What to do today (Feb 9, 2023)?

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

Part IV. Modern Biostatistical Approaches

Part IV.1 Incomplete Data Analysis

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

Part IV.2 Some Other Important Topics (Chp 8 - 18, Koepsell and Weiss, 2003)

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 "complete data": $\{Y_i : i = 1, \ldots, n\}$
- 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

- *Missing Completely at Random* (MCAR): the probability of missingness is independent of the variable. (i.e. $R \perp\!\!\!\perp Y$)

- *Missing at Random* (MAR): conditional on the auxiliary covariate, the probability of missingness does not depend on the primary variable: (i.e. $R \perp\!\!\!\perp Y \big| W$)

- *Not Missing at Random* (NMAR/MNAR): the probability of missingness depends on the variable. (i.e. $R \not\perp\!\!\!\perp Y \big| X, W$)

$\Longrightarrow$ understanding the missingness and then making inference about Y's distn by the observed data accounting for the missing. For example,
Likelihood Methods: Assume
$(Y, W) \sim f_{Y,W}(y, w) = f_{Y|W}(y|w; \gamma_1) f_W(w; \gamma_2)$. 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 the MLE of $\mu = E(Y)$. *practical challenges?*

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

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

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

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

consistent if either of the two models is specified correctly

(Why?)

Part IV.1.3A Coarsened Data Analysis: Coarsening vs Missing

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)

Part IV.1.3A Coarsened Data Analysis: Coarsening vs Missing

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

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

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

Usually, $\mathcal{C} = \infty$ is used to indicate an observation of Z:
$G_\infty(Z) = Z$
the complete-case data are $\{Z_i : \mathcal{C}_i = \infty, i = 1, \dots, n\}$

Missing is a special case of coarsening.

Part IV.1.3B Coarsened Data Analysis: Coarsening Mechanisams

- *Coarsening completely at random* (CCAR)

$$P(\mathcal{C} = r|Z) = \pi(r), \forall r; \quad i.e., \mathcal{C} \perp Z$$

- *Coarsening at random* (CAR)

$$P(\mathcal{C} = r|Z) = \pi(r, G_r(Z)), \forall r; \quad i.e., \mathcal{C} \perp Z | G_{\mathcal{C}}(Z)$$

- *Not coarsening at random* (NCAR)
  There are $z_1 \neq z_2$ such that $G_r(z_1) = G_r(z_2)$ but
  $P(\mathcal{C} = r|Z = z_1) \neq P(\mathcal{C} = r|Z = z_2)$

Part IV.1.3B Coarsened Data Analysis: Coarsened Data Likelihood

Suppose $(\mathcal{C}, Z) \sim f_{\mathcal{C},Z}(r, z; \psi, \beta, \eta) = P(\mathcal{C}|Z = z; \psi) f_Z(z; \beta, \eta)$
With CAR,

$$
\begin{aligned}
& (\mathcal{C}, G_{\mathcal{C}}(Z)) \sim f_{\mathcal{C}, G_{\mathcal{C}}(Z)}(r, g_r; \psi, \beta, \eta) \\
= & \int_{z: G_r(z) = g_r} P(\mathcal{C} = r | Z = z; \psi) f_Z(z; \beta, \eta) dz = \pi(r, g_r; \psi) f_{G_r(Z)}(g_r; \beta, \eta)
\end{aligned}
$$

(the above notation for discrete/continuous Z ... ...) the likelihood

function of (ψ, β, η) with the observed (coarsened) data:

$$
\prod_{i=1}^{n} \pi(r_i, g_{r_i}; \psi) \prod_{i=1}^{n} f_{G_{\mathcal{C}}(Z)}(g_{r_i}; \psi, \beta, \eta)
$$

$\Longrightarrow$ the likelihood based approaches: estm and testing
computationally not easy ... ...

Part IV.1.4 Measurement Error

(Refs: "Measurement Error in Nonlinear Models" by Carroll, Ruppert and Stefanski, 1995;
"Measurement Error in Nonlinear Models: A Modern Perspective" by Carroll, Ruppert, Stefanski and Crainiceanu, 2006)

- This section focuses on an introduction to the problem of (quantitative!) predictors measured with errors.
- Misclassification, discussed in Chp 10 of Koepsell and Weiss (2003), will be covered in a section of **Part IV.2**

Part IV.1.4A Measurement Error: Introduction

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)
- 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

Part IV.1.4A Measurement Error: Introduction

Why it is needed to account for measurement error?
Let's see a simple example ... ...

Simple Linear Regression with Additive Error:

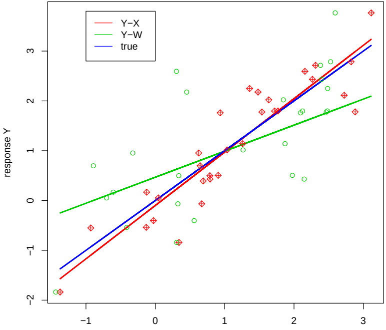
- Consider $Y = \beta_0 + \beta_1 X + \epsilon$, $X \perp \epsilon$ and $E(X) = \mu_x$, $V(X) = \sigma_x^2$, $E(\epsilon) = 0$, $V(\epsilon) = \sigma^2$.
- Suppose X cannot be observed and instead one observes $W = X + U$, with $U \perp X$ and $E(U) = 0$, $V(U) = \sigma_U^2$. [the classical additive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

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

Part IV.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: $Y = 0.4677 + 0.5226X$

Part IV.1.4A Measurement Error: Introduction

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)

Part IV.1.4A Measurement Error: Introduction

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,

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

Remark. Sometimes $\hat{\lambda} = (\hat{\sigma}_w^2 - \hat{\sigma}^2)/\hat{\sigma}_w^2$ can be negative. Further discussions are needed.

Part IV.1.4A Measurement Error: Introduction

How to “correct” the bias?

- *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 $\beta_0, \beta_1, X_1, \ldots, X_n$.

Remarks.

- η needs to be estm; if not properly specified, it may lead to “over correction”.
- The resulting estm of β_0, β_1 are the functional MLE with $X_1, \ldots, X_n$ as unknown fixed constants, assuming $(\epsilon_i, U_i) \sim$ normal, iid.

Part IV.1.4B Measurement Error: Modeling and Inference

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

Part IV.1.4B Measurement Error: Modeling and Inference

Given a specification of $[X, W|S]$ (or in the form of $[X|W, S]$, or $[W|X, S]$), procedures for making inference about $[Y|X, S]$:

Likelihood or Pseudo-Likelihood Approaches, or their variations

- parametric, semi-parametric, semi-nonparametric
- with Y continuous, or categorical (binary, count)
- with coarsened response data (e.g. censored survival times), with some X_i observed, ...

Remark:

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

Part IV.1.4C Measurement Error: vs Coarsening?

Measurement error as a missing data problem, or, more general, a coarsened data problem?

Recall the simple example in **Part IV.1.4A**: *Simple Linear Regression with Additive Error*

- Consider $Y = \beta_0 + \beta_1 X + \epsilon$, $X \perp \epsilon$ 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 $U \perp X$ and $E(U) = 0$, $V(U) = \sigma_U^2$.

We have ... ...

- the full data: $Z_i = (Y_i, X_i)$, $i = 1, \ldots, n$
- the observed data: $Z_i^* = (Y_i, W_i)$, $i = 1, \ldots, n$

Part IV.1.4C Measurement Error: vs Coarsening?

Any appropriate $\mathcal{C}_i$ *(observable) and* $G_{\mathcal{C}}(\cdot)$ *such that* $Z_i^* = G_{\mathcal{C}_i}(Z_i)$*?*

Recall $W_i = X_i + U_i$ depends on U_i, something unobservable.

$\Longrightarrow$ viewing $G_{\mathcal{C}}(\cdot)$ as a stochastic mapping, instead of a deterministic one, with a given $\mathcal{C}$?

an extended version of coarsening ... ...

What to study next class?

Part IV. Modern Biostatistical (Analytic Epidemiologic) Approaches

Part IV.1 Incomplete Data Analysis (*supplementary*)

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

Part IV.2 Some Other Important Topics (Chp 8 - 18, Koepsell and Weiss, 2003)