Missing data lecture 12: Missing-not-at-random

MNAR missingness

The past few lectures have assumed that we have had ignorable missingness, so that inferences under:

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

were equivalent to inferences under

$$L_{\text{ign}}(\theta \mid \tilde{y}_{(0)}) \propto \int_{\mathcal{Y}_{(1)}} f_Y(Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y_{(1)} \mid \theta) dy_{(1)}$$

This was true as long as θ and ϕ were variationally independent, and that our missingness process was MAR:

$$P(M = \tilde{m} \mid Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y_{(1)}, \phi) = P(M = \tilde{m} \mid Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y^{\star}_{(1)}, \phi)$$

for any two $y_{(1)}, y^{\star}_{(1)}$ for all ϕ.

If instead we have MNAR missingness:

$$P(M = \tilde{m} \mid Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y_{(1)}, \phi) \neq P(M = \tilde{m} \mid Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y^{\star}_{(1)}, \phi)$$

for some $y_{(1)}, y^{\star}_{(1)}$ such that $y_{(1)} \neq y^{\star}_{(1)}$ and some ϕ, the inferences won't be the same.

MAR is often a very strong assumption. For example, suppose we are running a survey that includes the question: "How many cigarettes have you smoked in the last month?" Given the social attitudes around smoking, it might not be reasonable to assume that $P(M = 1 \mid y_i = 100, \phi) = P(M = 1 \mid y_i = 0, \phi)$.

Unfortunately, for us there are problems when our data are MNAR. For all but several special classes of models, we won't be able to identify the parameters of the missingness mechanism. That means that no matter how many observations we have we will never be able to learn some subset of the ϕ parameters with any certainty. This stands in contrast to most of the models we are used to working with. Formally, identifiability of a parametric model is defined as in Rothenberg (1971):

Definition 1. Let $\theta \in \Theta$ be a parameter indexing a parametric density function $f(y \mid \theta)$. θ is identifiable if there does not exist a parameter value $\theta' \in \Theta, \theta' \neq \theta$ for which the density $f(y \mid \theta) = f(y \mid \theta')$ for all observations y.

A simple example of nonidentifiability is the following model:

$$y_i \overset{\text{iid}}{\sim} \text{Normal}(\mu + \eta, \sigma^2)$$

Any pair (μ, η) and (μ', η') such that $c = \mu + \eta = \mu' + \eta'$ will lead to the same observational density. In this problem we would say that μ and η are nonidentifable or unidentifiable.

In our case, if our data are independent and y_i is one-dimensional, we can identify $P(M_i = 0 \mid y_i, \phi)$, but not $P(M_i = 1 \mid y_i, \phi)$, except in special cases, which we'll talk about later.

Approaches to modeling MNAR data: Pattern-mixture and selection models

Because we can't separate inference on θ from the missingness process, we'll need to use a joint model for Y, M. That entails some decomposition of $P(M = m, Y_{(0)} = y_{(0)}, Y_{(1)} = y_{(1)} \mid \theta, \phi)$.

The most straightforward decomposition is the one we used above, which is called a selection model:

$$f_Y(Y_{(0)} = y_{(0)}, Y_{(1)} = y_{(1)} \mid \theta) P(M = m \mid Y_{(0)} = y_{(0)}, Y_{(1)} = y_{(1)} \phi)$$

For now, we will assume that our units are independent conditional on a fully-observed covariate that varies with unit, so we'll focus on

$$f_Y(Y_{(0)} = y_{(0)}, Y_{(1)} = y_{(1)} \mid X = x, \theta) P(M = m \mid Y_{(0)} = y_{(0)}, Y_{(1)} = y_{(1)}, X = x, \phi) =$$
$$\prod_i f_Y(Y_{i(0)} = y_{i(0)}, Y_{i(1)} = y_{i(1)} \mid X_i = x_i, \theta) P(M = m_i \mid Y_{i(0)} = y_{i(0)}, Y_{i(1)} = y_{i(1)}, X_i = x_i, \phi)$$

A different decomposition is called the pattern-mixture model:

$$f_Y(Y_{(0)} = y_{(0)}, Y_{(1)} = y_{(1)} \mid M = m, \xi) P(M = m \mid \omega)$$

which, under an independence assumption, with a covariate x_i paired with each observation simplifies:

$$f_Y(Y_{(0)} = y_{(0)}, Y_{(1)} = y_{(1)} \mid M = m, X = x, \xi) P(M = m \mid X = x, \omega) =$$
$$\prod_i f_Y(Y_{i(0)} = y_{i(0)}, Y_{i(1)} = y_{i(1)} \mid M_i = m_i, X_i = x_i, \xi) P(M_i = m_i \mid X_i = x_i, \omega)$$

Example: Univariate nonresponse

Continuing with our example above, let's suppose that in addition answers to the how many cigarettes in the past month question, we have age data for each respondent, as well as state of residence. Let y_{i1} represent age of respondent i, and let y_{i2} be how many cigarettes individual i smoked in the past month, and let x_i be state of residence for each respondent. Suppose that we have arranged our data such that respondents $i = 1, \dots, r$ have complete data, whereas for $i = r+1, \dots, n$, respondents are missing y_{i2}.

The pattern mixture likelihood for this dataset would take the form:

$$f_Y(M = m, Y_{(0)} = y_{(0)}, Y_{(1)} = y_{(1)} \mid X = x, \xi) = \prod_{i=1}^{r} f_Y(y_{i1}, y_{i2} \mid m_i = 0, x_i, \xi) P(m_i = 0 \mid x_i, \omega) \times \prod_{i=r+1}^{n} f_Y(y_{i1} \mid m_i = 1, x_i, \xi) P(m_i = 1 \mid x_i, \omega)$$

Under MAR, we have that

$$f_Y(y_{i2} \mid y_{i1}, m_i = 1, x_i, \xi) = f_Y(y_{i2} \mid y_{i1}, m_i = 0, x_i, \xi)$$

which identifies the joint distribution of y_{i1}, y_{i2} for nonrespondents. For MNAR, the distribution $f_Y(y_{i2} \mid y_{i1}, m_i = 1, x_i, \xi)$ is that which is unknown.

The selection model approach to this is

$$f_Y(M = m, Y_{(0)} = y_{(0)}, Y_{(1)} = y_{(1)} \mid X = x, \xi) = \prod_{i=1}^{r} f_Y(y_{i1}, y_{i2} \mid x_i, \theta) P(m_i = 0 \mid y_{i1}, y_{i2}, x_i, \phi) \times \prod_{i=r+1}^{n} f_Y(y_{i1} \mid x_i, \theta) \int_{\mathcal{Y}} f_Y(y_{i2} \mid y_{i1}, x_i, \theta) P(m_i = 1 \mid y_{i1}, y_{i2}, x_i, \phi) dy_{i2}$$

You can see the appeal of pattern-mixture models from the two formulations: The pattern mixture model doesn't involve integrals of the missingness mechanism against an unknown density, and it is clear where the information is missing, namely

Identified selection models

Heckman selection model

If we're willing to assume a specific distribution for $(y_{i2}, m_i) \mid y_{i1}$, we can have identifiability of selection models. One choice is the following latent variable model, where

$m_i = \mathbb{1}(z_i \leq 0)$:

$$\begin{aligned} y_{i2} &= x_i^T \beta + \alpha y_{i1} + \sigma \epsilon_{i1} \\ z_i &= x_i^T \gamma + \phi_1 y_{i1} + \epsilon_{i2} \\ (\epsilon_{i1}, \epsilon_{i2}) &\sim \text{Normal}(0, \Sigma) \end{aligned}$$

where

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{bmatrix}$$

$$(y_{i2}, z_i) \sim \text{Normal}(x_i^T \beta + \alpha y_{i1}, x_i^T \gamma + \phi_1 y_{i1}, \Sigma)$$

If $\rho \neq 0$, $m_i \not\perp\!\!\!\perp y_{i2} \mid y_{i1}$. To see why:

$$z_i \mid y_{i2} \sim \text{Normal}(x_i^T \gamma + \phi_1 y_{i1} + \frac{\rho}{\sigma}(y_{i2} - (x_i^T \beta + \alpha y_{i1})), 1 - \rho^2)$$

This means that the likelihood for someone with $m_i = 0$ is:

$$f(y_{i1} \mid x_i, \theta) \text{Normal}(y_{i2} \mid x_i^T \beta + \alpha y_{i1}, \sigma^2) \Phi\left(\frac{x_i^T \gamma + \phi_1 y_{i1} + \frac{\rho}{\sigma}(y_{i2} - (x_i^T \beta + \alpha y_{i1}))}{\sqrt{1 - \rho^2}} \right)$$

Then the likelihood for someone with missing observations of y_{i2} is:

$$f_Y(y_{i1} \mid x_i, \theta)(1 - \Phi(x_i^T \gamma + \phi_1 y_{i1}))$$

The conditional formulation makes it clear that this model is equivalent to the following model:

$$\begin{aligned} y_{i1} &\sim \text{Normal}(x_i^T \beta + \alpha y_{i1}, \sigma^2) \\ P(m_i = 1 \mid y_{i2}) &= \Phi(x_i^T \gamma' + \phi_1' y_{i1} + \phi_2 y_{i2}) \end{aligned}$$

Rothenberg, Thomas J. 1971. "Identification in Parametric Models." *Econometrica* 39 (3): 577–91. https://doi.org/10.2307/1913267.