

Imputation in EU-SILC survey

Andris Fisenko and Vita Kozirkova
Central Statistical Bureau of Latvia

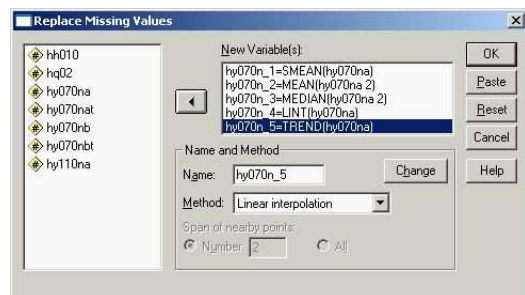
Structure

- Households level
 - Income per year is asked.
 - If refusal, then ask average and how often.
 - If answer is “ I don’t know” then IMPUTATION
- Personal level
 - Differences for persons level
 - Average in one time or interval.

Imputation methods

- RMV (Replace Missing Values)
- Hot deck & Cold deck
- Mean/Median imputation
- Nearest Neighbour
- Regression
- Other

SPSS RMV- 1



RMV - results

		Statistics					
		SMEAN(HY070NA)	MEAN(HY070NA)	MEDIAN(HY070NA,2)	LINT(HY070NA)	TREND(HY070NA)	
N	Valid	53	58	58	58	58	58
	Missing	5	0	0	0	0	0
Mean		\$35.66	\$35.66	\$35.30	\$35.22	\$35.25	\$35.65
Std. Error of Mean		\$3.38	\$3.09	\$3.13	\$3.13	\$3.14	\$3.09
Median		\$30.00	\$32.50	\$30.00	\$30.00	\$30.00	\$30.00
Std. Deviation		\$24.63	\$23.52	\$23.82	\$23.83	\$23.92	\$23.56
Variance		\$606.62	\$553.41	\$567.61	\$567.87	\$572.28	\$555.16
Sum		\$1,889.92	\$2,068.21	\$2,047.17	\$2,042.92	\$2,044.42	\$2,067.51

SMEAN (HY070NA)	MEAN (HY070NA,2)	MEDIAN (HY070NA,2)	LINT (HY070NA)	TREND (HY070NA)
\$35.66	\$39.50	\$42.50	\$28.00	\$28.83
\$35.66	\$10.25	\$9.00	\$9.00	\$32.18
\$35.66	\$33.00	\$35.00	\$35.00	\$36.29
\$35.66	\$28.75	\$27.50	\$30.00	\$39.63
\$35.66	\$45.75	\$39.00	\$52.50	\$40.66
\$35.66	\$31.45	\$30.60	\$30.90	\$35.52

Cold-deck

- Does not bring new information
- Very good benchmarking method

- There is not data from the past
- Method will be analysed in the future

Random donor (Hot-deck)

- The donor is chosen randomly within same group
- Random overall imputation is not sensible
- But random donor within imputation classes (or clusters, groups, ..) is important method

Random donor (Hot-deck)

- Firstly, the data is divided into homogenous sub groups by a suitable method, e.g. easily sorting into categories by some explanatory variable. Secondly, the donors are chosen randomly within these new groups.
- Many clustering methods are using the idea of nearest neighbours when determining a new grouping of the data.

Multiple imputation

- Creates m imputed data sets.
- Hot deck+Multiple imputation

Some results

Number of simulations	Percentage of missing data	Mean	Standar Deviation
0	0	1673.3	1511.5
20	1	1670.8	1511.2
100	1.5	1672.9	1515.7
50	2	1670.7	1508.7
21	3	1677.6	1529.4
100	4	1676.3	1529.1
50	5	1670.7	1521.9
100	5	1670.1	1531.4
50	7	1682.1	1534.4
20	10	1659.4	1499.8
20	15	1657.5	1457.4
50	20	1693.4	1567.8

Some results

Percentage of missing data	Value
1%	0.1
1.5%	0.2
2%	0.3
3%	1.4
4%	1.3
5%	0.8
5%	1.4
7%	2.0
10%	1.5
15%	4.5
20%	5.0

What's next

- Software
- Linear regression
- Auxiliary information from
 - State Revenue Service
 - State Social Insurance Agency
 - Regional governments
 - Banks

