Missing data lecture 2

Recap

We left off with characterizations of MCAR, MAR, and MNAR:

1. MCAR: $f_{M|Y}(m_i \mid y_i, \phi) = f_{M|Y}(m_i \mid y_i^*, \phi)$ for all ϕ, i, y_i, y_i^* such that $y_i \neq y_i^*$.
2. MAR: $f_{M|Y}(m_i \mid y_{(0)i}, y_{(1)i}, \phi) = f_{M|Y}(m_i \mid y_{(0)i}, y_{(1)i}^*, \phi)$ for all $\phi, i, y_{(1)i}, y_{(1)i}^*$ such that $y_{(1)i} \neq y_{(1)i}^*$.
3. MAR: $f_{M|Y}(m_i \mid y_{(0)i}, y_{(1)i}, \phi) \neq f_{M|Y}(m_i \mid y_{(0)i}, y_{(1)i}^*, \phi)$ for some $\phi, i, y_{(1)i}, y_{(1)i}^*$ such that $y_{(1)i} \neq y_{(1)i}^*$.

We'll come back to the "Always" versions of these missingness mechanisms later in the lecture.

MAR with covariates

A common extension of MAR is MAR given covariates. Let the *completely observed* covariate matrix be X with n rows and p columns. Let the i^{th} row of X be denoted x_i. Then MAR with covariates is as follows: The equality:

$$f_{M|Y}(m_i \mid x_i, y_{(0)i}, y_{(1)i}\phi) = f_{M|Y}(m_i \mid x_i, y_{(0)i}, y_{(1)i}^*\phi)$$

holds for all i, $y_{(1)i}$, $y_{(1)i}^*$, and ϕ.

Joint density Y, M

The past lecture focused only on the missingness mechanism, the distribution of M given Y. Of course, y_i has a distribution (with a density) that we'd like to learn, typically through inferring parameters θ: $f_Y(y_i \mid \theta)$. Let the sample space for y_i be $\mathcal{Y}$. Putting the distribution of Y together with the density corresponding to the conditional distribution $M \mid Y$ will give us a joint density for the pairs (y_i, m_i):

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

This leads to the joint density for all n observations, assuming y_i are independent and m_i are conditionally independent given y_i and observations are identically distributed:

$$f_{Y,M}(y_1, m_1, \dots, y_n, m_n \mid \theta, \phi) = \prod_{i=1}^{n} f_Y(y_i \mid \theta) f_{M|Y}(m_i \mid y_i, \phi) \tag{1}$$

When there is a single variable, y_i is one dimensional and m_i is a binary random variable.

What we've written above isn't something we can evaluate, because we don't observe y_i when $m_i = 1$. The expression (Equation 1) will need to be modified in order to evaluate the density for all n:

$$f_{Y_{(0)},M}(y_{(0)}, m_1, \dots, m_n \mid \theta, \phi) = \prod_{i=1}^{n} (\int_{\mathcal{Y}} f_Y(y_i \mid \theta) f_{M|Y}(m_i \mid y_i, \phi) dy_i)^{m_i} (f_Y(y_i \mid \theta) f_{M|Y}(m_i \mid y_i, \phi))^{1-m_i} \tag{2}$$

This is pretty ugly! When we have a missing observation, the density needs to be integrated over all possible values of y_i to properly account for the fact that the observation is missing.

Under MAR in the univariate setting the following must be true:

$$f_{M|Y}(m_i \mid y_i, \phi) = f_{M|Y}(m_i \mid \phi)$$

In this case, $f_{M|Y}(m_i \mid \phi)$ is just a Bernoullli distribution with ϕ governing the probability of missingness.

Which leads to the following simplification for the observed data density:

$$f_{Y_{(0)},M}(y_{(0)}, m_1, \dots, m_n \mid \theta, \phi) = \prod_{i=1}^{n} (\int_{\mathcal{Y}} f_Y(y_i \mid \theta) dy_i)^{m_i} f_Y(y_i \mid \theta))^{1-m_i} f_{M|Y}(m_i \mid \phi) \tag{3}$$

This looks much better: the integral $\int_{\mathcal{Y}} f_Y(y_i \mid \theta) dy_i = 1$, so the observed density is:

$$f_{Y_{(0)},M}(y_{(0)}, m_1, \dots, m_n \mid \theta, \phi) = \prod_{i=1}^{n} f_Y(y_i \mid \theta))^{1-m_i} f_{M|Y}(m_i \mid \phi) \tag{4}$$

Now that we have something that we can evaluate, what do we do with the density (Equation 4)?

Likelihood-based inference

This bit of the notes follows §6.1 pretty closely.

For the moment, let's forget about missing data, and move to the simpler case where we have no missing observations and we're just focused on learning the unknown parameters θ in the density $f_Y(y_i \mid \theta)$. This unknown parameter is going to live in the space Ω_θ. For example, if $\theta = (\mu, \sigma^2)$ and $f_Y(y_i \mid \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2}$, then Ω_θ would be $(\mathbb{R}, \mathbb{R}^+)$.

We know densities are functions of y_i for fixed values of θ. If we instead fix y_i and let θ vary, we get a *likelihood function*.

Let the likelihood be defined for a fixed y_i as

$$L_Y(\theta \mid y_i) \propto \begin{cases} f(y_i \mid \theta) & \theta \in \Omega_\theta \\ 0 & \theta \notin \Omega_\theta \end{cases}$$

This is a bit odd, because the expression means that anything proportional to the density of Y such that the factor by which $L_Y(\theta \mid y_i)$ differs from $f(y_i \mid \theta)$ is constant in θ.

Typically, we'll have more than one observation, and under independence of y_i we get the likelihood for the full sample:

$$L_Y(\theta \mid y_1, \dots, y_n) \propto \begin{cases} \prod_{i=1}^n f(y_i \mid \theta) & \theta \in \Omega_\theta \\ 0 & \theta \notin \Omega_\theta \end{cases}$$

We'll often work with the log-likelihood, which is denoted in our book as:

$$\ell_Y(\theta \mid y_1, \dots, y_n) = \log(L_Y(\theta \mid y_1, \dots, y_n))$$

When we have independence, we get

$$\ell_Y(\theta \mid y_1, \dots, y_n) = \sum_{i=1}^n \log f(y_i \mid \theta) + C$$

where C is any term that doesn't depend on θ.

Moving forward with the normal example from above, the likelihood of n observations from the normal distribution is:

$$\begin{aligned} L_Y(\mu, \sigma^2 \mid y_1, \dots, y_n) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\textstyle\sum_i (y_i - \mu)^2\right) \end{aligned}$$

Our definition of likelihood means that we can drop the factor of $(2\pi)^{-\frac{n}{2}}$ from the front of the expression. Taking logs and dropping that constant leads to the log-likelihood:

$$\ell_Y(\mu, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_i (y_i - \mu)^2$$

which is a bivariate function of μ and σ^2.

The way that we'll use the likelihood is to use the intuition that for two parameter values θ' and θ'' if $L(\theta' \mid y) = 2L(\theta'' \mid y)$, then there is evidence against θ'' being the parameter that generated the dataset.

If we were to find a $\hat{\theta}$ such that $L_Y(\hat{\theta} \mid y) \geq L_Y(\theta^* \mid y)$ for all $\theta* \neq \hat{\theta}$, then this would be evidence against θ being anything other than $\hat{theta}$. This is the intuition behind *maximum likelihood*, or finding the value of the unknown parameters θ that maximizes the likelihood function (or, equivalently, the log-likelihood).

We'll say that the maximum likelihood estimator (MLE) of θ is the value $\hat{\theta} \in \Omega_\theta$ that maximizes the log-likelihood.

In order to maximize the likelihood, we need to find the point at which the gradient of the log-likelihood, also known as the score function, is zero, or $\frac{\partial \ell_Y(\theta|y)}{\partial \theta} \mid_{\theta=\hat{\theta}} = 0$. If θ is d-dimensional, then there are d equations that need to be solved. In addition, the Hessian of the log-likelihood needs to be checked to see if it is negative semidefinite at the MLE:

$$z^T \frac{\partial^2 \ell_Y(\theta \mid y)}{\partial\theta\, \partial\theta^T} \mid_{\theta=\hat{\theta}} z \leq 0 \; \forall \, z \in \mathbb{R}^d$$

This ensures that we've found a maximum. Note that we can have several maxima, all of which lead to the same log-likelihood value.

Simple MLE example

Let's say that $y_i \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$. Then the likelihood will be

$$L(\lambda \mid y_1, \ldots, y_n) = \lambda^{-n} e^{-\frac{1}{\lambda}\sum_i y_i}$$

Then the score equation is:

$$-\frac{n}{\lambda} + \sum_i \frac{y_i}{\lambda^2} = 0$$

This leads to $\hat{\lambda} = \bar{y}$ or the sample mean.

More involved MLE example: Normal with unknown mean and variance

Suppose $y_i \overset{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$ for $i = 1, \ldots, n$.

We wrote the log-likelihood above as:

$$\ell_Y(\mu, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \textstyle\sum_i (y_i - \mu)^2$$

This leads to two likelihood equations:

$$\begin{aligned} \frac{\partial \ell}{\partial \mu} &= \frac{1}{\sigma^2} \textstyle\sum_i (y_i - \mu)^2 \\ \frac{\partial \ell}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \textstyle\sum_i (y_i - \mu)^2 \end{aligned}$$

Setting these to zero and solving for (μ, σ^2) gives the MLEs $\hat{\mu} = \bar{y}$ and $\hat{\sigma}^2 = \frac{1}{n} \sum_i (y_i - \bar{y})^2$.

It's straightforward to show that $\mathbb{E}[\hat{\mu}]$ under the data generating process above is equal to μ. It is less straightforward and somewhat distressing that $\mathbb{E}\left[\widehat{\sigma^2}\right]$ not equal to σ^2.

$$\begin{aligned} \mathbb{E}\left[\widehat{\sigma^2}\right] &= \frac{1}{n} \mathbb{E}\left[\sum_i (y_i^2 - 2 y_i \bar{y} + \bar{y}^2)\right] \\ &= \frac{1}{n} \left(\mathbb{E}\left[\textstyle\sum_i y_i^2 - 2n\bar{y}^2 + n\bar{y}^2\right]\right) \\ &= \frac{1}{n} \textstyle\sum_i \mathbb{E}[y_i^2] - \mathbb{E}[\bar{y}^2] \\ &= \mu^2 + \sigma^2 - \frac{1}{n^2} \mathbb{E}\left[\textstyle\sum_i y_i^2 + 2 \sum_{i<j} y_i y_j\right] \\ &= \mu^2 + \sigma^2 - \frac{1}{n^2} \left(\textstyle\sum_i \mathbb{E}[y_i^2] + 2 \sum_{i<j} \mathbb{E}[y_i y_j]\right) \\ &= \mu^2 + \sigma^2 - \frac{1}{n^2} \left(n(\mu^2 + \sigma^2) + n(n-1)\mu^2\right) \\ &= \mu^2 + \sigma^2 - \mu^2 - \frac{1}{n}\sigma^2 \\ &= \frac{n-1}{n} \sigma^2 \end{aligned}$$

Sorta odd that MLEs can give us biased estimates, but, you might say, as $n \to \infty$ all is fine and we recover σ^2. Indeed, you can show that the MLE for σ^2 is *consistent*, which means that as $n \to \infty$ $\hat{\sigma}^2 \to \sigma$ in probability.

More complicated still: Neyman-Scott problem

Is this always the case, that the MLE is consistent? The answer is no, and the MLE can fail pretty spectacularly. Consider the following problem:

$$y_{i1}, y_{i2} \sim \text{Normal}(\mu_i, \sigma^2)$$

Let's say that we don't care about inferring μ_i but we care only about σ^2. One way we could come up with an estimator is to difference the two observations for each group, so $z_i = y_{i1} - y_{i2}$ and :

$$z_i \sim \text{Normal}(0, 2\sigma^2)$$

Then we could reuse our work from above to solve for σ^2:

$$2\sigma^2 = \frac{1}{n} \sum_i z_i^2$$

which leads to an estimator for σ^2 of:

$$\hat{\sigma}^2 = \frac{1}{2n} \sum_i z_i^2$$

But this isn't quite maximum likelihood because we've transformed the data, and then used the transformed data to derive an MLE for σ^2 using z_i. What if we just use y_{i1}, y_{i2}?

The likelihood is straightforward:

$$L_Y(\{\mu_i\}, \sigma^2 \mid y) = \prod_{i=1}^{n} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{1}{2\sigma^2}\left((y_{i1} - \mu_i)^2 + (y_{i2} - \mu_i)^2\right)\right)$$

The likelihood equations for μ_i are

$$\frac{\partial \ell_Y(\{\mu_i\}, \sigma^2 \mid y)}{\partial \mu_i} = \frac{1}{\sigma^2}\left((y_{i1} - \mu_i) + (y_{i2} - \mu_i)\right)$$

Which just leads to the estimator $\hat{\mu}_i = \frac{y_{i1}+y_{i2}}{2} = \bar{y}_i$.
Let's write out the likelihood equations for σ^2 after plugging in our $\hat{\mu}_i = \bar{y}_i$:

$$\frac{\partial \ell_Y(\{\mu_i\}, \sigma^2 \mid y)}{\partial \sigma^2} = -\frac{n}{\sigma^2} + \frac{1}{2\sigma^2} \sum_i \sum_{j=1}^{2} (y_{ij} - \bar{y}_i)^2$$

This leads to an MLE of

$$\hat{\sigma}^2 = \frac{1}{2n} \sum_i \sum_{j=1}^{2} (y_{ij} - \bar{y}_i)^2$$

Which seems reasonable enough. But let's write the inner sum in terms of z_i:

$$\begin{aligned}
(y_{i1} - \bar{y}_i)^2 + (y_{i2} - \bar{y}_i)^2 &= y_{i1}^2 + y_{i2}^2 + 2\bar{y}_i^2 - 2\bar{y}_i(y_{i1} + y_{i2}) \\
&= y_{i1}^2 + y_{i2}^2 + \frac{(y_{i1} + y_{i2})^2}{2} - (y_{i1} + y_{i2})^2 \\
&= y_{i1}^2 + y_{i2}^2 - \frac{(y_{i1} + y_{i2})^2}{2} \\
&= \frac{1}{2}(y_{i1}^2 + y_{i2}^2 - 2y_{i1}y_{i2}) \\
&= \frac{1}{2}z_i^2
\end{aligned}$$

This means that the MLE is equal to

$$\hat{\sigma}^2 = \frac{1}{4n}\sum_i z_i^2$$

This estimator will never approach σ^2, as $n \to \infty$.

This is due to the fact that the dimension of the parameter space grows linearly with the sample size as $n \to \infty$.

The point is that MLEs (and any other sort of likelihood-based inference) can lead you astray if you're not careful, and they are not a panacea.