Answers to HW #3

Problem 4.8. Note that the table shows large sample bias, not the exact bias defined as difference between the expectation of the estimate and the true value. Suppose the sample size is \( n \) and the first \( r \) observations are complete cases. Denote the complete-case estimates for regression parameters are \((b_{20.1}, b_{21.1}, s_{22.1})\) and \((b_{10.2}, b_{12.2}, s_{11.2})\); \((\bar{y}_1, s_{11})\), \((\bar{y}_2, s_{22})\) are the complete-case means and variances.

1. \(U\text{mean.} \) Imputing the missing values via complete-case mean: \(\bar{y}_{i2} = \bar{y}_2, \quad i = r + 1, \ldots, n.\)
   (a) \(\hat{\mu}_2 = \bar{y}_2, \text{ so bias}=0.\)
   (b) \(\hat{\sigma}_{22} = \frac{1}{n} \sum_{i=1}^{r} (y_{i2} - \bar{y}_2)^2 \approx \frac{r}{n} \sigma_{22}. \text{ So bias}= -\lambda \sigma_{22}.\)
   (c) \(\hat{\beta}_{21.1} = \frac{\sum_{i=1}^{r} (y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)}{\sum_{i=1}^{r} (y_{i1} - \bar{y}_1)^2} \approx \frac{r}{n} \beta_{21.1}.\)
   (d) \(\hat{\beta}_{12.2} = \text{complete case estimate. So bias}=0.\)

2. \(U\text{draw.} \) Imputing via: \(\bar{y}_{i2} = \bar{y}_2 + e_i, \) where \(e_i \sim N(0, s_{22}).\) As \( n \to \infty,\)
   (a) \(\hat{\mu}_2 = \bar{y}_2 + \frac{1}{n} \sum_{i=r+1}^{n} e_i := \bar{y}_2 + \frac{n-r}{n} \bar{e} \to \mu_2 + 0 = \mu_2.\)
   (b) \(\hat{\sigma}_{22} = \frac{\sum_{i=1}^{r} (y_{i2} - \bar{y}_2 - \bar{e})^2 + \sum_{i=r+1}^{n} (e_i - \bar{e})^2}{n-1} \to \frac{r \sigma_{22} + (n-r)\sigma_{22}}{n-1} \to \sigma_{22}.\)
   (c) \(\hat{\beta}_{21.1} \approx \frac{\sum_{i=1}^{r} (y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2) + \sum_{i=r+1}^{n} (y_{i1} - \bar{y}_1)e_i}{\sum_{i=1}^{r} (y_{i2} - \bar{y}_2)^2 + \sum_{i=r+1}^{n} e_i^2} \approx \frac{r \beta_{21.1}}{n \sigma_{11}} + \frac{r}{n} \beta_{21.1}, \text{ since } E(e_i) = 0 \) and \( e_i \) is independent from \( y_{1i}'s.\)
   (d) \(\hat{\beta}_{12.2} \approx \frac{\sum_{i=1}^{r} (y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2) + \sum_{i=r+1}^{n} (y_{i1} - \bar{y}_1)e_i}{\sum_{i=1}^{r} (y_{i2} - \bar{y}_2)^2 + \sum_{i=r+1}^{n} e_i^2} \to \frac{r \beta_{12.2} + 0}{r \sigma_{22} + (n-r)\sigma_{22}} = \frac{r}{n} \beta_{12.2}.\)

3. \(C\text{mean.} \) Impute via: \(\bar{y}_{i2} = b_{20.1} + b_{21.1}y_{i1} = \bar{y}_2 + b_{21.1}(y_{i1} - \bar{y}_1).\)
   (b) \(\hat{\sigma}_{22} \approx r \sigma_{22} + (n-r)\beta_{21.1}^2 \sigma_{11} = r \sigma_{22} + (n-r)\rho^2 \sigma_{22}.\)

4. \(C\text{draw.} \) Imputing the missing values by \(\bar{y}_{i2} = \bar{y}_2 + b_{21.1}(y_{i1} - \bar{y}_1) + e_i, \) where \(e_i \sim N(0, s_{22.1}).\)
   (b) \(\hat{\sigma}_{22} \approx \frac{1}{n} \{r \sigma_{22} + (n-r)(\beta_{21.1}^2 \sigma_{11} + \sigma_{22.1})\} = \sigma_{22}, \text{ since } \sigma_{22.1} = (1-\rho^2)\sigma_{22}.\)
   (d) From (b), the denominator of \(\hat{\beta}_{12.2}\) is about \(n\sigma_{22};\) the numerator is about \(n\sigma_{12}\) since \(e_i\) is independent from \(y_{1i}'s\) and has mean 0.

Other scenarios can be derived similarly.

Problem 4.11. The HD-2 estimate can be represented as

\[
\bar{y}_{HD2} = \frac{1}{n}(r \bar{y}_R + k r \bar{y}_R + \sum_{i=1}^{r} H_i y_i),
\]
where $H_i$ is the indicator function that $y_i$ is selected for imputation. Since the selection is without replacement for the $t$ missing values, (i) $H_i \sim \text{Bernoulli}(\frac{t}{r})$; (2) $E(H_iH_j) = \text{pr}(H_i = H_j = 1) = \frac{t(t-1)}{r(r-1)}$, i.e., the probability of selecting both observations $i$ and $j$, for $i \neq j$. Then $\text{Var}(\bar{y}_{HD2} \mid Y_{obs}) = \frac{1}{n^2}\text{Var}(\sum_{i=1}^t H_iy_i \mid Y_{obs})$ can be calculated accordingly.

The proportionate variance increase is about

\[
\frac{tr-t}{n} \leq \frac{tn-t-t}{n} \quad \text{equivalent only if } k=0
\]

\[
= \frac{2t(n/2-t)}{n^2} \leq \frac{1}{4} \frac{(t+n/2-t)^2}{n^2}. \quad \text{equivalent only if } t = n/2 - t. \\
= \frac{1}{8}
\]

The equality holds only if $k = 0$ and $t = n/2 - t$. Thus $t = n/4$ and $r = n-t = 3/4t$. 