Survey Methodology
Part F

Fall 2015

Seppo Laaksonen

Statistical editing
Imputation

Statistical editing or data editing is a big and important part of a survey process. If raw data has not been edited in any way, editing may require much time and resources. This is typically the case if a paper questionnaire is used, but all other modes give opportunity to make some pre-editing and hence final editing is obviously easier. This part presents the core methods and tools for editing, but we start from the main purpose of it

Statistical editing thus is a crucial part of quality assurance of the survey data and also the survey process. The first point is close to quality control and improvement of the data, and we focus here on it. However, it is also needed to look forward and to document strong and weak points of the survey process, that is, to learn about them for future surveys. This has been too often forgotten and hence the same errors or mistakes are repeated. It is good also to look at other similar types of surveys and they have edited and to use this information in a best way.

Specific tasks of statistical editing are as follows:

- Evaluate and develop the survey process for the future, as learning by doing. It is good to follow what happens in other surveys similar to this particular one.
- Develop the system that helps in reducing manual work in editing, using selective editing, for instance. Its aim is to concentrate on detecting most fatal errors of the data, such whose impact on data quality is big.
- Detect, check and correct errors of micro level so that the results from a macro (aggregate) level are also plausible and reliable.
- Pay special attention to missing values so that they are coded with as many codes as it is possible, or given as missing. At the same time decide preliminarily what to do for missing values in data analysis. If you decide to impute those partially or entirely, this decision is good to do now since you also need to think which auxiliary variables could be used in imputation.
- Provide indicators that tell about changes made in editing and how some core estimates have been revised ('improved') due to editing. Estimate also the workload of the editing (and imputation).

Edit Rules

Edit rules are the rules for checking the correctness of individual variable values.

These may be more or less strict; this can be determined by gates given for each value checking. If the gate is narrow, it thus is more strictly checked than in the other cases, thus if the gate is broad. The workload of editing thus much depends on the broadness of gates in editing.

On the other hand, the quality is expected to be better if the gate is narrower. But: all depends on how well a possibly erroneous value can be corrected. If all suspect values are possible to check from the respondent (using the list of suspect values), it is very fine but it is not possible in most human surveys. Fortunately, it is often possible in the case of big businesses of business surveys.

Edit Rules

If suspect values cannot be checked against true values, the only strategy is to make them believable, plausible or logical. This means first that they should be at a correct level or (i) within the predetermined range of each variable. This is fairly easy to do at data entry already, thus in pre-editing.
It is possible to give different acceptable values for different sub-groups such as gender, age or education. This is (ii) the second edit rule, thus so that one value may depend on the value of another variable. This is a bit more demanding in computer-assisted surveys but possible. It is danger that a respondent does not like if his/her answer is not accepted due to his/her another answer. Hence it should be used carefully, but in the stage of post-fieldwork editing it should be done but the solution is not maybe always nice since it is needed to change one of both values in order to pass this edit rule.

Edit Rules

The edit rule thus can be one-, two-, three-, and multidimensional. The number of suspect values is obviously growing at the same time, consequently the workload for checking and correcting implausible values. Usually, the logics of different values seems to be most important than that the value is right absolutely.
For example:

- If the age of a person is 10, and his/her has a child, it is maybe best to change the age but keep the child there.
- If the age of a person is 20 and he/she is university professor, either the age is wrong or the occupation is wrong.
- If a person is unemployed but the wage is 5000 €, one of these values is obviously wrong. It is good to look other answers as well before correcting one value.

Edit Rules

A special multidimensional edit rule a linear or other multivariate regression model in which the dependent variable is that desired check, or the variable of interest and a number of explanatory variables are selected and then the model is estimated.

Finally, the residuals are calculated and ordered. Now the extreme residuals are first looked. It is possible that these are due an error in the dependent variable or in one or more explanatory variables. Of course, all extreme residuals are not errors but they are interesting in other meaning.

Other edit checks

Identifiers may have a big role in surveys, for survey institution people in particular. These are unique identifiers such as personal identity code or business entity code. An identifier may consists of several variables as well such as firstname, secondname, birthdate or birthyear. All these are confidential and cannot be given outsiders without the permission. On the other hand, they should be correct, and hence checked and corrected if needed. All correct identifiers should be maintained in the survey institution as long as needed.

Moreover, the confidential identifiers are converted to a protected form into the file given outsiders. This conversion rule should be saved since it is possibly needed later.

Other edit checks

Due to mistakes in data entry it is possible that after merging two data for example, two same identifiers = dublications are in a new file. One of these should be deleted. This is part of editing work.

Extreme or other exceptional values may be awkward. They are often called Outliers. On the other hand, the data file consists also of Inliers that look as ordinary values but if controlling them by one or more other variables, they do not look anymore ordinary. In editing, it is most important to detect those of these that are not correct. An erroneous outlier is called Out-Error, and the inlier In-Error, respectively.

Other edit tasks

Unless the questionnaire already and the data entry has not coded correctly missing values, they should be coded at this stage.

One rule is that a missing value should never be coded as ‘zero’ since it is usually a proper value.

Instead, the best codes for missingness would negative ones, e.g. -1, -2, -3, -4. The ESS and many other surveys use such codes that are far enough of the codes of proper codes, e.g. 7, 8, 9, 77, 88, 99, 6666.

Adding nonresponse in editing?

We have in Part B considered problems in replying, including the term ‘satisfycing.’ One consequence from satisfycing is ‘straightlining, and the other is ’item nonresponse’ without any real reason. These could be found if editing is made well but it is not necessarily easy.

What to do if detected definitely enough? One consequence could be to mark the whole answer as non-respondent since the answers are not plausible, thus coding as unit-nonresponse. If the half of the answers are not plausible, it is possible to change these as missing and to include the respective meta data code such as ‘deficient’.

Selective editing

Is much used in business surveys where the fact variables are common. It is possible to use also for other variables. There are several approaches to selective editing but the basic idea is to construct a model (that can be a statistical model or a mathematical model as a function) that predicts the probability that a certain value is erroneous (called error-localization). In editing, the values with highest probabilities are first checked and corrected most carefully, and those with low probability less carefully or even left as such or corrected automatically.

When developing the selective editing model, it is good to train it with the so-called training data set in which the workability of the model has been first checked against a smaller data set, often more manually. And when it has been found that the model works, it has been run over the whole data.

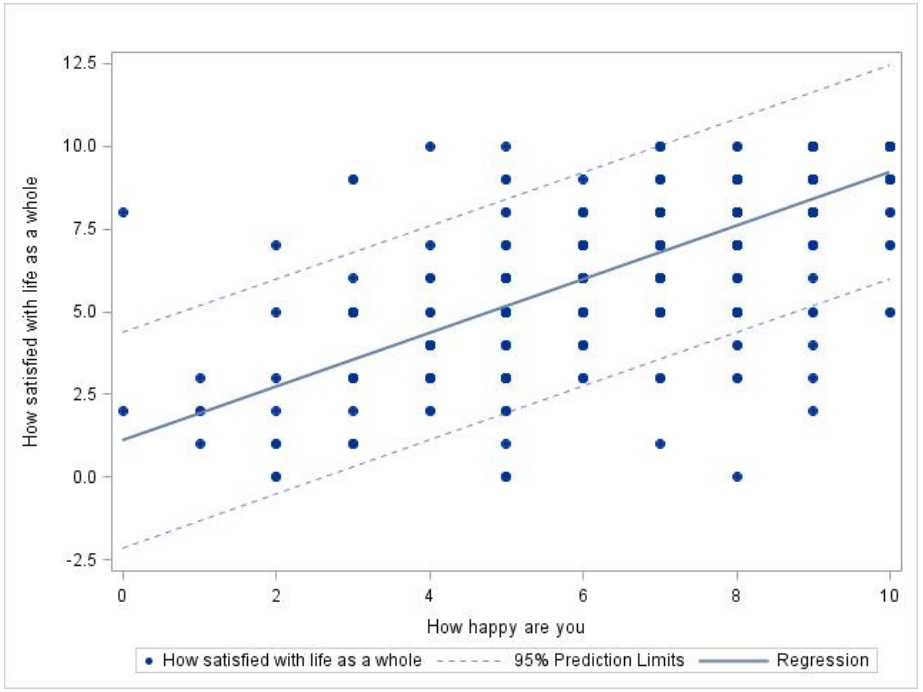
Graphical editing

Graphical editing is often useful, since it helps in observing outliers that can be out-errors, or in-errors. There exists nowadays more and more multidimensional statistical graphics that can identify possible errors. Unfortunately, it is not still easy to check whether just these values are erroneous or some others connected with them.

The scatter plot is a simple example from the ESS.

x-axis is for the happiness and y-axis for the life satisfaction that are well correlated (coefficient of correlation = 0.71).

The graph includes also the 95% confidence intervals of the linear predictions. It is obvious that the values outside them are erroneous.



Tabular editing

Here is the previous page graph in tabular form that may help when thinking whether to revise some values or not or how to handle further to analysis

Table of happy by stflife												
happy(How happy are you)	stflife(How satisfied with life as a whole)											
	0	1	2	3	4	5	6	7	8	9	10	Total
0	262 0.51	40 0.08	21 0.04	21 0.04	12 0.02	19 0.04	5 0.01	4 0.01	6 0.01	4 0.01	14 0.03	408 0.79
1	88 0.17	143 0.28	61 0.12	44 0.09	17 0.03	29 0.06	16 0.03	7 0.01	8 0.02	7 0.01	5 0.01	425 0.82
2	116 0.23	115 0.22	227 0.44	129 0.25	60 0.12	68 0.13	27 0.05	18 0.03	18 0.03	15 0.03	5 0.01	798 1.55
3	105 0.20	120 0.23	299 0.58	487 0.95	216 0.42	195 0.38	65 0.13	47 0.09	44 0.09	20 0.04	12 0.02	1610 3.12
4	57 0.11	65 0.13	191 0.37	407 0.79	532 1.03	356 0.69	134 0.26	107 0.21	49 0.10	20 0.04	16 0.03	1934 3.75
5	193 0.37	128 0.25	298 0.58	565 1.10	714 1.39	2135 4.14	694 1.35	485 0.94	281 0.55	89 0.17	123 0.24	5705 11.07
6	45 0.09	31 0.06	98 0.19	260 0.50	356 0.69	1057 2.05	1447 2.81	962 1.87	432 0.84	85 0.16	70 0.14	4843 9.40
7	36 0.07	35 0.07	93 0.18	232 0.45	314 0.61	948 1.84	1430 2.78	3487 6.77	1892 3.67	416 0.81	168 0.33	9051 17.56
8	42 0.08	28 0.05	58 0.11	138 0.27	178 0.35	640 1.24	706 1.37	2406 4.67	6421 12.46	1975 3.83	674 1.31	13266 25.74
9	24 0.05	23 0.04	18 0.03	57 0.11	51 0.10	179 0.35	188 0.36	499 0.97	2050 3.98	3899 7.57	1226 2.38	8214 15.94
10	46 0.09	16 0.03	31 0.06	58 0.11	57 0.11	228 0.44	120 0.23	263 0.51	585 1.14	817 1.59	3055 5.93	5276 10.24
Total	1014 1.97	744 1.44	1395 2.71	2398 4.65	2507 4.87	5854 11.36	4832 9.38	8285 16.08	11786 22.87	7347 14.26	5368 10.42	51530 100.00

Handling screening data in editing

It is good to create the two variables if on filter is used: one for the entire target population, and the other for the restricted one with the filter variable = yes. The table is from this case and continues from the example in the questionnaire designing part B.

Car stolen in the last 5 years			
	Prevalence, %		Number of stolen cars
	Un-weighted	Adjust-ment weights	Adjust-ment weights
The entire target population	2,25	2,34	268174
Households with one or more cars	2,55	2,66	268174

Editing not always complete in public use data

The ordinary editing is such that the range of individual values of each variable is checked and they are plausible. It is more difficult to edit the data conditionally, thus checking that the values of two or more variables are plausible. It means the end-user should make this checking as him/herself and thus to continue the editing if implausible values are found.

The next page example is from the ESS in which two variables are checked, thus checked how logical the values are.

As you see, certain values are not logical.

Cross-tabulation of the ESS data by TV watching

TV watching, total time on average weekday)	TV watching, news/politics/current affairs on average weekday)												
	0 Not at all	1 Less than 0.5 hours	2 0.5 to 1 hour	3 More than 1 hour up to 1.5 hours	4 More than 1.5 hours up to 2 hours	5 More than 2 hours up to 2.5 hours	6 More than 2.5 hours up to 3 hours	7 More than 3hours	66 Not appli-cable	77 Refusal	88 Don't know	99 No answer	Total
0	0	0	0	0	0	0	0	0	2342	0	0	0	2342
1	664	2204	85	24	14	3	3	9	0	1	15	0	3022
2	826	3350	2552	84	25	5	11	13	0	0	16	0	6882
3	605	2887	2867	948	58	17	5	8	0	0	14	4	7413
4	617	2744	3670	1075	581	26	15	8	0	0	21	2	8759
5	426	1784	2882	1283	489	313	31	16	0	0	15	2	7241
6	330	1378	2605	1225	618	257	277	27	0	0	14	2	6733
7	688	1801	3780	2272	1375	663	454	971	0	0	45	4	12053
77	0	0	4	0	0	0	0	0	0	1	0	0	5
88	20	46	39	11	6	2	0	4	0	0	57	0	185
99	2	6	11	2	1	1	0	1	1	0	0	13	38
Total	4178	16200	18495	6924	3167	1287	796	1057	2343	2	197	27	54673

We can wonder several things as for example:

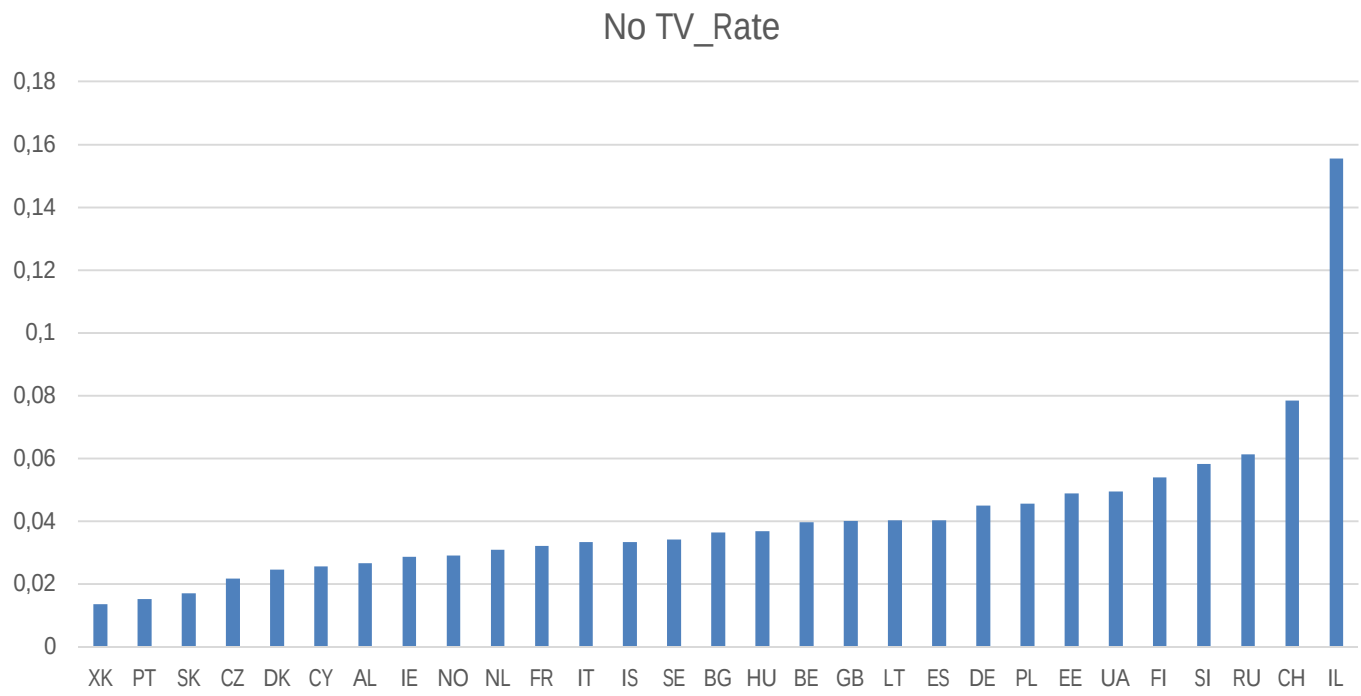
- Upper diagonal positive values are not logical
- What is 'Not applicable' since it is only in column; it is possible that they have no TV.
- Don't know for total TV watching looks strange if the answer is for TV news etc watching

What to do in editing for TV watching variables?

It is at least possible to change all values for TV total watching at the same level as that for TV news etc. watching.

It is possible to use specific categorization based on the subject matter targets of the study so that the missingness codes are in the group ‘Others’. If ‘Not applicable’ means that they have no access to TV, this could be a specific small scale study. I checked what this gives. The result in on next page.

ESS countries by the answer ‘Not applicable’ rate (No TV rate) for the question on watching TV news and politics and current affairs. The rate for Israel is fairly high.



Imputation

What is imputation, its purpose, concepts
Most common tools for missing item handling without real imputations
Missingness pattern
Targets for imputation
Imputation process
Imputation model
Imputation task
Single and multiple imputation (yksikertainen ja monikertainen imputointi)
Imputation model plus Imputation task in the case of the linear regression model
Imputation model plus Imputation task In the case of the response indicator model
Preserving associations in the case of missing data
General conclusion

What is imputation?

It is to insert a value into the data in a more or less fabricated way ('best proxy'). 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 this is desired to replace with a good unique value e.g. for distribution purposes.
- Since the existing value does not seem to be correct, and consequently, it is desired to get a more reliable value by replacing with a more plausible value.
- Since the current value seems to be too confidential, that is, and this individual unit should be disclosed. Motivation: the fabricated (imputed) value can be considered as less problematic even when told that it is no true value.

Imputation can be performed both for the macro and micro data but here 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, i.e. simpler methods are enough.

Purpose of imputation

To repeat: The purpose of imputation is twofold

-Either to replace a missing or partially missing or incorrect value with a such value that the estimate derived from this variable will be more valuable than without imputation. Thus if imputation is advantageous from an estimation point 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! You thus have to convince yourself or your client that imputation improves something.

- 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.

Information requirements for imputation

If any explanatory variable (auxiliary variable, covariate) does not exist, imputation can only be random based, i.e. guessing randomly missing values. This rarely works, but usually, it is needed auxiliary variables that predict missingness as well as possible. You can look the two previous pages and see which variables can be used, i.e. such variables that are non-missing. In panels or longitudinal data there are more such variables since e.g. the variables of the previous waves are available (a problem is still the fact that this variable may have been changed).

What can be imputed due to missingness?

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

(i) Under-coverage that requires a new up-to-date frame. Very seldom possible.
(ii) Those units that are not selected into the sample. Done in theoretical (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 earlier imputed correctly enough).

Most common tools for missing item handling without real imputation

(i) In the case of mass missingness, the weighting or the reweighting 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 in estimation). 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-non-response is marked with a good and well-covered code, e.g.:

- -1 = respondent candidate not contacted (a problem here may be that we do not know whether this unit belongs to the target population). These cases are rarely imputed.
- -2 = respondent refused to answer (main reason for imputation)
- -3 = respondent was not able to give a correct answer (possible to impute)
- -4 = missing for other reasons
- -6 = question was not asked from the respondent (imputation using logical rules, see below how to handle this screening case)
- -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)!

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. When going to impute, it is good to try a different imputation technique for each missingness code, since the nature of these units is different. I think that this is rarely applied in this way.

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 linearized models, among others. But if desired to use it as continuous, real imputation is required.

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 may vary by variable.

(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. It is called ‘case deletion.’ In think that this is still a fairly common strategy.

(v) Pair-wise analysis for multivariate purposes in such cases where e.g. the correlations are the basis for further analysis. This operation first computes pair-wise correlations like in case (iii) and when continues from the correlation matrix towards multivariate analysis. We lose less information here than in (iv).

Example: Item non-response

It is useful before imputation to examine how nonresponse vary. It is good to compute item nonresponse rates for a group of variables used in the analysis. Here is an example using the European Social Survey and selecting certain different variables intended to use in further analysis. We thus first create the new item response indicators for these five variables:

```
if happy>10 then happy_resp=0; else happy_resp=1;
if stflife>10 then stflife_resp=0; else stflife_resp=1;
if HINCTNTA>10 then HINCTNTA_resp=0; else HINCTNTA_resp=1;
if hincfel>4 then hincfel_resp=0; else hincfel_resp=1;
if imdfetn>4 then imdfetn_resp=0; else imdfetn_resp=1;
```

It is easiest to get the basic figures from these rates by calculating the means. The objective income HINCTNT has the lowest response rate, the variable imdfetn (allow immigrants from countries with different ethnic group to come) has the second lowest but fairly high, and nearly as high as for subjective income hincfel, for happiness and life satisfaction.

Variable	N	Mean
happy_resp	52177	0.9919696
stflife_resp	52177	0.9942120
HINCTNTA_resp	52177	0.7989344
hincfel_resp	52177	0.9873316
imdfetn_resp	52177	0.9569159

These rates are one-dimensional but it is often good to learn about missingness multidimensionally. In this case the pattern could be calculated by cross-classifying all response indicators. This leads to the item non-response pattern of these five variables.

Item non-response pattern of the five variables

	happy_resp	stflife_resp	HINCTNTA_resp	hincfel_resp	imdfetn_resp	Frequency Count	Percent of Total Frequency
1	0	0	0	0	0	5	0.0095827664
2	0	0	0	0	1	2	0.0038331065
3	0	0	0	1	0	7	0.0134158729
4	0	0	0	1	1	17	0.0325814056
5	0	0	1	0	0	3	0.0057496598
6	0	0	1	0	1	1	0.0019165533
7	0	0	1	1	0	19	0.0364145121
8	0	0	1	1	1	20	0.0383310654
9	0	1	0	0	0	7	0.0134158729
10	0	1	0	0	1	11	0.021082086
11	0	1	0	1	0	22	0.042164172
12	0	1	0	1	1	61	0.1169097495
13	0	1	1	0	0	1	0.0019165533
14	0	1	1	0	1	1	0.0019165533
15	0	1	1	1	0	73	0.1399083888
16	0	1	1	1	1	169	0.3238975027
17	1	0	0	0	0	5	0.0095827664
18	1	0	0	0	1	11	0.021082086
19	1	0	0	1	0	23	0.0440807252
20	1	0	0	1	1	65	0.1245759626
21	1	0	1	0	0	1	0.0019165533
22	1	0	1	0	1	1	0.0019165533
23	1	0	1	1	0	28	0.0536634916
24	1	0	1	1	1	94	0.1801560074
25	1	1	0	0	0	57	0.1092435364
26	1	1	0	0	1	446	0.8547827587
27	1	1	0	1	0	574	1.1001015773
28	1	1	0	1	1	9178	17.590125918
29	1	1	1	0	0	8	0.0153324262
30	1	1	1	0	1	101	0.1935718803
31	1	1	1	1	0	1415	2.7119228779
32	1	1	1	1	1	39751	76.18490906

How to impute if decided to impute when looking for the pattern

The variable HINCTNTA has most missing values. If this variable is important in analysis, it is good to try to impute at least some of its missing values. The row 28 could be important since the other variables here are complete and the missingness rate of this income group is high = 17.6 per cent. For imputing these values, there are all four auxiliary variables available, and many others. The entire response rate would increase from 76.2 per cent to 83.8 per cent that would be fairly high. This strategy thus started from a high item nonresponse, but it may be easy to start to impute the cases with low non-response as well, or a compromise of both strategies. Think your strategy for imputing missing values of this pattern.

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/true values. Success at individual level. Another point is that how to know how close they are, except in some cases. This may be often a too demanding 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 satisfactory and specifically in statistical institutes or in other survey institutes where such estimates as average, total, ratio, median, point of decile and standard deviation are typical.

(iv) Some users hope to get the order of imputed values as correctly as possible.

(v) Finally, success to preserve associations (like correlations) is also important in many studies.

The summary: it is most important to keep in mind the end use of the data set after imputation as well.

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 exploit 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.

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.

A strategy: First, try to impute using the first alternative as well as possible = logical imputation, and second, to impute using the second alternative the rest; naturally if you will impute at all.

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 (i) can cover all possible forms, categorical including binary and continuous but in case (ii) the dependent 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 the different units.

Note that in sequential imputation the number of non-respondents (missing value units) will be declining from one imputation to the next. In order to work well in this imputation, individual level success is important or such aggregate level that is important.

Imputation model 4

The model (i) is concerned a continuous variable

In this case the most common model is linear regression or its logarithmic version. Recently also mixed models are going to be applied and these models may be better than linear if the measurements are from two levels for example.

Regression models are easy to use and also the model fit (*R-square*) is a good indicator and it is good to look when searching for best auxiliary variables or covariates in the model specification phase. This will be the first real operation when going to imputation. Its result can be used in the imputation models (ii) as well. It is useful also for comparing different methods with each other.

Imputation model 5

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 either logit, probit. complementary log-log or log-log as discussed in Part C.

There are no dramatic differences in explaining models between those link functions but some of course. 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 6

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 modeling*. If the dependent variable is categorical, we speak about *classification trees* (*random forests* is its newer version), 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 or post-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. Imputation cells can be constructed with 'smart thinking', e.g. the
model (i) or (ii) can be estimated two times by gender if thought that the predictions vary by gender. Or regions and age groups can be good as well.

Imputation model 7

Both types of imputation models thus have been estimated in a best way in the sense that it predicts well so that the final target is imputation. The guru's of imputations have said that the imputation model should have a good predictability feature that is not necessarily easy to know what this means. We can say that this means at least that it is not necessary to concentrate on a model that is explaining well the dependent variable of the multivariate model. Naturally, it may be good if the estimated model coefficients of the explanatory (auxiliary) variables or covariates can be interpreted well since it helps in explaining for clients or reviewers why imputation is obviously working well. Keep still in mind the predictability. Hence we have to get the predicted values of the models before going on to the next step, imputation task.

Concluding points about imputation models

The predicted values will have a big role when going to impute, that is, in the stage of the imputation task. The big point is that the predicted values should be available both for the respondents and for the non-respondents, i.e. the auxiliary variables should be complete. All the previous predictions can be attempted. We have observed that there are many similarities but also essential differences and we cannot say surely which method is finally going to be the best if this will be found any way. However, it is expected that some methods are not good although used in real life. If the imputation model would be strong, that is, it is predicting well, most imputation task choices work quite well. Thus it does not matter which imputation task uses. But a usual real life application is not as easy and the imputation model thus does not fit very well. Nevertheless, imputations are good to perform.

Imputation task

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

(a) Model-donor approach (malliluovuttaja) in which case the imputed values are computed deterministically (or stochastically) from the predicted values (adding noise) of the model.
(b) Real-donor approach (vastaajaluovuttaja) in which case the predicted values (or with 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 (or they can be converted to possible values after preliminary imputation).

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

Imputation task 3

Comment:
You will find from imputation literature the term ‘hot deck’ or ‘hot decking.’ This mystic term is derived from 1950’s, I think, when certain US surveyors randomly selected a donor from the observed values. This looked like ‘a hot deck’ in which those donors were moving their place and suddenly one was selected to replace a missing value. I do not like this term. It is historical and it is good to know origin. Later, I think, the term has been used also even though the donor selection is not random. E.g. when these real-donors are sorted in a certain order. The title of my 2000 paper was e.g. ‘Regression-based nearest neighbor hot decking,’ but now this method could be ‘Nearest neighbor real-donor imputation when the imputation model is linear regression.’

We thus see that there is needed a certain near or nearest neighbor metrics for selecting a best donor whose observed value to be borrowed for imputing. We proceed to more details soon of this metrics.

Imputation task 4

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 is still required 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 violate your estimates.

Imputation task 5

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 here 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.

Post-Editing after the model-donor method possibly

As known, the real-donor methods give observed values that are (or should) valid values. Hence nothing needed to do before the use of the with imputations.

But the model-donor imputed values thus are calculated and it is guaranteed that they are valid in all meanings. Sometimes they can still be used as such, but not always. Some examples:

- If we for instance wish to impute happiness that obtains the integer values from 0 to 10. When using model-donor methods, the imputes will be in most cases in decimal values. Any user does not accept it. A simple solution and sometimes used is to round them to integers, but it is not necessarily any best solution.

Post-Editing after the model-donor method possibly 2

The variable HAPPY thus is categorical but in the cases of a real continuous variable, the post-editing can also be important but its influence in the final results is not necessarily big. However, most clients do not like e.g. incomes with several decimals as can be obtained using model-donor imputation. Such values also indicate clearly for an expert that these are imputed. Thus: if the confidentiality is important as is often, a rounding is a good solution but what is the best rounding? It is not clear. The mathematical rounding is not ideal but such statistical that makes rounding probabilistically.
For example, if the value is 4555.7 and the rounded values should in tens, there is the probability for rounding into 4550 = (10-5.7)/10 and probability for rounding to 4560 =(10-4.3)/10, respectively.

Nearness metrics of real-donor methods

The most common metrics is derived from the predicted values of the binary regression model (thus the link function should be chosen by the user). In the case of a stochastic selection, some random noise is needed to add but there are different options for this. We do not go to their details, but I want to mention a common tool from the Imputation book by Rubin:

- Classify the predicted values into a certain number of categories by their values, e.g. 10 to 20 categories, called imputation cells. These are fairly homogeneous and thus enough close to each other.
- Select randomly within each cell one observed value to replace a missing value. This method is called sometimes cell-based random hot deck.

The observations of this kind of imputation cells are called also 'donor pools.' There thus is a pool where to go to borrow a good value to replace a missing value. It is maybe good to create such donor pools in advance for imputing but the values of this pool should be from the same period at minimum.

Nearness metrics of real-donor methods 2

The other rational strategy in many situations is to use model-donor imputation values (that are predicted values of a regression model e.g.) over both the respondents and the non-respondents as the nearness metrics. This thus means that we impute technically the values for the respondents too, using the same strategy as for the non-respondents. It is not difficult. 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 randomize the procedure. Thus e.g. our nearness metrics is the previous model-donor output:

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

Nearness metrics of real-donor methods 3

To make the previous point “Thus e.g. our nearness metrics can be the previous model-donor output” more clear:

We can thus work so that we first perform imputations using model-donor methodology but in this case also for the respondents (observed units) in addition to the non-respondents (not observed). Now we have the nearness metrics that is used – to find the nearest neighbor (or a reasonably near neighbor) for each non-respondent from the respondents and

- to insert this value to this unit.

This also gives opportunity to compare both strategies easily when estimating some figures from the imputed data set.

Nearness metrics of real-donor methods 4

The imputed value of the real-donor method.

If the imputation model is based on the missingness/response indicator, the imputation is similar to that presented in previous pages, but now the values of the nearness metrics are thus within the interval (0,1). Now we have automatically these propensity values both for the respondents and for the non-respondents. There are still several options to work with these values. An interesting special case is such in which the variable being imputed is binary as well. Thus both variables (in imputation model and in analysis) are binary. This may arise confusion.

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. The rules are not always clear and hence criticized.

MI is in certain problems difficult to realize 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. Many details of MI are considered in the specific section of this course. MI is usually based on a Bayesian approach that is developed by Don Rubin (US), but non-Bayesian (called also repeated MI) is also used that I will prefer so far. Jan Björnstad (Norway) introduced this concept in 2007 (J. of Official Statistics).

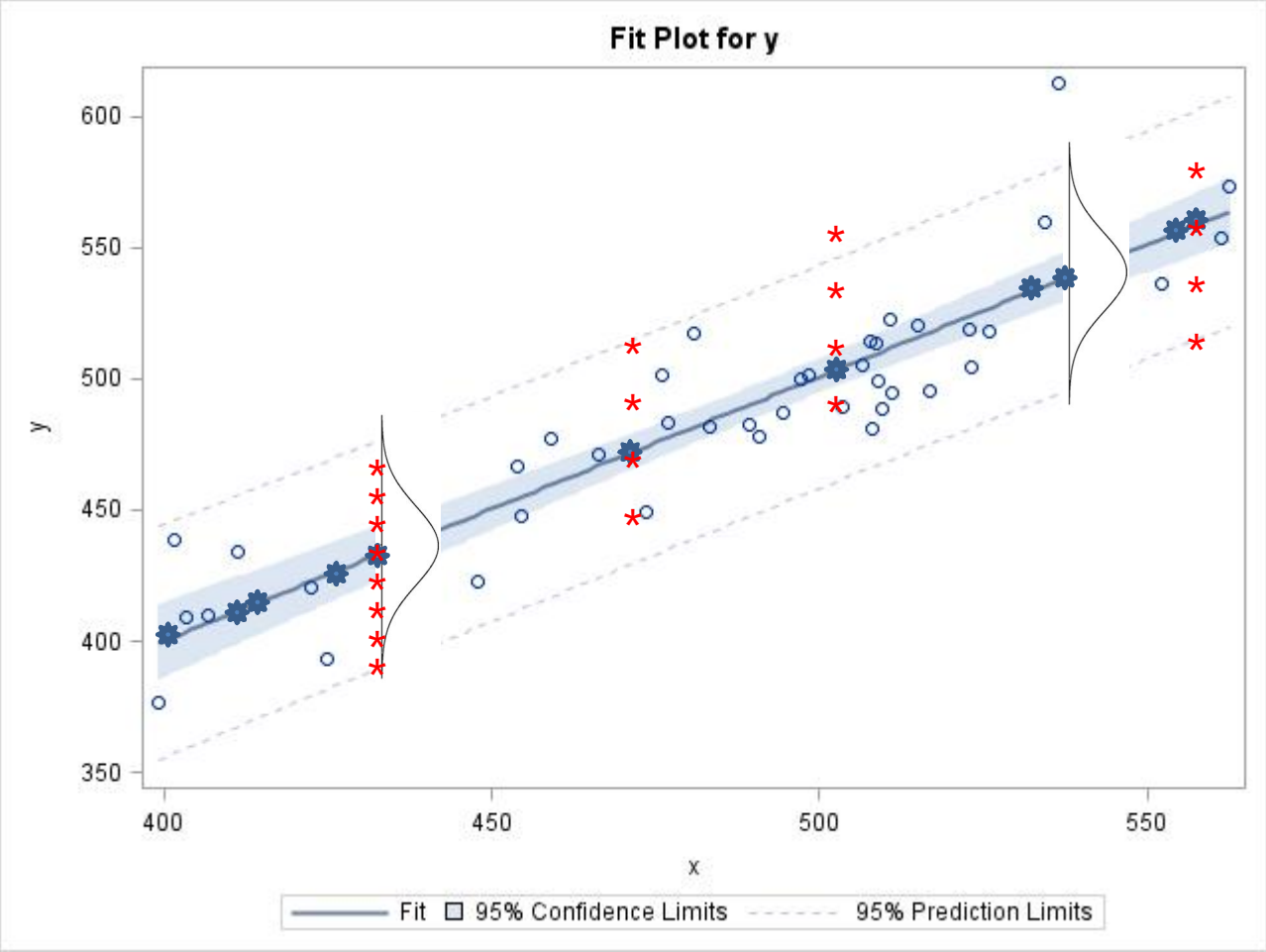
Summary: Imputation model plus Imputation task in the case of the linear regression model

	Deterministic Single	Stochastic Single Multiple
Model-Donor	A. Regression model estimated and its predicted values are used as imputed values for missing items	C. Adding to the A model the normally distributed random numbers with the zero mean and with the Root_Mean-Square_ Error standard deviation. Or to add observed residuals.
Real-Donor	B. Regression model as in A but those predicted values are computed both for the respondents and for the non-respondents but now these are used as a nearness metrics.	D. Like B but applying to the C model. Multiple imputation by using several seeds for random numbers. This is concerned C too.

Example, why and how to get adding to the A model the normally distributed random numbers with the zero mean and with the Root-Mean-Square-Error (RMSE) standard deviation.

This thus is derived from the model uncertainty (non fitting) that is simply measured by the residual and its standard deviation. As said above if assumed that they are normally distributed, it is possible that some ‘residuals’ are too big (i.e. above any observed residual): it that case it is good to think whether to truncate them.

Illustration of the model-donor imputation with a simple regression. The random noise term N(0, RMSE) is added to the predicted values. It is a danger that the imputes are outside the plausible limits.



- A predicted value = Deterministic impute

* A possible impute with noise

y = imputed if missing
x = auxiliary variable

Single and multiple imputation
Technics

Let
L = number of imputations *u*,
Θ = parameter being estimated,
and its point-estimate = *Q* (e.g. mean income and CV)
and variance estimate, respectively, = *B*
And then standard error of the mean = square root of the variance.

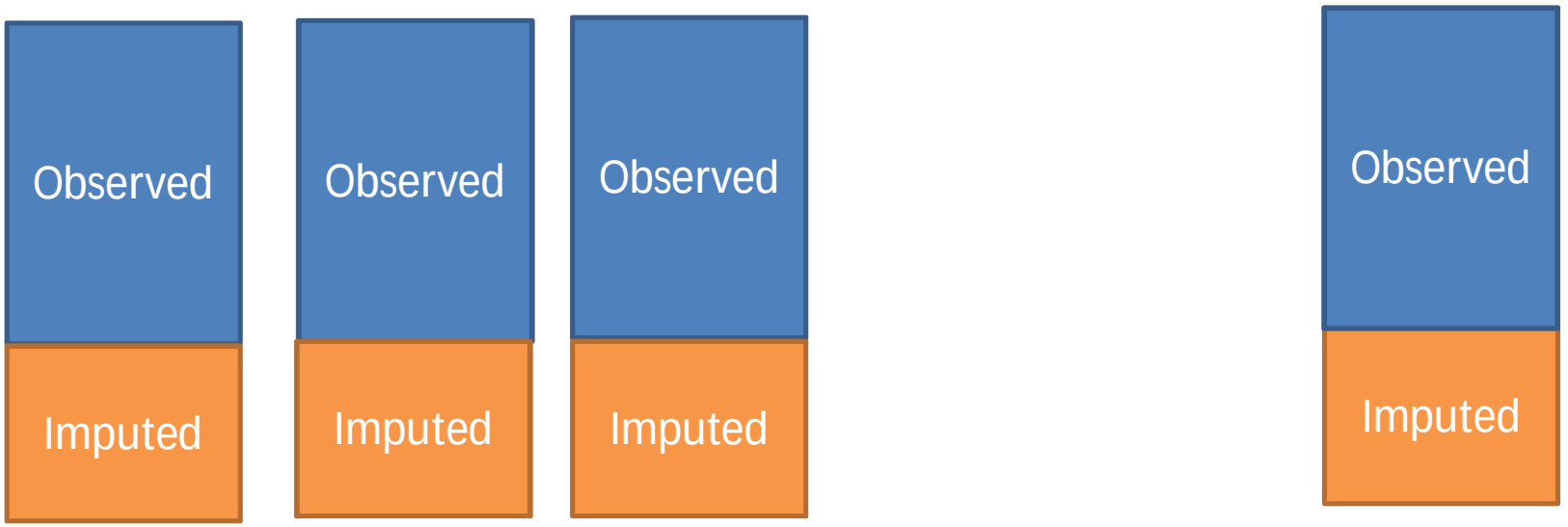
All these are calculated as usually so that the imputed values are included as such. The estimate may be whatever such as average, total, ratio, proportion, median, percentile, regression coefficient.

The number of imputations = *L* is in Rubin's initial book even as small as 3, but may work only with simple data sets. I think that *L*>=10 could be best to use in practice. Rubin's *L=3* is understood if remembering how inefficient the computers were in 1980's.

Single and multiple imputation 2

A simplified illustration of *L* single data sets with imputations (complete data)

Simple because the fractions of missing values may vary by variables. Here all have the same fraction.



Point and interval estimates from each data set

Single and multiple imputation 3

Now the multiply-imputed point-estimate is a simple average of multiply imputed estimates

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

Respectively, the variance can be calculated as the average of the variances of *L* complete data sets in which each variance is estimated using the formula that is valid for the sampling design of the survey. This is for the gross sample data set that also includes the units that are not needed to impute. But because a certain number is missing these are imputed and the average and the variance are calculated in a best way thus.

$$B = \frac{\sum_u B_u}{L}$$

Single and multiple imputation 4

The variance estimate is respectively $B_{MI} = \frac{\sum_u B_u}{L} + (k + \frac{1}{L}) \frac{1}{L-1} \sum_u (Q_u - Q_{MI})^2 =$

$k = \frac{1}{1-f}$ *f* = the fraction of missing and imputed values

If *k=1* or *f=0*, it is Rubin's formula, otherwise Björnstad's formula.

You see that the entire variance consists of the two components: (i) the average of variances (within-variance) and (ii) the between-variance that indicates how much multiply imputed estimates vary. If the variation is zero, this between-variance is zero too.

It is good to remind that multiple imputation is not any own imputation method but it consists of several single imputations. If single imputation is not working, multiple imputation is not either working. Some authors, unfortunately, are not speaking in this way. 'Multiple' requires thus a stochastic element.

Single and multiple imputation 5

The initial multiple imputation was developed by Donald Rubin. It was based on the Bayesian theory. This theory thus was reformulated by the Norwegian Jan Björnstad. A reason was that Rubin's strategy is not well working in many practical situations like in statistical offices. Hence he uses the term non-Bayesian.

It is not the only difference in these frameworks. The Bayesians use certain Bayesian rules in all imputation methods. Instead, the non-Bayesian framework uses simpler rules. A big question follows from this: How good are these frameworks in practice?
And are the Bayesian rules really useful and better? Note that these rules are developed by Rubin and a user thus have to trust in him or his specifications.

Specialities for imputation of a categorical variable

This same framework is workable for categorical variables as well but the

	(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

Alternatives of the first row are automatically different since the imputation model can not be ideally any linear regression model.

Fortunately, when using the binary missingness indicator as the dependent variable, the imputation task can be exactly similar as in the case of a continuous variable. That is, use the same nearness metrics in imputing missing values as above.

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 it does not matter if you do not like to publish this paper.
(ii) Try to use such analysis method that takes missingness into account (the Nobel winner economist Heckman has developed a much cited strategy).
(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. Note that if the first imputation is not good, the next one may be worse, etc. but try nevertheless.

End comments

This ‘story’ covers my approach to imputation. Many things have also been trained and concretized respectively. I hope that you will keep in mind these principles.

An alternative could be to use ‘a black box software’ package (SAS, SPSS, …) that gives your imputed values rather automatically. I would not be happy with such ‘boxes’ when working with real data since a client or a reviewer is demanding and not without convincing statements believe all completed data.