What to do today (2022/03/29)?

Part IV. Advanced Topics

- *Part IV.1 Counting Process Formulation* (Revisits to KM estm, Logrank test, and Cox PH model)
  - *IV.1.1 Theoretical Preparation*
  - *IV.1.2 Counting Process Formulation in LIDA and Applications: Revisits to KM, Logrank, Cox PH*

- **Part IV.2 Selected Recent Topics in LIDA**
  - *IV.2.1 Alternatives to Cox PH model*
  - *IV.2.2 Multivariate event times*
  - **IV.2.3 More unconventional data structures**
  - **IV.2.4 Analysis of incomplete data**

- *Part IV.3 Beyond Lifetime Data Analysis**

Part IV.2.3A More unconventional data structures in LIDA: Competing risks

What if to consider situations with J distinct causes of death?

- the ideal possibly available information on T: (T, j)
- envision T_j as the time to death due to jth cause, $j = 1, \ldots, J$

  $\Longrightarrow \quad T = \min(T_1, \ldots, T_J)$
  - often available is (U, δ) with $U = \min(T, C)$, and $\delta = 0$ for $T > C$ and $\delta = j$ for $T = T_j$

Problems of interest

- Estimate failure occurrence rates of specific types, and the relationship between specific failure types and covariates.
- Study interrelation between failure types.
- Estimate failure rates for certain types given the "removal" of some/all the other failure types.

How to achieve the goals?

- If $T_1, \ldots, T_J$ are $\perp\!\!\!\perp$, ...
- If $(T_1, \ldots, T_J) \sim f(t_1, \ldots, t_j; \theta)$, ...
- If it is neither of the cases above?

Part IV.2.3A Competing risks

Useful concepts

Recall *conditional hazard function* of T:

$$h(t|Z) = \lim_{\Delta t \to 0+} \frac{P(T \in [t, t+\Delta t) | T \geq t, Z)}{\Delta t}$$

$S(t|Z) = \exp(-\int_0^t h(u|Z)du)$

- **cause-specific hazard function**: for $j \geq 1$

$$h_j(t|Z) = \lim_{\Delta t \to 0+} \frac{P(T \in [t, t+\Delta t), \delta = j | T \geq t, Z)}{\Delta t}$$

$h(t|Z) = \sum_{j=1}^J h_j(t|Z)$; $f_j(t|Z) = h_j(t|Z)S(t|Z)$

- **sub-distribution**: for $j \geq 1$

$$P(T \leq t, \delta = j|Z) = \int_0^t f_j(u|Z)du$$

Part IV.2.3A Competing risks

Provided data of $\{(U_i, \delta_i, Z_i) : i = 1, \ldots, n\}$

Statistical inference

- to estm cause-specific hazard function:

$$\prod_{i=1}^{n} \Big(\prod_{j=1}^{J} h_j(u_i|z_i)^{I(\delta_i=j)} \Big) S(u_i|z_i)$$

identifiability of h_j?

- to estm the regression parameters β_j with $h_j(t|Z) = h_{0j}(t)e^{\beta_j Z}$: the partial likelihood function

$$L_P(\beta_1, \ldots, \beta_J) = \prod_{j=1}^{J} \prod_{i:\delta_i=j} \Big(\frac{e^{\beta_j z_i}}{\sum_{l \in \mathcal{R}(u_i)} e^{\beta_j z_l}} \Big)$$

Part IV.2.3A Competing risks

Provided data of $\{(U_i, \delta_i, Z_i) : i = 1, \ldots, n\}$

what if to study $(T_1, \ldots, T_J)$ jointly?
It's necessary to specify the dependence of $(T_1, \ldots, T_J)$ if only the competing risks data are available.

e.g. $J = 2$, assume $(T_1, T_2) \sim C(F_1(t_1), F_2(t_2))$ with
$C : [0, 1]^2 \to [0, 1]$, a copula function.
$F_j(t_j)$ is the marginal cdf of T_j

Part IV.2.3B Censoring mechanisms

Recall *right-censoring* ... ...
Consider an event time T. Its observation is subject to right-censoring if T is observed only when $T \leq C$, where C is the censoring time:
$U = \min(T, C)$ and $\delta = I(T \leq C)$.
Right-censored event times $\{(U_i, \delta_i) : i = 1, \ldots, n\}$

- If $T \perp\!\!\!\perp C$,
  - studied by conditional on C ...
  - or by modeling C ...

- What if $T \not\perp\!\!\!\perp C$? *identifiability problem!*
  - e.g. competing risks:
  - e.g. conditional indpt? $T \perp\!\!\!\perp C|Z$

Part IV.2.3B Censoring mechanisms

Left-censoring

e.g. the HIV RNA example: due to the lower detection limit of the "standard" assay

if HIV RNA ≤ 500, either no signal or unreliable

e.g. cost information in an insurance database

Part IV.2.3B Censoring mechanisms

Interval-censoring (cfs: Lawless, 2003; Sun, 2006)

- "the current status data": observed only $T \leq O$ or $T > O$
- "interval censoring": observed only $T \in (W, V]$ due to periodic observations

The observed data likelihood function: $\prod_{i=1}^{n}(F(V_i) - F(W_i))$

- parametric inference
- nonparametric inference (Turnbull, 1976 JRSSB)

Remarks:

- "coarsening"
- "panel counts"

Part IV.2.3C Truncation

Examples ...

- Lynden-Bell (1971, *Monthly Notices of the Royal Astronomical Society*)
  In an astronomical survey, a quantity, say, the luminosity (the brightness in comparison with that of the sun), of stars in a galaxy was observed as $Y_1, \ldots, Y_K$: what's the distn?
  the observational selection? (if $Y_i \geq O$?)

- Lagakos, et al (1988, *Biometrika*)
  In an AIDS study, the time between HIV infection and AIDS is of interest (Y), and the available data are (X_i, Y_i) for $i = 1, \ldots, n$, provided $Y_i + X_i \leq O_i$ (the observation times): what's the distn of Y?

Part IV.2.3C Truncation

Consider an event time T with information collected from a study

Recall, with censoring, available information on T_i is "coarsened" as $\min(T_i, C_i)$ for all i

The examples lead to ... ...

Truncation: the available data are $\{T_i : T_i \geq \tau_i\}$ or $\{T_i : T_i \leq \tau_i\}$ (left/right-truncated data)

Truncated data arise in many contexts ... ...
e.g. Car Warranty Claims (Hu and Lawless, 1996a,b)

Compared to censored data, truncated data provides less information on the target population.

Part IV.2.3C Truncation

Provided $\{T_i : T_i \leq \tau_i\}$ (left-truncated data)

- nonparametric approaches, e.g. Lynden-Bell and Woodroofe estimator; Woodroofe (1985)
  an identifiability problem when both nonpara models are for $Y, \mathcal{T}$: only $F_Y(\cdot)/F_Y(\tau_{max})$ is estimatable

- semiparametric approaches, e.g. Kalbfleisch and Lawless (1991); Wang (1989)
  "length bias sampling": in Lagakos's setting, if $X_i \sim$ a uniform distn (e.g. Qin and Shen, 2010)

- using additional (supplementary) info, e.g. Hu and Lawless (1996a,b)

Part IV.2.4A Analysis of incomplete data: introduction

Incomplete data are prevalent. What are incomplete data? Consider the following settings ... ...

Objective: Making inference about some aspect (parameter, finite/infinite dimensional) of a population, such as

- A. the distn of r.v. Y, or
- B. the relationship of r.v. Y with X,

based on a set of sample data: a random sample $S \subseteq \mathcal{P}$ is usually selected and Data A. $\{Y_i : i \in S\}$ or Data B. $\{(Y_i, X_i) : i \in S\}$ is designated to collect.

If the available data have less information than the designated ...
... **incomplete data**

Part IV.2.4A Analysis of incomplete data: introduction

Example A. $Y = T$, and Data A$= \{T_i : i = 1, \ldots, n\}$, iid observations on T: to estm Y's distn

- right-censored data $\{(U_i, \delta_i) : i = 1, \ldots, n\}$
- missing data $\{T_i : i \in S^*\}$, with $S^* \subset S = \{1, \ldots, n\}$

Example B. $Y = T$ and $X = Z$, and Data B$= \{(T_i, Z_i) : i = 1, \ldots, n\}$, iid observations on (T, Z): to estm $T|Z$'s conditional distn

- right-censored data $\{(U_i, \delta_i, Z_i) : i = 1, \ldots, n\}$
- missing data $\{T_i : i \in S^*\} \bigcup \{Z_i : i \in S\}$ or $\{T_i : i \in S\} \bigcup \{Z_i : i \in S^*\}$, with $S^* \subset S = \{1, \ldots, n\}$
- measurement errors $\{(T_i, m(Z)_i) : i = 1, \ldots, n\}$, with $E(m(Z)_i|Z_i) = Z_i$.

Part IV.2.4A Analysis of incomplete data: introduction

Inherent Problem. When data are incomplete, depending on how and why they are incomplete,

- our ability to make an inference may be compromised;
- not accounting for the incompleteness properly when analyzing the data can lead to severe biases.

A couple of examples ... ...

- sickle cell disease: neuro-cognitive damage? (Steen et al, 2002)
- TB contact study (Cook et al, 2011)

Most software packages, by default, delete records for which data are incomplete and conduct the "complete-case analysis".

Part IV.2.4B Analysis of incomplete data: models and methods for missing data

Consider a study to assess the efficacy of a new drug in reducing blood pressure for patients: the endpoint of interest is the decrease in blok pressure after six months.

- Y_i=subject i's reduction in blood pressure after six months
- R_i=1 or 0 corresponding to Y_i was taken or not
- $i = 1, \ldots, n$
- assume (Y_i, R_i) to be iid and the population mean $E(Y_i) = \mu$

Some terms:

- "complete data" (or full data): $\{(Y_i, R_i) : i = 1, \ldots, n\}$
- "observed data": $\{(R_i Y_i, R_i) : i = 1, \ldots, n\}$
- "complete-case data": $\{R_i Y_i : R_i = 1, i = 1, \ldots, n\}$

The sample mean with the full data: $\hat{\mu}_F = \sum_{i=1}^{n} Y_i/n$.

A natural estimator for μ with the observed data: $\hat{\mu}_C = \frac{\sum_{i=1}^{n} R_i Y_i}{\sum_{i=1}^{n} R_i}$, the complete-case sample average (observed sample mean).

As $n \to \infty$, by SLLN, a.s. $\hat{\mu}_F \to \mu$ and $\hat{\mu}_C \to \frac{E(RY)}{E(R)}$

Missing Completely at Random (MCAR): the probability of missingness is independent of the variable.

- If the data are MCAR, $R \perp\!\!\!\perp Y$ and $E(RY) = E(R)E(Y) \Longrightarrow \hat{\mu}_C$ is consistent (in fact, is also unbiased), provided $E(R) > 0$.
- How efficient is $\hat{\mu}_C$, compared to $\hat{\mu}_F$? $[\frac{\sigma^2}{nE(R)}]$
- What does it do imputing the missing observations with the average based on the observed?
- What if not MCAR?

Part IV.2.4B Analysis of incomplete data: models and methods for missing data

Not Missing at Random (NMAR): the probability of missingness depends on the variable.

With $E(R|Y) = P(R = 1|Y) = \pi(Y)$,

$$\hat{\mu}_C \to \frac{E(RY)}{E(R)} = \frac{E(Y\pi(Y))}{E(\pi(Y))} \neq E(Y) = \mu \quad (\textit{in general})$$

e.g. $\pi(y) \uparrow$ as $y \uparrow$, $\frac{E(Y\pi(Y))}{E(\pi(Y))} > \mu$.

- If NMAR, does it help imputing the missing observation with the average based on the observed?
- If NMAR, given the current formulation, no way (i) to know Y_i if $R_i = 0$ and (ii) to estm $\pi(y)$

$\Longrightarrow$ no way to find out whether MCAR or NMAR from the observed data (an inherent nonidentifiability problem). A third possibility to consider ...

...

Suppose there are additional observations W_i, $i = 1, \ldots, n$. [auxiliary covariates: they represent variables not of the primary interest for inference]
The "observed data" are now $\{(R_i Y_i, R_i, W_i) : i = 1, \ldots, n\}$.

Missing at Random (MAR): conditional on the auxiliary covariate, the probability of missingness does not depend on the primary variable:
$P(R_i = 1 | Y_i, W_i) = \pi(W_i)$, that is, $R_i \perp\!\!\!\perp Y_i | W_i$.

e.g. a survey on presidential election: gender, soci-economic status, race can be W, and the assumption of MAR ...

How to account for the missingness when MAR?

Likelihood Methods: Consider
$(Y, W) \sim f_{Y,W}(y, w) = f_{Y|W}(y|w; \gamma_1) f_W(w; \gamma_2)$.

$$\mu = E(Y) = E\{E(Y|W)\} = \int y f_{Y|W}(y|w; \gamma_1) f_W(w; \gamma_2) dy dw.$$

Since $[RY, R, W]$ is either $[Y|R = 1, W][R = 1, W]$ or $[R = 0, W]$, and $[Y|R = 1, W] = [Y|W]$ with MAR, the likelihood function

$$L(\gamma_1, \gamma_2|ObservedData) \propto \Big(\prod_{i=1}^{n} f_{Y|W}(y_i|w_i; \gamma_1)^{r_i} \Big) \Big(\prod_{i=1}^{n} f_W(w_i; \gamma_2) \Big).$$

$\Longrightarrow$ the MLE of γ_1, γ_2 and then the MLE of μ, say, $\hat{\mu}_{MLE}$.
Remark: γ_1 estm by the complete cases and γ_2 estm by all the data.
numerical challenge: computing? the EM algorithm?

Imputation:

With the "full data",
$\hat{\mu}_F = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n}\Big[R_i Y_i + (1-R_i)Y_i\Big]$.

With MAR, $E(Y_i|R_i = 0, W_i) = E(Y_i|W_i)$ is
$\int y f_{Y|W}(y|W_i;\gamma_1)dy = \mu(W_i;\gamma_1)$.

Using the MLE of γ_1, a consistent estm

$$\hat{\mu}_{IMP} = E\Big[\frac{1}{n}\sum_{i=1}^{n} Y_i|ObservedData;\hat{\gamma}_1\Big] = \frac{1}{n}\sum_{i=1}^{n}\Big[R_i Y_i + (1-R_i)\mu(W_i;\hat{\gamma}_1)\Big]$$

- Is $\hat{\mu}_{IMP}$ consistent?
- How about the efficiency of $\hat{\mu}_{IMP}$? How does it compare with $\hat{\mu}_C$ when MCAR?
- This is in fact $\hat{\mu}_{IMP}(W_i's)$; given $\hat{\gamma}_2$, an alternative:

$$\tilde{\mu}_{IMP} = \frac{1}{n}\sum_{i=1}^{n}\left[R_i Y_i + (1-R_i)\int \mu(w_i;\hat{\gamma}_1) f_W(w_i;\hat{\gamma}_2) dw_i\right]$$

- Other imputation techniques, such as to impute the missing Y_i using a random draw (or more) from $f_{Y|W}(y|W_i;\hat{\gamma}_1)$?

Inverse Probability Weighted (IPW) Complete-Case Estimator: With the "observed data", $R_i Y_i$ with $R_i = 1$ should present more than one but $1/P(R = 1|W_i)$ many individuals.
$\Longrightarrow$ another consistent estm $\hat{\mu}_{IPWCC} = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i Y_i}{\hat{\pi}(W_i)}$

This is because

$$E\Big[E\Big(\frac{RY}{\pi(W)}\Big|Y, W\Big)\Big] = E\Big[\frac{Y}{\pi(W)} E\Big(R\Big|Y, W\Big)\Big].$$

e.g. Hu, et al (2007): kindergarten readiness skills in children with sickle cell disease [cognitive impairment?]
where $\hat{\pi}(w) = \pi(w; \hat{\gamma})$ with $\hat{\gamma}$ obtained from
$\prod_{i=1}^{n} \pi(W_i; \gamma)^{R_i} \big(1 - \pi(W_i; \gamma)\big)^{1-R_i}$.

- $\hat{\mu}_{MLE}$ and $\hat{\mu}_{IMP}$ require to specify $f_{Y|W}(y|w;\gamma_1)$: what if it's misspecified?
- $\hat{\mu}_{IPWCC}$ requires to specify $P(R=1|w)=\pi(w;\gamma)$: what if it's misspecified?

$\Longrightarrow$ the following ... ...

Double Robust Estimator: an augmented inverse probability weighted complete-case estimator

$$\hat{\mu}_{AIPWCC}=\frac{1}{n}\sum_{i=1}^{n}\Big[\frac{R_iY_i}{\pi(W_i;\hat{\gamma})}+(1-\frac{R_i}{\pi(W_i;\hat{\gamma})})\mu(W_i;\hat{\gamma}_1)\Big].$$

consistent when MAR, if either of the two models is specified correctly (Why?)

- How about the efficiency of $\hat{\mu}_{AIPWCC}$?
  the optimal (most efficient) AIPWCC?
- Does it require MAR? How to check for the MAR assumption?
- What if none of the two specified models is appropriate?
- What if, when to consider a regression analysis of (Y, X), a portion of $\{X_i : i = 1, \ldots, n\}$ is missing?

Part IV.2.4B Analysis of incomplete data: models and methods for missing data

Example. To study the relationship between the concentration of HIV RNA, a viral biological marker, with a clinical outcome Y. Two blood samples of equal volume are drawn from each subject in a study. The full data are observations on (Y, X_1, X_2); however, to save on expense, some subjects' HIV RNA concentrations were obtained from the combined samples, and thus only available were the observations of $(Y, \frac{X_1+X_2}{2})$.

$\Longrightarrow$ the concentrations of those subjects are not missing but **coarsened**. (Heitjan and Rubin, 1991)

Coarsened Data: When the full data are $\{Y_i : i = 1, \ldots, n\}$, the observed data are

$$\{\mathcal{C}_i, G_{\mathcal{C}_i}(Y_i)\} : i = 1, \ldots, n$$

$\mathcal{C}$: the coarsening variable, specifying how the data are coarsened; $G_{\mathcal{C}}(Y)$ are the resulting data.

Missing, censoring are special cases of coarsening.

Part IV.2.4C Analysis of incomplete data: measurement errors (imperfectly measured data)

Example. Nutrition Studies the NHANES-I Epidemiologic Study Cohort (Jones, et al 1987)

- originally consisting of 8,596 women, interviewed about their nutrition habits and then later examined for evidence of cancer
- response Y indicates the presence of breast cancer
- predictor variables S (measured without significant error, such as age, poverty index, body mass index, etc), and predictor variables X (the nutrition variables, such as long-term saturated fat intake, known to be imprecisely measured): the measured W was a 24 hour recall and then X was computed
- the study modeled the measurement error structure using an external data set: parameters in the external study may differ from parameters in the primary study, leading to bias
- alternative: an internal subset? the Nurses' Health Study

Why it is needed to account for measurement error?

Let's see a simple example ... ...

Simple Linear Regression with Additive Error:

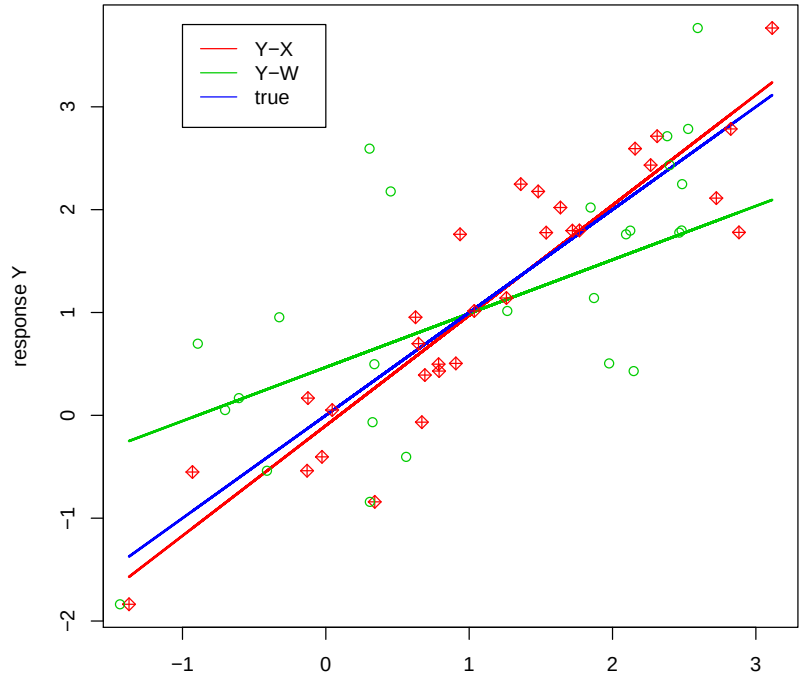
- Consider $Y = \beta_0 + \beta_1 X + \epsilon$, $X \perp\!\!\!\perp \epsilon$ and $E(X) = \mu_x$, $V(X) = \sigma_x^2$, $E(\epsilon) = 0$, $V(\epsilon) = \sigma^2$.
- Suppose X cannot be observed and instead one observes $W = X + U$, with $U \perp\!\!\!\perp X$ and $E(U) = 0$, $V(U) = \sigma_U^2$. [the classical additive measurement error model]

What if use W's observations as X's and fit the simple linear regression line? See a simulation... ...

Part III.1.4A Measurement Error: Introduction

For $i = 1, \ldots, 30$, indpt

- $X_i \sim N(1,1)$; $U_i \sim N(0,1)$; $\epsilon_i \sim N(0, .25)$
- $Y_i = 0 + 1 * X_i + \epsilon_i$



- blue line: $Y = X$; red line: $Y = 0.09955 + 1.07155X$; green line: $Y = 0.4677 + 0.5226X$

In general,

- An ordinary least squares regression of Y on W is a consistent estimator not of β_1 but $\beta_1^* = \lambda\beta_1$, where

$$\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} < 1$$

  λ: reliability ratio

- The residual variance of this regression of Y on W is

$$var(Y|W) = \sigma^2 + \frac{\beta_1^2\sigma_x^2\sigma_u^2}{\sigma_x^2 + \sigma_u^2}$$

$\Longrightarrow$ "Measurement error causes a double-whammy: not only is the slope attenuated, but the data are more noisy, with an increased error about the line" – Carroll et al (1995)

How to "correct" the bias?

Method of Moments. Note that $\beta_1 = \beta_1^*/\lambda$

- β_1^* can be estm consistently
- if λ, the reliability ratio, can be estimated?
  - $\hat{\sigma}_w^2$, the sample variance of W_i's
  - σ_u^2? If there're k_i replicate measurements of X_i,

$$\hat{\sigma}_u^2 = \frac{1}{\sum_i (k_i - 1)} \sum_i \sum_{j=1}^{k_i} (W_{ij} - \bar{W}_i)^2$$

Orthogonal Regression. If the ratio $\eta = \sigma^2/\sigma_u^2$ is known, minimize the weighted orthogonal distance of (Y, W) to the line $\beta_0 + \beta_1 X$

$$\sum_i \left[(Y_i - \beta_0 - \beta_1 X_i)^2 + \eta (W_i - X_i)^2 \right]$$

in the unknown parameters $\beta_0, \beta_1, X_1, \ldots, X_n$.

There are various models for measurement error. They may be categorized into two modeling classes:

Functional modeling.

- the classical functional models: X_i's are a sequence of unknown fixed constants
- extended to either fix or random: in the latter case no or at least minimal assumptions are made about the ditn

Structural modeling.

- the classical structural models: X_i's are regarded as r.v.s.
- usuallythe distn are parametric

Analysis of data with measurement errors:

Likelihood or Pseudo-Likelihood Approaches, or their variations

- something from Econometrics ...
  instrumental variables, *the generalized method of moments*

- data with measurement errors: an extended version of coarsening

Thank-you for your participation in this course!

What have we studied?

- *Part I. Preliminaries*
- *Part II. Parametric Interence in LIDA*
- *Part III. Nonparametric/Semi-parametric Inference*

  Part III.1. Introduction and Overview
  Part III.2. Kaplan-Meier Estimator
  Part III.3. Nonparametric Tests
  Part III.4. Cox Proportional Hazards Model

- *Part IV. Advanced Topics*

Please be friendly reminded ...

- The Presentations on March 31, April 5, and April 7.
- See the posted schedule in the course webpage/canvas page FYI.
- The final reports are due on Friday April 22 by 5:00pm.