Presentation (WP43)

UNECE Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

ON IMPUTATION OF BINARY VARIABLES IN REGISTERS

by

Thomas Laitila

Statistics Sweden and Department of Statistics, Örebro university

(thomas.laitila@scb.se)

Question

- Random imputation suggested for imputation in registers
- Imputation techniques, like random imputation and the MI technique by Rubin (1987), are developed in the context of sample surveys.
- What are their properties in Register surveys?

Estimators

- A population U of N units, real valued variables y_k and x_k
- Estimation of $t_{yx} = \sum_U y_k x_k$ using register information
- Random imputation estimator

$$\hat{t}_{yx} = \sum_{U_R} y_k x_k + \sum_{U_y} \hat{y}_k(x_k) x_k + \sum_{U_x} y_k \hat{x}_k(y_k) + \sum_{U_{yx}} \hat{y}_k \hat{x}_k(\hat{y}_k)$$

- Deterministic imputation estimator

$$\hat{t}^D_{yx} = \sum_{U_R} y_k x_k + \sum_{U_y} \mu_{yk}(x_k) x_k + \sum_{U_x} y_k \mu_{xk}(y_k) + \sum_{U_{yx}} \lambda_k$$

Results

- Random imputation yields estimators with high relative precision (in terms of cv) if the population is large (law of large numbers).
- The loss in efficiency can be substantial by using random instead of deterministic imputation
- The loss in efficiency is particularly pronounced if the imputation method yields unbiased or nearly unbiased estimates.
- Random imputation do not provide with information on estimator uncertainty

Efficiency

- We note that

$$\hat{t}_{yx}^{D} = E\left(\hat{t}_{yx}\right)$$

so

$$Bias(\hat{t}_{yx}) = Bias(\hat{t}_{yx}^{D})$$

Relative efficiency, measured in MSE terms, (generic)

$$\text{Reff}\left(\hat{t}_{yx};\hat{t}_{yx}^{D}\right)=\frac{Bias(\hat{t}_{yx})^2}{V(\hat{t}_{yx})+Bias(\hat{t}_{yx})^2}$$

Let

$$\beta=\frac{Bias(\hat{t}_{yx})}{t_{yx}}$$

then

$$\text{Reff}\left(\hat{t}_{yx};\hat{t}_{yx}^{D}\right)=\frac{\left(\beta/(1+\beta)\right)^2}{cv(\hat{t}_{yx})^2+\left(\beta/(1+\beta)\right)^2}$$

Table 1: Efficiency of $\hat{t}_{yx}$ compared with $\hat{t}_{yx}^{D}$ for different levels of relative bias and coefficient of variation.

	Coefficient of Variation			
Relative bias	0.01	0.05	0.1	0.2
-0.2	0.998	0.962	0.862	0.610
-0.1	0.992	0.832	0.552	0.236
-0.05	0.965	0.526	0.217	0.065
-0.01	0.505	0.039	0.010	0.003
0.01	0.495	0.038	0.010	0.002
0.05	0.958	0.476	0.185	0.054
0.1	0.988	0.768	0.452	0.171
0.2	0.996	0.917	0.735	0.410

Confidence intervals

The interval

$$\hat{t}_{yx} \pm 1.96 \cdot \hat{V}(\hat{t}_{yx})^{1/2}$$

gives an approximate 95% CI for $\hat{t}_{yx}^{D}$, not for t_{yx}!

An Illustration

The income register lacks information on Swedish citizens' earnings in Norway.

Information is available after publishing of income statistics.

Can we impute predicted values? (Snölilja, 2010)

- Prediction of work in Norway for people registered in three Swedish municipalities on the border.
- y_k = 1 if unit *k* have earnings in Norway, y_k = 0 otherwise ($x_k = 1$)
- Logit model derived from income statistics for 2006
- Predictions made for 2007

- Estimators:
  - $\hat{t}_y^D = \sum_{U_R} y_k + \sum_{U_y} \mu_k$
  - $\hat{t}_y = \sum_{U_R} y_k + \sum_{U_y} \hat{y}_k \ , \qquad \hat{y}_k \sim \text{bern}(\mu_k)$
- Register total in 2007
  - t_y = 2324
  - N = 17457 (Note $U_y = U$ and $U_R = \phi$)

Number of persons with income from Norway 2007, estimated and recorded values.

Estimator	Estimate	Rel. Bias	MSE
$\hat{t}_y^D$	2110	-10%	53824
$\hat{t}_y$ [a)]	2075 [a)]	-10%	54421 [b)]
t_y	2342 [c)]	---	---

[a)] One realization of the random imputation estimator.
[b)] Bias and MSE for the estimator $\hat{t}_y$.

The variance of the estimator $\hat{t}_y$ is

$$V(\hat{t}_y) = \sum_{U_y} \mu_k (1 - \mu_k) = 596.9$$

A 95% CI based on $\hat{t}_y$ yields the interval

$$2075 \pm 47.9$$

This interval covers $\hat{t}_y^D$ = 2110, but not the population value t_y = 2342.

Bias of interest, not variance!

THANKS FOR YOUR ATTENTION!