Use of hot-deck imputation in estimating the proportions of sexually/physically victimized runaway and homeless adolescents in midwest area

Mingue Park, David Johnson, Kurt Johnson

Korea University and University of Nebraska

The Korean association for Survey Research Meetings
December, 2006

Outline

- Example: Study of runaway and homeless adolescents
- Hot-deck imputation
  - Simple random imputation(SR)
  - Approximate Bayesian Bootstrap imputation (ABB)
  - Fully Efficient Fractional Imputation (FEFI)
- Example - Continued
- Conclusion

Study of runaway and homeless adolescents

- Longitudinal Study of runaway and homeless adolescents age 16-19
- CD, PTSD, MDE, Smoking, Drinking, Drug-use, Physical and sexual victimization, etc
- Study was conducted in 7 Midwestern cities
- Sexual/Physical Victimization coded as 0, 1
- Many observations missing after the initial survey. (Initial and second wave data were used for this study)

Notations and Assumptions

- Population: $U = U_1 \cup U_2 \cup \cdots \cup U_G$
- Sample: $A = A_R \cup A_M$
- n_g: number of sample elements in imputation cell g
- r_g: number of respondents in imputation cell g
- All elements in an imputation cell have the same probability of responding
- The responses are independent
- Unconditional distribution of a dichotomous variable Y

$$Y_i \quad iid \quad \text{Bernoulli}(p)$$

- A full sample estimator of p with an equal probability sampling design

$$\hat{p}_F = n^{-1} \sum n_g \hat{p}_{Fg}, \quad \hat{p}_{Fg} = n_g^{-1} \sum_{A \cap U_g} Y_i$$

Simple random imputation

- Within each imputation cell, draw $(n_g - r_g)$ observations with replacement from r_g respondents.
- Estimator:

$$\hat{p}_I = n^{-1} \sum w_i Y_i^*,$$

$$Y_i^* = \begin{cases} Y_i & \text{, if } i \text{ is not missing} \\ \sum_j Y_j I_{ij} & \text{, if } i \text{ is missing,} \end{cases}$$

$$I_{ij} = \begin{cases} 1 & \text{, if } j\text{-th element is imputed for the missing } i \\ 0 & \text{, elsewhere.} \end{cases}$$

w_i's are sampling weights.

Simple random imputation

- With an equal probability sampling design,

$$E(\hat{p}_I) = p,$$

$$Var(\hat{p}_I) = \sum_{g=1}^{G} \frac{n_g^2}{n^2} \frac{p(1-p)}{r_g} + \sum_{g=1}^{G} \frac{(n_g - r_g)^2}{n^2} \frac{(r_g - 1)p(1-p)}{r_g}$$

$$= V_1 + V_2,$$

$$V_1 = Var\left(\sum_{g=1}^{G} \frac{n_g}{n} \hat{p}_{r_g}\right), \quad V_2 = \text{Variance due to imputation}$$

Simple random imputation

- Variance estimation
  - Multiple imputation

$$\hat{V}_{MI} = \bar{U} + \frac{m+1}{m} B$$

$$\bar{U} = \frac{1}{m} U_i, \quad B = \frac{1}{m-1} \sum_i (\hat{p}_{I(i)} - \bar{p}_I)^2,$$

$$U_i = n^{-1} \hat{p}_{I(i)} (1 - \hat{p}_{I(i)})$$

    - Rubin & Schenker (1996): improper
  - Direct variance estimation [Johnson (2006)]

Approximate Bayesian Bootstrap imputation

[Rubin & Schenker (1986), Rubin (1987)]

- From each imputation cell,
  - First draw r_g observations with replacement from the r_g respondents, say A_{Ig}.
  - Then draw $n_g - r_g$ observations from A_{Ig}.
- Estimator:

$$\hat{p}_I = n^{-1} \sum w_i Y_i^*$$

- Mean and variance of the estimator are the same as the ones for simple random imputation
- Variance estimation: Multiple imputation

Fully Efficient Fractional Imputation

Kalton & Kish (1984), Fay (1996), Kim & Fuller (2004), Fuller & Kim (2005)

- Every responding unit in an imputation cell is used as a donor for every nonrespondent in the cell
- For element $j \in A_M \cap U_g$,

$$w_{ij}^* = \left[\sum_{A_R \cap U_g} w_s \right]^{-1} w_i$$

is the fraction of original weight of element j assigned to the value from donor i,

$$Y_{Ij} = \sum_{i \in A_R} w_{ij}^* Y_i$$

is the weighted mean of imputed values.

Fully Efficient Fractional Imputation

- Illustration: [Fuller & Kim (2005)]

obs	weight	Imputation Cell	x
1	1	1	1
2	1	1	2
3	1	1	3
4	1	1	m
5	1	1	1
6	1	2	2
7	1	2	3
8	1	2	3
9	1	2	2
10	1	2	m

Fully Efficient Fractional Imputation

- Illustration - Imputed data

		cell		replicate weight	
obs	weight	for x	x	1	2
1	1.00	1	1	0.00	1.11
2	1.00	1	2	1.11	0.00
3	1.00	1	3	1.11	1.11
4	**0.50**	1	**1**	0.37	0.74
4	**0.25**	1	**2**	0.37	0.00
4	**0.25**	1	**3**	0.37	0.37
⋮	⋮	⋮	⋮	⋮	⋮

Fully Efficient Fractional Imputation

- Any full sample computer program can be used to compute estimates of function of variables
  - Linear estimator: $\sum_g \sum_{j \in U_g \cap A} \sum_{i \in U_g \cap A_R} w_i w_{ij}^* Y_i$
  - Regression, ANOVA, etc
- Variance estimation - Jackknife method
  - Define k-th replicate weight for i-th unit $w_i^{(k)}$ and $w_{ij}^{*(k)}$
  - Define k-th replicate estimator using k-th replicate weight, say, $\hat{\theta}^{(k)}$
  - Variance estimator: $\hat{V} = \sum_k c_k \left(\hat{\theta}^{(k)} - \bar{\theta}\right)^2$
- FEFI data set is fully efficient for any function of the variable

Fully Efficient Fractional Imputation

Estimating a proportion

- Estimator

$$\hat{p}_I = \sum_g \sum_{j \in U_g \cap A} \sum_{i \in U_g \cap A_R} w_i w_{ij}^* Y_i$$
$$= \sum_g \frac{n_g}{n} \hat{p}_{r_g}, \qquad \text{equal prob. design}$$

- Jackknife replicate weights for imputed values

$$w_{ij}^{*(k)} = \left[\sum_{A_R \cap U_g} w_s^{(k)} \right]^{-1} \sum_{A_R \cap U_g} w_i^{(k)} Y_i, \quad \text{for } 1,$$
$$= \left[\sum_{A_R \cap U_g} w_s^{(k)} \right]^{-1} \sum_{A_R \cap U_g} w_i^{(k)} (1 - Y_i), \quad \text{for } 0$$

Estimation of proportion of sexually/physically victimized adolescents

- Imputation cell and percent of missing

Imputation cell	r_g	n_g	% missing
1 (CD=0, V1=0)	28	42	33.33
2 (CD=0, V1=1)	33	47	29.79
3 (CD=1, V1=0)	37	53	30.19
4 (CD=1, V1=1)	167	227	26.43

- Estimating
  - proportion of sexually/physically victimized adolescent at Wave 2
  - proportion of sexually/physically victimized CD adolescent at Wave 2
  - logistic regression coefficients regressing PV on V1 and CD

Estimation of proportion of sexually/physically victimized adolescents

- Proportion of physically victimized adolescents

	ABB $(m = 5)$	FEFI
Estimator	0.3431	0.3427
Var. Est	0.0008509	0.0008294

- Proportion of sexually victimized adolescents

	ABB $(m = 5)$	FEFI
Estimator	0.1501	0.1607
Var. Est	0.0004320	0.0003679

Estimation of proportion of sexually/physically victimized adolescents

- Proportion of physically victimized CD adolescents

	ABB ($m = 5$)	FEFI
Estimator	0.2867	0.2844
Var. Est	0.0008551	0.0007218

- Proportion of sexually victimized CD adolescents

	ABB ($m = 5$)	FEFI
Estimator	0.1268	0.1373
Var. Est	0.0003239	0.0003277

- Logistic regression coefficient estimate

Explanatory variable	ABB ($m = 5$)		FEFI	
	Estimate	Var. Est.	Estimate	Var. Est.
CD	-0.2848	0.1093	-0.2237	0.0857
V1	-2.0006	0.2292	-1.9537	0.2157

Conclusion

- Two hot-deck imputations were applied to estimate the proportion of sexually/physically victimized adolescent
- Multiple imputation with ABB is easy to implement
- FEFI is fully efficient
- For a dichotomous variable, defining the FEFI data set and jackknife replicate weight for the variance estimation are simple
- FEFI weights can be used directly for the analysis of statistical model