What to do today (2022/03/22)?

Part IV. Advanced Topics

- **Part IV.1 Counting Process Formulation** (Revisits to KM estm, Logrank test, and Cox PH model)
  - *Part IV.1.1 Theoretical Preparation*
  - **Part IV.1.2 Counting Process Formulation in LIDA and Applications: Revisits to KM, Logrank, Cox PH**
    - *Part IV.1.2A Formulation*
    - *Part IV.1.2B Revisit to Logrank Test*
    - *Part IV.1.2C Revisit to Cox's Partial Likelihood Approach*
    - **Part IV.1.2D Revisit to Nelson-Aalen Estimator and Breslow Estimator**
    - **Part IV.1.2E Revisit to Kaplan-Meier Estimator**
    - **Part IV.1.2F Others**
- *Part IV.2 Selected Recent Topics in LIDA*
- *Part IV.3 Beyond Lifetime Data Analysis*

Part IV.1.2D Revisit to Nelson-Aalen Estimator and Breslow Estimator

Recall the notation:

$$N_{\cdot}(t),\ Y_{\cdot}(t),\ A_{\cdot}(t),\ M_{\cdot}(t) = N_{\cdot}(t) - A_{\cdot}(t)$$

Thus $dM_{\cdot}(t) = dN_{\cdot}(t) - Y_{\cdot}(t)dH(t)$.

Assign $dM_{\cdot}(t) = 0 \Longrightarrow d\hat{H}(t) = \frac{dN_{\cdot}(t)}{Y_{\cdot}(t)}$:

Nelson-Aalen Estimator for the cumulative hazard function

$$\hat{H}(t) = \int_0^t \frac{1}{Y_{\cdot}(u)} dN_{\cdot}(u)$$

Part IV.1.2D Revisit to Nelson-Aalen Estimator and Breslow Estimator

Remarks:

- The observed failure times $0 \leq V_1 < \ldots < V_J$,

$$\hat{H}(t) = \sum_{V_j \leq t} \frac{\#\ failures\ at\ V_j}{\#at\ risk\ at\ V_j} = \sum_{V_j \leq t} \frac{d_j}{N_j}.$$

- Define $0/0 = 0$ or replace $1/Y_{\cdot}(u) = I(Y_{\cdot}(u) > 0)/Y_{\cdot}(u)$.
- $E\left[dN_i(t) | Y_i(t)\right] = Y_i(t)h(t)dt$

Nelson-Aalen Estimator's Asymptotics

$\left[\hat{H}(t) - H(t)\right]$ is

$$\begin{aligned} & \int_0^t \frac{I(Y_\cdot(u) > 0)}{Y_\cdot(u)} dN_\cdot(u) - \int_0^t \frac{I(Y_\cdot(u) > 0) Y_\cdot(u)}{Y_\cdot(u)} dH(u) - \int_0^t I(Y_\cdot(u) = 0) dH(u) \\ = & \int_0^t \frac{I(Y_\cdot(u) > 0)}{Y_\cdot(u)} dM_\cdot(u) - \int_0^t I(Y_\cdot(u) = 0) dH(u) \end{aligned}$$

By Martingale CLT, in distn

$$\sqrt{n} \int_0^t \frac{I(Y_\cdot(u) > 0)}{Y_\cdot(u)} dM_\cdot(u) \to \mathit{Gaussian}(0, \sigma^2(t))$$

and $\sqrt{n} \int_0^t I(Y_\cdot(u) = 0) dH(u) \to 0$ in Pr 1.

Thus $\sqrt{n}\left[\hat{H}(t) - H(t)\right] \to \mathit{Gaussian}(0, \sigma^2(t))$.
Plus $\sup_{t>0} |\hat{H}(t) - H(t)| \to 0$ in Pr 1.

Remarks:

- $\sigma^2(t) = \int_0^t \frac{1}{P(U \geq u)} dH(u)$ and $\hat{\sigma}^2(t) = \int_0^t \frac{nI(Y_\cdot(u)>0)}{Y_\cdot(u)} d\hat{H}_{NA}(u)$ with $n >> 1$.
- **Fleming-Harrington Estimator** for survivor function $\hat{S}(t) = \exp\{-\hat{H}(t)\}$:

$$\sqrt{n}\big[\hat{S}(t) - S(t)\big] \approx S(t)\sqrt{n}\big[\hat{H}(t) - H(t)\big]$$

- **Breslow Estimator** for the baseline hazard function in the Cox PH model:

$$\hat{H}_0(t;\beta) - H_0(t) = \int_0^t \frac{1}{\sum_{l=1}^n Y_l(t) e^{\beta z_l}} dN_\cdot(u)$$

with β replaced by the PMLE $\hat{\beta}$

  - Note that $\hat{H}_0(t;\hat{\beta}) = \big[\hat{H}_0(t;\hat{\beta}) - \hat{H}_0(t;\beta)\big] + \big[\hat{H}_0(t;\beta) - H_0(t)\big]$

Part IV.1.2D Revisit to Nelson-Aalen Estimator and Breslow Estimator: Confidence Band of $H(\cdot)$

Recall that $\hat{H}_{NA}(t) = \int_0^t \frac{I(Y_\cdot(u)}{Y_\cdot(u)} dN_\cdot(u) \sim Gaussian(H(t), \frac{\sigma^2(t)}{n})$ approximately.

Pointwise 95% CI: $\forall t > 0$, $\hat{H}_{NA}(t) \pm 1.96\sqrt{\frac{\hat{\sigma}^2(t)}{n}}$

95% Confidence Band: the lower, upper boundaries $L(t), U(t)$ satisfy $P\big(L(t) \leq H(t) \leq U(t) : t \in (0, \infty)\big) = 95\%$.

e.g. $\hat{H}_{NA}(t) \pm c\sqrt{\frac{\hat{\sigma}^2(t)}{n}}$ with c determined by

$$P\Big(\sup_{t>0} \Big|\frac{\hat{H}_{NA}(t) - H(t)}{\sqrt{\hat{\sigma}^2(t)/n}}\Big| \leq c\Big) = 95\%.$$

How to compute the critical value c?

- **Approach I** (based on the Brownian Motion)

Recall $\sqrt{n}\frac{\hat{H}_{NA}(\cdot)-H(\cdot)}{\hat{\sigma}(t)} \to W(\frac{\sigma^2(\cdot)}{\sigma^2(t)})$ in distn: $W(\cdot)$ the standard Brownian motion.
Thus, by the continuous mapping theorem,

$$\sup_{0\le s\le t} \sqrt{n}\Big|\frac{\hat{H}_{NA}(\cdot)-H(\cdot)}{\hat{\sigma}(t)}\Big| \to \sup_{0\le s\le t} \Big|W(\frac{\sigma^2(\cdot)}{\sigma^2(t)})\Big| = \sup_{0\le u\le 1} |W(u)|$$

in distribution.
From the table of Brownian Motion, find c such that $P(\sup_{0\le u\le 1} |W(u)| \le c) = 95\%$.

- **Approach II** (resampling) (Lin et al, 1993, Biometrika)

Recall to choose c such that $P(\sup_{0\le t<\infty}|Q_n(t)|\le c)=95\%$:
$Q_n(t)=\sqrt{n}\frac{\hat{H}_{NA}(\cdot)-H(\cdot)}{\hat{\sigma}(t)}$ is about

$$\frac{\sqrt{n}}{\hat{\sigma}(t)}\Big[\int_0^t \frac{I(Y_\cdot(u)>0)}{Y_\cdot(u)}(dN_\cdot(u)-Y_\cdot(u)dH(u))\Big]$$

$$= \frac{\sqrt{n}}{\hat{\sigma}(t)}\sum_{i=1}^n\int_0^t \frac{I(Y_\cdot(u)>0)}{Y_\cdot(u)}dM_i(u)$$

Define $\tilde{Q}_n(t)$ as follows

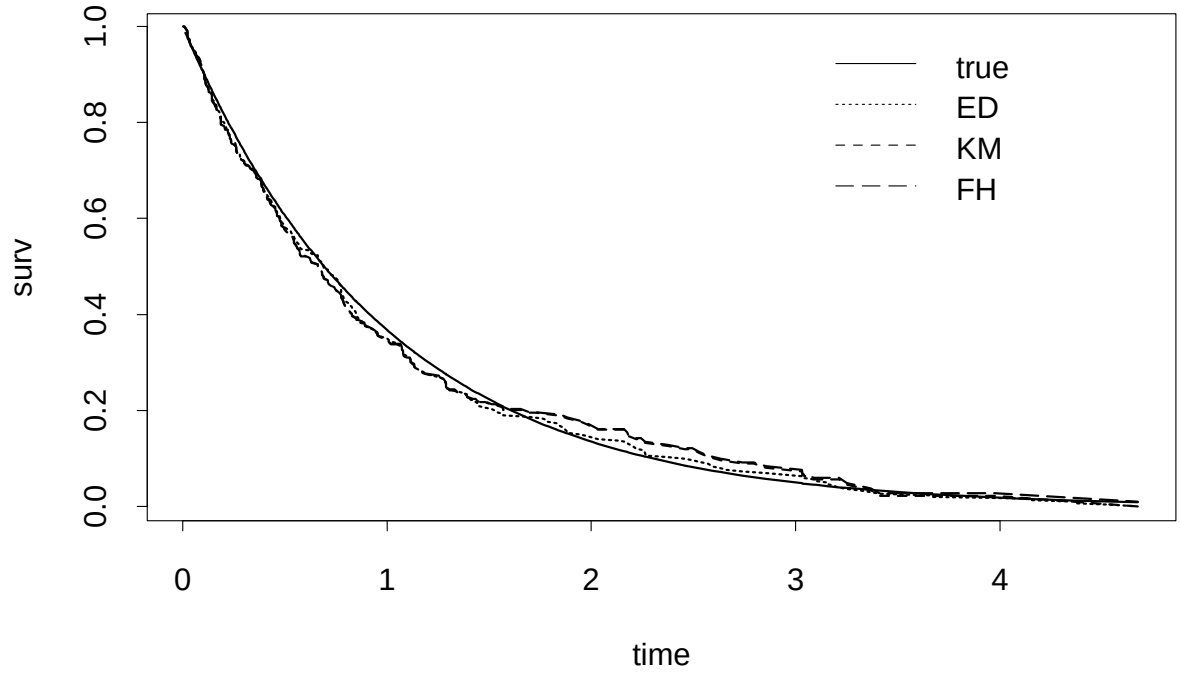
$$\frac{\sqrt{n}}{\hat{\sigma}(t)}\sum_{i=1}^n\int_0^t \frac{I(Y_\cdot(u)>0)}{Y_\cdot(u)}dN_i(u)Z_i$$

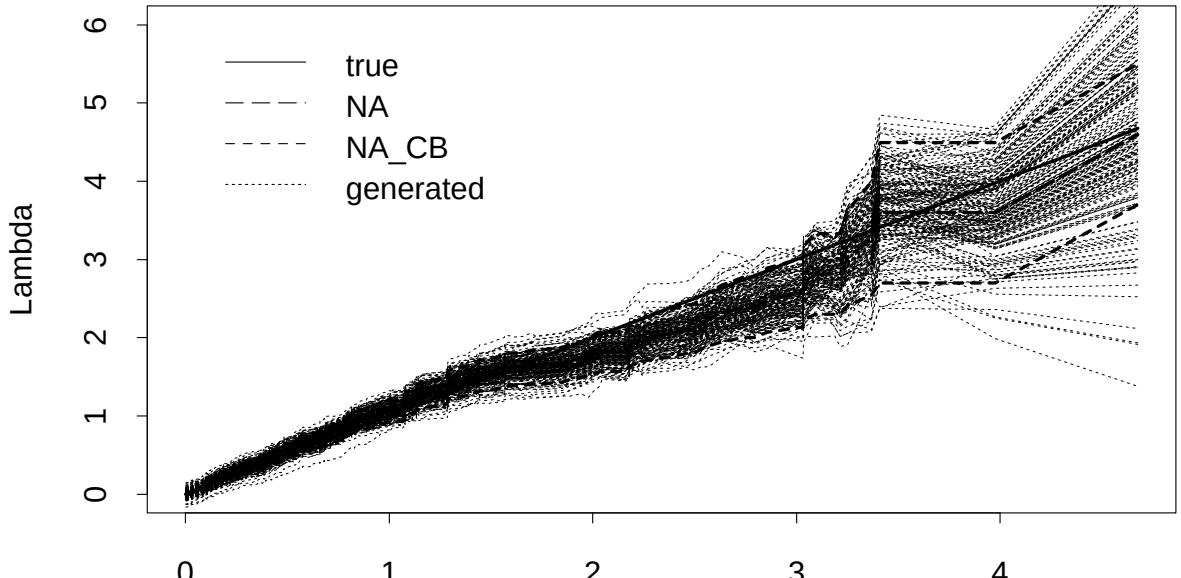
with $Z_1,\ldots,Z_n$ iid from $N(0,1)$ and $\perp\!\!\!\perp$ the data:
$\tilde{Q}_n(\cdot)$ converges weakly to $Gaussian(0,\sigma^2(t))$ (i.e. asymptotically equivalent to $\sigma(t)Q_n(\cdot)$).

Algorithm. At lth time, $l = 1, \ldots, M$,

- Step A. Generate $(Z_1, \ldots, Z_n)^{(l)}$ as an iid sample from $N(0, 1)$ and $\perp\!\!\!\perp$ the data.
- Step B. Evaluate $\tilde{Q}_n^{(l)}(\cdot)$ over $[0, T^*]$.
- Setp C. Calculate $W^{(l)} = \sup_{0 \le s \le T^*} |\tilde{Q}_n^{(l)}(s)|$.

Choose the critical c to be the 95% quantile of $W^{(1)}, \ldots, W^{(M)}$.


surv
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
time
true
ED
KM
FH


Lambda
0
1
2
3
4
5
6
0
1
2
3
4
true
NA
NA_CB
generated

Part IV.1.2E Revisit to Kaplan-Meier Estimator

Recall, with right-censored event times due to indpt censoring,

$$\hat{S}_{KM}(t) = \prod_{V_j<t} \left(1 - \frac{d_j}{N_j}\right)$$

Distinct observed event times $0 \leq V_1 < \ldots < V_J \leq V_{J+1}$: d_j=# of observed event time at V_j, N_j=# of individuals at risk at V_j. Plus $\hat{S}_{KM}(t) - S(t)$ is

$$\prod_{s<t} \left(1 - \frac{\Delta N_\cdot(s)}{Y_\cdot(s)}\right) - S(t) = -S(t) \int_0^t \frac{\hat{S}_{KM}(s-) I(Y_\cdot(s) > 0)}{S(s) Y_\cdot(s)} dM_\cdot(s) + B(t)$$

- $\sqrt{n} B(t) \to 0$ in Pr 1 as $n \to \infty$.
- $\sqrt{n}[\hat{S}_{KM}(t) - S(t)] \to Gaussian(0, \sigma^2(t))$ in distn.
  Thus $\sup_{t>0} |\hat{S}_{KM}(t) - S(t)| \to 0$ (*uniform consistency*).

Part IV.1.2E Revisit to Kaplan-Meier Estimator

Note that $\sqrt{n}[\hat{S}_{KM}(t) - S(t)]$ can be approximated by

$$Q_n(t) = -S(t) \int_0^t \frac{\sqrt{n} I(Y_\cdot(s) > 0)}{Y_\cdot(s)} dM_\cdot(s)$$

Define

$$\tilde{Q}_n(t) = -\hat{S}_{KM}(t) \sum_{i=1}^{n} \int_0^t \frac{\sqrt{n} I(Y_\cdot(s) > 0)}{Y_\cdot(s)} dN_i(s) Z_i$$

$Z_1, \ldots, Z_n \sim N(0, 1)$ iid and $\perp\!\!\!\perp$ the right-censored data.

$\Longrightarrow$ *the resampling method* to determine the critical value c for the CB *(alternative to the nonparametric bootstrap approach)*

Part IV.1.2E Revisit to Kaplan-Meier Estimator

Remarks:

- **Product-Integral**

$$\Pi_{s\in(0,t]}\Big(1+dX\Big)=\lim_{\max|t_i-t_{i-1}|\to 0}\prod\big(1+\Delta X(t_i)\big)$$

$$S(t)=\exp\{-\Lambda(t)\}=\Pi_{s\in(0,t]}\Big(1-d\Lambda(s)\Big)=\Pi_{s\in(0,t]}\Big(1-\lambda(s)ds\Big)$$

- **Applications of CB**
  - Model checking
  - Two-Sample Problem: $H_0 : S_1(\cdot) = S_2(\cdot)$
    - nonparametric comparison?
      CB of $\Delta(t) = S_1(t) - S_2(t)$ or $\Delta(t) = \Lambda_1(t) - \Lambda_2(t)$
      e.g. Hu and Lagakos (Biometrika, 1999); Zhao et al (Biostatistics, 2009)
    - semiparametric comparison?
      Assume $\lambda_2(t) = \theta\lambda_1(t)$ and test on $H_0 : \theta = 1$
  - K-Sample Problem?

What to study next?

Part IV. Advanced Topics

- *Part IV.1 Counting Process Formulation* (Revisits to KM estm, Logrank test, and Cox PH model)
- **Part IV.2 Selected Recent Topics in LIDA**
  - **Part IV.2.1 Alternatives to Cox PH model**
  - **Part IV.2.2 Multivariate event times**
  - *Part IV.2.3 More incomplete data structures*
  - *Part IV.2.4 Missing covariates in regression*
- *Part IV.3 Beyond Lifetime Data Analysis*