Missing data lecture 9: MAR vs. MAAR (again) and Data Coarsening

Clarifying MAR vs. unit MAR

Let there be n units, of which we're interested in measuring K variables. Let the $n \times K$ matrix of observations be denoted Y, with elements y_{ij} while a realization of this matrix is called $\tilde{y}$, with elements $\tilde{y}_{ij}$. Let the matrix of missingness indicators be denoted M, elements m_{ij}, with a particular realization $\tilde{m}$, with elements $\tilde{m}_{ij}$. Let $Y_{(0)} = \{y_{ij} \mid m_{ij} = 0, i = 1, \dots, n, j = 1, \dots, K\}$. Let $Y_{(1)} = \{y_{ij} \mid m_{ij} = 1, i = 1, \dots, n, j = 1, \dots, K\}$. Let $\mathcal{Y}_{(1)}$ be the sample space of the missing values. Let a realization of these sets of variables be $\tilde{y}_{(0)}$ and $\tilde{y}_{(1)}$.

The joint likelihood of the observed data and the missingness indicators is:

$$L_{\text{full}}(\theta, \phi \mid \tilde{y}_{(0)}, \tilde{m}) = \int_{\mathcal{Y}_{(1)}} f_Y(\tilde{y}_{(0)}, y_{(1)} \mid \theta) P(M = \tilde{m} \mid Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y_{(1)}, \phi) dy_{(1)}$$

The definition of MAR from the book is as follows:

$$f_{M|Y}(\tilde{m} \mid Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y_{(1)}, \phi) = f_{M|Y}(\tilde{m} \mid Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y^*_{(1)}, \phi)$$

for all $y_{(1)}, y^*_{(1)}, \phi$.

The definition of MAAR is:

$$f_{M|Y}(m \mid Y_{(0)} = y_{(0)}, Y_{(1)} = y_{(1)}, \phi) = f_{M|Y}(m \mid Y_{(0)} = y_{(0)}, Y_{(1)} = y^*_{(1)}, \phi)$$

for all $m, y_{(0)}, y_{(1)}, y^*_{(1)}, \phi$.

Unit missingness

When we have an assumption that observations and missingness for units can be considered conditionally independent given parameters θ and ϕ, we get unit MAR instead of MAR.

Let Y_i be the i^{th} row of the matrix Y, or the length K random vector representing observations for unit i, while $\tilde{y}_i$ is a particular realization of this random vector and y_i is a dummy vector. Similarly let M_i be the i^{th} row of the matrix M, with particular realization $\tilde{m}_i$ and m_i a dummy vector. Furthermore, let $Y_{i(0)}, Y_{i(1)}$ be the observed and missing random vectors of y_i, while $\tilde{y}_{i(0)}, \tilde{y}_{i(1)}$ are realizations of these vectors.

Then the joint distribution of observations and missingness is

$$f_{Y,M}(y, m \mid \theta, \phi) = \prod_{i=1}^{n} f_{Y_i}(y_i \mid \theta) f_{M_i|Y_i}(m_i \mid y_i, \theta)$$

For unit MAR, the condition becomes:

$$f_{M_i|Y_i}(\tilde{m}_i \mid Y_{i(0)} = \tilde{y}_{i(0)}, Y_{i(1)} = y_{i(1)}, \phi) = f_{M_i|Y_i}(\tilde{m}_i \mid Y_{i(0)} = \tilde{y}_{i(0)}, Y_{i(1)} = y^{\star}_{i(1)}, \phi)$$

for all $y_{i(1)}, y^{\star}_{i(1)}, \phi$ for all i.

Clarifying MAR vs. MAAR

The example given by Little (2021) is the following:

Suppose we have observations (y_i, x_i), where y_i is potentially missing and x_i is either 1 or 2, denoting group membership. Let $\theta = (\mu_1, \mu_2, \sigma^2)$, and the model for the observations be

$$y_i \mid x_i, \theta \sim \text{Normal}(\mu_{x_i}, \sigma^2)$$

The standard confidence interval for the difference in means is:

$$\bar{y}_2 - \bar{y}_1 \pm t_{\nu,0.975}(s\sqrt{1/n_1 + 1/n2})$$

This also corresponds to the Bayesian credible interval when using a flat prior on μ_1, μ_2 and $\log \sigma^2$. Suppose the missingness mechanism is as follows:

$$P(m_i = 1 \mid x_i, y_i, \phi) = \begin{cases} 0 & x_i = 1 \\ 0 & x_i = 2 \text{ and } y_i \leq \phi \\ 1 & x_i = 2 \text{ and } y_i > \phi \end{cases}$$

That is, for group 2, if the observation is above an unknown cutoff value, the value is not recorded.

Suppose that we have a dataset where there are no missing values. Then the data is MAR but not MAAR, because in repeated hypothetical samples there would be missing values that are MNAR. The Bayesian credible interval is still valid under MAR because we're conditioning on the dataset we have, whereas the Frequentist interval isn't valid because it couldn't be repeated for datasets where y_i is missing for some group 2 observations.

Coarsened data

Coarsened data is a generalization of missing data that includes other ways in which the resolution of data can be reduced. Examples include censoring, grouping, rounding, or heaping. Heaping is the phenomenon where there are varying levels of resolution reported in the same dataset. For example, on a questionnaire that asks for the the number of cigarettes smoked per day, some people will report exact numbers, and others will report multiples of packs. With rounded data, the coarsening is more deterministic, namely we know that an observation is exactly within the interval, say between $[\text{floor}(y), \text{floor}(y)+1]$ With coarsened data, there is still the complete data matrix $y = (y_i j)$, but there is now a coarsening variable c_{ij} that interacts with the true value to return the observed data.

Let w_{ij} be the observed data, and let $W(y_{ij}, c_{ij})$ be the function of the true value and the coarsening variable that returns some subset of $\mathcal{Y}_{ij}$ to which y_{ij} belongs. Thus $w_{ij} = W(y_{ij}, c_{ij})$ with the requirement that $y_{ij} \in W(y_{ij}, c_{ij})$. Let g_{ij} be the observed coarsening random variable that is governed by a function $G(y_{ij}, c_{ij})$ such that $c_{ij} \in G(y_{ij}, c_{ij})$. Just as w is a coarsened version of y, g is a coarsened version of c.

The simplest nontrivial example is the censored exponential data from above, though we will modify the scenario so that each individual has a potentially different censoring time c_i. Let y_i be the true time to failure, while c_i is the censoring time.

$$w_i = W(y_i, c_i) = \begin{cases} y_i & y_i \leq c_i \\ (c_i, \infty) & y_i > c_i \end{cases}$$

$$g_i = G(y_i, c_i) = \begin{cases} (y_i, \infty) & y_i \leq c_i \\ c_i & y_i > c_i \end{cases}$$

Let the realization of g and w be $\tilde{g}$ and $\tilde{w}$, with elements $\tilde{g}_i$ and $\tilde{w}_i$. Furthermore, let the distribution of interest for y_i be $f_Y(y_i \mid \theta)$, while we let the coarsening distributuion be $f_{C|Y}(c_i \mid y_i, \phi)$. Then we can write:

$$L_{\text{full}}(\theta, \phi \mid \tilde{g}, \tilde{w}) = \int \int f_{C|Y}(c \mid y, \phi) f_Y(y \mid \theta) \mathbb{1}\left(y \in \tilde{w}\right) \mathbb{1}\left(c \in \tilde{g}\right) dy\, dc$$

Another way to write this is by simplifying after the integration:

Let the vector $c_{(0)} = \{c_i \mid g_i = c_i, i = 1, \ldots, n\}$ and let $c_{(1)} = \{c_i \mid g_i \neq c_i, i = 1, \ldots, n\}$ Let $c = (c_{(0)}, c_{(1)})$ be the vector of coarsening values. Let $y_{(0)}$ be the set of values that we observe exactly, and $y_{(1)}$ be the set of values that are censored. Let $\tilde{w}_{(1)}$ be the set of subsets corresponding to the coarsened y's and the same for $\tilde{g}_{(1)}$.

Then the integral can be rewritten in terms of these variables:

$$L_{\text{full}}(\theta, \phi \mid \tilde{c}_{(0)}, \tilde{y}_{(0)}, \tilde{g}_{(1)}, \tilde{w}_{(1)}) = \int \int f_{C|Y}(\tilde{c}_{(0)}, c_{(1)} \mid \tilde{y}_{(0)}, y_{(1)}, \phi) f_Y(\tilde{y}_{(0)}, y_{(1)} \mid \theta) \mathbb{1}\left(y_{(1)} \in \tilde{w}\right) \mathbb{1}\left(c_{(1)} \in \tilde{g}\right) dy$$

The likelihood that ignores the coarsening process is:

$$L_{\text{ign}}(\theta \mid \tilde{y}_{(0)}, \tilde{w}_{(1)}) = \int f_Y(\tilde{y}_{(0)}, y_{(1)} \mid \theta) \mathbb{1}\left(y_{(1)} \in \tilde{w}_i\right) dy_{(1)}$$

This leads to a definition of coarsening at random, or CAR, that relates to conditions on the coarsening distribution:

$$f_{C|Y}(\tilde{c}_{(0)}, c_{(1)} \mid \tilde{y}_{(0)}, y_{(1)}, \phi) = f_{C|Y}(\tilde{c}_{(0)}, c^\star_{(1)} \mid \tilde{y}_{(0)}, y^\star_{(1)}, \phi)$$

For all $c_{(1)}, c^\star_{(1)}, y_{(1)}, y^\star_{(1)}, \phi$.

Each of these definitions has a unit-level variant, as MAR did above:

In the failure time example we have two contributions to the likelihood:

$$L_{\text{full}}(\theta, \phi \mid y_{(0)}, c_{(0)}) = \prod_{i \mid y_i \leq c_i} f(y_i \mid \theta) \int_{y_i}^{\infty} f_{C|Y}(c_i \mid y_i, \phi) dc \times \prod_{i \mid y_i > c_i} \int_{c_i}^{\infty} f_{C|Y}(c_i \mid y, \phi) f(y \mid \theta) dy$$

If we have that $f(c_i \mid y_i, \phi) = f(c_i \mid \phi)$ for all i, and that ϕ and θ are variationally independent, we can write the likelihood as the product of $L_{\text{ign}}(\theta \mid y_{(0)}, c_{(0)})$ and $L_{\text{rest}}(\phi \mid y_{(0)}, c_{(0)})$

$$\prod_{i \mid y_i \leq c_i} f(y_i \mid \theta) \prod_{i \mid y_i > c_i} \int_{c_i}^{\infty} f(y \mid \theta) dy \times \prod_{i \mid y_i \leq c_i} \int_{y_i}^{\infty} f_C(c_i \mid \phi) dc \prod_{i \mid y_i > c_i} f(c_i \mid \phi)$$

In this case, the censoring mechanism is CAR, but not MAR, as we saw earlier.

In the cigarette smoking example, let y_i be the true number of cigarettes smoked per day, and let c_i be an indicator for the precision of reporting. Then define w_i to be:

$$w_i = \begin{cases} [\text{floor}(y_i), \text{floor}(y_i) + 1] & c_i = 0 \\ [20 \times \text{floor}(y_i/20), 20 \times \text{floor}(y_i/20) + 20] & c_i = 1 \end{cases}$$

This assumes that people round down the number of cigarettes they smoke, rather than rounding to the nearest integer, like you'd do if there weren't a stigma around smoking.

Let the lower bound of the interval w_i be w_{iL}, then g_i be defined as:

$$g_i = \begin{cases} c_i & w_{iL} \text{ mod } 20 \neq 0 \\ \{0, 1\} & w_{iL} \text{ mod } 20 = 0 \end{cases}$$

Suppose that $f_{C|Y}(c_i \mid y_i, \phi) = \Phi(\phi_1 + \phi_2 y_i)^{c_i}(1 - \Phi(\phi_1 + \phi_2 y_i)^{1-c_i})$.

Then this example isn't CAR, because the coarsening is dependent on the number of cigarettes smoked.

What does the likelihood look like?

$$\begin{aligned}L(\theta,\phi \mid \tilde{w},\tilde{g}) = &\prod_{i \mid w_{iL} \bmod 20 \neq 0} \int_{w_{iL}}^{w_{iL}+1} f_Y(y \mid \theta)(1-\Phi(\phi_1+\phi_2 y))dy \\ &\times \prod_{i \mid w_i \bmod 20 = 0} \int_{w_{iL}}^{w_{iL}+20} (f_Y(y \mid \theta)\Phi(\phi_1+\phi_2 y))dy + \int_{w_{iL}}^{w_{iL}+1} (f_Y(y \mid \theta)(1-\Phi(\phi_1+\phi_2 y))dy\end{aligned}$$

Little, Roderick J. 2021. "Missing Data Assumptions." *Annual Review of Statistics and Its Application* 8 (1): 89–107. https://doi.org/10.1146/annurev-statistics-040720-031104.