Missing data lecture 8: Likelihood-based inference with incomplete data

Likelihood inference with incomplete data

We said that the likelihood function is really a set of functions that are proportional the probability density such that the constant of proportionality doesn't depend the parameters.

Missing data methods distinguish themselves from other methods by modeling the joint distribution of Y and M. Let y and m represent the outcome measurements and m representing the missingness indictors for all i units:

$$f_{Y,M}(y, m \mid \theta, \phi) = f_Y(y \mid \theta) f_{M|Y}(m \mid y, \phi)$$

When we have missing data, we can partition the matrix y into $y_{(1)}$ and $y_{(0)}$, representing the components of y that are missing and observed, respectively.

Let $\mathcal{Y}$ be the sample space for y_i and let $\mathcal{Y}_{(1)}$ and $\mathcal{Y}_{(0)}$ be the sample space for the missing and observed components of y.

Then the distribution of the observed data is:

$$\int_{\mathcal{Y}_{(1)}} f_{Y,M}(y, m \mid \theta, \phi) dy_{(1)} = \int_{\mathcal{Y}_{(1)}} f_Y(y_{(0)}, y_{(1)} \mid \theta) f_{M|Y}(m \mid y_{(0)}, y_{(1)}, \phi) dy_{(1)}$$

This joint density is proportional to what we'll call the full-data likelihood:

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

We can also compute the likelihood *ignoring* the missingness process:

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

We'll say that the missingness mechanism is *ignorable* if inferences based on $L_{\text{ign}}(\theta \mid y_{(0)})$ and $L_{\text{full}}(\theta, \phi \mid y_{(0)}, m)$ are the same given $m, y_{(0)}$.

Formally, the missingness mechanism is ignorable for direct likelihood inference if the likeihood ratios for any two θ, θ^* given $m, y_{(0)}$ are equal:

$$\frac{L_{\text{full}}(\theta, \phi \mid y_{(0)}, m)}{L_{\text{full}}(\theta^*, \phi \mid y_{(0)}, m_i)} = \frac{L_{\text{ign}}(\theta \mid y_{(0)})}{L_{\text{ign}}(\theta^* \mid y_{(0)})} \forall \theta, \theta^*, \phi$$

There are two sufficient conditions ensure ignorability:

1. Parameters θ and ϕ are *variationally independent*, i.e. the joint parameter space $\Omega_{\theta,\phi} = \Omega_\theta \times \Omega_\phi$
2. The full likelihood factorizes as

$$L_{\text{full}}(\theta, \phi \mid y_{(0)}, m) = L_{\text{ign}}(\theta \mid y_{(0)}) L_{\text{rest}}(\phi \mid y_{(0)}, m)$$

   The first condition is sufficient to ensure that the value of ϕ doesn't lead to a different likelihood value for θ vs. θ^*.

If the data are MAR, then we will satisfy the second condition:

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

for all $y_{(1)}, y^*_{(1)}, \phi$. Then we can write the full-likelihood as:

$$f_{M|Y}(m \mid y_{(0)}, \phi) \int_{\mathcal{Y}_{(1)}} f_Y(y_{(0)}, y_{(1)} \mid \theta) dy_{(1)} = f_{M|Y}(m \mid y_{(0)}, \phi) f_Y(y_{(0)} \mid \theta)$$

Then by the above theorem, parameter distinctness and MAR are sufficient for ignorability.

When we do Bayesian inference we need to ensure that the posterior for θ when using the ignorable likelihood is equal to the posterior for θ when using the full likelihood. Under the full likleihood, the posterior for (θ, ϕ) is:

$$p(\theta, \phi \mid y_{(0)}, m) \propto p(\theta, \phi) L_{\text{full}}(\theta, \phi \mid y_{(0)}, m)$$

and under the ignorable likelihood we have:

$$p(\theta \mid y_{(0)}, m) \propto p(\theta) L_{\text{ign}}(\theta \mid y_{(0)})$$

Thus, sufficient conditions for the posteriors to be equal is that 1. $p(\theta, \phi) = p(\theta)p(\phi)$, or the prior independence of θ and ϕ 2. the likelihood factorizes

$$L_{\text{full}}(\theta, \phi \mid y_{(0)}, m) = L_{\text{ign}}(\theta \mid y_{(0)}) L_{\text{rest}}(\phi \mid y_{(0)}, m)$$

Example: Incomplete exponential sample

Let $y_i \overset{\text{iid}}{\sim} \text{Exponential}(\theta)$ for $i = 1, \dots, n$. Let m_i be the missingness indicators, and suppose $r = \sum_{i=1}^n m_i$. The full likelihood is

$$f_Y(y \mid \theta) = \theta^{-n} \exp\left(-\sum_{i=1}^n y_i/\theta\right)$$

Let $y_{(0)} = (y_1, \dots, y_r)$ and $y_{(1)} = (y_{r+1}, \dots, y_n)$. The likelihood that ignores the likelihood is

$$L_{\text{ign}}(\theta \mid y_{(0)}) = \theta^{-r} \exp\left(-\sum_{i=1}^r y_i/\theta\right)$$

Let $m_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\phi)$, so

$$f_{M|Y}(m \mid y, \phi) = \phi^r (1-\phi)^{n-r}$$

Then $f(y_{(0)}, m \mid \theta, \phi) = \phi^r(1-\phi)^{n-r}\theta^{-r}\exp\left(-\sum_{i=1}^r y_i/\theta\right)$, which factorizes into a factor related to θ and a factor related to ϕ. This means we can base inferences on θ on $L_{\text{ign}}(\theta \mid y_{(0)}$ instead of the full likelihood. The MLE is $\hat{\theta} = \sum_{i=1}^n y_i/r$.

Now suppose we observe only observations for which $y_i \leq c$, so

$$f(m_i \mid y_i, \phi) = \mathbb{1}\left(y_i \geq c\right)^{m_i} \mathbb{1}\left(y_i < c\right)^{1-m_i}$$

Putting this together, the full likelihood is:

$$\prod_{i=1}^r f_Y(y_i \mid \theta)\mathbb{1}\left(y_i < c\right) \prod_{i=r+1}^n \int_{\mathbb{R}^+} \mathbb{1}\left(y_i \geq c\right) f(y \mid \theta) dy$$

Which of course simplifies to

$$\theta^{-r} \exp\left(-\sum_{i=1}^r y_i/\theta\right) \exp(-(n-r)c/\theta)$$

This shows that the missingness is nonignorable, because the full likelihood isn't equal to the ignorable likelihood we used in the first part of the problem.

The log-likelihood is:

$$\ell(\theta \mid y_{(0)}, m) = -r\log\theta - \sum_{i=1}^r y_i/\theta - (n-r)c/\theta$$

$$\frac{\partial \ell(\theta \mid y_{(0)}, m)}{\partial\theta} = -r/\theta + \sum_{i=1}^r y_i/\theta^2 + (n-r)c/\theta^2$$

Setting this equal to zero and solving for θ gives the MLE:

$$\hat{\theta} = \frac{\sum_{i=1}^r y_i + (n-r)c}{r}$$

This of course doesn't equal the ignorable MLE, $\bar{y}$ for the observed values.

Missing data example: Parameter distinctness

Let the model be defined as

$$
\begin{aligned}
y_{ij} \mid \mu_i, \theta &\sim \text{Normal}(\alpha_i, \sigma^2) \\
\alpha_i \mid \theta &\sim \text{Normal}(\mu, \tau^2)
\end{aligned}
$$

Let the missingness mechanism be:

$$f_{M|Y}(m_{ij} \mid y, \alpha_i, \phi) = \pi(\alpha_i, \phi) = (1 + e^{-(\phi_0 + \phi_1 \alpha_i)})^{-1}$$

The joint density of the observations and parameters, also known as the complete data likelihood, is:

$$\prod_{i=1}^{I} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_{ij}-\alpha_i)^2} \pi(\alpha_i, \phi)^{m_{ij}} (1 - \pi(\alpha_i, \phi))^{1-m_{ij}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\alpha_i - \mu)^2}$$

Because the α_i aren't observed, but do have a density, we need to integrate over them to compute the full likelihood:

$$\prod_{i=1}^{I} \int_{\mathbb{R}} \prod_{j=1}^{n_i} \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_{ij}-\alpha_i)^2} \right)^{1-m_{ij}} \pi(\alpha_i, \phi)^{m_{ij}} (1 - \pi(\alpha_i, \phi))^{1-m_{ij}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\alpha_i - \mu)^2} d\alpha_i$$

This shows that the missingness process isn't ignorable here, even though we don't technically have the distribution of missingness depending on missing obesrvable data, per se. This shows that in some sense, α_i is missing data, and, indeed, this is what our textbook considers missing data; namely anything that has a distribution that is unobserved. This makes the problem MNAR.

Compare this to the ANOVA model with the same missingness mechanism:

$$y_{ij} \mid \alpha_i, \theta \sim \text{Normal}(\alpha_i, \sigma^2)$$

Then the joint likelihood is

$$\prod_{i=1}^{I} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_{ij}-\alpha_i)^2} \pi(\alpha_i, \phi)^{m_{ij}} (1 - \pi(\alpha_i, \phi))^{1-m_{ij}}$$

This shows that the data are MAR, because the missingness mechanism doesn't depend on missing data. However, the parameters for the missingness mechanism and the observations don't satisfy the distinctness condition, so the missingness is nonignorable.

Partial MAR

Suppose we can partition θ into two pieces, θ_1 and θ_2 so that the parameter of interest is θ_1. The data are partially MAR for θ_1 if we can factorize the full likelihood:

$$L_{\text{full}}(\theta_1, \theta_2, \phi \mid y_{(0)}, m) = L_1(\theta \mid y_{(0)}) L_{\text{rest}}(\theta_2, \phi \mid y_{(0)}, m)$$

Example: Regression with missing data

An example of this is when we have covariates paired with each observation so that the complete data is $(y_i, x_i), i = 1, \ldots, n$ where $y_i \in \mathbb{R}^d$ and $x_i \in \mathbb{R}^p$. let $y_{(0)}, y_{(1)}$ be the observed and missing elements of y_i and $x_{(0)}, x_{(1)}$ are the observed and missing elements of x_i. Let m_i^Z be the missingness indicators for the covariates, Z, and let m_i^Y be the missingness indicators for the observations y_i. Let $m_i = (m_i^Y, m_i^Z)$ be the combined missingness indicators for unit i. Suppose that for $i = 1, \ldots, r$ x_i is fully observed, while at least one component of y_i is observed, and for the remaining $i = r+1, \ldots, n$ y_i is completely missing and each z_i has at least one missing component. Let y_i, x_i, z_i be unit iid, so:

$$f_{Y,X,M}(y_i, x_i, m_i \mid \theta_1, \theta_2, \phi) = f_{Y|X}(y_i \mid x_i, \theta_1) f_X(x_i \mid \theta_2) f_{M|Y,X}(m_i \mid y_i, x_i, \phi)$$

We'll assume the missingness mechanism takes the following form:

$$f_{M|Y,X}(m_i \mid x_{i(1)}, x_{i(0)}, y_i, \phi) = f_{M|Y,X}(m_i \mid x_{i(1)}, x_{i(0)}, y_i^\star, \phi)$$

for all $y_i, y_i^\star, x_{i(0)}, i = 1, \ldots, n$.

This missingness mechanism is MNAR because it depends on unobserved components of $x_{i(1)}$. Luckily we'll be able to factorize our likelihood so inference θ_1 is partially ignorable:

$$L_{\text{full}}(\theta_1, \theta_2, \phi \mid y_{(0)}, x_{(0)}, m) = L_{\text{p-ign}}(\theta_1 \mid y_{(0)}, x_{(0)}) L_{\text{rest}}(\theta_2, \phi \mid m, x_{(0)})$$

Let $\mathcal{Y}_i$ be the sample space corresponding to the missing $y_{i(1)}$. Then we can write the ignorable part of the likelihood:

$$L_{\text{p-ign}}(\theta_1 \mid y_{(0)}, x_{(0)}) = \prod_{i=1}^r \int_{\mathcal{Y}_i} f_{Y|X}(y_{i(0)}, y_{i(1)} \mid x_i, \theta_1) dy_{i(1)}$$

Let $\mathcal{X}_i$ be the sample space of the missing covariates for the i^{th} unit. Then the rest of the likelihood can be written as

$$L_{\text{rest}}(\theta_2, \phi \mid x_{i(0)}, m) = \prod_{i=1}^r f_X(z_i \mid \theta_2) f_{M|X}(m_i \mid x_i, \phi) \prod_{i=r+1}^n \int_{\mathcal{X}_i} f_X(x_{i(0)}, x_{i(1)} \mid \theta_2) f_{M|X}(m_i \mid x_{i(1)}, x_{i(0)}, \phi)$$

Note the book has a typo here.