Regression Fractional Hot deck Imputation

Jae Kwang Kim

February 16, 2006

Background : Small area income & Poverty estimates

- State-level estimates
  - Poverty estimates : Children under age 5 in poverty, Related children ages 5-17 in families in poverty, Children under age 18 in poverty, All people in poverty
  - Income estimates : Median household income
- County-level estimates : Same as State-level estimates except for "Children under age 5 in poverty"
- School District estimates : Compute only "Related children ages 5-17 in families in poverty"

Background : Task

- The above estimates are computed from the Census long form data using age, relation, income items.

- Parameters of interest : Grand mean, domain mean, proportion, population median

- There are eight income items and each item has non-negligible imputation rate with different missing pattern

- The main task is to compute the variances of the estimates incorporating the imputation information.

Outline

- Survey Sampling Setup
- Fractional Hot deck Imputation
- Extension to the regression imputation model
- Monte Carlo Study
- Conclusion

Survey Sampling

- Data : Generated (in part) by probability rules set by the survey sampler

- Product : Data set that "represents" population sampled. Data to be analyzed by others. "y" may be unknown.

Example 1: Simple Random Sample

Element	Sample Weight	Auxiliary Variable (House Rent)	Y (Income)
1	0.1	1	1
2	0.1	1	2
3	0.1	1	3
4	0.1	1	4
5	0.1	1	5
6	0.1	1	?
7	0.1	0	3
8	0.1	0	6
9	0.1	0	9
10	0.1	0	?

Assumptions

- Population model approach : modelling on Y_i

- Response model approach : modelling on R_i

$$R_i = \begin{cases} 1 & \text{if } Y_i \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

Universe of G cells

- Cell mean model

$$\begin{aligned} Y_{gi} &= \mu_g + e_{ig} \\ e_{gi} &\sim II\left(0, \sigma_g^2\right) \end{aligned}$$

- Cell response model

$$\begin{aligned} Pr\left(R_{gi} = 1\right) &= \pi_g \\ Pr\left(R_{gi} = 1, R_{sj} = 1\right) &= \pi_g \pi_s \end{aligned}$$

Weighted Data Set

Original Weight	New Weight	Auxiliary Variable	Y
0.1	0.120	1	1
0.1	0.120	1	2
0.1	0.120	1	3
0.1	0.120	1	4
0.1	0.120	1	5
0.1	0	1	?
0.1	0.133	0	3
0.1	0.134	0	6
0.1	0.133	0	9
0.1	0	0	?

$$\widehat{\mu}_Y = 0.6 \times 3.0 + 0.4 \times 6.0 = 4.2$$

Weighted Data Set

Original Weight	New Weight	Aux. Var.	Y	Z (County)
0.1	0.120	1	1	0
0.1	0.120	1	2	1
0.1	0.120	1	3	1
0.1	0.120	1	4	0
0.1	0.120	1	5	1
0.1	0	1	?	0
0.1	0.133	0	3	1
0.1	0.134	0	6	0
0.1	0.133	0	9	0
0.1	0	0	?	1

$$\widehat{P}\{Z = 1\} = 0.493 \quad \text{vs} \quad 0.500$$

Additional Assumption

- y_{gj} is independent of Z_{gj} given cell

Mean Imputed Data Set

Original Weight	Aux. Var.	Y	Z
0.1	1	1	0
0.1	1	2	1
0.1	1	3	1
0.1	1	4	0
0.1	1	5	1
0.1	1	3*	0
0.1	0	3	1
0.1	0	6	0
0.1	0	9	0
0.1	0	6*	1

$\widehat{\mu}_Y = 4.2,\ \widehat{P}(Z = 1) = 0.5,\ \widehat{\mu}_{Y|Z=1} = 3.8$

Mean Imputed Data Set

Original Weight	Aux. Var.	Y	Z
0.1	1	1	0
0.1	1	2	1
0.1	1	3	1
0.1	1	4	0
0.1	1	5	1
0.1	1	3*	0
0.1	0	3	1
0.1	0	6	0
0.1	0	9	0
0.1	0	6*	1

Weighted $\quad \hat{P}\{2 < Y < 7\} = 0.6267$

Mean Imputed $\quad \hat{P}\{2 < Y < 7\} = 0.7000$

Random Hot Deck Imputation

Element	Original Weight	Aux. Var.	Y	Z
1	0.1	1	1	0
2	0.1	1	2	1
3	0.1	1	3	1
4	0.1	1	4	0
5	0.1	1	5	1
6	0.1	1	4*	0
7	0.1	0	3	1
8	0.1	0	6	0
9	0.1	0	9	0
10	0.1	0	9*	1

$$\widehat{\mu}_Y = 4.6, \qquad \widehat{\mu}_{Y|Z=1} = 4.4, \qquad \widehat{P}\left(2 < Y < 7\right) = 0.6$$

$$E_I\left(\widehat{\mu}_Y\right) = 4.2, E_I\left\{\widehat{\mu}_{Y|Z=1}\right\} = 3.8, E_I\left\{\widehat{P}\left(2 < Y < 7\right)\right\} = 0.6267$$

Fully Efficient Fractional Imputation

Unit	Weight	X	Y	Z
1	0.100	1	1	0
2	0.100	1	2	1
3	0.100	1	3	1
4	0.100	1	4	0
5	0.100	1	5	1
6	0.020	1	1^*	1
	0.020	1	2^*	1
	0.020	1	3^*	1
	0.020	1	4^*	1
	0.020	1	5^*	1
7	0.100	0	3	1
8	0.100	0	6	0
9	0.100	0	9	0
10	0.033	0	3^*	1
	0.034	0	6^*	1
	0.033	0	9^*	1

Properties of FEFI

1. For the estimation of the grand mean, it is algebraically equivalent to imputing the cell mean.

2. Unlike the cell mean imputation, it provide unbiased estimates for the proportions.

3. For domain estimation, it often borrows strength.

Approximations

1. Fixed, small number of fractions

2. Efficient sampling

3. Regression for FE of some y

Fractional Imputation (M=3)

Unit	Weight	X	Y	Z
1	0.100	1	1	0
2	0.100	1	2	1
3	0.100	1	3	1
4	0.100	1	4	0
5	0.100	1	5	1
6	0.033	1	1*	1
	0.034	1	3*	1
	0.033	1	5*	1
7	0.100	0	3	1
8	0.100	0	6	0
9	0.100	0	9	0
10	0.033	0	3*	1
	0.034	0	6*	1
	0.033	0	9*	1

Variance estimation

- For simplicity, suppose that

$$\widehat{\theta}_n = (y_1 + y_2 + y_3 + y_4 + y_5)/5$$

- If y_2 is missing and imputed by $y_2^* = y_4$, then the imputed estimator will be

$$\widehat{\theta}_I = (y_1 + y_2^* + y_3 + y_4 + y_5)/5$$

- The variance is increased: Under the IID model,

$$V\left(\widehat{\theta}_I\right) = V\left(\widehat{\theta}_n\right) + 2Cov\left(y_2^*, y_4\right)/25$$

- Naive variance estimator (treating the imputed values as if observed) will estimate the variance of $\widehat{\theta}_n$, not the variance of $\widehat{\theta}_I$.

Variance Estimation for complete sample

- Replication variance estimator for complete sample

$$\widehat{V}_n = \sum_{k=1}^{L} c_k \left(\widehat{\theta}_n^{(k)} - \widehat{\theta}_n \right)^2$$

  where L is the number of replication, c_k is replication factor associated with k-th replication, $\widehat{\theta}_n^{(k)}$ is the k-th replicate of $\widehat{\theta}_n$.

- If $\widehat{\theta}_n = \sum_{i \in A} w_i y_i$, then $\widehat{\theta}_n^{(k)} = \sum_{i \in A} w_i^{(k)} y_i$.

- Useful for several θ's.

Example: Jackknife

Unit	Weight	X	Y	$w_i^{(1)}$	$w_i^{(2)}$	$\cdots$	$w_i^{(10)}$
1	0.1	1	1	0	0.111		0.111
2	0.1	1	2	0.111	0		0.111
3	0.1	1	3	0.111	0.111		0.111
4	0.1	1	4	0.111	0.111		0.111
5	0.1	1	5	0.111	0.111		0.111
6	0.1	1	?	0.111	0.111		0.111
7	0.1	0	3	0.111	0.111		0.111
8	0.1	0	6	0.111	0.111		0.111
9	0.1	0	9	0.111	0.111		0.111
10	0.1	0	?	0.111	0.111		0

Jackknife variance estimation

- Jackknife replicate

$$\begin{aligned}\widehat{\theta}_n^{(k)} &= (n-1)^{-1}\sum_{i\neq k}^{n} y_i = (n-1)^{-1}\left(n\bar{y}_n - y_k\right)\\ \widehat{\theta}_n^{(k)} - \widehat{\theta}_n &= (n-1)^{-1}\left(\bar{y}_n - y_k\right)\end{aligned}$$

- Jackknife variance estimator

$$\begin{aligned}\widehat{V}_n &= \frac{n-1}{n}\sum_{k=1}^{n}\left(\widehat{\theta}_n^{(k)} - \widehat{\theta}_n\right)^2\\ &= \frac{1}{n\,(n-1)}\sum_{k=1}^{n}\left(y_k - \bar{y}_n\right)^2 = \frac{1}{n}s_n^2\end{aligned}$$

Jackknife for FI

Unit	Weight	Y	Z	Rep. 1	Rep. 2	Rep. 3	Rep. 4	Rep. 5
1	0.100	1	0	0	0.111	0.111	0.111	0.111
2	0.100	2	1	0.111	0	0.111	0.111	0.111
3	0.100	3	1	0.111	0.111	0	0.111	0.111
4	0.100	4	0	0.111	0.111	0.111	0	0.111
5	0.100	5	1	0.111	0.111	0.111	0.111	0
6	0.033	1^*	1	$0.033 - \delta_1$	0.033	$0.033 + \delta_3/2$	0.033	$0.033 + \delta_5/2$
	0.034	3^*	1	$0.034 + \delta_1/2$	0.034	$0.034 - \delta_3$	0.034	$0.034 + \delta_5/2$
	0.033	5^*	1	$0.033 + \delta_1/2$	0.033	$0.033 + \delta_3/2$	0.033	$0.033 - \delta_5$
7	0.100	3	1	0.111	0.111	0.111	0.111	0.111
8	0.100	6	0	0.111	0.111	0.111	0.111	0.111
9	0.100	9	0	0.111	0.111	0.111	0.111	0.111
10	0.033	3^*	1	0.033	0.033	0.033	0.033	0.033
	0.034	6^*	1	0.034	0.034	0.034	0.034	0.034
	0.033	9^*	1	0.033	0.033	0.033	0.033	0.033

Jackknife for FI

Unit	Weight	Y	Z	Rep. 6	Rep. 7	Rep. 8	Rep. 9	Rep. 10
1	0.100	1	0	0.111	0.111	0.111	0.111	0.111
2	0.100	2	1	0.111	0.111	0.111	0.111	0.111
3	0.100	3	1	0.111	0.111	0.111	0.111	0.111
4	0.100	4	0	0.111	0.111	0.111	0.111	0.111
5	0.100	5	1	0.111	0.111	0.111	0.111	0.111
6	0.033	1^*	1	0	0.033	0.033	0.033	0.033
	0.034	3^*	1	0	0.034	0.034	0.034	0.034
	0.033	5^*	1	0	0.033	0.033	0.033	0.033
7	0.100	3	1	0.111	0	0.111	0.111	0.111
8	0.100	6	0	0.111	0.111	0	0.111	0.111
9	0.100	9	0	0.111	0.111	0.111	0	0.111
10	0.033	3^*	1	0.033	$0.033 - \delta_7$	$0.034 + \delta_8/2$	$0.034 + \delta_9/2$	0
	0.034	6^*	1	0.034	$0.034 + \delta_7/2$	$0.034 - \delta_8$	$0.034 + \delta_9/2$	0
	0.033	9^*	1	0.033	$0.033 + \delta_7/2$	$0.034 + \delta_8/2$	$0.034 - \delta_9$	0

Jackknife replicates for FI

- If applied to y variable,

$$\widehat{\theta}_I^{(k)} - \widehat{\theta}_I = \widehat{\theta}_{I,naive}^{(k)} - \widehat{\theta}_I - \delta_k \left(y_k^* - \bar{y}_{Ij}^{*(k)} \right)$$

  Thus, increasing δ_k will increase the expected value of $\left(\widehat{\theta}_I^{(k)} - \widehat{\theta}_I \right)^2$.

- If applied to z-variable, $\widehat{\theta}_I^{(k)}$ does not depend on δ_k's.

- Kim and Fuller (2004) propose a method for computing δ_k for unbiased variance estimation.

Properties - variance estimation for FI

If full sample estimator of variance is consistent, the jackknife variance estimator is

1. consistent for variance of imputed y

2. $\widehat{V}$ gives design variance for full sample variables

3. consistent for variance of subpopulations.

Extension to the regression model

Example 2: Simple Random Sample

Element	Sample Weight	Auxiliary Variable (Value of house)	Y (Income)
1	0.1	3	1
2	0.1	4	2
3	0.1	4	3
4	0.1	5	4
5	0.1	6	5
6	0.1	7	?
7	0.1	10	3
8	0.1	12	6
9	0.1	15	9
10	0.1	16	?

Basic Setup

- y_j^* : imputed value for y_j, $\quad j \in A_M = A \cap A_R^c$.

- Often, $y_j^* = \widehat{y}_j + \widehat{e}_i$, where
  - $\widehat{y}_j = \widehat{\beta}_0 + \widehat{\beta}_1 x_j$
  - $\widehat{e}_i$ is selected from $\widehat{\mathbf{e}} = \{\widehat{e}_i \ ; i \in A_R\}$

- Usually called (stochastic) regression imputation.

Basic Setup - continued 2

- Reasons for the regression imputation
  - preserve the correlation structure (unlike hot deck imputation ignoring x).
  - unbiased (for the marginal mean of y and for the regression coefficient) if $E(y_i \mid x_i) = \beta_0 + \beta_1 x_i$.
  - efficient if $V(y_i \mid x_i)$ is small.

Implementation using fractional imputation

- Start with a simple model : $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$.

- Compute the residuals $\widehat{e}_i = y_i - \widehat{y}_i$.

- The imputed values for a missing unit j are $\widehat{y}_j + \widehat{e}_j^*$, where $\widehat{e}_j^*$ are taken from the respondents in the same cell.

Fractional regression imputation

Unit	Weight	X	Y
1	0.100	x_1	y_1
2	0.100	x_2	y_2
3	0.100	x_3	y_3
4	0.100	x_4	y_4
5	0.100	x_5	y_5
6	0.033	x_6	$\widehat{y}_6 + \widehat{e}_1$
	0.034	x_6	$\widehat{y}_6 + \widehat{e}_3$
	0.033	x_6	$\widehat{y}_6 + \widehat{e}_5$
7	0.100	x_7	y_7
8	0.100	x_8	y_8
9	0.100	x_9	y_9
10	0.033	x_{10}	$\widehat{y}_{10} + \widehat{e}_7$
	0.034	x_{10}	$\widehat{y}_{10} + \widehat{e}_8$
	0.033	x_{10}	$\widehat{y}_{10} + \widehat{e}_9$

Fractional regression hot deck imputation

Unit	Weight	X	Y
1	0.100	x_1	y_1
2	0.100	x_2	y_2
3	0.100	x_3	y_3
4	0.100	x_4	y_4
5	0.100	x_5	y_5
6	$0.1 \times w^*_{16}$	x_6	y_1
	$0.1 \times w^*_{36}$	x_6	y_3
	$0.1 \times w^*_{56}$	x_6	y_5
7	0.100	x_7	y_7
8	0.100	x_8	y_8
9	0.100	x_9	y_9
10	$0.1 \times w^*_{7,10}$	x_{10}	y_7
	$0.1 \times w^*_{8,10}$	x_{10}	y_8
	$0.1 \times w^*_{9,10}$	x_{10}	y_9

where w^*_{ij} satisfy $\sum_{i \in A_R} w^*_{ij} = 1$ and $\sum_{i \in A_R} w^*_{ij} x_i = x_j$.

Construction of regression fractional weights

- Fractional weight w_{ij}^* : the factor applied to the original weight for element j when y_i is used as a donor for element j.

- If w_{ij}^* satisfy

$$\sum_{i \in A_R} w_{ij}^* \left(1, x_i\right) = \left(1, x_j\right), \qquad (*)$$

where A_R be the set of indices for the respondents in y, then

$$\sum_{i \in A_R} w_{ij}^* \left(\widehat{y}_j + \widehat{e}_i\right) = \sum_{i \in A_R} w_{ij}^* y_i.$$

Thus, the regression fractional imputation can be written as a fractional hot deck imputation.

- Regression weighting method can be used to construct a fractional weights satisfying the constraints:

$$w_{ij}^{*} = w_{ij0}^{*} + \left(x_j - \bar{x}_{Ij0}\right) S_{xx,j}^{-1} w_{ij0}^{*} \left(x_i - \bar{x}_{Ij0}\right)$$

  where w_{ij0}^{*} is the initial fractional weight ($= 1/M$) and

$$\begin{aligned} S_{xx,j} &= \sum_{i \in A_R} w_{ij0}^{*} (x_i - \bar{x}_{Ij})^2, \\ \bar{x}_{Ij0} &= \sum_{i \in A_R} w_{ij0}^{*} x_i. \end{aligned}$$

- Variance estimation can be performed similarly by creating the replicated fractional weights $w_{ij}^{*(k)}$ satisfying (*).

Robustness

- If the complete sample estimator is of the form $\widehat{\theta}_n = \sum_{i \in A} w_i y_i$, then the FEFI estimator can be written

$$\widehat{\theta}_{FI} = \sum_{g=1}^{G} \sum_{i \in A_g} w_i \left(\widehat{y}_i + \widehat{e}_{Ij} \right)$$

  where $\bar{e}_{Ij} = \sum_{i \in A_{Rg}} w_{ij}^* \widehat{e}_i$.

- "Imputed value" = "Predictive value " + "Local mean of the residual"

- If the cells are well chosen, the "local mean of the residual" can make up for the model bias.

Multiple Imputation

Unit	Weight	Aux. Var.	Imp. 1	Imp. 2	Imp. 3
1	0.1	1	1	1	1
2	0.1	1	2	2	2
3	0.1	1	3	3	3
4	0.1	1	4	4	4
5	0.1	1	5	5	5
6	0.1	1	4*	2*	4*
7	0.1	0	3	3	3
8	0.1	0	6	6	6
9	0.1	0	9	9	9
10	0.1	0	3*	6*	6*

Multiple Imputation

- $\left(\widehat{\theta}_{I,t}, \widehat{V}_{I,t}\right)$: (Point Estimator, Variance estimator) using imputed data set t

- Point estimator : $\widehat{\theta}_{MI} = M^{-1} \sum_{t=1}^{M} \widehat{\theta}_{I,t}$

- Variance estimator : $\widehat{V}\left(\widehat{\theta}_{MI}\right) = W_M + \left(1 + M^{-1}\right) B_M,$

  where $W_M = M^{-1} \sum_{t=1}^{M} \widehat{V}_{I,t}$ and $B_M = (M-1)^{-1} \sum_{t=1}^{M} \left(\widehat{\theta}_{I,t} - \widehat{\theta}_M\right)^2.$

Multiple Imputation under the normal linear regression model

1. Generate parameters

$$\begin{aligned} \sigma^{*2} &\sim S_{xx}/\chi^2_{r-1} \\ \beta_1^* &\sim N\left(\hat{b}_1, \sigma^{*2}/S_{xx}\right) \\ \beta_0^* &\sim N\left(\bar{y}_r + \beta_1^* \bar{x}, \sigma^{*2}/r\right) \end{aligned}$$

   where $S_{xx} = \sum_{i \in A_R} (x_i - \bar{x}_r)^2$.

2. Generate sampling errors : $e_j^* \sim N\left(0, \sigma^{*2}\right)$

3. Imputed values are

$$y_j^* = \beta_0^* + \beta_1^* x_j + e_j^*.$$

Monte Carlo Study

- Sampling distribution:

  Linear : $y_i = 2 + x_i + e_i$

  Quadratic : $y_i = 2 + \sqrt{0.5}\left(x_i^2 - 1\right) + e_i$

  where $x_i \sim N(0,1)$ and $e_i \sim N(0,1)$.

- In addition to (x_i, y_i), $z_i \sim U(0,1)$ are generated. Independently of y_i.

- Uniform response mechanism with 65 % response rate.

- $B = 5000$ Monte Carlo Samples of size $n = 100$.

- Imputation Method

  1. Fractional Regression Hot deck imputation ($M = 5$) using a linear regression model. The cells are formed by using similar x-values.

  2. Multiple imputation ($M = 5$) under the normal linear regression model

- Parameters of Interest

$$\begin{aligned} \theta_1 &= \text{mean of y} \\ \theta_2 &= \text{mean of y among the units with z less 0.25} \\ \theta_3 &= \text{proportion of y less than 1.0} \\ \theta_4 &= \text{slope of y on x} \end{aligned}$$

Mean and variance of the point estimators under the linear model

Parameter	Method	Mean	Variance
Mean	Complete sample	2.00	0.020
	FI	2.00	0.028
	MI	2.00	0.026
Domain	Complete sample	2.00	0.082
	FI	2.00	0.082
	MI	2.00	0.078
Mean	Complete sample	0.24	0.0018
	FI	0.24	0.0027
	MI	0.24	0.0020
Mean	Complete sample	1.00	0.0103
	FI	1.00	0.0220
	MI	1.00	0.0176

Relative mean and t-statistics for the variance estimators under linear model

Parameter	Method	Rel. Mean	t-statistic*
Mean	FI	100.4	0.11
	MI	104.4	2.20
Domain	FI	108.1	3.93
	MI	134.9	16.62
Proportion	FI	102.4	1.19
	MI	122.1	11.08
Slope	FI	104.8	2.00
	MI	101.1	0.53

Mean and variance of the point estimators under the quadratic model

Parameter	Method	Mean	Variance
Mean	Complete sample	2.00	0.020
	FI	2.00	0.027
	MI	2.00	0.033
Domain	Complete sample	2.00	0.080
	FI	2.00	0.080
	MI	2.00	0.078
Mean	Complete sample	0.23	0.0018
	FI	0.23	0.0028
	MI	0.23	0.0022
Mean	Complete sample	0.00	0.0574
	FI	0.00	0.0714
	MI	0.00	0.0903

Relative mean and t-statistics for the variance estimators under quadratic model

Parameter	Method	Rel. Mean	t-statistic*
Mean	FI	101.8	0.90
	MI	103.3	-0.13
Domain	FI	111.3	5.32
	MI	161.4	27.45
Proportion	FI	101.0	0.49
	MI	127.3	13.28
Slope	FI	98.3	-0.80
	MI	63.0	-18.87

Discussion : Fractional Imputation

- More efficient than single imputation
- Unbiased variance estimation possible
- Can handle auxiliary variables
- Can be made robust against the failure of the imputation model