$Y = (y_1, \cdots, y_n)$ and $M = (m_1, \cdots, m_n)$; M indicates missingness in Y

$E(Y) = \mu$ population mean

$E(Y \mid M=0) = \bar{Y}$ sample mean ; $E[\bar{Y}] = E\left[\frac{1}{n}\sum_1^n Y_i\right] = \frac{1}{n}\sum E[Y_i] = E(Y) = \mu$

$M_{i\in\{1,\cdots,n\}} \sim Bern(expit(y_i))$; $P(m_i = 1) = expit(\beta_0 + \beta_1 \cdot y_i)$; note that M depends on Y

Suppose we have k missing values where $0 \le k \le n$. Y^* indicates Y after imputed

$$y_i^* = \begin{cases} y_i & (m_i = 0) \quad \text{total } n-k \text{ entries} \\ \bar{Y} & (m_i = 1) \quad \text{total } k \text{ entries ; impute } \bar{Y} \text{ (sample mean) when } M=1 \end{cases}$$

$$Y^* = \frac{1}{n}\sum_{i=1}^n y_i^* = \frac{1}{n}\left\{\sum_{i=k+1}^n y_i + k\cdot\bar{Y}\right\} = \frac{1}{n}\left\{(n-k)\cdot\bar{Y} + k\cdot\bar{Y}\right\} = \bar{Y}$$

Thus, $E[\bar{Y}] = E[Y^*]$

$Bias(Y^*) = E(Y^*) - Y$

$E(Y^*) = E(\bar{Y}) = \mu$ $\quad\therefore$ Thus, Bias in imputed for mean imputation is $Bias(y_i^*) = \mu - y_i$

OR

$$E(\bar{Y}) = E[Y \mid M=0] \quad \text{by definition}$$

By definition, $E[X \mid A] = \dfrac{E[X;A]}{P(A)}$

$$= \frac{E[Y;M=0]}{P(M=0)}$$

$$= \frac{E[Y;M=0]}{P(M=0)} = \frac{E[Y\cdot I(M=0)]}{P(M=0)}$$

Let $R = 1-M$ where R = non-missing indicator, then

$$E(\bar{Y}) = \frac{E[Y\cdot I(R=1)]}{P(R=1)} = \frac{E[Y\cdot R]}{E(R)} = \frac{E_Y[E_{R|Y}[y_i r_i \mid y_i]]}{E_Y[E_{R|Y}[r_i \mid y_i]]} = \frac{E_Y[y_i \cdot E_{R|Y}[r_i \mid y_i]]}{E_Y[E_{R|Y}[r_i \mid y_i]]}$$

note $E_{R|Y}[r_i \mid y_i] = 1 - expit(\cdot)$

$$= \frac{E_Y\left[y_i \cdot \frac{1}{1+\exp(\beta_0+\beta_1 y_i)}\right]}{E_Y\left[\frac{1}{1+\exp(\beta_0+\beta_1 y_i)}\right]}$$

$$Y_i^* = Y_i\,[M_i=0] + \overset{\text{constant}}{c}\,[M_i=1]$$

$$P(M_i=1) = \text{expit}(\beta_0 + \beta_1 Y_i)$$

$$E(Y_i^*) = E\left[Y_i[M_i=0] + c\,[M_i=1]\right]$$

$$= E_Y\left[E_{M|Y}\{Y_i[M_i=0] + c\,[M_i=1]\}\right]$$

$$= E_Y\Big[\,Y_i\cdot\underbrace{(1-\text{expit}(\beta_0+\beta_1 Y_i))}_{P(M=0)} + c\cdot\underbrace{\text{expit}(\beta_0+\beta_1 Y_i)}_{P(M=1)}\,\Big]$$

$$= \int y\cdot\{1-\text{expit}(\beta_0+\beta_1 y)\}\,f_Y(y)\,dy + c\cdot\int \text{expit}(\beta_0+\beta_1 y)\,f_Y(y)\,dy$$

Assume $Y \sim \text{normal}(\mu, \sigma^2)$; $f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$

$$= \int_{-\infty}^{\infty} y\cdot\{1-\text{expit}(\beta_0+\beta_1 y)\}\cdot\frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}dy + c\cdot\int_{-\infty}^{\infty}\text{expit}(\beta_0+\beta_1 y)\,\frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}dy$$

Let $y = \mu + \sigma\sqrt{2}\,t$, $dy = \sigma\sqrt{2}\,dt$ $\Rightarrow$ $\frac{y-\mu}{\sigma\sqrt{2}} = t$, $\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2 = t^2$

Also $dy = \sigma\sqrt{2}\,dt$ So, $\sigma\sqrt{2\pi}$ in the denominator would now be $\sqrt{\pi}$

$$= \int_{-\infty}^{\infty}\underbrace{(\mu+\sigma\sqrt{2}t)\{1-\text{expit}(\beta_0+\beta_1(\mu+\sigma\sqrt{2}t))\}\cdot\frac{1}{\sqrt{\pi}}}_{f_1(t_i)}\,e^{-t^2}dt + \underbrace{c\int_{-\infty}^{\infty}\text{expit}(\beta_0+\beta_1(\mu+\sigma\sqrt{2}t))\,\frac{1}{\sqrt{\pi}}}_{f_2(t_i)}\cdot e^{-t^2}dt$$

$\approx \sum_{i=1}^{n} w_i\cdot f(t_i)$ by Gauss-Hermite quadrature

If we let $k = \beta_0 + \beta_1(\mu+\sigma\sqrt{2}t)$, $f_1(t_i) = (\mu+\sigma\sqrt{2}t_i)\{1-\text{expit}(k)\}\frac{1}{\sqrt{\pi}}$, $f_2(t_i) = c\cdot\text{expit}(k)\cdot\frac{1}{\sqrt{\pi}}$

We can calculate numerical bias if we plug in parameter values $\beta_0, \beta_1, \mu, \sigma, c$

Also, t_i and w_i can be computed with R

Note ***

If we let $P[M_i=1] = \pi_i$

Then, $E(Y_i^*) = E_Y[\,Y_i\cdot(1-\pi_i) + c\cdot\pi_i\,] = (1-\pi_i)\,E_Y(Y_i) + \pi_i\cdot c$ if MCAR $(Y_i \perp\!\!\!\perp \pi_i)$

π_i Missingness not related to any other values (including Y_i)

$$= \mu - \mu\pi_i + c\,\pi_i = \mu\cdot(1-\pi_i) + c\cdot\pi_i = \mu + (c-\mu)\cdot\pi_i$$

bias term added (under MCAR)

$\Rightarrow$ bias increases with the increase in absence rate π_i

$\Rightarrow$ but, our study assumes MAR (missingness depends on observed)