What to do today (Feb 7, 2023)?

Part I. Introduction
Part II. Epidemiologic Concepts and Designs
Part IV. Clinical Trials

Part IV. Modern Biostatistical Approaches

Part IV.1 Incomplete Data Analysis
- **IV.1.1 Introduction**
- **IV.1.2 Models and Methods for Missing Data**
- *IV.1.3 Coarsened Data Analysis*
- *IV.1.4 Truncation*
- *IV.1.5 Measurement Errors*

Part IV.2 Other Important Topics

Part IV. Modern Biostatistical (Analytic Epidemiologic) Approaches

Part IV.1 Incomplete Data Analysis
(*supplementary*; Ref: Tsiatis, 2006 *"Semiparametric Theory and Missing Data"*)

- **Part IV.1.1 Introduction**
  Incomplete data are prevalent in practice:
  - Many studies set out in advance to collect data following a "nice" plan but do not work out quite as intended, especially when the studies involve human beings.
  - Many studies even begin with the knowledge that the desired information is not affordable.

Part IV.1.1 Introduction: Some examples of incomplete data

- Nonresponse in sampling survey

e.g. We send out questionnaries to a sample of randomly chosen individuals: some may provide only a partial answer or no answer to some questions, or, many not return the questionnaire at all.

- Dropout or noncompliance in clinical trial

e.g. In a randomized clinical trial, subjects are enrolled and then randomly assigned to one of the treatment arms: some subjects may "drop out" of the study – failing to show up for any clinic visit after a certain point, and some others may miss clinic visits occasionally or quit taking their assigned treatment.

Part IV.1.1 Introduction: Some examples of incomplete data

- Surrogate measurements

e.g. In some studies, the response of interest or some important covariate may be very expensive to obtain, such as the daily average percentage fat intake of a subject. A cheaper measurement (surrogate) is to have subjects recall the food they ate in the past 24 hurs.

- Observations on biomarkers

e.g. In AIDS studies, the time to AIDS since HIV infection has become not desirable endpoint for it takes long to collect enough information on it. Many recent AIDS studies use biomarkers, such as CD4 counts and HIV RNA as study responses. Efforts should be made to establish the association of time to AIDS with the biomarkers.

Original Objective: making an inference about some aspect (parameter, finite/infinite dimensional) of the distribution of the "full data" (i.e., the data that would have been observed if there is no data incompleteness)

Inherent Problem: when data are incomplete, depending on how and why they are missing, our ability to make an inference may be compromised. Moreover, not accounting for incomplete data properly when analyzing the data can lead to severe biases.

Most software packages, by default, delete records for which data are incomplete and conduct the "complete-case analysis". e.g. Cook et al (2011)

Serious attempts since 1980s to address the problem ... ...

Part IV.1.2 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:

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

Part IV.1.2 Models and Methods for Missing Data

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

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, $R \perp Y$ and $E(RY) = E(R)E(Y) \Longrightarrow$ $\hat{\mu}_C$ is consistent (in fact, is unbiased)
- What if not MCAR?

Part IV.1.2 Models and Methods for Missing Data

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

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

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

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

If NMAR, 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 ... ...

Part IV.1.2 Models and Methods for Missing Data

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, $R_i \perp Y_i | W_i$.

$\Longrightarrow$ understanding the missingness and then making inference about Y's distn with the observed data ... ...

For example, consider $P(R = 1 | W = w) = \pi(w; \gamma)$, say, a logistic regression model, and estm γ by maximizing
$\prod_{i=1}^{n} \pi(W_i; \gamma)^{R_i} \big(1 - \pi(W_i; \gamma)\big)^{1-R_i}$

Part IV.1.2 Models and Methods for Missing Data

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

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

$\Longrightarrow$ 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_1 estm by all the data.
numerical challenge: computing? the EM algorithm?

Part IV.1.2 Models and Methods for Missing Data

A small simulation study ...

$T \sim NE(\theta)$ with $E(T) = \theta = 3$

- iid $T_1, \ldots, T_n$ with $n = 100$
  $\Longrightarrow \hat{\theta}_F = \sum_{i=1}^{n} T_i/n$
- iid $C_i \sim NE(\phi)$ and $U_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$: right-censored observations $\Longrightarrow$
  - $\sum_{i=1}^{n} U_i/n$
  - $\hat{\theta}_{MLE} = \sum_{i=1}^{n} U_i / \sum_{i=1}^{n} \delta_i$
- iid $R_i \sim B(1, p)$ and then $R_i T_i$: MCAR $\Longrightarrow$
  - $\hat{\theta}_{MCAR} = \sum_{i=1}^{n} R_i T_i / \sum_{i=1}^{n} R_i$

Generated $m = 1000$ *sets of data to examine the performance of the estimators*

Part IV.1.2 Models and Methods for Missing Data

A small simulation study ... (cont'd)

- iid $W_1, \ldots, W_n$ with $n = 100$ from $B(1, 0.5)$
- $T_i \sim NE(\theta_1)$ if $W_i = 1$; $T_i \sim NE(\theta_0)$ if $W_i = 0$
  $E(T) = \frac{1}{2}\theta_1 + \frac{1}{2}\theta_0$
  $\Longrightarrow \hat{\theta}_F = \sum_{i=1}^{n} T_i/n$
- iid $R_i \sim B(1, p_1)$ if $W_i = 1$; $R_i \sim B(1, p_0)$ if $W_i = 0$
  - $\sum_{i=1}^{n} R_i T_i / \sum_{i=1}^{n} R_i$
  - $\frac{1}{2}\sum_{i:W_i=1}(R_i/p_1)T_i/n_1 + \frac{1}{2}\sum_{i:W_i=0}(R_i/p_0)T_i/n_0$

Generated $m = 1000$ *sets of data to examine the performance of the estimators*

Part IV.1.2 Models and Methods for Missing Data

Imputation: With the "full data",

$$\hat{\mu}_F = \frac{\sum_{i=1}^n Y_i}{n} = \frac{1}{n}\sum_{i=1}^n R_i Y_i + (1 - R_i) Y_i.$$

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

Using the MLE of γ_1, a consistent estm

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

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

Part IV.1.2 Models and Methods for Missing Data

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 $\hat{\mu}_{IPWCC} = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i Y_i}{\hat{\pi}(W_i)}$
$\hat{\pi}(w)$ is obtained from $\prod_{i=1}^{n} \pi(W_i;\gamma)^{R_i}\big(1-\pi(W_i;\gamma)\big)^{1-R_i}$.

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?]

Part IV.1.2 Models and Methods for Missing Data

- $\hat{\mu}_{MLE}$ 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 if either of the two models is specified correctly (Why?)

What to study next?

Part I. Introduction
Part II. Epidemiologic Concepts and Designs
Part III. Clinical Trials

Part IV. Modern Biostatistical (Analytic Epidemiologic) Approaches

- **Part IV.1 Incomplete Data Analysis**
  - *IV.1.1 Introduction*
  - **IV.1.2 Models and Methods for Missing Data**
  - **IV.1.3 Coarsened Data Analysis**
  - **IV.1.4 Measurement Errors**
  - *IV.1.5 Truncation*
- **Part IV.2 Other Important Topics**
  - *IV.2.1 Measures of Risks*
  - *IV.2.2 Measurement Error Revisit*
  - *IV.2.3 Confounding and Its Control*
  - *IV.2.4 Causation vs Association*
  - *What other topics that interest you?*