Statistical Inference of Interval-censored Failure Time Data

Jinheum Kim[1]

[1]Department of Applied Statistics, University of Suwon

May 28, 2011

Outline

- Interval censoring
- Nonparametric estimation
- Comparison of the survival functions
- Regression analysis
- Miscellaneous topics
- Summary

Interval-censored data

- T : survival time of interest
- An observation on T is **interval-censored** if instead of observing T exactly, only an interval $(L, R]$ is observed such that

$$T \in (L, R]$$

- *cf.* Grouped data: intervals for any two subjects either are completely identical or have no overlapping

Types of interval-censored data

- Case I interval-censored data or current status data
  - T is only known to be larger or smaller than an observed monitoring time, C
  - Either $L = 0$ or $R = \infty$
  - Observed data: $\{(C, \delta = I(C \leq T))\}$
  - eg, Cross-sectional studies or tumourigenicity experiments
- Case II interval-censored data
  - Include at least one interval $(L, R]$ with both L and R
  - In experiments with two monitoring times, U and V, with $U \leq V$, $T \leq U$, $U < T \leq V$, or $T > V$
- Case K interval-censored data
  - In longitudinal studies with periodic follow-up and K monitoring times, $M_1, \ldots, M_K$, the event is only observed between two consecutive inspecting times, M_l and M_{l+1}, and the observed data reduced to $(M_l, M_{l+1}]$

Non-informative interval censoring

- Censoring times are independent of the survival time completely or given covariates
- Except for the fact that T lies between l and r, the interval $(L, R]$ does not provide any extra information for T, i.e.,

$$P(T \leq t | L = l, R = r, L < T \leq R) = P(T \leq t | l < T \leq r),$$

- In the existence of covariates, Z,

$$P(T \leq t | L = l, R = r, L < T \leq R, Z = z)$$

$$= P(T \leq t | l < T \leq r, Z = z)$$

Notation

- Observed data: $\mathcal{O} = \{(L_i, R_i]; i = 1, \ldots, n\}$
- Want to estimate $S(t) = P(T > t)$ or $F(t) = 1 - S(t)$
- $\{t_j\}_{j=0}^{m+1}$: unique ordered elements of $\{0, \{L_i\}_{i=1}^n, \{R_i\}_{i=1}^n, \infty\}$, i.e.,

$$0 = t_0 < t_1 < \cdots < t_m < t_{m+1} = \infty$$

- Define

$$\alpha_{ij} = I((t_{j-1}, t_j] \subset (L_i, R_i])$$

and

$$p_j = S(t_{j-1}) - S(t_j), j = 1, \ldots, m+1$$

Non-parametric MLE

- Likelihood function for $\mathbf{p} = (p_1, \ldots, p_{m+1})'$:

$$L_S(\mathbf{p}) = \prod_{i=1}^{n} [S(L_i) - S(R_i)] = \prod_{i=1}^{n} \sum_{j=1}^{m+1} \alpha_{ij} p_j$$

  - Depend on S only through $\{S(t_j)\}_{j=1}^{m}$
- NPMLE, $\hat{S}$, of S: Maximize $L_S(\mathbf{p})$ under $\sum_{j=1}^{m+1} p_j = 1$ and $p_j \geq 0$
  - $\hat{S}$: Right-continuous step function, i.e., $\hat{S}(t) = \hat{S}(t_{j-1}), t_{j-1} \leq t < t_j$
- Remark
  - Some elements of $\hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_{m+1})'$ could be 0 and it could help to know these zero components before running a determination process
  - $\hat{p}_j$ could be non-zero only if $t_{j-1} = L_i$ for some i and $t_j = R_k$ for some possibly k, $i, k = 1, \ldots, n$ (Turnbull (JRSSB, 1976)'s approach)

Illustrative example

Subject number	L_i	R_i
1	0	7
2	0	8
3	6	10
4	7	16
5	7	14
6	17	∞
7	37	44
8	45	∞
9	46	∞
10	46	∞

- $t_0 = 0, t_1 = 6, t_2 = 7, t_3 = 8, t_4 = 10, t_5 = 14, t_6 = 16, t_7 = 17, t_8 = 37, t_9 = 44, t_{10} = 45, t_{11} = 46, t_{12} = \infty$

Turnbull intervals

Sub. #	t_0	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}
1	0		7										
2	0			8									
3		6			10								
4			7				16						
5			7			14							
6								17					∞
7									37	44			
8											45		∞
9												46	∞
10												46	∞
		(6	7]	(7	8]				(37	44]		(46	∞)

Likelihood: complete data-based

- Suppose $\{T_i\}_{i=1}^n$ are available
- Log-likelihood for the complete data

$$l_S(\mathbf{p}; T_1, \ldots, T_n) = \log[\prod_{i=1}^{n} dF(T_i)] = \sum_{j=1}^{m+1} (\log p_j) \sum_{i=1}^{n} I(T_i = s_j)$$

EM algorithm: E-step

- Proposed by Turnbull (JRSSB, 1976) using EM algorithm (Dempster et al., JRSSB, 1977)
- E-step: Compute the conditional expectation of the complete-data log-likelihood given the observed data $\mathcal{O}$ & $\mathbf{p} = \hat{\mathbf{p}}^{(s)}$

$$E[l_S(\mathbf{p}; T_1, \ldots, T_n)|\mathcal{O}, \hat{\mathbf{p}}^{(s)}] = \sum_{j=1}^{m+1} (\log p_j) \sum_{i=1}^{n} E[I(T_i = s_j)|\mathcal{O}, \hat{\mathbf{p}}^{(s)}]$$

$$= \sum_{j=1}^{m+1} (\log p_j) \sum_{i=1}^{n} \frac{\alpha_{ij} \hat{p}_j^{(s)}}{\sum_{l=1}^{m+1} \alpha_{il} \hat{p}_l^{(s)}}$$

EM algorithm: M-step

- M-step: Maximize the conditional expectation wrt $\mathbf{p}$ subject to $\sum_{j=1}^{m+1} p_j = 1$ and $p_j \geq 0$

$$L_S(\mathbf{p}, \lambda) = \sum_{j=1}^{m+1} (\log p_j) \sum_{i=1}^{n} \frac{\alpha_{ij} \hat{p}_j^{(s)}}{\sum_{l=1}^{m+1} \alpha_{il} \hat{p}_l^{(s)}} + \lambda (1 - \sum_{j=1}^{m+1} p_j)$$

- NPMLE at the $(s+1)$th step:

$$\hat{p}_j^{(s+1)} = \frac{1}{n} \sum_{i=1}^{n} \frac{\alpha_{ij} \hat{p}_j^{(s)}}{\sum_{l=1}^{m+1} \alpha_{il} \hat{p}_l^{(s)}}, j = 1, \ldots, m+1$$

- Iterate E & M steps until convergence
- EM algorithm leads to a self-consistency estimate if the iteration converges

Self-consistency algorithm

- Rewriting $\hat{p}_j^{(s+1)} = \frac{1}{n} E[\sum_{i=1}^{n} I(T_i = s_j)|\mathcal{O}, \hat{\mathbf{p}}^{(s)}]$ in its cumulative form in terms of distribution function,

$$\hat{F}^{(s+1)}(t) = \frac{1}{n} E[\sum_{i=1}^{n} I(T_i \leq t)|\mathcal{O}, \hat{F}^{(s)}],$$

  where $\hat{F}^{(s)}(t) = \sum_{j:s_j \leq t} \hat{p}_j^{(s)}$
- Easily implemented, but slowly converge and do not guarantee NPMLE
- Another algorithms: ICM(iterative convex minorant, Groeneboom & Wellner, 1992), Hybrid algorithm(EM-ICM, Wellner & Zhan, JASA, 1997)

Asymptotic behavior of NPMLE

- Unlike with right-censored data, the use of the counting process technique is quite difficult and as a consequence, the martingale theory cannot be applied
- $\hat{S}$ is strongly consistent, but the convergence rate is $O_p(n^{-1/3})$ (Geskus & Groeneboom, Statistica Neerlandica, 1997) and its limiting distribution is non-normal because of lack of information (Groeneboom, 1996)
- But, linear functionals of $\hat{S}$ are asymptotically normal with the usual $n^{1/2}$-rate such as the estimated mean failure time, $\hat{E}(T) = \int t d\hat{F}(t)$, where $\hat{F} = 1 - \hat{S}$ (Geskus & Groeneboom, Statistica Neerlandica, 1997)

Example

- Two treatments for breast cancer, radiation (Rad, n=46), and radiation with chemotherapy (RadChem, n=48)
- Response: Time in months until breast retraction (Finkelstein & Wolfe, BCS, 1985)
- Use R package **interval**: **icfit** function
- icfit function calculates NPMLE by EM algorithm
- > library(interval)
  > data(bcos)
  > head(bcos)

	left	right	treatment
1	45	Inf	Rad
2	6	10	Rad
3	0	7	Rad
4	46	Inf	Rad
5	46	Inf	Rad
6	7	16	Rad

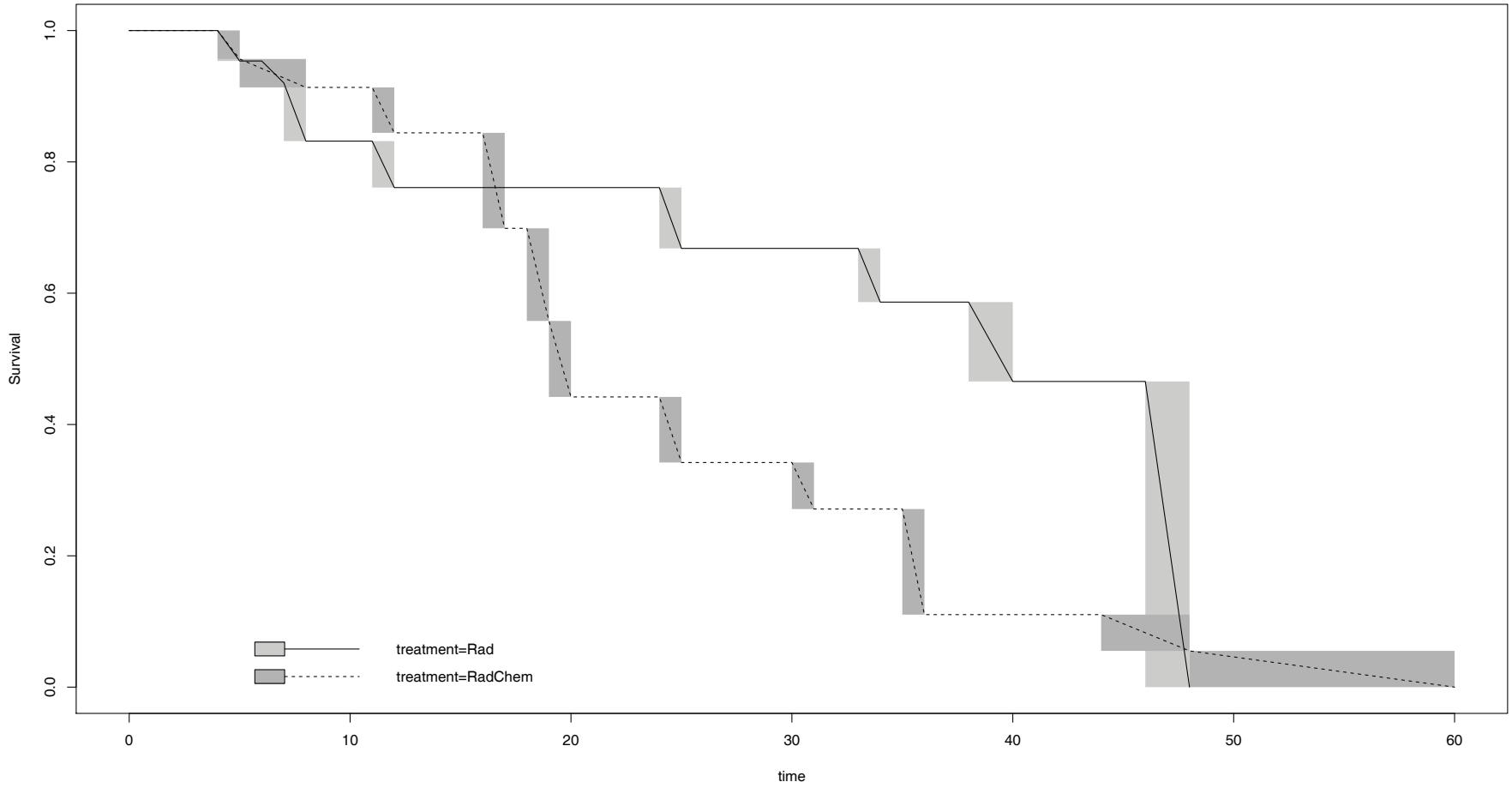
Results: NPMLE by EM algorithm

- > fit<-icfit(Surv(left,right,type="interval2")~ treatment,data=bcos)

```
> fit<-icfit(Surv(left,right,type="interval2")~treatment,data=bcos)
> summary(fit)
treatment=Rad:
   Interval Probability
1     (4,5]      0.0463
2     (6,7]      0.0334
3     (7,8]      0.0887
4    (11,12]     0.0708
5    (24,25]     0.0926
6    (33,34]     0.0818
7    (38,40]     0.1209
8    (46,48]     0.4656
treatment=RadChem:
    Interval Probability
1      (4,5]      0.0433
2      (5,8]      0.0433
3     (11,12]     0.0692
4     (16,17]     0.1454
5     (18,19]     0.1411
6     (19,20]     0.1157
7     (24,25]     0.0999
8     (30,31]     0.0709
9     (35,36]     0.1608
10    (44,48]     0.0552
11    (48,60]     0.0552
```

Estimated survival curves

- >plot(fit)



Objective: Hypothesis testing

- $S^{(k)}(t)$: Survival function of the kth arm with $k = 1, \ldots, K$
- Want to test

$$H_0 : S^{(1)}(t) = \cdots = S^{(K)}(t), \forall t$$

- With right censoring,
  - Rank-based tests: Rely on differences between the estimated hazard functions, eg, log-rank test
  - Survival-based tests: Rely on differences between the estimated survival functions
- Generalize to the case of interval-censored data

Notation

- $\hat{S}_0$: NPMLE of the $S^{(k)}$'s under H_0
- $\delta_i = 0$ if right-censored and 1 otherwise
- $\rho_{ij} = I(\delta_i = 0, L_i \geq t_j\}$, i.e., $\rho_{ij} = 1$ if T_i is right-censored and subject i is still ar risk at t_j-
- Define the estimates of the total observed failures and risk numbers, respectively, as

$$d_j = \sum_{i=1}^{n} \delta_i \frac{\alpha_{ij}[\hat{S}_0(t_j-) - \hat{S}_0(t_j)]}{\sum_{l=1}^{m+1} \alpha_{il}[\hat{S}_0(t_l-) - \hat{S}_0(t_l)]}, j = 1, \ldots, m,$$

$$n_j = \sum_{r=j}^{m+1} \sum_{i=1}^{n} \delta_i \frac{\alpha_{ir}[\hat{S}_0(t_r-) - \hat{S}_0(t_r)]}{\sum_{l=1}^{m+1} \alpha_{il}[\hat{S}_0(t_l-) - \hat{S}_0(t_l)]} + \sum_{i=1}^{n} \rho_{ij}, j = 1, \ldots, m$$

- Similarly, define d_{jk} and n_{jk} from subjects in arm $k = 1, \ldots, K$

Weighted log-rank tests

- Define $\mathbf{U} = (U_1, \ldots, U_K)'$ with

$$U_k = \sum_{j=1}^{m} (d_{jk} - n_{jk} \frac{d_j}{n_j})$$

- Estimation of the covariance matrix of $\mathbf{U}$: employ resampling methods such as multiple imputation, bootstrap and permutation procedures
- Remark
  - When $K = 2$, with $W(t) = 1$,

$$U_1 = \int_0^{\infty} W(t) \frac{Y_1(t) Y_2(t)}{Y_1(t) + Y_2(t)} [d\hat{\Lambda}_1(t) - d\hat{\Lambda}_2(t)],$$

    where $Y_k(t) = \sum_{j:t_j \le t} n_{jk}$ and $\hat{\Lambda}_k(t) = \sum_{j:t_j \le t} d_{jk}/n_{jk}$
  - General weight process: $W(t) = \hat{S}_0(t-)^{\rho}[1 - \hat{S}_0(t-)]^{\gamma}$ with $\rho, \gamma > 0$

Survival-based tests

- Zhang et al. (BKA, 2001) and Fang et al. (StatSinica, 2002): Based on

$$\int_0^\tau W(t)[\hat{S}^{(1)}(t) - \hat{S}^{(2)}(t)]dt$$

  - $\hat{S}^{(k)}$: NPMLE of $S^{(k)}$, separately
  - τ : longest follow-up time
  - Specially, when $W(t) = 1$, reduced to the difference of the estimated sample means
- Supremum-type test based on the difference between the estimated survival functions (Yuen et al., BKA, 2006)

Example(continued): Logrank test

- Use R package **interval**: **ictest** function
- > Suntest<-ictest(Surv(left,right,type="interval2")~treatment,scores="logrank1",data=bcos)

```
> Suntest<-ictest(Surv(left,right,type="interval2")~treatment,scores="logrank1",data=bcos)
> Suntest

        Asymptotic Logrank two-sample test (permutation form), Sun's scores

data:  Surv(left, right, type = "interval2") by treatment
Z = -2.6684, p-value = 0.007622
alternative hypothesis: survival distributions not equal

                    n Score Statistic*
treatment=Rad      46         -9.141846
treatment=RadChem 48          9.141846
* like Obs-Exp, positive implies earlier failures than expected
>
```

Proportional hazards model

- Data: $\{(L_i, R_i], \mathbf{Z}_i; i = 1, \ldots, n\}$
  - $\mathbf{Z}_i$: p-dimensional vector of covariates
- Model: $\lambda(t|\mathbf{Z}) = \lambda_0(t)e^{\boldsymbol{\beta}'\mathbf{Z}}$
  - $\lambda_0(t)$: unknown baseline hazard function
  - $\boldsymbol{\beta}$: vector of unknown regression parameters
- Unlike right-censored data, estimating $\boldsymbol{\beta}$ under interval censoring involve estimation of both $\boldsymbol{\beta}$ and the cumulative baseline hazard function, $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$

ML approach

- Likelihood function: $L = \prod_{i=1}^{n} [S(L_i; \mathbf{Z}_i) - S(U_i; \mathbf{Z}_i)]$
  - $S(t; \mathbf{Z}) = S_0(t)^{\exp(\beta' \mathbf{Z})}$: Survival function for a subject with covariates $\mathbf{Z}$
- Log-likelihood: Assuming that $L_i < R_i, \forall i$,

$$l(\beta, S_0) = \sum_{i=1}^{n} \log\{[S_0(L_i)^{\exp(\beta' \mathbf{Z}_i)} - S_0(R_i)^{\exp(\beta' \mathbf{Z}_i)}]\}$$

  - S_0 : baseline survival function, $S_0(t) = e^{-\Lambda_0(t)}$
- Focus on estimation of S_0 at the different observation time points, i.e., $t_0 = 0 < t_1 < \cdots < t_{m+1} = \infty$, of the form,

$$S_0(t) = \prod_{j: t_j \le t} e^{-\exp(\alpha_j)} = e^{-\sum_{j: t_j \le t} \exp(\alpha_j)}$$

  - $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)'$: unknown parameters

ML approach (continued)

- $l(\boldsymbol{\beta}, S_0)$ can be rewritten as

$$l(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \log\{\sum_{j=1}^{m+1} \alpha_{ij}[e^{-\exp(\boldsymbol{\beta}'\mathbf{Z}_i)a_{j-1}} - e^{-\exp(\boldsymbol{\beta}'\mathbf{Z}_i)a_j}]\}$$

  - $a_j = \sum_{k=1}^{j} \exp(\alpha_k)$
- Use the Newton-Raphson algorithm to determine the MLE of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ (Finkelstein, BCS, 1986)
- Asymptotic properties (Huang & Wellner, 1997)
  - $\hat{S}_0$: strongly consistent
  - $\hat{\boldsymbol{\beta}}$: asymptotically normal with the usual $\sqrt{n}$-convergence rate and efficient

Alternative approaches

- Marginal likelihood approach
  - Based on the likelihood given by the sum over all rankings of the underlying and unobserved failure times that are consistent with the observed censoring intervals (Kalbfleish & Prentice, BKA, 1973)
  - NOT require estimation of S_0, but involve much computational works
  - Use Gibbs sampling procedure for generating underlying rankings and stochastic approximation for solving the score equations (Sattern, BKA, 1996)
  - MCMC EM algorithm for determination of regression parameter estimates (Giggins et al., BCS, 1998)
- Multiple imputation procedure (Pan, BCS, 2000)

Proportional odds model

- $\frac{S(t;\mathbf{Z})}{1-S(t;\mathbf{Z})} = e^{-\beta'\mathbf{Z}} \frac{S_0(t)}{1-S_0(t)} \Leftrightarrow S(t;\mathbf{Z}) = \frac{1}{1+\exp[H(t)+\beta'\mathbf{Z}]}$
  - $H(t) = -\mathrm{logit}[\mathrm{S_0(t)}]$: baseline log-odds function
- As in PH model case, directly maximize a likelihood function wrt β and H or S_0
- Other sieve ML approaches: piecewise linear function (Huang & Rossini, JASA, 1997), monotone spline (Shen, BKA, 1998)
- Alternative method: based on the approximate conditional likelihood (Rabinowitz et al., BKA, 2000)
  - Free of estimation of H or S_0

Accelerated failure time model

- $\log T = \beta' \mathbf{Z} + e$
  - e : error variable with an unknown distribution function F
  - e is independent of L, R, and $\mathbf{Z}$
  - Directly specify the linear relationship between $\log T$ and $\mathbf{Z}$
- Likelihood

$$L(\beta, F) = \prod_{i=1}^{n} [F(R_i(\beta)) - F(L_i(\beta))]$$

  - $R_i(\beta) = \log(R_i) - \beta' \mathbf{Z}$, $L_i(\beta) = \log(L_i) - \beta' \mathbf{Z}$,
- Unlike the previous two models, ML approach is relatively hard because β and F are tangled in the likelihood.

Linear rank-based approach

- Key idea: each individual's censoring times, L and R, can be translated into a sequence of current status observations
- Denote L_i by X_{i1} and R_i by X_{i2}
- Likelihood

$$L^*(\beta, F) = \prod_{i=1}^{n} \prod_{j=1}^{2} F(X_{ij}(\beta))^{1-\delta_{ij}} [1 - F(X_{ij}(\beta))]^{\delta_{ij}},$$

  where $\delta_{ij} = I(X_i \leq T_{ij}), j = 1, 2$
- $\hat{F}_{\mathbf{b}}$: NPMLE of F based on $L^*(\beta, F)$ with $\beta = \mathbf{b}$ applied to $\{(\delta_{ij}, X_{ij}(\mathbf{b})); i = 1, \ldots, n, j = 1, 2\}$

Linear rank-based approach (continued)

- Propose a linear rank statistic (Betensky et al, BKA, 2001)

$$S(\mathbf{b}) = \sum_{i=1}^{n} \sum_{j=1}^{2} [Y_{ij} - \hat{F}_{\mathbf{b}}(X_{ij}(\mathbf{b}))] Z_i$$

  - $\hat{\beta}$: value of $\mathbf{b}$ for which $S(\mathbf{b})$ is closest to zero
- Rabinowitz et al. (BKA, 1995): based on $L(\beta, F)$ instead of $L^*(\beta, F)$
  - Asymptotic efficiency is achieved, but not to be practical

Additive risk model

- Model: $\lambda(t;\mathbf{Z}) = \lambda_0(t) + \boldsymbol{\beta}'\mathbf{Z}$
- Likelihood

$$\prod_{i=1}^{n}[S_0(L_i)e^{-(\boldsymbol{\beta}'\mathbf{Z}_i)L_i} - S_0(R_i)e^{-(\boldsymbol{\beta}'\mathbf{Z}_i)R_i}]$$

- As in PH model case, directly maximize a likelihood function wrt $\boldsymbol{\beta}$ and S_0
- $\hat{S}_0$: non-increasing step function with jumps only at the observed examination times
- $\hat{S}_0$ almost surely converge to S_0 and $\hat{\boldsymbol{\beta}}$ is asymptotically normal

Alternative approaches

- Transformation approach (Zhu et al., LIDA, 2008)
  - Define two sets of current status data such as $\Omega_L = \{(L_i, \delta_{1i} = I(T_i \leq L_i), \mathbf{Z}); i = 1, \ldots, n\}$, $\Omega_R = \{(R_i, \delta_{2i} = I(T_i > R_i), \mathbf{Z}); i = 1, \ldots, n\}$
  - For observed interval-censored data, $\Omega = \{(L_i, R_i, \delta_{1i}, \delta_{2i}, \mathbf{Z}); i = 1, \ldots, n\}$,

$$\Omega = \Omega_L \cup \Omega_R$$

  - Denote $U_C(\beta, \Omega)$ by the estimating function based on the current status data Ω
  - Estimate β based on the estimating equation

$$U(\beta, \Omega) = U_C(\beta, \Omega_L) + U_C(\beta, \Omega_R) = 0$$

- Multiple imputation approach (Chen & Sun, CommStat, 2010)

Miscellaneous topics

- Multivariate interval-censored data
  - When happens? A survival study involves several related survival variables of interest and each of them suffers interval censoring
  - Need to take into account the correlation among the survival variables and make inference the association between the survival variables
- Doubly censored data
  - Motivated by AIDS research
  - De Gruttola & Lagakos (BCS, 1989) proposed a SC algorithm for estimating of the survival function of AIDS latency time
- Competing risks analysis
  - When needed? The failure on an individual may be one of several distinct failure types
  - For the current status data, Groeneboom et al. (AnnStat, 2008)
- Truncation
- Parametric procedures

Informative interval-censored data

- Define the contribution of an interval censored observation to the likelihood by

$$\Pr(L = l, R = r, T \in (L, R]) = \Pr(T \in (L, R] | L = l, R = r) dG(l, r),$$

  where dG : joint density function of (L, R)
- Under the independent interval censoring, replace by

$$\Pr(l < T \leq r),$$

  called as a simplified likelihood (i.e., can ignore the censoring mechanism or the observation process)
- Under what conditions we have or what types of observation processes give independent interval censoring?
  - When a censoring model satisfies a constant-sum condition (Oller et al., CanadJ, 2004), for all $t \in \{t : dF(t) > 0\}$,

$$\int_{\{(l,r):t \in (l,r]\}} \frac{\Pr(L \in dl, R \in dr, T \in (L, R])}{\Pr(T \in (L, R])} = 1$$

Informative interval-censored data

- Common way to deal with the informative censoring are to jointly model the survival variable and the censoring variables or assume that the complete observation process is known. For the latter, it is important to have follow-up beyond the failure.
- For current status data, van der Laan & Robins (JASA, 1998), Zhang et al. (StatMed, 2005)
- For interval-censored data, Finkelstein et al. (BCS, 2002), Betensky & Finkelstein (BCS, 2002), Zhang et al. (StatMed, 2007)

Estimating equation approach

- With a regression analysis of current status data (Zhang et al., StatMed, 2005),
- Data: $\{(C_i, \delta_i = I(C_i \leq T_i, \mathbf{Z}_i); i = 1, \ldots, n\}$
- Assumption: Given $\mathbf{Z}$, T & C are correlated
- So, assume that the relationship between T and C can be characterized by a random effects u, and given u, T and C are independent.
- For T_i, assume an additive risk model

$$\lambda^T(t|u_i, \mathbf{Z}_i) = \lambda_{t0}(t) + u_i + \beta' \mathbf{Z}_i$$

- For C_i, assume the Cox PH model

$$\lambda^C(t|u_i, \mathbf{Z}_i) = \lambda_{c0}(t) \exp(u_i + \gamma' \mathbf{Z}_i)$$

Estimating equation approach (continued)

- Define a counting process as

$$N_i(t) = \delta_i I(C_i \leq t)$$

- Note that

$$E(dN_i(t)|\mathbf{Z}_i) = e^{\gamma'\mathbf{Z}_i - \beta'\mathbf{Z}_i t} d\Lambda_0^*(t),$$

where $d\Lambda^*(t) = e^{-\Lambda_{t0}(t)} E[e^{(1-t)u_i}] d\Lambda_{c0}(t)$. So,

$$M^*(t) = N_i(t) - \int_0^t Y_i(s) e^{\gamma'\mathbf{Z}_i - \beta'\mathbf{Z}_i s} d\Lambda_0^*(s) : \text{ martingales}$$

- For inference about β & γ, apply the partial likelihood method

Full likelihood approach

- With a regression analysis of interval-censored data (Zhang et al., StatMed, 2007),
- Assumption: Conditional on $\mathbf{Z}_i$ & $\mathbf{b}_i$, T_i, L_i, & $W_i = R_i - L_i$ are independent
- Let $\mathbf{b}_i = (b_{i1}, b_{i2}, b_{i3})' \sim \text{MVN}(0, \Sigma)$: a vector of unobserved or latent variables
- For T_i, L_i, & $W_i = R_i - L_i$, assume the Cox PH models, respectively, as follows:

$$\lambda^T(t|\mathbf{b}_i, \mathbf{Z}_i) = \lambda_{t0}(t)\exp(b_{i1} + \beta'\mathbf{Z}_i)$$

$$\lambda^L(t|\mathbf{b}_i, \mathbf{Z}_i) = \lambda_{l0}(t)\exp(b_{i2} + \gamma'\mathbf{Z}_i)$$

$$\lambda^W(t|\mathbf{b}_i, \mathbf{Z}_i) = \lambda_{w0}(t)\exp(b_{i3} + \xi'\mathbf{Z}_i)$$

Full likelihood approach (continued)

- Let $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_m)'$, where $\zeta_j = \log \Lambda_{t0}(s_j)$
- Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\xi}', \boldsymbol{\zeta}', \sigma_{kl}, 1 \le k \le l \le 3)'$
- Let $\boldsymbol{\Delta}_i = (\delta_{1i}, \delta_{2i})'$, where $\delta_{1i} = I(T_i \le L_i)$ and $\delta_{2i} = I(L_i < T_i \le R_i)$
- Let $\Psi_i = I(W_i < \infty)$
- Full likelihood based on the observed data, $\mathcal{O} = \{\mathbf{O}_1, \ldots, \mathbf{O}_n\}$, where $\mathbf{O}_i = (\boldsymbol{\Delta}_i, \Psi_i, L_i, W_i, \mathbf{Z}_i')'$, is given by

$$L_{\mathcal{O}}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \int L_{\boldsymbol{\Delta}|L_i, W_i, \mathbf{b}_i}(\boldsymbol{\theta}) L_{L_i|\mathbf{b}_i}(\boldsymbol{\theta}) L_{W_i|\mathbf{b}_i}(\boldsymbol{\theta}) f(\mathbf{b}_i, \Sigma) d\mathbf{b}_i$$

- But, difficult to maximum this full likelihood since the $\mathbf{b}_i$'s are unknown
- Use EM-algorithm for making inference about $\boldsymbol{\theta}$

Summary

- Interval censoring
- Nonparametric estimation
  - Self-consistent (EM) algorithm
  - Asymptotic behavior of NPMLE
- Comparison of the survival functions
  - Rank-based tests
  - Survival-based tests
- Regression analysis
  - Cox PH model
  - Proportional odds model
  - Accelerated failure time model
  - Additive risk model
- Miscellaneous topics
  - Informative censoring
  - Estimating equation approach
  - Full likelihood approach

Thank you!