BIOS 755: Missing Data in Longitudinal Studies II

Alexander McLain

March 27, 2023

Incomplete Data Models

- Distinguish between dependent variable $\boldsymbol{y} = \boldsymbol{y}^O$ for $R = 1$ and $\boldsymbol{y} = \boldsymbol{y}^M$ for $R = 0$. and independent variables (all observed): X time, group,
- GEE assumes special case of MCAR

$$P(\boldsymbol{R}|\boldsymbol{y}, \boldsymbol{X}) = P(\boldsymbol{R}|\boldsymbol{X}) \text{ for all } \boldsymbol{y}$$

$\Rightarrow$ conditional on covariates, R is independent of both $\boldsymbol{y}^O$ and $\boldsymbol{y}^M$ ("covariate-dependent missingness")

Incomplete Data Models

- Likelihood-based methods assume MAR

$$P(\boldsymbol{R}|\boldsymbol{y}, \boldsymbol{X}) = P(\boldsymbol{R}|\boldsymbol{X}, \boldsymbol{y}^O) \text{ for all } \boldsymbol{y}^M$$

  ⇒ conditional on covariates and observed values of the dependent variable, $\boldsymbol{R}$ is independent of $\boldsymbol{y}^M$ "ignorable non-response" (Laird, 1988)
- But if baseline covariates are missing, all their $\boldsymbol{y}$ data are missing from our analyses (whether they were observed or not).
- In that situation, we'd need to assume that the probability they are missing is independent of the $\boldsymbol{y}$ data.

Simulation Study

- Data from 5000 subjects were simulated according to:

$$y_{ij} = \beta_0 + \beta_1 T_j + \beta_2 G_i + \beta_3 (G_i \times T_j) + \upsilon_{0i} + \upsilon_{1i} T_j + \varepsilon_{ij}$$

$T_j = 0, 1, 2, 3, 4$ for five timepoints
$G_i =$ dummy-code (0 or 1) with half in each group

Regression coefficients:
$\beta_0 = 25$, $\beta_1 = -1$, $\beta_2 = 0$, and $\beta_3 = -1$

$\Rightarrow$ the population means are:
25, 24, 23, 22, and 21 for $Grp = 0$
25, 23, 21, 19, and 17 for $Grp = 1$

Variance parameters:
$\sigma^2_{\upsilon_0} = 4$, $\sigma^2_{\upsilon_1} = .25$, $\sigma_{\upsilon_{01}} = -.1$ $(\rho = -.1)$, $\sigma^2 = 4$

Simulation Study

- The population variance-covariance matrix,

$$V(\boldsymbol{y}) = \boldsymbol{Z}\Sigma_{\upsilon}\boldsymbol{Z}' + \sigma^2\boldsymbol{I}$$

$$V(\boldsymbol{y}) = \begin{bmatrix} 8.00 & 3.90 & 3.80 & 3.70 & 3.60 \\ 3.90 & 8.05 & 4.20 & 4.35 & 4.50 \\ 3.80 & 4.20 & 8.60 & 5.00 & 5.40 \\ 3.70 & 4.35 & 5.00 & 9.65 & 6.30 \\ 3.60 & 4.50 & 5.40 & 6.30 & 11.20 \end{bmatrix}$$

Scenarios

- **Complete data:** no missing data.
- **50% random missing:** 50% missing data at every timepoint; completely random and unrelated to any variable.
- **Time related dropout:** dropout rates of 0%, 25%, 50%, 75%, 87.5% for the five timepoints. If a subject was missing at a timepoint, then they were also missing at all later timepoints; these rates indicate the percentage of the original sample that were missing at each of these timepoints.

Scenarios

- **Group by time related dropout:** dropout rates of
  0%, 23%, 46%, 70% and 83% for G=0
  0%, 27%, 55%, 81% and 91% for G=1
  ⇒ notice that these missing data scenarios are all MCAR (as long as analysis model includes time, G, and G by time)

Results

- MCAR Simulation Results Mixed effects Model estimates (standard errors)

	β_0	β_1	β_2	β_3	$\sigma^2_{\upsilon 0}$	$\sigma_{\upsilon 01}$	$\sigma^2_{\upsilon 1}$	σ^2
simulated value:	25	-1	0	-1	4	-.1	.25	4
complete data	24.969	-.994	-.001	-.986	3.918	-.057	.239	3.991
	(.050)	(.016)	(.071)	(.023)	(.129)	(.032)	(.014)	(.046)
50% random	24.991	-1.024	-.087	-.933	3.811	-.020	.199	4.070
missing	(.063)	(.023)	(.089)	(.032)	(.193)	(.056)	(.025)	(.083)
Time-related	24.989	-.968	.019	-1.021	3.853	-.062	.229	4.000
dropout	(.053)	(.028)	(.075)	(.040)	(.150)	(.060)	(.032)	(.074)
Group by Time	24.991	-.977	.041	-1.014	3.872	-.048	.234	3.994
related dropout	(.053)	(.026)	(.075)	(.041)	(.150)	(.059)	(.031)	(.073)

MAR and MNAR Scenarios

- **MAR(a):** if the value of the dependent variable was lower than 23, then the subject dropped out at the next timepoint (i.e., they were missing at the next and all subsequent timepoints).
- **MAR(b):** the MAR specification was different for the two groups. For Grp $= 1$, if the dependent variable was lower than 23, then the subject dropped out at the next timepoint, however for Grp $= 0$, if the dependent variable was greater than 25.5 then the subject dropped out at the next timepoint.
- **MNAR:** after the first timepoint, if the value of the dependent variable was lower than 21.5, then the subject was missing at that timepoint and all subsequent timepoints.

MAR and MNAR Simulation Results - Estimates (standard errors)

	β_0	β_1	β_2	β_3	$\sigma^2_{\upsilon_0}$	$\sigma_{\upsilon_{01}}$	$\sigma^2_{\upsilon_1}$	σ^2
simulated value:	25	-1	0	-1	4	-.1	.25	4
MAR(a)								
MRM	24.996	-1.039	-.010	-.969	3.981	-.064	.233	3.873
	(.053)	(.025)	(.075)	(.041)	(.158)	(.065)	(.032)	(.078)
GEE1	25.281	-1.164	.019	-1.001				
	(.058)	(.037)	(.097)	(.085)				
MAR(b)								
MRM	24.999	-1.003	-.016	-1.004	4.050	-.082	.229	3.812
	(.053)	(.022)	(.075)	(.039)	(.154)	(.064)	(.027)	(.073)
GEE1	24.635	-.714	.634	-1.532				
	(.055)	(.030)	(.097)	(.090)				
MNAR								
MRM	24.956	-.233	.027	-.552	3.856	-.943	.319	3.020
	(.049)	(.020)	(.070)	(.035)	(.131)	(.051)	(.025)	(.053)
GEE1	25.051	-.386	.016	-.583				
	(.049)	(.020)	(.071)	(.034)				

Misspecification of Variance-Covariance

- Re-analysis of the simulated MAR-generated data, however using only a random-intercepts model:

$$y_{ij} = \beta_0 + \beta_1 Time_j + \beta_2 Grp_i + \beta_3 (Grp_i \times Time_j) + v_{0i} + e_{ij}$$

- This is a misspecified model for the variance-covariance structure because the random slope term was omitted from the analysis.

Mis-specified MAR Simulation Results - Estimates (standard errors)

	β_0	β_1	β_2	β_3	σ^2_{v0}	σ_{v01}	σ^2_{v1}	σ^2
simulated value:	25	-1	0	-1	4	-.1	.25	4
MRM with MAR(a)	24.938	-.891	-.009	-.949	3.722			4.329
	(.053)	(.021)	(.075)	(.036)	(.136)			(.072)
MRM with MAR(b)	24.998	-1.048	-.069	-.805	3.880			4.288
	(.053)	(.018)	(.076)	(.035)	(.138)			(.070)

Results

- These random-intercepts analyses yield biased results, in particular for the time-related parameters β_1 and β_3.
- Performing any full-likelihood analysis, even with missing data following an MAR mechanism, does not guarantee that the correct results will be obtained (need to have the mean structure and variance-covariance structure of $\boldsymbol{y}$ correctly modeled).

Testing MCAR

- If MCAR, then either GEE or Mixed Effects models are fine (provided that the covariate matrix $\boldsymbol{X}_i$ includes predictors of missingness).
- If MAR, then GEE does not perform well, whereas Mixed Effects models analysis is acceptable (as long as the mean and variance-covariance structures are correctly modeled).
- It is useful to determine whether MCAR is acceptable or not.
- Distinction between MCAR and MAR is that missingness cannot depend on observed values of the dependent variable, $\boldsymbol{y}_i^O$, in the former, but can in the latter.
- Tests of MCAR are based on analyses involving $\boldsymbol{y}_i^O$.

Testing MCAR in 2-timepoint study

- Suppose all subjects have data at time 1, but some are missing at time 2
- Define $D_i = 0$ for subjects with data at both timepoints, $D_i = 1$ for subjects with data at first timepoint only
- Compare y_1 between these two groups ($D_i = 0$ vs. $D_i = 1$); for MCAR y_1 data should not differ between the groups.
- Logistic regression model

$$\log\left\{\frac{P(D_i = 1)}{1 - P(D_i = 1)}\right\} = \alpha_0 + \alpha_1 y_{i1} + \boldsymbol{\alpha}_2 \boldsymbol{x}_i + \boldsymbol{\alpha}_2 (y_{i1} \times \boldsymbol{x}_i)$$

- α_1 and α_3 are vectors of regression coefficients for $\boldsymbol{x}_i$ and their interactions with y_{i1} $\Rightarrow$ MCAR dictates that $\alpha_1 = \alpha_3 = 0$.

Testing MAR vs MNAR

- The only way to test the MAR assumption (i.e., that $\boldsymbol{R}$ is independent of $\boldsymbol{y}^M$ given $\boldsymbol{y}^O$ and $\boldsymbol{X}_i$) is to have some measure on the missing data.
- This can be accomplished by a follow-up with phone calls to a group of the non-respondents.
- Mostly we will not have any data.
- The only quantitative solution in this case is to do the analysis with an MAR model and an MNAR model (i.e., selection, pattern-mixture, or shared parameter).
- The most common way to "test" the MAR assumption is to use your scientific knowledge of the data and the field.