5. Fractional Hot deck Imputation

1 Introduction

• Suppose that we are interested in estimating $\theta_1 = E(Y)$ or even $\theta_2 = Pr(Y < c)$ where $y \sim f(y \mid x)$ where $x$ is always observed and $y$ is subject to missingness.

• Assume MAR in the sense that $Pr(\delta = 1 \mid x, y)$ does not depend on $y$.

• We do not want to make strong parametric assumptions on $f(y \mid x)$.

• If $x$ is a categorical variable with support $\{1, \cdots, G\}$, then $f(y \mid x = g)$ can be described as

$$y \mid (x = g) \overset{i.i.d.}{\sim} (\mu_g, \sigma^2_g)$$

which is sometimes called cell mean model.

• Even when $x$ is not categorical, one can consider an approximation

$$f(y \mid x) \cong \sum_{g=1}^{G} \pi_g(x) f_g(y)$$

where $\pi_g(x) = P(z = g \mid x)$ and $f_g(y) = f(y \mid z = g)$. Such approximation is accurate when

$$f(y \mid x, \text{group}) = f(y \mid \text{group}).$$

• Group $g$ is often called imputation cell.

• Under MAR, one imputation method is to take a random sample from the set of respondents in the same imputation cell.

• Hot deck imputation
– Partition the sample into $G$ groups: $A = A_1 \cup A_2 \cup \cdots \cup A_G$.

– In group $g$, we have $n_g$ elements and $r_g$ respondents. We need to impute $y$ values for the nonrespondents.

– For each group $A_g$, select $m_g = n_g - r_g$ imputed values from $r_g$ respondents with replacement (or without replacement).

• Example 1

– $A_g = A_{Rg} \cup A_{Mg}$ with $A_{Rg} = \{ i \in A_g; \delta_i = 1 \}$ and $A_{Mg} = \{ i \in A_g; \delta_i = 0 \}$.

– Imputation mechanism: $y^*_j \sim Uniform \{ y_i; i \in A_{Rg} \}$. That is, $y^*_j = y_i$ with probability $1/r_g$ for $i \in A_{Rg}$ and $j \in A_{Mg}$.

– Note that

$$E_I(y^*_j) = \frac{1}{r_g} \sum_{i \in S_{Rg}} y_i \equiv \bar{y}_{Rg}$$

$$V_I(y^*_j) = \frac{1}{r_g} \sum_{i \in S_{Rg}} (y_i - \bar{y}_{Rg})^2 \equiv \bar{S}_{Rg}^2$$

where $E_I(\cdot)$ is the expectation taken with respect to the imputation mechanism.

– Imputed estimator of $\theta = E(Y)$:

$$\hat{\theta}_I = \frac{1}{n} \sum_{i=1}^{n} \{ \delta_i y_i + (1 - \delta_i) y^*_i \}$$

– Variance

$$V(\hat{\theta}_I) = V \left\{ E_I(\hat{\theta}_I) \right\} + E \left\{ V_I(\hat{\theta}_I) \right\}$$

$$= V \left\{ n^{-1} \sum_{g=1}^{G} n_g \bar{y}_{Rg} \right\} + E \left\{ n^{-2} \sum_{g=1}^{G} m_g \bar{S}_{Rg}^2 \right\}.$$ 

• Under the model

$$y_i \mid (i \in S_g) \sim i.i.d. \left( \mu_g, \sigma_g^2 \right),$$

the variance can be written

$$V(\hat{\theta}_I) = V \left\{ n^{-1} \sum_{g=1}^{G} n_g \mu_g \right\} + E \left\{ n^{-2} \sum_{g=1}^{G} \left( n_g + 2m_g + \frac{m_g(m_g - 1)}{r_g} \right) \sigma_g^2 \right\}.$$
Note that
\[ V\left( \hat{\theta}_n \right) = V\left\{ n^{-1} \sum_{g=1}^{G} n_g \mu_g \right\} + E\left\{ n^{-2} \sum_{g=1}^{G} n_g \sigma_g^2 \right\}. \]
Thus, the variance is increased.

- Variance is increased after imputation. Two sources:
  1. Reduced sample size.
  2. Randomness due to imputation mechanism in stochastic imputation.
- Variance estimation after imputation: A Naive approach is treating the imputed values as if observed and applying the standard variance estimation formula to the imputed data. Such naive approach underestimates the true variance!
- **Lemma 5.1** If \( E(y_i^*) = E(y_i) \) and \( V(y_i^*) = V(y_i) \), the naive variance estimator \( \hat{V}_I = n^{-1}S_I^2 \) has expectation
  \[ E\{\hat{V}_I\} = V(\hat{\theta}_n) - \frac{1}{n-1} \left\{ V(\hat{\theta}_n) - V(\hat{\theta}_I) \right\} \cong V(\hat{\theta}_n). \] (3)
  A proof of Lemma 5.1 is given in Appendix A.
- Example 1 (Continued): Note that
  \[ V(y_i^*) = E\{V_I(y_i^*)\} + V\{E_I(y_i^*)\} \]
  \[ = V(\bar{y}_{Rg}) + E\{S_{Rg}^2\} \]
  \[ = r^{-1}_g \sigma^2_g + (1 - r^{-1}_g)\sigma^2_g = \sigma^2_g = V(y_i). \]
Thus, the assumptions for (3) are satisfied and obtain
\[ E\{\hat{V}_I\} = V\left\{ n^{-1} \sum_{g=1}^{G} n_g \mu_g \right\} + \frac{1}{n(n-1)} E\left\{ \sum_{g=1}^{G} \left( n_g - \frac{n_g}{n} - 2 \frac{m_g}{n} - \frac{m_g(m_g - 1)}{nr_g} \right) \right\} \sigma^2_g. \]
Thus, we can write
\[ V\left( \hat{\theta}_I \right) = V\left( \hat{\theta}_n \right) + E\left\{ \sum_{g=1}^{G} c_g \sigma^2_g \right\} \]
for some \( c_g \). The (approximate) bias-corrected estimator is
\[ \hat{V} = \hat{V}_I + \sum_{g=1}^{G} c_g S_{Rg}^2. \]
2 Fractional Hot deck imputation

- For each \( j \in A_Mg \), select \( M \) imputed values from \( A_Rg \) using SRS (simple random sampling) without replacement. Let \( \bar{y}_{Ij} = M^{-1} \sum_{i \in A_R} d_{ij} y_i \) be the mean of the \( M \) imputed values for \( y_j \) where \( d_{ij} = 1 \) if \( y_i \) is selected as one of the imputed values for missing \( y_j \).

- The resulting fractional hot deck imputation (FHDI) estimator of \( \theta_1 = E(Y) \) is

\[
\hat{\theta}_{FHDI,1} = \frac{1}{n} \sum_{j=1}^{n} \left\{ \delta_j y_j + (1 - \delta_j) \bar{y}_{Ij} \right\} = \frac{1}{n} \sum_{j=1}^{n} \left\{ \delta_j y_j + (1 - \delta_j) \sum_{i \in A_R} w_{ij}^* y_i \right\} \tag{4}
\]

where \( w_{ij}^* = d_{ij}/M \). Also, the FHDI estimator of \( \theta_2 = P(Y < 1) \) is

\[
\hat{\theta}_{FHDI,2} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \delta_i I(y_i < 1) + (1 - \delta_i) \sum_{j=1}^{M} w_{ij}^* I(y_{ij}^* < 1) \right\}.
\]

Thus, one set of FHDI data can be used to estimate several parameters.

- How to estimate the variance of FHDI estimators?

- Kim and Fuller (2004) considered a modified jackknife method where the fractional weights are modified to estimate the variance correctly.

- Jackknife method for complete data: For \( \hat{\theta}_n = \bar{y}_n \), the jackknife variance estimator is defined by

\[
\hat{V}_{JK}(\bar{y}_n) = \frac{n}{n} \sum_{k=1}^{n-1} \left( \bar{y}_n^{(k)} - \bar{y}_n \right)^2
\]

where

\[
\bar{y}_n^{(k)} = \sum_{i=1}^{n} w_i^{(k)} y_i
\]

and

\[
w_i^{(k)} = \begin{cases} 0 & \text{if } i = k \\ (n-1)^{-1} & \text{otherwise.} \end{cases}
\]

Note that

\[
\bar{y}_n^{(k)} - \bar{y}_n = -\frac{y_k - \bar{y}_n}{n-1}
\]
and so
\[
\hat{V}_{JK}(\bar{y}_n) = \sum_{k=1}^{n} \frac{n-1}{n} (\bar{y}^{(k)}_n - \bar{y}_n)^2 \\
= \frac{1}{n(n-1)} \sum_{k=1}^{n} (y_k - \bar{y}_n)^2 \\
= n^{-1} \sigma_n^2
\]
and it is unbiased for \( V(\bar{y}_n) \). Similarly, the jackknife variance estimator of \( \hat{\theta}_{n,2} = n^{-1} \sum_{i=1}^{n} I(y_i < 1) \) is computed by
\[
\hat{V}_{JK}(\hat{\theta}_{n,2}) = \sum_{k=1}^{n} \frac{n-1}{n} \left( \hat{\theta}_{n,2}^{(k)} - \hat{\theta}_{n,2} \right)^2
\]
where
\[
\hat{\theta}_{n,2}^{(k)} = \sum_{i=1}^{n} w_i^{(k)} I(y_i < 1).
\]
Thus, jackknife is easy to implement as we have only to repeat the same estimation procedure using \( w_i^{(k)} \) instead of \( w_i \).

- Now, to estimate the FHDI estimator in (4), first note that the FHDI estimator of \( \theta_1 = E(Y) \) can be written as
\[
\hat{\theta}_{FHDI,1} = \sum_{i=1}^{n} \delta_i \alpha_i y_i
\]
where
\[
\alpha_i = \frac{1}{n} \left( 1 + \frac{1}{M} \sum_{j \in A_M} d_{ij} \right)
\]
and \( d_{ij} = 1 \) if \( y_i \) is selected as one of the imputed values for missing \( y_j \). The variance of \( \hat{\theta}_{FHDI,1} \) is then
\[
V \left( \hat{\theta}_{FHDI,1} \right) = V \left\{ n^{-1} \sum_{g=1}^{G} n_g \mu_g \right\} + E \left\{ \sum_{g=1}^{G} \sum_{i \in A_{R_g}} \alpha_i^2 \sigma_y^2 \right\}. \tag{5}
\]
- The modified jackknife method of Kim and Fuller (2004) is computed by
\[
\hat{V}_{JK}(\hat{\theta}_{FHDI,1}) = \sum_{k=1}^{n} \frac{n-1}{n} \left( \hat{\theta}_{FHDI,1}^{(k)} - \hat{\theta}_{FHDI,1} \right)^2, \tag{6}
\]
where
\[ \hat{\theta}_{FHDI,1}^{(k)} = \sum_{j=1}^{n} w_{ij}^{(k)} \left( \delta_j y_j + (1 - \delta_j) \sum_{i \in A_R} w_{ij}^{*(k)} d_{ij} y_i \right) := \sum_{i \in A_R} \alpha_i^{(k)} y_i \]
and
\[ w_{ij}^{*(k)} = \begin{cases} w_{ij}^{*} - \phi_k & \text{if } i = k \text{ and } d_{kj} = 1 \\ w_{ij}^{*} + \phi_k(M - 1)^{-1} & \text{if } i \neq k \text{ and } d_{kj} = d_{ij} = 1 \\ w_{ij}^{*} & \text{otherwise} \end{cases} \]

Here, \( \phi_k \in (0, 1/M) \) is a constant to be determined. (Note that \( \sum_j w_{ij}^{*(k)} = 1 \).)

- The expected value of the modified jackknife variance estimator in (6) is given by
\[ E\{ \hat{V}_{JK}(\hat{\theta}_{FHDI,1}) \} = V \left\{ n^{-1} \sum_{g=1}^{G} n_g \mu_g \right\} + E \left\{ \sum_{k=1}^{n} \frac{n - 1}{n} \sum_{g=1}^{G} \sum_{i \in A_R g} \left( \alpha_i^{(k)} - \alpha_i \right)^2 \sigma_g^2 \right\}. \]

- To obtain the variance of FHDI estimator correctly, we want to achieve
\[ \sum_{k=1}^{n} \sum_{i \in A_R g} \left( \alpha_i^{(k)} - \alpha_i \right)^2 = \frac{n}{n - 1} \sum_{i \in A_R g} \alpha_i^2. \]

- There are two approaches of computing the suitable set of replicated fractional weights that satisfy (8). One is a bisection method and the other is a closed form solution described in Appendix A.

- To describe the bisection method, we use the following steps:
  1. Set \( \phi_k^{(0)} = 1/(2M) \).
  2. Using the current value of \( \phi_k^{(t)} \), compute
\[ Q_g^{(t)}(\phi^{(t)}) = \sum_{k=1}^{n} \sum_{i \in A_R g} \left( \alpha_i^{(k)} - \alpha_i \right)^2 - \frac{n}{n - 1} \sum_{i \in A_R g} \alpha_i^2. \]
  3. If \( Q_g^{(t)}(\phi^{(t)}) < 0 \), then set \( \phi_k^{(t+1)} = \phi_k^{(t)} + (1/2)^{t+1}(1/M) \) for all \( k \in A_R g \).
    If \( Q_g^{(t)}(\phi^{(t)}) > 0 \), then set \( \phi_k^{(t+1)} = \phi_k^{(t)} - (1/2)^{t+1}(1/M) \) for all \( k \in A_R g \).
    Continue this update for \( g = 1, \cdots , G \).
  4. Continue the process until \( |Q_g^{*}(\phi^{(t)})| < \epsilon \) for a sufficiently small \( \epsilon > 0 \).


3 Fully efficient fractional imputation

- Under the cell mean model (1), the best imputation is to use the cell mean imputation:

\[
\hat{\theta}_{Id} = n^{-1} \sum_{g=1}^{G} \sum_{i \in A_g} \{\delta_i y_i + (1 - \delta_i) \bar{y}_{RG}\},
\]

which is algebraically equivalent to fully efficient fractional imputation (FEFI) estimator

\[
\hat{\theta}_{FEFI} = n^{-1} \sum_{g=1}^{G} \sum_{j \in A_g} \left\{\delta_j y_i + (1 - \delta_j) \sum_{i \in A_{RG}} w_{ij}^* y_i\right\}, \tag{9}
\]

with \( w_{ij}^* = 1/r_g \). The FEFI estimator uses all the respondents in the same cell as the imputed values for each missing value. Note that (9) is equivalent to

\[
\hat{\theta}_{FEFI} = n^{-1} \sum_{g=1}^{G} \frac{n_g}{r_g} \sum_{i \in A_{RG}} y_i, \tag{10}
\]

which is equivalent to using \( \hat{\pi}_g^{-1} = n_g/r_g \) as the nonresponse adjustment factor that is multiplied to the original weight of the respondents in the cell \( g \).

- For the fractional hot deck imputation estimator \( \hat{\theta}_{FHDI} \) in Section 2, we can obtain

\[
V(\hat{\theta}_{FHDI}) = V(\hat{\theta}_{FEFI}) + V(\hat{\theta}_{FHDI} - \hat{\theta}_{FEFI}) \tag{11}
\]

because we have \( E_I(\hat{\theta}_{FHDI}) = \hat{\theta}_{FEFI} \). The second term is the variance due to random selection in the fractional hot deck imputation and will be zero if \( M \to \infty \).

- How to reduce the second term in (11) ?

1. Increase \( M \).
2. Use calibration weighting.
3. Use a balanced sampling mechanism to select donors.

- Calibration weighting approach (Fuller and Kim, 2005):
1. For each \( j \in A_{Mg} \), we select \( M \) elements from \( \{ y_i; i \in A_{Rg} \} \) at random with equal probability.

2. The initial fractional weight for each donor is \( w_{ij}^{*(0)} = 1/M \) for \( d_{ij} = 1 \).

3. The fractional weights are modified to satisfy
\[
\sum_{i \in A_{Rg}} y_i + \sum_{j \in A_{Mg}} \sum_{i \in A_{Rg}} w_{ij}^* d_{ij} y_i = n_g \bar{y}_{Rg}
\]
and
\[
\sum_{i \in A_{Rg}} w_{ij}^* d_{ij} = 1.
\]

One solution is
\[
w_{ij}^* = w_{ij}^{*(0)} + (\bar{y}_{Rg} - \bar{y}_{Ig}) T_g^{-1} w_{ij}^{*(0)} d_{ij} (y_i - \bar{y}_{Ij})
\] (12)
where \( \bar{y}_{Ij} = \sum_{i \in A_{Rg}} w_{ij}^{*(0)} y_i \), \( T_g = \sum_{j \in A_{Mg}} \sum_{i \in A_{Rg}} w_{ij}^{*(0)} d_{ij} (y_i - \bar{y}_{Ij})^2 \), and
\[
\bar{y}_{Ig} = \frac{1}{n_g} \left\{ \sum_{i \in A_{Rg}} y_i + \sum_{j \in A_{Mg}} \bar{y}_{Ij} \right\}.
\]

• Modifying the weights to satisfy certain constraints is a popular problem in statistics. Such weighting is often called calibration weighting. Regression weighting is of the form in (12) and is extensively discussed in survey sampling courses (Stat 521, Stat 621).

• Balanced imputation approach (Chauvet et al., 2011): Apply a balanced sampling technique to achieve that
\[
\sum_{i \in A_{Rg}} y_i + \sum_{j \in A_{Mg}} \left\{ M^{-1} \sum_{i \in A_{Rg}} d_{ij} y_i \right\} = n_g \bar{y}_{Rg}.
\]

• Note: Balanced sampling is a set of sampling method that satisfies some constraints:
\[
\sum_{i \in A} w_i z_i = \sum_{i \in U} z_i
\]
where \( w_i \) is the design weight ( = inverse of the first-order inclusion probability) and \( z_i \) is the design variable. Stratified sampling is one example of balanced
sampling when $z_i$ is categorical. For more general case, we may use Cube method (Deville and Tillé, 2004) or rejective method (Fuller, 2009).

- For variance estimation, the first term of (11) is easy to estimate because we can easily take into account of the sampling variability of $\hat{\pi}_g$ into estimation. That is, writing $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_G)$, we can express $\hat{\theta}_{FEFI} = \hat{\theta}_{FEFI}(\hat{\pi})$. To estimate the variance of $\hat{\theta}_{FEFI}$, we need to incorporate the sampling variability of $\hat{\pi}$ in $\hat{\theta}_{FEFI}$. Either Taylor linearization or replication method can be used.

- Linearization method: Let $U_g(\pi_g) = \sum_{i \in A_g} (\delta_i - \pi_g)$ be an estimating function for $\pi_g$. (i.e. $\hat{\pi}_g$ is obtained by solving $U_g(\hat{\pi}_g) = 0$ for $\pi_g$.) Now, using

$$\hat{\theta}_{FEFI}(\hat{\pi}) \approx \hat{\theta}_{FEFI}(\pi) + \sum_{g=1}^G E \left( \frac{\partial \hat{\theta}_{FEFI}}{\partial \pi_g} \right) (\hat{\pi}_g - \pi_g)$$

and

$$0 = U_g(\hat{\pi}_g) \approx U_g(\pi_g) + E \left( \frac{\partial U_g}{\partial \pi_g} \right) (\hat{\pi}_g - \pi_g),$$

we have

$$\hat{\theta}_{FEFI}(\hat{\pi}) \approx \hat{\theta}_{FEFI}(\pi) - \sum_{g=1}^G E \left( \frac{\partial \hat{\theta}_{FEFI}}{\partial \pi_g} \right) E \left( \frac{\partial U_g}{\partial \pi_g} \right)^{-1} U_g(\pi_g)$$

$$= \frac{1}{n} \sum_{g=1}^G \sum_{i \in A_{\pi_g}} \frac{1}{\pi_g} y_i + \frac{1}{n} \sum_{g=1}^G \left( n_g - \frac{r_g}{\pi_g} \right) \mu_g$$

$$= \frac{1}{n} \sum_{g=1}^G \sum_{i \in A_{\pi_g}} \left\{ \mu_g + \frac{\delta_i}{\pi_g} (y_i - \mu_g) \right\}$$

$$:= \frac{1}{n} \sum_{i=1}^n \eta_i$$

where $\eta_i = \mu_g + (\delta_i/\pi_g)(y_i - \mu_g)$ if $i \in A_g$. Once $\eta_i$ is calculated, we can apply a standard variance formula to $\hat{\eta}_i = \hat{\mu}_g + (\delta_i/\hat{\pi}_g)(y_i - \hat{\mu}_g)$ to obtain the linearized variance estimator.

- Replication method (such as jackknife) is straightforward.
4 FHDI method using a parametric model

• Assume $y \sim f(y \mid x)$ where $x$ is always observed and $y$ is subject to missingness.

• Now suppose that (2) does not hold and the cell mean model (1) is not satisfied. (That is, we does not create imputation cells.) But, we still want to take the real observations as the imputed values.

• Kim and Yang (2014) approach: Three steps

  1. Fully efficient fractional imputation (FEFI) by choosing all the respondents as donors. That is, we use $M = r$ imputed values for each missing unit, where $r$ is the number of respondents in the sample. Compute the fractional weights.

  2. Use a systematic PPS sampling to select $m (<< r)$ donors from the FEFI using the fractional weights as the size measure.

  3. Use a calibration weighting technique to compute the final fractional weights (which lead to the same estimates of FEFI for some items).

• Step 1: FEFI step

  – Want to find the fractional weights $w_{ij}^*$ when the $j$-th imputed value $y_i^{*(j)}$ is taken from the $j$-th value in the set of the respondents.

  – Without loss of generality, we assume that the first $r$ elements respond and write $y_i^{*(j)} = y_j$.

  – Recall that

    $$w_{ij}^* \propto f(y_i^{*(j)} \mid x_i; \hat{\theta})/h(y_i^{*(j)} \mid x_i)$$

    when $y_i^{*(j)}$ are generated from $h(y \mid x_i)$.

  – We have only to find $h(y_i^{*(j)} \mid x_i)$ when we use $y_i^{*(j)} = y_j$.

  – We can treat $\{y_i; \delta_i = 1\}$ as a realization from $f(y \mid \delta = 1)$, the marginal distribution of $y$ among respondents.
– Now, we can write
\[
f(y_j|δ_j = 1) = \int f(y_j | x, δ_j = 1) f(x | δ_j = 1)dx
= \int f(y_j | x) f(x | δ_j = 1)dx
\approx \frac{1}{r} \sum_{k=1}^n δ_k f(y_j | x_k).
\]

– Thus, the fractional weight for \( y_i^*(j) = y_j \) becomes
\[
w_{ij}^* \propto \frac{f(y_j | x_i; \hat{\theta})}{\sum_{k=1}^n δ_k f(y_j | x_k; \hat{\theta})}
\tag{13}
\]
with \( \sum_{j \in A_R} w_{ij}^* = 1 \), where \( \hat{\theta} \) is computed from
\[
\sum_{i=1}^n δ_i S(\theta; x_i, y_i) = 0
\]
with \( S(\theta; x, y) = \partial \log f(y | x; \theta) / \partial \theta \).

• Step 2: Sampling Step

– FEFI uses all the elements in \( A_R \) as donors for each missing \( i \).

– Want to reduce the number of donors to, say, \( m = 10 \).

– For each \( i \), we can treat the FEFI donor set as the weighted population and apply a sampling method to select a smaller set of donors.

– Fractional weights (13) for FEFI can be used as the selection probabilities for the PPS sampling.

– That is, our goal is to obtain a (systematic) PPS sample \( D_i \) of size \( m \) from the FEFI donor set of size \( M = r \), using \( w_{ij}^* \) as the selection probability assigned to the \( j \)-th element in \( A_R \). (Note that \( w_{ij}^* \) satisfies \( \sum_{j=1}^M w_{ij}^* = 1 \) and \( w_{ij}^* > 0 \).)

• Step 3: Calibration Step

– After we select \( D_i \) from the complete set of respondents, the selected donors in \( D_i \) are assigned with the initial fractional weights \( w_{ij0}^* = 1/m \).
– The fractional weights are further adjusted to satisfy
\[
\sum_{i=1}^{n}\{(1 - \delta_i) \sum_{j \in D_i} w_{ij,c}^* q(x_i, y_j)\} = \sum_{i=1}^{n}\{(1 - \delta_i) \sum_{j \in A_R} w_{ij}^* q(x_i, y_j)\}, \quad (14)
\]
for some \( q(x_i, y_j) \), and \( \sum_{j \in D_i} w_{ij,c}^* = 1 \) for all \( i \) with \( \delta_i = 0 \), where \( w_{ij}^* \) is the fractional weights for FEFI method, as defined in (13).

– Regarding the choice of the control function \( q(x, y) \) in (14), we can use \( q(x, y) = (y, y^2)' \), which will lead to fully efficient estimates for the mean and the variance of \( y \).

- For variance estimation, replication method can be used. The imputed values are not changed, only the fractional weights are changed for each replication. (Details skipped)

Reference


Appendix A: Proof of Lemma 5.1

Write \( \eta_i = y_i \) if \( \delta_i = 1 \) and \( \eta_i = y_i^* \) if \( \delta_i = 0 \). Since,

\[
\hat{V}_I = \frac{1}{n(n-1)} \sum_{i=1}^{n} \eta_i^2 - \frac{1}{n-1} \left( \frac{1}{n} \sum_{i=1}^{n} \eta_i \right)^2,
\]

we have

\[
E\{\hat{V}_I\} = \frac{1}{n(n-1)} \sum_{i=1}^{n} E(\eta_i^2) - \frac{1}{n-1} \left\{ \left( \frac{1}{n} \sum_{i=1}^{n} E(\eta_i) \right)^2 + V(\hat{\theta}_I) \right\}
\]

\[
= \frac{1}{n(n-1)} \sum_{i=1}^{n} E(y_i^2) - \frac{1}{n-1} \left\{ \left( \frac{1}{n} \sum_{i=1}^{n} E(y_i) \right)^2 + V(\hat{\theta}_I) \right\}
\]

\[
= \frac{1}{n(n-1)} \sum_{i=1}^{n} E(y_i^2) - \frac{1}{n-1} \left\{ E \left( \frac{1}{n} \sum_{i=1}^{n} y_i \right)^2 + V(\hat{\theta}_I) - V(\hat{\theta}_n) \right\}
\]

\[
= V\{\hat{\theta}_n\} - \frac{1}{n-1} \left\{ V(\hat{\theta}_I) - V(\hat{\theta}_n) \right\}.
\]

Appendix B: A closed form solution to (8)

To find a set of \( \phi_k \) that satisfies (8), note first that we can express the left side of (8) as

\[
\sum_{k \in A_M} \sum_{i \in \mathcal{A}_R} \left( \alpha^{(k)}_i - \alpha_i \right)^2 + \sum_{k \in A_M} \sum_{i \in \mathcal{A}_R} \left( \alpha^{(k)}_i - \alpha_i \right)^2.
\]

For \( k \in A_M \), we have

\[
\alpha^{(k)}_i = \begin{cases} (n-1)^{-1}(1 + M^{-1}d_i) & \text{if } d_{ik} = 0 \\ (n-1)^{-1}(1 + M^{-1}(d_i - d_{ik})) & \text{if } d_{ik} = 1 \end{cases}
\]

and

\[
\sum_{i \in \mathcal{A}_R} \left( \alpha^{(k)}_i - \alpha_i \right)^2 = \sum_{i \in \mathcal{A}_R} \left( \alpha_i \frac{n}{n-1} - \alpha_i - \frac{1}{(n-1)M}d_{ik} \right)^2
\]

\[
= \frac{1}{(n-1)^2} \sum_{i \in \mathcal{A}_R} \left( \alpha_i - \frac{d_{ik}}{M} \right)^2.
\]
Thus, equation (8) reduces to

$$\sum_{k \in A} \sum_{i \in A_R} (\alpha_i^{(k)} - \alpha_i)^2 = \frac{n}{n-1} \sum_{i \in A_R} \alpha_i^2 - \frac{1}{(n-1)^2} \sum_{k \in A_M} \sum_{i \in A_R} \left( \alpha_i - \frac{d_{ik}}{M} \right)^2.$$  \hspace{1cm} (15)

If we define

$$Q_k = \frac{n}{n-1} \alpha_k^2 - \frac{1}{(n-1)^2} \sum_{j \in A_M} \left( \alpha_k - \frac{d_{kj}}{M} \right)^2,$$

a sufficient condition to (15) is to find $\phi_k$ such that

$$\left( \alpha_k^{(k)} - \alpha_k \right)^2 + \sum_{i \in R(k)} \left( \alpha_i^{(k)} - \alpha_i \right)^2 + \sum_{i \notin R(k)} \left( \alpha_i^{(k)} - \alpha_i \right)^2 = Q_k$$ \hspace{1cm} (16)

where $R(k) = \{ i; \sum_{j \in A_M} d_{ij}d_{kj} > 0, i \neq k \}$. Note that

$$\left( \alpha_k^{(k)} - \alpha_k \right)^2 = \left\{ 0 + \frac{1}{n-1} \left( \frac{1}{M} - \phi_k \right) d_k - \frac{1}{n} - \frac{1}{n M} \right\}^2$$

$$= \left( \frac{1}{n-1} \right)^2 \left\{ \frac{n-1}{n} + \phi_k d_k - \frac{1}{n M} \right\}^2$$

where $d_k = \sum_{j \in A_M} d_{kj}$ and, for $i \in R(k)$,

$$\left( \alpha_i^{(k)} - \alpha_i \right)^2 = \left\{ \frac{1}{n-1} + \frac{1}{(n-1)} \left( \frac{d_i}{M} + \frac{\phi_k b_{ik}}{M-1} \right) - \frac{1}{n} - \frac{1}{n M} d_i \right\}^2$$

$$= \left( \frac{1}{n-1} \right)^2 \left\{ \alpha_i + \frac{\phi_k b_{ik}}{M-1} \right\}^2$$

where $b_{ik} = \sum_{j \in A_M} d_{ij}d_{kj}$. For $i \notin R(k)$, we have

$$\left( \alpha_i^{(k)} - \alpha_i \right)^2 = \left( \frac{1}{n-1} \right)^2 \alpha_i^2.$$

Thus, (16) reduces to

$$(1 - \alpha_k + \phi_k d_k)^2 + \sum_{i \in R(k)} \left\{ \alpha_i + \phi_k b_{ik}/(M-1) \right\}^2 + \sum_{i \notin R(k)} \alpha_i^2 = (n-1)^2 Q_k,$$

which is approximately equal to

$$(1 + \phi_k d_k)^2 + \phi_k^2 \frac{\sum_{i \in R(k)} b_{ik}^2}{(M-1)^2} = \left( 1 + \frac{d_k}{M} \right)^2 - \frac{d_k}{M^2}.$$ \hspace{1cm} (17)

for sufficiently large $n$. 

14