Missing data lecture 13: Identified and nonidentified models for MNAR data

Selection models with exclusion restrictions

We mentioned last time that the following model:

$$\begin{aligned} y_{i1} &\sim \text{Normal}(x_i^T\beta, \sigma^2) \\ P(m_i = 1 \mid y_{i1}) &= \Phi(x_i^T\gamma + \phi_2 y_{i1}) \end{aligned}$$

was identified only by the implied joint normality of the errors in the equivalent model:

$$\begin{aligned} y_{i1} &= x_i^T\beta + \sigma\epsilon_{i1} \\ z_i &= x_i^T\gamma' + \epsilon_{i2} \\ (\epsilon_{i1}, \epsilon_{i2}) &\sim \text{Normal}\left(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right) \\ m_i &= \mathbb{1}(z_i > 0) \end{aligned}$$

Why is this the case that identification follows from the normality assumption?

The standard procedure for estimating these models is via a two step procedure:

1. Fit a probit model to the missingness indicators to learn γ'.
2. Fit a linear regression model for y_{i1} on x_i and the conditional expectation of $z_i \mid z_i < 0$, which we get from the probit model.

Let's look at the conditional expectation of $y_{i1} \mid z_i$ given $z_i < 0$. We know the conditional expectation of $y_{i1} \mid z_i = z$:

$$x_i^T\beta + \rho\sigma(z_i - x_i^T\gamma')$$

What is the conditional expectation of $z \mid z < 0$? This is called the inverse Mills ratio:

$$\mathbb{E}[z \mid z < 0] = x_i^T\gamma' - \frac{\phi(-x_i^T\gamma')}{\Phi(-x_i^T\gamma')}$$

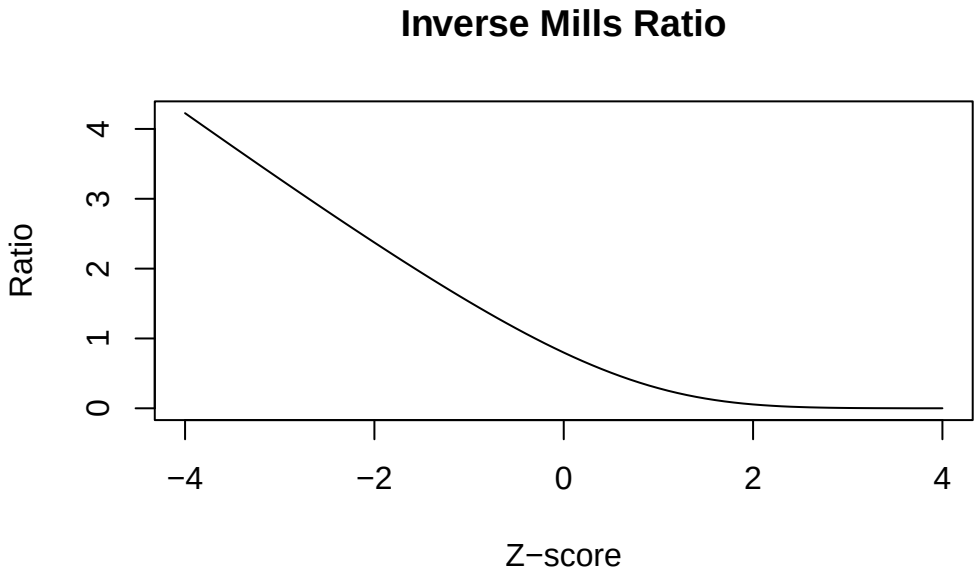
Plugging this in above gives:

$$x_i^T\beta - \rho\sigma\frac{\phi(-x_i^T\gamma')}{\Phi(-x_i^T\gamma')}$$

Thus, this is design matrix for the model we would fit to the units in which y_{i1} is observed:

$$\begin{bmatrix} x_1^T & \phi(-x_1^T\hat{\gamma}')/\Phi(-x_1^T\hat{\gamma}') \\ x_2^T & \phi(-x_2^T\hat{\gamma}')/\Phi(-x_2^T\hat{\gamma}') \\ \vdots & \vdots \\ x_n^T & \phi(-x_n^T\hat{\gamma}')/\Phi(-x_n^T\hat{\gamma}') \end{bmatrix}$$

To the extent that the function $\phi(\cdot)/\Phi(\cdot)$, which is called the inverse Mills ratio, is linear in its arguments, this extra term in the regression will be collinear with $x_i^T\beta$ and the parameter ρ won't be well-identified. Let's plot the function to see what it looks like:



Thus, the regression for y_{i1} on x_i and the Inverse Mills Ratio in the selected units is identified solely by the shape of the Inverse Mills ratio.

If we can write the model instead like:

$$y_{i1} \sim \text{Normal}(x_i^T\beta, \sigma^2)$$
$$P(m_i = 1 \mid y_{i1}) = \Phi(x_i^T\gamma + w_i^T\psi + \phi_2 y_{i1}),$$

the point estimates for β are more robust to deviations from the normality assumption. The reason for this is from above: we have another source of variation in the inverse Mills ratio, so the term won't be collinear with the predictors in the $y_{i1} \mid z_i$ regression.

Here's an example from our textbook where the study design makes it possible to exclude the w_i from the model for y_{i1}.

Example 1. Estimating HIV incidence via household surveys @janssens2014refusal analyzes data from a survey designed to infer population level incidence of HIV in Namibia. In this study, the researchers estimated population-level HIV incidence by sending a nurse to randomly selected households, and interviewing the household participants about demographic information, attitudes towards HIV and prevention. The nurse also would take a saliva sample to assess HIV status. As with any survey, there were respondents who declined to give a saliva sample, and there was concern that propensity to refuse to give a sample is related to HIV status. Crucially, the study design randomly assigned nurses to households. The model the researchers would like to fit is the following:

$$\begin{aligned} P(y_{i1} = 1) &= \Phi(x_i^T \beta, \sigma^2) \\ P(m_i = 1 \mid y_{i1}) &= \Phi(x_i^T \gamma + w_i^T \psi + \phi_2 y_{i1}), \end{aligned}$$

where x_i are sociodemographic variables, while w_i is vector of dummy variables corresponding to nurse ID. Because nurses were randomly assigned, it is plausible that nurse ID does not affect HIV status, but could impact the probability of refusing to give a sample.

The model can be formulated as a latent bivariate normal model:

$$\begin{aligned} z_{i1} &= x_i^T \beta + \epsilon_{i1} \\ z_{i2} &= x_i^T \gamma' + w_i^T \psi' + \epsilon_{i2} \\ (\epsilon_{i1}, \epsilon_{i2}) &\sim \text{Normal}\left(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right) \\ y_{i1} &= \mathbb{1}\left(z_{i1} \geq 0\right), \\ m_i &= \mathbb{1}\left(z_{i2} \geq 0\right) \end{aligned}$$

If $\psi \neq 0$, the model is robust to deviations from the joint latent normality assumption. The plausibility of the

The probability of each outcome is as follows, where we let $\mu_i^Y = x_i^T \beta$ and $\mu_i^M = x_i^T \gamma' + w_i^T \psi'$:

$$\begin{aligned} P(y_{i1} = 0, m_i = 0) &= P(\mu_i^Y + \epsilon_{i1} < 0, \mu_i^M + \epsilon_{i2} < 0) \\ P(y_{i1} = 1, m_i = 0) &= P(\mu_i^Y + \epsilon_{i1} \geq 0, \mu_i^M + \epsilon_{i2} < 0) \\ P(y_{i1} = 0, m_i = 1) &= P(\mu_i^Y + \epsilon_{i1} < 0, \mu_i^M + \epsilon_{i2} \geq 0) \\ P(y_{i1} = 1, m_i = 1) &= P(\mu_i^Y + \epsilon_{i1} \geq 0, \mu_i^M + \epsilon_{i2} \geq 0) \end{aligned}$$

This expression can be written in terms of the bivariate normal CDF: $\Phi(a, b, \rho)$, which equals $P(\mathcal{Z}_1 \leq a, \mathcal{Z}_2 \leq b)$, where $\mathcal{Z}_1, \mathcal{Z}_2$ are bivariate normal with mean zero, standard deviations 1, and correlation ρ:

$$\begin{aligned} P(y_{i1} = 0, m_i = 0) &= P(\mu_i^Y + \epsilon_{i1} < 0, \mu_i^M + \epsilon_{i2} < 0) \\ &= P(\epsilon_{i1} < -\mu_i^Y, \epsilon_{i2} < -\mu_i^M) \\ &= \Phi(-\mu_i^Y, -\mu_i^M, \rho) \end{aligned}$$

Also note that if $\mathcal{Z}_1, \mathcal{Z}_2$ are bivariate normal with correlation ρ, then $(-\mathcal{Z}_1, \mathcal{Z}_2)$ are bivariate normal with correlation $-\rho$. Then

$$\begin{aligned} P(y_{i1} = 1, m_i = 0) &= P(\mu_i^Y + \epsilon_{i1} \geq 0, \mu_i^M + \epsilon_{i2} < 0) \\ &= P(\epsilon_{i1} > -\mu_i^Y, \epsilon_{i2} < -\mu_i^M) \\ &= P(-\epsilon_{i1} < \mu_i^Y, \epsilon_{i2} < -\mu_i^M) \\ &= \Phi(\mu_i^Y, -\mu_i^M, -\rho) \end{aligned}$$

All of these events are written

$$\begin{aligned} P(y_{i1} = 0, m_i = 0) &= \Phi(-\mu_i^Y, -\mu_i^M, \rho) \\ P(y_{i1} = 1, m_i = 0) &= \Phi(\mu_i^Y, -\mu_i^M, -\rho) \\ P(y_{i1} = 0, m_i = 1) &= \Phi(-\mu_i^Y, \mu_i^M, -\rho) \\ P(y_{i1} = 1, m_i = 1) &= \Phi(\mu_i^Y, \mu_i^M, \rho) \end{aligned}$$

This model is called the bivariate probit model, and you will fit this model in Stan for the next HW to a similar dataset.

The researchers found that incidence estimates for differet subgroups could be impacted by refusal bias, so it was necessary to perform this adjustment.

There is another bivariate normal model that is identified by exclusion restrictions.

Example 2. Bivariate Normal pattern-mixture model Imagine we're running a household survey on income, so we can observe the address at which someone lives, but we might not learn household income due to refusals.

Let Y_{i1} be the log-household value, which we obtain from Zillow, or from property tax assessments, and Y_{i2} be the log of the response to a question about household income on a survey. Let $M_i = 1$ when Y_{i2} is missing and 0 when it is observed. Let $i = 1, \ldots, r$ be the cases for which we have observed log-income, and let $i = r+1, \ldots, n$ be the observations that are missing log-income. We assume that we have log-houshold value for all survey respondents.

We can use the pattern-mixture model for this scenario:

$$\begin{aligned} L(\mu, \Sigma \mid Y_{(0)} = \tilde{y}_{(0)}, M = \tilde{m}) = &\prod_{i=1}^{r} (1-\omega) \mathrm{N}(y_{i1}, y_{i1} \mid, m_i = 0, \mu_0, \Sigma_0) \\ &\times \prod_{i=r+1}^{n} \omega \mathrm{N}(y_{i1} \mid m_i = 1, \mu_1, \sigma_1^2) \end{aligned}$$

We'll suppose that the missingness mechanism has the following form:

$$P(M_i = 1 \mid Y_{i1} = y_{i1}, Y_{i1} = y_{i2}) = P(M_i = 1 \mid Y_{i1} = y_{i2}).$$

We would like to learn the following marginal expection for y_{i2}:

$$\mathbb{E}[Y_{i2}] = \mathbb{E}[Y_{i2} \mid M_i = 0](1-\omega) + \mathbb{E}[Y_{i2} \mid M_i = 1]\,\omega$$

where we can't calculate $\mathbb{E}[Y_{i2} \mid M_i = 1]$ directly. However, we can use the fact that the missingness mechanism does not depend on y_{i1} to identify our model.

$$\begin{aligned} f_Y(y_{i1} = y \mid Y_{i2} = y_{i2}, M_i = m) &= \frac{f_Y(y_{i1} = y \mid Y_{i2} = y_{i2}) P(M_i = m \mid y_{i1} = y, Y_{i2} = y_{i2})}{P(M_i = m \mid Y_{i2} = y_{i2})} \\ &= \frac{f_Y(y_{i1} = y \mid Y_{i2} = y_{i2}) P(M_i = m \mid Y_{i2} = y_{i2})}{P(M_i = m \mid Y_{i2} = y_{i2})} \\ &= f_Y(y_{i1} = y \mid Y_{i2} = y_{i2}) \end{aligned}$$

This means that the following holds:

$$f_Y(y_{i1} = y \mid Y_{i2} = y_{i2}, M_i = 1) = f_Y(y_{i1} = y \mid Y_{i2} = y_{i2}, M_i = 0)$$

This means that we can infer the joint distribution of Y_{i2}, Y_{i1} for $M_i = 1$ if the joint normality assumption holds. It comes from the following observation. Let $\beta^{(0)}_{10\cdot 2}$ $\beta^{(0)}_{12\cdot 2}$ be the intercept and slope from the regression of y_{i1} on y_{i2} in the complete units.

$$\mu_1^{(0)} = \beta^{(0)}_{10\cdot 2} + \beta^{(0)}_{12\cdot 2}\mu_2^{(0)}$$

Using the assumption that $\beta^{(1)}_{10\cdot 2} = \beta^{(0)}_{10\cdot 2}$ $\beta^{(1)}_{12\cdot 2} = \beta^{(0)}_{12\cdot 2}$ gives us the following relationship:

$$\mu_1^{(1)} = \beta^{(0)}_{10\cdot 2} + \beta^{(0)}_{12\cdot 2}\mu_2^{(1)}$$

Then we can solve for $\mu_2^{(1)}$:

$$\mu_2^{(1)} = \frac{\mu_1^{(1)} - \beta^{(0)}_{10\cdot 2}}{\beta^{(0)}_{12\cdot 2}}$$

We observe $\mu_1^{(1)}$, and have an MLE of $\bar{y}_1^{(1)}$ and the MLEs for

$$\widehat{\beta}^{(0)}_{10\cdot 2} = \bar{y}_1^{(0)} - \frac{s_{12}}{s_{22}}\bar{y}_2^{(0)}, \widehat{\beta}^{(0)}_{12\cdot 2} = \frac{s_{12}}{s_{22}}$$

Thus we get the following MLE for the mean for y_{i2} in the missing units:

$$\widehat{\mu}_2^{(1)} = \bar{y}_2^{(0)} + \frac{\bar{y}_1^{(1)} - \bar{y}_1^{(0)}}{\frac{s_{12}}{s_{22}}}$$