What to do today (Feb 14, 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

- *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 Some Other Important Topics (Chp 8 - 18, Koepsell and Weiss, 2003)

- **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*

Part IV.3 Selected Widely-Used Algorithms

Part IV.1.4A Measurement Error: Introduction

Why it is needed to account for measurement error?

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 ... ...

Part IV.1.5A Truncation: Introduction

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.1.5 Truncation: as Coarsening?

- the full data: $Z_i = (Y_i, \mathcal{T}_i)$ with $i \in \mathcal{P}$
- the observed data: $\{Z_i : Y_i \geq \mathcal{T}_i, i \in \mathcal{P}\}$

Any observed coarsening variable $\mathcal{C}$ and $G_{\mathcal{C}}(\cdot)$ presents the observation selection?
Recall that no information about individual i, if $Y_i < \mathcal{T}_i$ ... ...

$\Longrightarrow$ mechanism of incompleteness, different from what studied before

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

Part IV.1.5 Truncation: Analysis of 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), and Qin and Shen (2010)
  length bias sampling: in Lagakos's setting, if $X_i \sim$ a uniform distn

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

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

Part IV.2.1 Measure of Risk

Example. Crib death of SIDS (Sudden Infant Death Syndrome)
Cumulative incidence of crib death was recorded based on usual sleeping position of 2607 one-month old Tasmanian infants born 1988-1991 (Dwyer et al., 1991): **to study how X affects Y?** $\Longrightarrow$ measure of risk to SIDS with a sleeping position?

Cumulative Incidence:

Usual Sleeping (X) Position	SIDS Death? (Y) yes	no	Total
prone	$n_{11} = 9$	$n_{12} = 837$	$n_{1+} = 846$
other	$n_{21} = 6$	$n_{22} = 1755$	$n_{2+} = 1761$
Total	$n_{+1} = 15$	$n_{+2} = 2592$	$n_{++} = 2607$

Part IV.2.1 Measure of Risk

A 2×2 **table with the row and column variables** X **and** Y**, both binary:** for $i = 1, 2$ and $j = 1, 2$,

- the joint and marginal prob
  $\pi_{ij} = P(X = i, Y = j)$; $\pi_{i+} = P(X = i)$; $\pi_{+j} = P(Y = j)$;
- the conditional prob
  $P(X = i | Y = j) = \pi_{ij}/\pi_{+j}$; $P(Y = j | X = i) = \pi_{ij}/\pi_{i+}$

Probabilities:

Usual Sleeping (X) Position	SIDS Death? (Y) yes	no	Total
prone	π_{11}	π_{12}	π_{1+}
other	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	$\pi_{++} = 1$

Part IV.2.1 Measure of Risk

A 2×2 table with the row and column variables X and Y, both binary:

Probabilities:

Usual Sleeping (X) Position	SIDS Death? (Y) yes	no	Total
prone	π_{11}	π_{12}	π_{1+}
other	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	$\pi_{++} = 1$

the MLE of the prob:

- with the multinomial sampling (fixed n_{++}, e.g. cohort study): $\hat{\pi}_{ij} = n_{ij}/n_{++}$ and thus $\hat{\pi}_{i+}$ etc.
- with the purposive sampling (fixed n_{+j}, e.g. case-control study): $\widehat{\frac{\pi_{ij}}{\pi_{+j}}} = n_{ij}/n_{+j}$

Part IV.2.1 Measure of Risk

Probabilities:

Usual Sleeping (X) Position	SIDS Death? (Y) yes	no	Total
prone	π_{11}	π_{12}	π_{1+}
other	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	$\pi_{++} = 1$

Measures of Risk

- **excess risk:** $ER = P(Y=1|X=1) - P(Y=1|X=2)$
  $= \frac{\pi_{11}}{\pi_{1+}} - \frac{\pi_{21}}{\pi_{2+}}$ [attributable risk to the exposed]
- **relative risk:** $RR = P(Y=1|X=1) \big/ P(Y=1|X=2) = \frac{\pi_{11}}{\pi_{1+}} \Big/ \frac{\pi_{21}}{\pi_{2+}}$
- **odds ratio:** $OR = \frac{P(Y=1,X=1)}{P(Y=2,X=1)} \Big/ \frac{P(Y=1,X=2)}{P(Y=2,X=2)} = \frac{\pi_{11}}{\pi_{12}} \Big/ \frac{\pi_{21}}{\pi_{22}}$
  - $RR \approx OR$ when $Y = 1$ is a rare event

Part IV.2.1 Measure of Risk

Estimation for the measures of risk:

Cumulative Incidence:

Usual Sleeping (X) Position	SIDS Death? (Y) yes	no	Total
prone	n_{11}	n_{12}	n_{1+}
other	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n_{++}

the MLE of the Measures of Risk: with the multinomial sampling

- **excess risk:** $\hat{ER} = \frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}} = I_E - I_O$ [diff of cumulative incidences with E vs O]
- **relative risk:** $\hat{RR} = \frac{n_{11}}{n_{1+}} \Big/ \frac{n_{21}}{n_{2+}} = I_E / I_O$
- **odds ratio:** $\hat{OR} = \frac{n_{11}}{n_{12}} \Big/ \frac{n_{21}}{n_{22}} = \frac{n_{11} n_{22}}{n_{12} n_{21}} = \frac{I_E/(1-I_E)}{I_O/(1-I_O)}$

Part IV.2.1 Measure of Risk

Estimation for the measures of risk:

Cumulative Incidence:

Usual Sleeping (X) Position	SIDS Death? (Y) yes	no	Total
prone	n_{11}	n_{12}	n_{1+}
other	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n_{++}

the MLE of the Measures of Risk: with the purposive sampling

- **excess risk:** $\hat{ER} = \frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}}$
- **relative risk:** $\hat{RR} = \frac{n_{11}}{n_{1+}} \Big/ \frac{n_{21}}{n_{2+}}$
- **odds ratio:** $\hat{OR} = \frac{n_{11}/n_{1+}}{n_{12}/n_{1+}} \Big/ \frac{n_{21}/n_{2+}}{n_{22}/n_{2+}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$

Part IV.2.1 Measure of Risk

Estimation for the measures of risk:

Cumulative Incidence:

Usual Sleeping (X) Position	SIDS Death? (Y) yes	no	Total
prone	n_{11}	n_{12}	n_{1+}
other	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n_{++}

Confidence Intervals for the Measures of Risk: for example,

- **odds ratio:** $\hat{OR} \pm 1.96 SE_{\hat{OR}}$
- **odds ratio:** $\exp\left\{\log(\hat{OR}) \pm 1.96 SE_{\log(\hat{OR})}\right\}$ with
  $SE^2_{\log(\hat{OR})} = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$
  $\log(OR)$=the coef to X in the logistic regression model of Y on X

Part IV.2.1 Measure of Risk

Revisit to Example of Crib death of SIDS

Cumulative Incidence:

Usual Sleeping (X) Position	SIDS Death? (Y) yes	no	Total
prone	$n_{11} = 9$	$n_{12} = 837$	$n_{1+} = 846$
other	$n_{21} = 6$	$n_{22} = 1755$	$n_{2+} = 1761$
Total	$n_{+1} = 15$	$n_{+2} = 2592$	$n_{++} = 2607$

- a cohort study
  - $\hat{ER} = \frac{9}{846} - \frac{6}{1761} = 0.723\%$
  - $\hat{RR} = \frac{9}{846} / \frac{6}{1761} = 3.122$ [Does exposure cause disease?], with 95% CI $(1.12, 8.74)$
  - $\hat{OR} = 3.145$ [Does exposure cause disease?], with 95% CI $(1.1158, 8.8654)$

Part IV.2.1 Measure of Risk

Revisit to Example of Crib death of SIDS

Cumulative Incidence:

Usual Sleeping (X) Position	SIDS Death? (Y) yes	no	Total
prone	$n_{11} = 9$	$n_{12} = 837$	$n_{1+} = 846$
other	$n_{21} = 6$	$n_{22} = 1755$	$n_{2+} = 1761$
Total	$n_{+1} = 15$	$n_{+2} = 2592$	$n_{++} = 2607$

- relative risk vs risk difference?
- study design?

What to study next?

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

Part IV. Analytic Epidemiology

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

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

IV.2.1 Measures of Risks
IV.2.2 Measurement Error Revisit
IV.2.3 Confunding and Its Control
IV.2.4 Causation vs Association

Part IV.3 Selected Widely-Used Algorithms

IV.3.1 Bootstrap and Related
IV.3.2 EM Algorithm and Related