BIOSTAT 880 HW1 Solution, Fall 2024

Yize Hao

Solution compiled on September 17, 2024

1.6

Consider the full data for the ith observation:

$$(m_{i1}, m_{i2}, m_{i3}, y_{i1}, y_{i2}, u_i)$$

where (m_{i1}, m_{i2}, m_{i3}) is the missing indicator vector for (y_{i1}, y_{i2}, u_i).

And

$$y_{i1} = 1 + z_{i1},$$

$$y_{i2} = 5 + 2 \times z_{i1} + z_{i2},$$

$$u_i = a \times (y_{i1} - 1) + b \times (y_{i2} - 5) + z_{i3},$$

By assumption, Y_1 is fully observed and U is fully unobserved. Thus only m_{i2} is random whose distribution is given by:

$$\Pr(m_{i2} = 1 \mid y_{i1}, y_{i2}, u_i, \phi) = \begin{cases} 1, & \text{if } u_i < 0, \\ 0, & \text{if } u_i \geq 0. \end{cases} \tag{1}$$

Note that there exists a choice of $y_i = (y_{i1}, y_{i2}, u_i)$, and $y_i^* = (y_{i1}, y_{i2}, u_i^*)$ with $u_i < 0$ and $u_i^* \geq 0$ s.t.

$$\Pr(m_{i2} = 1 \mid y_{i1}, y_{i2}, u_i, \phi) = 1,\ \Pr(m_{i2} = 0 \mid y_{i1}, y_{i2}, u_i, \phi) = 0$$

And

$$\Pr(m_{i2} = 1 \mid y_{i1}, y_{i2}, u_i^*, \phi) = 0,\ \Pr(m_{i2} = 0 \mid y_{i1}, y_{i2}, u_i^*, \phi) = 1$$

i.e. m_{i2} has different distributions under different choices of the unobserved variable u_i, which by definition, is MNAR.

Note that if u_i is independent of all the other variables in the data set, then including or excluding U in the analysis is distributional-wise meaningless. So we can choose whether to consider the latent variable accordingly.

3.1

Using the complete data:

$$\hat{\beta_1}^{cc} = \frac{\sum_{i=1}^{n} r_i \cdot y_{2i}(y_{1i} - \bar{y}_1^{cc})}{\sum_{i=1}^{n} r_i(y_{2i} - \bar{y}_2^{cc})^2}$$

where $\bar{y}_i^{cc} = \frac{1}{r}\sum_{j=1}^{n} r_j \cdot y_{ij}, i = 1, 2, j = 1, ..., n,$ and $r_i = 1$ if observed, $r_i = 0$ otherwise, and $r = \sum_{j=1}^{n} I_{\{r_j=1\}}$ i.e. the number of observed.

Then:

$$\begin{aligned}
E[\hat{\beta_1}^{cc}|Y_2] &= E\left[\frac{\sum_{i=1}^{n} r_i \cdot y_{2i}(y_{1i} - \bar{y}_1^{cc})}{\sum_{i=1}^{n} r_i(y_{2i} - \bar{y}_2^{cc})^2} \mid Y_2\right] \\
&= E\left[\frac{\sum_{i=1}^{n} r_i \cdot (\beta_0 + \beta_1 y_{2i} + \epsilon_i)(y_{1i} - \bar{y}_1^{cc})}{\sum_{i=1}^{n} r_i(y_{2i} - \bar{y}_2^{cc})^2} \mid Y_2\right] \\
&= E\left[\frac{\beta_0 \sum_{i=1}^{n} r_i(y_{2i} - \bar{y}_2^{cc})}{\sum_{i=1}^{n} r_i(y_{2i} - \bar{y}_2^{cc})^2} + \frac{\beta_1 \sum_{i=1}^{n} r_i y_{2i}(y_{2i} - \bar{y}_2^{cc})}{\sum_{i=1}^{n} r_i(y_{2i} - \bar{y}_2^{cc})^2} + \frac{\sum_{i=1}^{n} r_i(y_{2i} - \bar{y}_2^{cc}) \cdot \epsilon_i}{\sum_{i=1}^{n} r_i(y_{2i} - \bar{y}_2^{cc})^2} \mid Y_2\right] \\
&\overset{(i)}{=} \beta_0 E\left[E[\frac{\sum_{i=1}^{n} r_i\left(y_{2i} - \bar{y}_2^{cc}\right)}{\sum_{i=1}^{n} r_i\left(y_{2i} - \bar{y}_2^{cc}\right)^2} \mid Y_2, R]\right] + \beta_1 E\left[E[\frac{\sum_{i=1}^{n} r_i y_{2i}\left(y_{2i} - \bar{y}_2^{cc}\right)}{\sum_{i=1}^{n} r_i\left(y_{2i} - \bar{y}_2^{cc}\right)^2} \mid Y_2, R]\right] \\
&\quad + E\left[\sum_{i=1}^{n} \frac{r_i\left(y_{2i} - \bar{y}_2^{cc}\right)}{\sum_{j=1}^{n} r_i\left(y_{2j} - \bar{y}_2^{cc}\right)^2} \cdot \epsilon_i \mid Y_2\right] \\
&\overset{(ii)}{=} 0 + \beta_1 + \sum_{i=1}^{n}\left(E\left[\frac{r_i\left(y_{2i} - \bar{y}_2^{cc}\right)}{\sum_{j=1}^{n} r_i\left(y_{2j} - \bar{y}_2^{cc}\right)^2}|Y_2\right] \cdot E\left[\epsilon_i \mid Y_2\right]\right) \\
&\overset{(ii)}{=} \beta_1
\end{aligned}$$

(i) The first two terms are due to the Law of Total Expectation.

(ii) The first two terms can be derived directly as in your linear regression class;
Given Y_2 since r_i only depends on Y_2, $[r_i \mid Y_2] \perp\!\!\!\perp [\epsilon_i \mid Y_2], \forall i$.

(iii) The minimum assumption in linear regression is $E\left[\epsilon_i \mid Y_2\right] = 0, \forall i = 1, ..., n$

Therefore, $\hat{\beta_1^{cc}}$ is unbiased.
Similarly:

$$\begin{aligned}
E[\bar{y}_1^{cc} - \hat{\beta_1}^{cc}\bar{y}_2^{cc}|Y_2] &= E[\beta_0 + \beta_1\bar{y}_2^{cc} - \hat{\beta_1}^{cc}\bar{y}_2^{cc}|Y_2] \\
&= \beta_0
\end{aligned}$$

3.3

The count table:

	$Y_1 = 1$	$Y_1 = 0$
$Y_2 = 1$	a	b
$Y_2 = 0$	c	d

$$a = \sum_{i=1}^{n} I[Y_{i1} = 1, Y_{i2} = 1]$$

$$b = \sum_{i=1}^{n} I[Y_{i1} = 0, Y_{i2} = 1]$$

$$c = \sum_{i=1}^{n} I[Y_{i1} = 1, Y_{i2} = 0]$$

$$d = \sum_{i=1}^{n} I[Y_{i1} = 0, Y_{i2} = 0]$$

Complete data: ($R_i = 1$ if both Y_{i1} and Y_{i2} were observed, $R_i = 0$ otherwise.)

	$Y_1 = 1$	$Y_1 = 0$
$Y_2 = 1$	a_c	b_c
$Y_2 = 0$	c_c	d_c

$$a_c = \sum_{i=1}^{n} I[Y_{i1} = 1, Y_{i2} = 1] \cdot R_i$$

$$b_c = \sum_{i=1}^{n} I[Y_{i1} = 0, Y_{i2} = 1] \cdot R_i$$

$$c_c = \sum_{i=1}^{n} I[Y_{i1} = 1, Y_{i2} = 0] \cdot R_i$$

$$d_c = \sum_{i=1}^{n} I[Y_{i1} = 0, Y_{i2} = 0] \cdot R_i$$

By SLLN:

$$a_c/n \xrightarrow{p} E\left[I[Y_{i1}=1,Y_{i2}=1]\cdot R\right]$$

$$b_c/n \rightarrow E\left[I[Y_{i1}=0,Y_{i2}=1]\cdot R\right]$$

$$c_c/n \rightarrow E\left[I[Y_{i1}=1,Y_{i2}=0]\cdot R\right]$$

$$d_c/n \rightarrow E\left[I[Y_{i1}=0,Y_{i1}=0]\cdot R\right]$$

$$\frac{a_c\cdot d_c}{b_c\cdot c_c} \rightarrow \frac{E\left[I[Y_{i1}=1,Y_{i2}=1]\cdot R\right]\cdot E\left[I[Y_{i1}=0,Y_{i1}=0]\cdot R\right]}{E\left[I[Y_{i1}=1,Y_{i2}=0]\cdot R\right]\cdot E\left[I[Y_{i1}=0,Y_{i2}=1]\cdot R\right]}$$

By assumption:

$$\begin{aligned}
E\left[I[Y_{i1}=y_{i1},Y_{i2}=y_{i2}]\cdot R\right] &= E\left[I[Y_{i1}=y_{i1},Y_{i2}=y_{i2}]\cdot I_{\{R=1\}}\right] \\
&= P\left(Y_{i1}=y_{i1},Y_{i2}=y_{i2},R=1\right) \\
&= P\left(Y_{i1}=y_{i1},Y_{i2}=y_{i2}\right)P\left(R=1\mid Y_{i1}=y_{i1},Y_{i2}=y_{i2}\right) \\
&= P\left(Y_{i1}=y_{i1},Y_{i2}=y_{i2}\right)e^{\eta_1(y_{i1})+\eta_2(y_{i1})}
\end{aligned}$$

where $\eta_i(\cdot)$ is an arbitrary function of $y_i. =$ for $i=1,2$.
Therefore:

$$\begin{aligned}
\frac{a_c\cdot d_c}{b_c\cdot c_c} &\rightarrow \frac{E\left[I[Y_{i1}=1,Y_{i2}=1]\cdot R\right]\cdot E\left[I[Y_{i1}=0,Y_{i1}=0]\cdot R\right]}{E\left[I[Y_{i1}=1,Y_{i2}=0]\cdot R\right]\cdot E\left[I[Y_{i1}=0,Y_{i2}=1]\cdot R\right]} \\
&= \frac{P\left(Y_{i1}=1,Y_{i2}=1\right)e^{\eta_1(1)+\eta_2(1)}\cdot P\left(Y_{i1}=0,Y_{i2}=0\right)e^{\eta_1(0)+\eta_2(0)}}{P\left(Y_{i1}=0,Y_{i2}=1\right)e^{\eta_1(0)+\eta_2(1)}\cdot P\left(Y_{i1}=1,Y_{i2}=0\right)e^{\eta_1(1)+\eta_2(0)}} \\
&= \frac{P\left(Y_{i1}=1,Y_{i2}=1\right)\cdot P\left(Y_{i1}=0,Y_{i2}=0\right)}{P\left(Y_{i1}=0,Y_{i2}=1\right)\cdot P\left(Y_{i1}=1,Y_{i2}=0\right)}
\end{aligned}$$

is consistent.