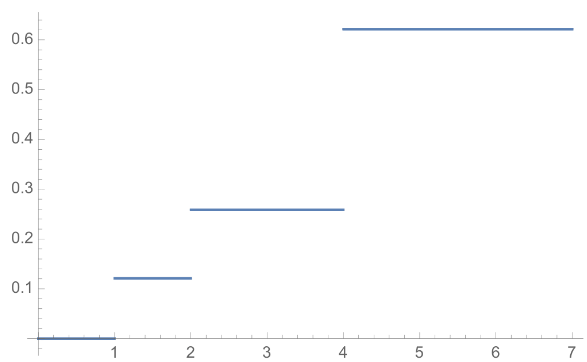
So the partial likelihood is $\frac{e^{2\beta}}{(3+3e^{\beta})(2+3e^{\beta})(1+e^{\beta})}$,we get the maximum is 0.0186189 when $\beta = 0.564351$ by mathematica.Namely, $\hat{\beta} = 0.564$. Then

$$e^{\widehat{\beta} Z} = \begin{cases} 1 & z = 0 \\ 1.7577 & z = 1 \end{cases}$$

Now we compute the Breslow estimator $\hat{\Lambda}_0(t) = \sum_{x<t} \left[\frac{dN(x)}{\sum_{i=1}^{n} \exp\left(\hat{\beta}^T Z_i\right) Y_i(x)}\right]$

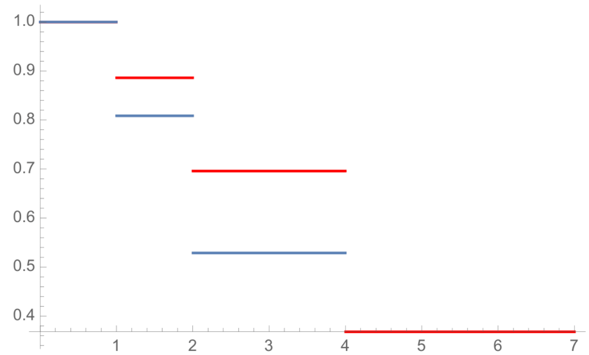
$$\begin{aligned} \widehat{\Lambda}_0(\tau_1) &= \tfrac{1}{3(1)+3(1.7577)} = 0.1209 \\ \widehat{\Lambda}_0(\tau_2) &= \tfrac{1}{2(1)+3(1.7577)} = 0.1376 \\ \widehat{\Lambda}_0(\tau_3) &= \tfrac{1}{1+1.7577} = 0.3629 \end{aligned} \quad \Rightarrow \quad \widehat{\Lambda}_0(t) = \begin{cases} 0 & t < 1 \\ 0.1209 & 1 \le t < 2 \\ 0.2585 & 2 \le t < 4 \\ 0.6214 & 4 \le t \end{cases}$$

We plot it as follow.



(b)

The estimated survival curve $\hat{S}(t \mid z) = \exp\left[-\hat{\Lambda}_0(t) \exp\left(\hat{\beta} z\right)\right]$.We let red curve be S(t | z = 0),blue curve be S(t | z = 1).Noted that curves are overlapping when S equals 1 and 0.


1.0
0.9
0.8
0.7
0.6
0.5
0.4
1
2
3
4
5
6
7