SURVIVAL ANALYSIS

(Lecture Notes)

by Qiqing Yu

Version 6/8/2010

This course will cover parametric, non-parametric and semi-parametric maximum likelihood estimation, in the Cox regression model and the linear regression model, with complete data and various types of censored data. The right censorship model, double censorship model, the mixed case interval censorship model, the mixture censorship model and the left-truncation model will be used to formulate the censored or truncated-data. Multivariate censorship models and the two-sample problem may also be introduced.

Homework due: Tuesday;

Final: closed book exam.

Grading: $70\%hw + 30\%final$;

Reference Textbooks:

1. Analysis of survival data, by Cox and Oakes.
2. Survival Analysis, by Rupert G. Miller, JR.
3. The Statistical Analysis of Interval-censored Failure Time Data, by J. Sun.

Chapter 1. Introduction.

§1. **Two main characters of survival analysis**.

Suppose X is a random variable,
with the cumulative distribution function (cdf),

$$F(x) = P(X \le x).$$

Assume X_1, ..., X_n are independently and identically distributed
(i.i.d.) from F.

Statistical inferences: $\begin{cases} estimation \\ testing \ \ hypotheses, \end{cases}$

on $S = 1 - F$ or the parameters in a certain model.

In particular, there are three different types:

parametric, *i.e.* $F(x) = F_o(x;\theta)$, where $\theta \in \Theta$, a parameter space;

nonparametric, *i.e.*, F is only known to be a c.d.f.

semi-parametric, *i.e.*, in between the above two.

In survival analysis, X is often

time to death of a patient after a treatment,

time to failure of a part of a system, etc.

We assume $X \ge 0$, referred as *survival time or failure time.*

By S, it is much intuitive for doctors to compare different treatments
or different systems,

the larger the S, the better.

Main character of survival analysis:

Incomplete data.

Definition: An observation on X_i is called

$$\begin{cases} \text{complete (exact, or uncensored)} & \text{if the exact value of } X_i \text{ is observed;} \\ incomplete & \text{o.w.;} \end{cases}$$

A data set is called $\begin{cases} complete & \text{if all observations are exact;} \\ \text{incomplete} & \text{otherwise.} \end{cases}$

An incomplete observation on X_i is called *interval censored* (IC) if $X_i \in I_i$, an interval with endpoints L_i and R_i being observed.

An IC observation can thus be represented by an **extended** random vector (L_i, R_i).

(L_i, R_i) is called

$$\begin{cases} right\ censored\ (RC) & \text{if } R_i = \infty \\ left\ censored\ (LC) & \text{if } L_i = -\infty \\ strictly\ interval\ censored\ (SIC) & \text{if } 0 < L_i < R_i < \infty. \end{cases}$$

§1.2. **Right censoring**.

§1.2.1.

Representations of an RC observation:

(L_i, R_i) —- a vector,
$I_i = (L_i, R_i)$ —- an interval,
$(L_i, 0)$ —- a vector,
L_i+,

Representations of an exact observation:

$(L_i, R_i) = (X_i, X_i)$
$I_i = [X_i, X_i]$ —- an interval,
$(L_i, 1)$ —- a vector,
L_i.

Definition: A RC data set —- observations are either exact or RC, but there exist exact observations.
Representations of RC data:

(L_i, R_i), $i = 1, ..., n$, —- random vectors,
where $(L_i, R_i) = \begin{cases} (X_i, X_i) & \text{if the observation is exact,} \\ (L_i, \infty) & \text{if the observation is RC,} \end{cases}$
I_i, $i = 1, ..., n$, —— random intervals,
where $I_i = \begin{cases} [X_i, X_i] & \text{if the observation is exact,} \\ (L_i, \infty) & \text{if the observation is RC.} \end{cases}$
(L_i, δ_i), $i = 1, ..., n$, —- random vectors,
where $\delta_i = \mathbf{1}_{(\text{the i-th observation is exact})}$, and
$\mathbf{1}_A = \begin{cases} 1 & \text{if } A \text{ happens} \\ 0 & \text{o.w.} \end{cases}$ is the indicator function of the event A.
L_i+ or L_i, $i = 1, ..., n$.

Example of RC data:

1. Mortality data (population census,
   for computing the life expectancy of the population).
   Let X_i be the age at which the i-th person died.
   Then at a census, we either knew X_i if the person died
   or knew L_i+ if he/she was alive, where L_i was his/her age then.
2. Type I censoring.
   Each individual was followed by a fixed time c.
   Each X was recorded unless $X > c$.

$$X_{(1)},\ ...,\ X_{(i)}, c+,\ ...,\ c+$$

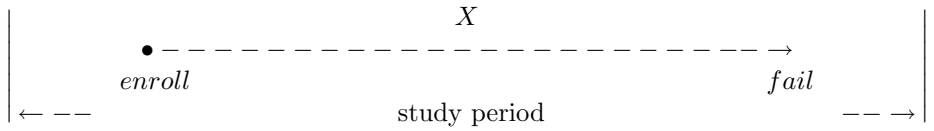
   where $X_{(1)} \le \cdots \le X_{(i)}$ are order statistics of observed exact values of X_1, ..., X_n.
3. Type II censoring.
   Observation ceases after a predetermined number d of failures.

$$X_1, \cdots,\ X_d,\ c,\ ...,\ c, c+,\ ...,\ c+$$

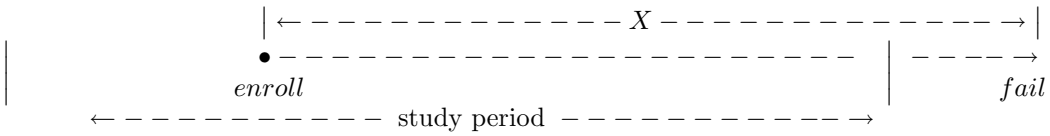
   where $X_d = c$.
4. Random censoring.
   In a medical follow-up study of 5 years, n cancer patients are enrolled (not necessary from the beginning). X is the time to death of a patient since a certain treatment (after the enrollment). We either know X or know $X > 5 - B$, where B is the beginning time of the treatment for the individual since the start of the study.

Graphical illustration for X and Y

Exact observation



Right-censored observation



Leukaemia data

Gehan, 1965 recorded times of remission of leukaemia patients.
Some were treated with drug 6-mercaptopurine (6-MP),
the others were serving as a control.
Table 1.1 (Cox and Oakes (1984) (pages 7,8)). Time of remission (weeks).
Group 0 (6-MP): 6+, 6, 6, 6, 7, 9+, 10+, 10, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+ (m=21),
Group 1 (control): 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23 (n=21).

§1.2.2. **The right censorship model (RC model)** (Kaplan and Meier, 1958, JASA).

Assume:

Y — random censoring variable.

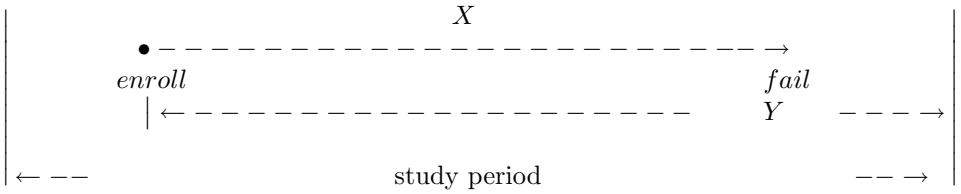
X and Y are independent ($X \perp Y$) *i.e.,* $P(X \le x, Y \le y) = P(X \le x)P(Y \le y)$.

Observable random vector $(L, R) = \begin{cases} (X, X) & \text{if } X \le Y \\ (Y, \infty) & \text{if } X > Y. \end{cases}$.

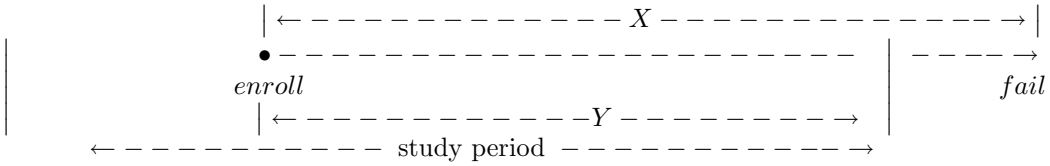
Equivalent observations: (Z, δ) $(= (\min(X, Y), \mathbf{1}_{(X \le Y)}))$..

Graphical illustration for Y and X

Exact observation



Right-censored observation



Note: If $Y = c$ w.p.1, it becomes type I censoring.

§1.2.3. Two incorrect approaches for RC data:

Method 1. Discard all RC observations;

Method 2. Treat RC observations as exact observations.

Question: What is wrong?

Reasoning:

For complete data, X_1, ..., X_n (i.i.d. from X),
if $\mu = E(X)$ exists, an estimator of μ is $\overline{X}$.
Its properties:

$E(\overline{X}) = \mu$ (unbiased);

$\overline{X} \to \mu$ w.p.1. (strongly consistent);

A nonparametric estimator of F and S are

$$\hat{F}(x) = \sum_{i=1}^{n} \mathbf{1}_{(X_i \le x)}/n \ (= \overline{W}) \text{ and } \hat{S}(x) = \sum_{i=1}^{n} \mathbf{1}_{(X_i > x)}/n \ (= \overline{Z})$$

—-Empirical distribution function (edf) and Empirical survival function, respectively.

Their properties:

$E(\hat{F}(x)) = F(x)$ (unbiased);

$\hat{F}(x) \to F(x)$ w.p.1. (strongly consistent);

Using method 1, we have

$m = \sum_{i=1}^{n} \mathbf{1}_{(X_i \le Y_i)}$ exact observations.

So the "sample mean" is

$$\tilde{\mu} = \frac{\sum_{i=1}^{n} X_i \mathbf{1}_{(X_i \le Y_i)}}{\sum_{i=1}^{n} \mathbf{1}_{(X_i \le Y_i)}}$$

Question:

$$E(\tilde{\mu}) = \mu? \qquad (*)$$

Or

$$E(\tilde{\mu}) = \frac{E(\sum_{i=1}^{n} X_i \mathbf{1}_{(X_i \le Y_i)})}{E(\sum_{i=1}^{n} \mathbf{1}_{(X_i \le Y_i)})} = \frac{E(X_i \mathbf{1}_{(X_i \le Y_i)})}{E(\mathbf{1}_{(X_i \le Y_i)})} \ ? \qquad (**)$$

Counterexample: Suppose $n = 2$, X_i, Y_i are i.i.d. from bin(1,1/2)+1. Then $E(\tilde{\mu})$ can be computed as in the table:

case	X_1	Y_1	X_2	Y_2	$\tilde{\mu}$	p	$p \times \tilde{\mu}$
1	1	1	1	1	1	1/16	1/16
2	2	1	2	1	∞	1/16	∞
.	.	.	.	.	.	.	.
2^4	.	.	.	.	.	.	.
sum							∞

Thus $\tilde{\mu}$ is biased, and both (*) and (**) fail.

Remark. Here we define $\frac{0}{0} = \infty$. If we define $\frac{0}{0} = 1$, we can also show $\tilde{\mu}$ is biased.

Question: $\tilde{\mu} \to \mu$?

$$\begin{aligned}
\lim_{n\to\infty} \tilde{\mu} =& \frac{\lim_{n\to\infty} \sum_{i=1}^{n} X_i \mathbf{1}_{(X_i \le Y_i)}/n}{\lim_{n\to\infty} \sum_{i=1}^{n} \mathbf{1}_{(X_i \le Y_i)}/n} \\
=& \frac{\lim_{n\to\infty} \overline{U}}{\lim_{n\to\infty} \overline{V}} \\
=& E(U)/E(V) \\
=& \frac{E(X \mathbf{1}_{(X \le Y)})}{E(\mathbf{1}_{(X_i \le Y_i)})} \ w.p.1.
\end{aligned}$$

Second counterexample. If $Y \equiv 1$ and X has a df $f(x) = \begin{cases} 1/3 & \text{if } x \in (0,1) \\ 2/3 & \text{if } x \in (1,2). \end{cases}$, then

(1) $\mu = 7/6$ but w.p.1 $\lim_{n\to\infty} \tilde{\mu} = \frac{1/6}{1/3} = 1/2$

(2) The "edf" is

$$\tilde{F}(x) = \frac{\sum_{i=1}^{n} \mathbf{1}_{(X_i \le Y_i, X_i \le x)}}{\sum_{i=1}^{n} \mathbf{1}_{(X_i \le Y_i)}}$$

W.p.1. its limit is

$$\frac{E(\mathbf{1}_{(X \le Y, X \le x)})}{P(X \le Y)} \ (= \frac{P(X \le 1, X \le x)}{P(X \le 1)}).$$

When $x = 1$, $\lim_{n\to\infty} \tilde{F}(1) = 1 \neq 1/3 = F(1)$.

Thus both estimators are inconsistent.

Homework : Verify (1) and (2).

Remark. The example illustrates the importance of a correct approach in dealing with censored data.

§1.2.4.**Homework**:

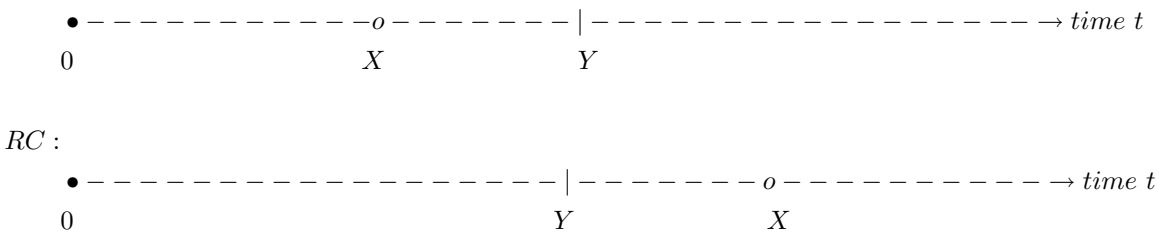
1. Using the second method, are the modified edf and sample mean consistent? Justify your statement.

§1.3. **Case 1 interval-censoring.**

Definition. If a data set only contains RC observations and LC observations, it is called *case 1 IC data* (C1 data) or *current status data*.

Example. Consider an animal sacrifice study in which a laboratory animal has to be dissected to check whether a tumor has developed. In this case, X is the onset of tumor and Y is the time of the dissection, and we only can infer at the time of dissection whether the tumor is present or has not yet developed.

Other examples are mentioned in Ayer *et al.* (1955), Keiding (1991) and Wang and Gardiner (1996).



The case 1 interval censorship model:

Assume:

Y is a random inspection time;

X and Y are independent;

The observable random vector is

$$(L, R) = \begin{cases} (-\infty, Y) & \text{if } X \leq Y \\ (Y, \infty) & \text{if } X > Y. \end{cases}$$

Equivalent forms:

Interval: $I = \begin{cases} (-\infty, Y] & \text{if LC} \\ (Y, \infty) & \text{if RC.} \end{cases}$

vector: (Y, δ), where $\delta = \mathbf{1}_{(X \leq Y)}$.

§1.4. **Double censoring**

Definition. If a data set contains RC, LC and exact observations, but not SIC observations, it is called a *doubly-censored data* (DC data).

Example. Leiderman *et al.* (1973) presented a study on the time needed for an infant to learn to perform a particular task (crawling) during the first year. The sampled infants were all born within 6 months of the start of the study. At the time of the start of the study, some children had already known how to perform the task; so their observed times were left-censored. Some children learned the task during the time-span of the study, and their ages were recorded. At the end of the study, some of the children had not yet learned the task, and hence their observed times were right-censored.

The double censorship model:

Assume

(Z, Y) is a random censoring vector;

$Z \leq Y$ w.p.1.;

X and (Z, Y) are independent;

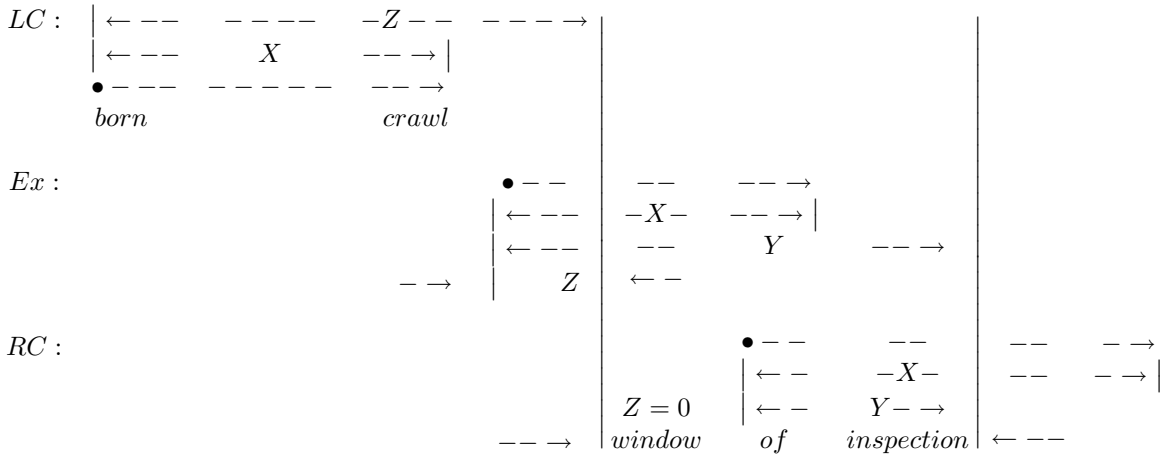
The observable random vector is

$$(L, R) = \begin{cases} (-\infty, Z) & \text{if } X \le Z \\ (X, X) & \text{if } Z < X \le Y \\ (Y, \infty) & \text{if } X > Y. \end{cases}$$

Equivalent forms:

Interval: $I = \begin{cases} (-\infty, Z] & \text{if } X \le Z \\ [X, X] & \text{if } Z < X \le Y \\ (Y, \infty) & \text{if } X > Y. \end{cases}$

Vector (U, δ), where $U = \max\{Z, \min\{X, Y\}\}$, $\delta = \begin{cases} 1 & \text{if exact} \\ 2 & \text{if RC} \\ 3 & \text{if LC.} \end{cases}$



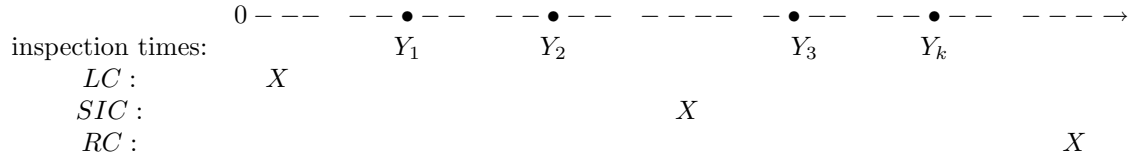
Remark. We could also set Z = time at birth - time to starting study, which may take negative values.

§1.5. **Case 2 interval censoring**.

Definition If a data set contains SIC observations and/or RC or LC observations, but not exact observations, it is called a *case 2 IC data* (C2 data).

Example. In medical research when each patient had several follow-ups and the event of interest was only known to take place either before the first follow-up, or between two consecutive follow-ups, or after the last one.

Graphical illustration



Examples of C2 data can be found in breast cancer research (Finkelstein and Wolfe, 1985) and AIDS studies (Becker and Belbye, 1991).

Possible models:

1.5.1. A simple model:

Groeneboom and Wellner (1992) proposed the following case 2 interval-censorship model.

Assume:

1. U and V are random inspection times such that $U < V$ w.p.1;
2. X and (U, V) are independent;
3. The observable random vector is

$$(L, R) = (-\infty, U)\mathbf{1}_{(X \le U)} + (U, V)\mathbf{1}_{(U < X \le V)} + (V, \infty)\mathbf{1}_{(X > V)}.$$

Remark. In a follow-up study, each patient has $\mathcal{N}$ visits, where $\mathcal{N} \ge 1$ is a random integer. The inspection times are $Y_1 < \cdots < Y_{\mathcal{N}}$. It is reasonable to assume that X and $(\mathcal{N}, \{Y_i : \ i \ge 1\})$ are independent. Then, on the event $\{\mathcal{N} = k\}$, define

$$(U, V) = (Y_1, Y_2)\mathbf{1}_{(X \le Y_1)} + (Y_{k-1}, Y_k)\mathbf{1}_{(X > Y_k)} + \sum_{i=2}^{k} (Y_{i-1}, Y_i)\mathbf{1}_{(Y_{i-1} < X \le Y_i)},$$

where $Y_0 = 0$. Then (U, V) and X are not independent. In other words, the case 2 model is simple, but its assumption is not realistic.

1.5.2. An alternative model: Wellner realized the drawback and proposed the Case k IC model, in which each patient has exactly k visits. Case 1 and Case 2 are special case of the case k models. However, it is not realistic except the case 1 model.

1.5.3. Another model: Petroni and Wolfe (1994) assume that inspection times Y_j's can only be taken at y_i's, where

$y_1 < y_2 < \cdots < y_k$,

which are predetermined (can be viewed as the reservation time), and $q(y_i) = P$(a patient keeps the appointment at time $y_i) \in (0, 1]$.

This results in inspection times $Y_1 < \cdots < Y_N$, together with their distribution, where N is again a random integer. Define $Y_0 = -\infty$ and $Y_{N+1} = \infty$.

Assume that X and $(Y_1, Y_2, ...)$ are independent.

Then $(L, R) = (Y_{j-1}, Y_j)$ if $X \in (Y_{j-1}, Y_j]$ for some j.

This model assumes that Y_j's are discrete, taking only k values.

The real data are actually continuous.

1.5.4. A realistic model (mixed case model, Schick and Yu (2000)):

Assume:

N is a random positive integer;

$Y_1 < Y_2 < \cdots < Y_k < \cdots$ are inspection times;

Conditional on N, X and $\{Y_1, ..., Y_N\}$) are independent

(or for simplicity, X, N, and $(Y_i, i \ge 1\}$ are independent);

The observable random vector is

$$(L, R) = \begin{cases} (-\infty, Y_1) & \text{if } X \le Y_1 \\ (Y_i, Y_{i+1}) & \text{if } Y_i < X \le Y_{i+1},\ i = 1, ..., N-1 \\ (Y_N, \infty) & \text{if } X > Y_N. \end{cases}$$

Remark .

1. When $N = k$ w.p.1, it is called a case k model, where $k = 1, 2, ...$
2. The mixed case model can be viewed as a mixture of various case k models.
3. When Y_i's are discrete, then the mixed case model becomes the model in §1.5.3.

Example of generating 100 observations under the mixed case model through simulation.

Main idea:

generate $X \sim f_X$;

generate $N \sim f_N$;

generate $Y_1, ..., Y_N \sim f_{\mathbf{Y}}$:

find j such that $Y_{j-1} < X \le Y_j$ to obtain (L, R).

repeat 100 times.

Example of generateing Y_i's:

(1) Generate N $Z_i \sim exp(\lambda)$, $Y_1 = Z_1, Y_2 = Y_1 + Z_2$, ..., $Y_N = Y_{N-1} + Z_N$.

(2) generate $Y_i \sim U(2i, 2i + 2)$, $i = 1$, ..., N.

Assume $N \sim$ Poisson(5), $X \sim exp(3)$ (mean=3), $Y_i \sim U(2i, 2i + 2)$.

```
L=rep(0,100) # initialize L
R=rep(0,100) # initialize R
for (i in 1 : 100) { # loop for 100 data
N=rpois(1,5) + 1 # generate 1 random variable from Poisson(5) +1
X=rexp(1,3) # generate 1 random variable from exp(3)
J = 1
if (N>1)
for (j in 2:N) J = c(J,j) # J is a vector of (1,2,...,N)
Y=runif(N,0,2)+2*J # generate N rv from U(2j,2j+2), j=1,...,N
if (X <= Y[1]) {
L[i] = 0
R[i] = Y[1]
}
else {
if (X > Y[N]) {
L[i] = Y[N]
R[i] = 1000
}
else {
j=2
while (X > Y[j] & j <= N) j=j+1
L[i] = Y[j-1]
R[i] = Y[j]
} } }
U=c(L,R)
dim(U)=c(100,2) # matrix of dimension 100 x 2
U # print the matrix
```

Question: What is wrong with the following scheme ?

```
X=rexp(100)
L=X-1
R=X+1
```

§1.6. **Mixed IC censoring**.

Definition. If a data set contains both exact and SIC observations, and/or RC or LC observations, it is called a *mixed IC data* (MIC data). It is also called partly IC data.

Example. (the National Longitudinal Survey of Youth 1979-98 (NLSY)). The 1979-98 cross-sectional and supplemental samples consist of 11,774 respondents, who were between the ages of 14 and 22 in 1979. Interviews were conducted yearly from 1979 through 1994; since then data were recorded bi-annually. One entry is the age at first marriage. There are SIC, exact, RC and LC observations in the data.

Possible models:

A simple model (MIC model (1), Yu *et al.* (2000)):

Assume:

1. (U, V) is an extended random censoring vector such that $U < V$ w.p.1;
2. X and (U, V) are independent;
3. The observable random variable is

$$(L, R) = \begin{cases} (X, X) & \text{if } X \notin (U, V] \\ (U, V) & \text{if } X \in (U, V]. \end{cases}$$

Remark . In reality, U and V is the $-\infty$ and the left censoring variable, respectively, if left censoring occurs; the two consecutive inspection times if SIC occurs; the right censoring variable and ∞ if right censoring occurs. Then assumption 2 in the model is false according to the interpretation. However, like the case 2 model for case 2 data, the model is very simple and easy to interpret for their variables.

A realistic model (MIC model (2), Yu *et al.* (2001)):

Assume:

N is a random integer;
T, $Y_1 < Y_2 < \cdots < Y_k < \cdots$ are inspection times, $Y_0 = -\infty$;
Conditional on N, X and $(T, Y_1, ..., Y_N)$ are independent;
$\mathrm{P}(N = 0) > 0$ and $\mathrm{P}(N > 1) > 0$;
The observable random vector is

$$(L, R) = \begin{cases} (X, X) & \text{if } X \le T \text{ and } N = 0 \\ (T, \infty) & \text{if } X > T \text{ and } N = 0 \\ (-\infty, Y_1) & \text{if } X \le Y_1 \text{ and } N \ge 1 \\ (Y_i, Y_{i+1}) & \text{if } Y_i < X \le Y_{i+1},\ i = 1, ..., N-1 \text{ and } N \ge 1 \\ (Y_N, \infty) & \text{if } X > Y_N \text{ and } N \ge 1. \end{cases}$$

The model can be viewed as a mixture of a RC model and a mixed case interval censorship model. That is,

$$F_{L,R}(l, r) = \sum_{k \ge 0} F_{L,R|N}(l, r|k) f_N(k).$$

Other models

Petroni and Wolfe (1994) and Huang (1999) construct two different models for the mixed IC data. Huang's model is basically a mixture of an uncensored model and a case k model, and thus is a special case of our MIC model (2) with $\mathrm{P}(N = i) = 0$ for $i \neq 0$ or k and with $T \equiv \infty$. The formulation of Petroni and Wolfe's model is basically the model described in §1.5.3 with the additional assumption that X is discrete as well. Thus it limits its extension to the continuous cases. Huang's model requires that X may be observed in the whole range of X, which is often not the case in reality.

§1.7. **Left censoring**.

Easy.

§1.8.

Table 1. Classification of IC data

observations :	*LC*	*SIC*	*RC*	*exact*
RC data			+	+
LC data	+			+
*C*1 *data*	+		+	
DC data	+		+	+
*C*2 *data*	+	+	+	
MIC data	+	+	+	+

§**1.9. Homework**: Generate a set of C2 data under the mixed case interval censorship model with a size of 100 and $\mathrm{P}(N = i) > 0$, $i = 1, ..., 8$. What will you do if you want to estimate the $F(x)$? (For example, one may consider the following treatment: Let $X_i^* = \begin{cases} \frac{L_i + R_i}{2} & \text{if SIC,} \\ L_i & \text{if RC,} \\ R_i & \text{if LC,} \end{cases}$ then pretend that X_i is observed and its value is X_i^*. Finally, estimate $F(t)$ by

$$\tilde{F}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(X_i^* \le t)}.)$$

What do you expect for the asymptotic property (consistency) of your or the above estimator ? Use the simulated data to compute the above estimate and compare to $F(t)$ (repeat 10 times) and the limiting value of the estimator (you can select only one specific t).

BIBLIOGRAPHY

* Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T. and Silverman, E. (1955). An empirical distribution function for sampling incomplete information. *Ann. Math. Statist.* 26, 641-647.

* Becker, N. and Melbye, M. (1991). Use of a log-linear model to compute the empirical survival curves from interval-censored data, with application to data on tests for HIV positivity. *Austral. J. Statistics*, 33, 125-133.
* Cox, D.R. and Oakes, D. (1984). Analysis of Survival Data. Chapman & Hall NY, 70-71.
* Finkelstein, D.M. and Wolfe, R.A. (1985). A semi-parametric model for regression analysis of interval-censored failure time data. *Biometrics*, 41, 933-945.
* Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily single-censored samples. *Biometrika*, 52, 203-23.
* Groeneboom, P. and Wellner, J. A. (1992). Information bounds and nonparametric maximum likelihood estimation. *Birkhäuser Verlag, Basel.*
* Huang, J. (1999). Asymptotic properties of nonparametric estimation based on partly interval-censored data. *Statistica Sinica*, 9, 501-520.
* Huang, J.L., Lee, C.S. and Yu, Q.Q. (2007). A generalized log-rank test for interval-censored failure time data via multiple imputation. *Statistics in Medicine*, (accepted).
* Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Stat. Assoc.*, 53, 457-481.
* Keiding. N. (1991) Age-specific incidence and prevalence: A statistical perspective (with discussion) *J. Roy. Statist. Soc. Ser. A*, 154, 371-412.
* Leiderman, P.H., Babu, D., Kagia, J., Kraemer, H.C. and Leiderman, G.F. (1973). African infant precocity and some social influences during the first year. *Nature*, 242, 247-249.
* Miller Jr., R. G. (1981). Survival analysis. *Wiley* NY. Odell, P.M., Anderson, K.M. and D'Agostino, R.B. (1992). Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics*, 48, 951-959.
* Petroni, G. R. and Wolfe, R. A. (1994). A two-sample test for stochastic ordering with interval-censored data. *Biometrics*, 50, 77-87.
* Schick, A. and Yu, Q.Q. (2000). Consistency of the GMLE with mixed case interval-censored data. *Scand. J. Statist.*, 27, 45-55.
* Sun, Jianguo (2006). The statistical analysis of interval-censored failure time data *Spring.*
* Wang, Z. and Gardiner, J. C. (1996). A class of estimators of the survival function from interval-censored data. *Ann. Statist.*, 24, 647-658.
* Yu, Q.Q., Li, L.X. and Wong, G.Y.C. (2000). On consistency of the self-consistent estimator of survival functions with interval censored data. *Scan. J. of Statist.* Vol 27 35-44.
* Yu, Q. Q., Wong, G. Y. C. and Li, L. X. (2001). Asymptotic properties of self-consistent estimators with mixed interval-censored data. *Ann. Inst. Statist. Math.*. 53 469-486.

Chapter 2. Distribution of failure time

§2.1. Hazard.

Suppose that X is r.v. with cdf F and density function f.

Definition. $S(t) = P(X > t)$ is often called the *survival function* of a r.v. X.

$S(t-) = P(X \geq t)$ is sometimes called the *survival function* of X, *e.g.* in Cox and Oakes (1984).

If X is continuous, $S(t-) = S(t)$.

In general, $S(t-) = \lim_{u \uparrow t} S(u)$.

The d.f. of X,

$$f(t) = \begin{cases} -S'(t) & \text{if } X \text{ is continuous} \\ S(t-) - S(t) & \text{if } X \text{ is discrete.} \end{cases}$$

$$S(t) = \begin{cases} \int_t^\infty f(x)dx & \text{if } X \text{ is continuous} \\ \sum_{x>t} f(x) & \text{if } X \text{ is discrete.} \end{cases}$$

.

Definition. $h(t) = f(t)/S(t-)$ is called the *hazard function* of X.

$H(t) = -\log S(t)$ is called the *integrated (or cumulative) hazard* of X.

Interpretation of h and H:

$$h(x) = \begin{cases} \lim_{\triangle \to 0} \frac{P(X \in [x, x+\triangle) | X \geq x)}{\triangle} & \text{if } X \text{ is continuous} \\ P(X = x | X \geq x) & \text{if } X \text{ is discrete.} \end{cases}$$

If X is continuous,

$$\begin{aligned} H'(t) =& (-\log S(t))' \\ =& f(t)/S(t) \\ =& h(t) \end{aligned}$$

That is

$$H(t) = \int_{-\infty}^{t} h(x)dx$$

integrated hazard.

If X is discrete,

$$\begin{aligned} S(t) &= \prod_{x_i \le t, x_i \in D_f} \frac{S(x_i)}{S(x_i-)} \\ &= \prod_{x_i \le t} (1 - h(x_i)) \end{aligned}$$

where $D_f = \{x : \ f(x) > 0\}$ and $\prod_{a \in \emptyset} a = 1$. Thus

$$H(t) = -\log \prod_{x_i \le t} (1 - h(x_i)) \qquad (\approx \sum_{x_i \le t} h(x_i) \text{ if } h(x_i) \approx 0 \text{ for all } i).$$

§2.2. **Some distributions**

For the purpose of parametric analysis later on, as well as for possible simulation studies, we need to be familiar with certain distributions which are related to the survival analysis. Thus we study some typical distributions here. Note that they all correspond to some positive random variables. We thus use the following conventions.

$f(x)$ ($F(x)$, $h(x)$, or $H(x)$), $x > 0$ implies that $f(x) = 0$ etc. for $x \le 0$;

$S(x)$, $x > 0$ implies that $S(x) = 1$ for $x \le 0$.

§2.2.1. **Exponential distribution**

$f(x) = \rho e^{-\rho x}$, $x > 0$, $\rho > 0$;

$S(x) = e^{-\rho x}$, $x > 0$; (**easy way to remember**)

$h(x) = \frac{f(x)}{S(x)} = \rho$, $x > 0$ (constant hazard);

$H(x) = \int_0^x h(t)dt = \rho x$, $x > 0$ (or $= -\ln S(x)$);

$E(X) = 1/\rho$.

Note that $E(X) = \int_0^\infty S(t)dt$.

Reason: If X is continuous, then

$$E(X) = \int_0^\infty x f(x)dx = \int_0^\infty \int_0^x f(x)dtdx = \int_0^\infty \int_t^\infty f(x)dxdt = \int_0^\infty S(t)dt.$$

In general,

$$E(X) = E(\int_0^X dt) = E(\int_0^\infty \mathbf{1}_{(t<X)}dt) = \int_0^\infty E(\mathbf{1}_{(t<X)})dt = \int_0^\infty P(X > t)dt = \int_0^\infty S(t)dt$$

§2.2.2. **Weibull distribution**

$f(x) = \kappa\rho(\rho x)^{\kappa-1}e^{-(\rho x)^\kappa}$, $x > 0$, $\rho, \kappa > 0$;

$S(x) = e^{-(\rho x)^\kappa}$, $x > 0$; (**easy way to remember**)

$h(x) = \kappa\rho(\rho x)^{\kappa-1}$, $x > 0$;

$H(x) = (\rho x)^\kappa$, $x > 0$;

§2.2.3 **Log normal distribution**

Let $Z \sim N(\log \rho^{-1}, \tau^2)$, $\rho, \tau > 0$, and $X = e^Z$, then X has a log normal distribution.

$$f(x) = f_Z(\log x)|\frac{dz}{dx}| = \frac{1}{\sqrt{2\pi}\tau x} \exp(\frac{-(\log x - \log \rho^{-1})^2}{2\tau^2}) = \frac{1}{\sqrt{2\pi}\tau x} \exp(\frac{-(\log(x\rho)^2}{2\tau^2}), \ x > 0$$

S, h and H cannot be simplified.

§2.2.4 **Log-logistic distribution**

If $X = e^T$ and T has logistic distribution, then X has the log logistic distribution.

A logistic distribution has a d.f

$$f_T(x) = \frac{\tau^{-1}\exp(\frac{x-\nu}{\tau})}{(1+\exp(\frac{x-\nu}{\tau}))^2}, \quad \tau > 0.$$

$$f(x) = f_T(\log x)\Big|\frac{dt}{dx}\Big| = \frac{\tau^{-1}\exp(\frac{\log x-\nu}{\tau})}{(1+\exp(\frac{\log x-\nu}{\tau})^2 x}, \ x \geq 0;$$

Let $\nu = -\log\rho$ and $\kappa = 1/\tau$ (reparameterization),

$$\begin{aligned} f(x) =& \frac{\kappa\exp(\log(\rho x)^\kappa)}{[1+\exp(\log(\rho x)^\kappa)]^2 x} \\ =& \frac{\kappa(\rho x)^\kappa}{[1+(\rho x)^\kappa]^2 x} \\ =& \frac{\kappa\rho^\kappa x^{\kappa-1}}{[1+(\rho x)^\kappa]^2}, \ x \geq 0. \end{aligned}$$

It yields S.

Another smart way to derive S.

A logistic survival function is

$S_T(x) = \frac{1}{1+\exp(\frac{x-\nu}{\tau})}$.

If $X = e^T$, then X has the log logistic distribution, with

$S(x) = S_X(x) = P\{X > x\} = P\{T > \ln x\} = \frac{1}{1+\exp(\frac{\ln x-\nu}{\tau})}$.

Let $\nu = -\log\rho$ and $\kappa = 1/\tau$ (reparameterization),

then

$$S(x) = \frac{1}{1+(\rho x)^\kappa}, \ x \geq 0.(\textbf{easy way to remember})$$

$$h(x) = \frac{\kappa\rho^\kappa x^{\kappa-1}}{1+(\rho x)^\kappa}, \ x \geq 0.$$

$$H(x) = \log(1+(\rho x)^\kappa), \ x \geq 0.$$

§2.2.5. Gompertz-Makeham distribution.

$$\begin{aligned} h(t) =& \rho_0 + \rho_1 e^{\rho_2 t}, \ t > 0, \ \rho_0, \rho_1 > 0; (\textbf{easy way to remember}) \\ H(t) =& \int_0^t h(x)dx = [\rho_0 t + \frac{\rho_1}{\rho_2}(e^{\rho_2 t} - 1)], \ t > 0; \\ S(t) =& e^{-H(t)} = e^{-[\rho_0 t + \frac{\rho_1}{\rho_2}(e^{\rho_2 t}-1)]}, \ t > 0; \\ f(t) =& -S'(t) = h(t)S(t), \ t > 0; \end{aligned}$$

Remark:

1. It is an exponential distribution if $\rho_1 = 0$.
2. It is called Gompertz distribution if $\rho_0 = 0$.

§2.2.6. Compound exponential distribution.

Suppose that each individual survival time is exponentially distributed

but that the rate varies randomly between individuals.

To represent this let P be a random variable

with density f_P

and the conditional density of X given $P = p$ is

$$f_{X|P}(x|p) = pe^{-px},\ x > 0.$$

Then $f(x) = \int pe^{-px} f_P(p)dp$. If P has gamma distribution, say,

$$f_P(p) = \frac{p^{\alpha-1}e^{-p/\beta}}{\Gamma(\alpha)\beta^\alpha},\ p,\ \alpha,\ \beta > 0.$$

Letting $\rho = \alpha\beta$, (the mean), then

$$\begin{aligned} f(x) =& \int_0^\infty pe^{-px}\frac{p^{\alpha-1}e^{-\alpha p/\rho}}{\Gamma(\alpha)(\rho/\alpha)^\alpha}dp = \int_0^\infty \frac{p^\alpha e^{-p(x+\alpha/\rho)}}{\Gamma(\alpha)(\rho/\alpha)^\alpha}dp \\ =& \frac{\Gamma(\alpha+1)}{(x+\alpha/\rho)^{\alpha+1}\Gamma(\alpha)(\rho/\alpha)^\alpha} \\ =& \alpha(\alpha/\rho)^\alpha(x+\alpha/\rho)^{-\alpha-1}. \end{aligned}$$

The latter df is called the **Pareto distribution**, with

$$\begin{aligned} S(x) =& (\alpha/\rho)^\alpha(x+\alpha/\rho)^{-\alpha}. \\ H(x) =& \alpha(-\ln(\alpha/\rho) + \ln(x+\alpha/\rho)). \\ h(x) =& \alpha/(x+\alpha/\rho)(\textbf{easy way to remember}). \end{aligned}$$

§2.2.7. Discrete distributions.

The common discrete random variables do not have concise forms for h and H. Thus so far, we only consider continuous r.v.s. Now consider a binomial distribution. $X \sim bin(2, p)$.

$f(x) = \binom{2}{x}p^x(1-p)^{2-x}$, $x \in \{0, 1, 2\}$.

$$S(x) = \begin{cases} 1 & \text{if } x < 0 \\ 1-(1-p)^2 & \text{if } x \in [0,1), \\ p^2 & \text{if } x \in [1,2), \\ 0 & \text{if } x \geq 2. \end{cases}$$

$S(0-) = 1$, $S(1-) = 1-(1-p)^2$, $S(2-) = p^2$,

$$h(x) = \begin{cases} (1-p)^2 & \text{if } x = 0, \\ \frac{2(1-p)p}{1-(1-p)^2} & \text{if } x = 1, \\ p^2/p^2 & \text{if } x = 2, \end{cases} = \begin{cases} (1-p)^2 & \text{if } x = 0, \\ \frac{2(1-p)}{2-p} & \text{if } x = 1, \\ 1 & \text{if } x = 2, \end{cases}$$

$$H(x) = \begin{cases} 0 & \text{if } x < 0 \\ -\ln(1-(1-p)^2) & \text{if } x \in [0,1), \\ -2\ln p & \text{if } x \in [1,2), \\ \infty & \text{if } x \geq 2. \end{cases}$$

§2.2.8. Proportional hazards (PH) model.

An advantage of defining hazard functions is the introduction of the PH model. Define

$$\tau = \tau_T = \sup\{t : F_T(t) < 1\}$$

for a random variable T.

Definition. Let $(\mathbf{X}, Y)$ be a random vector, where $\mathbf{X} \in \mathcal{R}^p$. Given $\mathbf{X} = \mathbf{x}$, the hazard of $Y|\mathbf{x}$ is

$$h(y|\mathbf{x}) = h_o(y)c(\mathbf{x}),\ \text{for } y < \tau, \tag{1}$$

where $c(\mathbf{x}) \geq 0$, $c(\cdot)$ takes at least two distinct values, and h_o is a hazard function. Then we say $(\mathbf{X}, Y)$ follows a proportional hazards (PH) model or Cox's regression model (Cox, 1972).

Remark 1. It is common to set $c(\mathbf{x}) = exp(\beta\mathbf{x})$ so that $c(\mathbf{x}) \geq 0$, where $\beta\mathbf{x} = \beta'\mathbf{x}$.

Remark 2. If the random variable is discrete, then the choice of $c(\mathbf{x}) = e^{\beta \mathbf{x}}$ may cause problem. For instance, if $h_o(y) = 0.5$, $c(\mathbf{x}) = 3$, $P(Y = y|Y \geq y, \mathbf{X} = \mathbf{x}) = h(y|\mathbf{x}) = h_o(y)c(\mathbf{x}) > 1$.

Two alternatives:

(1) either choose $c(\mathbf{x}) = exp(-e^{\beta \mathbf{x}})$ to ensure that $c(\mathbf{x})$ is between 0 and 1,

(2) or restrict the parameter space $\mathcal{B}$, the set that β belongs to.

For continuous random variable, we only need $c(\mathbf{x}) \geq 0$, as the hazard does not need to belong to $[0, 1]$, as long as it belongs to $[0, \infty]$.

Remark 3. In the original definition of the PH model (see Cox and Oakes (1984)), there is no a restriction $y < \tau$. We shall show in Example 1 that if h_o corresponds to a discrete random variable, statement (1) without the restriction does not define a hazard function.

Hereafter, take $p = 1$.

Example 1. (Counterexample of Eq. (1) without $y < \tau$). We shall consider an example of discrete random variables. If T is discrete and $P\{T = \tau) = f_T(\tau) > 0$, then

$$h_T(\tau) = f_T(\tau)/S_T(\tau-) = f_T(\tau)/f_T(\tau) = 1$$

which is always true. It follows that for such a discrete random variable statement (1) does not hold at τ, as

$$h(\tau|x) = 1 \neq 1 \times c(x) = h_o(\tau)c(x) \text{ as } c(x) \neq 1 \text{ for some } x.$$

It does not happen for continuous random variables, as one can eliminate τ from the support.

Example 2. If $S_o(t)$ is a survival function of a continuous random variable, then

$$S(t|x) = (S_o(t))^{e^{\beta x}} \text{ satisfies the PH model.} \tag{2}$$

Reason: $f(t|x) = -S'(t|\mathbf{x}) = -e^{\beta x}(S_o(t))^{e^{\beta x}-1}S_o'(t) = e^{\beta x}S(t|x)\frac{f_o(t)}{S_o(t)}$,

$h(t|x) = \frac{f(t|x)}{S(t|x)} = e^{\beta x}h_o(t)$.

Special cases:

a. Weibull: $S_o(t) = e^{-t^\gamma}$, $t > 0$. $S(t|x) = exp(-e^{\beta x}t^\gamma)$, $t > 0$.
$h(t|x) = e^{\beta x}\gamma t^{\gamma-1}$, $h_o(t) = \gamma t^{\gamma-1}$, $t > 0$.

b. Log-logistic: $S_o(t) = \frac{1}{1+t^\kappa}$. $S(t|x) = (\frac{1}{1+t^\kappa})^{e^{\beta x}}$, $t > 0$,
$h(t|x) = \exp(\beta x)\frac{\kappa t^{\kappa-1}}{1+t^\kappa}$. $h_o(t) = \frac{\kappa t^{\kappa-1}}{1+t^\kappa}$, $t > 0$.

c. Logistic: $S_o(t) = \frac{1}{1+e^t}$. $S(t|x) = (\frac{1}{1+e^t})^{e^{\beta x}}$,
$h(t|x) = \exp(\beta x)\frac{1}{1+e^{-t}}$, $h_o(t) = \frac{1}{1+e^{-t}}$.

Remark.

The distribution generated from Case a is still a Weibull distribution for each (β, x).

The distribution generated from Case b may not be a log-logistic distribution unless $\beta = 0$.

The distribution in Case c corresponds to a random variable X with negative values in its domain, though it is often in survival analysis that we only consider the non-negative domain. However, it still satifies the PH model.

Definition. The family of the survival functions $S(t|\mathbf{x})$ satisfies statement (2) for all possible β is called a Lehmann family, or we can say that the distribution is from a proportional integrated hazards (PIH) model, as

$$H(t|x) = -\ln S(t|x) = e^{\beta x}(-\ln S_o(t)) = e^{\beta x}H_o(t).$$

Statement (2) does not hold for discrete random variable. When Cox proposes the PH model, he distinguishes the model from the Lehmann family or the PIH model. However, later in the literature, the PIH model and the PH model are mistaken to be the same (see Sun (2006) p.18). Example 3 shows that they are different.

Example 3. Suppose $Y_0 \sim bin(2, p)$. Then its hazard function is

$$h_o(t) = \begin{cases} (1-p)^2 & \text{if } t = 0, \\ \frac{2(1-p)}{2-p} & \text{if } t = 1, \\ 1 & \text{if } t = 2. \end{cases}$$

Suppose $h(y|x) = h_o(y)c(x)$ for $y = 0$ or 1.
Then
$h(0|x) = (1-p)^2c(x)$ yields $f(0|x) = (1-p)^2c(x)$ as $S(0-|x) = 1$.
$h(1|x) = \frac{2(1-p)}{2-p}c(x)$ yields $f(1|x) = \frac{2(1-p)}{2-p}c(x)(1-(1-p)^2c(x))$.
$f(2|x) = 1 - f(0|x) - f(1|x)$.
Verify that $f(\cdot|\mathbf{x})$ defines a discrete density function.
It follows that

$$S(0|x) = 1 - (1-p)^2c(x).$$

However, if $p = 0.2$ and $c(x) = 0.3$,

$$(S_o(0))^{c(x)} = (1-(1-p)^2)^{c(x)} \approx 0.7360219 \neq 0.808 \approx 1-(1-p)^2c(x) = S(0|x).$$

It indicates that if h_o is a hazard function of a discrete random variable, and $h(t|x) = h_o(t)c(x)$, its cdf may not be of the form $S(t|x) = (S_o(t))^{c(x)}$ or Equation (2).

§2.2.9. Accelerated lifetime (AL) model.

Definition. If S_o is a survival function and given a p dimensional vector $\mathbf{X} = \mathbf{x}$, $Y|\mathbf{x}$ has a survival function $S(y|\mathbf{x}) = S_o(y/\exp(\beta\mathbf{x}))$, $\beta \in \mathcal{R}^p$, then we say $Y|\mathbf{x}$ is from an accelerated lifetime model.

If $Y|\mathbf{x}$ is from an AL model, then $\ln Y = \beta\mathbf{x} + \epsilon$, where e^ϵ has the survival function S_o.

Examples.

Weibull: $S(y|\mathbf{x}) = \exp(-y^\kappa e^{\alpha\mathbf{x}})$, $y > 0$. $\alpha = -\beta\kappa$.
Log-logistic: $S(y|\mathbf{x}) = \frac{1}{1+(y\exp(\alpha\mathbf{x}))^\kappa}$. $\alpha = -\beta$.

§2.3. **Homework**:

A.1. Let X_1, ..., X_n be independent continuous nonnegative random variables with hazard functions $h_1(\cdot)$, ..., $h_n(\cdot)$. Prove that $X = \min\{X_1, ..., X_n\}$ has hazard function $\sum_{j=1} h_j(t)$.

A.2. In a compound exponential distribution, let the rate be represented by the random variable P. Prove that $E(X) = E(1/P)$ and $\mathrm{Var}(X) = 2E(1/P^2) - [E(1/P)]^2$.

A.3. Derive the integrated hazard function $H(t|x)$ under the PH model in Example 3 of §2.2.8.

A.4. Show that the two examples in §2.2.9 are indeed from the AL model. Moreover, one is from the PH model and the other is not.

References.

* Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Scociety*, Series B 34, 187-220.

Chapter 3. Parametric Analysis

§3.1. Introduction

Assume that
the failure time X has a df $f = f_o(x;\theta)$,
where f_o is known, θ is a parameter, $\theta \in \Theta$ (parameter space),
the censoring vector has a df $g(\cdot)$, which does not depend on θ,
(L, R) is an extended observable random vector,
(L_i, R_i, X_i), $i = 1$, ..., n, are i.i.d. copies of (L, R, X).

Note that $L \le X \le R$ and $L_i \le X_i \le R_i$,

$$(L,R) \text{ is } \begin{cases} RC & \text{if } R = \infty \\ LC & \text{if } L = -\infty \\ SIC & \text{if } 0 < L < R < \infty \\ exact & \text{if } L = R. \end{cases}$$

Recall that if we observed $X_1 = x_1$, ..., $X_n = x_n$,
the likelihood of a complete data is

$$l(\phi) = \prod_{i=1}^{n} f_o(x_i;\phi).$$

Note: θ is a parameter and ϕ is a variable in Θ, as we do not know θ.

In particular,

$$l(\theta)\begin{cases} = \prod_{i=1}^n P(X_i = x_i) & \text{if } X \text{ is discrete} \\ \approx \prod_{i=1}^n \frac{P\{X_i \in [x_i, x_i+\triangle)\}}{\triangle} & \text{if } X \text{ is continuous,} \end{cases} \quad \text{where } \triangle \approx 0,$$

as $P\{X \in [x, x+\triangle)\} \approx f(x)\triangle$. The MLE of θ maximizes $l(\phi)$, $\phi \in \Theta$. We shall modify this idea for IC data.

§**3.2. Likelihood functions**.

If (L, R) is discrete, extending the definition for complete data, we could call $l(\phi)$ the likelihood function of the IC data $(L_i, R_i) = (l_i, r_i)$, $i = 1, ..., n$, where

$$l(\theta) = \prod_{i=1}^n P((L_i, R_i) = (l_i, r_i))$$

Example 1 (RC model). Denote f_Y, F_Y and S_Y the df, cdf and survival function of Y, respectively.

$$P((L,R) = (l,r)) = \begin{cases} P(X = l \le Y) & \text{if exact} \\ P(Y = l < X) & \text{if RC.} \end{cases} = \begin{cases} f(l)S_Y(l-) & \text{if exact} \\ S(l)f_Y(l) & \text{if RC.} \end{cases}$$

Let $\mathbf{1}_e$, $\mathbf{1}_r$, $\mathbf{1}_l$ and $\mathbf{1}_s$ be the indicator functions of the events that the observation is exact, RC, LC and SIC, respectively. Denote $\mathbf{1}_{e,i}$, $\mathbf{1}_{r,i}$, $\mathbf{1}_{l,i}$ and $\mathbf{1}_{s,i}$ in an obvious way.

$$P((L,R) = (l,r)) = [f(l)S_Y(l-)]^{\mathbf{1}_e}[S(l)f_Y(l)]^{\mathbf{1}_r}.$$

Note that we ignored the variable θ in the above expression for simplicity,

as $f = f_o(x;\theta)$, $S = S_o(x;\theta)$ and $F = F_o(x;\theta)$.

Thus, ignoring ϕ, we have

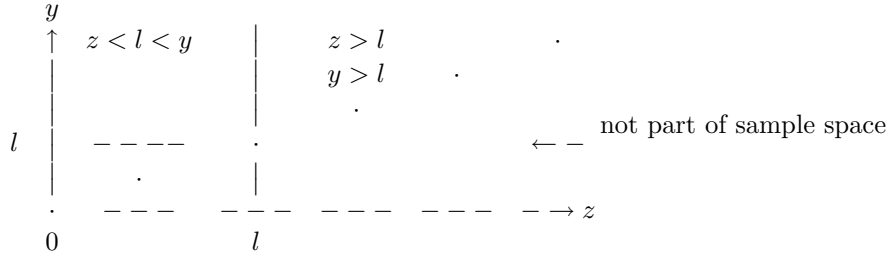
$$\begin{aligned} l(\phi) =& \prod_{i=1}^n \left\{[f(l_i)S_Y(l_i-)]^{\mathbf{1}_{e,i}}[S(l_i)f_Y(l_i)]^{\mathbf{1}_{r,i}}\right\} \\ =& \Big\{\prod_{i=1}^n \{[f(l_i)]^{\mathbf{1}_{e,i}}[S(l_i)]^{\mathbf{1}_{r,i}}\}\Big\}\Big\{\prod_{i=1}^n \{[(S_Y(l_i-)]^{\mathbf{1}_{e,i}}[f_Y(l_i)]^{\mathbf{1}_{r,i}}\}\Big\}. \end{aligned}$$

Example 2 (DC model).

$$P((L,R) = (l,r)) = \begin{cases} P(X = l, Z < l \le Y) & \text{if exact} \\ P(X > l = Y) & \text{if RC} \\ P(X \le r = Z) & \text{if LC} \end{cases} = \begin{cases} f(l)P(Z < l \le Y) & \text{if exact} \\ S(l)f_Y(l) & \text{if RC} \\ F(r)f_Z(r) & \text{if LC.} \end{cases}$$

Note $\{Z \ge l\} \subset \{Y \ge l\}$ as $Z \le Y$ w.p.1.

Thus $P(Z < l \le Y) = P(Y \ge l) - P(Z \ge l) = S_Y(l-) - S_Z(l-)$,



$$P((L,R)=(l,r))=[f(l)(S_Y(l-)-S_Z(l-))]^{\mathbf{1}_e}[S(l)f_Y(l)]^{\mathbf{1}_r}[F(r)f_Z(r)]^{\mathbf{1}_l}.$$

Thus $l(\phi)$

$$=\prod_{i=1}^{n}\{[f(l_i)(S_Y(l_i-)-S_Z(l_i-))]^{\mathbf{1}_{e,i}}[S(l_i)f_Y(l_i)]^{\mathbf{1}_{r,i}}[F(r_i)f_Z(r_i)]^{\mathbf{1}_{l,i}}\}$$
$$=\{\prod_{i=1}^{n}([f(l_i)]^{\mathbf{1}_{e,i}}[S(l_i)]^{\mathbf{1}_{r,i}}[F(r_i)]^{\mathbf{1}_{l,i}})\}\prod_{i=1}^{n}([S_Y(l_i-)-S_Z(l_i-)]^{\mathbf{1}_{e,i}}[f_Y(l_i)]^{\mathbf{1}_{r,i}}[f_Z(r_i)]^{\mathbf{1}_{l,i}}).$$

Example 3 (Case k model, $k \geq 2$).

$$P((L,R)=(l,r))=\begin{cases}P\{l<X\leq r,(Y_{j-1},Y_j)=(l,r),j\in\{2,...,k\}\} & \text{if SIC}\\ S(l)f_V(l) & \text{if RC}\\ F(r)f_U(r) & \text{if LC}\end{cases}$$
$$=\begin{cases}(S(l)-S(r))\sum_{j=2}^{k}f_{Y_{j-1},Y_j}(l,r) & \text{if it is SIC}\\ S(l)f_{Y_k}(l) & \text{if it is right censored}\\ F(r)f_{Y_1}(r) & \text{if it is left censored}\end{cases}$$

$$P((L,R)=(l,r))=[(S(l)-S(r))\sum_{j=2}^{k}f_{Y_{j-1},Y_j}(l,r)]^{\mathbf{1}_s}[S(l)f_{Y_k}(l)]^{\mathbf{1}_r}[F(r)f_{Y_1}(r)]^{\mathbf{1}_l}.$$

Thus $l(\phi)$

$$=\prod_{i=1}^{n}\{[(S(l_i)-S(r_i))\sum_{j=2}^{k}f_{Y_{j-1},Y_j}(l,r)]^{\mathbf{1}_{s,i}}[S(l_i)f_V(l_i)]^{\mathbf{1}_{r,i}}[F(r_i)f_U(r_i)]^{\mathbf{1}_{l,i}}\}$$
$$=\{\prod_{i=1}^{n}([S(l_i)-S(r_i)]^{\mathbf{1}_{s,i}}[S(l_i)]^{\mathbf{1}_{r,i}}[F(r_i)]^{\mathbf{1}_{l,i}})\}$$
$$\{\prod_{i=1}^{n}([\sum_{j=2}^{k}f_{Y_{j-1},Y_j}(l,r)]^{\mathbf{1}_{s,i}}[f_V(l_i)]^{\mathbf{1}_{r,i}}[f_U(r_i)]^{\mathbf{1}_{l,i}})\}.$$

Since the effect of the censoring vector can be factored out separately, and only the first factor in $l(\phi)$ depends on ϕ we can discard the second factor.

Definition The likelihood function of the IC data is defined to be

$$Ł(\phi)=\{\prod_{i=1}^{n}\{[f(l_i)]^{\mathbf{1}_{e,i}}[S(l_i)-S(r_i)]^{\mathbf{1}_{s,i}}[S(l_i)]^{\mathbf{1}_{r,i}}[F(r_i)]^{\mathbf{1}_{l,i}}\},$$

where $f(x)=f_o(x;\phi)$ in the parametric analysis.

Note that while we start our discussion under the assumption that (L,R) is discrete, the definition of Ł does not require that the (L,R) be discrete.

In an obvious way, we write

$$Ł(\phi)=\prod_{i:\ ex}f(l_i)\prod_{i:\ rc}S(l_i)\prod_{i:\ lc}F(r_i)\prod_{i:\ ic}(S(l_i)-S(r_i)),$$

where $f(x)=f_o(x;\phi)$ in the parametric analysis.

§**3.2.2. Homework** Mimic examples 1-3 for the mixed case IC model and MIC model (1).

§**3.3. MLE**.

Definition. Suppose $\phi=\hat{\theta}$ maximizes $Ł(\phi)$, $\phi\in\Theta$. Then $\hat{\theta}$ is called the maximum likelihood estimator (MLE) of θ.

Example. Suppose that X has an exponential distribution, i.e., $f(x) = \rho e^{-\rho x}$, $x > 0$. Find the MLE of ρ under the RC model.
Solution: Observe (Z_1, δ_1), ..., (Z_n, δ_n), where $Z_i = \min\{X_i, Y_i\}$ and $\delta_i = 1_{(R_i \neq \infty)} = \mathbf{1}_{(X_i \leq Y_i)}$ for all i

$$\begin{aligned}\log Ł(\rho) &= \sum_{i=1}^{n} \log\{[\rho e^{-\rho Z_i}]^{\delta_i}[e^{-\rho Z_i}]^{1-\delta_i}\} \quad (= \log(\prod_{i:\ ex} f(l_i) \prod_{i:\ rc} S(l_i))) \\ &= \sum_{i=1}^{n} \log\{\rho^{\delta_i} e^{-\rho Z_i}\} \\ &= \sum_{i=1}^{n} \delta_i \log \rho - \rho \sum_{i=1}^{n} Z_i.\end{aligned}$$

Taking derivative and letting it equal 0 yield

$$\hat{\rho} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} Z_i}.$$

Since $\frac{\partial^2 \mathcal{L}}{\partial \rho^2} = -\sum_{i=1}^{n} \delta_i \rho^{-2} < 0$, $\hat{\rho}$ is the MLE of ρ, unless $\sum_{i=1}^{n} \delta_i = 0$, as $\rho > 0$. In the latter case, $\rho = 0$ uniquely maximizes Ł, but it is not the MLE, as $0 \notin (0, \infty)$. We define $\hat{\rho} = 0$ so that $\hat{\rho}$ is well defined.
Remark. Formally, we shall write Ł(p) instead of Ł(ρ). However, in deriving the MLE, it does not matter.

Recall that if we have complete data, i.e., $\delta_i \equiv 1$, the MLE is $\frac{n}{\sum_{i=1}^{n} X_i}$. Let $h(x) = 1/x$, then $\frac{n}{\sum_{i=1}^{n} X_i} = h(\overline{X})$ is strongly consistent, as h is continuous. Furthermore, $\frac{n}{\sum_{i=1}^{n} X_i}$ is also asymptotically normally distributed as $\sigma_X < \infty$, h' is continuous and $h'(\rho) = -\rho^{-2} \neq 0$. Here we used the following asymptotic results:

1. By the the strong law of large number, if μ_X is finite and h is a continuous function,

$$h(\overline{X}) \text{ converges to } h(\mu_X) \text{ with probability one.}$$

2. By the central limit theorem and Corollary of the Slutsky's Theorem, if σ_X is finite, h' is continuous at μ_X and $h'(\mu_X) \neq 0$,

$$\frac{\sqrt{n}(h(\overline{X}) - h(\mu_X))}{\sigma_X |h'(\mu_X)|} \text{ converges in distribution to } N(0,1).$$

The above statements are valid even if X is an $m \times 1$ random vector. In particular,

$$\frac{\sqrt{n}(h(\overline{X}) - h(\mu_X))}{\sqrt{(\frac{\partial h(\mu_X)}{\partial \mu})^t \Sigma \frac{\partial h(\mu_X)}{\partial \mu}}} \text{ converges in distribution to } N(0,1).$$

where μ is an $m \times 1$ vector, Σ is the covariance matrix of X and μ^t is the transpose of μ, provided Σ is nonsingular, $\frac{\partial h}{\partial \mu}$ is continuous at $\mu = \mu_X$ and is not a zero vector at $\mu = \mu_X$.
§**3.3.2. Homework**:

1. Show that $\hat{\rho}$ in the example is consistent under the RC model.
(Hint: $\sum_{i=1}^{n} Z_i = \sum_{i=1}^{n} X_i \mathbf{1}_{(X_i \leq Y_i)} + \sum_{i=1}^{n} Y_i \mathbf{1}_{(X_i > Y_i)}$, or derive the df of Z.)
2. Show that $\hat{\rho}$ is asymptotically normally distributed under the RC model. That is

$$\sqrt{n}(\hat{\rho} - \rho) \text{ converges in distribution to a normal variate.}$$

3. Give a 99% approximate confidence interval for ρ when $n = 100$ and give a 99% confidence interval for ρ when $n = 1$ with $\rho > 1/\ln 50$.
4. Show that if X has a continuous cdf F with integrated hazard $H(\cdot)$ and $Z = \min\{X, c\}$, where c is a fixed constant, then $E(H(Z)) = F(c)$.

5. In the above example, if one uses the second incorrect approach to deal with the RC data, the MLE will be $\tilde{\rho} = \frac{n}{\sum_{i=1}^{n} Z_i}$. Derive the limit of the estimator and show that it is not a consistent estimator of ρ.
6. Suppose that a random sample of size n is generated from a type I RC model with $X \sim U(\theta, 4)$, $\theta \in (0, 4)$ and $P(Y = 3) = 1$. Find the MLE of θ. Derive the mean and variance of the MLE. Can we use the Cramer-Rao Lower Bound as the estimator of the variance of the MLE ?

§**3.4. Newton-Raphson method**.

If $Ł(\phi)$ is continuous and Θ is compact, then the MLE of ϕ exists. However, the MLE may not have closed form solution.

We use the Newton Raphson method to derive it. Denote

$$\mathcal{L} = \log Ł.$$

This is an iterative algorithm:
Step 1. Assign an initial value $\phi_{(1)}$ to ϕ.
Step $k+1$, $k \geq 1$. Given $\phi_{(k)}$, up-date ϕ by

$$\phi_{(k+1)} = \phi_{(k)} - \left(\frac{\partial^2 \mathcal{L}}{\partial \phi \partial \phi^t} \big|_{\phi=\phi_{(k)}} \right)^{-1} \frac{\partial \mathcal{L}}{\partial \phi} \big|_{\phi=\phi_{(k)}}. \tag{3.1}$$

Stop when $||\phi_{(k+1)} - \phi_{(k)}|| < \epsilon$, where ϵ is sufficiently small and $||\mathbf{z}|| = \max_i |z_i|$ or $= \sqrt{\sum_i z_i^2}$.
Reason: By the first order Taylor expansion, (**under certain condition on** $\mathcal{L}$),

$$\frac{\partial \mathcal{L}}{\partial \phi} - \frac{\partial \mathcal{L}}{\partial \phi} \big|_{\phi=\hat{\theta}} \approx \frac{\partial^2 \mathcal{L}}{\partial \phi \partial \phi^t} \big|_{\phi=\hat{\theta}} (\phi - \hat{\theta})$$

Since $\frac{\partial \mathcal{L}}{\partial \phi} \big|_{\phi=\hat{\theta}} = 0$ under certain regularity conditions (as $\hat{\theta}$ maximizes $\mathcal{L}$),

$$\frac{\partial \mathcal{L}}{\partial \phi} \approx \frac{\partial^2 \mathcal{L}}{\partial \phi \partial \phi^t} \big|_{\phi=\hat{\theta}} (\phi - \hat{\theta})$$

$$\hat{\theta} \approx \phi - \left(\frac{\partial^2 \mathcal{L}}{\partial \phi \partial \phi^t} \big|_{\phi=\hat{\theta}} \right)^{-1} \frac{\partial \mathcal{L}}{\partial \phi}.$$

Equation (3.1) is based on the last equation with $\left(\frac{\partial^2 \mathcal{L}}{\partial \phi \partial \phi^t} \big|_{\phi=\hat{\theta}} \right)^{-1}$ replaced by $\left(\frac{\partial^2 \mathcal{L}}{\partial \phi \partial \phi^t} \right)^{-1}$.

Drawbacks of the algorithm:

1. (Convergence). It may need some regularity assumptions to guarantee the convergence of the algorithm. For example, the algorithm may not converge even in the case that we generate complete data from Weibull(2,2).
2. (Uniqueness). It may converge to a local extreme point, unless $-\mathcal{L}$ is convex in ϕ.
3. (Feasibility). It is often difficult to obtain the inverse matrix $\left(\frac{\partial^2 \mathcal{L}}{\partial \phi \partial \phi^t} \right)^{-1}$ if the dimension of ϕ is large.

Example 1 (Log normal under the mixed case model)

Let $U_i = \log X_i$ be a $N(\mathbf{z}_i \beta, \sigma^2)$ variate, $i = 1, ..., n$, where $\mathbf{z}_i$ are $p \times 1$ covariate vector (real vectors), and are known, and β is a $p \times 1$ parameter vector. Under the mixed case model, X_i and thus U_i are not observed and we only observe (L_i, R_i) $(U_i \in (L_i, R_i])$ and $\mathbf{z}_i$, $i = 1, ..., n$. We shall estimate β.

We consider U_i rather than X_i because X_i is nonnegative, while a normal variate can be negative. The problem arises from linear regression for complete data. If $U_1, ..., U_n$ are observed, the MLE of β, which is also called the least squares estimator, is

$$\hat{\beta} = (\mathbf{Z}^t \mathbf{Z})^{-1} \mathbf{Z}^t \mathbf{U}, \text{ where } \mathbf{Z} = \begin{pmatrix} \mathbf{z}_1^t \\ \cdots \\ \mathbf{z}_n^t \end{pmatrix} \text{ and } \mathbf{U} = \begin{pmatrix} U_1 \\ \cdots \\ U_n \end{pmatrix}.$$

Here σ^2 does not matter, even though, it is unknown.

Under the mixed case IC model, $\theta = (\beta, \sigma)$ and the log likelihood function is

$$\begin{aligned}\mathcal{L}(\theta) =& \log \prod_{i=1}^{n} \int_{l_i}^{r_i} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\beta^t \mathbf{Z}_i)^2}{2\sigma^2}} dx \text{ (if treating as independent r.v.'s)} \\ =& \log \prod_{i=1}^{n} (\Phi(\frac{r_i - \beta' \mathbf{Z}_i}{\sigma}) - \Phi(\frac{l_i - \beta' \mathbf{Z}_i}{\sigma})) \qquad \text{(if treating as i.i.d.),}\end{aligned}$$

where Φ is the cdf of the $N(0,1)$, the standard normal random variable. Since

$(\int_{-\infty}^{b(x)} f(t)dt)'_x = f(b(x))b'(x)$, taking partial derivatives yields

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial \sigma} =& \sum_{i=1}^{n} \frac{\frac{\partial}{\partial\sigma}\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma}) - \frac{\partial}{\partial\sigma}\Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma})}{\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma}) - \Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma})} \\ =& -\frac{\sigma^{-2}}{\sqrt{2\pi}} \sum_{i=1}^{n} \frac{(r_i-\beta^t\mathbf{Z}_i)e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}} - (l_i-\beta^t\mathbf{Z}_i)e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma}) - \Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma})} \\ =& -\frac{\sigma^{-1}}{\sqrt{2\pi}} \sum_{i=1}^{n} \frac{\frac{(r_i-\beta^t\mathbf{Z}_i)}{\sigma}e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}} - \frac{(l_i-\beta^t\mathbf{Z}_i)}{\sigma}e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma}) - \Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma})}\end{aligned}$$

If $l_i = -\infty$, then $\Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma}) = 0$ and thus $\Phi'_\sigma(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma}) = 0$. In such a case, we can nevertheless write $\Phi'_\sigma(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma}) = -\frac{(l_i-\beta^t\mathbf{Z}_i)}{\sigma^2}e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}$, as it still equals 0 for $l_i = -\infty$. A similar argument can be applied to the case $r_i = \infty$. In this way, $\frac{\partial \mathcal{L}(\theta)}{\partial \sigma}$ can have a simpler form.

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial \beta} =& -\sum_{i=1}^{n} \left(\frac{e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}} - e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma}) - \Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma})} \frac{\sigma^{-1}}{\sqrt{2\pi}} \mathbf{Z}_i \right) \\ =& -\frac{\sigma^{-1}}{\sqrt{2\pi}} \sum_{i=1}^{n} \left(\frac{e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}} - e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma}) - \Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma})} \mathbf{Z}_i \right).\end{aligned}$$

Here

$$\frac{\partial \mathcal{L}}{\partial \beta} = \begin{pmatrix} \frac{\partial \mathcal{L}}{\beta_1} \\ \cdots \\ \frac{\partial \mathcal{L}}{\beta_p} \end{pmatrix} \text{ and } \frac{\partial \mathcal{L}}{\partial \beta^t} = (\frac{\partial \mathcal{L}}{\beta_1} \ \cdots \ \frac{\partial \mathcal{L}}{\beta_p}).$$

There is no closed form solution to the equation $\frac{\partial \mathcal{L}}{\partial \beta} = \mathbf{0}$ and $\frac{\partial \mathcal{L}}{\partial \sigma} = 0$. So we shall use the Newton-Raphson method to obtain the MLE of $\begin{pmatrix} \beta \\ \sigma \end{pmatrix}$:

$$\begin{pmatrix} \beta_{(k+1)} \\ \sigma_{(k+1)} \end{pmatrix} = \begin{pmatrix} \beta_{(k)} \\ \sigma_{(k)} \end{pmatrix} - \begin{pmatrix} \frac{\partial^2 \mathcal{L}}{\partial\beta\partial\beta^t} & \frac{\partial^2 \mathcal{L}}{\partial\sigma\partial\beta} \\ \frac{\partial^2 \mathcal{L}}{\partial\beta^t\partial\sigma} & \frac{\partial^2 \mathcal{L}}{\partial\sigma^2} \end{pmatrix}^{-1} \Big|_{\beta=\beta_{(k)},\sigma=\sigma_{(k)}} \begin{pmatrix} \frac{\partial \mathcal{L}}{\partial\beta} \\ \frac{\partial \mathcal{L}}{\partial\sigma} \end{pmatrix} \Big|_{\beta=\beta_{(k)},\sigma=\sigma_{(k)}}.$$

Here

$$\begin{aligned}\frac{\partial^2 \mathcal{L}(\theta)}{\partial\beta\partial\beta^t} =& -\sum_{i=1}^{n} \left[\frac{\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma^2}e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}} - \frac{l_i-\beta^t\mathbf{Z}_i}{\sigma^2}e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{(\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma}) - \Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma}))\sqrt{2\pi\sigma^2}} + \left(\frac{e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}} - e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma}) - \Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma})} \right)^2 \frac{1}{2\pi\sigma^2} \right] \mathbf{Z}_i\mathbf{Z}_i^t \\ =& -\sum_{i=1}^{n} \left[\frac{\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma}e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}} - \frac{l_i-\beta^t\mathbf{Z}_i}{\sigma}e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{(\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma}) - \Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma}))\sqrt{2\pi}} + \left(\frac{e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}} - e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma}) - \Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma})} \right)^2 \frac{1}{2\pi} \right] \frac{\mathbf{Z}_i\mathbf{Z}_i^t}{\sigma^2}.\end{aligned}$$

$$
\begin{aligned}
\frac{\partial^2 \mathcal{L}(\theta)}{\partial\sigma\partial\beta} =& -\frac{\sigma^{-2}}{\sqrt{2\pi}}\sum_{i=1}^{n}\mathbf{z}_i\Big\{\frac{\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{\sigma^2}e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{\sigma^2}e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma})-\Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma})}-\frac{e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}-e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma})-\Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma})}\\
&+\frac{(r_i-\beta^t\mathbf{Z}_i)e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}-(l_i-\beta^t\mathbf{Z}_i)e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma})-\Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma})}\cdot\frac{e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}-e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma})-\Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma})}\cdot\frac{\sigma^{-1}}{\sqrt{2\pi}}\Big\}\\
=& -\frac{\sigma^{-2}}{\sqrt{2\pi}}\sum_{i=1}^{n}\mathbf{z}_i\Big\{\frac{\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{\sigma^2}e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{\sigma^2}e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma})-\Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma})}-\frac{e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}-e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma})-\Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma})}\\
&+\frac{\frac{(r_i-\beta^t\mathbf{Z}_i)}{\sigma}e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}-\frac{(l_i-\beta^t\mathbf{Z}_i)}{\sigma}e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{(\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma})-\Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma}))^2}\cdot\frac{e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}-e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{\sqrt{2\pi}}\Big\}.\\
\frac{\partial^2 \mathcal{L}(\theta)}{\partial\sigma^2} =& 2\frac{\sigma^{-3}}{\sqrt{2\pi}}\sum_{i=1}^{n}\frac{(r_i-\beta^t\mathbf{Z}_i)e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}-(l_i-\beta^t\mathbf{Z}_i)e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma})-\Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma})}\\
&-\frac{\sigma^{-2}}{\sqrt{2\pi}}\sum_{i=1}^{n}\frac{\frac{(r_i-\beta^t\mathbf{Z}_i)^3}{\sigma^3}e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}-\frac{(l_i-\beta^t\mathbf{Z}_i)^3}{\sigma^3}e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma})-\Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma})}\\
&-(\frac{\sigma^{-2}}{\sqrt{2\pi}})^2\sum_{i=1}^{n}\left(\frac{(r_i-\beta^t\mathbf{Z}_i)e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}-(l_i-\beta^t\mathbf{Z}_i)e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{(\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma})-\Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma})}\right)^2\\
=& 2\frac{\sigma^{-2}}{\sqrt{2\pi}}\sum_{i=1}^{n}\frac{\frac{(r_i-\beta^t\mathbf{Z}_i)}{\sigma}e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}-\frac{(l_i-\beta^t\mathbf{Z}_i)}{\sigma}e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma})-\Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma})}\\
&-\frac{\sigma^{-2}}{\sqrt{2\pi}}\sum_{i=1}^{n}\frac{\frac{(r_i-\beta^t\mathbf{Z}_i)^3}{\sigma^3}e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}-\frac{(l_i-\beta^t\mathbf{Z}_i)^3}{\sigma^3}e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma})-\Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma})}\\
&-\frac{\sigma^{-2}}{2\pi}\sum_{i=1}^{n}\left(\frac{\frac{(r_i-\beta^t\mathbf{Z}_i)}{\sigma}e^{-\frac{(r_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}-\frac{(l_i-\beta^t\mathbf{Z}_i)}{\sigma}e^{-\frac{(l_i-\beta^t\mathbf{Z}_i)^2}{2\sigma^2}}}{(\Phi(\frac{r_i-\beta^t\mathbf{Z}_i}{\sigma})-\Phi(\frac{l_i-\beta^t\mathbf{Z}_i}{\sigma})}\right)^2.
\end{aligned}
$$

Some useful asymptotic results.

Under certain regularity conditions, the following asymptotic properties are valid and are used for testing statistical hypotheses and constructing confidence intervals or confidence regions for $\theta\in\Theta\subset R^m$.

A. $\left(-\frac{\partial^2\mathcal{L}}{\partial\phi\partial\phi^t}\big|_{\phi=\hat{\theta}}\right)^{1/2}(\hat{\theta}-\theta)$ is approximately $N(0,I_m)$ distributed if n is large, where I_m is a $m\times m$ identity matrix.

B. $(\hat{\theta}-\theta)^t\left(-\frac{\partial^2\mathcal{L}}{\partial\phi\partial\phi^t}\big|_{\phi=\hat{\theta}}\right)(\hat{\theta}-\theta)$ is approximately $\chi^2(m)$ distributed if n is large,

C. For testing H_0: $\theta\in\Theta_0$ v.s. H_1: $\theta\notin\Theta_0$, an asymptotic test is the likelihood ratio test and another test is a score test. Here
the likelihood ratio test is $\mathbf{1}_{(-2\log\frac{\mathrm{L}(\tilde{\theta}_0)}{\mathrm{L}(\hat{\theta})}>\chi^2_{\alpha,d})}$ where $\hat{\theta}$ and $\tilde{\theta}_0$ are the MLEs of θ in the space Θ and Θ_0, respectively, $d=||\Theta||-||\Theta_0||$ and $||\Theta_0||$ is the dimension of Θ_0;
The score test is based on $\frac{\partial\mathcal{L}}{\partial\phi}\big|_{\phi=\tilde{\theta}_0}$. In particular, if $\Theta_o=\{\theta_o\}$, then the test is $\mathbf{1}_{(T>\chi^2_{\alpha,d})}$, where $T=\left(\frac{\partial\mathcal{L}}{\partial\phi}^t(\frac{-\partial^2\mathcal{L}}{\partial\phi\partial\phi^t})^{-1}\frac{\partial\mathcal{L}}{\partial\phi}\right)\big|_{\phi=\theta_o}$.

§**3.4.2. Homework**:

1. Generate a RC data set of size 100 from a Weibull distribution with $\kappa=2$ and $\rho=1$. $S(t)=exp(-(\rho t)^\kappa)$.
2. Estimate the MLE of the parameter using the Newton-Raphson method (you could use the command in Splus:
   y=rweibull(100,1/2,1) #kappa = 2 scale = 1
   x=(1:100)
   c=runif(100,0,2)

d=as.numeric($y <= c$)
m=d*y+(1-d)*c
yy=survReg(Surv(m,d)$\sim$ x-1, dist="weibull")
but you need at least to derive the iteration formula.)

3. Estimate the covariance matrix of the MLE.
4. Derive a 95% confidence interval for ρ.
5. Test H_0: $\kappa = 1$ v.s. H_1: $\kappa \neq 1$ using the data you generated with size $\alpha = 0.1$. This is to test whether the distribution is actually from an Exp(ρ) rather from the Weibull distribution.
6. Is the result in problem 5 as what you expected ? Why you say so?
7. If you have a sample of size 4, do you still expect to see what you expect in Problem 6 ?

§3.5. Consistency and Asymptotic Normality:
Two issues:

1. Is the MLE consistent ?
2. Is the MLE asymptotically normally distributed ?

They need to be verified for each problem. It is easy to verify if the MLE has a closed form expression, e.g., in the example considered in §3.3. In general, it is not so trivial. We shall illustrate the usual approach through the problem of estimating the parameter of an exponential distribution under the C2 model. In particular, we assume:

1) $X \sim Exp(\rho_o)$,
2) (U, V) has a joint c.d.f. $G(u, v)$, where $0 \leq U < V$ w.p.1.,
3) X and (U, V) are independent;
4) $(L, R) = \begin{cases} (U, V) & \text{if } X \in (U, V], \\ (V, +\infty) & \text{if } X > V, \\ (-\infty, U) & \text{if } X \leq U, \end{cases}$

Let $(X_i, U_i, V_i, L_i, R_i)$, $i = 1, ..., n$, be i.i.d. copies of (X, U, V, L, R).
The log likelihood function

$$\mathcal{L}(\rho) = \sum_{i:\ lc} \log F_o(R_i; \rho) + \sum_{i:\ ic} \log[S_o(L_i; \rho) - S_o(R_i; \rho)] + \sum_{i:\ rc} \log S_o(L_i; \rho). \tag{1.1}$$

Eq. (1.1) yields that

$$\mathcal{L}(\rho) = \sum_{i,lc} \log(1 - e^{-\rho R_i}) + \sum_{i,ic} \log(e^{-\rho L_i} - e^{-\rho R_i}) - \rho \sum_{i,rc} L_i. \tag{1.2}$$

To find the MLE, we look at the normal equation $\frac{\partial \mathcal{L}}{\partial \rho} = 0$:

$$\sum_{i:lc} \frac{R_i \exp(-\rho R_i)}{1 - \exp(-\rho R_i)} - \sum_{i:lc} \frac{L_i \exp(-\rho L_i)}{\exp(-\rho L_i)} - \sum_{i:lc} \frac{L_i \exp(-\rho L_i) - R_i \exp(-\rho R_i)}{\exp(-\rho L_i) - \exp(-\rho R_i)} = 0.$$

It is easy to verify that there is no closed form expression for the MLE of ρ_o. Thus we need to use the Newton-Raphson method to derive it.

§3.5.1. Existence of the MLE.

In some extreme case, the MLE may not exist. However, one can always modify the MLE so that it is well defined.

Theorem 1. Suppose that $X \sim \text{Exp}(\rho_o)$. Under the C2 model, the MLE $\hat{\rho}$ exists and $\hat{\rho} \in (0, +\infty)$ unless $\sum_{i,rc} 1 = \sum_i \mathbf{1}_{(R_i=\infty)} = n$ or $\sum_{i,lc} 1 = \sum_i \mathbf{1}_{(L_i=-\infty)} = n$.

Proof. First assume that $\sum_{rc} 1 < n$ and $\sum_{lc} 1 < n$. Verify from (1.2) that

(i) $\lim_{\rho \to 0^+} \mathcal{L}(\rho) = -\infty$;
(ii) $\lim_{\rho \to +\infty} \mathcal{L}(\rho) = -\infty$;
(iii) $\mathcal{L}(\rho)$ is continuous in ρ.

It follows that the MLE $\hat{\rho}$ exists and $\hat{\rho} \in (0, +\infty)$. □

Remark. If $\sum_{rc} 1 = n$ then $\mathcal{L}(\rho)$ $(= -\rho \sum_{rc} V_i)$ is maximized uniquely by $\rho = 0 \notin (0, \infty)$. Thus the MLE does not exist and we define $\hat{\rho} = 0$. If $\sum_{lc} 1 = n$, then $\mathcal{L}(\rho)$ $(= \sum_{lc} \log(1 - e^{-\rho U_i}))$ is maximized uniquely by $\rho = \infty \notin (0, \infty)$. Thus the MLE does not exist and we define $\hat{\rho} = +\infty$. In this way, $\hat{\rho}$ is properly defined in the whole sample space and thus we can study its properties.

§**3.5.2. Consistency of the MLE.**

Hereafter, abusing notations, we write $S = S_o$ etc.. The normalized log likelihood function

$$\begin{aligned}
l(\rho) =& \frac{1}{n}\{\sum_{i:\ ic} \log[S(L_i;\rho) - S(R_i;\rho)] + \sum_{i:\ rc} \log S(L_i;\rho) + \sum_{i:\ lc} \log F(R_i;\rho)\} \\
=& \frac{1}{n}\{\sum_{i:\ ic} \log[S(U_i;\rho) - S(V_i;\rho)] + \sum_{i:\ rc} \log S(V_i;\rho) + \sum_{i:\ lc} \log F(U_i;\rho)\}. \qquad (2.1) \\
=& (\sum_{lc} \log(1 - e^{-\rho U_i}) + \sum_{ic} \log e^{-\rho U_i} + \sum_{ic} \log(1 - e^{-\rho(V_i - U_i)}) - \rho \sum_{rc} V_i)/n \\
=& (\sum_{i=1}^{n} \log(1 - e^{-\rho U_i}) \mathbf{1}_{(X_i \le U_i)} + \sum_{i=1}^{n} \log(1 - e^{-\rho(V_i - U_i)}) \mathbf{1}_{(X_i \in (U_i, V_i])} \\
& - \rho \sum_{i=1}^{n} U_i \mathbf{1}_{(X_i \in (U_i, V_i])} - \rho \sum_{i=1}^{n} V_i \mathbf{1}_{(X_i > V_i)})/n
\end{aligned}$$

Theorem 2. *Suppose that $X \sim Exp(\rho_o)$. Under the C2 model, the MLE $\hat{\rho}$ of ρ is strongly consistent.*

Proof of Theorem 2. Let

$$\mu(\rho) = E(l(\rho)).$$

We shall show

$$\mu(\rho) \le \mu(\rho_o) \text{ with equality iff } \rho = \rho_o. \qquad (2.2)$$

$$\text{Let } \rho^* \text{ be a limiting point of } \hat{\rho}, \text{ then } \mu(\rho^*) \ge \mu(\rho_o) \text{ with probability 1.} \qquad (2.3)$$

(2.2) and (2.3) imply that $\mu(\rho^*) = \mu(\rho_o)$ and thus $\rho^* = \rho_o$ by (2.2). Moreover, since ρ^* is an arbitrary limiting point of $\hat{\rho}$, it implies that $\hat{\rho} \to \rho_o$ w.p.1. In other words, (2.2) and (2.3) imply that $\hat{\rho}$ is strongly consistent.

Proof of (2.2): Now

$$\begin{aligned}
&\mu(\rho)(= E(l(\mathbf{L}, \mathbf{R}, \rho))) && \mathbf{L} = (L_1, ..., L_n), \quad \mathbf{R} = (R_1, ..., R_n) \\
=& E(E(l(\mathbf{L}, \mathbf{R}, \rho)|\mathbf{U}, \mathbf{V})) \\
=& E(E[\mathbf{1}_{(X \le U)} \log F(U;\rho) + \mathbf{1}_{(X > V)} \log S(V;\rho) + \mathbf{1}_{(U < X \le V)} \log(S(U;\rho) - S(V;\rho))|U, V]). \\
=& E[F(U;\rho_o) \log F(U;\rho) + S(V;\rho_o) \log S(V;\rho) + (S(U;\rho_o) - S(V;\rho_o)) \log(S(U;\rho) - S(V;\rho))].
\end{aligned}$$

Define $\log 0 = -\infty$, $\log \infty = \infty$ and $0 \log \infty = 0$. Then $\mu(\rho)$ is uniquely maximized by $\rho = \rho_o$, due to the following three reasons:

(a) For each (u, v),
$F(u;\rho_o) \log F(u;\rho) + (S(u;\rho_o) - S(v;\rho_o)) \log(S(u;\rho) - S(v;\rho)) + S(v;\rho_o) \log S(v;\rho)$
as a function of $F(\cdot, \rho)$ is uniquely maximized by
$F(u, \rho) = F(u, \rho_o)$,
$S(u, \rho) - S(v, \rho) = S(u, \rho_o) - S(v, \rho_o)$
and $S(v, \rho) = S(v, \rho_o)$,

(b) the last equation in (a) implies $\rho = \rho_o$ as $S(x, \rho) = e^{-\rho x}$.

(c) $|\mu(\rho_o)| < +\infty$.

Remark. (a) is equivalent to say that $g(\cdot, \cdot)$, defined by

$$g(q_1, q_2) = p_1 \log q_1 + p_2 \log q_2 + (1 - p_1 - p_2) \log(1 - q_1 - q_2),\ q_i, p_i \ge 0,\ p_1 + p_2 \le 1,\ q_1 + q_2 \le 1,$$

is uniquely maximized by $q_1 = p_1$ and $q_2 = p_2$. (Hint: take partial derivatives and set to zero, etc., or by the Shannon-Kolmogorov inequality).

To prove (c) note that
each summand in $g(p_1, p_2)$ is of the form $x\ln x$,
$0 \geq x \log x \geq -1/e$, for $x \in [0, 1]$,
and thus $g(p_1, p_2) \geq -3/e$,
Thus (c) follows.
(a), (b) and (c) imply that

$$\mu(\rho_o) > \mu(\rho) \text{ if } \rho \neq \rho_o \text{ that is, (2.2) holds.}$$

Prove (2.3). By definition of the MLE, $l(\hat{\rho}) \geq l(\rho_o)$. Thus

$$\liminf_{n\to\infty} l(\hat{\rho}) \geq \lim_{n\to\infty} l(\rho_o) = \mu(\rho_o) \tag{2.4}$$

a.s. (by SLLN).

$$\begin{aligned} l(\rho) = &\sum_{i=1}^{n} \frac{\log(1 - e^{-\rho U_i})\mathbf{1}_{(X_i \leq U_i)}}{n} + \sum_{i=1}^{n} \frac{\log(1 - e^{-\rho(V_i - U_i)})\mathbf{1}_{(X_i \in (U_i, V_i])}}{n} \\ &- \rho \sum_{i=1}^{n} \frac{U_i \mathbf{1}_{(X_i \in (U_i, V_i])}}{n} - \rho \sum_{i=1}^{n} \frac{V_i \mathbf{1}_{(X_i > V_i)}}{n}. \end{aligned} \tag{2.5}$$

For each ρ, the four summations in (2.5) all converges almost surely to their means, respectively. However, it is not clear that it is true uniformly for all $\rho > 0$ (with probability one). Thus, we let K be the set of all positive rational numbers and ρ_o, and let Ω^* be the event

* such that (2.4) holds, and for all $\rho \in K$, these four summations in (2.5) converges a.s. to their means, respectively.

Then $\mathrm{P}(\Omega^*) = 1$ as K is countable.

To emphasize that the MLE $\hat{\rho}$ is a function of n, we write $\hat{\rho} = \hat{\rho}_n$. For each ω in Ω^*, let ρ^* be a limiting point of $\hat{\rho} = \hat{\rho}_n(\omega)$ in the sense that $\hat{\rho}_{n_j}(\omega) \to \rho^*$ for a subsequence of $\{\hat{\rho}_n\}$, where ρ^* may be $+\infty$ or 0. We shall show next that

$$\mu(\rho_o) \leq \mu(\rho^*), \tag{2.6}$$

which implies $\mu(\rho_o) = \mu(\rho^*)$ by Eq. (2.2) and thus $\rho^* = \rho_o$ by Eq. (2.2) . Since ω and $\rho*$ are arbitrary, it implies that $\hat{\rho} \to \rho_o$ a.s..

Assume $\omega \in \Omega^*$. If $\rho^* = +\infty$, then Eq. (2.5) implies that

$$l(\hat{\rho}_{n_j}(\omega)) \leq -\hat{\rho}_{n_j}(\omega) \sum_{i=1}^{n} V_i \frac{\mathbf{1}_{(X_i > V_i)}}{n} \to -\infty \tag{2.7}$$

Then Eq. (2.7) and inequality (2.4) imply that

$$-\infty = \lim_{n_j \to \infty} l(\hat{\rho}(\omega)) \geq \mu(\rho_o).$$

It reaches a contradiction as $\mu(\rho_o)$ is finite by (2.2). Thus $\rho^* = +\infty$ is impossible.

If $\rho^* = 0$, the first two sums in Eq. (2.5) tend $-\infty$ for large enough n as $\log 0 = -\infty$ by our convention. Thus it can be shown $\rho^* = 0$ is impossible too.

Then for $\omega \in \Omega^*$, $\rho^* \in (0, +\infty)$. For any $m, M \in K$ satisfying $m < \rho^* < M$,
if n_j is large enough, then

$$\begin{aligned} \sum_{i=1}^{n_j} \log(1 - e^{-mU_i(\omega)})\mathbf{1}_{(X_i(\omega) \leq U_i(\omega))}/n_j \leq &\sum_{i=1}^{n_j} \log(1 - e^{-\hat{\rho}_{n_j}(\omega)U_i(\omega)})\mathbf{1}_{(X_i(\omega) \leq U_i(\omega))}/n_j \\ \leq &\sum_{i=1}^{n_j} \log(1 - e^{-MU_i(\omega)})\mathbf{1}_{(X_i(\omega) \leq U_i(\omega))}/n_j. \end{aligned}$$

In an obvious way, rewrite the above inequalities as

$$\Psi_{n_j}(m,\omega) \le \Psi_{n_j}(\hat{\rho}_{n_j}(\omega),\omega) \le \Psi_{n_j}(M,\omega). \tag{2.8}$$

Then

$$\begin{aligned} &\liminf_{j\to\infty}\Psi_{n_j}(m,\omega) \le \liminf_{j\to\infty}\Psi_{n_j}(\hat{\rho}_{n_j}(\omega),\omega) \\ \le &\limsup_{j\to\infty}\Psi_{n_j}(\hat{\rho}_{n_j}(\omega),\omega) \le \limsup_{j\to\infty}\Psi_{n_j}(M,\omega). \end{aligned} \tag{2.9}$$

Note that since $m, M \in K$,

$$\liminf_{j\to\infty}\Psi_{n_j}(m,\omega) = \lim_{n\to\infty}\Psi_n(m,\omega) = E(\log(1-e^{-mU})\mathbf{1}_{(X\le U)}). \tag{2.10}$$

$$\limsup_{j\to\infty}\Psi_{n_j}(M,\omega) = \lim_{n\to\infty}\Psi_n(M,\omega) = E(\log(1-e^{-MU})\mathbf{1}_{(X\le U)}). \tag{2.11}$$

Recall the monotone convergence theorem:

If f_n is a monotone convergent sequence and $f_n \to f$ and are all integrable, then $\int f_n(x) \to \int f(x)$.

Since $g(\rho,\omega) = \mathbf{1}(X_i(\omega) \le U_i(\omega)\log(1-e^{-\rho U_i}(\omega))$ is a monotone function of ρ, $g(m,\cdot)$ (or $g(M,\cdot)$) is an increasing function in m (or M) as $m \uparrow \rho^*$ (or $M \downarrow \rho^*$), and $E(g(m,\omega) = \int g(m,\omega)dP$, by the monotone convergence theorem, taking limits as $m, M \to \rho^*$ yields

$$E(g(m,\cdot)) \to E(g(\rho^*,\cdot)) \text{ and } E(g(M,\cdot)) \to E(g(\rho^*,\cdot)). \tag{2.12}$$

Then it follows from (2.8) through (2.12) that

$$\sum_{i=1}^{n_j}\log(1-e^{-\hat{\rho}_{n_j}(\omega)U_i(\omega)})\mathbf{1}_{(X_i(\omega)\le U_i(\omega))}/n_j \to E(\log(1-e^{-\rho^* U})\mathbf{1}_{(X\le U)}). \tag{2.13}$$

(**Remark: In (2.13), since we are dealing $\lim_{n\to\infty}\sum_{i=1}^n g(\hat{\rho}_n)$, we may not have**

$$\lim_{n\to\infty}\sum_{i=1}^n g(\hat{\rho}_n) = \lim_{n\to\infty}\sum_{i=1}^n \lim_{n\to\infty} g(\hat{\rho}_n),$$

even if all the limits make sense.)

Since $\log(1-e^{-\rho(V_i-U_i)})$ is also a monotone function of ρ, it can be shown that

$$\begin{aligned} &\sum_{i=1}^{n_j}\log(1-e^{-\hat{\rho}_{n_j}(\omega)(V_i-U_i)(\omega)})\mathbf{1}_{(X_i(\omega)\in(U_i(\omega),V_i(\omega)])}/n_j \\ &\to E(\log(1-e^{-\rho^*(V-U)})\mathbf{1}_{(X\in(U,V])}). \end{aligned}$$

Thus, $\liminf_{n_j\to\infty} l_{n_j}(\hat{\rho}_{n_j}(\omega)) = \mu(\rho^*)$, since the last two summations in (2.5) do not involve $\hat{\rho}$ and converge almost surely to their means, respectively. Then inequality (2.4) yields

$$\mu(\rho_o)) \le \lim_{n_j\to\infty} l_{n_j}(\hat{\rho}_{n_j}(\omega)) = \mu(\rho^*),\ \forall\ \omega \in \Omega^*,$$

which is Eq. (2.6). This concludes our proof. □

Remark. The standard theory for proving the consistency of the MLE does not work here, as it usually requires that the MLE belongs to compact set, which is not the case here.

Remark.

1. $P(U < V) = 1 \not\Rightarrow U \not\perp V$.
   e.g. If $(U,V) \sim U((0,1)\times(1,2))$, then $U \perp V$ and $P(U<V)=1$.
2. $P(U<V)=1$ and $\sup\{t:\ F_U(t)<1\} > \inf\{t:\ F_V(t)>0\} \Rightarrow U \not\perp V$.

§**3.5.3. Asymptotic Normality of the MLE**.

Hereafter, we prove the asymptotic normality under the assumption given in §3.5.

Theorem 3. *Under the assumptions in Theorem 2, the MLE $\hat{\rho}$ of ρ satisfies that $\sqrt{n}(\hat{\rho}-\rho_o)$ converges in distribution to $N(0,\sigma^2)$, where $\sigma^2 = -1/E(\frac{\partial^2 l(\rho)}{\partial \rho^2})\big|_{\rho=\rho_o}$.*

Proof. We shall show for each t,

$$P\{\frac{\sqrt{n}(\hat{\rho}-\rho_o)}{\sigma} \le t\} \to \Phi(t),\ as\ n\to\infty.$$

Let Ω^o be the event such that $\hat{\rho}\to\rho_o$, $\lim_{n\to\infty}\sum_{lc} 1/n < 1$ and $\lim_{n\to\infty}\sum_{rc} 1/n < 1$. Then $\mathrm{P}(\Omega^o)=1$.

Now for each $\omega\in\Omega^o$ (and supress ω in the expressions), Eq. (2.1) yields

$$\begin{aligned}
\frac{\partial l(\rho)}{\partial \rho} =& \frac{1}{n}\sum_{i=1}^n \frac{U_i e^{-\rho U_i}}{1-e^{-\rho U_i}} 1_{(X_i\le U_i)} + \frac{1}{n}\sum_{i=1}^n \frac{(V_i-U_i)e^{-\rho(V_i-U_i)}}{1-e^{-\rho(V_i-U_i)}} 1_{(U_i<X_i\le V_i)} \\
& - \frac{1}{n}\sum_{i=1}^n U_i 1_{(U_i<X_i\le V_i)} - \frac{1}{n}\sum_{i=1}^n V_i 1_{(X_i>V_i)} \\
=& \frac{1}{n}\sum_{i=1}^n (-U_i + \frac{U_i}{1-e^{-\rho U_i}}) 1_{(X_i\le U_i)} + \frac{1}{n}\sum_{i=1}^n (-(V_i-U_i) + \frac{(V_i-U_i)}{1-e^{-\rho(V_i-U_i)}}) 1_{(U_i<X_i\le V_i)} \\
& - \frac{1}{n}\sum_{i=1}^n U_i 1_{(U_i<X_i\le V_i)} - \frac{1}{n}\sum_{i=1}^n V_i 1_{(X_i>V_i)}
\end{aligned}$$

and

$$-\frac{\partial^2 l(\rho)}{\partial \rho^2} = \frac{1}{n}\sum_{i=1}^n \frac{U_i^2 e^{-\rho U_i}}{(1-e^{-\rho U_i})^2} 1_{(X_i\le U_i)} + \frac{1}{n}\sum_{i=1}^n \frac{(V_i-U_i)^2 e^{-\rho(V_i-U_i)}}{(1-e^{-\rho(V_i-U_i)})^2}) 1_{(U_i<X_i\le V_i)}, \tag{3.1}$$

which are continuous in $\rho\in(0,+\infty)$.

For each $\omega\in\Omega^o$, for n large enough, $\sum_{lc} 1/n < 1$ and $\sum_{rc} 1/n < 1$, thus the MLE $\hat{\rho}(\omega)\in(0,+\infty)$ by Theorem 1. Since $\frac{\partial^2 l(\rho)}{\partial\rho^2}(\omega)$ is continuous in ρ,

$$\frac{\partial l(\hat{\rho})}{\partial \rho}\ (= \frac{\partial l(\rho)}{\partial \rho}(\omega)\Big|_{\hat{\rho}(\omega)})\ = 0. \tag{3.2}$$

By the first order Taylor expansion,

$$\frac{\partial l(\rho)}{\partial \rho}(\omega)\Big|_{\rho_o} - \frac{\partial l(\rho)}{\partial \rho}(\omega)\Big|_{\hat{\rho}(\omega)} = \frac{\partial^2 l(\rho)}{\partial \rho^2(\omega)}\Big|_{\rho^*(\omega)} (\rho_o - \hat{\rho}(\omega)), \tag{3.3}$$
$$\text{provided that } \omega\in\Omega^o,$$

where $\rho^*(\omega)$ is between ρ_o and $\hat{\rho}(\omega)$, and thus by the assumption on Ω^o,

$$\rho^*(\omega)\to\rho_o,\ \forall\ \omega\in\Omega^o. \tag{3.4}$$

Denote

$$p(X,U,V,\rho) = F(U;\rho)^{\mathbf{1}_{(X\le U)}} \times (S(U;\rho)-S(V;\rho))^{\mathbf{1}_{(U<X\le V)}} \times S(V;\rho)^{\mathbf{1}_{(V<X)}}$$

and

$$Z = \frac{\partial}{\partial\rho}\log p(X,U,V,\rho)\Big|_{\rho=\rho_o}.$$

Then

$$l(\rho) = \frac{1}{n}\sum_{i=1}^n \log p(X_i,U_i,V_i,\rho),$$

$$\frac{\partial l}{\partial \rho}\Big|_{\rho=\rho_o} = \overline{Z}.$$

By (3.2) and (3.3),

$$-\sqrt{n}(\hat{\rho}(\omega)-\rho_o)\frac{\partial^2 l(\rho^*)}{\partial \rho^2}(\omega) = \sqrt{n}\cdot\overline{Z}(\omega) \text{ if } \hat{\rho}(\omega)\in(0,\infty) \text{ (or } \omega\in\Omega^o). \tag{3.5}$$

By the CLT,

$$\sqrt{n}(\overline{Z}-E(Z)) \xrightarrow{\mathcal{D}} N(0,\sigma_Z^2). \tag{3.6}$$

We claim:

$$E(Z)=0 \text{ and } \sigma_Z^2 = E(Z^2) = E((\frac{\partial}{\partial\rho}\ln p(X,U,V,\rho_o))^2) = -E(\frac{\partial^2 \ln p(X,U,V,\rho_o)}{\partial\rho^2}). \tag{3.7}$$

Then

$$\sqrt{n}(\hat{\rho}-\rho_o) \xrightarrow{\mathcal{D}} N(0,\tau^2)$$

for some $\tau^2 = \sigma_Z^2/(E(\frac{\partial^2 l(\rho_o)}{\partial\rho^2}))^2$, provided

$$\frac{\partial^2 l(\rho)}{\partial\rho^2}\Big|_{\rho^*} \to E(\frac{\partial^2 l(\rho_o)}{\partial\rho^2}) \text{ a.s..} \tag{3.8}$$

(Here we utilize the following results: by (3.5)

$$\{-\sqrt{n}(\hat{\rho}-\rho_o)\frac{\partial^2 l(\rho^*)}{\partial\rho^2} = \sqrt{n}\overline{Z}\} \supset \{\hat{\rho}\in(0,\infty)\},$$

$$\textit{thus } \{-\sqrt{n}(\hat{\rho}-\rho_o)\frac{\partial^2 l(\rho^*)}{\partial\rho^2} \neq \sqrt{n}\overline{Z}\} \subset \{\hat{\rho}\notin(0,\infty)\},$$

$$|P\{\sqrt{n}(\hat{\rho}-\rho_o)/\tau_n \le t\} - P\{\sqrt{n}\frac{\overline{Z}}{\sigma_Z}\le t\}| \qquad \tau_n = -\sigma_Z/\frac{\partial^2 l(\rho^*)}{\partial\rho^2}$$

$$= \Big|P\{\sqrt{n}(\hat{\rho}-\rho_o)/\tau_n \le t\}$$

$$- P\{\sqrt{n}(\hat{\rho}-\rho_o)/\tau_n \le t, \hat{\rho}\in(0,\infty)\} + P\{\sqrt{n}\frac{\overline{Z}}{\sigma_Z}\le t, \hat{\rho}\in(0,\infty)\}$$

$$(=0 \text{ by } (3.5))$$

$$- P\{\sqrt{n}\frac{\overline{Z}}{\sigma_Z}\le t\}\Big|$$

$$\le P\{\hat{\rho}\notin(0,\infty)\} + P\{\hat{\rho}\notin(0,\infty)\} \qquad (|P(A)-P(A\cap B)|\le P(B^c))$$

$$\to 0, \textit{ as } n\to\infty \text{ due to consistency of } \hat{\rho}.$$

$$P\{\sqrt{n}\frac{\overline{Z}}{\sigma_Z}\le t\}\to\Phi(t) \textit{ as } n\to\infty, \qquad \text{where } \Phi \text{ is the cdf of } N(0,1) \text{ (due to CLT)}$$

$$\textit{thus}$$

$$P\{\sqrt{n}(\hat{\rho}-\rho_o)/\tau_n\le t\} = P\{\sqrt{n}(\hat{\rho}-\rho_o)/\tau_n\le t\} - P\{\sqrt{n}\frac{\overline{Z}}{\sigma_Z}\le t\} + P\{\sqrt{n}\frac{\overline{Z}}{\sigma_Z}\le t\}\to\Phi(t)$$

Moreover, by Slustky's theorem

($W_n \xrightarrow{\mathcal{D}} W$ *and* $T_n \xrightarrow{\mathcal{D}} t$ *imply that* $W_nT_n \xrightarrow{\mathcal{D}} Wt$),

letting $W_n = \sqrt{n}(\hat{\rho}-\rho_o)/\tau_n$ and $T_n = \tau_n/\tau$,

we have $P\{\sqrt{n}(\hat{\rho}-\rho_o)/\tau\le t\}\to\Phi(t)$ or $\sqrt{n}(\hat{\rho}-\rho_o)/\tau \xrightarrow{\mathcal{D}} N(0,\tau^2)$)

Hereafter, we prove (3.7) and (3.8). Verify that

$$\begin{aligned}
E(Z) =&E(\frac{\partial \log p(X,U,V,\rho)}{\partial\rho}\big|_{\rho=\rho_o})\\
=&E(\frac{\frac{\partial p(X,U,V,\rho)}{\partial\rho}}{p(X,U,V,\rho_o)}\big|_{\rho=\rho_o})\\
=&E(E(\frac{\frac{\partial p(X,U,V,\rho)}{\partial\rho}}{p(X,U,V,\rho)}\big|_{\rho=\rho_o}|(U,V)))\\
=&E\big\{\frac{\frac{\partial F(U;\rho)}{\partial\rho}\big|_{\rho=\rho_o}}{F(U;\rho_o)}F(U;\rho_o)+\frac{\frac{\partial S(V;\rho)}{\partial\rho}\big|_{\rho=\rho_o}}{S(V;\rho_o)}S(V;\rho_o)\\
&+\frac{\frac{\partial(S(U;\rho)-S(V;\rho))}{\partial\rho}\big|_{\rho=\rho_o}}{(S(U;\rho_o)-S(V;\rho_o))}(S(U;\rho_o)-S(V;\rho_o))\big\}\\
=&E[\frac{\partial F(U;\rho)}{\partial\rho}\big|_{\rho=\rho_o}+\frac{\partial S(V;\rho)}{\partial\rho}\big|_{\rho=\rho_o}+\frac{\partial(S(U;\rho)-S(V;\rho))}{\partial\rho}\big|_{\rho=\rho_o}]\\
=&E[\frac{\partial(F(U;\rho)+S(V,\rho)+S(U,\rho)-S(V,\rho))}{\partial\rho}\big|_{\rho=\rho_o}]\\
=&E(\frac{\partial 1}{\partial\rho}\Big|_{\rho_o})\\
=&0.
\end{aligned}$$

Thus (3.7) holds. Verify that (by (3.1) and formula $\frac{x}{1-x}=-1+\frac{1}{1-x}$),

$$\begin{aligned}
-\frac{\partial^2 l(\rho)}{\partial\rho^2}\Big|_{\rho^*} =&\frac{1}{n}\sum_{i=1}^n\frac{-U_i^2}{1-e^{-\rho U_i}}1_{(X_i\le U_i)}\Big|_{\rho^*}+\frac{1}{n}\sum_{i=1}^n\frac{U_i^2}{(1-e^{-\rho U_i})^2}1_{(X_i\le U_i)}\Big|_{\rho^*}\\
&+\frac{1}{n}\sum_{i=1}^n\frac{-(V_i-U_i)^2}{1-e^{-\rho(V_i-U_i)}})1_{(U_i<X_i\le V_i)}\Big|_{\rho^*}+\frac{1}{n}\sum_{i=1}^n\frac{(V_i-U_i)^2}{(1-e^{-\rho(V_i-U_i)})^2})1_{(U_i<X_i\le V_i)}\Big|_{\rho^*}
\end{aligned}\tag{3.9}$$

Applying the technique in proving Eq. (2.9) in the proof of Theorem 2 to the four summands in (3.9), we can show

$$\frac{\partial^2 l(\rho)}{\partial\rho^2}\Big|_{\rho^*}\to E(\frac{\partial^2 l(\rho_o)}{\partial\rho^2})\text{ a.s..}\tag{3.10}$$

Thus (3.8) holds. It is worth mentioning that in (3.10),

$$\frac{\partial^2 l}{\partial\rho^2}=\int\int\int\frac{\partial^2}{\partial\rho^2}\ln p(x,u,v,\rho)dF_n(x,u,v)\text{ and}$$

$$E(\frac{\partial^2 l}{\partial\rho^2})=\int\int\int\frac{\partial^2}{\partial\rho^2}\ln p(x,u,v,\rho)dF_o(x,u,v),$$

where F_n is the edf of F_o and F_o is the cdf of (X,U,V).

It is easy to show that

$$E(-\frac{\partial^2 l(\rho_o)}{\partial\rho^2})=E(\frac{U^2e^{-\rho_o U}}{(1-e^{-\rho_o U})^2}1_{(X\le U)})+E(\frac{(V-U)^2e^{-\rho_o(V-U)}}{(1-e^{-\rho_o(V-U)})^2})1_{(U<X\le V)}).$$

Finally, verify that the Fisher information number

$$\sigma_Z^2=E((\frac{\partial\log p(X,U,V,\rho_o)}{\partial\rho})^2)=-E(\frac{\partial^2 l(\rho_o)}{\partial\rho^2})=1/\tau^2.\tag{3.11}$$

Thus

$$\sqrt{n}(\hat{\rho} - \rho_o) \xrightarrow{\mathcal{D}} N(0, \tau^2). \quad \square$$

Comment. This proof can be replaced by the general theory, *e.g.* Cramér's theorem. However, we still need to verify conditions required by the theory. In particular, Cramér's theorem requires that there is a function $K(x, u, v)$ such that $E_{\rho_o}(K(X, U, V)) < \infty$ and $|\frac{d^2 l(\rho; x, u, v)}{d\rho^2}|$ is bounded by $K(x, u, v)$ uniformly in some neighborhood of ρ_o.

§3.5.4. Homework:

(1) Prove (3.10).
(2) Prove Equation (3.11).
(3) Check the existence of the MLE of ρ under the DC model
(4) Under the assumption in Theorem 2, compute $\mu(\rho)$ when $(U, V) = (i, i+2)$ w.p.1/2, $i = 1, 2$.

Chapter 4. Univariate nonparametric estimation

§4.1. Introduction.

Suppose that

the failure time $X \sim F_o$ (cdf),
(X_i, L_i, R_i), $i = 1$, ..., n are i.i.d. from an extended random vector (X, L, R).

Question: Observed (L_i, R_i)s,

$$F_o = ?$$

This is called a nonparametric estimation problem.

The parameter space can be viewed as

$$\Theta_o = \{F : \ F \text{ is a cdf}\}.$$

However, it is more convenient to viewed the parameter space as

$$\Theta = \{F : \ F(t) \uparrow, \ F : [-\infty, \infty] \to [0, 1], \ F(-\infty) = 0 \text{ and } F(\infty) = 1. \ \}. \tag{1.1}$$

Define an interval $I_i = \begin{cases} (L_i, R_i] & \text{if } L_i < R_i \\ [L_i, R_i] & \text{if } L_i = R_i. \end{cases}$

Let $\mu_F(\cdot)$ be a measure induced by F, i.e.,

$$\mu_F(I_i) = \begin{cases} F(R_i) - F(L_i) & \text{if } L_i < R_i \\ F(R_i) - F(L_i-) & \text{if } L_i = R_i. \end{cases}$$

Definition. The nonparametric likelihood function based on data (L_i, R_i), $i = 1$, ..., n, is

$$Ł(F) = \prod_{i=1}^{n} \mu_F(I_i), \ F \in \Theta.$$

The above definition is the same as the parametric definition, except that f is replaced by $f(t) = F(t) - F(t-)$. In other words, it is assumed that F is discrete, though F_o maybe continuous.

Definition. The generalized maximum likelihood estimator (GMLE) of F_o is an $F = \hat{F} \in \Theta$ such that

$$\hat{F} \text{ maximizes } Ł(F) \text{ over } \Theta.$$

Remark.

1. The GMLE is also called the nonparametric MLE.
2. If F is discrete, $Ł(F)$ is the same as the definition of parametric likelihood function $Ł(\phi)$ in §3.2.

§4.1.2. Homework.

Prove the following statement: If the data is complete, but X_i's are not necessarily distinct (there could be ties), then the GMLE of F_o is given by

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(X_i \le x)}.$$

$\hat{F}$ above is called the empirical distribution function (edf).

Properties of the edf:

1. Given observations, it is a discrete cdf.
2. It gives equal weight $\frac{1}{n}$ to each observation.
3. The df at t is $\hat{f}(t) = \frac{\sum_{i=1}^{n} \mathbf{1}_{(X_i=t)}}{n}$ and $\hat{F}(t) = \sum_{x \le t} \hat{f}(x)$.
4. $\overline{X} = \sum_t t\hat{f}(t)$.

§**4.2. The RC model and the product-limit estimator (PLE)**.

Assume the RC model, i.e.,
$Y \sim G$,
X and Y are independent,
observe $(Z, \delta) = (X \wedge Y, \mathbf{1}_{(X \le Y)})$. Then

$$\begin{aligned}\mathcal{L}(F) &= \log Ł(F) = \log \prod_{i=1}^{n} \mu_F(I_i) \\ &= \sum_{i:\ ex} \log f(Z_i) + \sum_{i:\ rc} \log S(Z_i), \text{ where } f(x) = F(x) - F(x-).)\end{aligned}$$

The GMLE of F_o under the RC model is $\hat{F}_{pl} = 1 - \hat{S}_{pl}$, where

$$\hat{S}pl(t) = \prod_{t \ge Z_{(i)}} (1 - \frac{1}{n-i+1})^{\delta_{(i)}},$$

where $Z_{(1)} \le \cdots \le Z_{(n)}$ are order statistics of Z_is and $\delta_{(i)}$ is the δ_j associated with $Z_{(i)}$. Here we use the convention that $x < x+$ in our ordering. The GMLE is called the PLE or Kaplan-Meier estimator after Kaplan and Meier (1958).

4.2.1. Understanding the PLE.

1. Redistribution to the right algorithm.

At each time point t, the PLE redistributes the weight $\hat{S}_{pl}(t-)$ equally to each of the observations to the right of t including t.

In particular, let $a_1 < \cdots < a_m$ be all the distinct points among uncensored Z_1, ..., Z_n, the PLE only puts weights on these a_ks. Let

$d_k = \sum_{i=1}^{n} \mathbf{1}_{(L_i = R_i = a_k)}$, # of deaths at a_k;
$r_k = \sum_{i=1}^{n} \mathbf{1}_{(Z_i \ge a_k)}$, # of people in risk at a_k;

a. The PLE $\hat{S}_{pl}(t)$ is constant at $[a_{k-1}, a_k)$, $k = 1$, ..., m, where $a_0 = -\infty$.
b. For $t < a_1$, $\hat{S}_{pl}(t) = 1$.
c. For $t = a_1$, $\hat{S}_{pl}$ distributes the mass 1 equally to each of the r_1 observations to the right of a_1 (including a_1).
   Total r_1 in risk, thus each with probability $\frac{1}{r_1}$.
   Since d_1 deaths at a_1,
   $\Rightarrow \hat{S}_{pl} = 1 - \frac{d_1}{r_1}$ $(= \hat{P}(X > a_1))$.
d. At a_k, total of $\hat{S}_{pl}(a_{k-1})$ mass remains on $[a_k, +\infty]$.
   $\hat{S}_{pl}$ redistributes the mass $\hat{S}_{pl}(a_{k-1})$ equally to each observation in risk to the right,
   thus each has probability $\frac{\hat{S}_{pl}(a_{k-1})}{r_k}$
   $\Rightarrow \hat{S}_{pl}(a_k) = \hat{S}_{pl}(a_{k-1})(1 - \frac{d_k}{r_k})$.

2. The above algorithm results in the expression

$$\hat{S}_{pl}(t) = \prod_{k:\ t \geq a_k} (1 - \frac{d_k}{r_k}).$$

Note that

$$\hat{S}_{pl}(t) = \hat{S}_{pl}(a_m), \quad t > a_m.$$

Thus $\hat{F}_{pl} = 1 - \hat{S}_{pl}$ may not be a cdf. In particular,

$$\lim_{t \to \infty} \hat{F}_{pl}(t) = \hat{F}_{pl}(Z_{(n)}) < 1 \text{ if } \delta_{(n)} = 0. \tag{2.1}$$

We define

$$\hat{S}_{pl}(+\infty) = 0,$$

so that $\hat{F}_{pl} \in \Theta$, where $\hat{F}_{pl} = 1 - \hat{S}_{pl}$. This means that the PLE puts weight $\hat{S}_{pl}(Z_{(n)})$ at ∞.

3. The PLE $\hat{F}_{pl}(t)$ is nondecreasing in t, but may not be a proper cdf as (2.1) may hold. Some people use the convention that

$$\hat{S}_{pl}(t) = 0 \ \ if\ t > Z_{(n)}. \tag{2.2}$$

However, it can be shown that such a definition does not have optimal asymptotic properties.

The following table calculates the PLE using the Leukaemia data group 0, as in Table 1.2 of §2.3.

Table 1. Calculation of PLE

Remission Time	Reverse Order (K)	$(1 - \frac{1}{n-i+1})^{\delta_{(i)}}$	$(1 - \frac{d_k}{r_k})$	$\hat{S}_{pl}(a_k)$
6	21	20/21		
6	20	19/20		
6	19	18/19	18/21	18/21
6+	18	1		18/21
7	17	16/17	16/17	$\frac{18}{21} \cdot \frac{16}{17}$
9+	16	1		$\frac{18}{21} \cdot \frac{16}{17}$
10	15	14/15	14/15	$\frac{18}{21} \cdot \frac{16}{17} \cdot \frac{14}{15}$
10+	14	1	1	$\frac{18}{21} \cdot \frac{16}{17} \cdot \frac{14}{15}$
11+	13	1	1	$\frac{18}{21} \cdot \frac{16}{17} \cdot \frac{14}{15}$
13	12	11/12	11/12	$\frac{18}{21} \cdot \frac{16}{17} \cdot \frac{14}{15} \cdot \frac{11}{12}$
16	11	10/11	10/11	$\frac{18}{21} \cdot \frac{16}{17} \cdot \frac{14}{15} \cdot \frac{10}{12}$
17+	10	1		$\frac{18}{21} \cdot \frac{16}{17} \cdot \frac{14}{15} \cdot \frac{10}{12}$
19+	9	1		$\frac{18}{21} \cdot \frac{16}{17} \cdot \frac{14}{15} \cdot \frac{10}{12}$
20+	8	1		$\frac{18}{21} \cdot \frac{16}{17} \cdot \frac{14}{15} \cdot \frac{10}{12}$
22	7	6/7	6/7	$\frac{6 \cdot 10 \cdot 14 \cdot 16 \cdot 18}{7 \cdot 12 \cdot 15 \cdot 17 \cdot 21}$
23	6	5/6	5/6	$\frac{5 \cdot 10 \cdot 14 \cdot 16 \cdot 18}{7 \cdot 12 \cdot 15 \cdot 17 \cdot 21}$
25+	5	1		$\frac{5 \cdot 10 \cdot 14 \cdot 16 \cdot 18}{7 \cdot 12 \cdot 15 \cdot 17 \cdot 21}$
32+	4	1		
32+	3	1		$\frac{5 \cdot 10 \cdot 14 \cdot 16 \cdot 18}{7 \cdot 12 \cdot 15 \cdot 17 \cdot 21}$
34+	2	1		$\frac{5 \cdot 10 \cdot 14 \cdot 16 \cdot 18}{7 \cdot 12 \cdot 15 \cdot 17 \cdot 21}$
35+	1	1		$\frac{5 \cdot 10 \cdot 14 \cdot 16 \cdot 18}{7 \cdot 12 \cdot 15 \cdot 17 \cdot 21}$

§**4.2.1.2. Homework**:

1. Apply the redistribution-to-the-right (RTR) method to the last 13 data from Table 1 and compute by hand the PLE. (Now $n = 13$.)
2. Suppose that $X \sim Bin(3, 1/3)$, $Y \sim Bin(1, 0.4)$. There are two observations under the RC model. Compute the mean of $\hat{S}_{pl}(t)$ for $t \leq 1$.

Remark: Since we do not have observation beyond 1, we do not expect that we can make decent inference on $S_o(t)$ for $t > 1$. However, we can estimate S_o for $t \le 1$.

3. In # 2, is the PLE of S_o an unbiased estimator for $t \le 1$?

Theorem 1. (Johanson (1978)). *Under the assumption in this section, $\hat{S}_{pl}$ is the GMLE of S_o $(= 1 - F_o)$.*

We only prove the theorem under a special case.

Suppose that there are only 3 distinct Z_is, say $a_1 < a_2 < a_3$, among n observations.

Let $[-\infty, a_1)$, $[a_1, a_1]$, (a_1, a_2), $[a_2, a_2]$, (a_2, a_3), $[a_3, a_3]$ and $(a_3, \infty]$ be a partition of $[-\infty, +\infty]$,
let the measure assigned by an $F \in \Theta$ to these intervals be p_1, ,..., p_7, where $\sum_{i=1}^7 p_i = 1$.
Denote $d_k = \sum_{i=1}^n \mathbf{1}_{(Z_i = X_i = a_k)}$
and $r_k = \sum_{i=1}^n \mathbf{1}_{(Z_i \ge a_k)}$, and
denote $c_k = \sum_{i=1}^n \mathbf{1}_{(X_i \ne Z_i = a_k)}$, the number of people censored at a_k. Note that

$$
\begin{aligned}
Ł(F) &= \prod_{k=1}^3 (\mu_F([a_k, a_k]))^{\sum_{i=1}^n \mathbf{1}_{(Z_i = X_i = a_k)}} \prod_{k=1}^3 (\mu_F((a_k, +\infty]))^{\sum_{i=1}^n \mathbf{1}_{(X_i \ne Z_i = a_k)}} \\
&= \prod_{k=1}^3 p_{2k}^{d_k} \prod_{k=1}^3 (p_{2k+1} + \cdots + p_7)^{c_k} \\
&= \prod_{k=1}^3 p_{2k}^{d_k} (p_3 + p_4 + p_5 + p_6 + p_7)^{c_1} (p_5 + p_6 + p_7)^{c_2} (p_7)^{c_3} \\
&\le \prod_{k=1}^3 s_{2k}^{d_k} (0 + s_4 + 0 + s_6 + s_7)^{c_1} (0 + s_6 + s_7)^{c_2} (s_7)^{c_3} \ \text{(if } p_1 + p_2 = s_2, \\
&\qquad\qquad p_3 + p_4 = s_4,\ p_5 + p_6 = s_6,\ p_7 = s_7 \text{ and } s_i = 0,\ i = 1, 3, 5) \\
&= (\prod_{k=1}^3 s_{2k}^{d_k}) (s_4 + s_6 + s_7)^{c_1} (s_6 + s_7)^{c_2} s_7^{c_3} \\
&= \prod_{k=1}^3 s_{2k}^{d_k} \prod_{k=1}^3 (s_{2k+2} + \cdots + s_7)^{c_k}.
\end{aligned}
$$

Note $s_2 + s_4 + s_6 + s_7 = 1$. Then

$$
\begin{aligned}
Ł(F) &\le s_2^{d_1} (s_4 + s_6 + s_7)^{c_1} s_4^{d_2} (s_6 + s_7)^{c_2} s_6^{d_3} s_7^{c_3} \\
&= s_2^{d_1} \big(1 - s_2\big)^{c_1 + d_2 + c_2 + d_3 + c_3} \big(\frac{s_4}{1 - s_2}\big)^{d_2} \big(\frac{s_6 + s_7}{1 - s_2}\big)^{c_2} \big(\frac{s_6}{1 - s_2}\big)^{d_3} \big(\frac{s_7}{1 - s_2}\big)^{c_3} \\
&= s_2^{d_1} \big(1 - s_2\big)^{c_1 + d_2 + c_2 + d_3 + c_3} \big(\frac{s_4}{1 - s_2}\big)^{d_2} \big(1 - \frac{s_4}{1 - s_2}\big)^{c_2 + d_3 + c_3} \\
&\quad \big(\frac{s_6}{1 - s_2 - s_4}\big)^{d_3} \big(1 - \frac{s_6}{1 - s_2 - s_4}\big)^{r_3 - d_3} \\
&= s_2^{d_1} \big(1 - s_2\big)^{r_1 - d_1} \big(\frac{s_4}{1 - s_2}\big)^{d_2} \big(1 - \frac{s_4}{1 - s_2}\big)^{r_2 - d_2} \big(\frac{s_6}{1 - s_2 - s_4}\big)^{d_3} \\
&\quad \big(1 - \frac{s_6}{1 - s_2 - s_4}\big)^{r_3 - d_3} \qquad\qquad (2.3) \\
&\qquad\qquad (as\ r_1 = n,\ r_2 = d_2 + c_2 + d_3 + c_3 \text{ and } r_3 = d_3 + c_3)
\end{aligned}
$$

where $s_2 + s_4 + s_6 + s_7 = 1$ and $s_i \ge 0$. Verify that

(a) (2.3) equals $(s_2^*)^{d_1} \big(1 - s_2^*\big)^{r_1 - d_1} (s_4^*)^{d_2} \big(1 - s_4^*\big)^{r_2 - d_2} (s_6^*)^{d_3} \big(1 - s_6^*\big)^{r_3 - d_3}$; where s_i^*'s are defined in an obvious way, satisfying $s_i^* \in [0, 1]$.

(b) The transformation from (s_2, s_4, s_6, s_7) to $(s_2^*, s_4^*, s_6^*, s_7^*)$ is one-to-one and onto.

Thus, $Ł(F)$ is maximized by setting $s_{2i}^* = d_i / r_i$ for $i = 1, 2, 3$.

It follows from the relation between s_i and s_i^* that
$Ł(F)$ is maximized by $s_2 = d_1 / r_1 = \frac{d_1}{n}$, $s_4 = \frac{r_1 - d_1}{r_1} \frac{d_2}{r_2}$, $s_6 = \cdots$.
Consequently, $\hat{S}_{pl}(a_1) = \frac{r_1 - d_1}{r_1}$, $\hat{S}_{pl}(a_2) = \frac{r_1 - d_1}{r_1} \frac{r_2 - d_2}{r_2}$, $\hat{S}_{pl}(a_3) = \frac{r_1 - d_1}{r_1} \frac{r_2 - d_2}{r_2} \frac{r_3 - d_3}{r_3}$.

§**4.2.1.3. Homework**:

4. Extend the proof of the GMLE from the case $a_m = a_3$ to the general case by induction on the number of distinct Z_is.

§**4.2.2. Properties of the PLE**.

We shall first state the main results on the properties of the PLE and present some simpler proofs under the assumption that the random variables take on finitely many values.

Let $\tau = \sup\{t:\ P(X \le t) < 1,\ P(Y \le t) < 1\}$.

Theorem 1. (Yu and Li (1994)). For any arbitrary F_o and G,

$\lim_{n\to\infty} \sup_{t<\tau} |\hat{S}_{pl}(t) - S_o(t)| = 0$ a.s.

Moreover,

$\lim_{n\to\infty} \sup_{t\le\tau} |\hat{S}_{pl}(t) - S_o(t)| = 0$ a.s. unless $F_o(\tau-) < F_o(\tau)$ and $G(\tau-) = 1$.

Remark. $F_o(t)$, $t > \tau$, is not estimable as there is no observation beyond τ, unless $F_o(\tau-) = 1$.

Several weaker results on the consistency were established earlier by Peterson (1977), Phadia and Van Ryzin (1980), Shorack and Wellner (1986), Wang (1987) and Stute and Wang (1993).

Theorem 2. (Breslow and Crowley (1974) and Gill (1983)). *Under the assumption given in this section*

$$Z_n(t) = \sqrt{n}\left(\frac{\hat{S}_{pl}(t) - S_o(t)}{S_o(t)}\right) \xrightarrow{D} N(0, \sigma^2) \text{ for } t < \tau, .$$

The asymptotic covariance of $\hat{S}_{pl}(t)$ *and* $\hat{S}_{pl}(s)$ *is*

$$nCov(\hat{S}_{pl}(t), \hat{S}_{pl}(s)) \approx S_o(t)S_o(s)\int_0^{t\wedge s} \frac{1}{S_o(x-)S_Y(x-)S_o(x)} dF_o(x), \quad t, s < \tau,$$

where $S_Y(x) = P(Y > x)$. *The above two statements are also right for all* $t, s \le \tau$, *if*

$$\int_0^{\tau} \frac{1}{1 - G(t-)} dF_o(t) < \infty.$$

It follows that

$$\sigma^2_{\hat{S}_{pl}(t)} \approx \frac{1}{n}(S_o(t))^2 \int_0^t \frac{1}{S_o(x-)S_Y(x-)S_o(x)} dF_o(x), \quad t < \tau.$$

It can be estimated by

$$\begin{aligned}
n(\hat{\sigma}_{\hat{S}_{pl}(t)})^2 =& (\hat{S}_{pl}(t))^2 \int_0^t \frac{1}{\hat{S}_{pl}(x-)\hat{S}_Y(x-)\hat{S}_{pl}(x)} d\hat{F}_{pl}(x) \\
=& (\hat{S}_{pl}(t))^2 \sum_{k:\ a_k \le t} \frac{\hat{F}_{pl}(a_k) - \hat{F}_{pl}(a_k-)}{\hat{S}_{pl}(a_k-)\hat{S}_Y(a_k-)\hat{S}_{pl}(a_k)} \\
=& (\hat{S}_{pl}(t))^2 \sum_{k:\ a_k \le t} \frac{\hat{h}(a_k)}{\hat{S}_Y(a_k-)\hat{S}_{pl}(a_k)} \\
=& (\hat{S}_{pl}(t))^2 \sum_{k:\ a_k \le t} \frac{\hat{f}(a_k)}{\hat{S}_Z(a_k-)\hat{S}_{pl}(a_k)},
\end{aligned}$$

where $\hat{h}$ is the sample hazard function and $\hat{S}_Y$ is the PLE of S_Y.

$$\hat{S}_Y(t) = \prod_{i:\ Z_{(i)} \le t} (1 - \frac{1 - \delta_{(i)}}{n - i + 1}).$$

Explain using RTR method. Note

$$\hat{S}_{pl}(t) = \prod_{i:\ Z_{(i)} \le t} (1 - \frac{\delta_{(i)}}{n - i + 1}).$$

$$\hat{S}_Z(t) = \hat{S}_Y(t)\hat{S}_{pl}(t) = \prod_{i:\ Z_{(i)} \le t} (1 - \frac{1}{n-i+1}) = \sum_{i=1}^{n} \frac{\mathbf{1}_{(Z_i > t)}}{n}.$$

§**4.2.2.2. Homework**:

1. Suppose that $X \sim Bin(3, 1/3)$, $Y \sim Bin(1, 0.4)$. Under the RC model, what is the limit of $\hat{S}_{pl}(2)$?
2. Suppose $X \sim Exp(\rho)$, where $E(X) = 1/\rho$, and $Y \sim U(0, 4)$. Select a ρ_o. Generate 100 RC data, plot, on the same figure, the PLE $\hat{S}_{pl}$, the true survival function $S(\cdot, \rho_o)$ and the $S(\cdot, \hat{\rho})$, where $\hat{\rho}$ is the parametric MLE of ρ. **Moreover, add the parametric MLE of the survival function of a normal distribution $N(\mu, 1)$ based on the 100 RC data using Monte Carlo method and the parametric MLE of the survival function of a $U(\theta, 5)$.** Make comments on their deviations from the true $S(t, \rho)$ at each time point t.
3. Using the data in problem #2, we can derive the confidence intervals for $S_o(t)$ based on $\hat{S}_{pl}(t)$ and based on the other two estimates of S_o. They are

$$(\hat{S}_{pl}(t) - z_{\alpha/2}\hat{\sigma}_{\hat{S}_{pl}(t)}, \hat{S}_{pl}(t) + z_{\alpha/2}\hat{\sigma}_{\hat{S}_{pl}(t)})$$

and

$$(S(t, \hat{\rho} + z_{\alpha/2}\hat{\sigma}_{\hat{\rho}}), S(t, \hat{\rho} - z_{\alpha/2}\hat{\sigma}_{\hat{\rho}})),$$

etc., respectively, where $\Phi(-z_\alpha) = \alpha$ and Φ is the cdf of $N(0, 1)$ (Actually, for the $U(\theta, 5)$, one can use the Bootstrap method to get SD of the MLE). Both the ends are curves and each pair of the curves induced by the ends of the confidence interval is called a confidence band. Plot on the same figure the **three** confidence bands and the true survival function $S(, \rho)$.

R program for putting the two graphs on one paper.

```
par(mfrow=c(3,1))
x=(-35:45)/10
y=1-pnorm(1.65-x)
plot(x,y,type="l",lty=1 ,xlim=c(-3.5,4.5), ylim=c(0,1.0))
y=1-pnorm(1.65-2*x)
lines(x,y,lty=2)
y=1-pnorm(1.65-3*x)
lines(x,y,lty=3,col=1)
y=1-pnorm(1.65-4*x)
lines(x,y,lty=4,col=2)
leg.names< −c("x1","x2", "x3", "x4")
legend(-2.5, 0.77, leg.names, lty=c(1,2,3,4),cex=1.5)
#————————————
x=rexp(100,0.5)
mean(x)
[1] 1.503257
z=rnorm(200,1.5,3)
u=rnorm(200,1.5,3)
v=3*z+4*u
u[max(v)]
[1] 2.364634
z[max(v)]
[1] -0.1354150
```

Idea for Monte Carlo Method in computing MLE of μ assuming $N(\mu, 1)$ and based on some data:

1. Guess the range of μ and then generate 1000-10000 number in that range.
2. Compare the likelihoods with these numbers being μ

Splus code for Monte Carlo method:

```
x=rexp(100)
d=rbinom(100,1,0.9)
```

```
n=1000
m=runif(n,-1,4)
l=-10000
j=-1
for(i in 1:n){
y=m[i]
L=sum(d*log(dnorm(x,y,1)) + (1-d)*log(1-pnorm(x,y,1)))
if (L>l) {
l=L
j=i
}
}
m[j]
```

4. Suppose there are 6 right-censored observations (Z_i, δ_i): (1,1), (2,0), (3,1), (2,1), (2,1), (4,1). Compute the PLE $\hat{S}_X$ and $\hat{S}_Y$ of S_X and S_Y respectively, on $[0, \infty)$. Show their product is the empirical survival function $\tilde{S}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(t<Z_i)}$. That is, $\tilde{S}(t) = \hat{S}_X(t)\hat{S}_Y(t)$ for all t.
5. Suppose that (Z_i, δ_i), $i = 1, ..., 100$, is a random sample of right-censored data and suppose that the failure time X has an exponential distributions with the parameter ρ and the censoring time $Y = 2$ with probability 1. The PLE $\hat{S}_{pl}$ and $\hat{S}_1 = e^{-\hat{\rho}t}$ are both estimators of S_X. Compute their asymptotic standard deviations explicitly. Which is bigger ? **Hint: you may draw their graphs first.**
6. If you want to estimate a cdf F based on a random sample of right-censored data (Z_i, δ_i), $i = 1, ..., n$, and you believe the cdf is an exponential distribution, what is your estimator?
7. Suppose $X \sim Exp(\rho)$, where $E(X) = 1/\rho$, and $Y \equiv 2$. Select a ρ_o. Generate 3 RC data. Compute the PLE of $S(t)$ and its Variance. Construct a CI for $S(t)$ with the significance level you chosen and the 3 data you generated. Notice that your CI is a statistic, depends only on the data.

We shall give a proof of Theorems 1 and 2 under the assumption that Z takes on finitely many values, with the largest values of Y is τ. Let $a_1 < \cdots < a_m$ be all the possible values of $X \le \tau$ and $\tau = a_m$. Assume that Y can only take values among a_js. Let d_k, c_k and r_k be defined as before. The likelihood function becomes

$$\mathrm{L}(\vec{p}) = \prod_{i=1}^{m} p_i^{d_i} (\sum_{j>i} p_j)^{c_i},$$

Then the problem reduces to a parametric problem of multinomial distribution

$$(Y_1, ..., Y_{2m-1}) \sim M(n, c\theta_1, ..., c\theta_{2m-1}),$$

where for $i = 1, ..., m$, $Y_i = d_i$, $Y_{m+i} = c_i$, $\theta_i = p_i$, $\theta_{m+i} = \sum_{j>i} p_j$, $p_i = P\{X = a_i\}$ and $c = 1/\sum_{i=1}^{2m} \theta_i$. Here θ_i's are function of $\vec{p} = (p_1, ..., p_{m-1})$. The multinomial distribution belongs to the exponential family and its consistency and asymptotic normality can be proved easily by standard approach such as Cramér's theorem (see, e.g., Ferguson (1996)) However, it is not easy to verify the expression for the variance of $\hat{S}_{pl}$ from the inverse of the Fisher information matrix. Thus we use a different approach.

The following consistency proof helps understanding the PLE.

A simple proof of Theorem 1.

The PLE can be written as

$$\hat{S}_{pl}(t) = \prod_{k:\ t \ge a_k} (1 - \frac{d_k}{r_k}),\ .$$

Since m is finite,

$$\begin{aligned} \frac{d_k}{n} &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(Z_i = X_i = a_k)} \\ &\to E(\mathbf{1}_{(Z=X=a_k)}) \quad \text{a.s. by SLLN.} \end{aligned} \tag{2.4}$$

$$= P(X = a_k \le Y) = P(X = a_k)P(Y \ge a_k),$$

$$\frac{r_k}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(Z_i \ge a_k)}$$

$$\to E(\mathbf{1}_{(Z \ge a_k)}) \text{ a.s. by SLLN.}$$

$$= P(Y \ge a_k, X \ge a_k) = P(X \ge a_k)P(Y \ge a_k). \tag{2.5}$$

Prove

$$\hat{S}_{pl}(t) \to S_o(t) \text{ for } t \in [a_{k-1}, a_k) \text{ a.s.} \tag{2.6}$$

by induction on $k = 1, ..., m$, where $a_0 = -\infty$.

$k = 1$, $S_o(t) = 1$ and $\hat{S}_{pl}(t) = 1$ for $t < a_1$. Thus $\hat{S}_{pl}(t) \to S_o(t)$ for $t \in [a_0, a_1)$ a.s..

Assume that (2.6) holds for $k < m$.

For $t \in [a_k, a_{k+1})$,

$$\begin{aligned}
\hat{S}_{pl}(t) =& \hat{S}_{pl}(a_k) = \hat{S}_{pl}(a_{k-1})(1 - \frac{d_k}{r_k}) \\
=& \hat{S}_{pl}(a_{k-1})(1 - \frac{d_k/n}{r_k/n}) \\
\to& S_o(a_{k-1})(1 - \frac{P(X = a_k)P(Y \ge a_k)}{P(X \ge a_k)P(Y \ge a_k)}) \text{ a.s.} \\
& \text{(by induction assumption on } k \text{ and by (2.4) and (2.5))} \\
=& S_o(a_{k-1})(1 - \frac{P(X = a_k)}{P(X \ge a_k)}) \\
=& P(X > a_{k-1})\frac{P(X > a_k)}{P(X \ge a_k)} \\
=& P(X > a_k) = S_o(a_k) = S_o(t).
\end{aligned} \tag{2.7}$$

Thus (2.6) holds for $k + 1$ as well.

This completes the induction proof.

It is easy to show that $\hat{S}_{pl}(a_m) \to S_o(a_m)$ replacing k by m in (2.5) and (2.7). Note that $\tau = a_m$. Since S_o takes finitely many values, point-wise strong consistent implies uniform strong consistency. Thus it completes the proof of Theorem 1 under the finite assumption. □

The following proof makes the expression of $\sigma^2_{\hat{S}_{pl}(t)}$ more explicit than the inverse of the Fisher information matrix.

A simple proof of Theorem 2. We shall now give a proof under the simple assumption that X takes on finitely many values before τ. Assume $f(a_k) > 0$ for each k. Then

$$\hat{S}_{pl}(t) = \prod_{k:\ a_k \le t} (1 - \frac{d_k/n}{r_k/n}),$$

which is a function of $d_1, ..., d_m$ and $r_1, ... , r_m$. In particular, for $t = a_j$, $j < m$, we can write

$$\ln\hat{S}_{pl}(t) = g(d_1/n, r_1/n, ..., d_j/n, r_j/n) = \sum_{i=1}^{j} \ln(1 - \frac{d_i/n}{r_i/n}).$$

That is

$$g(\mathbf{w}) = \ln(1 - \frac{w_1}{w_2}) + \cdots + \ln(1 - \frac{w_{2j-1}}{w_{2j}}), \ where\ \mathbf{w}^t = (w_1, w_3, w_4, w_5, ..., w_{2j}),$$

as $w_2 = r_1/n = n/n = 1$ is a constant. Write

$$\overline{\mathbf{W}} = (\overline{W}_1, \overline{W}_3, \overline{W}_4, ..., \overline{W}_{2j-1}, \overline{W}_{2j})^t = (d_1/n, d_2/n, r_2/n, ..., d_j/n, r_j/n)^t,$$

as $\overline{W}_1 = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(X_i = a_1 \le Y_i)$, $\overline{W}_3 = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(X_i = a_2 \le Y_i)$, $\overline{W}_4 = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(Z_i \ge a_2)$, ...

Then it follows from a corollary of Slutsky's theorem that the $\ln\hat{S}_{pl}(t)$ is asymptotic normally distributed with an asymptotic variance

$$\sigma^2_{\ln\hat{S}_{pl}(t)} \approx \frac{\partial g}{\partial \mathbf{w}^t}\Big|_{\mathbf{w}=E(\overline{\mathbf{W}})} \Sigma \frac{\partial g}{\partial \mathbf{w}}\Big|_{\mathbf{w}=E(\overline{\mathbf{W}})},$$

where Σ is the covariance matrix of $\overline{\mathbf{W}}$.

$$\sqrt{n}(g(\overline{\mathbf{W}}) - g(E(\overline{\mathbf{W}}))) \xrightarrow{D} N(0,\sigma^2).$$

$$\sqrt{n}(\ln\hat{S}_{pl}(t) - \ln S_o(t)) \xrightarrow{D} N(0,\sigma^2).$$

$$\sqrt{n}(h(\ln\hat{S}_{pl}(t)) - h(\ln S_o(t))) \xrightarrow{D} N(0,\sigma_2^2) \text{ where } h(t) = e^t.$$

$$\sqrt{n}(\hat{S}_{pl}(t) - S_o(t)) \xrightarrow{D} N(0,\sigma_2^2).$$

That is, $\hat{S}_{pl}(t)$ is asymptotic normally distributed with an asymptotic variance

$$\begin{aligned}\sigma_2^2/n \approx \sigma^2_{\hat{S}_{pl}(t)} \approx & \frac{\partial e^x}{\partial x}\Big|_{x=\ln S_o(t)} \sigma^2_{\ln\hat{S}_{pl}(t)} \frac{\partial e^x}{\partial x}\Big|_{x=\ln S_o(t)} \\ = & (S_o(t))^2 \sigma^2_{\ln\hat{S}_{pl}(t)},\end{aligned}$$

as $\hat{S}_{pl}(t) = e^{\ln(\hat{S}_{pl}(t))}$. For simplicity, for now, we let $j = 2$. Then $\mathbf{w}^t = (w_1, w_3, w_4)$,

$$\frac{\partial g}{\partial \mathbf{w}} = \begin{pmatrix} \frac{1}{-w_2(1-\frac{w_1}{w_2})} \\ \frac{1}{-w_{2j}(1-\frac{w_{2j-1}}{w_{2j}})} \\ \frac{\frac{w_{2j-1}}{w_{2j}^2}}{(1-\frac{w_{2j-1}}{w_{2j}})} \end{pmatrix} = \begin{pmatrix} \frac{-1}{w_2 - w_1} \\ \frac{-1}{w_{2j} - w_{2j-1}} \\ \frac{w_{2j-1}}{w_{2j}(w_{2j} - w_{2j-1})} \end{pmatrix}$$

$$\frac{\partial g}{\partial \mathbf{w}}\Big|_{\mathbf{w}=E(\overline{\mathbf{W}})} = \begin{pmatrix} \frac{-1}{S_o(a_1)S_Y(a_1-)} \\ \frac{-1}{S_o(a_j)S_Y(a_j-)} \\ \frac{f_o(a_j)}{S_o(a_j-)S_o(a_j)S_Y(a_j-)} \end{pmatrix} \text{ (by (2.4) and (2.5))},$$

$$\sum = Cov(\overline{\mathbf{W}}) = \frac{1}{n}Cov(\mathbf{W}) = \frac{1}{n}Cov(\mathbf{1}_{(X=Z=a_1)}, \mathbf{1}_{(X=Z=a_2)}, \mathbf{1}_{(Z\ge a_2)}),$$

Denote

$$n\Sigma = (s_{ih})_{(2j-1)\times(2j-1)}.$$

Note that $s_{ij} = s_{ji}$. Then

$$\begin{aligned} s_{11} = & E([\mathbf{1}_{(X=Z=a_1)}]^2) - [E(\mathbf{1}_{(X=Z=a_1)})]^2 = f_o(a_1)S_Y(a_1-)(1 - f_o(a_1)S_Y(a_1-)) \\ & \begin{pmatrix} s_{12} & s_{13} \\ s_{22} & s_{23} \\ \cdot & s_{33} \end{pmatrix} \\ = & \begin{pmatrix} -E(\mathbf{1}_{(X=Z=a_1)})E(\mathbf{1}_{(X=Z=a_2)}) & -E(\mathbf{1}_{(X=Z=a_1)})E(\mathbf{1}_{(Z\ge a_2)}) \\ E(\mathbf{1}_{(X=Z=a_2)}) - [E(\mathbf{1}_{(X=Z=a_2)})]^2 & E(\mathbf{1}_{(X=Z=a_2)}) - E(\mathbf{1}_{(X=Z=a_2)})E(\mathbf{1}_{(Z\ge a_2)}) \\ \cdot & E(\mathbf{1}_{(Z\ge a_2)}) - [E(\mathbf{1}_{(Z\ge a_2)})]^2 \end{pmatrix} \\ = & \begin{pmatrix} -f_o(a_1)S_Y(a_1-)f_o(a_2)S_Y(a_2-) & -f_o(a_1)S_Y(a_1-)S_o(a_2-)S_Y(a_2-) \\ f_o(a_2)S_Y(a_2-)(1 - f_o(a_2)S_Y(a_2-)) & f_o(a_2)S_Y(a_2-)(1 - S_o(a_2-)S_Y(a_2-)) \\ \cdot & S_o(a_2-)S_Y(a_2-)(1 - S_o(a_2-)S_Y(a_2-)) \end{pmatrix} \end{aligned}$$

Denote

$$\frac{\partial g}{\partial \mathbf{w}}\Big|_{\mathbf{w}=E(\overline{\mathbf{W}})} = (p_1, p_2, p_3)^t \quad (= \mathbf{p}^t).$$

Then

$$n\sigma^2_{\ln\hat{S}_{pl}(t)} \approx n\mathbf{p}^t \sum \mathbf{p} = (1,1) \begin{pmatrix} p_1 & 0 & 0 \\ 0 & p_2 & p_3 \end{pmatrix} \begin{pmatrix} s_{11} & s_{12} & s_{13} \\ s_{12} & s_{22} & s_{23} \\ s_{13} & s_{23} & s_{33} \end{pmatrix} \begin{pmatrix} p_1 & 0 \\ 0 & p_2 \\ 0 & p_3 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Verify that

$$(\, p_1 \quad 0 \quad 0\,) \begin{pmatrix} s_{11} & s_{12} & s_{13} \\ s_{12} & s_{22} & s_{23} \\ s_{13} & s_{23} & s_{33} \end{pmatrix} \begin{pmatrix} 0 \\ p_2 \\ p_3 \end{pmatrix} = 0. \tag{2.8}$$

By symmetry

$$\begin{aligned} n\sigma^2_{\ln\hat{S}_{pl}(t)} \approx & (1,1) \begin{pmatrix} p_1 & 0 & 0 \\ 0 & p_2 & p_3 \end{pmatrix} \begin{pmatrix} s_{11} & 0 & 0 \\ 0 & s_{22} & s_{23} \\ 0 & s_{32} & s_{33} \end{pmatrix} \begin{pmatrix} p_1 & 0 \\ 0 & p_2 \\ 0 & p_3 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ = & (1,1) \begin{pmatrix} p_1 p_1 s_{11} & 0 \\ 0 & p_2 p_2 s_{22} + p_2 p_3 s_{23} + p_3 p_2 s_{23} + p_3 p_3 s_{33} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ = & p_1 p_1 s_{11} + p_2 p_2 s_{22} + p_2 p_3 s_{23} + p_3 p_2 s_{23} + p_3 p_3 s_{33} \end{aligned} \tag{2.9}$$

Then it yields

$$\sigma^2_{\ln\hat{S}_{pl}(t)} \approx \frac{1}{n} \sum_{a_k \le t} \frac{f_o(a_k)}{S_o(a_k-) S_Y(a_k-) S_o(a_k)}. \tag{2.10}$$

§**4.2.2.3. Homework**:

8. Verify (2.8).
9. Using (2.9) to verify (2.10).

References.

[*] Breslow, N.E. and Crowley, J. (1974). A large-sample study of the life table and product limit estimates under random censorship. *Ann. Statist.* 2 437-453.

[*] Ferguson, T. S. (1996). A course in large sample theory. p. 119, Cahpman & Hall. New York.

[*] Gill, R. (1983). Convergence of the product limit estimator on the entire half line. *Ann. Statist.* 11 49-59.

[*] Johanson, S. (1978). The product limit estimator as maximum likelihood estimator. *Scan. J. Statist.*, 5, 195-199.

[*] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Stat. Assoc.*, 53, 457-481.

[*] Miller Jr., R. G. (1981). Survival analysis. *Wiley.* p. 62.

[*] Peterson, A. V. (1977). Expressing the Kaplan-Meier estimator as a function of the empirical subsurvival functions. *JASA*. 72, 854-858.

[*] Phadia, E.G. and Van Ryzin, J (1980). A note on the convergence rates for the product limit estimator. *Ann. Statist.* 8, 673-678.

[*] Shorack, G. R. and Wellner, J. A. (1986). Empirical Processes with applications to statistics. *Wiley*, New York.

[*] Stute, W. and Wang J.L. (1993). The strong law under random censorship. *Ann. Statist.* 21 1591-1607.

[*] Wang, J. G. (1987). A note on the uniform consistency of the Kaplan-Meier estimator. *Ann. Statist.* 15, 1313-1316.

§**4.3. C1 model and GMLE**.

Assume the case 1 interval censorship model, i.e.,
Each patient is followed once at time $Y \sim G$;
X and Y are independent;
Observable random vector is

$$(L, R) = \begin{cases} (-\infty, Y) & \text{if } X \le Y \\ (Y, +\infty) & \text{if } X > Y, \end{cases}$$

or

$$(Y, \delta) \quad \text{where } \delta = \mathbf{1}_{(X \le Y)}.$$

It is more convenient to use (Y, δ). It corresponds to an interval

$$I = \begin{cases} (-\infty, Y] & \text{if } X \le Y \\ (Y, +\infty) & \text{if } X > Y. \end{cases}$$

Given a random sample of size n, say (Y_i, δ_i), $i = 1$, ..., n, let $a_1 < \cdots < a_m$ be all the distinct values of Y_is. Let

$$N_-(a_j) = \sum_{i=1}^{n} \mathbf{1}_{(X_i \le Y_i = a_j)}$$

$$N_+(a_j) = \sum_{i=1}^{n} \mathbf{1}_{(X_i > Y_i = a_j)}$$

$$N(a_j) = \sum_{i=1}^{n} \mathbf{1}_{(Y_i = a_j)}.$$

The likelihood function is

$$\begin{aligned} Ł(F) &= \prod_{i=1}^{n} \mu_F(I_i) \\ &= \prod_{j=1}^{m} [(F(a_j))^{N_-(a_j)} (S(a_j))^{N_+(a_j)}], \ F \in \Theta, \end{aligned}$$

where Θ is defined in (1.1) of §4.1, i.e.,

$$\Theta = \{F : \ F \text{ is a nondecreasing function on } [-\infty, +\infty], \ F(-\infty) = 0 \text{ and } F(+\infty) = 1 \ \}.$$

That is

$$F(a_1) \le F(a_2) \le \cdots \le F(a_m).$$

Let $s_j = F(a_j)$, the log likelihood is

$$\mathcal{L}(F) = \sum_{j=1}^{m} [N_-(a_j) \log s_j + N_+(a_j) \log(1 - s_j)], \quad 0 \le s_1 \le \cdots \le s_m \le 1. \tag{3.1}$$

$$\frac{\partial \mathcal{L}}{\partial s_j} = \frac{N_-(a_j)}{s_j} - \frac{N_+(a_j)}{1 - s_j} = 0$$

$$\Rightarrow s_j = \tilde{F}(a_j) = \frac{N_-(a_j)}{N(a_j)}. \tag{3.2}$$

If $\tilde{F} \in \Theta$, or $\quad 0 \le s_1 \le \cdots \le s_m \le 1$, it is the GMLE (in fact, each summand looks like the log likelihood of a binomial distribution $\text{Bin}(N(a_j), s_j)$). Otherwise, it is not and the GMLE will be on the boundary $s_i = s_j$, $i \ne j$.

Example 1. A random sample of $n = 4$. The observations (Y_i, δ_i)s are $(1, 1)$, $(2, 0)$, $(2, 1)$, $(3, 1)$. Then

$$\tilde{F}(t) = \begin{cases} 0 & \text{if } t < 1 \\ 1 & \text{if } t \in [1, 2) \\ \frac{1}{2} & \text{if } t \in [2, 3) \\ 1 & \text{if } t \ge 3. \end{cases}$$

$F \notin \Theta$. Thus $\tilde{F}$ is not a GMLE of F. To be completed in a homework.

Theorem 1. *A GMLE of F_o under C1 model is*

$$\hat{F}(t) = \hat{F}(Y_{(j)}) \quad if \;\; t \in [Y_{(j)}, Y_{(j+1)}),\; j = 0, 1, ..., n,$$

where $Y_{(0)} = -\infty$, $Y_{(n+1)} = +\infty$, and

$$\hat{F}(Y_{(j)}) = \max_{i \le j} \min_{k \ge j} \frac{\sum_{i \le h \le k} \delta_{(h)}}{k - i + 1},\; j = 1, ..., n, \tag{3.3}$$

and $\delta_{(h)}$ is the δ_i that is associated with the order statistic $Y_{(h)}$, i.e. $Y_i = Y_{(h)}$, provided $\hat{F} \in \Theta$.

For proof, we refer to Groeneboom and Wellner (1992) or Ayer *et al.* (1955).

Remark.

1. The GMLE $\hat{F}(t)$ is uniquely determined at each Y_i.
2. The GMLE is not uniquely determined in the interval $(-\infty, a_1)$ and (a_i, a_{i+1}), where $i \ge 0$, unless $\hat{F}(a_i) = \hat{F}(a_{i+1})$.
3. It can be shown that the following definitions of $\hat{F}(Y_{(j)})$ are equivalent to (3.3) (homework).

$$\hat{F}(Y_{(j)}) = \min_{k \ge j} \max_{i \le j} \frac{\sum_{i \le h \le k} \delta_{(h)}}{k - i + 1}, \tag{3.4}$$

$$\hat{F}(Y_{(j)}) = \max_{i \le j} \min_{k \ge i} \frac{\sum_{i \le h \le k} \delta_{(h)}}{k - i + 1}, \tag{3.5}$$

$$\hat{F}(Y_{(j)}) = \min_{k \ge j} \max_{i \le k} \frac{\sum_{i \le h \le k} \delta_{(h)}}{k - i + 1}. \tag{3.6}$$

Example 2. For a random sample of 4 C1 observations, (1,0), (2,1), (3,0) and (3,1). The GMLE is uniquely determined on the set $(-\infty, 1] \cup [2, 3]$. $\hat{F}(i) = 0$, 2/3, 2/3, respectively, and it arbitrary on $(1, 2) \cup (3, \infty)$.

$$\hat{F}(Y_{(j)}) = \max_{i \le j} \min_{k \ge j} \frac{\sum_{i \le h \le k} \delta_{(h)}}{k - i + 1},\; j = 1, ..., n, \tag{3.3}$$

Let $A_{ik} = \frac{\delta_{(i)} + \cdots + \delta_{(k)}}{k - i + 1}$.

$B_{ij} = \min_{k \ge j} \frac{\delta_{(i)} + \cdots + \delta_{(k)}}{k - i + 1}$, $i \le j$.

Order data: (1,0), (2,1), (3,1), (3,0).

$i \backslash k$	1	2	3	4	B_{ij}	$\hat{F}(Y_{(j)})$
1	$\frac{\delta_{(1)}}{1}$	$\frac{\delta_{(1)}+\delta_{(2)}}{2}$	$\frac{\delta_{(1)}+\delta_{(2)}+\delta_{(3)}}{3}$	$\frac{\delta_{(1)}+\delta_{(2)}+\delta_{(3)}+\delta_{(4)}}{4}$		
2		$\frac{\delta_{(2)}}{1}$	$\frac{\delta_{(2)}+\delta_{(3)}}{2}$	$\frac{\delta_{(2)}+\delta_{(3)}+\delta_{(4)}}{3}$		$max_{i \le j} B_{ij}$
3			$\frac{\delta_{(3)}}{1}$	$\frac{\delta_{(3)}+\delta_{(4)}}{2}$		
4				$\frac{\delta_{(4)}}{1}$		

$i \backslash k$	1	2	3	4	B_{i1}	$\hat{F}(Y_{(1)})$	B_{i2}	$\hat{F}(Y_{(2)})$	B_{i3}	$\hat{F}(Y_{(3)})$	B_{i4}	$\hat{F}(Y_{(4)})$
1	0	1/2	2/3	2/4	0	0	1/2		1/2		1/2	
2		1	2/2	2/3			2/3	2/3	2/3		2/3	
3			1/1	1/2					1/2	2/3	1/2	
4				0							0	2/3

In fact, it is more convenient to use formula (3.5) (or (3.6)) which results in the following matrix.

$$\hat{F}(Y_{(j)}) = \max_{i \le j} \min_{k \ge i} \frac{\sum_{i \le h \le k} \delta_{(h)}}{k - i + 1}, \tag{3.5}$$

$$\hat{F}(Y_{(j)}) = \min_{k\geq j} \max_{i\leq k} \frac{\sum_{i\leq h\leq k} \delta_{(h)}}{k-i+1}. \tag{3.6}$$

$i\backslash k$	1	2	3	4	$\min_k A_{ik}$	$\hat{F}(Y_{(i)})$
1	$\frac{\delta_{(1)}}{1}$	$\frac{\delta_{(1)}+\delta_{(2)}}{2}$	$\frac{\delta_{(1)}+\delta_{(2)}+\delta_{(3)}}{3}$	$\frac{\delta_{(1)}+\delta_{(2)}+\delta_{(3)}+\delta_{(4)}}{4}$	m_1	m_1
2		$\frac{\delta_{(2)}}{1}$	$\frac{\delta_{(2)}+\delta_{(3)}}{2}$	$\frac{\delta_{(2)}+\delta_{(3)}+\delta_{(4)}}{3}$	m_2	$\max\{m_1, m_2\}$
3			$\frac{\delta_{(3)}}{1}$	$\frac{\delta_{(3)}+\delta_{(4)}}{2}$	m_3	$\max\{m_1, m_2, m_3\}$
4				$\frac{\delta_{(4)}}{1}$	m_4	$\max_i m_i$
$\max_i A_{ik}$	M_1	M_2	M_3	M_4		
$\hat{F}(Y_{(k)})$	$\min_k M_k$	$\min\{M_2, M_3, M_4\}$	$\min\{M_3, M_4\}$	M_4		
	Eq. (3.6)					

$i\backslash k$	1	2	3	4	$\min_k A_{ik}$	$\hat{F}(Y_{(i)})$	*Eq.*(3.5)
1	0	1/2	2/3	2/4	0	0	
2		1	2/2	2/3	2/3	2/3	
3			1/1	1/2	1/2	2/3	
4				0	0	2/3	
$\max_i A_{ik}$	0	1	1	2/3			
$\hat{F}(Y_k)$	0	2/3	2/3	2/3			
Eq. (3.6)							

Obtain the GMLE directly from Ł in (3.1):

$\mathcal{L} = \ln(1-s_1) + \ln s_2 + \ln s_3 + \ln(1-s_3)$, $0 \leq s_1 \leq s_2 \leq s_3 \leq 1$

Boundary: $s_1 = 0$; ($\Leftarrow$ $s_2 = 0$; $s_3 = 0$;)

$s_3 = 1$; ($\Leftarrow$ $s_2 = 1$; $s_1 = 1$;)

$s_1 = s_2$; $s_2 = s_3$; $\Leftarrow$ $(s_1 = s_3)$.

$\mathbf{s}:$	*critical pt* $(0,1,1/2)$	$s_1 = 0$	$s_3 = 1$	$s_1 = s_2$	*others*
$\mathcal{L}:$	**violating the constraint**	no need to check			

$s_1 = 0$:

$\mathbf{s}:$	*critical point* $(0,1,1/2)$	$s_2 = s_3$	$s_2 = 0$	$s_3 = 1$
$\mathcal{L}:$	*violating the constraint*	no need to check		

$s_2 = s_3$:

$\mathbf{s}:$	*critical point* $(0, 2/3, 2/3)$	$s_2 = 0$	$s_3 = 1$	$s_1 = s_2 = s_3$
$\mathcal{L}:$	$\ln(\frac{2^2}{3^3})$	$-\infty$	$-\infty$	$\ln\frac{1}{2^4}$

$s_1 = s_2$:

$\mathbf{s}:$	$(1/2, 1/2, 1/2)$	$s_1 = 0$	$s_3 = 1$
$\mathcal{L}:$	$\ln\frac{1}{2^4}$	$-\infty$	$-\infty$

Other boundaries can be skipped.

Obtain the GMLE through $L(F)$ directly: The distinct observations Z_i's, 1,2,3, partition $(-\infty, \infty)$ as 4 disjoint intervals. Let the weight assigns by an F to these intervals as p_1, ..., p_4. The the likelihood function

$$\begin{aligned} L =& (p_2+p_3+p_4)(p_1+p_2)(p_4)(p_1+p_2+p_3) \\ \leq& (q_2+q_4)(q_2)(q_4)(q_2) \\ & \qquad\qquad\qquad\qquad\qquad\qquad \textit{setting } p_1+p_2+p_3 = q_2,\ p_4 = q_4 \\ =& (q_2+q_4)(q_2)^2(q_4) \qquad\qquad\qquad\qquad\qquad\qquad q_2+q_4 = 1 \\ =& (q_2)^2(q_4) \end{aligned}$$

Thus $q_2 = 2/3$, $q_4 = 1/3$.

§**4.3.1.2. Homework**.

1. Derive the GMLE of F using data from Example 1. Use two approaches: (1) use formula (3.3); (2) derive directly from the likelihood function in (3.1).
2. Assuming Y is discrete, show that the estimator $\tilde{F}$ in (3.2) is a consistent estimator of $F_o(a_j)$ and find its asymptotic variance.

3. Show that the definitions (3.4), (3.5) and (3.6) of $\hat{F}(Y_{(j)})$ are equivalent to (3.3). **Hint:** Inspect the matrix $(A_{ik})_{j\times(n-j+1)}$, where $A_{ik} = \frac{\sum_{i\le h\le k}\delta_{(h)}}{k-i+1}$. Try first $n = 3$ or 4 and use induction argument on j.

In many practical situations, consistency of the GMLE is only relevant for $t \le \tau$, where

$$\tau = \sup\{t:\ G(t) < 1\}.$$

Consistency of the GMLE has been investigated by Ayer *et al.* (1955), Groeneboom and Wellner (1992), Yu *et al.* (1998), and Schick and Yu (2000).

Theorem 2. *Under the C1 model, The GMLE $\hat{F}$ satisfies*
(1) $\lim_{n\to\infty}\int|\hat{F}(t) - F_o(t)|dG(t) = 0$ *a.s. (Schick and Yu (2000));*
(2) $\hat{F}(a) \to F_o(a)$ *almost surely for each* $a \in \mathcal{A}$, *where* $\mathcal{A} = \{a : P(Y = a) > 0\}$ *(Yu et al. (1998));*
(3) If F_o is continuous in $(0,\tau]$, $P\{Y = \tau\} > 0$ or $F_o(\tau) = 1$, and the range of Y is dense in $[0,\tau]$, then $\hat{F}$ is uniformly strongly consistent on $(-\infty,\tau]$, i.e., $\sup_{x\le\tau}|\hat{F}(x) - F_o(x)| \to 0$ *a.s. (Schick and Yu (2000)).*

In clinical follow-ups, the studies typically last for a certain period of time, say $[0,\tau]$. It is often that $F_o(\tau) < 1$. Gentleman and Geyer (1994) claimed a vague convergence result in their Theorem 2 and Huang (1996) claimed a uniform strong consistency result in his Theorem 3.1. Both of their results as stated imply the uniform strong consistency of the GMLE on $[0,\tau]$ in the case 1 model, if F_o is continuous and $G'(x) > 0$ on $[0,\tau]$. The following example shows that this is not true.

Example 3. Consider C1 data $(Y_1, \mathbf{1}_{(X_1\le Y_1)}), \ldots, (Y_n, \mathbf{1}_{(X_n\le Y_n)})$, where the survival times $X_1, \ldots, X_n$ are uniformly distributed on $[0,3]$ and the inspection times $Y_1, \ldots, Y_n$ are uniformly distributed on $[0,2]$. Then F_0 is the uniform distribution function on $[0,3]$. Note that on the event $\cup_{j=1}^n B_j$, where $B_j = \{X_j \le 1 \le Y_j, Y_j > Y_i, i = 1,\ldots,n, i \ne j\}$ we have $\hat{F}_n(2) = \hat{F}_n(2-) = \hat{F}(Y_{(n)}) = 1$ (as $Y_{(n)} < 2$ and $max_{i\le n}min_{k\ge n}\frac{\sum_{i\le h\le k}\delta_{(h)}}{k-i+1} = \max_{i\le n}\frac{\sum_{i\le h\le n}\delta_{(h)}}{n-i+1} = 1$ (see (3.1)). The event has probability $1/3 - \frac{1}{3\cdot 2^n} \ge \frac{1}{6}$, as it equals

$$\begin{aligned}
&P\{\cup_{j=1}^n B_j\}\\
=&nP\{X_1 \le 1 \le Y_1, Y_1 > Y_j,\ j = 2, ..., n\}\\
&\qquad\qquad (B_j s \text{ are disjoint events with the same probability})\\
=&nP\{X_1 \le 1\}P\{1 \le Y_1, Y_1 > Y_j,\ j = 2, ..., n\}\\
=&\frac{n}{3}\int_1^2 \frac{1}{2}P\{Y_j < y_1, j = 2, ..., n\}dy_1\\
=&\frac{n}{3}\frac{1}{n}(\frac{y}{2})^n\Big|_1^2\\
=&\frac{1}{3}(1 - 2^{-n}).
\end{aligned}$$

That is $P(\hat{F}(2-) = 1) \ge P(\cup_j B_j) = \frac{1}{3} - \frac{1}{32^n}$.

Since $F_o(2) = F_0(2-) = 2/3$, we see that the following two statements are false:
$\hat{F}(2-)$ converges to $F_o(2-)$ almost surely ($P(\hat{F}(2-) \to F_o(2-)) = 1$),
and $\hat{F}(x)$ converges to $F_o(x)$ almost surely for $x = 2$ ($P(\hat{F}(2) \to F_o(2)) = 1$).

This shows that point-wise convergence on the closed interval $[0,2]$ to a continuous F_o is not implied by the condition: $\frac{dG}{dx} > 0$ for all $x \in [0,2]$.

Remark. The GMLE under the assumptions in Example 2 is consistent for all $t < \tau$. It follows from Proposition 3.2 in Schick and Yu (2000).

Question: Is $\hat{F}(t)$ consistent for all $t < \tau$?

Example 4. Suppose that $X \sim Exp(1)$ and Y has a Poisson distribution with mean 1. Then $\tau = +\infty$.
$\lim_{n\to\infty}\hat{S}(1) = S_o(1)$ a.s. ? ($S_o(1) = e^{-1}$)
$\lim_{n\to\infty}\hat{S}(1.5) = S_o(1.5)$ a.s. ? ($S_o(1.5) = e^{-1.5}$)

Answer:
$\lim_{n\to\infty}\hat{S}(1) = S_o(1)$ a.s. !
$\lim_{n\to\infty}\hat{S}(1.5) \ne S_o(1.5)$ a.s..

$\lim_{n\to\infty} \hat{S}(1.5) = S_o(1)$ a.s..

Thus the GMLE is not consistent at $t \in (i, i+1)$ for all integers i.

Note that under the RC model, the PLE $\hat{S}_{pl}(t)$ is always consistent for $t < \tau$!

Peto (1973) and Turnbull (1976) conjectured that for arbitrary F_o and G, the GMLE is asymptotically normal at the usual $n^{1/2}$ rate. It was, however, shown by Groeneboom and Wellner (1992) that this conjecture is false if F_o and G satisfy certain smoothness assumptions. Indeed, their Theorem 3 below establishes that under differentiability assumptions on F_o and G the convergence is at the slower $n^{1/3}$ rate and the limiting distribution is not normal.

Theorem 3. *Let t_o be such that $0 < F_o(t_o), G(t_o) < 1$, and let F_o and G be differentiable at t_o, with strictly positive derivatives $f_o(t_o)$ and $g(t_o)$, respectively. Then we have, as $n \to \infty$,*

$$n^{1/3} \frac{\hat{F}(t_o) - F_o(t_o)}{\{\frac{1}{2} F_o(t_o) S_o(t_o) f_o(t_o)/g(t_o)\}^{1/3}} \xrightarrow{D} 2Z,$$

where $Z \equiv \text{argmin}(W(t) + t^2)$ and W is the two-sided Brownian motion starting from 0.

Here, for each ω in the sample space, $Z(\omega) = t_o$, where $W(t_o)(\omega) + t_o^2 \le W(t)(\omega) + t^2$ for $t \ge 0$.

Remark. The main assumption in Theorem 3 is that G is differentiable, i.e. Y is continuous.

There are many practical situations in which Y is discrete. In medical research, for example, the data are often recorded as integers (to represent number of days, weeks etc). Then the conclusion is different.

Let $\mathcal{A}_* = \mathcal{A} \cup \{-\infty, \infty\}$. For $x \in (-\infty, +\infty)$, set

$$x_- := \sup\{a \in \mathcal{A}_* : a < x\} \text{ and } x_+ := \inf\{a \in \mathcal{A}_* : a > x\}.$$

We say x is a *regular point*, if x belongs to $\mathcal{A}$, x_- and x_+ belong to $\mathcal{A}_*$, $x_- < x < x_+$ and $F_0(x_-) < F_0(x) < F_0(x_+)$. It is worth mentioning that there may be infinitely many regular points. For example, if F_0 is strictly increasing and $\mathcal{A}$ is the set of all positive integers, then every positive integer is a regular point.

Theorem 4. (Yu *et al.* (1998)). *Let x be a regular point. Then $n^{1/2}(\hat{F}(x) - F_o(x))$ is asymptotically normal with mean 0 and variance $F_o(x)(1 - F_o(x))/g(x)$. This asymptotic variance can be consistently estimated by $\hat{F}(x)(1 - \hat{F}(x))/N(x)$. Also, if $x_1 < \ldots < x_m$ are regular points, then $n^{1/2}(\hat{F}(x_1) - F_o(x_1), \ldots, \hat{F}(x_m) - F_o(x_m))$ is asymptotically normal with mean vector 0 and diagonal covariance matrix.*

Remark. Theorems 3 and 4 both allow X takes on infinitely many values.

Remark. Suppose that F_o is strictly monotone and Y takes on finitely many values, say at $a_1 < \cdots < a_m$, with $F(a_1) > 0$ and $F(a_m) < 1$. For n large enough,

$$\hat{F}(a_i) = \tilde{F}(a_i), \quad i = 1, ..., m-1.$$

Reason: Let $\epsilon = \min\{F_o(a_1), F_o(a_i) - F_o(a_{i-1}),\ i = 2, ..., m,\ S_o(a_m)\}$. Then $\epsilon > 0$. Let Ω be the event that $N_-(a_j)/n$ and $N(a_j)/n$ converge for $j = 1, ..., m$. Then for each $\omega \in \Omega$, for n large enough, $|\tilde{F}(a_j) - F_o(a_j)| < \epsilon/3$ for all j. It follows that

$$0 \le \tilde{F}(a_1) < F_o(a_1) + \epsilon/3 < F_o(a_2) - \epsilon/3 < \tilde{F}(a_2) < \cdots \tilde{F}(a_m) \le 1. \tag{3.7}$$

As a consequence, $\tilde{F}$ is a GMLE.

In fact, the asymptotic covariance matrix of $(\hat{F}(a_1), ..., \hat{F}(a_m))$ can be estimated by the expression

$$\hat{\Sigma} = \left(-\frac{\partial^2 \mathcal{L}}{\partial \mathbf{s} \partial \mathbf{s}^t} \Big|_{\mathbf{s}^t = (\hat{F}(a_1), ..., \hat{F}(a_m))} \right)^{-1}.$$

Now

$$\frac{\partial \mathcal{L}}{\partial \mathbf{s}^t} = \left(\frac{N_-(a_1)}{s_1} - \frac{N_+(a_1)}{1 - s_1}, ..., \frac{N_-(a_m)}{s_m} - \frac{N_+(a_m)}{1 - s_m}\right).$$

$$\frac{\partial^2 \mathcal{L}}{\partial \mathbf{s} \partial \mathbf{s}^t} = -\begin{pmatrix} \frac{N_-(a_1)}{s_1^2} + \frac{N_+(a_1)}{(1-s_1)^2} & 0 & \cdot & 0 \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & \frac{N_-(a_m)}{s_m^2} + \frac{N_+(a_m)}{(1-s_m)^2} \end{pmatrix}.$$

$$\frac{1}{n}\left(-\frac{\partial^2\mathcal{L}}{\partial\mathbf{s}\partial\mathbf{s}^t}|_{\mathbf{s}^t=(\hat{F}(a_1),...,\hat{F}(a_m))}\right)$$

$$\to\begin{pmatrix}\frac{F_o(a_1)g(a_1)}{F_o^2(a_1)}+\frac{S_o(a_1)g(a_1)}{(S_o(a_1))^2} & 0 & \cdot & 0\\ \cdot & \cdot & \cdots & \cdot\\ 0 & 0 & \cdots & \frac{F_o(a_m)g(a_m)}{F_o^2(a_m)}+\frac{S_o(a_m)g(a_m)}{(S_o(a_m))^2}\end{pmatrix}\text{ a.s.}$$

$$=\begin{pmatrix}\frac{g(a_1)}{F_o(a_1)}+\frac{g(a_1)}{S_o(a_1)} & 0 & \cdot & 0\\ \cdot & \cdot & \cdots & \cdot\\ 0 & 0 & \cdots & \frac{g(a_m)}{F_o(a_m)}+\frac{g(a_m)}{S_o(a_m)}\end{pmatrix}$$

$$=\begin{pmatrix}\frac{g(a_1)}{F_o(a_1)S_o(a_1)} & 0 & \cdot & 0\\ \cdot & \cdot & \cdots & \cdot\\ 0 & 0 & \cdots & \frac{g(a_m)}{F_o(a_m)S_o(a_m)}\end{pmatrix}$$

$$n\hat{\Sigma}\to\begin{pmatrix}\frac{F_o(a_1)S_o(a_1)}{g(a_1)} & 0 & \cdot & 0\\ \cdot & \cdot & \cdots & \cdot\\ 0 & 0 & \cdots & \frac{F_o(a_m)S_o(a_m)}{g(a_m)}\end{pmatrix}\ a.s.$$

Theorem 4 is an extension of the above observation.

However, if F_o is not strictly increasing on $\{a_1,...,a_m\}$, it is not true that

$$\tilde{F}(a_i)=\hat{F}(a_i)\ i=2,...,m-1,\text{ if } n \text{ is large enough,}$$

as (3.7) does not hold. In such a case, Theorem 4 is false.

Question. In application, how to determine when the censoring distribution G is continuous or discrete ?

The important feature of a discrete Y is that there are ties among Y_i's.

If there are few ties, then one should consider Y is continuous.

Otherwise, discrete.

Question. If Y is continuous, how to construct a confidence interval (CI) for $F(t_o)$?

Note that by Theorem 3 and Theorem 2.4 of Banerjee and Wellner (2001),

$$n^{1/3}(\hat{F}(t_o)-F_o(t_o))\xrightarrow{D}hZ,$$

where

$$h=\{4F_o(t_o)S_o(t_o)f_o(t_o)/g(t_o)\}^{1/3}.$$

Thus, a 95% CI for $F_o(t_o)$ is

$$\hat{F}(t_o)\pm n^{-1/3}\hat{h}Q_{0.025}=\hat{F}(t_o)\pm n^{-1/3}\hat{h}0.99818$$

where Q_α is the $100(1-\alpha)$ quantile of the distribution of Z ($P\{Z>Q_\alpha\}=\alpha$), which is provided in Groeneboom and Wellner (2001), and $\hat{h}$ is an estimate of h, *e.g.*,

$$\hat{h}=\{4\hat{F}(t_o)\hat{S}(t_o)\hat{f}(t_o)/\hat{g}(t_o)\}^{-1/3}$$

$$\hat{f}(t)=\int\frac{1}{w_n}K(\frac{x-t}{w_n})d\hat{F}(x),$$

w_n is the window width, $w_n\to 0$, $K(\cdot)$ is a kernel (*e.g.*, $K(x)=\frac{1}{2}\mathbf{1}_{(-1<x\le 1)}$ and thus

$$\hat{f}(t)=\frac{\hat{F}(t+w_n)-\hat{F}(t-w_n)}{2w_n}),$$

$$\hat{g}(t)=\int\frac{1}{w_n}K(\frac{x-t}{w_n})d\hat{G}(x)$$

$$\hat{G}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(Y_i \le x)}.$$

§**4.3.1.3. Homework**.

4. Let F_o be $Exp(\rho)$ (with mean $1/\rho$) and $Y \sim U(0, 3/\rho)$.
   4.a. Generate $n = 400$ case 1 data and compute $\hat{F}(1/\rho)$.
   4.b. Repeat 4.a 50 times and compute the sample mean $\overline{\hat{F}}(1/\rho)$ and sample variance S_n^2.
   4.c. Discuss the region on which the GMLE is consistent.
   4.d. Repeat 4.a and 4.b, with $n = 100$ and compare $\frac{S_{100}}{S_{400}}$ to $\sqrt{400/100}$ and to $(\frac{400}{100})^{1/3}$, which is closer ?
5. Let Y be a discrete random variable taking values 1, 2, 3 and 4; and let $X \sim U(0, 5)$.
   5.a. Generate 400 case 1 data and compute $\hat{F}(3)$.
   5.b. Repeat 4.a 50 times and compute the sample mean $\overline{\hat{F}}(3)$ and sample variance S_n^2.
   5.c. Discuss the region on which the GMLE is consistent.
   5.d. Repeat 5.a and 5.b, with $n = 100$ and compare $\frac{S_{100}}{S_{400}}$ to $\sqrt{400/100}$ and to $(\frac{400}{100})^{1/3}$, which is closer ?

References.

* Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T. and Silverman, E. (1955). An empirical distribution function for sampling incomplete information. *Ann. Math. Statist.* 26, 641-647.
* Bannerjee, M and Wellner, J.A. (2001). Likelihood ratio tests for monotone functions. *Ann. Statist.* 29 1699-1731.
* Gentleman, R. and Geyer, C. J. (1994). Maximum likelihood for interval censored data: Consistency and computation. *Biometrika*, 81, 618-623.
* Groeneboom, P. and Wellner, J.A. (2001). Computing Chernoff's distribution. *J. Comp. and Graphical Statist.* 10, 388-400.
* Huang, J. (1996). Efficient estimation for proportional hazards models with interval censoring. *Ann. Statist.*, 24, 540-568.
* Schick, A. and Yu, Q. Q. (2000). Consistency of the GMLE with mixed case interval-censored data. *Scan. J. of Statist.* Vol. 27 45-55.
* Yu, Q. Q., Schick, A., Li, L. X. and Wong, G. Y. C. (1998). Asymptotic properties of the GMLE in the case 1 interval-censorship model with discrete inspection times. *Canadian Journal of Statistics*, 26, No. 4, 619-627.

§**4.4. Self-consistent (SC) algorithm**.

Under the RC model, C1 model and LC model, the GMLE of F_o has a closed form solution. However, this is not so in general. There are several numerical methods to compute the GMLE:

1. The Newton-Raphson algorithm (Peto (1973));
2. The self-consistent algorithm (Turnbull (1976));
3. The convex minorant algorithm (Groeneboom and Wellner (1992)).

The self-consistent algorithm is easy to implemented, so we introduce it in this section.

Let I_1, ..., I_n denote the observed intervals. Suppose that there are g distinct observations and the first g are distinct. Then the IC data consist of N_1 (l_1, r_1)'s, N_2 (l_2, r_2)'s, ..., and N_g (l_g, r_g)'s, where $N_i \ge 0$ and $N_1 + \cdots + N_g = n$. Define *innermost intervals* $A_j, j = 1, 2, \ldots, m$, induced by I_1, ..., I_n to be all the disjoint intervals which are non-empty intersections of these I_i's (*e.g.* $I_k = I_k \cap I_k$ is an intersection of I_i's) such that

$$A_j \cap I_i = \emptyset \text{ or } A_j \ \ \forall\ i \text{ and } \ j.$$

Let the endpoints of the innermost intervals be a_j and b_j, $j = 1$, ..., m, where

$$a_1 \le b_1 \le a_2 \le b_2 \le \cdots \le a_m \le b_m.$$

Let $\delta_{ij} = \mathbf{1}_{(A_j \subset I_i)}$. The following example illustrates the procedure for finding innermost intervals.

Example 2.1. Suppose that there are five observed intervals $(1, 4]$, $[2, 2]$, $(2, 6]$, $[5, 5]$, and $(1, 6]$. Then there are two exact observations, $I_1 = [2, 2]$ and $I_2 = [5, 5]$, and three censored intervals, $I_3 = (1, 4]$, $I_4 = (2, 6]$ and $I_5 = (1, 6]$. Furthermore, there are three innermost intervals, $A_1 = [2, 2]$, $A_2 = [5, 5]$, and $A_3 = (2, 4]$.

Peto (1973) shows that the GMLE of F_o assigns weight, say $s_1, ..., s_m$, to the corresponding innermost intervals $A_1, ..., A_m$ only. The generalized likelihood function Ł can be simplified as

$$Ł = Ł(s_1, ..., s_m) = \prod_{i=1}^{n} [\sum_j \delta_{ij} s_j], \tag{4.1}$$

where $\mathbf{s}^t = (s_1, ..., s_{m-1}) \in D_s$ is the transpose of $\mathbf{s}$, $D_s = \{\mathbf{s};\ s_i \geq 0, s_1 + \cdots + s_{m-1} \leq 1\}$ and $s_m = 1 - s_1 - \cdots - s_{m-1}$. Turnbull (1976) proposes a self-consistent algorithm for obtaining the GMLE via an iterative procedure as follows.

At step 1, let $s_j^{(1)} = 1/m$ for $j = 1, ..., m$.

At step h,

$$s_j^{(h)} = \sum_{i=1}^{n} \frac{1}{n} \frac{\delta_{ij} s_j^{(h-1)}}{\sum_{k=1}^{m} \delta_{ik} s_k^{(h-1)}}, \quad j = 1, ..., m,\ h \geq 2.$$

Stop when $\mathbf{s}^{(h)}$ converges, *i.e.*, $||\mathbf{s}^{(h)} - \mathbf{s}^{(h-1)}||$ is small enough.

He shows that, as $h \to \infty$, $s_j^{(h)}$ converges to the GMLE, $\hat{s}_j$, which maximizes Ł and satisfies the system of self-consistent equations

$$s_j = \sum_{i=1}^{n} \frac{1}{n} \frac{\delta_{ij} s_j}{\sum_{k=1}^{m} \delta_{ik} s_k}, \quad j = 1, ..., m, \mathbf{s} \in D_s. \tag{4.2}$$

A solution $\mathbf{s} = \hat{\mathbf{s}}$ to (4.2) is called a self-consistent estimator (SCE) of $\mathbf{s}$. An estimate $\hat{F}(t)$ of $F(t)$ can be uniquely defined for $t \in [b_i, a_{i+1}]$ by $\hat{F}(b_i) = \hat{F}(a_{i+1}-) = \hat{s}_1 + \cdots + \hat{s}_i$, but is not uniquely defined for t being in a non-singleton innermost interval (Peto, 1973; Turnbull, 1976). To avoid this ambiguity we define

$$\hat{F}(t) = \sum_{j:\ A_j \subset (-\infty, t]} \hat{s}_j. \tag{4.3}$$

where $(a, b]$ is an empty set if $a = b$. Under the RC model, the definition of $\hat{F}$ given by (4.3) reduces to the PLE. Under the C1 model, it reduces to the max-min solution.

Remark. An SCE of $\mathbf{s}$ can be viewed as a local extreme point of Ł($\mathbf{s}$), subject to the constraint on $\mathbf{s}$. The reason is as follows. Given $j \in \{1, ..., m\}$, let $\mathbf{s}(\epsilon) = (s_1(\epsilon), ..., s_{m-1}(\epsilon))$ be defined by

$$s_k(\epsilon) = \begin{cases} \frac{s_k}{1+\epsilon} & \text{if } k \neq j \\ \frac{s_j + \epsilon}{1+\epsilon} & \text{if } k = j, \end{cases}$$

Write $\Lambda_j(\epsilon) = \ln L(\mathbf{s}(\epsilon))$, then verify that

$$\left.\frac{\partial \Lambda_j(\epsilon)}{\partial \epsilon}\right|_{\epsilon=0} = -n + \sum_{i=1}^{n} \frac{\delta_{ij}}{\sum_{k=1}^{m} \delta_{ik} s_k}.$$

If $\mathbf{s}$ is the GMLE, then the value $\epsilon = 0$ maximizes $\Lambda_j(\epsilon)$ for each j subject to $\epsilon \geq 0$. In other words, we have

$$\left.\frac{\partial \Lambda_j(\epsilon)}{\partial \epsilon}\right|_{\epsilon=0} = -n(1 - \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_{ij}}{\sum_{k=1}^{m} \delta_{ik} s_k}) \leq 0$$

with equality unless $s_j = 0$. Consequently, (4.2) holds. That is, the GMLE satisfies (4.2). Of course the fact that (4.2) holds for $\mathbf{s}$ does not implies that $\mathbf{s}$ is the GMLE, but it implies that $\mathbf{s}$ is a critical point of the likelihood.

§**4.4.2. Homework**.

1. Suppose our data consist of 5 intervals: $(-\infty, 5]$, [2,2], (4,7], (3,8], $(9, +\infty)$. Find all the innermost intervals. Compute the GMLEs of $\mathbf{s}$ and F by two methods: directly from differentiation and by the SC algorithm. Verify that it is a solution of the self-consistent equation (4.2).

(4.2) is in the form of density function as $\hat{s}_i = \hat{f}_X(b_i)$. Its cumulative form is

$$\begin{aligned} H(x) &= \sum_{i=1}^{n} \frac{1}{n} \frac{\mu_H(I_i \cap (-\infty, x])}{\mu_H(I_i)} \\ &= \sum_{i=1}^{n} \frac{1}{n} \mathbf{1}_{(l_i \le x < r_i)} \frac{H(x) - H(l_i)}{H(r_i) - H(l_i)} + \sum_{i=1}^{n} \frac{\mathbf{1}_{(r_i \le x)}}{n}, \ H \in \Theta. \end{aligned}$$

which is equivalent to

$$H(x) = \int_{l \le x < r} \frac{H(x) - H(l)}{H(r) - H(l)} d\hat{Q}(l, r) + \int_{r \le x} \mathbf{1}_{(r \in \mathcal{S}_H)} d\hat{Q}(l, r), \quad H \in \Theta, \tag{4.4}$$

where $Q(l, r) = P(L \le l, R \le r)$,

$$\hat{Q}(l, r) = \sum_{i=1}^{n} \frac{\mathbf{1}_{(l_i \le l, r_i \le r)}}{n} \text{ (the empirical estimator of } Q),$$

$\mathcal{S}_H$ is the support set of H, namely, $x \in \mathcal{S}_H$ if $|H(x + t_k) - H(x)| > 0$ for a sequence $\{t_k\}_{k \ge 1}$ satisfying $t_k \to 0$.

Definition A solution H (or $\mathbf{s}$) to (4.4) (or (4.2)) is called a self-consistent estimator (SCE) of F_o (or $\mathbf{s}$).

Remark GMLE $\Rightarrow$ SCE, but SCE $\not\Rightarrow$ GMLE.

Example 1 Six case 2 IC data: two (1,5] and (3,7], one $(-\infty, 3]$ and $(5, +\infty)$.

3 IIs: $(1, 3]$, (3,5] and (5,7]. Solving equation (4.2) yields two solutions for (s_1, s_2, s_3): (1/3, 1/3, 1/3) and (1/2, 0,1/2). The two SCEs of $\mathbf{s}$ yield two SCEs of F_o:

$$\hat{F}(t) = \begin{cases} 0 & \text{if } t < 3 \\ 1/3 & \text{if } 3 \le t < 5 \\ 2/3 & \text{if } 5 \le t < 7 \\ 1 & \text{if } t \ge 7; \end{cases} \qquad \text{and} \qquad \tilde{F}(t) = \begin{cases} 0 & \text{if } t < 3 \\ 1/2 & \text{if } 3 \le t < 7 \\ 1 & \text{if } t \ge 7. \end{cases}$$

In this example, without using any algorithm, there are **two ways to find the GMLE**: (1) derive directly from Ł(F), or (2) find the SCE that has the largest Ł(F). Verify that

$$\text{Ł}(\hat{F}) > \text{Ł}(\tilde{F}). \tag{4.5}$$

Thus $\hat{F}$ is a GMLE but $\tilde{F}$ is not a GMLE. $\hat{F}$ and $\tilde{F}$ both satisfy the summation equation in (4.4). Thus $\hat{F}$ and $\tilde{F}_2$ are SCEs.

Remark Gu and Zhang (1993) defined an SCE H of F_o to be a solution to the equation in (4.4) without the restriction $H \in \Theta$. They thought that the solution of (4.4) will belong to Θ, as their proofs need the restriction. The following example show that their thought is wrong and there is an $H \notin \Theta$ but the H is a solution to (4.4).

Example 2 Six DC data.

$N_1 = \sum_{i=1}^{n} \mathbf{1}_{(L_i = R_i = 2)} = 1$, $N_2 = \sum_{i=1}^{n} \mathbf{1}_{(L_i = R_i = 5)} = 3$,
$N_3 = \sum_{i=1}^{n} \mathbf{1}_{((L_i, R_i) = (-\infty, 4))} = 1$ and $N_4 = \sum_{i=1}^{n} \mathbf{1}_{((L_i, R_i) = (2, +\infty))} = 1$.

A solution H to the SC equation in (4.4) gives mass $(1/2, -1/4, 3/4)$ to points 2, 4 and 5. Since H has jumps only at 2, 4 and 5, it suffices to verify (4.4) at $-\infty$, 2, 4 and 5. Verify as follows.

$$\begin{aligned} &H(-\infty) = 0, \quad \text{RHD of (4.4)} = 0 \text{ trivially} \\ &H(2) = 1/2, \quad \text{RHD of (4.4)} = \frac{1}{6} + \frac{3}{6} \cdot 0 + \frac{1}{6} \frac{1/2 - 0}{1/4 - 0} = 1/2 \\ &H(4) = 1/4, \quad \text{RHD of (4.4)} = \frac{1}{6} + \frac{3}{6} \cdot 0 + \frac{1}{6} \cdot 1 + \frac{1}{6} \frac{1/4 - 1/2}{1 - 1/2} = 1/4 \\ &H(5) = 1, \quad \text{RHD of (4.4)} = \frac{1}{6} + \frac{3}{6} + \frac{1}{6} + \frac{1}{6} \frac{1 - 1/2}{1 - 1/2} = 1. \end{aligned}$$

Remark. The above examples illustrate that

1. A solution to the SC equation is a critical point under the constraint that $\sum_{i=1}^m s_i = 1$.
2. Moreover, the solution to the SC algorithm with initial point $s_i \geq 0$ and $\sum_{i=1}^m s_i = 1$ is a critical point under the constraint that $s_i \geq 0$ and $\sum_{i=1}^m s_i = 1$.
3. Finally, the solution to the SC algorithm with initial point $s_i = 1/m$ is the GMLE.

§**4.4.3. Homework**.

2. Solve the two SCEs directly from Eq. (4.2) and (4.3) using the data in Example 1. Derive the GMLE in two ways.
3. Define a second GMLE of F_o in Example 1. Where are the GMLE uniquely defined in the example ?
4. Derive an SCE in Example 2 using SC algorithm.

§**4.5. SCE under DC model** Assume the DC model. That is,

1. X and the censoring vector (Z, Y) are independent;
2. $Z < Y$ w.p.1.;
3. Observe $(L, R) = (-\infty, Z)\mathbf{1}_{(X \leq Z)} + (X, X)\mathbf{1}_{(Z < X \leq Y)} + (Y, +\infty)\mathbf{1}_{(X > Y)}$.

Turnbull (1974), Chang and Yang (1987), Chang (1990) Gu and Zhang (1993), and Yu and Li (2001) established consistency and asymptotic normality of the SCE with DC data under sets of regularity conditions that are weaker and weaker.

Define $\tau_l = \sup\{x : \max(S_X(x), S_Z(x)) = 1\}$ and
$\tau_r = \inf\{x : \min(S_X(x), S_Y(x)) = 0\}$.

It is obvious that if $F_o(\tau_l) > 0$ (or $F_o(\tau_r) < 1$), $F_o(x)$ is not identifiable for $x < \tau_l$ (or $x > \tau_r$). Thus in such cases, we only consider the estimation of $F_o(x)$ for $x \in [\tau_l, \tau_r]$.

Denote $P_c(x) = P\{X \text{ is not censored} | X = x\}$ and $K(x) = P\{Z \leq x < Y\}$.

$K(x-) = P_c(x)$ if $P_c(x)$ exists, where $K(x-) = P\{Z < x \leq Y\}$, as $K(x-) = \lim_{y \uparrow x} P\{Z \leq y < Y\}$.

Some assumptions that have been made in the literature are listed as follows.

$(AS1)$ The probability $\mathrm{P}\{X \in (\tau_l, \tau_r] \text{ and } K(X-) = 0\} = 0$.

$(AS2)$ $\mathrm{P}\{L = \tau_r\} > 0$ if $F_o(\tau_r-) < 1$ and $\mathrm{P}\{R = \tau_l\} > 0$ if $F_o(\tau_l) > 0$.

$(AS3)$ $$\int_{\tau_l < u < \tau_r} \frac{-d[S_Y(u) + S_Z(u)]}{S_Y(u) - S_Z(u)} < \infty;$$

$(AS4)$ $K(x-) = P_c(x) > 0$ for all $x \in (\tau_l, \tau_r)$;

$(AS5)$ Z and Y take on finitely many values, say $b_1 < \cdots < b_N$, and $0 < F_o(b_1) < \cdots < F_o(b_N) < 1$.

$(AS6)$ There are at most m IIs for each sample size n and $\mu_{F_o}(A_j) > 0$ for each II A_j

Theorem 1 (Yu and Li (2001)) *Suppose that (AS1) and (AS2) hold, the SCE $\hat{F}$ satisfies*

$$\lim_{n \to \infty} \sup_{x \in [\tau_l, \tau_r]} |\hat{F}(x) - F_o(x)| = 0 \text{ almost surely},$$

$$\lim_{n \to \infty} \sup_{x} |\hat{F}(x) - F_o(x)| = 0 \text{ almost surely if } F_o(\tau_l) = 0 \text{ and } F_o(\tau_r) = 1.$$

Theorem 2 (Yu and Li (2001)) *Suppose that either AS5 or AS6 holds; or AS2, AS3 and AS4 hold, $\tau_l < \tau_r$. Then the GMLE $\hat{F}$ satisfies that $\sqrt{n}(\hat{F} - F_o)$ converges in distribution to a Gaussian process $Z(x)$ on $[\tau_l, \tau_r]$. Furthermore, the GMLE is asymptotically efficient.*

Theorem 3 (Turnbull (1974)) *Suppose that the assumptions in Theorem 2 hold. Let $b_1 < \cdots < b_{k+1}$ be the distinct right endpoints of all the innermost intervals induced by the observed intervals. Then the covariance matrix of the SCE $(\hat{F}(b_1), ..., \hat{F}(b_k))$ can be estimated by $(J(\hat{F}))^{-1}$, where $J(F) = -\left(\frac{\partial^2 \mathcal{L}}{\partial F(b_i) \partial F(b_j)}\right)_{k \times k}$ and $\mathcal{L} = \ln L$. Furthermore, J^{-1} is of the form*

$$J^{-1} = \begin{pmatrix} c_1 & d_1 & 0 & 0 & \cdots & 0 & 0 \\ d_1 & c_2 & d_2 & 0 & \cdots & 0 & 0 \\ 0 & d_2 & c_3 & d_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & d_{k-1} & c_k \end{pmatrix}.$$

Remark Under AS6 and under all the IC models discussed so far, the GMLE of the weights on the II's are consistent and let $b_1 < \cdots < b_{k+1}$ be the right endpoints of the all possible II's, the $(\hat{F}(b_1), ..., \hat{F}(b_k))$ are asymptotically normal and their asymptotic covariance matrix is the inverse of the Fisher information matrix $(E(J(F_o)))^{-1}$ and can be estimated by the inverse of the empirical Fisher information matrix $(J(\hat{F}))^{-1}$.

We shall refer the proofs of the theorems to the literature. By means of the following example, we explain the assumptions AS1-AS6.

Example 1 Let (Z, Y) takes values $(0.5, 2)$, $(2, 4)$ and $(4, 8)$, with equal probabilities g_1, g_2 and g_3, respectively, where $g_1+g_2+g_3 = 1$, and let $F_o(x) = p_1\mathbf{1}(x \geq 1)+p_2\mathbf{1}(x \geq 5)$, where $p_1 = p_2$. Derive all possible SCEs of F_o based on Eq. (4.2) in §4.4. Derive the GMLE of F_o. Are they consistent and asymptotically normal ?

Solution We first solve for SCE of $\mathbf{s}$. Example 2 in §4.4 is a special case.

$\tau_l = 1$ and $\tau_r = 5$.

Possible observed intervals: [1,1], $(-\infty, 2]$, [5,5], $(4, +\infty)$, $(-\infty, 4]$, $(2, +\infty)$.

Possible IIs: [1,1], (2,4], [5,5] (when n is large enough), with weights s_1, s_2, s_3.

Suppose that there are N_1 $(1, 1)$'s or $(-\infty, 2)$'s, N_2 $(5, 5)$'s or $(4, +\infty)$'s, N_3 $(2, +\infty)$'s and N_4 $(-\infty, 4)$'s among n doubly-censored observations.

Eq. (4.2) of §4.4 becomes

$$s_1 = \frac{N_1}{n}\frac{s_1}{s_1} + \frac{N_2}{n}\cdot 0 + \frac{N_3}{n}\cdot 0 + \frac{N_4}{n}\frac{s_1}{s_1+s_2}$$
$$s_2 = \frac{N_1}{n}\cdot 0 + \frac{N_2}{n}\cdot 0 + \frac{N_3}{n}\cdot\frac{s_2}{s_2+s_3} + \frac{N_4}{n}\frac{s_2}{s_1+s_2}$$

(we skip 1 equation in (4.2) as $m = 3$ but $s_1 + s_2 + s_m = 1$).

Solving the equations yields two solutions of $\mathbf{s}$: $(s_1, s_2) = (\frac{N_1+N_4}{n}, 0)$ and $(s_1, s_2) = (\frac{N_1}{N_1+N_3}, U_n)$, where

$$U_n = \frac{N_4}{N_2+N_4} - \frac{N_1}{N_1+N_3}.$$

The first one is an SCE of $\mathbf{s}$ as $s_i \geq 0$ and $s_1 + s_2 \leq 1$. The second one is an SCE if and only if $U_n \geq 0$.

Let

1. $H_1 = \frac{N_1+N_4}{n}\mathbf{1}_{(x\geq 1)} + \frac{N_2+N_3}{n}\mathbf{1}_{(x\geq 5)}$.
2. $H_2 = \frac{N_1}{N_1+N_3}\mathbf{1}_{(x\geq 1)} + U_n\mathbf{1}_{(x\geq 4)} + \frac{N_2}{N_2+N_4}\mathbf{1}_{(x\geq 5)}$.

Thus there are possible two SCEs of F_o: H_1 or H_2 if $U_n \geq 0$. The GMLE is $\hat{F} = \begin{cases} H_1 & \text{if } U_n < 0 \\ H_2 & \text{if } U_n \geq 0. \end{cases}$ The reason is as follows.

Theorem 4 *Given interval-censored data (C1, C2, DC or MIC data), let A_i, $i = 1, ..., m$ be all the distinct IIs. Suppose $\hat{s}_i$, $i = 1, ..., m$ are weights assigned to these IIs by an SCE, respectively, and satisfies $\hat{s}_j = 0 \Rightarrow \tilde{s}_j = 0$ for any other SCE with weights $\tilde{s}_i$, $i = 1, ..., m$, then $\hat{\mathbf{s}}$ is a GMLE.*

The proof of the theorem utilizes the convexity of $-\ln Ł$ in $(s_1, ..., s_k)$.

Note that GMLE is also an SCE. If $U_n < 0$, there is only one SCE of $\mathbf{s}$, which induces H_1, thus H_1 is a GMLE if $U_n < 0$. If $U_n \geq 0$, there are exactly two SCEs, which induce H_1 and H_2. Then it follows either from Theorem 4 or from $Ł(H_2) > Ł(H_1)$ that H_2 is a GMLE if $U_n \geq 0$.

To prove consistency. Verify

AS1: $K(x-) = P_c(x) = P\{X \text{ is not censored}|X = x\} > 0 \ \forall\ x \in \{5\} \subset (\tau_l, \tau_r]$. Thus AS1 holds.

AS2: $F(\tau_l) = 1/2 > 0$ and $P\{L = \tau_l\} = P\{X = 1, (Z, Y) = (0.5, 2)\} = 1/6 > 0$. $F(\tau_r-) = 1/2 < 1$ and $P\{R = \tau_r\} = P\{X = 5, (Z, Y) = (4, 8)\} = 1/6 > 0$. Thus AS2 holds.

Thus the SCEs are consistent.

To discuss the asymptotic normality, verify that

(1) AS5 and AS6 fail, as $F_o(2) = p_1 \not< F_o(4) = p_1$ and

(2) AS4 fails, as $K(3-) = P\{Z < 3 \leq Y\} = g_2 > 0$, but $P_c(3)$ is not defined.

Thus Theorem 2 is not valid. You will verify that the asymptotic covariance matrix of $\hat{F}$ is not J^{-1} given in Theorem 3 with $k = 2$. We now establish the asymptotic normality directly. Note that

$$E(N_1/n) = P\{X = 1,\ (Z,Y) = (0.5,2) \text{ or } (2,4)\} = p_1(g_1 + g_2)$$
$$E(N_2/n) = P\{X = 5,\ (Z,Y) = (4,8) \text{ or } (2,4)\} = p_2(g_2 + g_3)$$
$$E(N_3/n) = P\{X = 5,\ (Z,Y) = (0.5,2)\} = p_2 g_1$$
$$E(N_4/n) = P\{X = 1,\ (Z,Y) = (4,8)\} = p_1 g_3.$$

We now derive the asymptotic covariance matrix of $(\hat{F}(1), \hat{F}(4))$. By SLLN,

$$U_n \to \frac{p_1 g_3}{p_1 g_3 + p_2(g_2 + g_3)} - \frac{p_1(g_1 + g_2)}{p_1(g_1 + g_2) + p_2 g_1},$$

which < 0 as $p_1 = p_2$ and $g_1 = g_2 = g_3$. Thus for n large enough, there is only one SCE and one GMLE. That is H_1. It is easy to show by SLLN or Theorem 1 that H_1 is consistent. It is also easy to show that

$$\frac{H_1(t) - F_o(t)}{\sigma_{H_1(t)}} \xrightarrow{D} N(0,1), \quad t \in [1,5), \tag{5.1}$$

by the corollary of Slutsky's theorem. Here

$$\sigma^2_{H_1(t)} = Var(\sum_{i=1}^{n} \mathbf{1}_{(X_i=1)}/n) = p_1 p_2/n,\ t \in [1,5). \tag{5.2}$$

$$\mathcal{L} = N_1 \ln F(1) + N_2 \ln(1 - F(4)) + N_3 \ln(1 - F(1)) + N_4 \ln F(4), \text{ where } F(5) = 1.$$

Verify the Fisher information matrix is

$$J/n = -\frac{1}{n}\frac{\partial^2 \mathcal{L}}{\partial F(1) \partial F(4)} = \begin{pmatrix} \frac{N_1}{nF^2(1)} + \frac{N_3}{n(1-F(1))^2} & 0 \\ 0 & \frac{N_4}{nF^2(4)} + \frac{N_2}{n(1-F(4))^2} \end{pmatrix}$$
$$\to \begin{pmatrix} \frac{g_1+g_2}{p_1} + \frac{g_1}{p_2} & 0 \\ 0 & \frac{g_3}{p_1} + \frac{g_2+g_3}{p_2} \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

If $(E(J))^{-1}$ is the covariance matrix of $(H_1(1), H_1(4))$, it contradicts (5.2).

§4.5.2. Homework

1. Let (Z,Y) takes values $(0.5,2)$, $(2,4)$ and $(4,8)$, with positive probabilities g_1, g_2 and g_3, respectively, where $g_1 = g_2 = g_3 = 1/3$, and let $F_o(x) = p_1 \mathbf{1}_{(x \geq 1)} + p_2 \mathbf{1}_{(x \geq 3)} + p_3 \mathbf{1}_{(x \geq 5)}$, where $p_1 = p_2 = p_3 = 1/3$. In the following, you may assume n is sufficiently large.
   1.a. Derive all possible SCEs of F_o based on Eq. (4.2) in §4.4.
   1.b. Find the limits of the SCEs directly.
   1.c. Derive the GMLE of F_o.
   1.d. Are the SCEs consistent and asymptotic normal ? (Prove or disprove them).
   1.e. Derive the asymptotic covariance matrix of the SCEs $(\hat{F}(1), \hat{F}(3))$.
2. Consider Example 1.
   2.a. What are the limits of H_1 and H_2 ? (Note that H_i is a function of both sample size n and $t \in R^1$.
   2.b. Are they consistent estimators of F_o ?
   2.c. Check whether AS1, AS2 and AS3. hold. AS3 can be interpreted as

$$\sum_{i:\ u_i \in (\tau_l, \tau_r)} \frac{f_Y(u_i) + f_Z(u_i)}{S_Y(u_i) - S_Z(u_i)} < \infty.$$

   2.d. Compute the Fisher information matrix $-H = -E\left(\frac{\partial^2 \mathcal{L}}{\partial F(b_i) \partial F(b_j)}\big|_{F=F_o}\right)_{2\times 2}$, where $(b_1, b_2) = (1,4)$ and $\mathcal{L} = \ln \text{Ł}$.

2.e. Compute the covariance matrix of the vector $(H_1(1), H_1(4))$, denoted by Σ. Does $\Sigma = -H^{-1}$? Does it contradicts Theorem 3 ?

3. Prove that under the DC model, $H = F_o$ satisfies the self-consistent equation

$$H(x) = \int_{l \le x < r} \frac{H(x) - H(l)}{H(r) - H(l)} dQ(l, r) + P\{R \le x\}, \ H \in \Theta.$$

References

* Chang, M. N. (1990). Weak convergence of a self-consistent estimator of the survival function with doubly censored data. *Ann. Statist.*, 18, 391-404.
* Chang, M.N. and Yang, G. (1987). Strong consistency of a self-consistent estimator of the survival function with doubly censored data. *Ann. Statist.*, 15, 1536-1547.
* Gu, M.G. and Zhang, C-H. (1993). Asymptotic properties of self-consistent estimator based on doubly censored data. *Ann. Statist.*, 21, 611-624.
* Tsai, W. and Crowley, J. (1985). A large sample study of the generalized maximum likelihood estimators from incomplete data via self-consistency. *Ann. Statist.*, 13, 1317-1334.
* Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *JASA.* 69, 169-173.
* Yu, Q.Q. and Li, L.X. (1994). On the strong consistency of the product limit estimator. *Sankhya*, A. 56, 416-430.
* Yu, Q. Q. and Li, L.X. (2001). Asymptotic properties of the GMLE of self-consistent estimators with doubly-censored data. *Acta Math. Sinica*. 17, 581-594.

§**4.6. SCE under MIC model** Consider MIC model (2). Assume:

(1) N is a random integer;
(2) T, $Y_1 < Y_2 < \cdots < Y_k < \cdots$ are inspection times;
(3) X and $(N, T, Y_i, i \ge 1\}$ are independent;
(4) $\mathrm{P}(N = 0) > 0$ and $\mathrm{P}(N > 1) > 0$;
(5) The observable random vector is

$$(L, R) = \begin{cases} (X, X) & \text{if } X \le T \text{ and } N = 0 \\ (T, \infty) & \text{if } X > T \text{ and } N = 0 \\ (-\infty, Y_1) & \text{if } X \le Y_1 \text{ and } N \ge 1 \\ (Y_i, Y_{i+1}) & \text{if } Y_i < X \le Y_{i+1},\ i = 1, ..., N-1 \text{ and } N \ge 1 \\ (Y_N, \infty) & \text{if } X > Y_N \text{ and } N \ge 1. \end{cases}$$

Let (L_i, R_i), $i = 1, ..., n$, be an i.i.d. copies of (L, R).

Remark If one replaces (4) by $N = 0$ or 1 w.p.1, it can be used to formulate the DC data.

Hereafter, denote H_n an SCE of F_o. Define

$$\tau = \sup\{x : \ F_o(x) < 1 \text{ and } F_T(x) < 1\},$$

$$\tau_Y = \sup\{t : \ F_{Y_N}(t) < 1\}.$$

$$\tau_N = \sup\{i : \ f_N(i) > 0\}.$$

A point x is called *a support point of a function* F if $|F(y_j) - F(x)| > 0$ for a sequence of points $y_j \to x$. Denote $\mathcal{S}_F$ the set of all support points of F. We need the following assumptions:

(AS1) $\tau_Y \le \tau$;
(AS2) $\mathrm{P}\{T \text{ or } Y_N = \tau\} > 0$ if $F_o(\tau-) < 1$.
(AS3) H_n is right continuous and $\mathcal{S}_{H_n} \subset \{R_1, ..., R_n\}$;
(AS4) $F_o(\tau) > 0$ and $\cup_{i \le \tau_N} \mathcal{S}_{F_{Y_i}} \subset \mathcal{S}_{F_o}$;
(AS5) $\cup_{i \le \tau_N} \mathcal{S}_{F_{Y_i}}$ is a finite set and F_o is strictly increasing on $\cup_{i \le \tau_N} \mathcal{S}_{F_{Y_i}}$.
(AS6) There are at most m IIs for each sample size n and $\mu_{F_o}(A_j) > 0$ for each II A_j.

Theorem 1 (Yu, Li and Wong (1998,2000)) *Suppose that AS1 and AS2 hold. Then the SCE H_n satisfies*

$$\lim_{n\to\infty}\sup_x|H_n(x)-F_o(x)|=0 \text{ a.s. if } F_o(\tau)=1 \text{ and } \lim_{n\to\infty}\sup_{x\le\tau}|H_n(x)-F_o(x)|=0 \text{ a.s..}$$

Theorem 2 (Yu, Li and Wong (1998,2000)) *Suppose that AS1 and AS2 hold. Moreover, either AS5 holds or AS6 holds or AS3 and AS4 hold. Then*

$$\sqrt{n}(H_n(x)-F_o(x))\xrightarrow{D}Z(x) \text{ for } x\le\tau,$$

where Z is a Gaussian process on $[0,\tau]$. Let $b_1<\cdots<b_{k+1}$ be the right endpoints of all the distinct IIs induced by the observed intervals. The covariance matrix of $(H_n(b_1),...,H_n(b_k))$ can be estimated by $-\left(\frac{\partial^2\mathcal{L}}{\partial F(b_i)\partial F(b_j)}\right)^{-1}_{k\times k}\Big|_{F=\hat{F}}$ and $\mathcal{L}=\ln Ł$.

Remark. The assumptions AS1, AS4, AS5 and AS6 eliminate the case in which $\mu_{F_o}(A_j)=0$ for an II A_j. The following case is a situation when A4-A6 are violated.

X= 3,6, N= 0, 1, T= 8, Y_1 = 4,5,

Then an II is $(4,5]$ and $\mu_{F_o}((4,5])=0$.

Remark. For IC data, $-\ln Ł$ is strictly convex in the interior of $\mathbf{s}\in D_s$, where

$D_s=\{\mathbf{s}:\ \vec{s}=(s_2,...,s_m), s_i\ge 0, \sum_{i=2}^m s_i\le 1\}$, $s_1=1-s_2-\cdots-s_m$.

Thus there is a unique GMLE of $\mathbf{s}$. Moreover, if $\hat{\mathbf{s}}$ is the GMLE and $\tilde{\mathbf{s}}\ne\hat{\mathbf{s}}$ is an SCE, then

(1) $\hat{s}_j=0$ implies $\tilde{s}_j=0$ for each j;

(2) there is at least one j such that $\tilde{s}_j=0<\hat{s}_j$.

A function g is strictly convex if $g(ax+(1-a)y)<ag(x)+(1-a)g(y)$ for each possible pair of (x,y) and $a\in(0,1)$. Since $\sum_{j=1}^m\delta_{ij}s_j$ is linear in $\mathbf{s}\in D_s$ and $-\ln x$ is strictly convex in $x\in R^1$, as $\frac{d^2(-\ln x)}{dx^2}=\frac{1}{x^2}>0$, $-\ln Ł$ is strictly convex.

§4.6.2. Homework

1. Consider the MIC model (2). Suppose that

   X takes values 1, 3 and 5 w.p.1.;

   N takes values 0, 1 and 2 w.p.1.;

   $T=7$,

   Y_1 takes values 2 and 4 w.p.1;

   $Y_2=Y_1+2$.

   1.a. Describe all possible observed intervals.

   1.b. Derive all possible SCEs of F_o. (**Hint**: If you properly group the observations, the estimators are similar to H_1 and H_2 in Example 1 or Homework problem 1 of §4.5.)

§4.7. Case 2 data and the GMLE

The mixed case model is used to formulate IC data which do not involve exact observations. Let $\mathbf{Y}=\{Y_j: j\ge 1\}$ be a sequence of random variables such that $Y_1<Y_2<\cdots$ and let K be a positive random integer such that $(K,\mathbf{Y})$ and X are independent. On the event $\{K=k\}$, let (L,R) denote the endpoints of that random interval among $(-\infty,Y_1],(Y_1,Y_2],\ldots,(Y_k,\infty)$ which contains X.

Let ν be the measure that is the sum of the two measures induced by the marginal distributions of observable extended random variables L and R.

Theorem 1. (Schick and Yu (2000)) *Suppose $E(K)<\infty$. Let $A=\cup_{i\le\tau_K}\mathcal{S}_{F_{Y_i}}$, where $\tau_K=\sup\{i:\ P\{K=i\}>0\}$, and $\hat{F}$ be a GMLE of F_o. Then*

1. *$\int|\hat{F}-F_o|\,d\nu\to 0$ almost surely;*
2. *If $P(Y_i=a)>0$ for some $i\le\tau_K$, then $\hat{F}(a)\to F_o(a)$ a.s.;*
3. *If A is dense in $[0,\infty)$ and F_o is continuous, then $\sup_x|\hat{F}(x)-F_o(x)|\to 0$ a.s..*

Question: Why not SCE ?

With case 2 IC data, the SCE may not be consistent even at discrete inspection times. See the following example:

Example 1. Let X take on values 2, 4, 6 with probability 1/3 for each value. Let $K = 2$ w.p.1 and (Y_1, Y_2) take on values (1,5) and (3,7) with probabilities 1/2 and 1/2, respectively. If n is large enough, the observations are

N_1 $(-\infty, 3]$, N_2 $(5, \infty)$, N_3 (3,7] and N_4 (1,5].

IIs: (1,3], (3,5] and (5,7]

As in Example 1 of §4.5, there are two solutions to equation (4.2): They induce two SCEs H_1 and $\hat{F}$ exactly the same as in Example 1 of §4.5. Note that under the assumption there $\hat{F} = H_1$ if n is large enough. However, in the current situation, $\hat{F} \neq H_1$ if n is large enough, as pointed out next.

Note Y_i's are discrete, thus we expect a GMLE is consistent at 1, 3, 5, 7.

The limit of $H_1(3)$:

$$\lim_{n\to\infty} \frac{N_1 + N_4}{n} = P\{X = 2, (Y_1, Y_2) = (3, 7)\} + P\{X \leq 4, (Y_1, Y_2) = (1, 5)\} = 1/2 \neq 1/3.$$

Thus H_1 is not consistent at 3.

Theorem 2. (Yu, Schick, Li and Wong (1998)) *Under the mixed case model, if there are only $k+1$ II's with right endpoints b_i for each sample size, F_o is strictly increasing on b_i's, then*

$$\sqrt{n} \begin{pmatrix} \hat{F}(b_1) - F_o(b_1) \\ \cdots \\ \hat{F}(b_k) - F_o(b_k)) \end{pmatrix} \xrightarrow{\mathcal{D}} N(0_k, \Sigma_k)$$

as $n \to \infty$, where

$$\Sigma_k = -n \left(E(\frac{\partial^2 \mathcal{L}}{\partial F(b_i) \partial F(b_j)}\big|_{F=F_o}) \right)^{-1}_{k\times k}$$

and the asymptotic covariance matrix of $\hat{F}$ can be estimated by $-\left(\frac{\partial^2 \mathcal{L}}{\partial F(b_i)\partial F(b_j)}\right)^{-1}_{k\times k}\big|_{F=\hat{F}}$.

The convergence rate for the GMLE under the assumption of Theorem 2 is $n^{1/2}$. However, the convergence rate of the GMLE is conjectured to be $(n \log n)^{1/3}$ under another set of continuity assumptions given by Groeneboom and Wellner (1992), and is proved to be $n^{1/3}$ under a set of continuity assumptions given by Groeneboom (1996).

Under the case 2 model (i.e., $K = 2$ w.p.1), Groeneboom and Wellner (1992) establish the following asymptotic result.

Suppose that F_o and G have continuous derivatives, with their densities $f_o(x_o) > 0$ and $g(x_o, x_o) > 0$, and let $\tilde{F}$ be the estimator of F_o, obtained at the first step of the iterative convex minorant algorithm, starting the iterations with F_o. Then the statistic $(n\ln n)^{1/3} \frac{\tilde{F}(x_0) - F_o(x_0)}{\{\frac{3}{4}(f_o(x_0))^2 / g(x_0, x_0)\}^{1/3}} \xrightarrow{\mathcal{D}} 2Z^*$, where Z^* is the last time where standard two - sided Brownian motion minus the parabola $y(t) = t^2$ reaches its maximum.

Conjecture (G&W (1992, p. 108)): The GMLE $\hat{F}$ has the same asymptotic distribution as $\tilde{F}$. Thus, the convergence rate of the GMLE is conjectured to be $(n \log n)^{1/3}$ under the same conditions mentioned above.

Remark. $\tilde{F}$ is not an estimator, because F_o is unknown and thus it is impossible to start the iterations from F_o except in simulation.

Theorem 3. (Groeneboom (1996)). *Let F_o be continuous with a bounded derivative f_o on $[0, M]$, satisfying $f_o(x) \geq c_o > 0$, $x \in (0, M)$, for some constant $c_o > 0$. Let (Y_1, Y_2) be the two continuous random inspection times in the Case 2 model, with df $g(\cdot, \cdot)$. Let g_1 and g_2 be the first and second marginal density of g, respectively. Suppose that the following conditions are satisfied*

(S1) g_1 and g_2 are continuous, with $g_1(x) + g_2(x) > 0 \ \forall\ x \in [0, M]$;

(S2) $g(\cdot, \cdot)$ is continuous, with uniformly bounded partial derivatives, except at a finite number of points, where left and right (partial) derivatives exist;

(S3) $P\{Y_2 - Y_1 < \epsilon_o\} = 0$ for some ϵ_o with $0 < \epsilon_o \leq 1/2M$, so g does not have mass close to the diagonal.

Then we have at each point $t_o \in (0, M)$

$$n^{1/3}\{2a(t_o)/f_o(t_o)\}^{1/3}\{\hat{F}(t_o) - F_o(t_o)\} \xrightarrow{\mathcal{D}} 2Z^*,$$

where Z^* is defined as in Theorem 2, and

$$a(t_o) = \frac{g_1(t_o)}{F_o(t_o)} + k_1(t_o) + k_2(t_o) + \frac{g_2(t_o)}{1 - F_o(t_o)},$$

$$k_1(u) = \int_u^M \frac{g(u,v)}{F_o(v) - F_o(u)} dv \text{ and } k_2(v) = \int_0^v \frac{g(u,v)}{F_o(v) - F_o(u)} du.$$

Remark. The main differences between the assumptions in the above theorems that the convergence rate varies are as follows.

1. The main assumption in Theorem 2 is that K is finite and Y_i, $i = 1, ..., K$, takes on finitely many values. Then the convergence rate is $n^{1/2}$.
2. The main assumption in the conjecture of G&W (1992) is that $\mathrm{P}\{(Y_1, Y_2) \in N(x_o, x_o)\} > 0$ for each neighborhood $N(x_o, x_o)$ of (x_o, x_o), in addition to smoothness. Then the convergence rate is $(n \ln n)^{1/3}$.
3. The main assumption in Theorem 3 is that (Y_1, Y_2) does not fall along a strip near the diagonal $y_1 = y_2$, in addition to smoothness. Then the convergence rate is $n^{1/3}$.

§**4.7.2. Homework**

1. a. Derive the limits of the SCEs in Example 1 and compare to F_o.
   b. Derive the asymptotic variance of the GMLE in Example 1.
   c. Give an estimate of the variance in part b.
2. Suppose that $F \sim Exp(\rho)$, $K = 2$ w.p.1., $Y_1 \sim Exp(\rho)$ and $Y_2 = Y_1 + Z$, where $Z \sim Exp(\rho)$ and $Z \perp Y_1$. let S_n^2 be the sample variance of the GMLE $\hat{F}(2)$ based on 1000 simulations of random samples of size n. What do you expect S_{100}^2/S_{400}^2 to be ? State your reasoning.
3. Suppose that $F \sim Exp(\rho)$, $K = 2$ w.p.1., Y_1 has a discrete uniform distribution on the set $\{1, 2, ..., 9\}$ and and $Y_2 = Y_1 + 1$. Let S_n^2 be the sample variance of the GMLE $\hat{F}(2)$ based on 1000 simulations of random samples of size n. What do you expect S_{100}^2/S_{400}^2 to be ? State your reasoning.

Reference.

* Groeneboom, P. and Wellner, J. A. (1992). Information bounds and nonparametric maximum likelihood estimation. *Birkhäuser Verlag, Basel.*
* Groeneboom, P. (1996). Lecture on inverse problems. In P. Bernard. (Ed.), *Lectures on probability and statistics*. p. 157. Berlin, Springer-Verlag.
* Schick, A. and Yu, Q.Q. (2000). Consistency of the GMLE with mixed case interval-censored data. *Scand. J. Statist.*, 27, 45-55.
* Yu, Q. Q., Schick, A., Li, L. X. and Wong, G. Y. C. (1998). Asymptotic properties of the GMLE in the case 1 interval-censorship model with discrete inspection times. *Canadian Journal of Statistics*, 26, No. 4, 619-627.

§**4.8. Application to model justification.**

It is desirable to use parametric models instead of nonparametric models. However, one needs to justify why a certain type of model can be applied to the given data, either from science (physics, biology, etc.) or from empirical data.

A naive method is to plot the empirical cdf against the targetting parametric cdf, replacing the parameters by their MLEs. We shall introduce several other methods here.

§**4.8.1. Q-Q plot**

For a complete data set

X_1, ..., X_n from a cdf F_o.

If we suspect F_o belongs to a parametric distribution family, say

H_0: $F_o = F(\cdot, \theta)$,

we can use probability plot to check.

First estimate the parameter θ by $\hat{\theta}$ (MLE etc.),

Plot of sample quantiles v.s. quantiles of $F(\cdot, \hat{\theta})$ (Q-Q plot).

Let $X_{(1)} \le \cdots \le X_{(n)}$ be order statistics;

They are $100\frac{1}{n+1}$, ..., $100\frac{n}{n+1}$-th sample percentiles (quantiles) of the sample.

Let $y_i = \sup\{u : \ F(u, \hat{\theta})) < \frac{i}{n+1}\}$, $i = 1, ..., n$.
— estimated population percentiles (quantiles).
Plot $(y_i, X_{(i)})$, $i = 1, ..., n$.
If the plot is close to the line $y = x$, then $F \approx F_o$.
qqplot() is a function in Splus.
It only needs the sample and F_o.
Recall the the $100p$-th quantile of a cdf F is a

$$q = F^{-1}(p), \text{ where } F^{-1}(p) = \sup\{q : \ F(q) < p\}.$$

We first explain the reasoning of QQ-plot with complete data.

Complete data. Let $a_1 < \cdots < a_m$ be distinct points among $X_1, ..., X_n$. Since the edf $\hat{F} \to F_o$ a.s. on $(0, \infty)$, we expect the quantile functions $(\hat{F})^{-1}$ and F_o^{-1} are close. These $(a_i, F_o^{-1}(\hat{F}(a_i)))$, $i = 1, ..., m$ are around the line $y = x$ or at least around $y = \sigma x + \mu$ if $\sigma \neq 1$ or $\mu \neq 0$ under H_0.

The following are two QQ-plot graphs. The first one is 100 observations from Exp(1) plotting against normal distribution using qqnorm(). The second is a QQ-plot of 100 observations from normal distribution plotting against normal distribution.

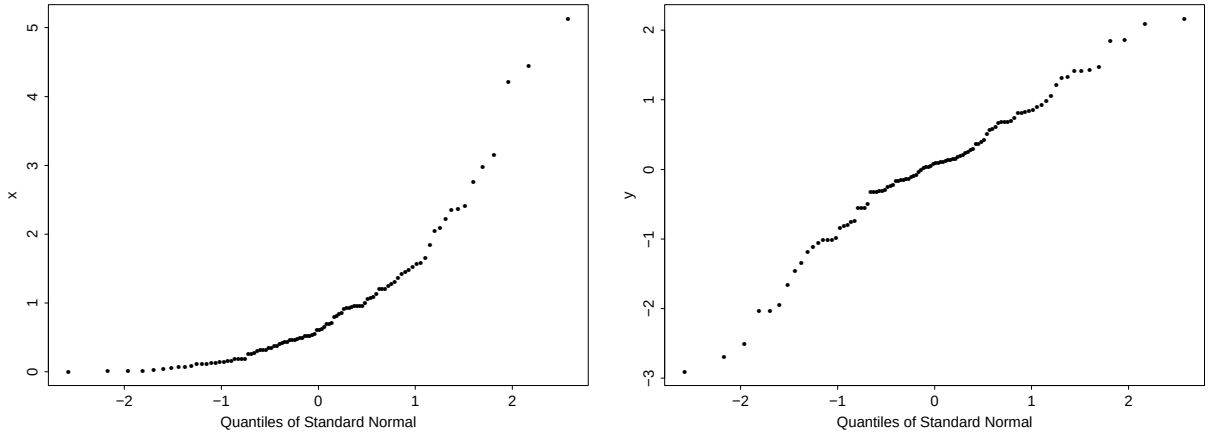


Fig. 1. QQ-plot

For RC or IC data Suppose that $X \sim F_o$ and (L_i, R_i)'s are iid IC observations. If we suspect that F_o is some of given form, say $F_o = F_*(\cdot, \theta)$, we can also do QQ-plot as follows.

1. Obtain the IIs based on observations (L_i, R_i), $i = 1, ..., n$, denoted by A_j, $j = 1, ..., m$.
2. Obtain the NPMLE of the distribution function F_o, denoted by $\hat{F}_*$. The GMLE $\hat{F}_*$ is redefined to be linear on each non-singleton A_j rather than a right continuous step function with jumps only at the right endpoints of the A_j's, provided the A_j is a finite set.
3. Obtain the MLE of the parameter θ, say $\hat{\theta}$, in F_*.
4. Denote the midpoint of the finite A_j's by m_i's. Plot $(m_i, F_*^{-1}(\hat{F}(m_i), \hat{\theta}))$'s for all i that A_i is finite. which ideally should also be around $y = x$ or $y = \sigma x + \mu$, in the latter case, it suggests that $F_o(t) = F_*(\frac{t-\mu}{\sigma}, \theta)$.

§**4.8.2. Hazard plot for RC data**

(Z_i, δ_i), $i = 1, ..., n$.

Plot sample integrated hazard v.s. integrated hazards of $F_o(\cdot, \hat{\theta})$.

To get sample integrated hazards:

a. Order Z_i's as $Z_{[n]} \leq \cdots \leq Z_{[1]}$
(reverse order);

b. Denote $\delta_{[i]}$, $i = 1, ..., n$, correspondingly;

c. For each $\delta_{[k]} = 0$ (RC ones), the sample hazard is 0;

d. For each $\delta_{[k]} = 1$ (exact ones),
the sample hazard $\hat{h}(Z_{[k]})$ is $1/k$;

e. The sample integrated hazard at $Z_{[k]}$ is

$$\hat{H}(Z_{[k]}) = - \sum_{i:\ Z_{[i]} \le Z_{[k]}, \delta_{[i]}=1} \log[1 - \hat{h}(Z_{[i]})].$$

f. Plot $(\hat{H}(Z_{[k]}), H_o(Z_{[k]}))$, where $\delta_{[k]} = 1$ and $H_o(x) = -\log S_o(x, \hat{\theta}))$.

Example. Hazard plot. Use Sample 0 in Table 1.1 v.s. exponential distribution.

Table 1.2. Calculation of sample hazards

Patient Number	Remission Time	Reverse Order (K)	$h(x)$ $= f(x)/S(x-)$	Integrated Hazard
2	6	21	1/21	0.15
3	6	20	1/20	0.15
4	6	19	1/19	0.15
1	6+	18		
5	7	17	1/17	0.21
6	9+	16		
8	10	15	1/15	0.28
7	10+	14		
9	11+	13		
10	13	12	1/12	0.37
11	16	11	1/11	0.47
12	17+	10		
13	19+	9		
14	20+	8		
15	22	7	1/7	0.62
16	23	6	1/6	0.80
17	25+	5		
18	32+	4		
19	32+	3		
20	34+	2		
21	35+	1		

It is easy to see that $6+ > 6$.

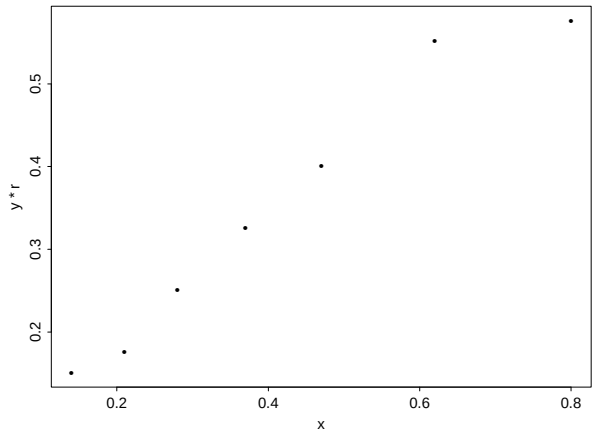


Fig. 2. (Left) : Hazards plot.

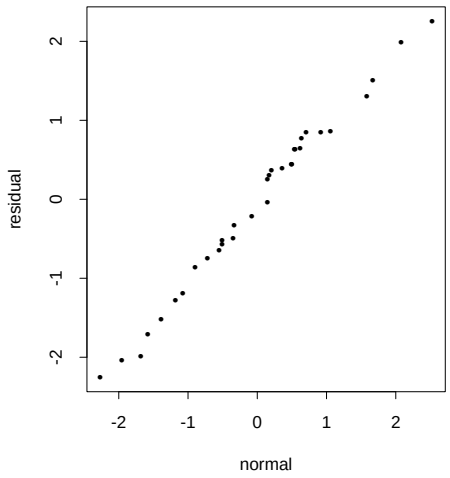


Fig 3. (Right): Q-Q plot with IC data

Note that the $\hat{H}(6)$ appears three times as there are three exact observations at $x = 6$. However, they are obviously the same. Thus we plot

$(0.15, 6\rho)$, $(0.21, 7\rho)$, $(0.28, 10\rho)$, $(0.37, 13\rho)$, $(0.47, 16\rho)$, $(0.62, 22\rho)$, $(0.80, 23\rho)$,

where $\rho = \sum_{i=1}^{n} \delta_i / \sum_{i=1}^{n} Z_i$.

We explain the reasoning of Hazard plot with RC data.

Let $a_1 < \cdots < a_m$ be distinct point among all exact observations. Since $\sum_{t<\tau} \hat{F}_{pl} \to F_*$ a.s., we expect sample cumulative hazard $\hat{H} \to H_o$ a.s. (population cumulative hazard) a.s., thus $(\hat{H}(a_i), H_o(a_i)$'s should be around the line $y = x$ if we know F_o completely.

§4.8.3. Diagnostic plot under regression set-up with IC data

In parametric analysis, we often assume that H_0: $F_o(t) = F_*(\frac{t-\mu}{\sigma})$ $\forall$ t, where F_* is known but the location-scale parameters μ and σ may or may not be known.

In general, we can also use the following QQ-plot to check whether the IC data satisfy a certain parametric distribution F_*. We shall consider this in regression set-up. We first explain this method via the following example.

Example 1. If we suspect that conditional $\mathbf{Z}$, $\ln X \sim N(\beta^t \mathbf{Z}, \sigma^2)$, where β is a $p \times 1$ parameter and X is interval censored, we can use the following procedure to check the assumption, based on our observations $(L_i, R_i, \mathbf{z}_i)$, $i = 1, ..., n$.

1. Compute the MLE of β, say $\hat{\beta}$, based on observations $(L_i, R_i, \mathbf{z}_i)$.
2. Obtain the IIs based on $(\ln L_i - \hat{\beta}^t \mathbf{z}_i, \ln R_i - \hat{\beta}^t \mathbf{z}_i)$, $i = 1, ..., n$, denoted by A_j, $j = 1, ..., m$.
3. Obtain the NPMLE of the distribution function based on $(\ln L_i - \hat{\beta}^t \mathbf{z}_i, \ln R_i - \hat{\beta}^t \mathbf{z}_i)$, $i = 1, ..., n$, denoted by $\hat{F}_*$. The GMLE $\hat{F}_*$ is redefined to be linear on each non-singleton A_j rather than a right continuous step function with jumps only at the right endpoints of the A_j's, provided the A_j is a finite set.
4. Plot $(m_i, F_*^{-1}(\hat{F}(m_i)))$ for all possible i.

If the assumption is correct, we expect a roughly linear plot.

The justification for the method is as follows. $\epsilon = \ln X - \beta^t \mathbf{Z}$ conditional on $\mathbf{Z} = \mathbf{z}$ is normal distribution $N(0, \sigma^2)$. Since X_i is interval censored by the random interval $(L_i, R_i]$, ϵ_i $(= \ln X_i - \beta^t \mathbf{z}_i)$ is interval censored by the random interval $(\ln L_i - \beta^t \mathbf{z}_i, \ln R_i - \beta^t \mathbf{z}_i]$, which can be estimated by $(\ln L_i - \hat{\beta}^t \mathbf{z}_i, \ln R_i - \hat{\beta}^t \mathbf{z}_i]$. Since the MLE of β and the NPMLE of F are consistent under a certain regularity assumptions (see §3 and 4), it should be approximately normally distributed.

Fig 3. is a result from a simulation study. We assume that conditional on $\mathbf{Z} = \mathbf{z}$, $\ln X$ has a normal distribution $(N(\beta^t \mathbf{z}, 1))$, where $\beta = (1, 1, 1)$. $\mathbf{Z} = (Z_1, Z_2, Z_3)$, and Z_1, Z_2 and Z_3 equal 0 and ± 1 with a certain probabilities. X is under a mixed case interval censorship model. The number of follow-up times K is a discrete uniform distribution on $\{1, 2, ..., 28\}$. Conditional on $K = k$, the follow-up time Y_i, $i = 1, ...$, k, satisfy $\ln Y_i = -5 + \sum_{i=1}^{k} U_i$, where U_i are i.i.d. from uniform distribution $U(0, 1)$. Here $n = 374$. The MLE of (β, σ) is $(0.92, 1.02, 1.02, 1.06)$. The resulting QQ-plot supports the normal regression model as we expected.

The method can be viewed as a pivotal method by choosing a function T of $(X, \mathbf{Z}, \beta)$ so that $T = T(X, \mathbf{Z}, \beta)$ has a distribution function F_T which does not depend on $\mathbf{Z}$ and T is strictly increasing in X. For lognormal,

$$T = \log X - \beta^t \mathbf{Z}.$$

In the above cases, $T \sim N(0, \sigma^2)$.

In general, the procedure is as follows:

1. Find a pivotal function T described as above.
2. Find the MLE $\hat{\beta}$ of parameter β.
3. Obtain the IIs based on $(T(L_i, \mathbf{z}_i, \hat{\beta}), T(R_i, \mathbf{z}_i, \hat{\beta}))$, $i = 1, ..., n$, denoted by A_j, $j = 1, ..., m$.
4. Obtain the GMLE of the distribution function based on $(T(L_i, \mathbf{z}_i, \hat{\beta}), T(R_i, \mathbf{z}_i, \hat{\beta}))$, $i = 1, ..., n$, denoted by $\hat{F}_*$. The GMLE $\hat{F}_*$ is redefined to be linear on each non-singleton A_j, provided the A_j is a finite set.
5. Plot $(m_i, F_*^{-1}(\hat{F} * (m_i)))$, for all possible i.

If the assumption is correct, we expect a roughly linear plot.

Example 2 Suppose that conditional on $\mathbf{Z} = \mathbf{z}$, X has a Weibull distribution. That is,

$$S_X(x) = e^{-x^\kappa e^{\beta^t \mathbf{Z}}}, \ x > 0.$$

One of such pivoting functions is

$$T = X^{\kappa} e^{\beta^t \mathbf{Z}},$$

where T has an Exponential distribution with survival function $S_T(t) = e^{-t}$, $t > 0$. If X_i is interval censored by $(L_i, R_i]$, then T_i is interval censored by $(T(L_i, \mathbf{z}_i, (\beta, \kappa)),\ T(R_i, \mathbf{z}_i, (\beta, \kappa))] = (L_i^{\kappa} e^{\beta^t \mathbf{z}_i}, R_i^{\kappa} e^{\beta^t \mathbf{z}_i}]$. Since (β, κ) are unknown and neither are $L_{i,\beta,\kappa}$ and $R_{i,\beta,\kappa}$, we replace β and κ by their MLE. Let $\hat{F}_T$ be the NPMLE of F_T based on the pivoted data $(T(L_i, \mathbf{z}_i, (\hat{\beta}, \hat{\kappa})), T(R_i, \mathbf{z}_i, (\hat{\beta}, \hat{\kappa}))]$, $i = 1, ..., n$. Finally we plot $\hat{F}_*$ against F_T, the exponential distribution Exp(1).

§**4.8.3.2. Homework**:

In the following, make comments on whether the plots suggest $F_o(x) = F_*(\frac{x-\mu}{\sigma})$.

1. The Weibull distribution in Example 2 can also be re-parametrized as a location-scale parameter family. Find the pivot function T and derive the distribution function F_T. What is the revision of the procedure for a diagnostic plot in Example 2 ?
2. Do a QQ-plot using sample 0 in Leukaemia data on page 2 v.s. exponential distribution Exp(1). Do you think the exponential distribution is appropriate ? If so, what is your guess of ρ according to the slope of fitting straight line ?
3. Q-Q plot:
   3.a. Use sample 1 in Leukaemia data on page 2 v.s. exponential distribution.
   3.b. Generated a random sample of size 100 from an exponential distribution v.s. exponential distribution, normal distribution.
4. Hazard plot. Generate a random sample of size 100 from a RC model, say
   $X \sim$ Weibull distribution (nontrivial one),
   $Y \sim$ Uniform distribution,
   Draw a hazard plot v.s. Weibull (that you used) and a hazard plot v.s. a normal distribution with the mean and variance equal those of the Weibull distribution you used.

Chapter 5. Semi-parametric Analysis

§5.1. Introduction

Suppose

X is a random survival time,
$\mathbf{Z}$ is a $p \times 1$ covariate (explanatory) (random) vector
(which sometimes is assumed to be nonrandom);
X is subject to interval censoring;
Observable random vector is now $(L, R, \mathbf{Z})$.

Example 1. Cancer research. In addition to observe the failure time of a patient, we also observe $\mathbf{Z}^t = (Z_1, Z_2, Z_3, Z_4)$, where

Z_1 — # of relatives who had cancer;
Z_2 — age of the patient;
Z_3 — tumor size;
Z_4 — smoking habit.

Example 2. Two-sample problem. There are two independent samples,

$\mathbf{Z} = \mathbf{1}_{(\text{patient is from sample 1})}$.

Two models are considered.

1. Proportional hazards (PH) model: Conditional on $\mathbf{Z} = \mathbf{z}$, the hazard function

$$h(t|\mathbf{z}) = \psi(\beta, \mathbf{z}) h_o(t),\ t < \tau, \tag{1.1}$$

where $\tau = \sup\{t : S_o(t) > 0\}$, h_o is a (baseline) hazard function, and ψ is a function of $(\beta, \mathbf{z})$.

If S_o is a survival function of a continuous random variable, then

$$S_{X|\mathbf{Z}}(t|\mathbf{z}) = (S_o(t))^{\psi(\beta, \mathbf{Z})} \quad (H_{X|\mathbf{Z}}(t|\mathbf{z}) = \psi(\beta, \mathbf{z}) H_o(t)) \tag{1.2}$$

where S_o is a (baseline) survival function.

The PH model is also called the *Cox regression model*.

For an arbitrary random variable,
Eq. (1.1) defines a PH family, and
Eq. (1.2) defines a Lehmann family or proportional integrated hazards family.
If S_o is continuous, then equations (1.1) and (1.2) are the same.
2. Accelerated lifetime model: Conditional on $\mathbf{Z} = \mathbf{z}$,

$$X = X_o/\psi(\beta, \mathbf{z}) \ \ (\ln X = \ln X_o - \ln\psi(\beta, \mathbf{z})). \tag{1.3}$$

$$S_{X|\mathbf{Z}}(t|\mathbf{z}) = S_o(\psi(\beta, \mathbf{z})t).$$

Under either model, it is a parametric problem, if S_o is known or is of a parametric form; otherwise, it is a semi-parametric one (*e.g.*, S_o is an arbitrary survival function).
Most usual forms of ψ: ($\psi \geq 0$)
$e^{\beta^t \mathbf{Z}}$ ——- log linear;
$log(1 + e^{\beta^t \mathbf{Z}})$ ——- logistic.

§5.2. PH model with RC data.

§5.2.1. Continuous RC data.

Assume:
Conditional on $\mathbf{Z} = \mathbf{z}$, $X \sim F(\cdot|\mathbf{z})$ with its hazard satisfies (1.1);
$Y \sim G$;
$(X, \mathbf{Z})$ and Y are independent;
F and G are absolutely continuous;
Observe $(M, \delta, \mathbf{Z}) = (\min\{X, Y\}, \mathbf{1}_{(X \leq Y)}, \mathbf{Z})$.
Let $(M_i, \delta_i, \mathbf{z}_i)$, $i = 1, ..., n$ be i.i.d. copies of $(M, \delta, \mathbf{Z})$.
The log likelihood function is

$$\begin{aligned}
\mathcal{L}(\beta) =& \ln \prod_{i=1}^{n} (f(M_i|\mathbf{z}_i))^{\mathbf{1}_{e,i}} (S(M_i|\mathbf{z}_i))^{\mathbf{1}_{r,i}} \\
=& \ln[\prod_{i:\ ex} h(M_i|\mathbf{z}_i) \prod_{i=1}^{n} S(M_i|\mathbf{z}_i)] \\
=& \sum_{i:\ ex} \ln\psi(\beta, \mathbf{z}_i) + \sum_{i:\ ex} \ln h_o(M_i) + \sum_{i=1}^{n} \psi(\beta, \mathbf{z}_i) \ln S_o(M_i).
\end{aligned}$$

This approach needs to estimate β and S_o in the same time. Cox (1972) uses a conditional probability approach and a partial likelihood approach with some assumptions to define a new likelihood function, which only involves β. We first give the likelihood functions and then introduce the derivation.
Notation:
$a_1 < \cdots < a_g$ —- all the distinct exact observations.
By rearranging the index, assume $X_i = a_i$, $i = 1, ..., g$.
$\mathcal{R}_j = \mathcal{R}(a_j) = \{i : \ M_i \geq a_j\}$,
$r_j = |\mathcal{R}(a_j)|$;
$\phi(i) = \psi(\beta, \mathbf{z}_i)$,
$\mathcal{D} = \{i : \ \delta_i = 1, \ i = 1, ..., n\}$ (note that all exact observations are distinct),
Define a modified likelihood function

$$lik = \prod_{i \in \mathcal{D}} \frac{\phi(i)}{\sum_{k \in \mathcal{R}_i} \phi(k)}. \tag{2.1}$$

The log likelihood

$$l(\beta) = \ln lik = \sum_{i \in \mathcal{D}} [\ln\phi(i) - \ln \sum_{k \in \mathcal{R}_i} \phi(k)]. \tag{2.2}$$

The SMLE $\hat{\beta}$ of β is a value of b that maximizes $l(b)$.

Remark. Under certain assumptions, the SMLE $\hat{\beta}$ is consistent and asymptotically normally distributed. Its covariance matrix can be estimated by $-\left(\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^t}|_{\beta=\hat{\beta}}\right)^{-1}$.

Hereafter to the end of §5.2, let

$$\psi(\beta, \mathbf{z}) = e^{\beta^t \mathbf{z}}. \tag{2.3}$$

Then (2.2) becomes

$$l(\beta) = \ln lik = \sum_{i\in\mathcal{D}} [\beta^t \mathbf{z}_i - \ln \sum_{k\in\mathcal{R}_i} e^{\beta^t \mathbf{z}_k}].$$

Example 1. 5 observations $(M_i, \delta_i, \mathbf{z}_i)$'s: $(2.5, 1, 2), (2, 0, 5), (4, 0, 1), (1, 1, 1), (7, 1, 2)$. lik= ?

Sol: Reorder $(M_i, \delta_i, \mathbf{z}_i)$'s as

$(1, 1, 1), (2, 0, 5), (2.5, 1, 2), (4, 0, 1), (7, 1, 2)$,.

Rearrange (according to exact obs.) $(M_i, \delta_i, \mathbf{z}_i)$'s as

$(1, 1, 1), (2.5, 1, 2), (7, 1, 2), (2, 0, 5), (4, 0, 1)$,.

We use this order from now on.

3 exact observations: $(a_1, a_2, a_3) = (1, 2.5, 7)$.

$\mathcal{D} = \{1, 2, 3\}$,

$\mathcal{R}_1 = \mathcal{R}(1) = \{1, 2, 3, 4, 5\}$, $\mathcal{R}_2 = \mathcal{R}(2.5) = \{2, 3, 5\}$, $\mathcal{R}_3 = \mathcal{R}(7) = \{3\}$;

$r_1 = 5$, $r_2 = 3$, $r_3 = 1$;

$\phi(i)$'s are e^{β}, $e^{2\beta}$, $e^{2\beta}$, $e^{5\beta}$, e^{β},

$$lik = \frac{e^{\beta}}{2e^{2\beta} + e^{5\beta} + 2e^{\beta}} \cdot \frac{e^{2\beta}}{2e^{2\beta} + e^{\beta}} \cdot \frac{e^{2\beta}}{e^{2\beta}} = \frac{1}{(2 + e^{-\beta})(2e^{\beta} + e^{4\beta} + 2)}.$$

§5.2.1.2. Homework

1. Derive the estimate of the covariance matrix of the SMLE $\hat{\beta}$ under the assumptions in this section and assuming (2.3).
2. Derive the SMLE of β in Example 1.
3. Construct a level-0.10 two sided test for $H_0 : \beta_2 = 0$, where $\beta^t = (\beta_1, \beta_2)$. Give the expression as explicitly as possible.

§5.2.2. Discrete RC data.

If F_o and G are continuous, the order statistics of the observations satisfy

$$M_{(1)} < M_{(2)} < \cdots < M_{(n)}.$$

In this section, we consider the case that there are ties in the observations. Using the idea of conditional probability, Cox suggests a likelihood function for the discrete RC data as follows.

Notations:

$a_1 < \cdots < a_g$ are all distinct exact observations;

d_j is the # of deaths at a_j;

$\mathcal{S}_j$ is the collection of all the combinations of selecting d_j elements out of those in $\mathcal{R}(a_j)$;

$\mathcal{D}_j = \{i : \ \delta_i = 1, M_i = a_j\}$;

$\mathcal{R}(a_j)$, r_j, ϕ are the same as in \$ 5.2.1.

A log likelihood is defined as

$$l(\beta) = \ln \prod_{j=1}^{g} \frac{\prod_{i\in\mathcal{D}_j} \phi(i)}{\sum_{(i_1,\ldots,i_{d_j})\in\mathcal{S}_j} \phi(i_1)\cdots\phi(i_{d_j})}.$$

The SMLE $\hat{\beta}$ of β maximizes $l(\beta)$.

Remark. The likelihood $l(\beta)$ needs to be modified as follows.

$$l(\beta) = \begin{cases} \ln \prod_{j=1}^{g} \frac{\prod_{i\in\mathcal{D}_j} \phi(i)}{\sum_{(i_1,\ldots,i_{d_j})\in\mathcal{S}_j} \phi(i_1)\cdots\phi(i_{d_j})} & \text{if } \delta_{(n)} = 0; \\ \ln \prod_{j=1}^{g-1} \frac{\prod_{i\in\mathcal{D}_j} \phi(i)}{\sum_{(i_1,\ldots,i_{d_j})\in\mathcal{S}_j} \phi(i_1)\cdots\phi(i_{d_j})} & \text{if } \delta_{(n)} = 1. \end{cases}$$

as $h(x) = \psi h_o(x), \quad x < \tau$, where $\tau = \sup\{t : S_o(t) > 0\}$.

Remark. The form of log likelihood function

$$\mathcal{L}(\beta) = \sum_{i:\ ex} \ln\psi(\beta, \mathbf{z}_i) + \sum_{i:\ ex} \ln h_o(M_i) + \sum_{i=1}^{n} \psi(\beta, \mathbf{z}_i) \ln S_o(M_i)$$

is not applicable, as $h(t) = \frac{f(t)}{S(t-)} \neq \frac{f(t)}{S(t)}$ and $S(t|\mathbf{z}) \neq (S_o(t))^{\exp(\beta\mathbf{z})}$.

Remark. Under certain assumptions, the SMLE $\hat{\beta}$ is consistent and asymptotically normally distributed. Its covariance matrix can be estimated by $-\left(\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^t}|_{\beta=\hat{\beta}}\right)^{-1}$.

§**5.2.2.2. Homework**

1. Suppose that ψ is log linear and there are 6 data: $(M_i, \delta_i, (z_1, z_2))$'s are

$$(4,1,(1,0)),(1,0,(1,1)),(5,1,(0,1)),(4,1,(1,0)),(1,1,(0,0)),(4,0,(0,1)).$$

Compute the SMLE of $\hat{\beta}$.

§**5.2.3. Conditional probability approach in continuous case**.

By continuous assumption, there are no ties. Let $M_{(1)} < \cdots < M_{(n)}$ be order statistics of observations. For each uncensored time $a_j = M_{(i)}$, denote $\mathcal{R}_i = \mathcal{R}(a_j)$, $j = 1, ..., g$, where $g = \sum_{i=1}^{n} \delta_i$.

Cox made the following assumptions:

1. there is only one death at a_j and thus we can pretend that there is only one death in $[a_j, a_j + \epsilon)$ for a small ϵ;
2. death only occurs in $[a_j, a_j + \epsilon)$, $j = 1, ..., g$.

Then $P\{$the i-th person died in $[a_j, a_j + \epsilon)|$the i-th person survived time $a_j\}$

$\approx h(a_j|\mathbf{z}_i)\epsilon$, and

$$\begin{aligned}
& P\{\text{ a death in } [a_j, a_j+\epsilon)| \text{ each in } \mathcal{R}_i \text{ survives time } a_j\} && (1)\\
=& P\{\cup_{k\in\mathcal{R}_i} A_k| \cap_{h\in\mathcal{R}_i} C_h\} && (A_h = \{X_h \in [a_j, a_j+\epsilon)\} \text{ and } C_h = \{X_h \geq a_j\})\\
=& \sum_{k\in\mathcal{R}_i} P\{\text{ the } k\text{-th died in } [a_j, a_j+\epsilon)| \text{ each in } \mathcal{R}_i \text{ survives time } a_j\}\\
=& \sum_{k\in\mathcal{R}_i} P\{A_k| \cap_{h\in\mathcal{R}_i} C_h\} && (2)\\
=& \sum_{k\in\mathcal{R}_i} P\{A_k|C_k \cap \cap_{h\in\mathcal{R}_i, h\neq k} C_h\}\\
=& \sum_{k\in\mathcal{R}_i} P\{A_k|C_k\} && \text{(by independence of observations)}\\
=& \sum_{k\in\mathcal{R}_i} P\{\text{ the } k\text{-th died in } [a_j, a_j+\epsilon)| \text{ the } k\text{-th survived time } a_j\}\\
\approx& \sum_{k\in\mathcal{R}_i} h(a_j|\mathbf{z}_k)\epsilon \quad \left(h(t)\epsilon = \frac{f(t)\epsilon}{S(t-)} \text{ and P\{the k-th death in } [a_j, a_j+\epsilon)\} \approx f(a_j)\epsilon\right)
\end{aligned}$$

as any one in the risk set may die. Let

B_i the event {one death in $[a_j, a_j + \epsilon)$ from $\mathcal{R}_i$, and each in $\mathcal{R}_i$ survived a_j},

$$\begin{aligned}
& P\{A_i|B_i\}\\
=& P\{A_i|(\cup_{h\in\mathcal{R}_i} A_h) \cap (\cap_{k\in\mathcal{R}_i} C_k)\}\\
=& P\{A_i| \cap_{h\in\mathcal{R}_i} C_h\}/P\{\cup_{k\in\mathcal{R}_i} A_k| \cap_{h\in\mathcal{R}_i} C_h\}\\
& \quad \left(as\ \frac{P\{A_i\cap\cap_{h\in\mathcal{R}_i} C_h\}}{P\{\cap_{h\in\mathcal{R}_i} C_h\}} \frac{P\{\cap_{h\in\mathcal{R}_i} C_h\}}{P\{\cup_{k\in\mathcal{R}_i} A_k\cap\cap_{h\in\mathcal{R}_i} C_h\}} = \frac{P\{A_i\cap\cap_{h\in\mathcal{R}_i, h\neq i} C_h\}}{P\{\cup_{k\in\mathcal{R}_i} A_k\cap\cap_{h\in\mathcal{R}_i} C_h\}}\right)\\
=& \frac{h(a_j|\mathbf{z}_i)\epsilon}{\sum_{k\in\mathcal{R}_i} h(a_j|\mathbf{z}_k)\epsilon} && (\text{by } \quad (1) \text{ and } (2))\\
(P\{A_i|B_i\}) \quad =& \frac{\phi(i)}{\sum_{k\in\mathcal{R}_i} \phi(k)} \text{ (since } h(a_j|\mathbf{z}_i) = \phi(i)h_o(a_j)). && (3)
\end{aligned}$$

Cox defines the conditional likelihood to be

$$lik = \prod_{i\in\mathcal{D}} P\{A_i|B_i\} = \prod_{i\in\mathcal{D}} \frac{\phi(i)}{\sum_{k\in\mathcal{R}_i}\phi(k)}.$$

§5.2.4. Partial likelihood approach in continuous case

By rearranging the index, WLOG, we can assume that $a_j = M_j$, $j = 1, ..., g$. Let

$A_i = \{$the i-th person died in $[a_j, a_j + \epsilon)\}$ and

B_i the event $\{$one death in $[a_j, a_j + \epsilon)$ from $\mathcal{R}_i$, and each in $\mathcal{R}_i$ survived $a_j\}$, $i = 1, ..., g$. Then define a full likelihood under assumptions 1 and 2 as

$$\begin{aligned}
P\{A_1B_1\cdots A_gB_g\} =& P(B_1)P(A_1|B_1)P(B_2|A_1B_1)P(A_2|A_1B_1B_2)\times\cdots \\
&\times P(B_g|A_1\cdots A_{g-1}B_1\cdots B_{g-1})P(A_g|A_1\cdots A_{g-1}B_1\cdots B_g) \\
=& P(A_1|B_1)P(A_2|A_1B_1B_2)\cdots P(A_g|A_1\cdots A_{g-1}B_1\cdots B_g) \\
&\times P(B_1)P(B_2|A_1B_1)\cdots P(B_g|A_1\cdots A_{g-1}B_1\cdots B_{g-1}) \\
=& P(A_1|B_1)P(A_2|B_2)\cdots P(A_g|B_g) \qquad (= lik) \\
&\times P(B_1)P(B_2|A_1B_1)\cdots P(B_g|A_1\cdots A_{g-1}B_1\cdots B_{g-1}).
\end{aligned}$$

$$lik = P(A_1|B_1)P(A_2|B_2)\cdots P(A_g|B_g). \qquad (\text{(see (3) in §5.3.2)})$$

Thus lik is called the *partial likelihood* by Cox.

§5.2.5. Nonparametric estimation of S_o.

Under the PH model: $h(t|\mathbf{z}) = \psi(\beta, \mathbf{z})h_o(t)$ and if X is continuous then $S(t|\mathbf{z}) = (S_o(t))^{\psi(\beta,\mathbf{Z})}$.

Baseline integrated hazard: $H_o(t) = \int_0^t h_o(u)du$ or $\sum_{u\le t}\ln(1 - h_o(u)) \approx \sum_{u\le t} h_o(u)$.

$S_o(t) = exp(-H_o(t))$.

There is no MLE of S_o under Cox's assumption.

Cox proposes an estimator

$$\hat{S}_o(t) = exp(-\sum_{a_j\le t}\frac{d_j}{\sum_{l\in\mathcal{R}(a_j)}\hat{\phi}(l)}).$$

Note that if $\phi(i) = 1$,

$$\hat{S}_o(t) = exp(-\sum_{a_j\le t}\frac{d_j}{r_j}) = exp(-\sum_{a_j\le t}\hat{h}(a_j)).$$

Breslow (1972, JRSS,B) also proposes another estimator.

The third estimator can be the one which maximizes $\mathcal{L}(\beta)$ in §5.2.1.

$$\mathcal{L}(\beta) = \sum_{i:\ ex, \mathcal{M}_{(n)}<\tau} \ln e^{\beta^t\mathbf{Z}_i} + \sum_{i:\ ex, \mathcal{M}_{(n)}<\tau} \ln h_o(M_i) + \sum_{i=1}^{n} e^{\beta^t\mathbf{Z}_i}\ln S_o(M_i),$$

or

$$\mathcal{L}(\beta) = \sum_{i:\ ex, \mathcal{M}_{(n)}<\tau} \ln e^{\beta^t\mathbf{Z}_i} + \sum_{i:\ ex, \mathcal{M}_{(n)}<\tau} \ln f_o(M_i) + \sum_{i:\ rc} e^{\beta^t\mathbf{Z}_i}\ln S_o(M_i),$$

$$f_o(t) = S_o(t-) - S_o(t),$$

which corresponding to the PH model, and

$$Ł(F) = \prod_{i:\ rc}(S_o(M_i))^{\psi(\beta,\mathbf{Z}_i)} \prod_{i:\ ex}[(S_o(M_i-))^{\psi(\beta,\mathbf{Z}_i)} - (S_o(M_i))^{\psi(\beta,\mathbf{Z}_i)}].$$

which corresponding to proportional integrated hazards model.

§5.3. Extension of PH model with IC data

The conditional probability approach does not work for IC data under the PH model. Finkelstein (1986, Biometrics) first considers the following extension of the PH model to IC data. She defines the SMLE of β to be the one that maximizes the likelihood function $\text{Ł}(\beta, F_o) =$

$$\prod_{i:\ L_i<R_i} [(S_o(L_i))^{\psi(\beta,\mathbf{Z}_i)} - (S_o(R_i))^{\psi(\beta,\mathbf{Z}_i)}] \prod_{i:\ L_i=R_i} [(S_o(L_i-))^{\psi(\beta,\mathbf{Z}_i)} - (S_o(R_i))^{\psi(\beta,\mathbf{Z}_i)}]. \tag{5.3.1}$$

β and S_o need to be estimated simultaneously.

Remark. The likelihood corresponds to proportional integrated hazards model, not really the PH model, unless one assumes that X is continuous. If X is discrete, then the likelihood under the PH model is different.

Since $h(t) = f(t)/S(t-)$ and $S(t) = \prod_{x_i\le t, x_i\in D_f}(1-h(x_i))$,

$f(t|\mathbf{z}) = h(t|\mathbf{z})S(t-|\mathbf{z})$, where D_f is the set of points at which $f>0$),

the correct generalized likelihood of the PH model with discrete IC data is

$$\text{Ł}(\beta, h_o) = \prod_{i:\ L_i<R_i} [S(L_i|\mathbf{z}_i) - S(R_i|\mathbf{z}_i)] \times \prod_{i:\ L_i=R_i} [h(L_i|\mathbf{z}_i)S(L_i-|\mathbf{z}_i)] \tag{5.3.2}$$

where $h(t|\mathbf{z}) = \begin{cases} \exp(\beta\mathbf{z})h_o(t) & \text{if } t<\tau \\ 1 & \text{if } t=\tau \end{cases}$,

$S(t|\mathbf{z}) = \prod_{x_i\le t, x_i\in D_{f_o}}(1-h(x_i|\mathbf{z}))$, and

$S(t-|\mathbf{z}) = \prod_{x_i< t, x_i\in D_{f_o}}(1-h(x_i|\mathbf{z}))$.

Notice that $h_o(t) = f_o(t)/S_o(t-)$ and $S_o(t-) = \prod_{x_i<t, x_i\in D_{f_o}}(1-h_o(x_i))$.

Remark. The likelihood in (5.3.2) is actually applicable for both continuous and discrete $X|\mathbf{z}$, though (5.3.2) and (5.3.1) will result in different estimates.

In order to compute the GMLE of S_o in both the PIH model (5.3.1) or the PH model (5.3.2), let A_1, ..., A_m be the II's induced by I_i's, the observed intervals, and $p_j = \mu_F(A_j)$. Consider S_o of form

$$S_o(t) = \sum_{j:\ A_j\cap(t,\infty]\neq\emptyset} p_j.$$

The variance of the SMLE $\hat\beta$ can be estimated by a $p\times p$ matrix V_{11}, where

$$\begin{pmatrix} V_{11} & V_{12} \\ V_{12}^t & V_{22} \end{pmatrix} = J^{-1},$$

where J is the empirical Fisher information matrix, i.e., minus the second partial derivatives matrix of the log likelihood function with respect to $(\beta^t, p_1, ..., p_{m-1})$, with all parameters replaced by their SMLE.

Finkelstein (1986, Biometrics) suggests to use the Newton-Raphson method to compute the SMLE of (β, S_o). It turns out most of the time, the approach does not work. The reason is that it is often that the SMLE of some p_j is zero and the algorithm will lead to a point which is not an SMLE but $p_j = 0$ for some j. The derivative $\frac{\partial \text{Ł}}{\partial p_j} < 0$. Then the algorithm has to stop. The following is such a counterexample.

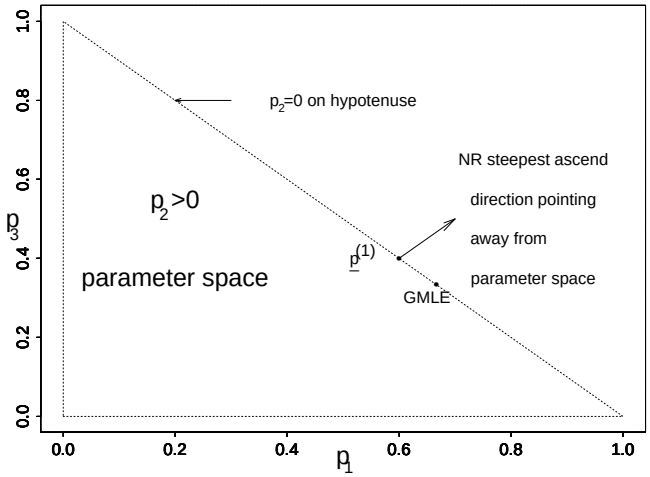
Consider fitting the PIH model with 5 observations (L_i, R_i, Z_i): (2,5,0), (3,4,0), (5,9,1), (1,6,1), (7,8,0). It can be viewed as data from two groups, corresponding to $Z_i = 0$ or 1. Then, the innermost intervals are (3,4], (5,6] and (7,8]. Let the weights on these innermost intervals be p_1, p_2 and p_3, with $p_1+p_2+p_3 = 1$ and $p_i \ge 0$. Note that the baseline survival function S satisfies $S(4-) = 1$, $S(4) = S(6-) = p_2+p_3$, $S(6) = S(8-) = p_3$ and $S(8) = 0$. For this example, it is more convenient to express the likelihood as a function of p_i's rather than S. The likelihood is $L = p_1^2 p_3(1-p_3^{e^\beta})(p_2+p_3)^{e^\beta}$. Since $p_1+p_2+p_3 = 1$, in view of L, it is simpler to write the log likelihood as

$$\mathcal{L} = \log[p_1^2 p_3(1-p_1)^{e^\beta}(1-p_3^{e^\beta})].$$

The parameter space is $\Omega = \{(\beta, p_1, p_3) : \beta\in(-\infty,\infty), p_1\ge 0, p_3\ge 0, p_1+p_3\le 1\}$ with $p_2 = 1-p_1-p_3$. For convenience, we write $\alpha = e^\beta$ hereafter. Thus,

$$\mathcal{L} = 2\log p_1 + \log p_3 + \alpha\log(1-p_1) + \log(1-p_3^\alpha).$$

Since the likelihood function has only three variables, it can be shown by direct derivation that the GMLE of (β, p_1, p_2, p_3) is approximately $(-0.461, 2/3, 0, 1/3)$.



A feasible algorithm for the GMLE. Abusing notations, we identify S with a vector $(S_1, ..., S_m)$. Similarly, we identify $S^{(i)}$ with $(S_1^{(i)}, ..., S_m^{(i)})$.

Step 0. Let $b^{(0)} = 0$ be the initial estimate of β and the GMLE of a survival function with observations (L_j, R_j), $j = 1, ..., n$ be the initial estimate of $S^{(0)}$.

Step $i+1$ ($i \geq 0$). Let $b^{(i)}$ and $S^{(i)}$ be the updated values of b and S at Step i. Do b=step and S-step as follows.

* (b**-step**) With $S = S^{(i)}$ fixed, find a b so that the log likelihood function $\mathcal{L}(S^{(i)}, \cdot)$ increases. Denote the up-dated estimate b by $b^{(i+1)}$. In particular, one can use the NR method to obtain the maximum point b of the log likelihood function with the given $S = S^{(i)}$.
* (S**-step**) With $b = b^{(i+1)}$ fixed, search a non-increasing S so that the log likelihood function $\mathcal{L}(\cdot, b^{(i+1)})$ is maximized (or increases). Since $S(\eta_i) = p_{i+1} + \cdots + p_m$, let $p^{(i+1),0} = p^{(i)}$. At Sub-step j ($j = 1, ..., m$), update $(p_1, ..., p_m)$ by $(p_1^{(i+1),j}, ...p_m^{(i+1),j})$, where $p_h^{(i+1),j} = p_{h,u_o}$ and

$$p_{h,u} = \begin{cases} \frac{p_h^{(i+1),j-1}+u}{1+u} & \text{if } h = j, \\ \frac{p_h^{(i+1),j-1}}{1+u} & \text{if } h \neq j, \end{cases}$$

$h = 1, ..., m$, $u_o > 0$ is a number maximizing $L(b^{(i+1)}, S_{\cdot,u})$ where $S_{\cdot,u} = (S_{1,u}, ..., S_{m,u})$ and $S_{i,u} = p_{i+1,u} + \cdots + p_{m,u}$.

Note: If such u_o is difficult to choose, one may choose a u_o satisfying

$$L(b^{(i+1)}, S^{(i+1),j}) > L(b^{(i+1)}, S^{(i+1),j-1}). \tag{3.1}$$

In particular, if $\frac{\partial}{\partial u} \ln L(b^{(i+1)}, S_{\cdot,u})\big|_{u=0} > 0$, $u_o = c^k \frac{\partial}{\partial u} \ln L(b^{(i+1)}, S_{\cdot,u})\big|_{u=0}$, where $S_{\cdot,u} = (S_{1,u}, ..., S_{m,u})$, $c \in (0.1)$, and k is the smallest non-negative integer such that Inequality (3.1) holds.

Stop at convergence.

Remark. The restriction $u > 0$ can be replaced by $u > -p_h^{(i+1),j-1}$.

If $X|\mathbf{z}$ is not continuous, (5.3.1) is not the likelihood function of the PH model. Now consider fitting Cox's regression model (5.3.2). First compute $S(L_i|\mathbf{z}_i) - S(R_i|\mathbf{z}_i)$ in the following table.

(L, R, z)	$S(L), S(R)$	$S(L\|z) - S(R\|z)$	*simplication*
$(2, 5, 0)$	$1, S(4\|0)$	$1 - (p_2 + p_3)$	p_1
$(3, 4, 0)$	$1, S(4\|0)$	$1 - (p_2 + p_3)$	p_1
$(5, 9, 1)$	$S(4\|1), 0$	$(1 - e^\beta p_1) - 0$	$1 - e^\beta p_1$
$(1, 6, 1)$	$1, S(6\|1)$	$1 - (1 - e^\beta p_1)(1 - e^{\beta \frac{p_2}{p_2+p_3}})$	$1 - (1 - e^\beta p_1)(1 - e^{\beta \frac{p_2}{1-p_1}})$
$(7, 8, 0)$	$S(6\|0), 0$	$p_3 - 0$	$1 - p_1 - p_2$

The likelihood is $\mathcal{L} = \ln \prod_i (S(L_i|\mathbf{z}_i) - S(R_i|\mathbf{z}_i)) =$

$$2\ln p_1 + \ln(1-p_1-p_2) + \ln[1-e^{\beta}p_1] + \ln[1-(1-e^{\beta}p_1)(1-e^{\beta}\frac{p_2}{1-p_1})]$$

§**5.3.2. Homework**.

1. Verify the GMLE of (β, p_1, p_2, p_3) for the data related to the figure
   is approximately $(-0.461, 2/3, 0, 1/3)$.
   You do not need to use the algorithm mentioned above.
2. Show that the GMLE of (β, p_1, p_2, p_3) under likelihood (5.3.2) is $(-0.288, 2/3, 0, 1/3)$ $(e^{\beta} = 3/4)$.

§**5.4. Accelerated lifetime model and regression analysis with IC data**

Assume that conditional on $\mathbf{Z} = \mathbf{z}$,

$$X = T/e^{\beta^t \mathbf{z}}.$$

For simplicity, we consider $p = 1$. Then conditional on $Z = z$, $\ln T = \ln X + \beta z$ has a distribution which does not depend on β and Z and has the mean α and variance σ^2. For simplicity, we shall replace β, $\ln X$ and $\ln T$ by $-\beta$, X and ϵ, respectively. That is, the ordinary linear regression set-up:

$$X = \beta z + \epsilon, \text{ where } E(\epsilon) = \alpha \text{ and } Var(\epsilon) = \sigma^2.$$

Here $\beta\mathbf{z}$ can be interpreted as $\beta'\mathbf{z}$. For simplicity, we only consider the univariate case.

Question: $(\alpha, \beta) = ?$

If one has complete data, (X_i, z_i), $i = 1, ..., n$, the least squares estimate of β is the one that minimizes

$$SS(\alpha, \beta) = \sum_{i=1}^{n} (X_i - \alpha - \beta z_i)^2.$$

The solution is

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (U^t U)^{-1} U^t \mathbf{X},$$

where $U = \begin{pmatrix} 1 & z_1 \\ \vdots & \vdots \\ 1 & z_n \end{pmatrix}$ and $\mathbf{X}^t = (X_1, ..., X_n)$. The least squares estimate can be viewed as the solution of

$$(\alpha, \beta) = \text{argmin } r(a, b), \text{ where } r(\alpha, \beta) = n \int t^2 d\hat{F}(t),$$

where $\hat{F}$ is the e.d.f. of F based on observations $X_i - \alpha - \beta z_i$, $i = 1, ..., n$.

Moreover, the LSE is also the MLE under the normal assumption. Thus it is the solution to the equation $\frac{\partial \ln \mathbb{L}}{\beta} = 0$ and $\frac{\partial \ln \mathbb{L}}{\partial \alpha} = 0$.

§5.4.1. **Miller Estimator** (1976, Biometrika). The estimator of (α, β) with RC data is $(\hat{\alpha}, \hat{\beta})$ that minimizes $r(a, b)$, where $\hat{F}$ is the PLE of F based on observations $(M_i - a - bz_i, \delta_i)$, $i = 1, ..., n$. Let s_i, ..., s_m be the weight assigned by $\hat{F}$ to the II's. If the largest observation is right censored, he suggested to pretend that it is exact to avoid assigning weight to $+\infty$ and thus $r(\alpha, \beta) = +\infty$. Note that s_i's are only functions of β not of α, as changing α only shifts the II's and the corresponding intervals, but not the weights. If we let $\hat{F}_0$ be the PLE based on $(M_i - \beta z_i, \delta_i)$'s and $\hat{F}_\alpha$ the one based on $(M_i - \alpha - \beta z_i, \delta_i)$, then $\hat{F}_\alpha(t) = \hat{F}_0(t + \alpha)$. Let η_1, ..., η_m be all the distinct exact observations based on $M_i - \beta z_i$'s.

$$\begin{pmatrix} \alpha: & 0 & c \\ II's: & \eta_j & \eta_j - c \\ weights: & s_j & s_j \end{pmatrix}$$

So we write $s_j = s_j(\beta)$ and $\eta_j = \eta_j(\beta)$. Denote $w_i(\beta)$ the weight assigned by $\hat{F}$ to each observation $(M_i - \beta z_i, \delta_i)$ (treating each observation is one unit, even if there are ties).

Note that we may have ties at η_j, say, there are h exact observations such that $X_{i_k} - \beta z_{i_k} = \eta_j$ for $k = 1, ..., h$. Then $w_{i_k} = s_j/h$ for each k. Then

$$r(\alpha, \beta) = n \sum_{i=1}^{n} (M_i - \alpha - \beta z_i)^2 w_i(\beta).$$

Taking derivative of $r(\alpha, \beta)$ w.r.t. α and setting the derivative to be zero yield

$$\alpha = \hat{\alpha}(\beta) = \sum_{i=1}^{n} w_i(\beta)(M_i - \beta z_i). \tag{1.0}$$

Thus it suffices to search $\hat{\beta}$ that minimizes

$$H(\beta) = \sum_{j=1}^{n} w_j(\beta)[M_j - \hat{\alpha}(\beta) - \beta z_i]^2. \tag{1.1}$$

Miller suggested the following iterative procedure:

1. Assign an initial value $\beta = \frac{\sum_{i:\ ex} X_i(z_i - \overline{z}_u)}{\sum_{j:\ ex}(z_j - \overline{z}_u)^2}$, where $\overline{z}_u$ ($\overline{X}_u$) is the average of z_i's (X_i's) corresponding to exact observations of X_i.
2. Obtain the II's, $\eta_j(\beta)$ and the GMLE of $s_j(\beta)$'s based on $(M_i - \beta z_i, \delta_i)$ with the given β, and compute $w_i(\beta)$.
3. Update β by $\frac{\sum_{i=1}^{n} w_i(\beta) X_i(z_i - \overline{z}_w)}{\sum_{j=1}^{n} w_j(\beta)(z_j - \overline{z}_w)^2}$, where $\overline{z}_w = \sum_{j=1}^{n} w_j z_j$.
4. Repeat steps 2 and 3 until β converges or oscillates between two values. In the latter case, take the midpoint as an estimate of β.
   The variance of β can be estimated by

$$\hat{\sigma}^2_{\hat{\beta}} = \frac{\sum_{j=1}^{n} w_j(\hat{\beta})(M_i - \hat{\alpha} - \hat{\beta} z_i)^2}{n \sum_{j=1}^{n} w_j(\hat{\beta})(z_i - \overline{z}_w)^2}, \text{ comparing to } \sigma^2 / \sum_{i=1}^{n} (z_i - \overline{z})^2.$$

   The consistency and asymptotic normality were considered under the assumption that

$$\mathrm{P}\{X \text{ is not censored} | X = t\} > 0 \text{ for all possible } t \tag{1.2}$$

and the censoring distribution is of form

$$G(y|z) = G_o(y - \beta z), \text{ where } G_o \text{ is a cdf.}$$

However, the estimator has not been proved to be asymptotically efficient even under the normal assumption.
Remark For IC data, if we replace the PLE by GMLE, the above procedure can be adopted, provided we define that the GMLE only has jumps at midpoints of the II's. However, the consistency and asymptotic normality have not been verified.
§**5.4.1.2. Homework**. There are 4 observations $(M_i, \delta_i, \mathbf{z}_i)$'s: $(3, 1, 2), (4, 0, 1), (1, 1, 1), (7, 1, 2)$. Find the Miller estimator of β under the linear regression model. $X_i = \beta z_i + \epsilon_i$. You are able to find the solution explicitly, because there are at most 6 distinct values of $\hat{S}_b(T_i(b))$ (as a function of b) for each fixed i, where $T_i = M_i - b z_i$.

§5.4.2. **Buckley-James Estimator** with RC data (1979, Biometrika).
Given RC data, $(M_i, \delta_i, \mathbf{z}_i)$s, denote $T_i = T_i(\mathbf{b}) = M_i - \mathbf{b}'\mathbf{z}_i$, where $\mathbf{b}$ is a $p-$ dimensional vector.

Denote $\hat{S}_{\mathbf{b}}$ the PLE of the survival function based on (T_i, δ_i), $i = 1, ..., n$. It is worth mentioning that $\hat{S}_{\mathbf{b}}$ depends on $\mathbf{b}$ as T_i's depend on $\mathbf{b}$. Denote $\hat{f}_{\mathbf{b}}$ and $\hat{F}_{\mathbf{b}}$ the PLE's of the density function and the distribution function, respectively. Let $\hat{\mathbf{X}}^* = \hat{\mathbf{X}}^*(\mathbf{b}) = (\hat{X}_1^*, ..., \hat{X}_n^*)'$, where

$$\hat{X}_i^* = M_i\delta_i + (1-\delta_i)[\mathbf{b}'\mathbf{z}_i + \frac{\sum_{t\in\mathcal{A}_i} t\hat{f}_{\mathbf{b}}(t)}{\hat{S}_{\mathbf{b}}(T_i)}], \tag{2.1}$$
$$\mathcal{A}_i = \{t: \ t > T_i, \hat{f}_{\mathbf{b}}(t) > 0\}.$$

Notice If the largest observation $T_{(n)}$ among T_i's is right censored, then Buckley and James suggest to treat $T_{(n)}$ as uncensored and define the Buckley and James estimator (BJE) $(\hat{\alpha}, \hat{\beta})$ by $\hat{\alpha} = \overline{\hat{X}^*} - \hat{\beta}'\overline{\mathbf{z}}$ and $b = \hat{\beta}$ being a solution to $H(b) = 0$, where $\overline{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i$ and

$$H(\mathbf{b}) = \sum_{i=1}^n (\hat{X}_i^*(\mathbf{b}) - \overline{\hat{X}^*}(\mathbf{b}) - \mathbf{b}'(\mathbf{z}_i - \overline{\mathbf{z}}))(\mathbf{z}_i - \overline{\mathbf{z}}).$$

Buckley and James point out that the solution may not exist. In the latter case, if $p = 1$, the BJE of β is modified as the zero-crossing point of $H(b)$ (see Lai and Ying (1991)). One says that $\hat{b}$ is a zero-crossing point of a function H if $H(b)$ changes its sign at $b = \hat{b}$.

Remark. With Buckley and James' modification, (2.1) defines X_i^* and is used in this note. One can also define $\hat{S}_{\mathbf{b}}(T_{(n)} + 1) = 0$ if $\delta_{(n)} = 0$. Then no need to modify $\delta_{(n)}$ is $\delta_{(n)} = 0$.

Two ways of reasoning of the BJE:

(1) Conditional expectation.

Note that

$$E(X_i|z_i) = \alpha + \beta z_i.$$

We only observe $M_i = X_i \wedge Y_i$ and

$$E(M_i|z_i) \neq \alpha + \beta z_i.$$

Define

$$X_i^* = X_i\delta_i + E(X_i|X_i > Y_i)(1-\delta_i) \text{ (not the same as (2.1))} \tag{2.2}$$

where $E(X_i|X_i > Y_i)$ is defined as $g(Y_i)$, and $g(y) = E(X_i|X_i > y)$. Then

$$E(X_i^*|z_i) = \alpha + \beta z_i, \tag{2.3}$$

as abusing notation, write $E(X_i) = E(X_i|\mathbf{z}_i)$ and $E(X_i^*) = E(X_i^*|\mathbf{z}_i)$, then

$$\begin{aligned}
&E(X_i^*)\\
=&E(E(X_i^*|\delta_i)))\\
=&P(\delta_i = 1)E(X_i^*|\delta_i = 1) + P(\delta_i = 0)E(X_i^*|\delta_i = 0)\\
=&P(\delta_i = 1)E(X_i|\delta_i = 1) + P(\delta_i = 0)E(E(X_i|X_i > Y_i)|\delta_i = 0)\\
=&P(\delta_i = 1)E(X_i|\delta_i = 1) + P(\delta_i = 0)E(E(X_i|\delta_i = 0)|\delta_i = 0)\\
=&P(\delta_i = 1)E(X_i|\delta_i = 1) + P(\delta_i = 0)E(X_i|\delta_i = 0)\\
=&E\{E(X_i|\delta_i)\}\\
=&E(X_i)
\end{aligned}$$

If we could observe X_1^*, ..., X_n^*, we can use

$$\hat{\alpha} = \overline{X}^* - \hat{\beta}\overline{z} \text{ and } \hat{\beta} = \frac{\sum_{i=1}^n X_i^*(z_i - \overline{z})}{\sum_{j=1}^n (z_i - \overline{z})^2}. \tag{2.4}$$

Since we cannot observe all X_i^*'s, we replace X_i^*'s by their predictors

$$\hat{X}_i^* = X_i\delta_i + \hat{E}(X_i|X_i > Y_i)(1-\delta_i)$$

(which is (2.1)), where

$$\hat{E}(X_i|X_i > Y_i) = \hat{\beta}z_i + \frac{\sum_{t>M_i-\hat{\beta}z_i} t\hat{f}_{\hat{\beta}}(t)}{\hat{S}(M_i - \hat{\beta}z_i)}, \tag{2.5}$$

and $\hat{S}$ is the PLE of S based on observations $(M_i - \hat{\beta}z_i, \delta_i)$'s.

(2) An M-estimator based on $N(\mu, \sigma)$.

The likelihood function for given RC data is

$$Ł = \prod_{i=1}^{n}(f(T_i))^{\delta_i}(S(T_i))^{1-\delta_i}.$$

$$\frac{\partial \ln Ł}{\partial \beta} = \sum_{i=1}^{n}\{\delta_i\frac{f'}{f}(T_i) + (1-\delta_i)\frac{-f}{S}(T_i)\}(-\mathbf{z}_i).$$

Under the normal assumption, $f(t) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(t-\mu)^2}{2\sigma^2})$,
$\frac{f'}{f}(t) = -\frac{t-\mu}{\sigma^2}$,
$f(t) = -\int_t^\infty f'(x)dx = -\int_t^\infty \frac{f'}{f}(x)f(x)dx = \int_t^\infty \frac{x-\mu}{\sigma^2}f(x)dx$. Then

$$\frac{\partial \ln Ł}{\partial \beta} = \sum_{i=1}^{n}\{\delta_i(T_i-\mu) + (1-\delta_i)\frac{\int_{T_i}^\infty (x-\mu)f(x)dx}{S(T_i)}\}(\mathbf{z}_i/\sigma^2). \tag{2.6}$$

The normal equation $H(\mathbf{b})$ is obtained by replacing f and S by their PLE $\hat{f}_{\mathbf{b}}$ and $\hat{S}_{\mathbf{b}}$, replacing μ by $\overline{T}^*$ and multiplying σ^2. In other words, (2.1) is the same as

$$H(\mathbf{b}) = \sum_{i=1}^{n}(T_i^* - \overline{T^*})\mathbf{z}_i, \text{ where } T_i^* = \hat{X}_i^* - \mathbf{b}\mathbf{z}_i.$$

The consistency and the asymptotic properties of the BJE have been established under continuous assumptions by Lai and Ying (1991) and under discontinuous assumptions by Kong and Yu (2006). In particular, if $\epsilon \sim N(\mu, \sigma^2)$, then the BJE is asymptotically efficient, just like the LSE. Otherwise, it is not efficient. Moreover, under certain discontinuous assumptions, the BJE may not have asymptotic normality (see Kong and Yu (2006)).

Estimation of covariance matrix of the BJE, say $\Sigma_{\hat{\beta}}$.

With complete data the BJE becomes the LSE. Under the assumption that ϵ_1, ..., ϵ_n are i.i.d. with variance σ^2, given U (or $\mathbf{z}_i$s), the covariance matrix of the LSE is

$$\Sigma_{\hat{\beta}} = \sigma^2(U'U)^{-1}. \tag{2.7}$$

Reason:

$$\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} \mathbf{z}_1' \\ \vdots \\ \mathbf{z}_n' \end{pmatrix}\beta + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \text{ or } \mathbf{X} = U\beta + \underline{\epsilon}.$$

$$\hat{\beta} = (U'U)^{-1}U'\mathbf{X} = (U'U)^{-1}U'(U\beta + \underline{\epsilon}) = (U'U)^{-1}U'U\beta + (U'U)^{-1}U'\underline{\epsilon})$$

$$Cov(\hat{\beta}) = (U'U)^{-1}U'Cov(\underline{\epsilon})((U'U)^{-1}U')' = \sigma^2(U'U)^{-1}$$

Formula (2.7) is valid no matter whether $X|\mathbf{z}$ is continuous or discrete.

Lai and Ying (1991, p.1389) present an estimator of $\Sigma_{\hat{\beta}}$ under smoothness assumptions with RC data. Kong and Yu present (2006)) present another estimator of $\Sigma_{\hat{\beta}}$ under discrete assumptions with RC data. Both expressions are complicated and are not given here.

An estimator under the normal assumption is their empirical Fisher information matrix $(I(\hat{\beta}))^{-1}$, where

$$I(\hat{\beta}) = \sum_{i=1}^{n} (T_i^* - \overline{T^*})(\mathbf{z}_i - \overline{\mathbf{z}})\{(T_i^* - \overline{T^*})(\mathbf{z}_i - \overline{\mathbf{z}})\}'/\hat{\sigma}^4$$

$\hat{\sigma}^2$ is an estimator of σ^2. For each i, denote $m = \sum_i \delta_i$. The parameter σ can be estimated in two ways. If the largest M_i is not censored, then one can estimate it by $\hat{\sigma}^2 = \sum_i T_i^2 \hat{f}_{\hat{\beta}}(T_i) - (\sum_i T_i \hat{f}_{\hat{\beta}}(T_i))^2$. Otherwise, we can use the least squares method as follows. We can find the quantiles of $\hat{F}(T_i)$ under $N(0,1)$, say q_i's. Then we find the least squares estimate of (μ, σ) that minimizes

$$\sum_{i=1}^{n} \delta_i (\frac{T_i - \mu}{\sigma} - q_i)^2.$$

It can be shown that the LSE is $\hat{\sigma} = \frac{\overline{Tq} - \overline{T}\cdot\overline{q}}{\overline{q^2} - (\overline{q})^2}$, where $\overline{T} = \frac{1}{m}\sum_{i=1}^n \delta_i T_i$, $\overline{q} = \frac{1}{m}\sum_{i=1}^n \delta_i q_i$, $\overline{q^2} = \frac{1}{m}\sum_{i=1}^n \delta_i q_i^2$ and $\overline{Tq} = \frac{1}{m}\sum_{i=1}^n \delta_i T_i q_i$.

An estimator of the variance of $\hat{\beta}$ given by Buckley and James is

$$\hat{\sigma}^2_{\hat{\beta}} = \frac{\hat{\sigma}_u^2}{\sum_{i:\ ex}(z_i - \overline{z}_u)^2}.$$

where

$$\hat{\sigma}_u^2 = \frac{1}{\sum_{i=1}^n \delta_i - 2} \sum_{i:\ ex} [X_i - \overline{X}_u - \hat{\beta}(z_i - \overline{z}_u)]^2,$$

$$\overline{X}_u = \frac{\sum_{i:\ ex} X_i}{\sum_{i=1}^n \delta_i} \text{ and } \overline{z}_u = \frac{\sum_{i:\ ex} z_i}{\sum_{i=1}^n \delta_i}.$$

An alternative estimator of the variance of $\hat{\beta}$ is

$$\tilde{\sigma}^2_{\hat{\beta}} = \frac{1}{\sum_{i=1}^n \delta_i - 2} \Sigma_z^{-1} \tilde{\sigma}^2_{\epsilon},$$

where

$$\tilde{\sigma}^2_{\epsilon} = \int (t - \mu_{\epsilon,u})^2 d\hat{F}_{\hat{\beta}}(t),\ \mu_{\epsilon,u} = \int t d\hat{F}_{\hat{\beta}}(t),$$

$$\tilde{\Sigma}^2_z = \int_{t<\infty} (z(t) - \mu_{z,u})(z(t) - \mu_{z,u})' d\hat{F}_{\hat{\beta}}(t),\ \mu_{z,u} = \int z(t) d\hat{F}_{\hat{\beta}}(t),$$

$\hat{F}_{\hat{\beta}}$ is the modified PLE that moves the weight from $+\infty$ to the largest observation, and $z(t)$ =average of z in $\{z_i : T_i(\hat{\beta}) = t, i = 1, ..., n\}$.

The last two estimators try to mimic the expression $\sigma^2(U'U)^{-1}$, but they are not consistent estimators.

The extension of Buckley-James estimator to the IC data are considered by Li and Pu (1999) and Rabinowitz, Tsiatis, and Aragon (1995).

An iteration algorithm for the BJE (Buckley and James (1979)).

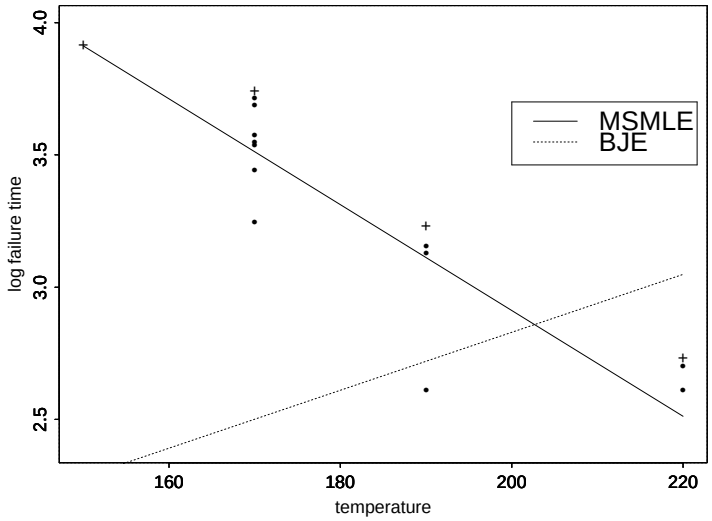
1. Give initial values to β.
2. Obtain $\hat{X}_i^*$'s using (2.1) with the given β.
3. Update β using (2.4) with the given $\hat{X}_i^*$'s.
4. Repeat steps 2 and 3 until β converges or oscillates between two values. In the latter case, take the midpoint as an estimate of β.

Remark.

1. The algorithm may not converges to a solution of the BJE even if the zero point of H exits (see Yu and Wong (2002)).
2. The BJE of β may not be unique. If there are both root and non-root zero-crossing point to $H(\mathbf{b})$, the iterative algorithm cannot present non-root zero-crossing-point of $H(\mathbf{b})$. In practice, it may **miss the most relevant estimate** as to be illustrated in Example 2.1.

Example 2.1. (Insulation data (Nelson 1973)). To evaluate a new Class-B insulation for electric motors, temperature-accelerated life testing was conducted on 40 motorettes. The main purpose was to estimate the distribution of insulation at the design temperature of 130^{o}C. Ten motorettes were put on test at each of four temperatures (150^{o}C, 170^{o}C, 190^{o}C, and 220^{o}C). Let z be the temperature (in ^{o}C) and X (or M) the logarithm of hours to failure of an insulation at temperature x. The data are plotted in Figure 1, "+" stands for right-censored observations and "·" stands for exact observations.

Fig. 5.1. MSMLE vs. BJE For Insulation Data



No matter what initial point is used, the existing algorithms always result in an estimate $\hat{\beta}_1 = 0.0109$, which is the unique solution to equation (2.3). The fitted line is plotted in Figure 1 in broken line. $\hat{\beta}_1$ does not make sense, as it should be negative according to the data (see Figure 1.). Yu and Wong (2002a) present an algorithm that can find all possible solutions for the BJE. Using this algorithm, we found that there are exactly 3 zero-crossing points: -0.0207, -0.0205 and -0.0193. They are approximately -0.02. We plot the BJE fitted line corresponding to $\hat{\beta} = -0.02$ in Figure 1 (in solid line). It appears to be a reasonable estimate and it is actually the semi-parametric MLE (SMLE).

Remark. After Yu and Wong's (2002a) paper was published, it is found out that the foregoing results of the BJE are due to treating $\hat{f}_b(M_{(n)}(b) = 0$ when $\delta_{(n)} = 0$, rather than setting $\hat{f}_{\mathbf{b}}(M_{(n)}(\mathbf{b})) = \hat{S}_{\mathbf{b}}(M_{(n)}(\mathbf{b}))$. Had the mistake not been made, there is just unique BJE and it is a root of H. Yu and Wong (2006) later found simulation data sets that have zero-crossing but not root of H.

A non-iterative algorithm for obtaining all BJE's (for $p = 1$) (Yu and Wong (2002a)):

1. Let b_{ij} be the solution to an equation $T_i(b) = T_j(b)$, where $z_i \neq z_j$ and $\delta_i \neq \delta_j$. Let $q_1 < \cdots < q_m$ be all the distinct solutions b_{ij}'s. Let $q_0 = -\infty$ and $q_{m+1} = \infty$.
2. (Case (1)). For each $h = 0, 1, ..., m$, first compute the PLE $\hat{S}_b$ for a $b \in (q_h, q_{h+1})$. For example, let b be the midpoint of the interval if $0 < i < m$, $b = q_1 - 1$ if $i = 0$, and $b = q_m + 1$ if $i = m$. Then compute $(M_i^*(b), z_i^*(b))$'s and

$$\hat{b}_h = \frac{\sum_{j=1}^n (z_j - \overline{z}) M_j^*}{\sum_{k=1}^n (z_k - \overline{z}) z_k^*}, \quad where$$

$$M_i^* = M_i \delta_i + (1 - \delta_i) \frac{\sum_{t > T_i(\mathbf{b})} \hat{f}_{\mathbf{b}}(t) \frac{\sum_{j=1}^n M_j \mathbf{1}_{(T_j(\mathbf{b}) = t, \delta_j = 1)}}{\sum_{k=1}^n \mathbf{1}_{(T_k(\mathbf{b}) = t, \delta_k = 1)}}}{\hat{S}_{\mathbf{b}}(T_i(\mathbf{b}))}$$

$$z_i^* = \delta_i z_i + (1-\delta_i)[\frac{\sum_{t>T_i(\mathbf{b})} \hat{f}_{\mathbf{b}}(t) \frac{\sum_{j=1}^n z_j \mathbf{1}_{(T_j(\mathbf{b})=t, \delta_j=1)}}{\sum_{k=1}^n \mathbf{1}_{(T_k(\mathbf{b})=t, \delta_k=1)}}}{\hat{S}_{\mathbf{b}}(T_i(\mathbf{b}))}]$$

If $\hat{b}_h \in (q_h, q_{h+1})$ then $\hat{b}_h$ is a root of $H(b)$ (see (2.7)) and thus is a BJE of β.

3. (Case (2)). Compute $H(q_i-)$, $H(q_i)$ and $H(q_i+)$, $i = 1, ..., m$. If $H(q_i-)H(q_i+) \le 0$, or $H(q_i-)H(q_i) \le 0$, or $H(q_i)H(q_i+) \le 0$, then q_i is a zero-crossing point of H and thus is a BJE of β.

Remark. In computing $H(b)$ is better off to use the following equivalent expression

$$H(b) = \sum_{i=1}^n (M_i^* - z_i^* b)(z_i - \overline{z}), \tag{2.7}$$

because even though (M_i^*, z_i^*) depends on b, it is constant in b on each interval (q_j, q_{j+1}), but $\hat{X}_i^*$ is not.

Remark. If data are complete, both the BJE $\hat{\beta}_{BJ}$ and the Miller estimator $\tilde{\beta}_M$ reduce to the LSE. In fact, $\tilde{\beta}_M = \text{argmin}_{\mathbf{b}} \int t^2 d\hat{F}_{\mathbf{b}}(t)$, where $\hat{F}_{\mathbf{b}}$ is the PLE of F_o based on $X_i - \overline{X} - \mathbf{b}'(z_i - \overline{z})$'s is an extension of $\text{argmin}_{a,\mathbf{b}} \sum_{i=1}^n (X_i - a - \mathbf{b}' z_i)^2/n$, and $\mathbf{b} = \hat{\beta}_{BJ}$ is the zero-crossing point of $H(\mathbf{b})$ and $H(b)$ is the extension of $\frac{d}{d\mathbf{b}} \sum_{i=1}^n (X_i - a - \mathbf{b}' z_i)^2$.

§5.4.2.2. Homework

1. There are 4 observations $(M_i, \delta_i, \mathbf{z}_i)$'s: $(3,1,2)$, $(4,0,1)$, $(1,1,1)$, $(7,1,2)$. Find the BJE of β under the linear regression model.
2. Under interval censoring, (2.2) can be rewritten as

$$X_i^* = E(X_i | X_i \in I_i), \text{ where } I_i \text{ is the } i\text{-th observed interval.}$$

2.a. Verify (2.3) under the mixed case model with continuous random vectors.

2.b. Give an estimator of X_i^* corresponding expressions for (2.5) and give a corresponding expressions for (2.6).

§5.4.3. An M-estimation approach

Recall that Huber (1964) proposed an M-estimator which is a zero point of a score function $\sum_{i=1}^n \psi(\theta, T_i)$, where θ is the parameter of interest and T_i's are observations. Modifying Huber's M-estimation (Huber, 1964), Zhang and Li (1996) consider another M-estimation approach with interval-censored data. The idea is to find a zero point of an *estimate* of the score function $\frac{\partial}{\partial \mathbf{b}} \ln Ł$ in b. We shall first illustrate via RC data.

Note that the likelihood function can be written as

$$Ł(b, S, f) = \prod_{i=1}^n (f(M_i - b' z_i))^{\delta_i} (S(M_i - b' z_i))^{1-\delta_i}.$$

Assuming that f and S are differentiable, the "MLE" of β is a critical point of Ł, where a critical point is a point that either Ł is not differentiable or $\frac{\partial Ł}{\partial b} = 0$. Moreover, to eliminate the effect of α, one needs to centralize z_i in Ł. Thus the derivative of $-\ln Ł$ is

$$\Phi = \sum_{i=1}^n (z_i - \overline{z}) \left(\delta_i (\frac{f'}{f})(M_i - b z_i) - (1-\delta_i)(\frac{f}{S})(M_i - b z_i) \right) \tag{3.1}$$

Since (under certain assumptions)

$$\int_{x>t} \frac{f'(x)}{f(x)} f(x) dx = \int_{x>t} df(x) = f(x)\Big|_t^\infty = -f(t)$$

as $f(\infty) = 0$, we have

$$\frac{f}{S}(t) = \frac{-\int_{x>t} \frac{f'(x)}{f(x)} f(x) dx}{S(t)} = \frac{-\int_{x>t} \frac{f'(x)}{f(x)} dF(x)}{S(t)} = \frac{\int_{x>t} \frac{f'(x)}{f(x)} dS(x)}{S(t)}$$

Thus

$$\Phi = \Phi(b, S, \frac{f'}{f}) = \sum_{i=1}^{n}(z_i - \overline{z})\left(\delta_i(\frac{f'}{f})(T_i(b)) - (1-\delta_i)\frac{\int_{x>T_i(b)}\frac{f'(x)}{f(x)}dS(x)}{S(T_i(b))}\right)$$

where $T_i(b) = M_i - bz_i$. Note that S, f and f' are all unknown. Thus they need to be estimated. If one replaces S by by its PLE and chooses $f = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ (thus $(f'/f)(x) = -x$), then

$$\Phi = -\sum_{i=1}^{n}(z_i - \overline{z})\left(\delta_i(T_i(b)) - (1-\delta_i)\frac{\int_{x>T_i(b)} x d\hat{S}_b(x)}{\hat{S}_b(T_i(b))}\right)$$

and the M-estimator reduces to the BJE. Thus the BJE is an M-estimator.

Zhang and Li suggest to look for a zero-crossing point of an estimate of $\Phi(b, S, \frac{f'}{f})$, say $\Phi(b, \hat{S}_b, \frac{\tilde{f}'_b}{\tilde{f}_b})$ where $\hat{S}_b$ is the PLE of S_o based on $(M_i - bz_i, \delta_i)$'s, and $\tilde{f}_b$ is a kernel estimator

$$\tilde{f}_b(t) = \frac{1}{h}\int K(\frac{x-t}{h})d\hat{F}_b(x), \text{ with } K \geq 0, \int K(x)dx = 1,$$

$$\text{the set } \{x : K'(x) \neq 0\} \text{ is not a null set, } n^{-1/2}/h \to 0$$

$$e.g., \ h = cn^{-1/5} \text{ and } c \text{ is a predetermined constant.} \tag{3.2}$$

Examples of such kernels are

$$K(x) = \frac{3}{4}(1-x^2)\mathbf{1}_{(|x|\leq 1)},$$

and

$$K(x) = (1-|x|)\mathbf{1}_{(|x|\leq 1)}.$$

Other examples can be found in Härdle W. (1990, p.45). It is obvious that an M-estimate can only be obtained by iterative algorithms. Zhang and Li show that under certain regularity conditions, the M-estimator is consistent and is asymptotically efficient if the initial point in the iterative algorithm is close enough to β. The asymptotic variance of the M-estimator is expected to be $(E(\Phi(\beta)\Phi'(\beta))^{-1}$ and can be estimated by $\hat{\Sigma} =$

$$\left(\sum_{i=1}^{n}\left(\delta_i[(\frac{\hat{f}'}{\hat{f}})(M_i - bz_i)]^2 - (1-\delta_i)(\frac{\hat{f}}{\hat{S}})^2(M_i - bz_i)\right)(z_i - \overline{z})(z_i - \overline{z})'\right)^{-1},$$

where $\hat{S}$, $\hat{f}$ and $\hat{f}'$ are estimates of S_o, f_o and f'_o, respectively.

However, in practice, there are several outstanding computational issues with this approach:

1. it is not clear how to select an initial point that is really "close" to β.
2. It is not clear how to select a constant c in (3.2).
3. It is not clear what is an optimal choice of the kernel K.
4. A similar phenomenon like the BJE, which is also a special case of M-estimators, may also occur. That is, there does not exist a zero point of Φ.
5. Even if one may obtain a zero point of Φ, it may not be the solution that is closed to the maximum point of the likelihood (*i.e.*, as point that is near a local maximum or even a local minimum of the likelihood. Consequently, the estimate is not good.

An algorithm for the case $p = 1$ maybe as follows under the assumption that $\Phi(-\infty)\Phi(\infty) < 0$.

1. Choose a tolerance number $\eta > 0$, *e.g.*, $\eta = 0.00001$.
2. Choose a $b_1 > 0$ and let $b_2 = -b_1$ such that $\Phi(b_1)\Phi(b_2) < 0$. If $\Phi(b_i) \in (-\eta, \eta)$ for an $i \in \{1, 2\}$, stop and let b_i be the "estimate" (**treat as a zero point of** Φ)). Otherwise, go to next step.
3. Let $b_3 = (b_2 + b_1)/2$. If $\Phi(b_3) \in (-\eta, \eta)$, stop and let b_3 be the "estimate". Otherwise, go to next step.

4. Set $b_{i+1} = \begin{cases} (b_{i-1}+b_i)/2 & \text{if } \Phi(b_{i-1})\Phi(b_i) < 0 \\ (b_{i-2}+b_i)/2 & \text{if } \Phi(b_{i-2})\Phi(b_i) < 0 \end{cases}$ for $i \geq 2$ iteratively until either $|b_i - b_{i+1}| < \eta$ (**zero point**), or $\Phi(b_{i+1}) \in (-\eta, \eta)$ (**zero-crossing**).

Remark. Verify that in the case of complete data, $\Phi = -2(\sum_{i=1}^n (z_i - \overline{z})(X_i - \overline{X}) - b\sum_{i=1}^n (z_i - \overline{z})^2)$. Thus $\Phi(-\infty) > 0$ and $\Phi(\infty) < 0$. One expects that the assumption $\Phi(-\infty)\Phi(\infty) < 0$ holds in general.

Another algorithm is to find a minimum point of $|\Phi(\mathbf{b})|$ by Monte Carlo method. That is, randomly select a sequence of values of $\mathbf{b}$ and find the up-to-date minimum point until it is stable.

§**5.4.3.2. Homework.**

1. Derive the expressions of Φ when f is the density of $U(0,\theta)$ and when $f(x) = e^{-x}$, $x > 0$.
2. Give the expression of Φ when data are mixed IC type.

§**5.4.4. An SMLE approach**

Yu and Wong (2003a,b,c) proposed the SMLE of (β, S_o), based on complete data, RC data and IC data, respectively. The semi-parametric likelihood function is $Ł = \prod_{i=1}^n \mu_F(I_i - bz_i)$, where F is a cdf, I_i's are the observed intervals and $I_i - c$ is a shift of the interval I_i by c units. Thus the SMLE of (β, F_o) is the value of (b, F) that maximizes $Ł(b, F)$.

Without loss of generality, we assume that the dimensions of b and $\mathbf{z}_i$ are 1. For fixed b, the likelihood function is maximized by the GMLE of F_o based on $I_i - bz_i$'s, denoted by $\hat{F}_b$. Thus it suffices to maximizes $Ł(b, \hat{F}_b)$. Since $\hat{F}_b$, as a function of b, only takes finitely many values, so is Ł. Let $a_1 < \cdots < a_m$ be all the solutions to equations of form $M_k - bz_k = M_j - bz_j$, $z_k \neq z_j$, $a_0 = -\infty$ and $a_{m+1} = \infty$. Note that the ranks of $L_i - bz_i$'s and $R_i - bz_i$'s are constant for $b \in (a_k, a_{k+1})$ and the GMLE $\hat{F}_b(L_j - bz_j)$ and $\hat{F}_b(R_j - bz_j)$'s only depends on the ranks of $L_i - bz_i$'s and $R_i - bz_i$'s. Thus for each fixed i, $Ł(b, \hat{F}_b)$ is constant in b on the interval (a_i, a_{i+1}).

Example 1. Consider a simple example, say,

3 (M_i, δ_i, z_i)s are $(1,1,1)$, $(2,0,1)$, $(3,1,0)$.

Let $T_i(b) = M_i - bz_i$. $M_k - bz_k = M_j - bz_j$ and $z_k \neq z_j$ yield

$1 - b = 3$ $(T_1(b) = T_3(b))$ and $2 - b = 3$ $(T_2(b) = T_3(b))$.

Thus $a_1 = -2$ and $a_2 = -1$.

Let (r_1, r_2, r_3) be the ranks of $(T_1, T_2, T_3)(b)$s (or $(1-b, 2-b, 3)$s).

$I_i(b)$s are $([T_1, T_1], (T_2, \infty), [T_3, T_3])$.

$b \in$	$(-\infty, -2)$	$\{-2\}$	$(-2, -1)$	$\{-1\}$	$(-1, \infty)$
$r_i s$	$(2,3,1)$	$(1.5, 3, 1.5)$	$(1,3,2)$	$(1,3,2)$	$(1,2,3)$
$\mu_{\hat{F}_b}(I_i(b))s$	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	$(\frac{2}{3}, \frac{1}{3}, \frac{2}{3})$	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	$(\frac{1}{3}, \frac{2}{3}, \frac{2}{3})$
$max Ł$	*no*	*yes*	*no*	*no*	*yes*

Each point in $\{-2\} \cup (-1, \infty)$ is an SMLE of β.

In the case of RC data with $p = 1$, there are finitely many such disjoint intervals and the PLE and Ł have explicit forms. Thus, the SMLE can be obtained by a non-iterative algorithm. Let $\mathcal{A}$ be the union of all a_i's, all midpoints $(a_i + a_{i+1})/2$, and $a_1 - 1$ and $a_m + 1$. A point in $\mathcal{A}$ that maximizes $Ł(b, \hat{F}_b)$ over all $b \in \mathcal{A}$ is an SMLE of β.

We summarize as an algorithm:

1. Derive $\mathcal{B} = \{b : \ b = \frac{M_i - M_j}{z_i - z_j}, z_i \neq z_j, 1 \leq i < j \leq n\}$. Let $a_1 < \cdots < a_m$ be the distinct elements of $\mathcal{B}$. Derive $\mathcal{A}$.
2. Derive $\mathcal{L}(b) = Ł(b, \hat{F}_b)$ for each $b \in \mathcal{A}$.
3. The maximizer of $\mathcal{L}(b)$, $b \in \mathcal{A}$ is an SMLE of β, denoted by $\hat{\beta}$. Moreover, if $a, \hat{\beta} \in (a_i, a_{i+1})$, then a is also an SMLE of β. The SMLE of S_o is $\hat{S}_{\hat{\beta}}$.

In the case of IC data, the GMLE may not have a closed form expression, one has to use iterative algorithm to obtain the GMLE and thus the SMLE may be obtained by an iterative algorithm.

The SMLE may not be unique, thus one can choose one that is the closest to the median of all SMLE's. Note that the median may not be an SMLE.

Properties of the SMLE with RC data:

1. If F_o is continuous and $P\{\delta = 1\} = 1$, there are inconsistent SMLE's and consistent SMLE's (see Yu and Wong (2003b)).

2. Simulation suggests that the SMLE and the BJE cannot dominate each other.
3. If F_o is discontinuous, then $\mathrm{P}\{\hat{\beta}_n \neq \beta \text{ infinitely often }\} = 0$ or $\mathrm{P}\{\hat{\beta}_n = \beta \text{ for large enough } n\} = 1$.

It is conjectured that the SMLE $\hat{\beta}$ (or $\hat{\beta}_n$) has the following properties:

4. If F_o is continuous and $\mathrm{P}\{\delta = 1\} \in (0,1)$, then the SMLE is consistent.
5. If F_o is continuous and $\mathrm{P}\{\delta = 1\} \in (0,1)$, then the SMLE is asymptotically normal, with estimated asymptotic covariance matrix

$$\hat{\Sigma} = \left(\sum_{i=1}^{n} (1-\delta_i) \left(\frac{\tilde{f}(T_i(\hat{\beta}))}{\hat{S}_{\hat{\beta}}(T_i(\hat{\beta}))} \right)^2 z_i z_i' \right)^{-1}, \tag{4.1}$$

where $\tilde{f}$ is a kernel estimate of the df of F_o. Note that

$$\ln(f^{\delta}(T(b))S^{1-\delta}(T(b))) = \delta \ln f(T(b)) + (1-\delta)\ln S(T(b))$$

$$\frac{\partial \ln(f^{\delta}(T(b))S^{1-\delta}(T(b)))}{\partial b} = [-\delta \frac{f_o'(T(b))}{f_o(T(b))} + (1-\delta)\frac{f_o(T(b))}{S_o(T(b))}]Z$$

Thus the first term is missing in $\hat{\Sigma}$.

6. If F_o is continuous, then the SMLE is not efficient. In fact, if $\hat{\Sigma}$ in (4.1) is true, this is obvious as the efficient covariance matrix is

$$E[(\delta \frac{f'(T(b))}{f(T(b))} + (1-\delta)\frac{f(T(b))}{S(T(b))})^2 ZZ']$$

We now verify property 1. Assume that $p = 1$, ϵ and $\mathbf{z}$ are both continuous random variables. Let (X_i, z_i), $i = 1, ..., n$ are observations. Then $\hat{\beta} = \frac{X_1 - X_2}{z_1 - z_2}$ is a continuous random variable. Let $\hat{\beta}_n = \max\{b : \ b \in \mathcal{B}\}$, then $\hat{\beta}_n \to \infty$ a.s., as $\min\{Z_1 - Z_2\} \to 0$ a.s.. That is, $\hat{\beta}_n$ is an inconsistent SMLE. On the other hand, the SMLE that is closest to the LSE is consistent.

We shall present some of the results in simulation studies. The main purpose is to study the properties of the SMLE when F_o is arbitrary, *i.e.*, continuous, or discontinuous but not necessarily discrete. We assume, for convenience, that Z, ϵ and Y are independent. We consider several cases in our simulation studies:

(1) F_o is continuous (Examples 4.3, 4.4 and 4.5), or discrete (Example 4.1), or discontinuous but not discrete (Examples 4.2).

(2) All the underlying distributions belong to the exponential family (Examples 4.3 and 4.5) or F_o does not belong to the exponential family (other examples).

Example 4.1. Suppose ϵ equals 0.5 and 25.5, with probabilities 0.5 and 0.5, respectively, $Z \sim U(2,3)$, $Y \sim U(24, 24.2)$, and $(\alpha, \beta) = (13, 1)$.

Example 4.2. Suppose ϵ is a mixture of $U(0, 0.5)$ and a unit point mass concentrated at 0.25, with probabilities 0.5 and 0.5, respectively, $Z \sim U(1,2)$, $Y \sim U(0,4)$, and $(\alpha, \beta) = (0.25, 1)$.

Hereafter denote $Exp(\mu, \sigma)$ a distribution with the df

$f(x) = \frac{1}{\sigma} e^{-[\frac{x-\mu}{\sigma}+1]} \mathbf{1}_{(x>\mu-\sigma)}$.

Example 4.3. Suppose ϵ, Y and $\mathbf{Z}$ have distributions $Exp(5,2)$, $Exp(3,4)$ and $Exp(2,2)$, respectively. $(\alpha, \beta) = (5, 1)$.

Example 4.4. Suppose $\epsilon \sim U(-1,1)$, $Z \sim Exp(0,19)$, $Y \sim Exp(0,25)$, $(\alpha, \beta) = (0,2)$.

Example 4.5. Suppose $\epsilon \sim N(0,1)$, $Y \sim N(0,6)$, $Z \sim N(2,1)$, $(\alpha, \beta) = (0,1)$.

The results of the above examples are summarized in Table 1. One can see that the SMLE is better than the BJE under the exponential distribution, but vice verse under the normal distribution.

Property 3 is proved for the following cases (1) complete-data and (2) right-censored discrete data (Yu and Wong (2003a,b)). The proof for right-censored discontinuous data is under preparing. The following example illustrates a proof under a simple assumption.

Example 4.6. Consider the simple linear regression ($p = 1$), with complete data. Suppose that ϵ and Z have the binomial distribution, $Bin(1, 0.5)$ and $\beta = 1$. The possible values of the observation (Z, X) are $(0,0)$, $(0,1)$, $(1,1)$ and $(1,2)$, denoted by (Z_i, X_i), $i = 1, 2, 3, 4$. Thus there are 4 possible values of $T(b)$,

say, $T_i(b) = X_i - bZ_i$, $i = 1, ..., 4$. Suppose a random sample of size n contains N_1, N_2, N_3 and N_4 of them. The empirical df

$$\hat{f}_b(T_i(b)) = \begin{cases} \frac{N_i}{n} & \text{if } T_i(b) \neq T_j(b) \ \forall \ j \neq i, \\ \frac{N_i+N_j}{n} & \text{if } T_i(b) = T_j(b) \text{ for only one } j \neq i, \end{cases} \tag{4.2}$$

where $i, j = 1, 2, 3, 4$. The possible finite solutions b to the equations $T_i(b) = T_j(b)$ are 0, 1 and 2.

For n large enough, $N_i \approx n/4$, and likelihood function (1.2) or (4.1) $\mathcal{L} = \prod_{i=1}^{n} \hat{f}_b(T_i(b)) =$

$$\begin{cases} (\frac{N_1+N_3}{n})^{N_1+N_3}(\frac{N_2+N_4}{n})^{N_2+N_4} \approx (0.5)^n & \text{if } b = 1, \\ (\frac{N_1}{n})^{N_1}(\frac{N_2+N_3}{n})^{N_2+N_3}(\frac{N_4}{n})^{N_4} \approx (0.5)^{n/2}(0.25)^{n/2} & \text{if } b = 0, \\ (\frac{N_1+N_4}{n})^{N_1+N_4}(\frac{N_2}{n})^{N_2}(\frac{N_3}{n})^{N_3} \approx (0.5)^{n/2}(0.25)^{n/2} & \text{if } b = 2, \\ (\frac{N_1}{n})^{N_1}(\frac{N_2}{n})^{N_2}(\frac{N_3}{n})^{N_3}(\frac{N_4}{n})^{N_4} \approx (0.25)^n & \text{otherwise,} \end{cases} \tag{4.3}$$

is maximized by $b = 1$. Thus the SMLE of β is $\tilde{\beta} = 1 = \beta$ for all large n (*i.e.*, (1.3) holds). The LSE $\tilde{\beta}_{LSE} = \beta + \frac{N_1N_4 - N_2N_3}{(N_1+N_4)(N_2+N_3)}$. Thus,

$\mathrm{P}(\tilde{\beta}_{LSE} \neq \beta \text{ i.o.}) = 1$.

The LSE $\tilde{\beta}_{LSE}$ satisfies

$\sqrt{n}(\tilde{\beta}_{LSE} - \beta) \xrightarrow{D} N(0, 1)$, and $n\sigma^2_{\tilde{\beta}_{LSE}} \to 1$ as $n \to \infty$,

as the asymptotic variance $\sigma^2_{\tilde{\beta}_{LSE}} = \frac{1}{n}$.

It can be shown that $\sqrt{n}(\tilde{\beta} - \beta) \xrightarrow{D} N(0, 0)$ and $n\sigma^2_{\tilde{\beta}} \to 0$. □

Table 1. Simulation Results on estimating (α, β)

		(α, β)	SMLE $(\hat{\alpha}, \tilde{\beta})$	BJE
Example 4.1 (discrete F_o).				
n=32	average	(13, 1)	(0.941, 1.000)	(-5.388, 1.749)
	SE		(0.044, 0.000)	(22.368, 6.702)
n=200	average	(13, 1)	(0.360, 1.000)	(1.388, 0.301)
	SE		(0.000, 0.000)	(1.075, 0.255)
Example 4.2 (discontinuous F_o).				
n=32	average	(0.25, 1)	(0.248, 1.001)	(0.258, 0.997)
	SE		(0.038, 0.021)	(0.142, 0.961)
n=200	average	(0.25, 1)	(0.250, 1.000)	(0.247, 1.002)
	SE		(0.010, 0.000)	(0.050, 0.033)
Example 4.3 (continuous F_o).				
n=32	average	(5, 1)	(3.693, 1.415)	(4.035, 0.884)
	SE		(3.784, 3.227)	(2.106, 1.708)
n=200	average	(5, 1)	(4.701, 0.957)	(4.701, 0.959)
	SE		(1.294, 0.693)	(0.798, 0.753)
Example 4.4 (continuous F_o).				
n=32	average	(0, 2)	(1.446, 2.129)	(0.008, 2.001)
	SE		(5.411, 0.464)	(0.291, 0.022)
n=200	average	(0, 2)	(0.821, 2.077)	(-0.002, 2.000)
	SE		(0.876, 0.080)	(0.093, 0.007)
Example 4.5 (continuous F_o).				
n=32	average	(0, 1)	(-0.2093, 1.1651)	(-0.0110, 1.0145)
	SE		(1.5933, 0.8971)	(0.7867, 0.3814)
n=200	average	(0, 1)	(-0.2598, 1.1516)	(0.0048, 0.9942)
	SE		(0.7068, 0.3831)	(0.2350, 0.1116)

Example 4.7. (Magazine advertising (Chatterjee and Price ((22), p. 257)). In a study of revenue from advertising, data were collected for 41 magazines in 1986. There was no censoring. Let Z denote the number of pages of advertising and X the advertising revenue.

The 41 data are plotted in Figure 1. Roughly speaking, there are three outliers in the data set. They are (25, 50), (15, 49.7), (77, 6.6). The SMLE of β is unique for this data set. The SMLE and the LSE are significantly different (see the first block of Table 2). The entries in the second block of Table 2 are results after deleting the three outliers. From Table 2, it is seen that the SMLE of β does not change after deleting outliers, though the estimate of α changes.

In Figure 1, we also plot the fitted straight lines with and without deleting those three outliers. We further plot the fitted line by the SMLE method without deleting the outliers. From Figure 1, it is seen that the fitted line by the SMLE approach using the original data is very close to the least squares fitted line after deleting outliers. This suggests that the SMLE is robust while the LSE is not.

Fig. 1. SMLE v.s. LSE (Journal Data)

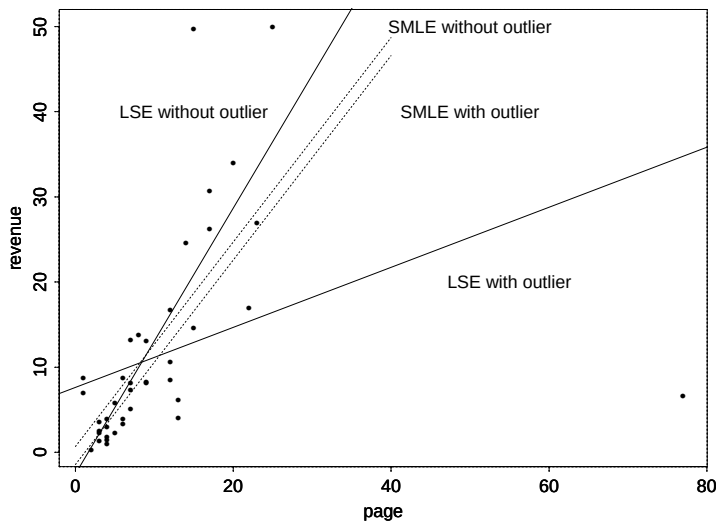


Table 2. Results on estimating (α, β)

		SMLE (SE)	LSE (SE)
with outliers	β	1.200 (0.196)	0.353 (0.1449)
	α	-1.427 (3.178)	7.604 (2.3466)
without outliers	β	1.200 (0.1379)	1.238 (0.138)
	α	-0.642 (1.410)	-0.962 (1.409)

Consistency is proved for the mixed case IC model (Yu and Wong (2006)) and is under preparation for the RC model. The latter case is quite complicated. Let $\hat{f}_{\mathbf{b}}(t) = \hat{S}_{\mathbf{b}}(t-) - \hat{S}_{\mathbf{b}}(t)$. If ϵ is continuous and there is no censoring, then

$$\hat{f}_{\mathbf{b}}(T_i(\mathbf{b})) = 1/n \ \forall\ i, \text{ except perhaps for two, at which } \hat{f}_{\mathbf{b}}(T_i(\mathbf{b})) = 2/n. \tag{2.2}$$

Consequently,

$$\text{Ł}(\hat{S}_{\hat{\beta}}, \hat{\beta}) = 2^2/n^n \text{ and } \text{Ł}(\hat{S}_{\beta}, \mathbf{b}) = 1/n^n \text{ if } \mathbf{b} \text{ is not an SMLE}$$

$$|\frac{1}{n}\ln\text{Ł}(\hat{S}_{\hat{\beta}}, \hat{\beta}) - \frac{1}{n}\ln\text{Ł}(\hat{S}_{\mathbf{b}}, \mathbf{b})| = \begin{cases} \frac{1}{n}\ln 4 & \text{if } \mathbf{b} \text{ is not an SMLE} \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

In other words, if ϵ is continuous, the inconsistent SMLE $\hat{\beta}$ also satisfies that

$$|\frac{1}{n}\ln\text{Ł}(\hat{S}_{\hat{\beta}}, \hat{\beta}) - \frac{1}{n}\ln\text{Ł}(\hat{S}_{\beta}, \beta)| \to 0$$

In a consistency proof, one may want to establish

$$\lim_{n\to\infty} \big[\frac{1}{n}\ln\text{Ł}(\hat{S}_{\hat{\beta}}, \hat{\beta}) - E\{\frac{1}{n}\ln L(S_o, \beta)\}\big] = 0 \ a.s..$$

However, it does not make sense here. because $E\{\frac{1}{n}\ln L(S_o,\beta)\} = -\infty$. We consider a modification of the above equality in the proof.

Property 5 is still very difficult to prove. We find that a variant of the SMLE with the RC data has similar properties with the SMLE. The estimator is a value of $\mathbf{b}$ that maximizes $\prod_{i=1}^n (\hat{S}_{\mathbf{b}}(T_i(\mathbf{b}))^{1-\delta_i}$ instead of $\prod_{i=1}^n (\hat{f}_{\mathbf{b}}(T_i(\mathbf{b}))^{\delta_i} (\hat{S}_{\mathbf{b}}(T_i(\mathbf{b}))^{1-\delta_i}$, and is called the partial likelihood SMLE (PSMLE). Table 3 presents simulation studies when ϵ, Y and Z $(= \mathbf{Z})$ have distributions $Exp(3,1)$, $Exp(1,1)$ and $Exp(0,1)$, respectively. $(\alpha,\beta) = (3,1)$.

Table 3. Simulation Results on estimating (α,β)

n		β $(\sigma_{\hat{\beta}})$	SMLE(SE)	PSMLE(SE)	BJE(SE)
200	sample mean	1	1.06	1.18	0.58
	sample SE	(0.33)	(0.47)	(1.47)	(1.23)
	est. of SE		0.48	0.51	
800	sample mean	1	1.00	1.01	0.86
	sample SE	(0.17)	(0.20)	(0.22)	(0.64)
	est. of SE		0.20	0.20	
1000	sample mean	1	1.02	1.03	0.91
	sample SE	(0.15)	(0.16)	(0.17)	(0.39)
	est. of SE		0.17	0.18	

Remark. Under the semiparametric model, $X = \beta'\mathbf{Z}+\epsilon$, the location parameter $\alpha = E(\epsilon)$ is not identifiable under censoring. This is the major reason why people do not consider the model $X = \alpha + \beta'\mathbf{Z} + e$, where $E(e) = 0$. In fact, if b is fixed, the likelihood function $\prod_{i=1}^n \mu_{\hat{F}_{a,b}}(I_i - a - b'\mathbf{z}_i)$ is constant in $a \in R$, where $\hat{F}_{a,b}$ is the GMLE of F_o based on $I_i - a - b'\mathbf{z}_i$'s

§5.4.4.2. Homework.

1. There are 4 observations $(M_i,\delta_i,\mathbf{z}_i)$'s: $(3,1,2),(4,0,1),(1,1,1),(7,1,2)$. Find the SMLE and PSMLE of β under the linear regression model.

§5.4.5. A modified SMLE approach. It is a little bit disappointed that the SMLE approach is not efficient when F_o is continuous, though it is super efficient if F_o is discontinuous. It may due to the reason that $f(t) = F(t) - F(t-)$ in the nonparametric likelihood.

A modification is to replace f in Ł by a smooth version, a kernel estimator of f_o, *i.e.*,

$$f(t) = \frac{1}{h}\int K((x-t)/h)dF(x),$$

where $K(\cdot)$ is a kernel. However, for most versions, it is difficult to find directly the maximum point of Ł, one can only find a critical point of Ł, or a root to $\frac{\partial \ln Ł}{\partial b}$. Such approach is called M-estimation approach, which requires that $\{x : K'(x) \neq 0\}$ is not a null set.

Yu and Wong (2005) propose a different modification. Let

$$f_F(x) = \int_{-\infty}^{\infty} \frac{1}{h_n} K(\frac{x-t}{h_n}) dF(t),\ K(x) = \tfrac{1}{2}\mathbf{1}_{(-1<x\leq 1)},\ h_n > 0,\ \lim_{n\to\infty} h_n = 0 \tag{5.1}$$

(*e.g.*, $h_n = O(n^{-1/5})$, as suggested in Härdle (1990, p.59 or p.91)), and $\hat{F}_{\mathbf{b}} = 1 - \hat{S}_{\mathbf{b}}$. Then $Ł(\hat{S}_{\mathbf{b}},\mathbf{b},f_{\hat{F}_{\mathbf{b}}}) =$

$$\prod_{i=1}^n \big[(\frac{1}{2h_n}[\hat{S}_{\mathbf{b}}(T_i(\mathbf{b}) - h_n) - \hat{S}_{\mathbf{b}}(T_i(\mathbf{b}) + h_n)])^{\delta_i} (\hat{S}_{\mathbf{b}}(T_i(\mathbf{b})))^{1-\delta_i}\big], \tag{5.2}$$

where $S(t-) = \lim_{x\uparrow t} S(x)$. We propose to estimate β by $\hat{\beta}$ which maximizes $Ł(\hat{S}_{\mathbf{b}},\mathbf{b},f_{\hat{F}_{\mathbf{b}}})$ over $\mathbf{b} \in \mathcal{R}^p$. We call $\hat{\beta}$ the MSMLE of β. Then $\hat{S}_{\hat{\beta}}(t)$ is an MSMLE of $S_o(t)$. Let $\alpha = E(\epsilon)$. α may not exist. Even if it

does, it is well known (see Buckley and James (1979)) that there is no consistent estimator of α under right censoring, unless some further assumptions are made. Nevertheless, a natural estimator of α is

$$\hat{\alpha} = a(\hat{\beta}), \text{ where } a(\hat{\beta}) = \frac{\int_{x \le T_{(n)}(\hat{\beta})} x d\hat{F}_{\hat{\beta}}(x)}{\int_{x \le T_{(n)}(\hat{\beta})} 1 d\hat{F}_{\hat{\beta}}(x)}, \tag{5.3}$$

where $T_{(1)} \le \cdots \le T_{(n)}$ are order statistics of T_i's (this is a least squares estimator). Though the MSMLE is motivated for continuous F_o, it has some nice properties for arbitrary F_o (continuous or discontinuous). Since $\frac{1}{2h_n}$ in (5.2) does not depend on $\mathbf{b}$, it suffices to maximize for $\mathbf{b} \in \mathcal{R}^p$,

$$l(\mathbf{b}) = \prod_{i=1}^{n} \big[[\hat{S}_{\mathbf{b}}(T_i(\mathbf{b}) - h_n) - \hat{S}_{\mathbf{b}}(T_i(\mathbf{b}) + h_n)]^{\delta_i} (\hat{S}_{\mathbf{b}}(T_i(\mathbf{b})))^{1-\delta_i}\big]. \tag{5.4}$$

For a similar argument as for the SMLE, the new likelihood function takes on finitely many values and is constant on intervals of the form (a_i, a_{i+1}), where $a_1 < \cdots < a_m$ are all the values of $b = \frac{M_i - M_j + kh_n}{z_i - z_j}$, $z_i \ne z_j$, $k = 0, \pm 1, \pm 2$. The argument is similar to that for the SMLE. Thus it can be obtained by a non-iterative algorithm. We need the following notation.

Let $\mathcal{O}$ be the collection of all points L_i or R_i, $i = 1, ..., n$,

where $(L_i, R_i) = \begin{cases} (M_i - h_n, M_i + h_n) & \text{if } \delta = 1 \\ (M_i, \infty) & \text{if } \delta_i = 0. \end{cases}$

Let $\mathcal{A}$ be the set of points $b = \frac{O_i - O_j}{z_i - z_j}$, where $O_i = L_i$, or R_i (if $\ne \infty$), and $z_i \ne z_j$, $i < j$.

Let $\mathcal{A}_1$ be the set of distinct elements of $\mathcal{A}$. Order them as $a_1 < \cdots < a_m$;

Let $a_0 = -\infty$ and $a_{m+1} = \infty$;

Let $\mathcal{A}_2$ be the set consisting of $a_1 - 1$, $a_m + 1$ and points $\frac{a_{i-1} + a_i}{2}$, $i = 2, ..., m$.

Non-iterative algorithm:

1. Obtain $\mathcal{O}$, $\mathcal{A}$, $\mathcal{A}_1$ and $\mathcal{A}_2$;
2. Compute $l(b)$ (see (5.4)) for each $b \in \mathcal{A}_1 \cup \mathcal{A}_2$.
3. Each b that maximizes $l(b)$ over $b \in \mathcal{A}_1 \cup \mathcal{A}_2$ is an MSMLE of β. Moreover, if b is an MSMLE and $b \in (a_i, a_{i+1})$, then each point in (a_i, a_{i+1}) is also an MSMLE.

Since $\mathcal{A}_1 \cup \mathcal{A}_2$ is finite and $l(\mathbf{b})$ has a closed-form expression, the algorithm is non-iterative.

The MSMLE $\hat{\beta}$ is consistent under certain regulation conditions (under preparation). Simulation suggests that it is also efficient. In particular, it suggests that if F_o is continuous, it attains the efficient lower bound of the variance, namely, $-\left(\frac{\partial \log \mathbf{L}}{\partial \beta \partial \beta'}\right)^{-1}$. If F_o is discontinuous, then $\mathrm{P}\{\hat{\beta} \ne \beta \text{ for large enough } n\} = 0$.

The following are simulation results supporting the above conjectures. In our simulation, for convenience, we assume that ϵ, $\mathbf{Z}$ and Y are independent.

We compare $\hat{\beta}$ to the BJE in several cases as follows.

(A) F_o is continuous (Examples 5.1, 5.3, 5.5 - 5.7), or is neither discrete nor continuous (Example 5.2).

(B) F_o is continuous but the regularity conditions in the Cramer-Rao theorem do not hold (Examples 5.1, 5.5 and 5.6), or all underlying distributions are exponential distributions so that they allow exchange of differentiation and integration (Example 5.7), or F_o is a normal distribution function (Examples 5.3 and 5.5).

(C) There is no censoring (Examples 5.1- 5.3), or there is censoring (Examples 5.4-5.7).

In our simulation, for each case, we repeated 1000 times and computed the sample mean and sample standard error (SE) of the 1000 estimates.

Example 5.1. Suppose $\epsilon \sim U(-1, 1)$ (the uniform distribution), $Z \sim U(0, 9)$ and $(\alpha, \beta) = (0, 2)$.

Example 5.2. Suppose ϵ is a mixture of $U(0, 0.9)$ and a constant 0.45, with probabilities (w.p.) 0.9 and 0.1, respectively, $Z \sim U(1, 2)$ and $(\alpha, \beta) = (0.45, 1)$.

Example 5.3. Suppose $\epsilon \sim N(0, 0.09)$, $Z \sim U(0, 9)$ and $(\alpha, \beta) = (0, 1)$.

The results of the above examples are summarized in Table 1. In the two rows of each block of Table 1, we present the sample averages and sample standard errors (SE) of the MSMLE and the BJE. It is seen that $\hat{\beta}$ dominates the LSE in the sense that $SE_{\hat{\beta}} \le SE_{LSE}$ in general, and $SE_{\hat{\beta}} < SE_{LSE}$ unless

$\epsilon \sim N(\mu, \sigma^2)$, provided $n \geq 200$. In the next 4 cases, there are right-censored data. Define $Y^c = Y - \beta'\mathbf{Z}$ and $\tau = \sup\{t : \ P(Y^c < t) < 1\}$.

Example 5.4. Suppose ϵ is a mixture of $U(0, 0.5)$ and 51, w.p. 0.5 and 0.5, respectively, $Z \sim U(1, 2)$, $Y \sim U(4, 4.1)$, and $(\alpha, \beta) = (25.625, 1)$.

Example 5.5. Suppose $\epsilon \sim N(0, 0.09)$, $Z \sim U(0, 9)$ and Y equals 0.5 and 39 w.p. 0.5 and 0.5, respectively. $(\alpha, \beta) = (0, 1)$.

Example 5.6. Suppose $\epsilon \sim U(-1, 1)$, $Z \sim Exp(0, 19)$, $Y \sim Exp(0, 25)$, $(\alpha, \beta) = (0, 2)$.

Example 5.7. Suppose ϵ, Y and $\mathbf{Z}$ have distributions $Exp(5, 2)$, $Exp(3, 4)$ and $Exp(2, 2)$, respectively. $(\alpha, \beta) = (5, 1)$.

Table 1. Simulation Results on estimating β without censoring.

	SMLE	parameter β	LSE	MSMLE $\hat{\beta}$
Example 5.1. (continuous F_o)				
n=32	Sample mean	2	1.996	1.993
	SE		0.040	0.045
n=200	Sample mean	2	1.999	1.998
	SE		0.016	0.014
Example 5.2. (discontinuous but not discrete F_o)				
n=32	1.001	1	1.000	1.003
	0.099		0.156	0.121
n=200	1.000	1	0.997	1.000
	0.000		0.060	0.000
Example 5.3. ($N(\mu, \sigma^2)$)				
n=32	Sample mean	1	1.000	0.998
	SE		0.022	0.025
n=200	Sample mean	1	1.0000	1.0000
	SE		0.0083	0.0083

Table 2. Simulation Results on estimating β with censoring.

	β	SMLE (SE)	BJE (SE)	MSMLE (SE) $\hat{\beta}$
Example 5.4 (discontinuous).				$F_o(\tau) < 1$.
n=32	1		0.290	0.981
			(0.700)	(0.181)
n=200	1		0.750	1.002
			(0.226)	(0.061)
Example 5.5 ($N(\mu, \sigma^2)$).				$F_o(\tau-) = 1$.
n=32	1		1.000	0.995
			(0.030)	(0.042)
n=200	1		1.000	0.994
			(0.011)	(0.013)
Example 5.6 (continuous).				$F_o(\tau-) = 1$.
n=32	2		2.001	1.999
			(0.022)	(0.029)
n=200	2	2.077	2.000	2.000
		(0.080)	(0.007)	(0.006)
Example 5.7 (Exp(μ, σ)).				$F_o(\tau-) = 1$.
n=32	1		0.930	1.298
			(1.565)	(1.832)
n=200	1	1.012	0.957	1.004
		(0.693)	(0.751)	(0.274)

The simulation results of Examples 5.4 - 5.7 are summarized in Table 2. In Table 3, we compare the sample variance of the MSMLE to the ELB under the exponential distribution (Example 5.7). We do not compare $\hat{\beta}$ to the ELB in other examples, as the ELB is not valid in Examples 5.1, 5.2, 5.4 and 5.6, and as $\tilde{\beta}_{BJE}$ is efficient in Examples 5.3 and 5.5. In Table 4, we give the empirical relative efficiency of $\hat{\beta}$ to the BJE, based on results not necessarily in Tables 3 and 4. The following are main observations from our simulation.

(1) All the 7 examples suggest that the MSMLE $\hat{\beta}$ is consistent, as the values of β are all within 2 SE's from the sample means and the SE's are decreasing in n.

(2) The results suggest that unless F_o is a normal distribution, in general the MSMLE $\hat{\beta}$ is asymptotically more efficient than the BJE as the SE's of $\hat{\beta}$ are uniformly smaller than those of the BJE when sample sizes are large in all but Examples 5.3 and 5.5, and the 7 examples include different types of distributions specified in cases (A), (B) and (C).

(3) If F_o is discontinuous, then $\mathrm{SE}_{\hat{\beta}} = 0$ for large sample sizes, while the SE of the BJE never equals 0. It suggests that (5.1) holds when F_o is neither discrete nor continuous, rather than only when F_o is discrete. Here we shall give a heuristic explanation as follows. For simplicity, let $h_n = 0$ in (2.4), and consider the case of complete data in Example 5.2. Then (2.4) becomes $l(b) = \prod_{i=1}^{n} \hat{f}_b(T_i(b))$, where $\hat{f}_b$ is the empirical density based on $T_i(b)$'s. Note $T_i(b) = M_i - bZ_i$ and $\epsilon_i = T_i(\beta)$. Let $n_1 = \sum_{i=1}^{n} \mathbf{1}_{(\epsilon_i = 0.45)}$. If n is large enough, one expects that $n_1 \geq 10$. If $b = \beta$, then there are n_1 $T_i(b)$'s that equal 0.45, thus $l(\beta) = \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{1}{n}\right)^{n-n_1}$. If $b \neq \beta$, then one expects that $T_i(b)$'s are all distinct and thus $l(b) = \left(\frac{1}{n}\right)^{n}$. Now it is easy to see that $b = \beta$ is the MSMLE of β when n is large.

(4) Simulation results in Examples 5.3 and 5.7 further suggest that the MSMLE is efficient. In Example 5.3, data are not censored and $\epsilon \sim N(\mu, \sigma^2)$. Thus the LSE is efficient. Since $SE^2_{\hat{\beta}}/SE^2_{\hat{\beta}_{LSE}} = 1$ (see Table 1), the MSMLE should also be efficient. In example 5.7, the efficient lower bound (ELB) of $\hat{\beta}$ is $Var(\epsilon)/(n \cdot Var(Z)) = 2.5^2/n$ (see (3.1)). It is seen from Table 3 that when $n = 800$, the MSMLE practically attains the ELB. Note that, in Table 3, $\hat{\sigma}^2_{\hat{\beta}}$ stands for the sample variance in the simulation.

Table 3. Comparison between the SE of the MSMLE and the ELB

n:	32	100	200	400	800	$\sqrt{n \cdot ELB}$
$\sqrt{n}\hat{\sigma}_{\hat{\beta}}$:	10.363	5.707	3.875	3.132	2.503 $\cdots$	2.5

(5) It is well known that if $\epsilon \sim N(\mu, \sigma^2)$ such as in Examples 5.3 and 5.5, the BJE is efficient. From Table 2, we note that the BJE is still better than $\hat{\beta}$ when $n = 200$ in Example 5.5, while $SE_{\hat{\beta}} = SE_{\tilde{\beta}_{BJE}}$ in Example 5.3. The results have two opposite interpretations: (5.a) $\hat{\beta}$ with right-censored data is not efficient, (5.b) $\hat{\beta}$ may be efficient but the sample size is not large enough. In fact, from our simulation results, $\frac{SE^2_{\hat{\beta}}}{SE^2_{\tilde{\beta}_{BJE}}} = 1.95$, 1.40, 1.23, 1.16 for $n = 32, 200, 300, 400$, respectively, in Example 5.5 and $\frac{SE^2_{\tilde{\beta}_{BJE}}}{SE^2_{\hat{\beta}}} = 1$ in Example 5.3. Thus (5.b) is a more logical explanation. If so, it also suggests that $\hat{\beta}$ is efficient in general.

Table 4. Estimates of the relative efficiency of $\hat{\beta}$ to the BJE.

Example:	5.1	5.2	5.3	5.4	5.5	5.6	5.7
$\frac{\hat{\sigma}^2_{\tilde{\beta}_{BJE}}}{\hat{\sigma}^2_{\hat{\beta}}}$:	1.3	∞	1.0	3.9	?	1.8	3.7
F_o:	*unif.*	*mixture*	*normal*	*mixture*	*normal*	*unif.*	*expon.*

See Observation (5) above for "?" in Table 4.

Example 5.8. (The Stanford heart transplant data). The data and detailed description can be found in Miller (1981, p.156). In this data, right-censored survival time, indicator of death, and five covariates including age of the recipient at the time of transplant were recorded. $n = 69$. For illustrated purpose, several methods are compared using the logarithm of time until death against age. The a priori guess H_1 under the AL model would be that younger patients fare better, that is $H_0: \beta < 0$. In Table 5, we compare the Miller estimator,

the BJE and the Cox procedure to the MSMLE. Note that the Cox model is $P(X > t|Z = z) = (S(t))^{e^{bz}}$, where S is a baseline survival function. Thus we expect $H_0 : \ b > 0$ rather than $\beta < 0$ as in the simple linear regression model.

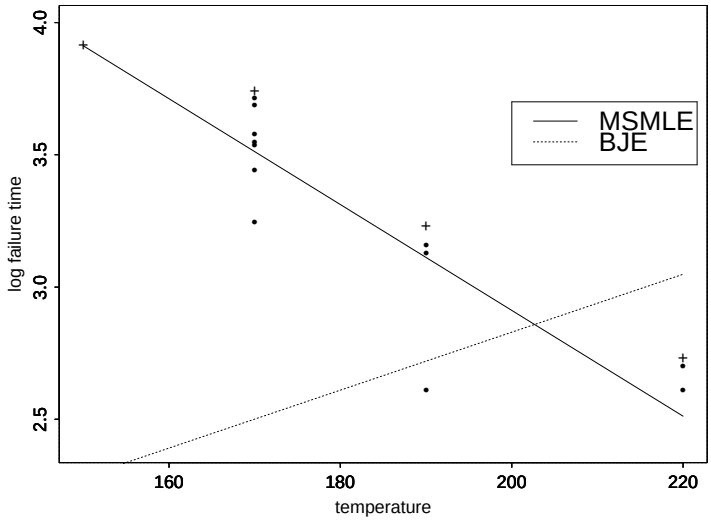
The entries in Table 5 related to the Miller estimator and the Cox procedure as well as their SE's are taken from Miller (1981, p.156). As commented by Miller (1981, p162), "The Cox method indicates there is a highly significant age effect. The Miller method says there is no effect due to age." The three BJE's in Table 2 basically suggest that there is no effect due to age. On the other hand, there is a unique MSMLE and is significantly negative and confirms with both the a priori guess and the Cox procedure. For this data set, taking $h_n = n^{-1/5}$ yields $\hat{\beta} = 0$, which does not lead to a satisfactory estimate. We choose $h_n = 3n^{-1/5}$.

Table 5. Regression analysis on the Stanford heart transplant data
age at transplant v.s. all death

	Miller (SE)	BJE (SE)	MSMLE (SE)	Cox (SE)
H_0 :	$\beta < 0$	$\beta < 0$	$\beta < 0$	$b > 0$
	-0.006 0.004 (0.017) 0.002 (0.016)	-0.028 (0.015)	-0.036 (0.017)	0.058 (0.023)

Example 2.1 of the section on BJE (Insulation data (Nelson 1973)). In this data there is unique MSMLE, which is -0.018, very close to the non-root zero-crossing point BJE solutions -0.02. Thus it is a reasonable estimator. The SMLE is -0.0416. It is also quite consistent with the trend.

Fig. 5.1. MSMLE vs. BJE For Insulation Data



§**5.4.5.2. Homework.**

1. There are 4 observations $(M_i, \delta_i, \mathbf{z}_i)$'s: $(3, 1, 2), (4, 0, 1), (1, 1, 1), (7, 1, 2)$. Find the MSMLE of β under the linear regression model.
2. Prove that the likelihood in (5.2) is a constant on the interval (a_i, a_{i+1}) as defined above based on the data in problem 1, thus there are only finitely many values.

References

* Finkelstein, D.M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* 42 845-854.
* Härdle W. (1990). Smoothing techniques, with implementation in S. *Springer-Verlag*, N.Y.
* Huber, P.J. (1964). Robust estimation of a location parameter". *Ann. Math. Statist.* 35, 73-101.
* Lai, T.L. and Ying, Z.L. (1991). Large sample theory of a modified Buckley-James estimator for regression-analysis with censored data. *Ann. Stat.* 19 1370-1402.
* Li, L.L. and Pu, Z.W. (1999). Regression models with arbitrarily interval-censored observations. *Comm. in Statist., Theory and Methods* 1547-1563.

* Rabinowitz, D. Tsiatis, A. and Aragon, J. (1995). Regression with interval-censored data. *Biometrika*, 82, 501-513.
* Yu, Q.Q. and Wong, G.Y.C. (2002a). How to find all Buckley-James estimates instead of just one ? *Journal of Statistical Computation and Simulation*, 72, 451-460.
* Yu, Q.Q. and Wong, G.Y.C. (2002b). Asymptotic properties of a modified semi-parametric MLE in linear regression analysis with right-censored data. *Acta Mathematica Sinica*, 18 405-416.
* Yu, Q.Q. and Wong, G.Y.C. (2003a). The semi-parametric MLE in linear regression with right-censored data. *Journal of Statistical Computation and Simulation*, 73 833-843.
* Yu, Q.Q. and Wong, G.Y.C. (2003b). Asymptotic properties of the generalized semi-parametric MLE in linear regression. *Statistica Sinica*, 13 311-325.
* Zhang, C.H. and Li, X. (1996). Linear regression with doubly censored data. *Ann. Statist.*, 24, 2720-2743.
* Yu, Q.Q. and Wong, G.Y.C. (2003c). Semi-parametric MLE in simple linear regression with interval-censored data. *Communications in Statistics-Simulation and Computation*, 32 147-164.
* Yu, Q.Q., Wong, G.Y.C. and Kong, F.H.(2006). Consistency of the semi-parametric MLE in linear regression with interval-censored data. Scan. J. Statist. 33 367-378.
* Yu, Q.Q. and Wong, G.Y.C. (2005). A modified semi-parametric MLE in linear regression analysis with complete data or right-censored data. *Technometrics*, 47 34-42.
* Kong, F.H. and Yu, Q.Q. (2006). Asymptotic distribution of the Buckley-James estimator under non-standard conditions *Statistica Sinica* (in press).
* Yu, Q.Q and Wong, G.Y.C. (2006) A data set that there are two BJEs. http://www.math.binghamton.edu/qyu.

Chapter 6. Testing (large sample tests)

There are two basic approaches in constructing a test:
(1) MLE approach: If $\hat\beta$ is the MLE, then under certain assumption, a test for H_0: $\beta=\beta_o$ is $\phi=\mathbf{1}((\hat\beta-\beta_o)'J(\hat\beta-\beta_o)|_{\beta=\hat\beta}$ or $_{\beta_o}>\chi^2_{\alpha,p})$, where $J=-\frac{\partial^2 \ln \mathbf{L}}{\partial\beta\partial\beta'}$, p is the dimension of β and $\chi^2_{\alpha,p}$ is the $(1-\alpha)100$-th percentile of the χ^2 distribution with degree freedom p.
(2) Score test approach: Under certain assumptions, a test for H_0: $\beta=\beta_o$ is $\phi=\mathbf{1}(U(\beta)'J^{-1}U(\beta)|_{\beta=\beta_o}>\chi^2_{\alpha,p})$, where $U(\beta)=\frac{\partial \ln \mathbf{L}}{\partial\beta}$.

§**6.1. One sample nonparametric test**

Hereafter denote $\hat F$ (or $\hat S$) the GMLE of F_o (or S_o).

(1) Two-sided level-α test for H_0: $S_o(t_0)=p_0$, where p_0 is known:
For RC data, under the RC model, $\phi=\mathbf{1}_{(\frac{|\hat S_{pl}(t_0)-p_0|}{\hat\sigma_{\hat S_{pl}(t_0)}}\ge z_{\alpha/2})}$,

where $\hat\sigma^2_{\hat S_{pl}(t_0)}$ is the estimate of

$$\sigma^2_{\hat S_{pl}(t)}\approx S_o(t)^2\int_0^t\frac{1}{S_o(x-)S_Y(x-)S_o(x)}dF_o(x)/n.$$

For IC data, the situation varies. If there are exact observations, in general, a test is

$$\phi=\mathbf{1}_{(\frac{|\hat S(t_0)-p_0|}{\hat\sigma_{\hat S(t_0)}}\ge z_{\alpha/2})},$$

where $\Phi(z_\alpha)=1-\alpha$, Φ is the cdf of $N(0,1)$, and $\hat\sigma^2_{\hat S(t_0)}$ is the estimate of $\sigma^2_{\hat S(t)}$ given in §4.

If there is no exact observation, then there are three possible cases corresponding to Theorems 2 and 3 and the conjecture in §4.7.

The convergence rates are $n^{-1/2}$, $n^{-1/3}$ and $(n\ln n)^{-1/3}$, respectively.

For each of the three cases, a test can be constructed, which is introduced in §4.7.

(2) Two-sided nonparametric level-α test for H_0: F_o is a given cdf:
In complete data case, we often use

$$\phi=\mathbf{1}_{(|T|>z_{\alpha/2})},\ \text{where } T=\frac{\overline X-\mu}{s/\sqrt n},\ \mu=E(X) \text{ and } s^2=\frac{\sum_{i=1}^n(X_i-\overline X)^2}{n-1}$$

Note $E(X) = \int S_o(x)dx$ for nonnegative X. Thus the test statistic

$$T = \frac{U_o}{\hat{\sigma}_{U_o}}, \text{ where } U_o = \int (\hat{S}_{pl}(x) - S_o(x))dx.$$

This motivates, for RC data, the weighted Kaplan-Meier (WKM) statistic by Pepe and Fleming (1989, Biometrics and 1991, JRSS. A.).

WKM test

$$T = \frac{U}{\hat{\sigma}_U}, \text{ where } U = \int W(x)(\hat{S}_{pl}(x) - S_o(x))dx, \tag{1.1}$$

and W is a weight function, which may depend on F_0 and G. For example, let $W(t)$ be an estimate of $P(X \wedge Y \geq t)$ or $1 - G(t)$ etc.. In such a case,

$$\sigma_U^2 \approx \frac{1}{n} \int_0^\tau \frac{[\int_t^\tau W(u)S_o(u)du]^2}{S_o(t-)S_o(t)S_Y(t-)} dF_o(t),$$

where $\tau = \sup\{t : \ F_o(t) < 1, G(t) < 1\}$. σ_U^2 can easily be estimated. The test ϕ with T given by (1.1) is a location test for H_0 above.

With IC data, under a discrete assumption or in the case that there exist exact observations, U in (1.1) has the following form: Suppose there are m finite innermost intervals with right end-points $b_1 < \cdots < b_{m+1}$. WLOG, we can assume that $F_o(b_m) < 1$. Suppose that $W(t) = \mathbf{1}_{(t<\tau)}$, where τ can be estimated by the largest finite observation, denoted by b_{m+1}. Then

$$U = \int_0^{\hat{\tau}} (\hat{S}(t) - S_o(t))dt = \sum_{i=0}^{m} \hat{S}(b_i)(b_{i+1} - b_i) - \int_0^\tau S_o(t)dt.$$

where $b_0 = 0$. U can be rewritten as

$$U = b_0 + (b_2 - b_1, ..., b_{m+1} - b_m)(\hat{S}(b_1), ..., \hat{S}(b_m))^t + \int_0^\tau S_o(t)dt.$$

$$\sigma_U^2 = (b_2 - b_1, ..., b_{m+1} - b_m)\Sigma(b_2 - b_1, ..., b_{m+1} - b_m)^t$$

where Σ is the covariance matrix of $(\hat{S}(b_1), ..., \hat{S}(b_m))$.

For continuous IC data, the test statistic is not simple and needs to be investigated, as it is the integration of a stochastic process and the convergence rate of the GMLE varies at least in three different cases.

Kolmogorov test: The test is $\phi = \mathbf{1}_{(U>u_\alpha)}$, where

$$U = \sup_t |\hat{F}(t) - F_o(t)|,$$

and the critical values u_α can be computed from tables for RC data (Hall and Wellner. (1980). *Biometrika*). We shall address the test for IC data after we introduce the next test.

Smirnov test: The test is $\phi = \mathbf{1}_{(U>u_\alpha)}$, where

$$U = \int (\hat{F}(t) - F_o(t))^2 dF_o(t) \text{ or } U = \int (\hat{F}(t) - F_o(t))^2 dW(t),$$

where W is a measure, and the critical values u_α can be computed from tables for RC data (Koriol and Green, (1976). *Technometrics*).

For IC data, one may use bootstrap method to find u_α for both the Kolmogorov test and the Smirnov test.

§6.2. Two-sample problem

Suppose that there are two independent random samples for sizes n_1 and n_2, from survival function S_1 and S_2, respectively. H_0: $S_1 = S_2$, v.s. H_1: $S_1(t) \geq S_2(t)$ for all t and $S_1 \neq S_2$.

(1) **WKM statistic:**

For RC data, a test is $\phi = \mathbf{1}_{(T>z_\alpha)}$, where

$$T = \frac{U}{\hat{\sigma}_U}, \text{ where } U = \int \hat{W}(t)[\hat{S}_1(t) - \hat{S}_2(t)]dt,$$

$$\hat{\sigma}_U^2 = \int_0^\infty \frac{[\int_t^\infty \hat{W}(u)\hat{S}_o(u)du]^2}{\hat{S}_o(t)\hat{S}_o(t-)[1-\hat{G}_1(t-)]} d\hat{F}_o(t)/n_1 + \int_0^\infty \frac{[\int_t^\infty \hat{W}(u)\hat{S}_o(u)du]^2}{\hat{S}_o(t)\hat{S}_o(t-)[1-\hat{G}_2(t-)]} d\hat{F}_o(t)/n_2,$$

$\hat{S}_o$ is PLE based on pooled-sample, G_1 and G_2 are cdf of samples 1 and 2, respectively, and $\hat{G}_1$ and $\hat{G}_2$ are their PLEs based on samples 1 and 2, respectively.

For discrete IC data, we only consider

$$T = \frac{U}{\hat{\sigma}_U}, \text{ where } U = \int_0^\tau [\hat{S}_1(t) - \hat{S}_2(t)]dt,$$

where τ is a fixed constant. Since $\sigma_U^2 = Var(\int_0^\tau \hat{S}_1(t)dt) + Var(\int_0^\tau \hat{S}_2(t)dt)$, it can be estimated using an approach similar to the one discussed in §6.1.

(2) Kolmogorov test: The test is $\phi = \mathbf{1}_{(U>u_\alpha)}$, where

$$U = \sup_t |\hat{F}_1(t) - \hat{F}_2(t)|.$$

We can use simulation to estimate u_α. For instance, Consider re-sample the pooled-sample (L_i, R_i), $i = 1$, ..., $n_1 + n_2$, to generate two independent samples of sizes n_1 and n_2. Compute the value of U with these two samples. Repeat it 100 times, use the upper $100(1-\alpha)$ percentile to be an estimate of u_α.

(3) Smirnov test Geskus and Groeneboom's asymptotic results (1999) on smooth functionals with IC data can be applied to the two-sample version of the test.

(4) **Gehan's generalized Wilcoxon test.**

For RC data :

Notations: (Use z or $z+$ instead of $(x \wedge y, \delta)$ representation).

n_1 observations in the first sample: a_i or a_i+'s;

n_2 observations in the second sample: b_i or b_i+'s.

$$U_{ij} = \begin{cases} 1 & \text{if } a_i > b_j \text{ or } a_i+ \geq b_j \text{ (we know for sure that obs-}i \text{ in sample 1} > \text{obs-}j \text{ in sample 2)}, \\ -1 & \text{if } a_i < b_j \text{ or } a_i \leq b_j+ \text{ (we know for sure that obs-}i \text{ in sample 1} < \text{obs-}j \text{ in sample 2)}, \\ 0 & \text{if not sure.} \end{cases}$$

$U = \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} U_{ij}$. A test is $\phi = \mathbf{1}_{(\frac{U}{\hat{\sigma}_U} > z_\alpha)}$.

Question: $\sigma_U^2 = ?$

σ_U^2 is derived (Gehan, 1965, Biometrika), but is very complicated. Thus it is not presented here.

Mantel (1967, Biometrics) considered a different sample space and derived a different but simpler variance:

1. Pool two samples together, and denoted by c_i or c_i+, $i = 1$, ..., $n_1 + n_2$.

Order the second sample after the first one:

$c_1 = a_1$ (or $c_1+ = a_1+$), ..., $c_{n_1} = a_{n_1}$ (or $c_{n_1}+ = a_{n_1}+$), ..., $c_{n_1+n_2} = b_{n_2}$ (or $c_{n_1+n_2}+ = b_{n_2}+$).

2. Denote

$$V_{kh} = \begin{cases} 1 & \text{if we know for sure obs-}k > \text{obs-}h, \\ -1 & \text{if we know for sure obs-}k < \text{obs-}h, \\ 0 & \text{if not sure,} \end{cases} \tag{2.1}$$

(for RC data, we know for sure obs-k > obs-h iff $c_k > c_h$ or $c_k+ \geq c_h$.)

$$V_k = \sum_{h=1}^{n_1+n_2} V_{kh}, \tag{2.2}$$

Let W be a random variable taking values $\sum_{i=1}^{n_1} V_{k_i}$, where $\{k_1, ..., k_{n_1}\}$ is a selection of n_1 distinct integers from $\{1, ..., n_1 + n_2\}$.

$$V = \sum_{k=1}^{n_1} V_k \tag{2.3}$$

is a value of W. Under this sample space (permutation sample space, each permutation has equal probability),

$$E(W) = 0 \text{ and } \sigma_W^2 = n_1 n_2 \sum_{k=1}^{n_1+n_2} \frac{V_k^2}{(n_1+n_2)(n_1+n_2-1)}. \tag{2.4}$$

$$\left(E(W^2) = \sum_{k_1,...,k_{n_1}} \frac{1}{\binom{n_1+n_2}{n_i}} \sum_{i=1}^{n_1} V_{k_i}^2 \right).$$

Mantel suggests that

$$\frac{W}{\sigma_W} \xrightarrow{D} N(0,1) \text{ as } n \to \infty.$$

We observe $W = V$.
If H_1 is true, V should be large. Thus a test is $\psi = \mathbf{1}_{(\frac{W}{\sigma_W} > z_\alpha)}$. That is $\psi = \mathbf{1}_{(\frac{V}{\sigma_W} > z_\alpha)}$. Under such a set-up, U is a value of W, in fact $U = V$.

$$\begin{aligned} V &= \sum_{k=1}^{n_1} \sum_{h=1}^{n_1+n_2} V_{kh} \\ &= \sum_{k=1}^{n_1} \sum_{h=1}^{n_1} V_{kh} + \sum_{k=1}^{n_1} \sum_{h>n_1}^{n_1+n_2} V_{kh} \\ &= \sum_{k=1}^{n_1} \sum_{h=1}^{n_1} V_{kh} + \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} U_{ij} \\ &= 0 + U \qquad (\text{as } V_{kh} = -V_{hk}). \end{aligned}$$

Thus σ_W^2 can be viewed as a pseudo variance of U (**not really** a variance of U) and the Gehan's generalized Wilcoxon test is

$$\psi = \mathbf{1}_{(\frac{U}{\sigma_W} > z_\alpha)}. \tag{2.5}$$

For IC data we are sure that observation-k > observation-h iff

$$\text{either } L_k > R_h \text{ or } R_k > L_k = R_h.$$

Using this interpretation in (2.1), (2.4) still holds. Thus the generalized Wilcoxon test can be extended to the IC data by using (2.5) directly.

(5) **Logrank test** (Mantel (1966)) The test is a common test in medical research.

Notations:

c_1, ..., c_{m+n} are pooled RC observations. Among the pooled-sample

$$\begin{cases} t_1 < \cdots < t_k & \text{are distinct exact times,} \\ d_{\cdot,i} & \text{is the \# of deaths at } t_i,\ i = 1, ..., k, \\ \mathcal{R}(t_i) & \text{is the set of individuals in risk at time } t_i. \end{cases}$$

$d_{j,i}$ = # of deaths from group j at time t_i.
$d_{j,\cdot}$ = # of deaths in group j.
$r_{j,i}$ = # of elements of group j in $\mathcal{R}(t_i)$.
$r_{\cdot,i} = r_{1,i} + r_{2,i}$.

$U = \sum_{j=1}^{k}[d_{1,j} - d_{\cdot,j}\frac{r_{1,j}}{r_{\cdot,j}}]$.

$$\frac{U}{\sigma_U} \xrightarrow{D} N(0,1), \text{ where } \sigma_U^2 = \sum_{j=1}^{k} d_j \frac{r_{2,j} r_{1,j}(r_{\cdot,j} - d_{\cdot,j})}{r_{\cdot,j}^2(r_{\cdot,j} - 1)} \tag{2.6}$$

The level-α logrank test is $\phi = \mathbf{1}_{(\frac{U}{\sigma_U} < -z_\alpha)}$.

Derivation of logrank test

Under continuous assumption and PH model,

$z_i = z_i(t) = w(t)\mathbf{1}_{(\text{observation } i \text{ is from group } 1)}$, where $w(t)$ is a weight function.

$\mathcal{D} = \{i: \text{ observation } i \text{ in pooled-sample died}\}$.

$S_i(t|z) = (S_o(t))^{e^{\beta z(t)}}$, $i = 1, 2$.

H_0: $S_1 = S_2$ is equivalent to $\beta = 0$.

$lik = \prod_{i\in\mathcal{D}} \frac{e^{\beta z_i(c_i)}}{\sum_{h\in\mathcal{R}(c_i)} e^{\beta z_h(c_i)}}$

$$U(\beta) = \frac{d\ \ln lik}{d\beta} = \sum_{i\in\mathcal{D}}[z_i(c_i) - \frac{\sum_{j\in\mathcal{R}(c_i)} e^{\beta z_j(c_i)} z_j(c_i)}{\sum_{h\in\mathcal{R}(c_i)} e^{\beta z_h(c_i)}}], \text{ } --- \text{ the score function.}$$

$$U(0) = \sum_{i\in\mathcal{D}}[z_i(c_i) - \frac{\sum_{j\in\mathcal{R}(c_i)} z_j(c_i)}{|\mathcal{R}(c_i)|}] = \sum_{i\in\mathcal{D}} z_i(c_i) - \sum_{i\in\mathcal{D}} \frac{\sum_{j\in\mathcal{R}(c_i)} z_j(c_i)}{|\mathcal{R}(c_i)|}]$$

$$U(0) = \sum_{j=1}^{k} w(t_j)[d_{1,j} - d_{\cdot,j}\frac{r_{1,j}}{r_{\cdot,j}}],$$

as $z_i = 0$ for $i > n_1$. If $w(t_i) = 1$, $U(0) = \sum_{i=1}^{n_1}[d_{1,j} - \frac{d_{\cdot,j} r_{1,j}}{r_{\cdot,j}}]$, corresponding to the logrank test.

The asymptotic variance of $U(0)$ can be obtained by

$$\begin{aligned}
\sigma_U^2 \approx -\frac{dU(\beta)}{d\beta}\Big|_{\beta=0} &= \sum_{i\in\mathcal{D}}[\frac{\sum_{j\in\mathcal{R}(c_i)} e^{\beta z_j} z_j^2}{\sum_{h\in\mathcal{R}(c_i)} e^{\beta z_h}} - \frac{\sum_{j\in\mathcal{R}(c_i)} e^{\beta z_j} z_j \sum_{l\in\mathcal{R}(c_i)} e^{\beta z_l} z_l}{(\sum_{h\in\mathcal{R}(c_i)} e^{\beta z_h})^2}]|_{\beta=0} \\
&= \sum_{i=1}^{k} (w(t_i))^2[\frac{r_{1,i}}{r_{\cdot,i}} - \frac{r_{1,i}^2}{r_{\cdot,i}^2}] \\
&= \sum_{i=1}^{k}[(w(t_i))^2 \frac{r_{1,i}}{r_{\cdot,i}} \frac{r_{2,i}}{r_{\cdot,i}}]
\end{aligned}$$

For $w(t_i) = 1$,

$$\sigma_U^2 = -\frac{dU(\beta)}{d\beta}\Big|_{\beta=0} = \sum_{i=1}^{k}[\frac{r_{1,i}}{r_{\cdot,i}} \frac{r_{2,i}}{r_{\cdot,i}}]$$

which equals (2.6) as $d_{\cdot,i} = 1$ by the continuity assumption.

A two-sided test $\psi = \mathbf{1}_{(|\frac{U(0)}{\sigma_U}| > z_{\alpha/2})}$ is called a linear rank test, where

$w(t_i) = r_{\cdot,i}$ — generalized Wilcoxon test, Gehan (1965);

$w(t_i) = 1$ —- logrank test Mantel (1965);

$w(t_i) = n\hat{S}_{pl}(t_i-)$ — Prentice (1978);

$w(t_i) = n(\hat{S}_{pl}(t_i-))^k$ — Harrington and Fleming (1982).

Example in Splus.

x=coxph(Surv(time)$\sim$ ag+log(wbc),data=leuk)

summary(x)

For IC data

Self and Grossman (1986) consider the linear regression problem with IC data for a given distribution with location-scale parameters and propose a marginal likelihood approach, the marginal distribution of the ranks of the underlying survival times. Note that their problem is essentially parametric even though they used a nonparametric approach. Rabinowitz *et al.* (1995) proposed a class of score statistics to estimate parameters of the accelerated failure time model with IC data. Finkelstein (1986, Biometrics) extended the logrank test to the IC data. Satten (1996) consider rank-based inference in the proportional hazards model with IC data. Both extended the logrank test for RC data.

Reference.

* Self, S.G. and Grossman, E.A. (1986). Linear rank tests for interval-censored data with application to PCB levels in adipose tissue of transformer repair workers. *Biometrics*, 42, 521-530.
* Gehan, E.A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biomerika* 52 203-223.
* Geskus, R. B. and Groeneboom, P. (1999). Asymptotically optimal estimation of smooth functionals for interval censoring, case 2. *Ann. Statist.* 27, 627-675.
* Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.*, 50, 163-170.
* Satten, G.A. (1996). Rank-based inference in the proportional hazards model for interval censored data. *Biometrika*, 83, 355-370.