BIOS 736, Fall 2023: Notes for Class # 4

SOME MISCELLANEOUS STRUCTURAL MEASUREMENT ERROR EXAMPLES

Examples we'll look at today:

a) Linear regression with a predictor (X) subject to normal measurement error, using replicates or longitudinal observations

b) Misclassification of the outcome (Y) in logistic regression (see slides)

a) One example of using "replicates" or "reproducibility study" observations to correct for structural measurement error

Suppose we have repeated measures (W_{i1}, W_{i2}, …,) of an error-prone surrogate for a true exposure (X_i). These repeated measures are taken on all (or at least a reasonable subset) of the study participants

Assume the following simple linear regression "TDM":

true exposure

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

Also, assume the W's relate to the X's in the following way:

MEM

$$W_{ij} = X_i + \varepsilon_{ij} \;,$$

where the errors ε_{ij} each have mean 0 and follow a compound-symmetric covariance structure

(PROCEED WITH HANDWRITTEN NOTES FROM HERE)

equivalently, consider a simple mixed linear model

$$W_{ij} = \mu + b_i + \epsilon_{ij} \sim \text{iid } N(0, \sigma_\epsilon^2)$$

$b_i \sim \text{iid } N(0, \sigma_b^2)$

b_i's and ϵ_{ij}'s mutually independent

$$\boxed{W_{ij} = \mu + b_i + \epsilon_{ij}}$$

SAS proc mixed
R

Let's start w/ balanced data ($j = 1, \ldots, n$)

Define "X_i" $= \boxed{\mu_i = \mu + b_i}$ "true" latent mean for subject i

TDM: $Y_i = \beta_0 + \beta_1 \mu_i + \epsilon_i$

of interest

Typical "surrogate" for μ_i = $\bar{W}_i = \frac{\sum_j W_{ij}}{n}$

Using $\bar{W}_i$ in the "naive" regressn incurs a bias in $\hat{\beta}_1$

How to charactize this bias?

$\Rightarrow$ can show $E(\mu_i | \bar{W}_i) = \gamma \bar{W}_i + (1-\gamma)\mu$

"shrinkage" toward μ dictated by

$$\boxed{\gamma = \frac{n\sigma_b^2}{n\sigma_b^2 + \sigma_w^2}} \qquad 0 < \gamma < 1$$

Let's try regression calibration (RC):

"Naive" OLS of Y vs. $\bar{W}_i$ $\Rightarrow$ what happens?

$$E(Y_i | \bar{W}_i) = \beta_0 + \beta_1 E(\mu_i | \bar{W}_i)$$

$$= \beta_0 + \beta_1 \left[\gamma \bar{W}_i + (1-\gamma)\mu \right]$$

$= \beta_0^* + \beta_1^* \bar{w}_i$ "link preserved"

$\beta_0^* \to \beta_0 + \beta_1(1-\gamma)\mu$

$\beta_1^* \to \hat{\beta}_1^* \to \beta_1 \gamma$ "attenuation"

$\Rightarrow \hat{\beta}_1 = \dfrac{\hat{\beta}_1^*}{\hat{\gamma}} \to$ plug in ML or REML estimates of $\sigma_b^2 + \sigma_\epsilon^2$

Brunekreef, Noy, Clausing, 1987 (AJE)
design considerations

(?) What if we had <u>unbalanced</u> data ($j = 1, .., n_i$)
$i = 1, .., k$ ⇓ # subjects

$\Rightarrow$ Now, $var(\mu_i | \bar{w}_i)$ is no longer constant across subjects.

$\Rightarrow$ No "clean" corrected closed-form $\hat{\beta}_1$?

<u>options</u>

1) Quasi-likelihood $\Rightarrow$ estimating eqns. weighted by $var^{-1}(\mu_i | \bar{w}_i)$
Liang + Liu (1991); Lyles + Kupper (1997)

2) <u>ML</u>: $\mathcal{L}(\underset{\sim}{\theta}, \underset{\sim}{\psi} ; \underset{\sim}{Y}, \underset{\sim}{W})$

$\underset{\sim}{\psi} \to (\mu, \sigma_b^2, \sigma_w^2) \to$ secondary parms (from MEM)

$\underset{\sim}{\theta} \to$ primary parms? $\to (\beta_0, \beta_1, \sigma_\epsilon^2)$ TDM

$$= \prod_{i=1}^{k} f(Y_i, \underset{\sim}{W}_i; \underset{\sim}{\theta}, \underset{\sim}{\psi}) = \prod_{i=1}^{k} \int_{-\infty}^{\infty} f(Y_i, \underset{\sim}{W}_i, \underset{\text{latent}}{b_i}; \underset{\sim}{\theta}, \underset{\sim}{\psi})\, db_i$$

$$= \prod_{i=1}^{k} \int_{-\infty}^{\infty} f(Y_i, \underset{\sim}{W}_i \mid b_i)\, f(b_i)\, db_i$$

$$= \prod_{i=1}^{k} \int_{-\infty}^{\infty} \underbrace{f(Y_i \mid b_i; \underset{\sim}{\theta})}_{\text{TPM}}\, \underbrace{f(\underset{\sim}{W}_i \mid b_i; \underset{\sim}{\psi})}_{\substack{W_{ij}\text{'s are independent given } b_i \\ (\text{MEM})}}\, \underbrace{f(b_i; \underset{\sim}{\psi})}_{\text{MEM}}\, db_i$$

*NLMIXED, SAS IML; R

Can we extend this to the "randomized regression" model:

$$\Rightarrow \underset{\text{longitudinal}}{W_{ij}} = (\alpha + \underset{\text{random int. deviation}}{a_i}) + (\beta + \underset{\text{random slope deviation}}{b_i})\underset{\text{time}}{t_{ij}} + \epsilon_{ij}$$

$i = 1, \dots, k$ subjects.
$j = 1, \dots, n_i$ obs. per subject

$$\begin{pmatrix} a_i \\ b_i \\ \epsilon_{ij} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \sigma_{\alpha\beta} & 0 \\ \sigma_{\alpha\beta} & \sigma_\beta^2 & 0 \\ 0 & 0 & \sigma_\epsilon^2 \end{pmatrix} \right)$$

"True" intercept $\alpha + a_i$ "True" slope $\beta + b_i$ (latent)

These could be exposures of interest to relate to health outcomes.

TDM : $Y_i = \beta_0 + \beta_1 \underset{\downarrow}{\alpha_i} + \beta_2 \underset{\downarrow}{\beta_i} + e_i$

$(\alpha + a_i) \qquad (\beta + b_i)$

If data are <u>balanced</u>, easy to correct

"naive" regression

$Y_i = \theta_0^* + \theta_1^* \hat{\alpha}_{i,OLS} + \theta_2^* \hat{\beta}_{i,OLS} + \epsilon_i^*$

OLS estimates

$\hat{\theta}_1^* \longrightarrow \lambda_1 \theta_1 \qquad \hat{\theta}_2^* \longrightarrow \lambda_2 \theta_2$

Lyles and McFarlane (2000)

These "clean" results require not just balanced data on W's, but equally-timed and equally-spaced W's.

What if <u>not</u>?

times

$\mathcal{L}(\underset{\sim}{\theta}, \underset{\sim}{\psi}; \underset{\sim}{Y}, \underset{\sim}{W} \mid \underset{\sim}{t})$

↓ steps

independent (MEM)

$$= \prod_{i=1}^{k} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \overbrace{f(Y_i \mid a_i, b_i, \underset{\sim}{t}_i; \underset{\sim}{\theta})}^{\text{TDM}} \times f(\underset{\sim}{W}_i \mid a_i, b_i; \underset{\sim}{\psi}) \times \underset{\downarrow\ \text{MEM}}{f(a_i, b_i)}\, da_i\, db_i$$

* NLMIXED

Wannemuehler et al., 2005 application Stats in Med.

Daowen Zhang + Marie Davidian (1994 ?)

$\boxed{\text{theory}}$

b) Misclassification of the outcome (Y) in logistic regression

- Slides to be presented from a 2012 CDC seminar, where the topic has to do with handling (potentially differential) misclassification of the outcome variable in logistic regression via the use of internal validation data. They summarize the contents of the following paper, later extended in various ways by former graduate Dr. Li Tang and others:

Lyles RH, Tang L, Superak HM, King CC, Celentano DD, Lo Y, Sobel JD. Validation data-based adjustments for outcome misclassification in logistic regression: An illustration. *Epidemiology* **22,** 589-598 (2011).