Real donor imputation

Selecting the (number of) potential donors

Nicklas Pettersson

Workshop in Valmiera (Latvia)

2012-08-24

Stockholm University

Outline

- Missing data
- Real donor (multiple) imputation
- Bias reduction features
- Selecting the real donor pool
- Simulations

Missing data

- Missing data is typically associated with nonresponse in surveys. (But in general, all data that is not known is missing.)
- There is no non-treatment solution to missing data, one always has to assume something about the missingness mechanism.
- Standard treatment involves weighting for unit nonresponse and imputation for item nonresponse.
- Typically we assume a missing at random (MAR) missing data mechanism, i.e. missingness depends on the observed data.

Imputation

(Little and Rubin, 2002) *"Imputations should generally be:*

(a) *Conditional on observed variables, to reduce bias due to nonresponse, improve precision, and preserve association between missing and observed variables;*

(b) *Multivariate, to preserve associations between missing variables;*

(c) *Draws from predictive distribution rather than means, to provide valid estimates of a wide range of estimands."*

- *The most important factor in imputation is access to variables (usually auxiliaries from registers and observed survey data) which are predictive of the missing values and the nonresponse propensity. These are used to model the data (including the missingness mechanism), i.e. (a).*
- *(b) is not relevant to this study (only univarate auxiliaries or study variables).*
- *(c) is achieved by using real donor multiple imputation via the (finite population) Bayesian bootstrap (Rubin, 1981; Lo, 1989).*

Real donor imputation

Categories of imputation, my bold selections;

"'0. Use of complete cases, when any missing items have not been imputed.
1. Deductive or logical imputation; there is a known function (identity equation) between certain observed values and missing values.
2. Imputed values are derived from a (behavioural) model, that is, imputed values may be non-observable in real life world. I call this the model-donor imputation methods family.
3. Imputed values are derived from a set of observed values, from a real donor respondent. This is called real-donor imputation. "' (Laaksonen, 2000)

Real donor imputation

A simple example;

i	X_i	$\lvert X_i - X_8 \rvert$	Y_i	R_i	
1	18	69	9	0	
2	23	57	7	0	
3	29	51	4	0	
4	40	40	1	0	
5	51	29	2	0	
6	55	25	4	0	$\leftarrow$ potential donor
7	69	11	6	0	$\leftarrow$ potential donor
8	80	*	0	1	$\leftarrow$ donee
9	89	10	8	0	$\leftarrow$ potential donor
..	..	..	..	..	
n	..	..	..	..	

X; Study variable

Y; Auxiliary variable

R; Missing data indicator

MAR missingness mechanism; $P(R|X, Y) = P(R|X)$

Real donor imputation

cont. citation from (Laaksonen, 2000), my bold selections;

"'The distinction between methods 2 and 3 is useful for better understanding the nature of imputations, since the latter one **always gives natural, possible values**, *whereas the former may provide impossible values as well. This feature is* **not always an advantage"'... "' if the observed values do not cover all potential values exhaustively**. *Real-donor imputation is* **impossible to apply if there are no respondents within some area**. *It is as well* **problematic to use with a low number of respondents**. *In such cases a modelling technique may be more helpful, provided that the model is estimable and predictable enough."'*

Real donor pool size

Some strategies for selecting the donor pool;

1 Always use the same number of donors for each donee.
   1. Automatically ensures that no donee ger zero donors.
   2. k-nearest neighbour (kNN)

2 Use donors that lies within a distance ϵ to the donee.
   1. Different number of donors for different donees.
   2. Parallell to a fixed kernel bandwidth, e.g. a 'rule-of-thumb' bandwidth (Silverman, 1986).

3 Other strategies, e.g. locally adapting the fixed version by increasing (decreasing) the maximum allowed distance if relatively few (many) donors are close to the donee.

Drawing real donors

1. Select a distance measure and rank the potential donors
2. Define the donor pool. In the example we took units with max distance $\epsilon = 25 \rightarrow$ k=3 potential donors
3. Decide how to allocate the donor selection probabilities, e.g. equal probablities 1/3
4. Draw a donor at random using these probabilities, and impute the value on the donee

Given (c) we should draw imputations from predictive distributions and do it multiple, say M, times. With small (large) donor pools, the M imputed values imputed on a single unit will be more (less) correlated since a single donor is likely to be drawn several (few) times. E.g. compare with sampling without accounting for clustering in the data, which can results in highly variable final estimates. On the other, larger donor pools instead reduces the quality of matches and may increase the bias of final estimates. We need to balance this trade-off when selecting the number of donors.

Bias reduction

The expected imputed auxilary value (in our example 1/3*(55+69+89)=71) is (almost) never equal the donee auxiliary value (80), thus the donor pool is imperfectly matched (80-71=9) to the donee. Matching is expected to be worst when the donee lies at the boundary with few (or none) potential donors that lies on one side of the donee. These "'individual matching biases"' may lead to bias of the final estimate. Therefore, try to reduce them by introducing the following features into the imputation process (Pettersson, 2012);

1. Since the closest donors provides a better match to the donee, assign them higher probability of being selected.
2. If there are only few or no donors to the left (or right) of the donee, remove the furthest (and thus worsed matched) potential donors.
3. When possible, calibrate the probabilities so that the expected imputed auxilary value equals the donee auxiliary value.

Simulation setup

- Draw 1000 samples of size 400 from population of 1600.
- Auxiliaries $X_{Uniform}$,X_{Normal},X_{Gamma} with range $(\pi/2, 2\pi)$
- Logistic missingness mechanism $logit(Pr(R = 1|X) = f(X)$. On average 25% nonresponse irrespetive of X.
- Estimate mean of study variable; Linear Y_X, nonlinear Y_{cosX}, mixed Y_{X+cosX} .
- Multiply impute $B = 20$ times.

Determine the number of donors in three ways.

1. Use $k = 2, ..., 30$ nearest neighbours (kNN).
2. Units with distance less than ϵ, which is proportional to the number of potential donors (fix).
3. Adapt fix so that ϵ is increased (decreased) if there are few (many) donors close to the donee ($adap$).

Also, either include or don't include the bias reduction feature.

Simulation results

Summary results on bias (see page 165 in the workshop book of "'lecture materials and contributed papers"')

- Bias correction $<$ No bias correction
- Nearest neighbour $<$ Fixed $\approx$ Adaptive
- Nearest neighbour increase with k

Summary results on variance (page 166)

- Bias correction $\approx$ no bias correction
- Nearest neighbour $\geq$ Fixed $\approx$ Adaptive
- Nearest neighbour decrease with k

Summary results on estimated variance (page 167)

- fixed and adaptive overestimate
- Nearest neighbour increase with k
- large kNN_b generally best, but
- small kNN_b underestimates variance

Simulation; Summary results for donor pool size

Always use the bias corrections !
If only concern is ...

- bias, use small donor pools (kNN).
- variance, use large donor pools (fixed/adaptive/large kNN_b).
- estimated variance, use not too few donors (large kNN_b, perhaps fixed/adaptive).

Given kNN (with bias correction), for ...

- bias, select small k.
- variance, select large k (at least not very small).
- estimated variance, select not too small k.

This is just a few examples with one study/auxiliary variable; basic rules on how to select the (number of) donors; one missingness mechanism, but the results are consistent with other results, e.g. various multivariate or multimodal auxiliaries, other estimators, etc.

Thank you for your attention!

References

Laaksonen, S., (2000). Regression-based nearest neighbour hot decking, Computational Statistics, 15(1), 65-71.

Little, R. J. A. and Rubin, D. B. (2002). Statistical analysis with missing data. John Wiley & Sons, New York.

Lo, A. Y. (1988). A bayesian bootstrap for a finite population. The annals of statistics, 16, 1684-1695.

Pettersson, N., (2012) Bias reduction of finite population imputation by kernel methods. To appear in Statistics in Transitions.

Rubin, D. B., (1981) The Bayesian bootstrap. The annals of statistics, 90, 130-134.

Silverman, B. W., (1986). Density estimation for statistics and data analysis. Chapman & Hall, London.