Midterm exam, Survival Analysis I, 2017 Spring [+27 points]

Name: 吴相花

+23

- **Not only answer but also derivations**

+4

Q1 [+4] Consider a two-parameter Weibull distribution whose survival function is given by $S_X(x) = \exp(-\lambda x^{\alpha})$, where $\lambda > 0, \alpha > 0$. Derive the likelihood function under the following cases.

+1

1) [+1] Left-truncated data ($y_{\ell i}, x_i$), subject to $y_{\ell i} \le x_i$, $i = 1, \ldots, n$, where $y_{\ell i}$ is a left-truncation time (likelihood must be simplified).

$$P(X = x_i \mid X \ge y_{\ell i}) = \frac{f(x_i)}{S(y_{\ell i})}$$

$$L = \prod_{i=1}^{n} \frac{f(x_i)}{S(y_{\ell i})} = \prod_{i=1}^{n} \frac{\lambda \alpha x_i^{\alpha-1} \exp(-\lambda x_i^{\alpha})}{\exp(-\lambda y_{\ell i}^{\alpha})} = \prod_{i=1}^{n} \lambda \alpha x_i^{\alpha-1} \exp(-\lambda(x_i^{\alpha} - y_{\ell i}^{\alpha}))$$

$$= \lambda^n \alpha^n \left(\prod_{i=1}^{n} x_i\right)^{\alpha-1} \exp\left(-\lambda \sum_{i=1}^{n} (x_i^{\alpha} - y_{\ell i}^{\alpha})\right)$$

+1

2) [+1] Interval-censored data (L_i, R_i), $i = 1, \ldots, n$.

$$P(X \in (L_i, R_i)) = S(L_i) - S(R_i)$$

$$L = \prod_{i=1}^{n} [S(L_i) - S(R_i)] = \prod_{i=1}^{n} [\exp(-\lambda L_i^{\alpha}) - \exp(-\lambda R_i^{\alpha})]$$

+1

3) [+1] Doubly-truncated data ($y_{\ell i}, x_i, y_{ri}$), subject to $y_{\ell i} \le x_i \le y_{ri}$, $i = 1, \ldots, n$, (likelihood must be simplified).

$$P(X = x_i \mid y_{\ell i} \le X \le y_{ri}) = \frac{f(x_i)}{S(y_{\ell i}) - S(y_{ri})}$$

$$L = \prod_{i=1}^{n} \frac{f(x_i)}{S(y_{\ell i}) - S(y_{ri})} = \prod_{i=1}^{n} \frac{\lambda \alpha x_i^{\alpha-1} \exp(-\lambda x_i^{\alpha})}{\exp(-\lambda y_{\ell i}^{\alpha}) - \exp(-\lambda y_{ri}^{\alpha})}$$

$$= \lambda^n \alpha^n \left(\prod x_i\right)^{\alpha-1} \exp\left(-\lambda \sum x_i^{\alpha}\right) \Big/ \prod_{i=1}^{n} [\exp(-\lambda y_{\ell i}^{\alpha}) - \exp(-\lambda y_{ri}^{\alpha})]$$

+1

4) [+1] n=4 patients whose age-at-death occurring in intervals (90, 120], (110, 115], (80, 100], (70, 75], subject to the entry condition Age>=50.

$$P(X \in (L_i, R_i) \mid X \ge y_{\ell i}) = \frac{S(L_i) - S(R_i)}{S(y_{\ell i})}$$

$$L = \frac{S(90) - S(120)}{S(50)} \cdot \frac{S(110) - S(115)}{S(50)} \cdot \frac{S(80) - S(100)}{S(50)} \cdot \frac{S(70) - S(75)}{S(50)}$$

$$= [\exp(-\lambda 90^{\alpha}) - \exp(-\lambda \cdot 120^{\alpha})] \cdot [\exp(-\lambda 110^{\alpha}) - \exp(-\lambda \cdot 115^{\alpha})]$$

$$\cdot [\exp(-\lambda 80^{\alpha}) - \exp(-\lambda 100^{\alpha})] \cdot [\exp(-\lambda 70^{\alpha}) - \exp(-\lambda \cdot 75^{\alpha})] \Big/ [\exp(-\lambda \cdot 50^{\alpha})]^4$$

+5

$S \prod_{i=} 1-$

Q2 [+6]

+1 1) [+1] Derive the product-limit form of a survival function $S(x) = \Pr(T > t)$.

$$S(t_i^*) = \frac{S(t_i^*)}{S(t_{i-1}^*)} \cdot \frac{S(t_{i-1}^*)}{S(t_{i-2}^*)} \cdot \cdots \cdot \frac{S(t_1^*)}{S(0)} \cdot S(0)$$

$$= P(T > t_i^* \mid T > t_{i-1}^*) \cdot P(T > t_{i-1}^* \mid T > t_{i-2}^*) \cdots P(T > t_1^* \mid T > 0) \quad \text{—— ①}$$

$$= P(T > t_i^* \mid T \geq t_i^*) \cdot P(T > t_{i-1}^* \mid T \geq t_{i-1}^*) \cdots P(T > t_1^* \mid T \geq t_1^*) \quad \text{—— ②}$$

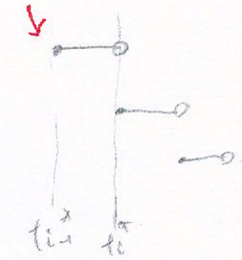
$$= \left[1 - \frac{f(t_i^*)}{S(t_i^*)}\right] \cdot \left[1 - \frac{f(t_{i-1}^*)}{S(t_{i-1}^*)}\right] \cdots \left[1 - \frac{f(t_1^*)}{S(t_1^*)}\right] = \prod_{j=1}^{i} \left[1 - \frac{f(t_j^*)}{S(t_j^*)}\right]$$

+1 2) [+1] What assumption is made to derive the form? Explain why the assumption is necessary.

$S(\cdot)$ is a right continuous, step function.

Without this assumption ① $\not\Rightarrow$ ②

Step function



+1 3) [+1] Define the Kaplan-Meier estimator for right-censored data (must define all notations clearly).

$$\hat{S}(t) = \prod_{t_i^* \leq t} \left[1 - \frac{d_i}{Y_i}\right]$$

d_i : numbers of deaths at time t_i^*.

Y_i : numbers of survivors who were at risk at time t_i^*

+1 4) [+1] Define the product-limit estimator for left-truncated data (define all notations).

$$\hat{S}(t) = \prod_{t_i^* \leq t} \left[1 - \frac{d_i}{Y_i}\right]$$

d_i : numbers of deaths at time t_i^*.

Y_i : numbers of survivors who has entered and were at risk at time t_i^*.

+1 5) [+1] Explain why the size of the risk set under left-truncation is small initially.

Formula can better explain Y_i.

In initial, a first patient enter into the study. But the other patients haven't entered into the study yet, so we cannot observe the event occurred from the other patients.

+0 6) [+1] Explain why one needs to estimate the conditional survival function under left-truncated data.

If event time $X < Y_L$, there is no information about X.

? Not clear enough

+5 **Q3 [+5]** The hazard function for the ovarian cancer patients follow the Cox model $h(t \mid z_1, z_2, z_3) = h_0(t)\exp(\beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3)$, where the gene expression values are

$$z_1 = \begin{cases} 1 & \text{high value of NCOA3} \\ 0 & \text{low value of NCOA3} \end{cases}, \qquad \beta_1 = 0.237$$

$$z_2 = \begin{cases} 1 & \text{high value of TEAD1} \\ 0 & \text{low value of TEAD1} \end{cases}, \qquad \beta_2 = 0.223$$

$$z_3 = \begin{cases} 1 & \text{high value of YWHAB} \\ 0 & \text{low value of YWHAB} \end{cases}, \qquad \beta_3 = 0.263$$

Compute the relative risk (RR).

+1 1) [+1] RR of (all three genes in high value) vs. (all three genes in low value).

$$RR = \frac{\exp(0.237 + 0.223 + 0.263)}{\exp(0)} = \exp(0.723) = 2.0606$$

+1 2) [+1] RR of (all three genes in high value) vs. (only NCOA3 in high value).

$$RR = \frac{\exp(0.237 + 0.223 + 0.263)}{\exp(0.237)} = \exp(0.486) = 1.6258$$

+1 3) [+1] RR of (only NCOA3 in high value) vs. (only YWHAB in high value).

$$RR = \frac{\exp(0.237)}{\exp(0.263)} = \exp(-0.026) = 0.974$$

+2 4) [+2] All RRs under different combinations risk factors (vs. the baseline risk). Make a table by sorting the RRs (from highest to lowest).

Order	RR	NCOA3	TEAD1	YWHAB
1	2.0606	High	High	High
2	1.6487	High	Low	High
3	1.6258	Low	High	High
4	1.5841	High	High	Low
5	1.301	Low	Low	High
6	1.267	High	Low	Low
7	1.2498	Low	High	Low
8	1	Low	Low	Low

Write -->>

+2 **Q4 [+4]** There are four stages of cancer (Stages I, II, III and IV). Define three different Cox models according to three different definitions of covariates.

1) [+1] Model 1

$$h(t|z) = h_0(t)\exp(\beta_2 z_2 + \beta_3 z_3 + \beta_4 z_4)$$

z_1 $z_2 = \begin{cases} 1, & \text{stage II} \\ 0, & \text{o.w} \end{cases}$

z_2 $z_3 = \begin{cases} 1, & \text{stage III} \\ 0, & \text{o.w.} \end{cases}$

z_3 $z_4 = \begin{cases} 1, & \text{stage IV} \\ 0, & \text{o.w} \end{cases}$

don't start from β_2
(start from β_1)

+1 RR (Stage IV vs. Stage I)= $\dfrac{\exp(\beta_4)}{\exp(0)} = \exp(\beta_4)$.

2) [+1] Model 2

$$h(t|z) = h_0(t)\exp(\beta_1 z_1 + \beta_3 z_3 + \beta_4 z_4)$$

$z_1 = \begin{cases} 1, & \text{stage I} \\ 0, & \text{o.w} \end{cases}$

All the same model as Model 1.

$z_3 = \begin{cases} 1, & \text{stage III} \\ 0, & \text{o.w} \end{cases}$

$z_4 = \begin{cases} 1, & \text{stage IV} \\ 0, & \text{o.w} \end{cases}$

RR (Stage IV vs. Stage I)= $\dfrac{\exp(\beta_4)}{\exp(\beta_1)} = \exp(\beta_4 - \beta_1)$.

3) [+1] Model 3

$$h(t|z) = h_0(t)\exp(\beta_1 z_1 + \beta_2 z_2 + \beta_4 z_4)$$

$z_1 = \begin{cases} 1, & \text{stage I} \\ 0, & \text{o.w} \end{cases}$

$z_2 = \begin{cases} 1, & \text{stage II} \\ 0, & \text{o.w.} \end{cases}$

$z_4 = \begin{cases} 1, & \text{stage IV} \\ 0, & \text{o.w.} \end{cases}$

$\dfrac{h(t|0,0,0)\cdot h(t|0,0,1)}{h(t|1,0,0)\, h(t|0,1,0)} =$

RR (Stage III-IV vs. Stage I-II)= $\dfrac{\exp(0)\exp(\beta_4)}{\exp(\beta_1+\beta_2)} = \exp(\beta_4 - \beta_1 - \beta_2)$

+1 4) [+1] Which model do you prefer? State your opinion.

Model. $h(t|z) = h_0(t)\exp(\beta_2 z_2 + \beta_3 z_3 + \beta_4 z_4)$.

$z_2 = \begin{cases} 1 & \text{stage II} \\ 0 & \text{o.w.} \end{cases}$ $z_3 = \begin{cases} 1 & \text{stage III} \\ 0 & \text{o.w.} \end{cases}$ $z_4 = \begin{cases} 1 & \text{stage IV} \\ 0 & \text{o.w.} \end{cases}$

Choose stage I as the baseline hazard.

It is nature to choose the lowest stage as the baseline hazard.

+7 **Q5 [+8]** Consider right-censored data (t_i, δ_i, z_i), $i = 1, \ldots, n$. Assume the Cox model $h(t \mid z_i) = h_0(t) e^{\beta z_i}$, where β is <u>one-dimensional</u> (β is a scalar). Define ordered times at deaths, $t_{(1)}^* \le t_{(2)}^* \le \cdots \le t_{(D)}^*$, where $D = \sum_{i=1}^{n} \delta_i$ (assume no ties).

+1 1) [+1] Define Cox's partial likelihood for β.

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta z_{(i)})}{\sum_{j \in R_i} \exp(\beta z_j)} \right)^{\delta_i}$$

$R_i = ?$

+1 2) [+1] Derive the score function

$$l(\beta) = \log L(\beta) = \sum_{i=1}^{n} \delta_i \left(\beta z_{(i)} - \log\left(\sum_{j \in R_i} \exp(\beta z_j) \right) \right)$$

$$U(\beta) = \frac{d}{d\beta} l(\beta) = \sum_{i=1}^{n} \delta_i \left(z_{(i)} - \frac{\sum_{j \in R_i} z_j \exp(\beta z_j)}{\sum_{j \in R_i} \exp(\beta z_j)} \right)$$

+1 3) [+1] Derive the information matrix

$$I(\beta) = -\frac{\partial}{\partial\beta} U(\beta) = \sum_{i=1}^{n} \delta_i \left[\frac{\left(\sum_{j \in R_i} z_j^2 \exp(\beta z_j)\right)\left(\sum_{j \in R_i} \exp(\beta z_j)\right) - \left(\sum_{j \in R_i} z_j \exp(\beta z_j)\right)^2}{\left(\sum_{j \in R_i} \exp(\beta z_j)\right)^2} \right]$$

$$= \sum_{i=1}^{n} \delta_i \left[\sum_{j \in R_i} z_j^2 \left(\frac{\exp(\beta z_j)}{\sum_{j \in R_i} \exp(\beta z_j)} \right) - \left(\sum_{j \in R_i} z_j \left(\frac{\exp(\beta z_j)}{\sum_{j \in R_i} \exp(\beta z_j)} \right) \right)^2 \right]$$

+1 4) [+1] Prove that the information matrix is nonnegative.

Define a discrete r.v $W_i \sim P(W_i = z_j) = \frac{\exp(\beta z_j)}{\sum_{j \in R_i} \exp(\beta z_j)} = P_j$

$\therefore \sum_{j \in R_i} z_j^2 P_j - \left(\sum_{j \in R_i} z_j P_j \right)^2 = Var(W_i) \ge 0 \quad \therefore I(\beta) = \sum_{i=1}^{n} \delta_i Var(W_i) \ge 0$

+2 4) [+2] Show that the log-rank statistics is equivalent to the score function at $\beta = 0$ under $z_i = 1$ (group 1) and $z_i = 0$ (group 2).

log-rank statistic

$= \sum_{i=1}^{D} \left[d_{i1} - Y_{i1} \frac{d_i}{Y_i} \right]$

$Y_{i1} \frac{d_i}{Y_i} \left(1 - \frac{d_i}{Y_i}\right) \cdot \frac{Y_i - Y_{i1}}{Y_i - 1}$

$$U(0) = \sum_{i=1}^{n} \delta_i \left(z_{(i)} - \frac{\sum_{j \in R_i} z_j}{\sum_{j \in R_i} 1} \right)$$

$$= \sum_{i=1}^{n} \delta_i \left(d_{i1} - \frac{Y_{i1}}{Y_i} \right)$$

$$= \sum_{i=1}^{D} \left[d_{i1} - \frac{Y_{i1}}{Y_i} \right] \quad (d_i = 1)$$

$=$ log-rank statistic of Z_1

+1 5) [+2] In the above setting, show that the variance of the log-rank statistics is equivalent to the information matrix.

$Var(Z_1)$

$\sum_{i=1}^{D} Y_{i1} \frac{d_i}{Y_i}\left(1 - \frac{d_i}{Y_i}\right) \cdot \frac{Y_i - Y_{i1}}{Y_i - 1}$

$= ?$

$$I(0) = \sum_{i=1}^{n} \delta_i \left[\sum_{j \in R_i} z_j^2 \cdot \frac{1}{\sum_{j \in R_i} 1} - \left(\sum_{j \in R_i} z_j \left(\frac{1}{\sum_{j \in R_i} 1} \right) \right)^2 \right]$$

$$= \sum_{i=1}^{n} \delta_i \left[\frac{Y_{i1}}{Y_i} - \left(\frac{Y_{i1}}{Y_i} \right)^2 \right]$$

$$= \sum_{i=1}^{n} \delta_i \left[\frac{Y_{i1} Y_{i2}}{Y_i^2} \right]$$

$$= \sum_{i=1}^{D} \frac{Y_{i1} Y_{i2}}{Y_i^2} \quad (d_i = 1)$$

$=$ log-rank statistic of $Var(Z_1)$ ← write clearly