Statistics Finland

Q2010 Training Courses

I M P U T A T I O N TECHNIQUES

Vantaa River
1 April 2010
5 km
from the airport

Statistics Finland

Imputation

- Time
  - Monday 3 May 2010
- Venue
  - Statistics Finland
  - Työpajankatu 13, FI-00580, Helsinki, Kalasatama
- Instructor for Imputation
  - Prof Seppo Laaksonen (University of Helsinki and Statistics Finland)
  - Seppo.Laaksonen@Helsinki.Fi

Statistics Finland

Content

What is imputation, its purpose, concepts
Missingness mechanisms
Most common tools for missing item handling without real imputations
Missingness pattern
Targets for imputation
Imputation process
Single and multiple imputation
Imputation model
Imputation task
Summary of imputation strategies
Examples including simple methods
Preserving associations in the case of missing data
General conclusion
Thanks
ANNEX: IMAI

What is imputation?

It is to insert a value into the data in a more or less fabricated way. **Why**?

- Since there is no value in this cell, that is, it is completely missing.
- Since the existing value is partially missing (like given as an interval) but it is desired replace this with a unique value.
- Since the existing value does not seem to be correct, and consequently, it is desired to get a more reliable value to replace this.
- Since the current value seems to be too confidential, that is, and this individual unit should be disclosed. Motivation: the fabricated value can be considered as less confidential.

Imputation can be performed both for the macro and micro data but during this course I only consider the imputation methods of **micro** data. However, basically the same methods can be applied to macro data but usually this imputation is more limited.

Purpose of imputation

The purpose of imputation is twofold

-**Either** to replace a missing or partially missing or incorrect value with a such value that the estimate behind this variable will be more valuable than without imputation. Thus if imputation is advantageous from the estimation points of view, use it. Naturally, there are in surveys several estimation tasks and can be possible that a certain imputation is not advantageous in all respects. Hence, it is possible that some estimates are computed without imputation and some others with imputation. On the other hand, a big question is which imputation is best for each estimation. It is good to notice also that a bad imputation may worsen the estimation. Be careful!

- **Or** to make data more confidential. This leads to create certain incorrect values into the data that is not difficult but this should not be a purpose but to impute the confidential values so that their pattern gives opportunity to get as the reliable estimates as possible.

Use of imputation increases

- Since missingness and data deficiencies have become more common and also statistical confidentiality is more important.

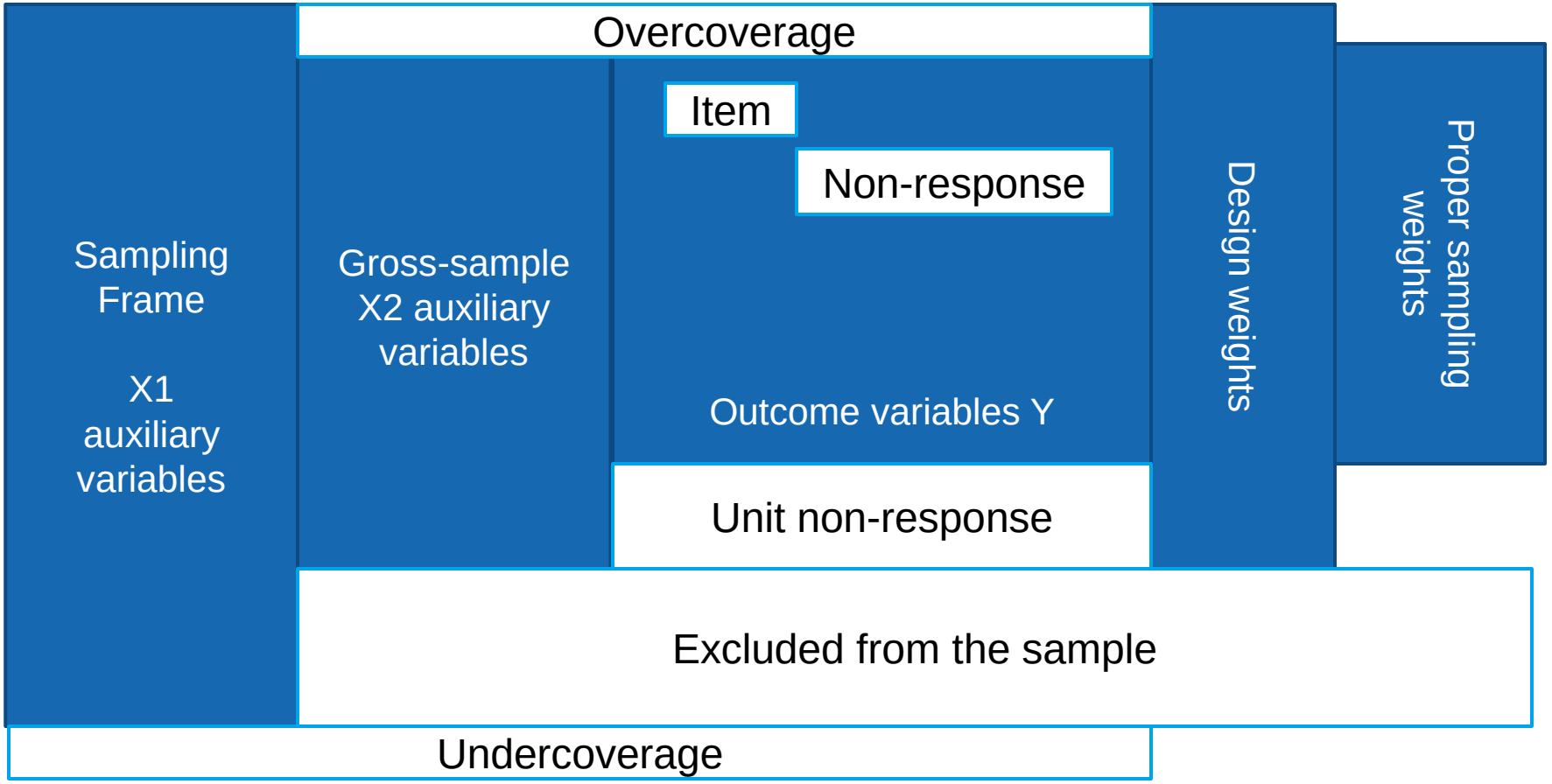
- Since methodology has been developed but its implementation into software is not satisfactory. Hence, many imputations in NSI's are still needed to do using specific programming. Some methods are fortunately easy to program but some not. Most methods presented here are not difficult to code using SAS that I have used.

- Also imputation research has been increased but not just recently. Early 2000's was a good time. See e.g. the Euredit project website, http://www.cs.york.ac.uk/euredit/

Statistics Finland

Micro data and Missingness

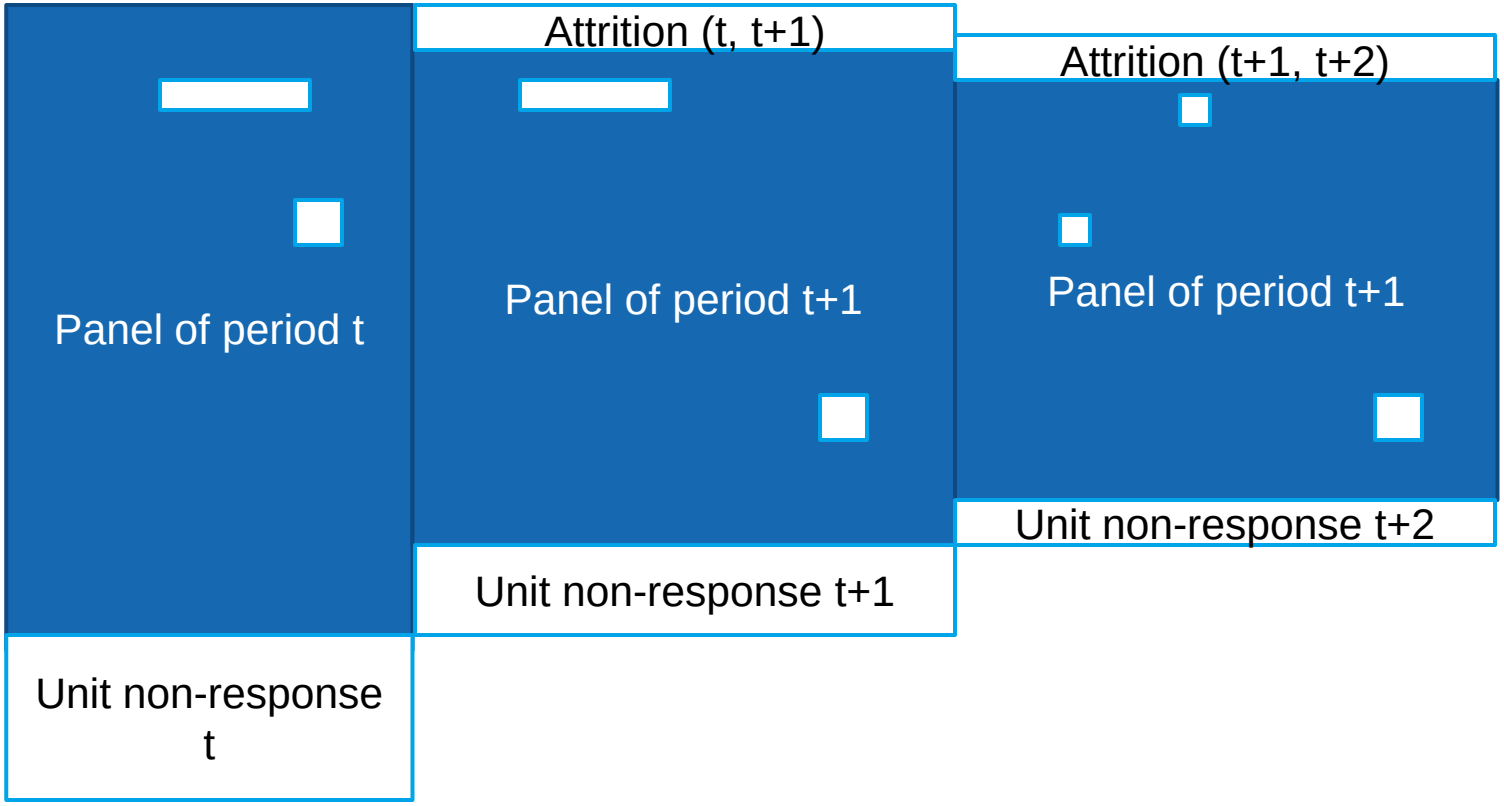
Now I focus on micro data where one can see various types of missingness. This is a cross-sectional case (white boxes are missing):



Statistics Finland

Micro data and Missingness

Cohort type of panel example



Statistics Finland

What can be imputed due to missingness

When looking for those schemes, we can find the following possible imputation affairs:

(i) Undercoverage that requires a new up-to-date frame. Very seldom possible.
(ii) Those units that are not selected into the sample. Done in threoretical (simulation) studies
(iii) Unit non-response, all or some variables. If done, called mass imputation. This is competitive to weighting methods.
(iv) Item non-response. This is the most common case.
(v) Deficient and sensitive values. Quite common.
(vi) Second, third etc wave missing values in cohort studies given that the previous value exists (or imputed correctly enough).

Missingness mechanisms 1

Imputation requires useful auxliary information. Without such data imputation is still possible but maybe bad. On the other hand, it is important to assess the missingness (or response) mechanim.

There are four basic mechanisms good to think and make assumptions before starting the imputation (usually only three of these are presented in literature):

MCAR (Missing Completely At Random): If this could be reality, it is rather easy to decide which methods to apply. Most methods are workable and you do not need auxiliary variables either.

MARS (Missing At Random Under Sampling Design): Now missingness only depends on the sampling design variables. This is often used so that one assume that MCAR holds true within strata (pre-strata, or even post-strata). Here imputation is performed by strata.

Missingness mechanisms 2

MAR (Missing At Random Conditionally): Now missingness depends on both the sampling design variables and all possible other auxiliary variables. This assumption is much used when good auxliary variables are available.

MNAR (Missing Not At Random): Unfortunately this is the most common case in real-life to some extent. So, when all the auxiliary variables have been exploited much help have been received but still it is rather clear that our results are not ideal. So, it is good to interpret possible biases in results against general knowledge and missingness of good auxiliaries.

Statistics Finland

Most common tools for missing item handling without real imputation

(i) In the case of mass missingness, the weighting or the reweighting is is mostly exploited. This is possible only for the respondents. The respective imputed data thus covers the non-respondents too (or those non-respondents desired to include). Note that one imputation strategy is a kind of weighting method but its weights are more flexible than the standard reweighted sampling weights.

Most common tools for missing item handling without real imputation 2

(ii) Item-nonresponse is marked with a good and well-covered code such as:

- -1 = respondent candidate not contacted
- -2 = respondent refused to answer
- -3 = respondent was not able to give a correct answer
- -4 = missing for other reasons
- -6 = question was not asked from the respondent
- -9 = question does not concern the respondent

These codes are not much used but such as 7, 8, 9, 66, 77, 88, 99 instead.

The negative values are easy to observe. Do not use a zero (0)!

Statistics Finland

Most common tools for missing item handling without real imputation 3

(ii) cont.

The good and illustrative codes are useful also when deciding the imputation methods itself. Thus a different method may be chosen for the different type of missingness.

Moreover, it is good to notice that the coded variable is full, without missing values. This kind of a categorical variable can be used as an explanatory variable in standard linear and linearised models, among others. But if desired to use it as continuous, real imputation is required.

Statistics Finland

Most common tools for missing item handling without real imputation 4

(iii) The values with missing codes are excluded from each analysis so that the observation number varies. This strategy does not give consistent results with each other.

(iv) Close to case (iii) but now the units with missing values have been excluded from each analysis. In this latter case, there are always the same number of observations. The standard multi-dimensional analysis makes this automatically for those variable patterns that are used in the multidimensional analysis. This strategy gives consistent results with each other.

(v) Pairwise analysis for multivariate purposes in such cases where e.g. the correlations are the basis for further analysis. This operation first computes pairwaise correlations like in case (iii) and when continues from the correlation matrix towards multivariate analysis.

Statistics Finland

Missingness pattern

Before imputation is necessary to examine missingness by variable, that is, compute e.g. the missingness pattern. Here is a simple example of some variables in an enterprise sample

exporter_res	profit_res	invest_res	R	COUNT	PERCENT
0	0	0	0	255	29,92958
0	1	1	0	12	1,408451
1	0	0	0	45	5,28169
1	0	1	0	15	1,760563
1	0	1	1	10	1,173709
1	1	1	0	261	30,6338
1	1	1	1	254	29,81221

Missingnes indicator for three variables

Variable R is their summary missingness indicator.

Statistics Finland

Analysis of missingness pattern

Logistic regression for the previous missingness pattern

Explanatory variables from the register

Model Information

Distribution	Binomial
Link Function	Logit
Dependent Variable	R

Ordered Value	R	Total Frequency
1	0	588
2	1	264

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	-0.0100	0.1219	-0.2489	0.2290	0.01	0.9349
nace_reg	Trade	1	1.1919	0.2636	0.6752	1.7086	20.44	<.0001
nace_reg	Traffic	1	2.1026	0.2976	1.5193	2.6860	49.91	<.0001
nace_reg	Other	1	2.3819	0.3702	1.6562	3.1075	41.39	<.0001
nace_reg	Service	1	1.6052	0.2516	1.1122	2.0983	40.72	<.0001
nace_reg	Constr	1	2.0429	0.3553	1.3465	2.7392	33.06	<.0001
nace_reg	Manufact	0	0.0000	0.0000	0.0000	0.0000	.	.
turnover_reg		1	-0.0000	0.0000	-0.0000	-0.0000	15.80	<.0001
employed_reg		1	0.0006	0.0005	-0.0004	0.0016	1.46	0.2276
Scale		0	1.0000	0.0000	1.0000	1.0000		

Targets for imputation should be specified clearly

It is rather clear (except when imputation aims at protecting data)

(i) That a user is happy if the imputed values are as close as possible to the correct values. Success at individual level. Another point is that how to know how close they are, except in some cases. This may be often the too demaning target and hence
(ii) A user is still fairly happy if the distribution of the imputed values is close to the distribution obtained from true values. Success at distributional level. Of course this is hard to check but however easier than case (i).
(iii) The target to succeed at aggregate level is also statisfactory and specificly in NSI's where such estimates as average, total, ratio, median, decile and standard deviation are typical.
(iv) Some users hope to get the order of imputed values as correct as possible.
(v) Finally, success to preserve associations (like correlations) is also important in many studies.

Targets for imputation should be specified clearly 2

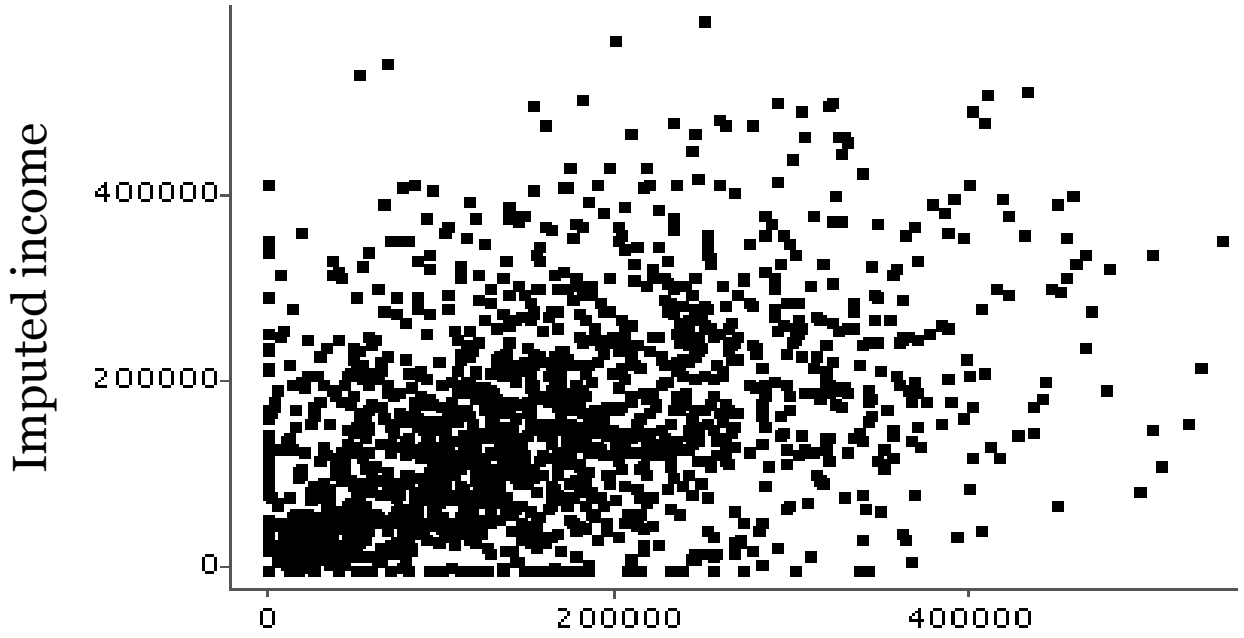
Target (i) is naturally very hard and it is not nessarily realistic to achieve. On the other hand, it is not clear what this means. I give some examples in which cases I have known the true values and I am thus able to check the success of imputations.

The first example on next page shows well that success at individual is not good when checking how absolute differences between imputed and true values. However, this imputation works very well
-At aggregate level where e.g. the difference in the overall average is about 0.6%.
-At distributional level where the standard deviation fits very well, 0.23% error (and Gini coefficient respectively and Kolmogorov-Smirnoff measure).

It should be noted that income differences are most interesting estimates in general although some aggregates are also important.

Statistics Finland

Example to impute incomes, source: My imputations for the Euredit project



Targets for imputation should be specified clearly 3

Individual level success may even be a bit funny target. This can be demonstrated well for categorical variables.

I tested with the same enterprise data as earlier how well the simplest possible imputation method works for the binary variable *invester* (1=the enterprise has invested, 0=not). The observed data show that the percentage was 77.2%. So, I imputed all the missing values as the majority values = 1. When I tested how well I succeeded:
-The mean for the imputed values = 100% but for the true ones give much less = 68.3%.
-The error at individual level occurred only in 31.7% of the cases, that is, the true values were imputed for those 68.3%:lle.
For comparing. my best imputation method was of course much better for the mean (68.9% vs 68.3%) but at individual level not (37% vs 31.7%).

Imputation process

Imputation is part of the data cleaning process. It can be considered to cover the following 6 **actions**:

(i) Basic data editing in which part the values desired to impute are also determined.

(ii) Auxiliary data acquisition and service incl. preliminary ideas to exploite these.

(iii) Imputation model(s): specification, estimation, outputs

(iv) Imputation task(s): use outputs of the model for imputation, possible re-editing if the imputed data are not clean and consistent.

(v) Estimation: point-estimates, variance estimation = sampling variance plus imputation variance.

(vi) Creation of the completed data (or several data): includes good meta data such as **flagging** of imputed values, documenting of the whole imputation procedure and deciding what to give outsiders.

Next I focus on the actions (iii) and (iv).

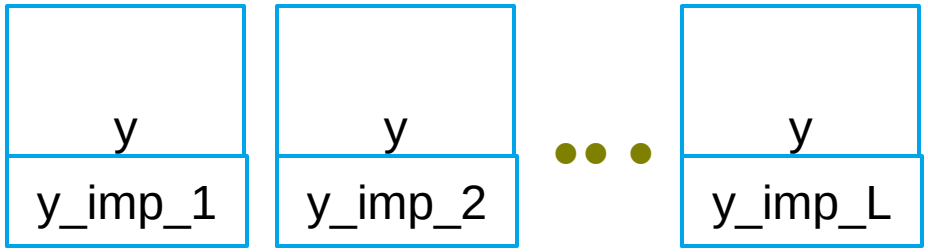
Single and multiple imputation

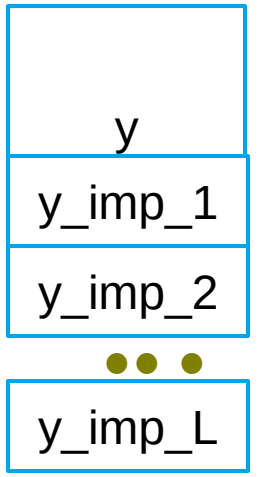
Imputation can be performed for each desired value of the non-complete variable just once, or several times. The first is called *single imputation (SI)* and the second *multiple imputation (MI)*. These are not the two different imputation methods as often said, since multiple imputation means that single imputation has been repeated several times. So, each single imputation should aim at succeeding as well as possible e.g. avoiding the bias. There are the strict rules how to repeat imputation properly. In this presentation these have not been much discussed. The rules are not always clear and hence often criticised.

MI is in certain NSI problems difficult to realise so that the users are happy. E.g. imputing values of large businesses this methodology may cause confusions. Instead, if imputation is concerned a big number of missing etc values for e.g. households and small/medium sized businesses (thus sample with large sampling weights) MI may be beneficial.

Single and multiple imputation 2

Technically: MI can be done in the two ways:
(i) Construct several completed files (minimum = 3 according to the 1987 Rubin book but today most require some more, even some tens ; that depends on how difficult is the estimand). This is the most common strategy.
or
(ii) Extend the initial file by new units based on imputations and balance the whole file so that each initial sampling weight is divided by the number of imputations. This is not well workable if the missingness pattern is complex.





Single and multiple imputation 3

The latter file (ii) can be analysed as a single file as far as point-estimates are concerned. The imputation variance will be partially taken into account using standard survey sampling procedures since there are the variation due to imputation. Strictly speaking some variation will not be included in these standard error estimates. I think that the literature is poor in this respect.

The first case, several files, has been much more studied. There is a consensus how to compute the point estimates that can be whatever types you need, thus e.g. total, average, median, ratio, decile, regression coefficient and other model estimates:

Let say

L = number of imputations u,

θ = parameter desired to estimate

Q = estimate of the parameter

B = variance estimate for the parameter

$$Q_{MI} = \frac{\sum_u Q_u}{L}$$

Single and multiple imputation 4

There is no consensus on the variance estimates even we forget problems to correctly construct proper imputations rules with complex data (like for skewed distribution and when the pattern of auxiliary variables is poor).

Two general variance estimators have been presented. Rubin's classical formula is almost only used. He says that it is derived from the Bayesian theory. At contrast, Björnstad gives in the 2007 JOS another formula that is not in conflict with Rubin's approach since Björnstad says that his formula is non-Bayesian. In fact there approaches are contradictory since Björnstad says also that Rubin's approach is not well workable with NSI data. So, what to use or not to use. A user will have difficulties. For me, the basic question is still how to arrange imputation properly (this I wrote in the 2007 JOS) so that the factual missingness mechanism can be followed well. It is in practical problems a big question. Moreover, the Björnstad revision is in my opinion logical since it takes concretely into account the 'imputation' rate.

Single and multiple imputation 5

Both formulas are similar in the sense that the formula consists of the two components, that is, of the within-imputation variance and the between-imputation variance. The first is the same in both formulas but the second differs.

Rubins' formula $$B_{MI}(DR) = \frac{\sum_u B_u}{L} + (1+\frac{1}{L})\frac{1}{L-1}\sum_u (Q_u - Q_{MI})^2$$

Björnstad's formula $$B_{MI}(JB) = \frac{\sum_u B_u}{L} + (k+\frac{1}{L})\frac{1}{L-1}\sum_u (Q_u - Q_{MI})^2$$

in which k is not equal to 1 but depends on the sampling design. Stratified random sampling gives $$k = \frac{1}{1-f}$$

where f = the 'imputation' rate. If all are imputed, the variance cannot be computed that is logical. Rubin can estimate it still.

Single and multiple imputation 6

Now I go forward to imputation methods. SI and MI are attending in those sessions. It is good to notice that MI always needs a stochastic element, that is, such techniques that include some stochasticity can be bases for MI unless the strategy automatically is workable for MI. As said this is a big question and I am not able to give the definite answer which stochasticity is reasonable for Bayesian and which for non-Bayesian MI. Discuss this question with your friends too.

Note that many researchers using MI are not especially interested in NSI types of simple estimates like totals, means, ratios and distributions but they are modelling and thus interested in those estimates and wish to obtain reasonable standards errors.

Imputation model

Imputation model should be integrated strictly to the next step, that is, to imputation task. There are two options to determine the specification of the imputation model:

• To determine the model using smart information so that it predicts well the case required to impute. The model may a deterministic (or stochastic function) like $y = f(x)$ (+ e) or a rule (like in editing) such as 'if so and so but not so then it is that.'

• To estimate the model using either the same data required to impute or other data that is similar (at least the structure) to the present data.

The previous models are often used in simple (conservative) imputations and in the same step as editing. Next I will focus on the latter models.

Imputation model 2

This second type of imputation model is always such in which it is purpose to predict something using auxiliary variables as independent variables.

The dependent variable of this imputation model can be of the two types only:

(i) either the variable being imputed itself

or

(ii) the missingness indicator of this variable.

Case (ii) can cover all possible forms, categorical including binary and continuous but in case (ii) the variable is binary.

Imputation model 3

These two models are estimated from the two different data sets:

(i) From the respondents (observed units)

(ii) Both from the respondents and the non-respondents.

But of course, the explanatory variables should be available from both the respondents and the non-respondents. Note my earlier comment that a categorical variable with the missingness codes may work reasonably in the imputation but many such variables maybe not unless these are concerned different units.

Imputation model 4

The model (ii) is concerned a binary variable (1 = responded, 0 = not) but the same model can used for the model (i) if the dependent variable is binary (e.g. 1 = employed, 0 = unemployed).

You know how to work with the binary model to predict. First you have to choose a link function, that can be:

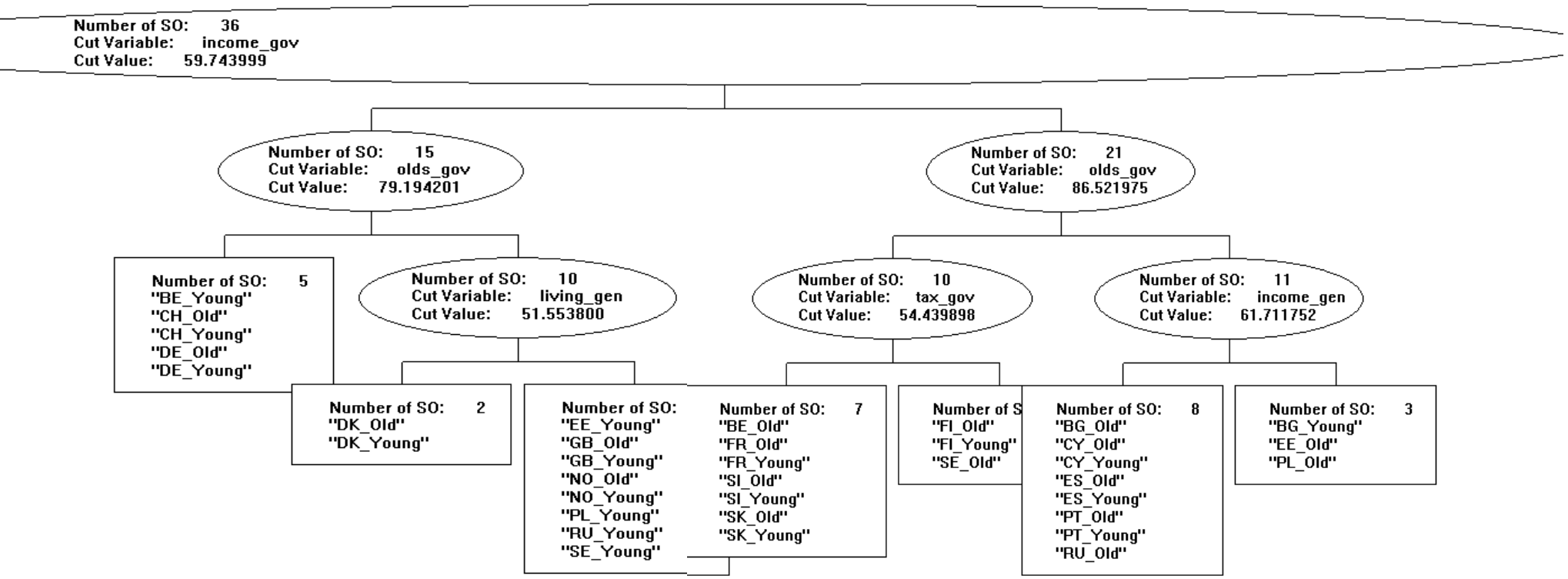
-logit
-probit
-complementary log-log
-log-log .

There are no dramatic differences in explaining models between those link functions but of course some. Imputation thus requires to use this model for predicting the response propensities for all units (respondents and non-respondents). That is, the first outputs are those values between (0, 1).

Imputation model 5

In addition to ordinary models such as linear regression or probit regression, the imputation model can be nonlinear and nonparametric. An interesting example of the latter ones is *tree modelling*. If the dependent variable is categorical, we speak about *classification tree*s, whereas the model for continuous variable is *regression tree*. Moreover, neural nets often create analogous groups of the gross sample. This kind of a group is called in imputation terminology as *imputation class* or imputation cell.

Imputation cells can also be constructed manually or using smart statistical thinking. For example, strata can be rather good imputation cells. Given that the imputation cells are homogenous from the imputational points of view (especially if MCAR holds true within cells), these offer many advantages.



This is an example of a classification tree in which case the cells have been created from three two variables, the country and the two age-groups. At bottom you see the terminal nodes of the tree. These could be optimal to use as imputation cells. You see e.g. the smallest cell consists or the Danishes including both old Danishes and young danishes. Source: European Social Survey, Round 4.

Imputation task

The two alternatives in general can be exploited after you have estimated the imputation model:

(a) Model-donor approach in which case the imputed values are computed deterministically (or stochastically) from the predicted values (adding noise) of the model.
(b) Real-donor approach in which case the predicted values (or adding noise) are used to find the nearest or a near neighbor of a unit with a missing value from whom an imputed value has been borrowed.

You see that the imputed values of case (b) are always observed values, observed at least once for respondents. The imputed values of case (a) are not necessarily observed except often for categorical variables.

Statistics Finland

Imputation task 2

To integrate model and task you see that we have the following options. So, the predicted values of the missingness indicator cannot be used for model-donor imputation directly.

	(a) Model-donor approach	(b) Real-donor approach
(i) either the variable being imputed itself	Yes	Yes
(ii) the missingness indicator of this variable	No	Yes

Statistics Finland

Imputation task 3

Comment:
I use the term donor as it is used by many others but it is not general to use the term like model-donor. This methodology is often quite different, even spoken about model imputation when meant a type of model-donor imputation like when the imputation model is regression model and the imputation task is the direct predicted value. This is for me confusing since regression model can be used also for real-donor imputation. Model imputation is also strange since imputation always needs a model; so all imputations are model imputations.

The same confusion has been met often when speaking about logit imputation or probit imputation since this model can be used in both imputation tasks.

My term donor in task (a) means that the lending is derived from a group (group donors) that is a factual situation when modelling. The donor in task (b) is a unit, an individual.

Statistics Finland

Imputation task 4

Comment:

Many other terms have been historically used in imputation literature and still they are used. A typical example is hot deck imputation or hot decking.

Wiki describes:

A once common method of imputation was **hot-deck imputation** where a missing values was imputed from a randomly selected similar record (The term "hot deck" dates back to the storage of data on punched cards, and indicates that the information donors come from the same dataset as the recipients; the stack of cards was "hot" because it was currently being processed. Cold-deck imputation, by contrast, selects donors from another dataset.)

You see that this term resembles some aspects of my real-donor imputations. For me the term is confusing and not needed to use after 60 years after inventing in strange circumstances. You are free to use it but I will not use.

Imputation task 5

Both imputation tasks use stochasticity or they can be applied deterministically. If stochasticity has been used in the imputation model, it follows that the imputation task should be automatically stochastic but it still requires to use certain random numbers in the imputation task. Stochasticity can be added also in the imputation task using appropriate random numbers. It is needed to assume how random numbers behave or what is their notional distribution (normal, lognormal, uniform). If the real life data do not behave so, your imputation may be violated.

Imputation task 6

The imputed value of the model-donor method is simply:
either
(•) Predicted value of the imputation model (*deterministic imputation)*
or
(••) Predicted value plus a noise term of the imputation model (*stochastic imputation)*.

I do not go to details of the noise term but when using regression model it is often assumed its distribution to be normal with the mean = zero and the standard deviation = root mean square error. A problem is that there can be outliers in random values and consequently in imputed values. It requires to truncate outliers in some way. Another option, less problematic, is to use a pattern of observed residuals estimated for the respondents and then randomly draw these residuals to the noise for non-respondents. This strategy thus is a kind of a real-donor method.

Imputation task 7

The imputed value of the real-donor method requires a metrics used to find an optimal unit donor from whom to borrow the imputed value.

This metrics can be derived from outside the data. Most typically, it is assumed that certain units (within an imputation cell, in particular) are as close to each other. This means that a donor has been selected randomly (within a cell). This is thus stochastic.

The other common strategy is to use a smartly chosen other metrics and search for the nearest or a near donor from the data set. This because it is assumed that the units close to each other are similar. Of course, the success depends on those variables in this metrics.

Imputation task 8

The imputed value of the real-donor method.
The third good and rational strategy in many situations is to use model-donor imputation values over both the respondents and the non-respondents as the nearness metrics. This thus mean that we impute technically the values for the respondents too, using the same strategy as for the non-respondents. This is not difficult technically. The next step is to work as in the previous case either to select the nearest donor, or a near donor that is usual when desired to randomise the procedure. Thus e.g. our nearness metrics can be the previous model-donor output:

(•) Predicted value of the imputation model (*deterministic imputation)*
or
(••) Predicted value plus a noise term of the imputation model (*stochastic imputation)*.

Summary of imputation methods

Here regression model has been used as imputation model.
For categorical variables the model may be binary or multinomial.
If the model is response indicator model-donor methods cannot be used, only **[1, 1]** and **[1, 2]**

	Deterministic Single	Stochastic Single Multiple
Real-donor	**[1, 1]** E. g. *y* is the dependent variable in imputation model ja a lot of auxiliary variables as explanatory ones. = predicted values as nearness metrics	**[1, 2] As** [1, 1] But the random normally distributed noise term added
Model-donor	**[2, 1] As** [1, 1] But predicted values used as imputed ones.	**[2, 2] As** [1, 2] But predicted values plus noise term used as imputed ones.

Example with very simple data

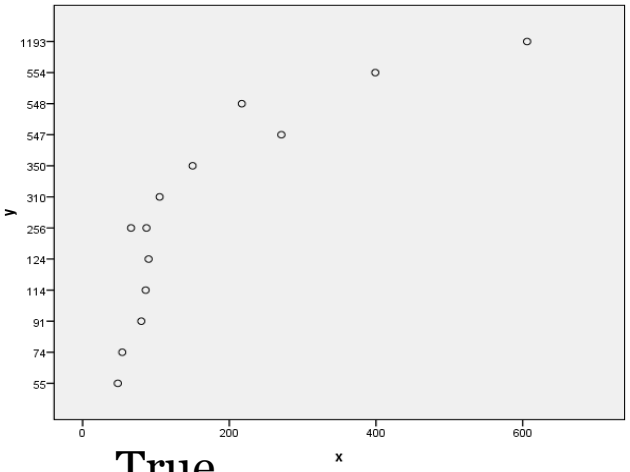
You have wondered that any commonly known imputation methods like mean imputation has not been mentioned in the text. This is due to my framework that covers of course those simple and usually inappropriate methods. The following example illustrates my framework in which the imputation model and imputation task is good to recognise even though the model is simple.

Example with very simple data 2

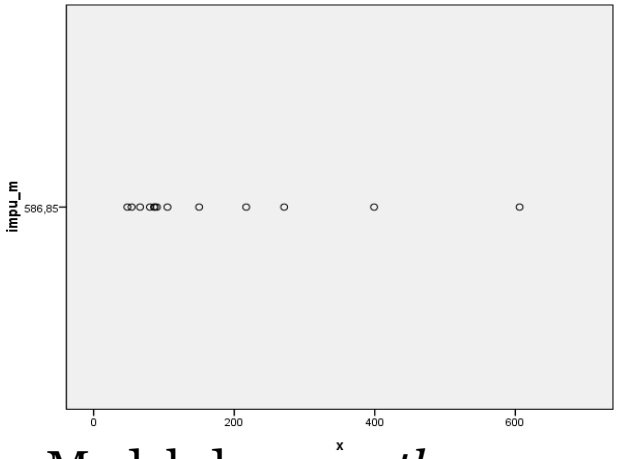
The data are artificial. Variable *y* is that required to impute to some extent. I have only one auxiliary variable *x*. These two variables are well correlated, *r*=0.92. The number of the units is 40, that of the non-respondents is 13. Missingess is not random, it is higher for small and large *y* values. Possibilities for successful imputation exist. My first imputation model is *y=the mean* but in the other four I tried the model *y=x,* also adding a normally distributed noise term. Results are below and on next page.

		Observations	Mean	Std deviation
True		40	507	317
Respondents		27	587	292
Model-donor				
Model	y= the mean	40	587	238
Model	y=x	40	519	279
Model	y=x+e	40	516	295
Real-donor				
Model	y=x	40	499	299
Model	y=x+e	40	534	299

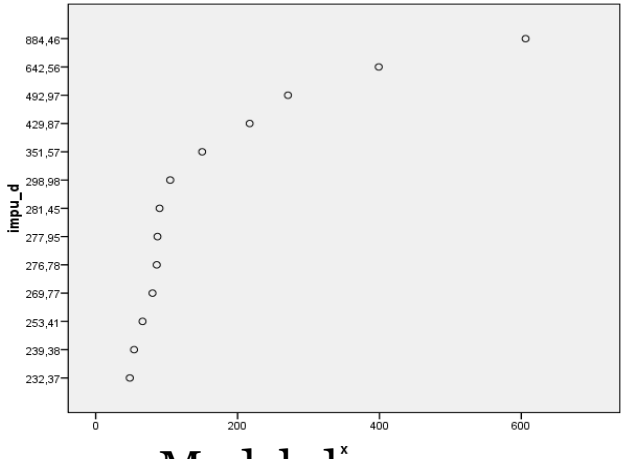
Statistics Finland



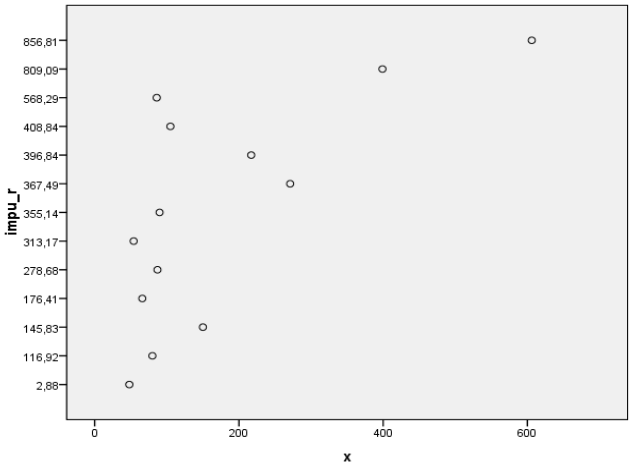
True



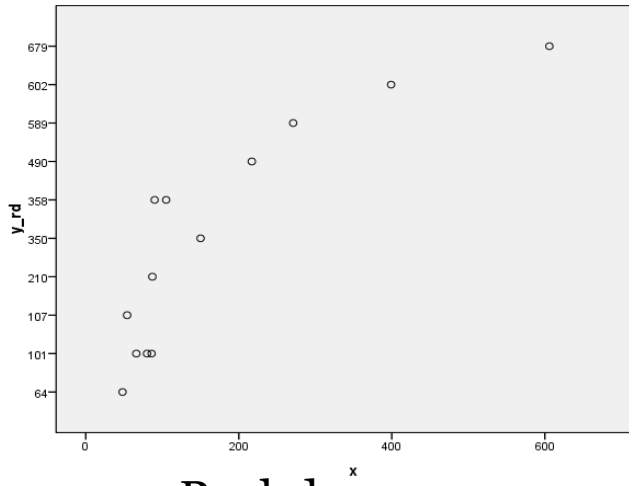
Model-donor *y=the mean*



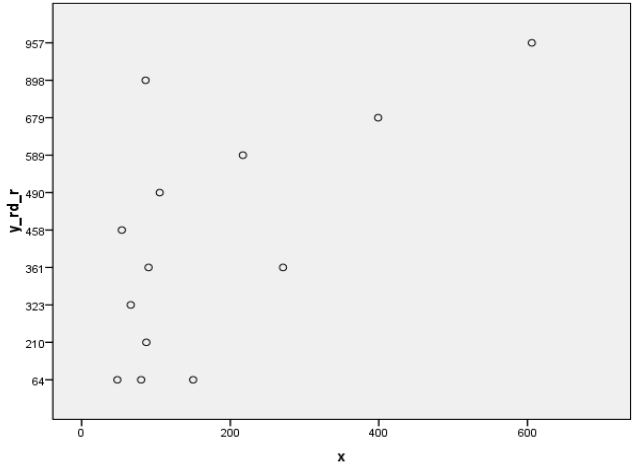
Model-donor *y=x*



Model-donor *y=x+e*



Real-donor *y=x*



Real-donor *y=x+e*

The scatters are for the imputed values. Thus compare with the true scatter.

Other simple imputation models

Deterministic model-donor imputations

Model can also be
y=median (or median of imputation cells); continuous and ordinal variables
y=mode (or mode of imputation cells); categorical variables
y=p% quantile that requires some other information (also for cells), continuous and ordinal variables
y=ratio
y=y_previous or *y=y_previous *inflator* (or deflator) in panels

Other simple imputation models 2

Stochastic model-donor imputations

Observed distribution for non-respondents, best by homogeneous imputation cells

You have observed
e.g. the following distribution:

Category	Frequency
A	10
B	40
C	35
D	15

Now you can impute
(r=uniformly distributed random number)
if r<=0.1 then y_imp='A';
else if r<=0.5 then y_imp='B';
else if r<=0.85 then y_imp='C';
else y_imp='D';
This does not ensure the individual preservation but the distribution may be satisfactory given the imputation cells are good.

Statistics Finland

Another example with simple imputations

Monthly business data so that imputation is made for preliminary quick estimates. Thus, we will know the real values some months later for most units. This is concerned short-term statistics and the variables being imputed are continuous.

This case gives opportunity to assess the success of imputations against recent real values, this advantage is good to exploit. Of course, there is still uncertainty since those two periods are not necessarily similar. If we however assume that the periods are reasonably similar, we can use the following strategy:

Example cont.

1. Perform several alternative imputations, say *k*, for the missing values of period *t*.

2. Using those *k* imputation methods impute the missing values for those units over some recent periods where the real values are already available. For example, these periods could be *t-2*, *t-3*, ..., *t-s* where *s* could be 6 or more.

3. Compare the success of each method against those real values.

4. Reject the method with a severe bias for a certain period (or many periods): e.g. if the maximum bias is over the certain 40 log%.

5. Choose the method with the smallest relative (log%) bias. Note that it is easiest if the same method has been chosen overall but you can choose a different method for the different domains (e.g. for different industry classes).

In this imputation method 'competition' you can test simple and complex methods such as on next page:

Example cont.

All methods can be applied by domain (industry class, size band). Also: there can be several cases for each if a unit does not exist in each period. The methods can be formulated also relatively and as differences, see method (i). Noise can be taken from a assumed theoretical distribution like normal distribution or from the observed residuals (deviations based on previous imputed values).

	Imputation model	Imputation task	Comments
(i)	$y_t = y_{t-1}$ $y_t / y_{t-1} = 1$ $log(y_t)-log(y_{t-1})=0$	Deterministic model-donor = md	Called also Last value carried forward
(ii)	$y_t = y_{t-1} + noise$	Stochastic md	
(iii)	y_t = *average of recent values*	Deterministic md	
(iv)	y_t = *average of recent values + noise*	Stochastic md	
(v)	$y_t = y_{t-12}$ (+*noise*)	Deterministic md (Stochastic)	One year earlier value carried forward
(vi)	y_t=*(average change of recent observed values)** y_{t-1} (+*noise*)	Deterministic md (Stochastic)	Assumed that the change for missings equal to non-missings
(vii)	y_t=*f(recent values, industry class, other auxiliary variables x)*	Deterministic (Stochastic) md or rd	Linear or loglinear regression etc.

Preserving associations in the case of missing data

Associations like correlations are in some cases good to preserve or not violate dramatically when handling missing data. Here are some strategies:

(i) Do not impute at all, thus use data deletion. You will lose observations and your standard errors are larger. Also your results are biased to some extent. But do not matter if you do not like to publish this paper.

(ii) Try to use such analysis method that takes missingness into account.

(iii) Adjust for missingness by a good reweighting method, also using auxiliary variables as much and well as possible.

(iv) Apply a real-donor methodology so that the whole (or essential) pattern of the variable values has been chosen from the same donor. You can put a bit random variation there, of course. This kind of pattern may also be relative such as relative distribution, not absolute values.

(v) Apply sequential imputation so that impute first variable y1, next impute y2 so that the imputed variable y1 is one additional auxiliary variable, and so on y3. ,,,, all variables that are interest for you in this respect.

Special example Q2010

3.5.13 Session 20
Wednesday, 5 May
Imputation
Meeting room 21
16:00 – 17:30
Chair: Ulrich Rendtel, Freie Universität Berlin
16:00 Imputing a binary variable with two alternative imputation models
Seppo Laaksonen – Statistics Finland and University of Helsinki

Statistics Finland

General conclusion

Imputation is a sexy job ("I keep saying that the sexy job in the next 10 years will be statisticians," says Google's chief economist). This means this job is desired everywhere due to increasing missing data in all types of data environments. Unfortunately, tools to make imputations well are limited unless …. (put here your recommendations).

On the other hand, this area is not yet well developed and even standardised terminology is missing. Join networks to improve the current situation.

Statistics Finland

Thanks a lot for your participation

Enjoy imputations

Loch Ness Monster removed to Vantaa River

Statistics Finland

ANNEX: My summary on the so-called IMAI approach, published on the Euredit website

IMAI = INTEGRATED MODELLING APPROACH TO IMPUTATION

A. Selection of training data and auxiliary variables for it

There should be a maximal potentiality of auxiliary variables with non-missing values or such values which have been considered as non-missing (like earlier imputed values or using missingness codes).

IMAI = INTEGRATED MODELLING APPROACH TO IMPUTATION

B. Construction or choice of imputation model

The two alternative target variables may be used:
(i) the target variable with missing values or
(ii) the missingness indicator of the target variable.

A model for each particular case may be of a whatever type, thus parametric or non-parametric, the model may be estimated from the same data, from another data or 'logically deducted.' The purpose for modelling is its high predictability. Note that: my model can also include a composition of edit rules (i.e. giving the limits for imputed values).

Statistics Finland

IMAI = INTEGRATED MODELLING APPROACH TO IMPUTATION

C. Choice of criteria for imputation

The criteria for imputation are of two types:
(i) assumptions for direct predictability or
(ii) metrics for nearness.
Typically, such a metrics is based on an Euclidean distance measure or other model-external solutions, often using such auxiliary variables which are not used in a model. Alternatively, the metrics can be taken from model results so that it can be basically a pattern of the imputed values of another approach.

D. Imputation task itself

If the modelled values (predicted with or without noise term) are used as imputed values, I speak about 'model-donor' methods, whereas if a model and a metrics have been used to find a good donor from whom an imputed value has been borrowed, I speak about 'real-donor' methods. Note that this technique may be used for finding a good observed residual (noise term), too.
That is, imputation can be a mixture or both approaches, too.