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

Left censoring

Concepts and Likelihood

Analysis of Grouped Survival Time

Arbitrary Interval Censoring

Left censoring

1. In right censored data, we observed the minimum of the survival time and the censoring time, along with the censoring indicator. That is, $X = \min(T, C), \delta = 1_{\{T \leq C\}}$.
2. In left censored data, we observed the maximum of the survival time and the censoring time, along with the censoring indicator. That is, $X = \max(T, C), \delta = 1_{\{T \geq C\}}$.
3. It can be seen that, if we use the inverse transformation, the left censored data can be transformed into right censored data. That is, $Y = 1/X = min(1/T, 1/C), \delta = 1_{\{T \geq C\}} = 1_{\{1/T \leq 1/C\}}$.
4. Assume T and C are independent.

Estimating survival function with left-censoring data

1. For the left censored data $(X_i, \delta_i), i = 1, \cdots, n$, we can use the inverse transformation to convert it into the right censored data $(Y_i, \delta_i), i = 1, \cdots, n$.
2. We can use the Kaplan-Meier approach to estimating the survival function for $1/T$.
3. To obtain the survival function for T, notice that

$$P(T > t) = P(1/T < 1/t) = 1 - P(1/T \geq 1/t).$$

As a result, if we get the survival function for $1/T$ from the converted right-censored data, denote by $\hat{S}_{1/T}(u)$. The survival function for T can be obtained by

$$\hat{S}_T(t) = 1 - \hat{S}_{\frac{1}{T}}\left(\frac{1}{t}-\right).$$

Confidence intervals for survival function with left censoring data

1. We can construct a confidence interval for the survival function $S_{1/T}(t)$ directly using the approaches with the right censored data.
2. For a given t, let the low and upp limits of the confidence interval be respectively $L_{1/T}(t)$ and $U_{1/T}(t)$.
3. The confidence interval for $S_{1/T}(t)$ can be converted into confidence interval for $S_T(t)$ based on the relationship

$$\hat{S}_T(t) = 1 - \hat{S}_{\frac{1}{T}}\left(\frac{1}{t}-\right).$$

4. A confidence interval for $S_T(t)$ thus obtained is

$$\left[1 - U_{1/T}\left(\frac{1}{t}-\right), 1 - L_{1/T}\left(\frac{1}{t}-\right)\right].$$

Regression model with left censored data

1. For left censored data with covariates, the observed data are $(X_i, \delta_1, Z_i), i = 1, \cdots, n$, where $X_i = \max(T_i, C_i), \delta_i = 1_{\{T_i \geq C_i\}}$.
2. All the regression models for the survival time conditional on covariates can be used.
3. In fitting such models, we can first convert the data into right censored data as $(Y_i, \delta_i, Z_i), i = 1, \cdots, n$, and then apply the estimation approaches for the right censored data, where $Y_i = 1/X_i$.
4. Once we get the survival function estimator for $S_{1/T}(t \mid z)$, the survival function for T given Z can be obtained by

$$S_T(t \mid z) = P(T > t \mid Z = z) = P(\frac{1}{T} < \frac{1}{t} \mid z) = 1 - S_{1/T}(\frac{1}{t} - \mid z).$$

Interval censored data

1. If a survival time is only known to fall within an interval, it is called interval censored.
2. Interval censoring is a general concept of censoring. Interval censored data are denoted by $(L, R]$.
3. Previously encountered censoring mechanisms all fit into interval censoring
   - A right censored survival time X can be regarded as interval censored in $(X, +\infty)$.
   - A observed failure time X can be regarded as interval censored in $[X, X]$ or $(X-, X]$.
   - A left censored survival time X can be regarded as interval censored in $(0, X]$.

Censoring mechanism

1. Let the true survival time be T. Let L and R be random times such that $L \leq R$ two examination times.
2. If $T \leq L$, the survival time is left censored. Denote by $[0, L]$.
3. If $T > R$, the survival time is right censored. Denote by $(R, +\infty)$.
4. If $L < T \leq R$, the survival time is interval censored. Denote by $(L, R]$.
5. Assume that T is independent of (L, R) possibly conditional on covariates.

Likelihood derivation

1. For observed (L_i, R_i, Z_i),

$$\begin{aligned} & P(L_i < T_i \leq R_i, Z_i) \\ & \qquad = \quad P(L_i < T_i \leq R_i \mid L_i, R_i, Z_i) p(L_i, R_i \mid Z_i) p(Z_i). \end{aligned}$$

2. Let the survival time model be:

$$S_T(t|Z, \theta) = \exp \left\{ - \int_0^t h(u|Z, \theta) du \right\}.$$

   where θ is the model parameter.

3. When T_i is independent of $L_i, R_i)$ given Z_i,

$$\begin{aligned} P(L_i < T_i \leq R_i \mid L_i, R_i, Z_i) &= P(T_i > L_i \mid L_i, R_i, Z_i) \\ & \quad -P(T_i > R_i \mid L_i, R_i, Z_i) \\ &= S_T(L_i \mid Z_i, \theta) - S_T(R_i \mid Z_i, \theta). \end{aligned}$$

Model and likelihood

1. Contribution of an observation $\{(L_i, R_i], Z_i\}$ with $L_i < R_i$ to the likelihood is

$$P(L_i < T \leq R_i | Z_i) \quad = \quad S_T(L_i | Z_i, \theta) - S_T(R_i | Z_i, \theta).$$

2. Contribution of an observed failure time $L_i = R_i$ to the likelihood is

$$h(T_i | Z_i, \theta) S(T_i | Z_i, \theta).$$

3. Likelihood is the product of the contributions from all subjects.

$$\begin{aligned} L(\theta) \quad = \quad & \prod_{i=1}^{n} \{S_T(L_i | Z_i, \theta) - S_T(R_i | Z_i, \theta)\}^{1_{\{L_i < R_i\}}} \\ & \times \{h(T_i | Z_i, \theta) S(T_i | Z_i, \theta)\}^{1_{\{L_i = R_i\}}} . \end{aligned}$$

For the exponential model

1. The exponential model has

$$P(T > t|Z) = \exp\{-t\exp(\beta_0 + \beta Z)\}.$$

2. Contribution of an observation $\{(L_i, R_i], Z_i\}$ with $L_i < R_i$ to the likelihood is

$$\begin{aligned} P(L_i < T \leq R_i|Z_i) &= \exp\{-L_i\exp(\beta_0 + \beta Z_i)\} \\ &\quad -\exp\{-R_i\exp(\beta_0 + \beta Z_i)\}. \end{aligned}$$

3. Contribution of an observed failure time $L_i = R_i$ to the likelihood is

$$\exp(\beta_0 + \beta Z_i)\exp\{-L_i\exp(\beta_0 + \beta Z_i)\}.$$

For the Weibull model

1. The Weibulll model has

$$P(T > t|Z) = \exp\left\{-t^{\frac{1}{\sigma}} \exp\left(\frac{\beta_0 + \beta Z}{\sigma}\right)\right\}.$$

2. Contribution of an observation $\{(L_i, R_i], Z_i\}$ with $L_i < R_i$ to the likelihood is

$$\begin{aligned} P(L_i < T \leq R_i|Z_i) &= \exp\left\{-L_i^{\frac{1}{\sigma}} \exp\left(\frac{\beta_0 + \beta Z_i}{\sigma}\right)\right\} \\ &\quad - \exp\left\{-R_i^{\frac{1}{\sigma}} \exp\left(\frac{\beta_0 + \beta Z_i}{\sigma}\right)\right\}. \end{aligned}$$

3. Contribution of an observed failure time $L_i = R_i$ to the likelihood is

$$\frac{1}{\sigma} L_i^{\frac{1}{\sigma}-1} \exp\left(\frac{\beta_0 + \beta Z_i}{\sigma}\right) \exp\left\{-L_i^{\frac{1}{\sigma}} \exp\left(\frac{\beta_0 + \beta Z_i}{\sigma}\right)\right\}.$$

For the Cox regression model

1. The Cox proportional hazards regression model has

$$P(T > t|Z) = \exp\{-\Lambda(t)\exp(\beta Z)\}.$$

2. Contribution of an observation $\{(L_i, R_i], Z_i\}$ with $L_i < R_i$ to the likelihood is

$$\begin{aligned} P(L_i < T \leq R_i|Z_i) &= \exp\{-\Lambda(L_i)\exp(\beta Z_i)\} \\ &\quad - \exp\{-\Lambda(R_i)\exp(\beta Z_i)\}. \end{aligned}$$

3. Contribution of an observed failure time $L_i = R_i$ to the likelihood is

$$\lambda(L_i)\exp(\beta Z_i)\exp\{-\Lambda(L_i)\exp(\beta Z_i)\}.$$

Parameter estimation and inference

1. In general:
   - The likelihood can be maximized to obtain the parameter estimator.
   - The variance of the parameter estimator can be obtained by the inverse of the information matrix from the likelihood.
   - Hypothesis tests can be performed by the likelihood ratio test, or Wald, or score test.
2. Cautions:
   - Conditions that restrict the arbitrariness of the censoring is required for the likelihood to bahave well.
   - For the Cox regression model, partial likelihood can no longer be obtained in general. The likelihood is relied upon.

Grouped survival times

If we can divide the time axis into consecutive disjoint intervals and any failure is known to occur in only one of the intervals, (i.e. no uncertainty with regard to which interval any failure in the data occurred), grouped survival time analysis may be used for interval censored data.

For example, if the time intervals are $(0, t_1], \cdots, (t_{K-1}, t_K]$, and (t_K, ∞). We can have interval censored data $(L_i, R_i]$ being one of the intervals. No such interval censored data as $(t_2, t_4]$ or $(0, t_2]$ or (t_2, ∞) unless $t_2 = t_K$, is allowed.

Grouped survival times (continuing)

1. Let $(0, t_1], \cdots, (t_{K-1}, t_K]$, (t_K, ∞) be the intervals.
2. Data are summarized in grouped form as follows

Interval	# of failues	# of non-failures
$(0, t_1]$	m_1	$n_1 = m_2 + m_3 + \cdots$
$\vdots$	$\vdots$	$\vdots$
$(t_{K-1}, t_K]$	m_K	$n_K = m_{K+1}$
(t_K, ∞)	m_{K+1}	$n_{K+1} = 0$

   where non-failure means survived beyond the right bound of the interval.

3. An interval censored data $(L_i, R_i] = (t_{j-1}, t_j]$ contributes 1 to the number of non-failures and 0 to the number of failures for each of the intervals $(0, t_1], \cdots, (t_{j-2}, t_{j-1}]$, 0 to the number of non-failures and 1 to the number of failures for the interval $(t_{j-1}, t_j]$, and 0 to the number of non-failures and 0 to the number of failures for each of the intervals $(t_j, t_{j+1}], \cdots, (t_K, \infty)$.

Analysis of grouped survival times

1. For the first time interval, a failure contributes to the likelihood

$$P(T \leq t_1 | Z) = 1 - \exp\left\{-\Lambda_0(t_1)e^{\beta z}\right\},$$

and a nonfailure contributes to the likelihood

$$P(T > t_1 | z) = \exp\left\{-\Lambda_0(t_1)e^{\beta z}\right\}.$$

2. For the second time interval, a failure contributes to the likelihood

$$P(T \leq t_2 | T > t_1, Z) = 1 - \exp\left[-\{\Lambda_0(t_2) - \Lambda_0(t_2)\}e^{\beta z}\right],$$

and a nonfailure contributes to the likelihood

$$P(T > t_2 | T > t_1, z) = \exp\left[-\{\Lambda_0(t_2) - \Lambda_0(t_1)\}e^{\beta z}\right].$$

Analysis of grouped survival times (continuing 1)

1. For the Kth time interval, a failure contributes to the likelihood

$$P(T \leq t_K | T > t_{K-1}, Z) = 1 - \exp\left[-\{\Lambda_0(t_K) - \Lambda_0(t_{K-1})\} e^{\beta z}\right],$$

and a nonfailure contributes to the likelihood

$$P(T > t_K | T > t_{K-1}, z) = \exp\left[-\{\Lambda_0(t_K) - \Lambda_0(t_{K-1})\} e^{\beta z}\right].$$

2. For the last time interval, a failure does not contribute anything to the likelihood because

$$P(T < \infty | T > t_K, Z) = 1.$$

Analysis of grouped survival times (continuing 2)

1. If we recode a failure at interval $(t_{k-1}, t_k]$ as $Y_k = 1$ and a non-failure as $Y_k = 0$, then

   $$P(Y_k = 1|Z) = 1 - \exp(-e^{\beta_{0k} + \beta Z}),$$

   where $\beta_{0k} = \log\{\Lambda_0(t_k) - \Lambda_0(t_{k-1})\}$.
2. The grouped survival data may be analyzed by a series of binary regression with the complementary log-log link function.
3. Such analyses can be performed one-by-one for each interval.

Analysis of grouped survival times (continuing 3)

1. The separate analyses are not efficient because for different intervals, different slopes are estimated. A better way is to combine the analyses for all the intervals.
2. To perform the combined analysis, let the first interval be the base-interval, and define a dummy variable to represent each of the subsequent intervales.

Interval	W_1	$\cdots$	W_{K-1}
$(0, t_1]$	0	$\cdots$	0
$(t_1, t_2]$	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(t_{K-1}, t_K]$	0	$\cdots$	1

Analysis of grouped survival times (continuing 4)

The data can now be regrouped in binary regression form as

Interval	Y	W_1	$\cdots$	W_{K-1}	frequency
$(0, t_1]$	1	0	$\cdots$	0	m_1
$(0, t_1]$	0	0	$\cdots$	0	n_1
$(t_1, t_2]$	1	1	$\cdots$	0	m_2
$(t_1, t_2]$	0	1	$\cdots$	0	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(t_{K-1}, t_K]$	1	0	$\cdots$	1	m_K
$(t_{K-1}, t_K]$	0	0	$\cdots$	1	n_K

Analysis of grouped survival times (continuing 5)

1. The model to fit to the regrouped data is

$$P(Y = 1|Z, W) = 1 - \exp\left\{-e^{g(W,Z)}\right\},$$

where

$$g(W, Z) = \beta_0 + \beta_{01} W_1 + \cdots + \beta_{0(K-1)} W_{K-1} + \beta Z.$$

2. Previous separate analyses correspond to fitting a model with

$$\begin{aligned} g(W, Z) \quad = \quad & \beta_0 + \beta_{01} W_1 + \cdots + \beta_{0(K-1)} W_{K-1} + \beta Z \\ & + \beta_{11} Z W_1 + \cdots + \beta_{1(K-1)} Z W_{K-1}. \end{aligned}$$

This model has the interpretation of time-varying coefficients for covariate Z.

Analysis of grouped survival times II

1. The survival distribution

$$P(T > t|Z) = \frac{1}{1 + \Lambda(t)e^{\beta Z}},$$

is called the proportional odds ratio model because

$$\log \frac{P(T \leq t|Z)}{P(T > t|Z)} = log\Lambda(t) + \beta Z.$$

2. If the original survival time follows the distribution, then

$$\begin{aligned} P(Y_k = 1|Z) &= P(T \leq t_k|T > t_{k-1}, Z) \\ &= 1 - P(T > t_k|T > t_{k-1}, Z) \\ &= \frac{\{\Lambda(t_k) - \Lambda(t_{k-1})\}e^{\beta Z}}{1 + \Lambda(t_k)e^{\beta Z}}. \end{aligned}$$

Analysis of grouped survival times II (continuing)

1. Let $\beta_{0k} = \log\{\Lambda(t_k) - \Lambda(t_{k-1})\}$, $k = 1, \cdots, K$. Then

   $$P(Y_k = 1|Z) = \frac{\exp(\beta_{0k} + \beta Z)}{1 + \exp(\beta_{01} + \beta Z) + \cdots + \exp(\beta_{0k} + \beta Z)}.$$

   and

   $$P(Y_k = 0|Z) = \frac{1 + \exp(\beta_{01} + \beta Z) + \cdots + \exp(\beta_{0(k-1)} + \beta Z)}{1 + \exp(\beta_{01} + \beta Z) + \cdots + \exp(\beta_{0k} + \beta Z)}.$$

2. If there are only two intervals, that is, $K = 1$, analysis of the grouped survival time by the proportional odds ratio model is equivalent to analyzing by the logistic regression model.
3. For $K \geq 2$, the proportional odds ratio model is not equivalent to logistic regression model for each Y_k.

Logistic regrssion analysis of grouped survival times

1. The logistic regression model may be applied directly to each Y_k as
$$P(Y_k = 1|Z) = \frac{\exp(\beta_{0k} + \beta Z)}{1 + \exp(\beta_{0k} + \beta Z)}.$$
2. Under such models,
$$\begin{aligned} P(T > t_k|Z) &= \prod_{j=1}^{k} P(T > t_j|T > t_{j-1}, Z) = \prod_{j=1}^{k} P(Y_j = 0|Z) \\ &= \left[\prod_{j=1}^{k} \{1 + \exp(\beta_{0j} + \beta Z)\}\right]^{-1}. \end{aligned}$$
3. This model can be interpreted as
$$\frac{d\Lambda(t|Z)}{1 - d\Lambda(t|Z)} = \frac{d\Lambda_0(t)}{1 - d\Lambda_0(t)} e^{\beta Z}$$
When the time interval becomes small, it reduces to the proportional hazards model.

Arbitrary Interval Censoring

1. In general, arbitrary interval censored data may not be reduced to grouped survival times. The grouped survival time analysis approach cannot be used.
2. There is no available software to handle the Cox regression model directly for general interval censored data. We have two options.
   - When only one discrete covariate is involved, there is a SAS macro to perform the analysis.
   - In general, we may manipulation Proc nlmix in SAS to fit a Cox regression model to interval censored data.

Fit Cox model using proc nlmixed:Data manipulation

Proc phreg does not allow interval censoring. Data can be manipulated to use proc nlmixed to fit a cox model to the interval censored data.

1. Split “true” interval censoring into two observations based on

   $$P(L < T < R|z) = P(T > L|z)\{1 - P(T > R|T > L, z)\}.$$

   One observation contribute to the likelihood with $P(T > L|z)$, which is equivalent to a right censoring with the lower limit of the interval censoring as the censoring time. The other observation contribute to the likelihood with $1 - P(T > R|T > L|z)$.
2. Define binary outcome $Y = 1$ for the split observation with $1 - P(T > R|T > L|z)$. Define binary outcome $Y = 0$ for the split observation with $P(T > L|z)$.

Fit Cox model using proc nlmixed:Data manipulation (continuing)

1. Let $0 = t_0 \leq t_1 \leq \cdots \leq t_K < +\infty$ denote all the possible limits of the interval, right, and left censoring in the data set. Model
$$P(T > t_k \mid T > t_{k-1}, z) = \exp\{-\lambda_k \exp(\beta z)\}.$$
2. For the observations having contribution of the form $P(T > L|z)$, rewrite
$$\begin{aligned} P(T > L|z) &= \prod_{k=1}^{K} \{P(T > t_k|T > t_{k-1}, z)\}^{1_{\{L \geq t_k\}}} \\ &= \exp\{-\sum_{k=1}^{K} \lambda_k 1_{\{L \geq t_k\}} \exp(\beta z)\}, \end{aligned}$$

Fit Cox model using proc nlmixed:Data manipulation (continuing 2)

1. Rewrite

$$
\begin{aligned}
P(T > R | T > L | z) &= \prod_{k=1}^{K} \{P(T > t_k | T > t_{k-1}, z)\}^{1_{\{R \geq t_k > L\}}} \\
&= \exp\{-\sum_{k=1}^{K} \lambda_k 1_{\{R \geq t_k > L\}} \exp(\beta z)\}.
\end{aligned}
$$

2. Define $d_{ik} = 1_{\{L_i \geq t_k\}}$ for the observations in the form $T > L$ (include the observation split from the "true" interval censoring). Define $d_{ik} = 1_{\{R_i \geq t_k > L_i\}}$ for the observation in the form $L < T < R$ (the one derived from the "true" interval censoring).
3. For right censoring, $Y = 0$. Define $d_{ik} = 1_{\{L_i \geq t_k\}}$.
4. For left censoring, $Y = 1$. $d_{ik} = 1_{\{R_i \geq t_k\}}$.

Fit Cox model using proc mixed for ulcer data

1. $t_0 = 0 < t_1 = 2 < t_2 = 3 < t_3 = 5 < t_4 = 6 < t_5 = 7 < t_6 = 10 < t_7 = 12 < +\infty$ and $K = 7$.
2. The "true" interval censoring that need to be split are subjects $1, 17, 18, 30, 32, 37$.
3. Outcome Y and covariates $d_1, \cdots, d_7$ need to be defined.
4. Examples of data manipulation:

Subject	Last visit	Result	Interval censoring
1	7	2	$(6, 7]$
2	12	1	$(12, +\infty)$
15	6	2	$(0, 6]$
16	6	1	$(6, +\infty)$
17	10	2	$(6, 10]$

Fit Cox model using proc nlmixed for ulcer data

The example data manipulations are as follow

Subject	Interval	Time	Y	$(d_1, d_2, d_3, d_4, d_5, d_6, d_7)$
1	$(6, 7]$	$(0, 6]$	0	(1,1,1,1,0,0,0)
		$(6, 7]$	1	(0,0,0,0,1,0,0)
2	$(12, +\infty)$	$(0, 12]$	0	(1,1,1,1,1,1,1)
15	$(0, 6]$	$(0, 6]$	1	(1,1,1,1,0,0,0)
16	$(6, +\infty)$	$(0, 6]$	0	(1,1,1,1,0,0,0)
17	$(6, 10]$	$(0, 6]$	0	(1,1,1,1,0,0,0)
		$(6, 10]$	1	(0,0,0,0,1,1,0)

The model to fit is

$$P(Y = 1|D, Z) = 1 - \exp\left\{ -\sum_{j=1}^{7} d_j \lambda_j \exp(\beta Z) \right\}.$$

SAS statement for fit proc nlmixed

Let $dur = duration - 1$, $trt = 1$ if $treatment = A$ and $trt = 0$ if $treatment = B$.

```
PROC      NLMIXED;
PARMS     lam1=0.01 lam2=0.01 lam3=0.01 lam4=0.2
          lam5=0.01 lam6=0.01 lam7=0.2
          beta1=0 beta2=0 beta3=0;
          sumlam=lam1*d1+lam2*d2+lam3*d3+lam4*d4
          +lam5*d5+lam6*d6+lam7*d7;
          sumbz=beta1*age+beta2*dur+beta3*trt;
          p=1-exp(-sumlam*exp(sumbz));
MODEL     y∼ binary(p);
BOUNDS    lam1-lam7> 0;
RUN;
```

Recurrence of an ulcer

1. Study design: Subjects are randomized into treatment groups A and B. Subjects are scheduled to be examined by endoscope at months 6 and 12 after randomization to see if ulcer occurred. If symptoms indicated an ulcer had occurred before the scheduled check-up time, an endoscope examine will be performed.
2. The duration indicator if the previous ulcer was occurred less than 5 years (denoted by 1)before the randomization or at least 5 years (denoted by 2).
3. Age is the age at the randomization.
4. Once a subject is detected to have an reoccurrence of ulcer, the subject is treated and removed from the study.
5. The result of the endoscope examination is negative (Result=1) at month 6 if the subject is examined at a time greater than month 6 (Result=1 or 2).

Parametric survival model for recurrence of an ulcer

1. SAS proc lifereg accommodates interval censoring in the model statement as
   Model (L,R)=covariates;
2. Left censored survival time is coded as L missing and Right censored survival time is coded as R missing in SAS Proc lifereg.
3. See SAS code for fitting models to the ulcer data.

Example: Breast retraction in treating Breast Cancer

1. In treatment of early stage breats cancer, a tumourectomy followed by radiation therapy is often used as an alternative to mastectomy. Chemotherapy may also be combined with radiation therapy. An adverse effect of the treatment is the breast retraction.
2. The study is to compare two treatment options (radiation therapy alone versus the radiation therapy combined with chemotherapy) in terms of the time lead to the breast retraction.
3. The data has 5 left censoring, 38 right censoring, and 51 "true" interval censoring.

Example: Breast retraction in treating Breast Cancer (continuing)

1. To fit parametric survival model using proc lifereg.
2. To fit Cox proportional regression model by SAS proc ICphreg. It works like the proc lifereg, but need to specify baseline smoothing method.
3. Alternatively, performing data manipulation and then using proc nlmixed.