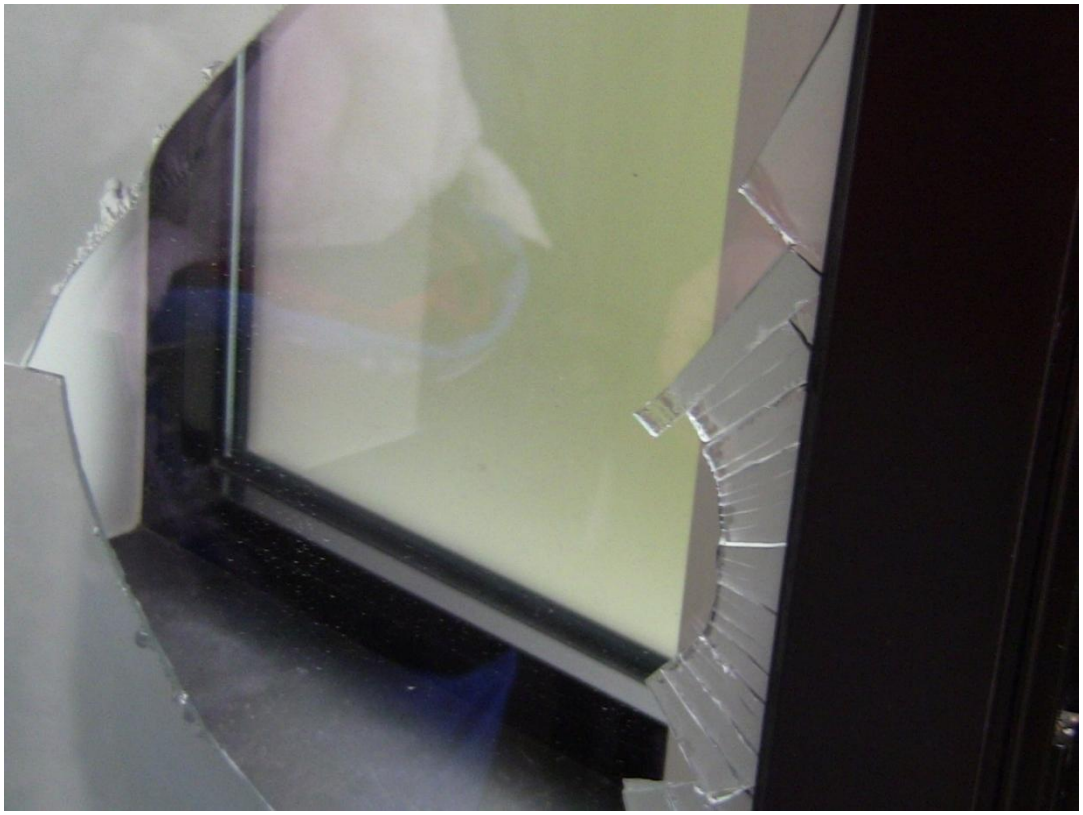
I M P U T A T I O N M E T H O D S

University of Helsinki 2017

Seppo Laaksonen

Missingness due to activity and bad luck



https://moodle.helsinki.fi/course/view.php?id=24026

Code: Imputation2017

Content of the introduction

What is imputation, its purpose, concepts
Missingness (puuttuneisuus) mechanisms
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 mod
Multiple Imputation with SPSS
Preserving associations in the case of missing data
General conclusion

After the introduction we go to details of each method.
And we will have Training all the time.

There are basically three techniques to deal with nonresponse or missing data:

(i) Weighting and reweighting
(ii) Analysis so that missingness has been taken into account by modelling
(iii) Imputation.

This course is on the last one but it is good to keep in mind the other alternatives.

The main ‘competitor’ of imputation is obviously ‘data deletion’ that could be considered as the base line method of imputations.

In this case the observed values only are used in analysis. In one-dimensional analysis we drop out the missing values of this particular variable, but the ordinary multivariate analysis includes such statistical units whose variable values are completely observed. Hence the data might reduce dramatically. Data deletion works only if the response mechanism is MCAR (Missing Completely at Random) although the standard errors (confidence intervals) will increase.

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 to replace this.
- 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 non-problematic but it is good to tell that it is no true value while the estimates can be trusted.

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, 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.

Use of imputation has increased

- 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 data institutions are still needed to do using a specific programming. Some methods are fortunately easy to program, some others not. Most methods I will present are not difficult to perform with SAS codes that I use.

- Imputation research was flourishing in 1990's and early 2000's but recently very little new things have been invented. Interestingly, the results of the Euredit project (http://www.cs.york.ac.uk/euredit/) are still useful. This project in which I was involved, tested a big number of imputation techniques, called traditional and new methods, respectively. I will concentrate mainly on traditional methods. Since new projects have been missing, less new ideas have been developed but certain techniques have been however implemented in software, like SAS MI, SPSS, Solas, MICE and R packages. Any general imputation software do not exist. Thus if imputation is wished to use, the understanding of its methodology is necessary. Do not believe any automatic software even though you might get results without problems.

One missing or other inappropriate value can be imputed once that is called single imputation (SI), or many times leading to varying imputes that is called multiple imputation (MI). We first present single imputation methods while multiple imputations after that.

We do not concentrate on imputation due to confidentiality but mainly thus on replacing a missing value with a best possible proxy.

The big question is this.

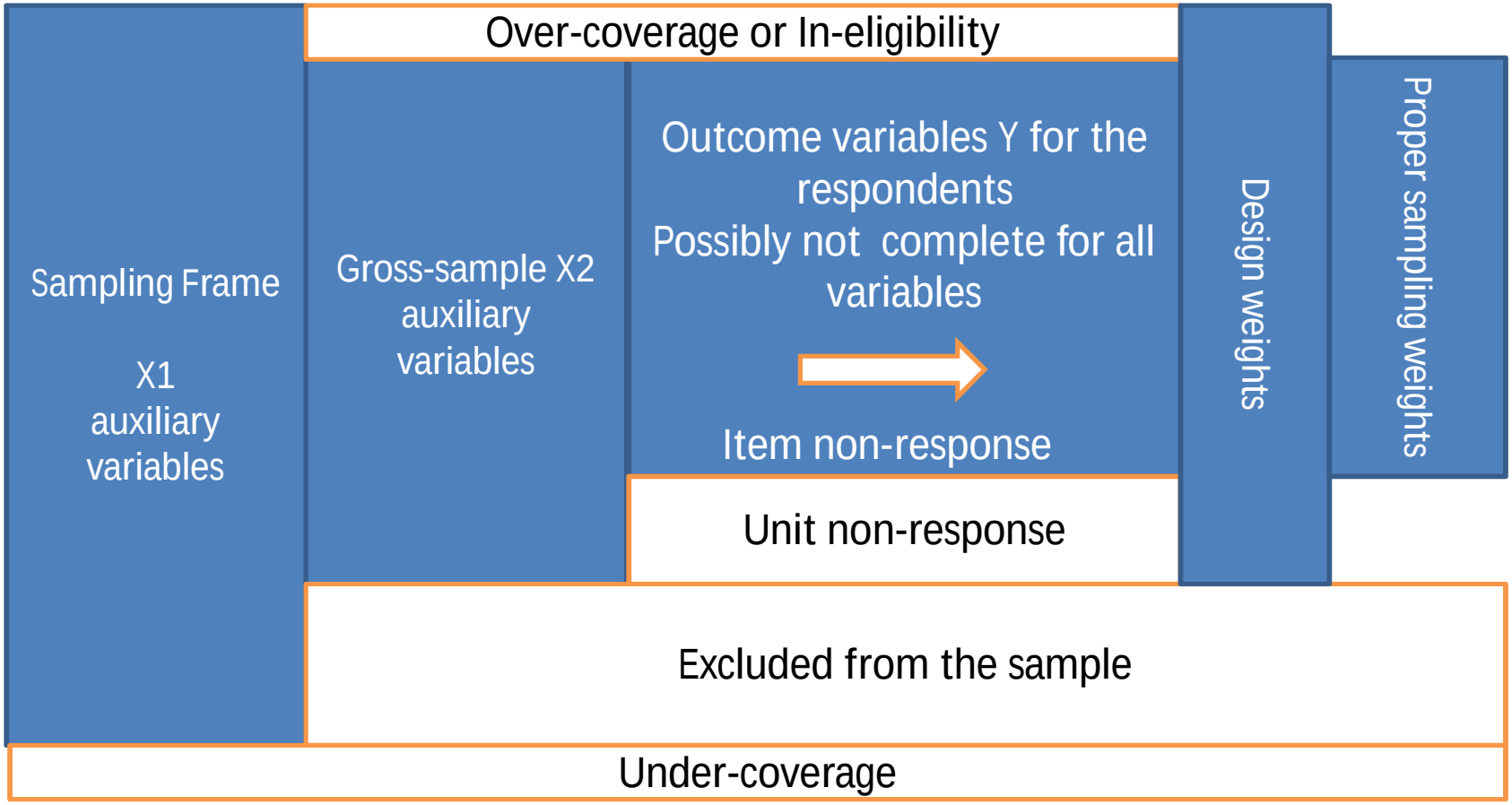
To impute or not to impute?

It is possible that one survey party is more willing to go to impute than another party. Imputation is however easier to do inside the survey institute that should take care of data quality. An outsider who is an end user at the same time, has still sometimes to impute if he/she is not otherwise happy. The reality is that the insiders have more auxiliary and other variables available also because some might be confidential and thus not possible to give outsiders. The insiders are also more familiar with the data process. However, the most important reason to impute is

The pattern of the imputed values should be as good that the estimate using this partially imputed variable will be more valuable than the data without imputation. Thus if imputation is advantageous from an estimation point of view, use it. This gives the certain requirements for the imputation methodology respectively.

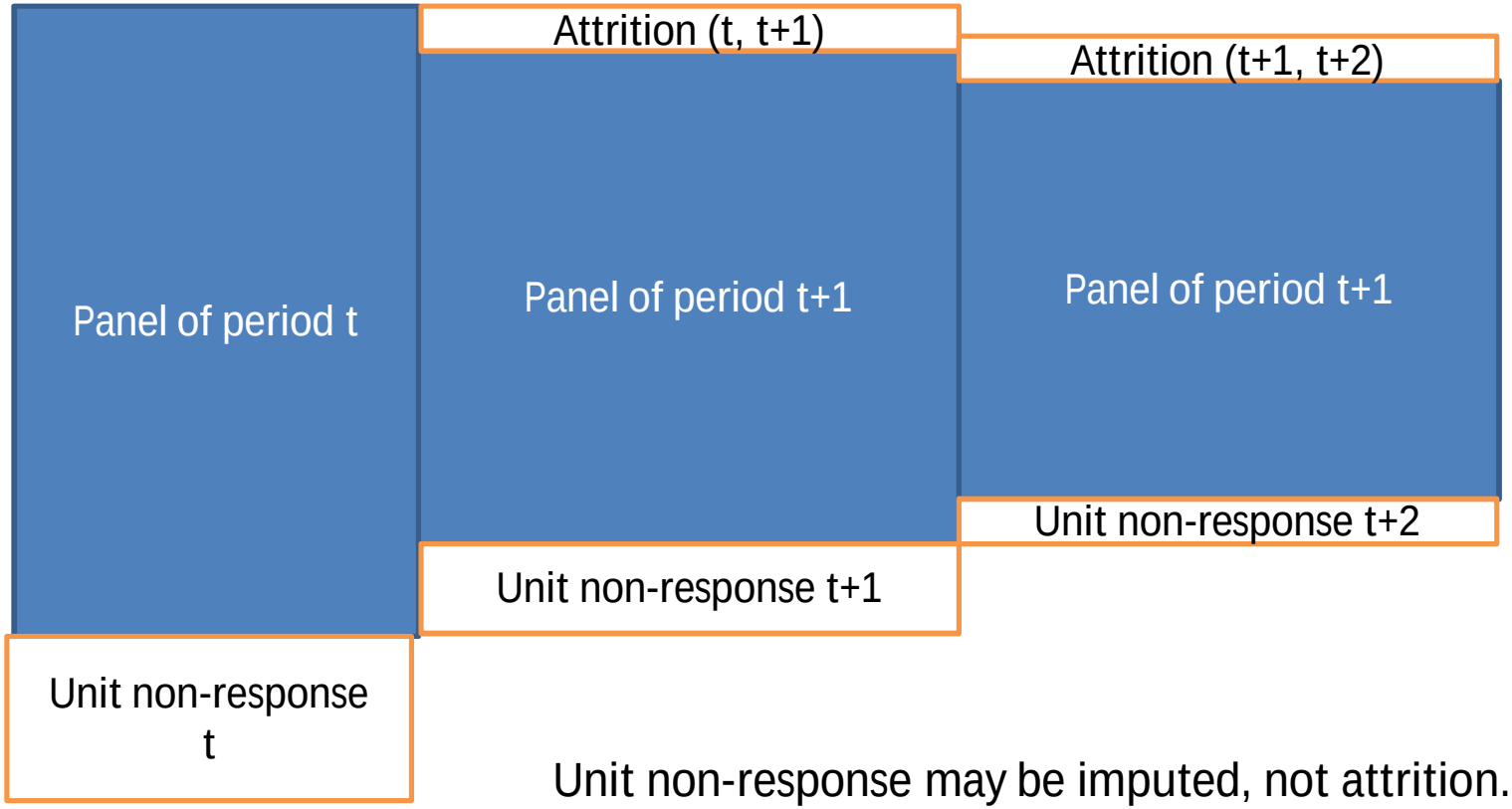
Micro data and Missingness

Now I focus on micro data where one can see various types of missingness. This is a cross-sectional case ('white boxes' are missing values and their values may be imputed using information from 'blue boxes'):



Micro data and Missingness

Cohort type of panel example



Unit non-response may be imputed, not attrition.

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. 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 might have been changed, and hence a new correct value should be known).

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 a pattern of 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).

Missingness mechanisms 1

Imputation requires useful auxiliary information. Without such data imputation is still possible but results are maybe bad. On the other hand, it is important to assess the missingness (or response) mechanism.
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. Simplest imputation methods follow this assumption.

Missingness mechanisms 2

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 or post-strata, or by other sub-groups.

MAR (Missing At Random (Conditionally)): Now missingness depends on both the sampling design variables and all possible other auxiliary variables that are often other survey variables without missing values (possibly since they are imputed using a good method). This assumption is much used when good auxiliary variables are available. It is a basic assumption in imputations when implementing an imputation model (later).

Missingness mechanisms 3

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, the quality of the estimates have been improved 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 lack of good auxiliaries (unfortunately).

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). Very seldom these cases are imputed.
- -2 = respondent refused to answer (main reason for imputation)
- -3 = respondent was not able to give a correct answer
- -4 = missing for other reasons
- -6 = question was not asked from the respondent (imputation using logical rules)
- -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 are different. I think that this is rarely applied in this way. Question: how to 'impute' cases with the codes -6 and -9?

Moreover, it is good to notice that the coded variable is full, without missing values (as imputed). This kind of a categorical variable can be used as an explanatory variable in standard linear and linearized models, among others. In this case, it is not good to impute the dependent variable. But if the variable is desired to use as continuous, proper 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. This strategy does not give consistent results with each other. Called ‘case deletion.’ In think that this is still a fairly common strategy.

Most common tools for missing item handling without real imputation 5

(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).

We cannot include these five cases in our training, but keep them in mind, and use if appropriate. One further strategy in modelling is not to include variables with high missingness rate.

Example: Item non-response

It is useful before imputation to examine how nonresponse vary. A pattern of missing values is good to compute. Here is an example using the European Social Survey Round 7 and selecting some different variables. The below example with SAS codes illustrate the computation, first for creating the item response indicators:

```
data ess7; set ess7b.ess7e02 ;

if hincfel<5 then sub_inc_resp=1; else sub_inc_resp=0;
if hinctnta<11 then income_resp=1; else income_resp=0;
if eisced<=5 then education_resp=1; else education_resp=0;
if happy<=10 then happy_resp=1; else happy_resp=0;
if imsmetn<=4 then immigration_resp=1; else immigration_resp=0;

run;
```

It is easiest to get the basic figures from these rates by calculating the means. The same might be concerned other variables in surveys like the satisfaction in job in which case whose not working is excluded.

The SAS System

The MEANS Procedure

Variable	N	Mean
sub_inc_resp	92362	0.9887724
income_resp	92362	0.8187999
education_resp	92362	0.7707174
happy_resp	92362	0.9936229
immigration_resp	92362	0.9662632

These rates are one-dimensional but it is often good to know the same multidimensionally. In this case the pattern could be calculated as the next page shows.

Using the SAS FREQ procedure is maybe the easiest way to get the whole pattern of the item response rates. The result can be seen from the file of ‘out=‘. In order to reduce the prints, ‘noprint’ option is used.

```
proc freq data=ess7; tables sub_inc_resp*
income_resp* education_resp* happy_resp*
immigration_resp
/noprint out=item_resp; Proc print; run;
```

The following page shows the ‘out’ file. If several variables is required to impute, this pattern helps in selecting the order. There is no definite order for this, but often it is good to start from variables that do not need many imputes and continue so that these imputed variables are used as covariates or auxiliary variables for the next variables being imputed. Another strategy is such in which best possible auxiliary variables can be used in each imputation. The compromise of both is maybe the ideal strategy. Think these questions but our practice later on does not include these sequential imputations.

Item non-response pattern for some variables

5! = 5*4*3*2*1=
120 = the maximum number of combinations
But now much less = 31

Obs	sub_inc_resp	income_resp	education_resp	happy_resp	immigration_resp	COUNT	PERCENT
1	0	0	0	0	0	6	0.0065
2	0	0	0	0	1	6	0.0065
3	0	0	0	1	0	19	0.0206
4	0	0	0	1	1	153	0.1657
5	0	0	1	0	0	9	0.0097
6	0	0	1	0	1	13	0.0141
7	0	0	1	1	0	67	0.0725
8	0	0	1	1	1	583	0.6312
9	0	1	0	0	1	2	0.0022
10	0	1	0	1	0	4	0.0043
11	0	1	0	1	1	35	0.0379
12	0	1	1	0	0	6	0.0065
13	0	1	1	0	1	1	0.0011
14	0	1	1	1	0	12	0.0130
15	0	1	1	1	1	121	0.1310
16	1	0	0	0	0	8	0.0087
17	1	0	0	0	1	42	0.0455
18	1	0	0	1	0	156	0.1689
19	1	0	0	1	1	2924	3.1658
20	1	0	1	0	0	26	0.0282
21	1	0	1	0	1	86	0.0931
22	1	0	1	1	0	624	0.6756
23	1	0	1	1	1	12014	13.0075
24	1	1	0	0	0	17	0.0184
25	1	1	0	0	1	54	0.0585
26	1	1	0	1	0	405	0.4385
27	1	1	0	1	1	17346	18.7805
28	1	1	1	0	0	83	0.0899
29	1	1	1	0	1	230	0.2490
30	1	1	1	1	0	1674	1.8124
31	1	1	1	1	1	55636	60.2369

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 NSI's or in other survey institutes where such estimates as average, total, ratio, median, point of decile and standard deviation are typical.

Targets for imputation should be specified clearly

(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.

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

It is not realistic succeed in all things. Hence I think that targets (ii) and (iii) are most important. This success is really possible if imputation is well done. Target (iii) is enough in official statistics concerning macro figures.

This is an example in which case imputation is successful at individual level, at least nearly. Necessary solution but not cheap. See the first page.

In our training

We are able to check everything since we know true values. Respectively we can look for the first three criteria, even though how well the imputation succeeds at individual level. For this purpose we compute the mean absolute error for the units with imputed values (the same can be done for the entire data but it is as illustrative.

$$MAE = \frac{\sum_{i=1}^{n-r} |y_{*i} - y_i|}{n - r}$$

in the formula y_* refers to an imputed value, y *to a respective true value,*

r = the number of the observed units and *n* = the number of all the units

The same in SAS codes (The data 'imp' include the imputed values):

```
data imp2; set  imp;
mae=abs(income_imp-income);
proc means n mean; var mae; run;
```

In our training 2

Our first variable being imputed is INCOME (yearly) but if you like you can look at variable HAPPY (from 0 to 10) as well, and later POOR (1=yes, 0=no). To make everything easier the missingness is equal for all although this is not realistic in real life.
The MAE is not however any very important indicator in our case since it is difficult to succeed well at individual level due to lack of excellent micro level auxiliary variables in our data set. This is usual in real life but sometimes it is possible to get a good tax variable that correlates well with income. The auxiliary variables for POOR are partially similar as for INCOME. Happiness is more unclear.

In our training 3

It is fortunate that the individual level success is not most important but the two other variable indicators instead:

- the average income

and

- income differences.

The latter one may be considered even more important than the average. Income differences can be measured by various indicators but the simplest is the coefficient of variation (CV). This basic statistic is well correlated with the Gini coefficient e.g. that is the mostly used. The income differences can be looked via different distributional statistics as well. The similar criteria are appropriate for happiness too.

Reasons and strategies for imputation

- The amount and impact of missing values without imputation. This is the most important practical question:

(i) If the missingness rate is high, let say above 50%, the data quality is in most cases bad with data deletion, and the users are unhappy. But if the missingness mechanism is ignorable, the results are obviously moderate. On the other hand, imputation would be easier in this case than in the case of a non-ignorable mechanism.

(ii) If the missingness rate is low, let say below 5%, but it has been found that one or more influential respondents are missing (e.g. big businesses in business surveys, or extremely rich people in income surveys), all possible should have been tried to do for improving the data quality. Imputation might be the only option.

(iii) The high missingness rate in categorical variables is not usually as awkward as in skew continuous variables given that the categories are determined optimally. For example, if the respondents with extremely high incomes are in the same category as the ordinary high income respondents, it is not fatal if some values are missing. On the other hand, these missing values can be fairly easily imputed into this high income category.

- Esthetic reasons in the sense, that the data file with an 'ugly' pattern of missing values does not convince users about the data quality at all. This can be a good point too, that is, if the quality is bad, the user might be more careful in his/her analysis. Of course, if the quality of imputations is high, it is the best thing.
- The worst strategy is to complete the data without taking care of the quality of imputations, and not telling at all which values are imputed.
- In all cases, the imputation methodology should be documented so that the user knows how much to trust in the data.

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 (internal and external variables are possible)

(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 be 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 its structure and core variables) to the present data.

Imputation model 2

The former 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 or deductive imputation so that the imputed value is true with high probability (E.g. if it is known the number of children and their ages, it is possible to logically impute fairly well the child benefit in Finland),
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 3

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 the variable being imputed.

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

Imputation model 4

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 that a categorical variable with the missingness codes may work reasonably in imputation.

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 5

The model (i) is concerned a continuous variable (as income in our first training).

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. In this course we do work with mixed models since our training data are from one level, i.e. it is concerned individuals.

Imputation model 6

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 7

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 or poor, 0 = unemployed or non-poor).
In the case of a binary model, predictions depend also on the link function used.
-logit
-probit
-complementary log-log
-log-log .
There are no dramatic differences in model estimates between those link functions but 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 within the interval (0, 1).

Imputation model 8

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 that seems to be popular), whereas the model for continuous variable is *regression tree*. In the case of Imputation model (ii), the classification tree is used.

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 model 9

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 assumed that the predictions vary by gender. Or regions and age groups can be good as well. If someone wishes to do so in our training that would be nice.

Imputation model 10

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 imputation guru's 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 trying explain well the dependent variable of the multivariate model, even though it is good.

Imputation model 11

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.

On next pages, I will give the basic SAS codes for both the linear regression model and for the binary regression model with the most common link functions, i.e. logit and probit.

Imputation model 12

SAS codes with predicted values in the output file

Symbols: y1 = continuous variable
y1_resp = response indicator of y1
x1, x2, x3, … = continuous auxiliary variables
z1, z2, z3, … = categorical auxiliary variables or those used categorically, * =interaction between two variables

<u>Linear regression</u>

```
proc glm data=a.impucomplete; class z1 z2 z3 z4 ; model
income2=z1 z2 z3 z3*z4 x1 x1*x1 /solution ;output
out=new p=predicted; run;
proc means n mean min max cv data=new; var predicted;
run;
```

Imputation model 13

SAS codes with predicted values in the output file

Logit regression or logistic regression

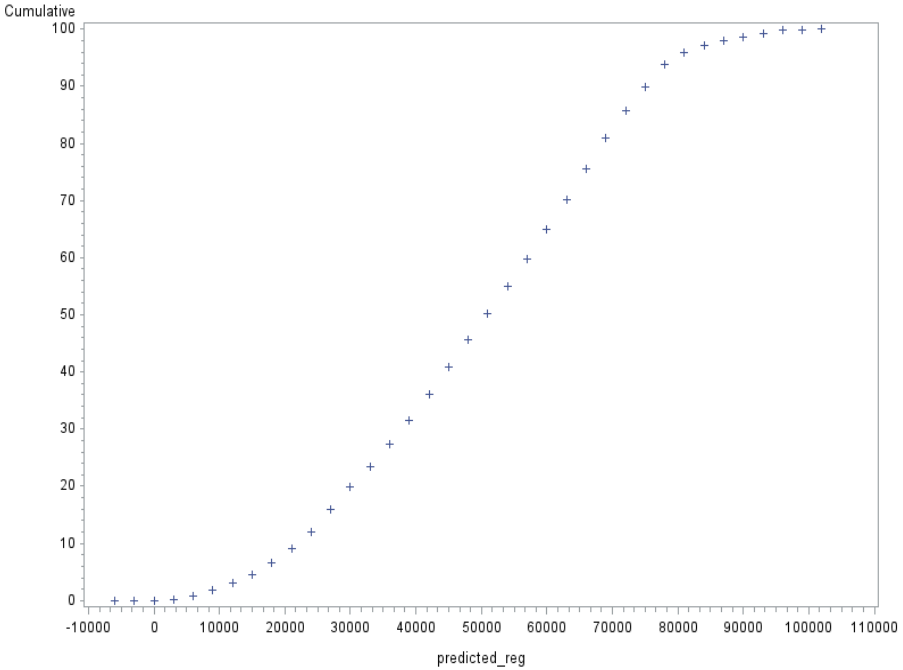
```
proc genmod data=a.impucomplete descending; class z1
z2 z3 z4 ; model income_resp=z1 z2 z3 z3*z4 x1 x1*x1
/link=logit dist=bin type3; output out=new2 p=predicted;
run;
proc means n mean min max cv data=new2; var
predicted; run;
```

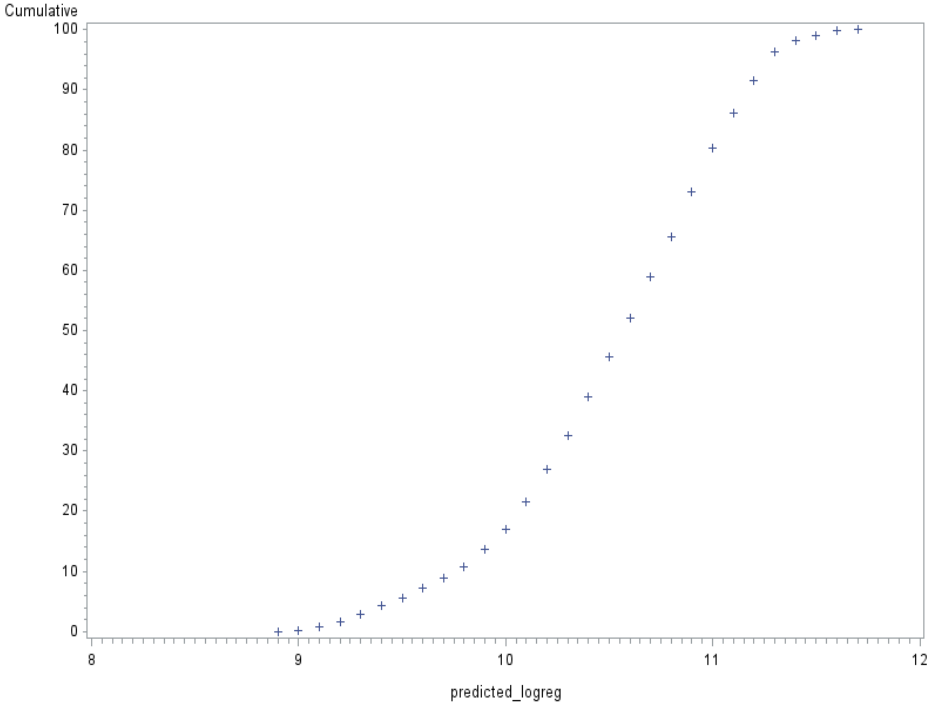
Probit regression as above but replace ‘logit’ with ‘probit’.

In our training, you thus have to choose the auxiliary variables. I hope that the same choice will be used in all models since it helps comparisons.

Graphs on predicted values on following pages

Cumulative frequencies of the predicted values for regression models

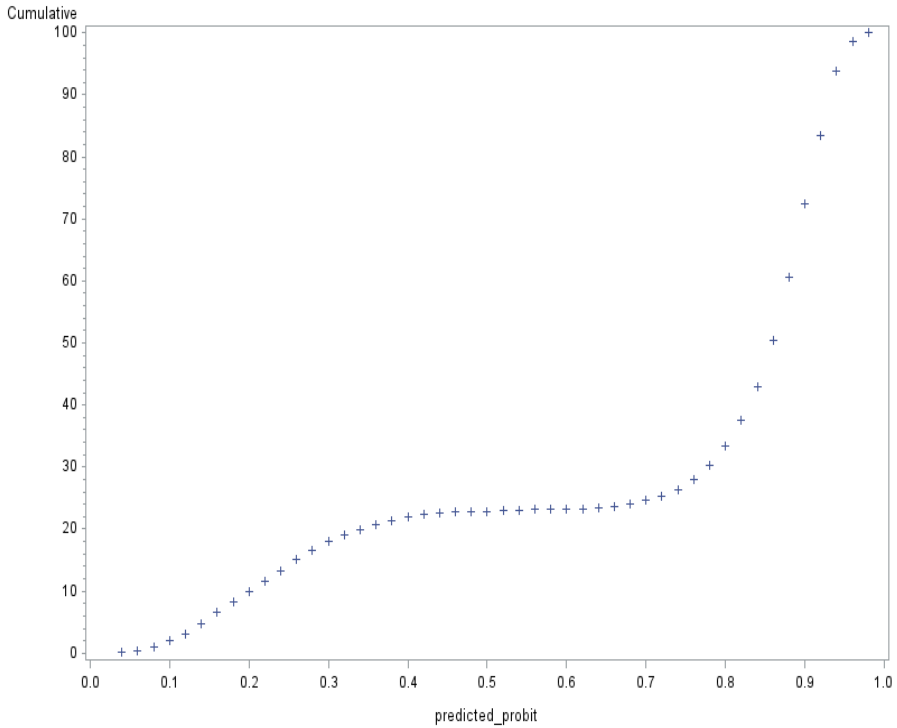


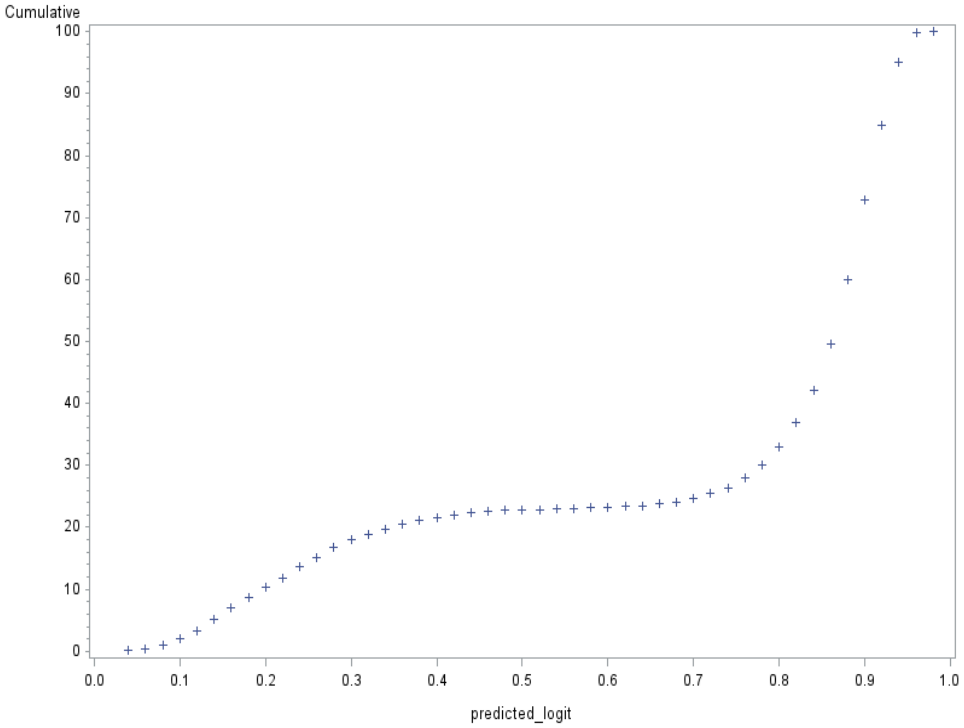


Linear regression

Loglinear regression

Cumulative frequencies of the predicted values for binary regression models



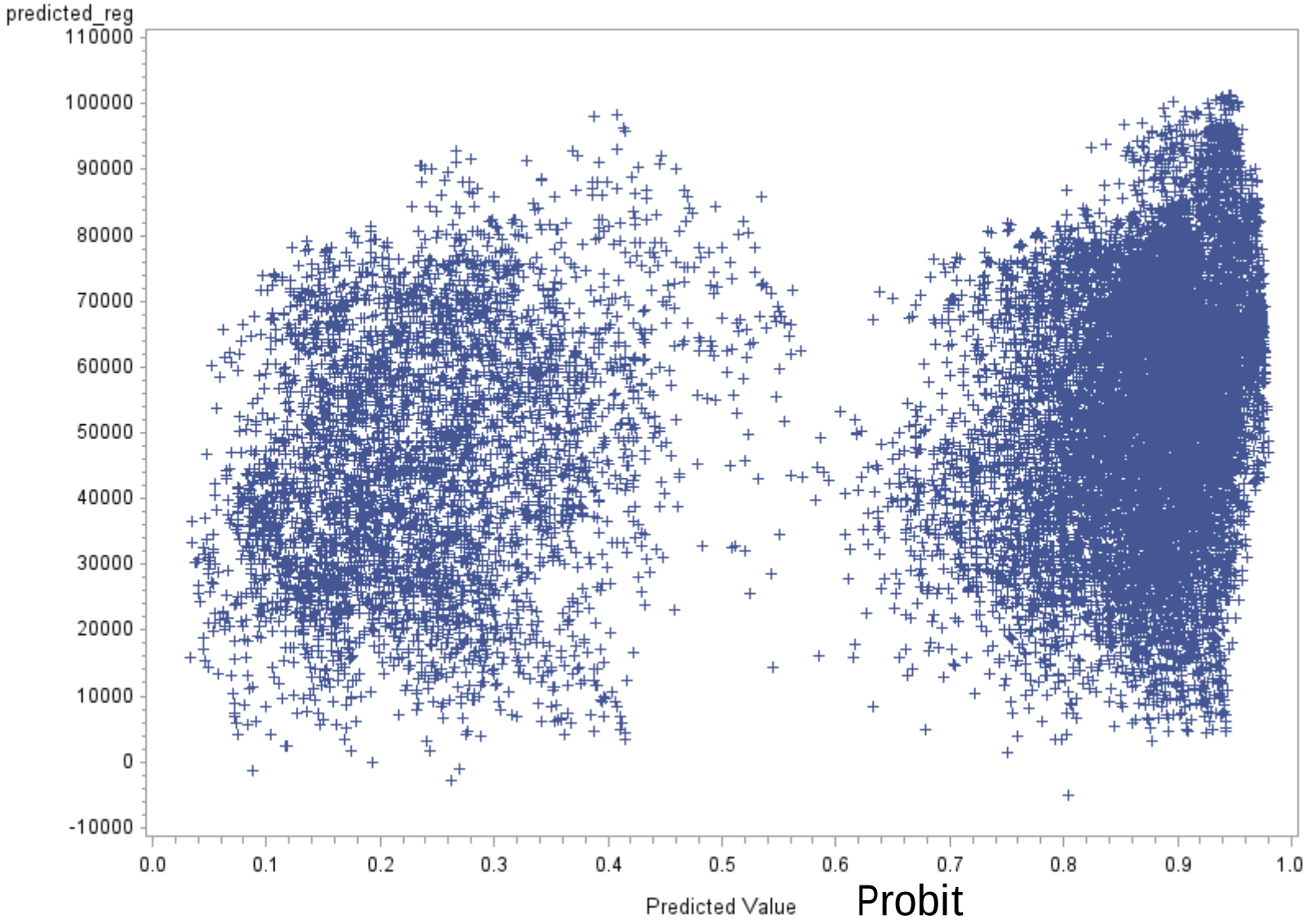


Probit regression

Logit regression

Scatter plots by predicted values: Both the respondent and the nonrespondents

Reg

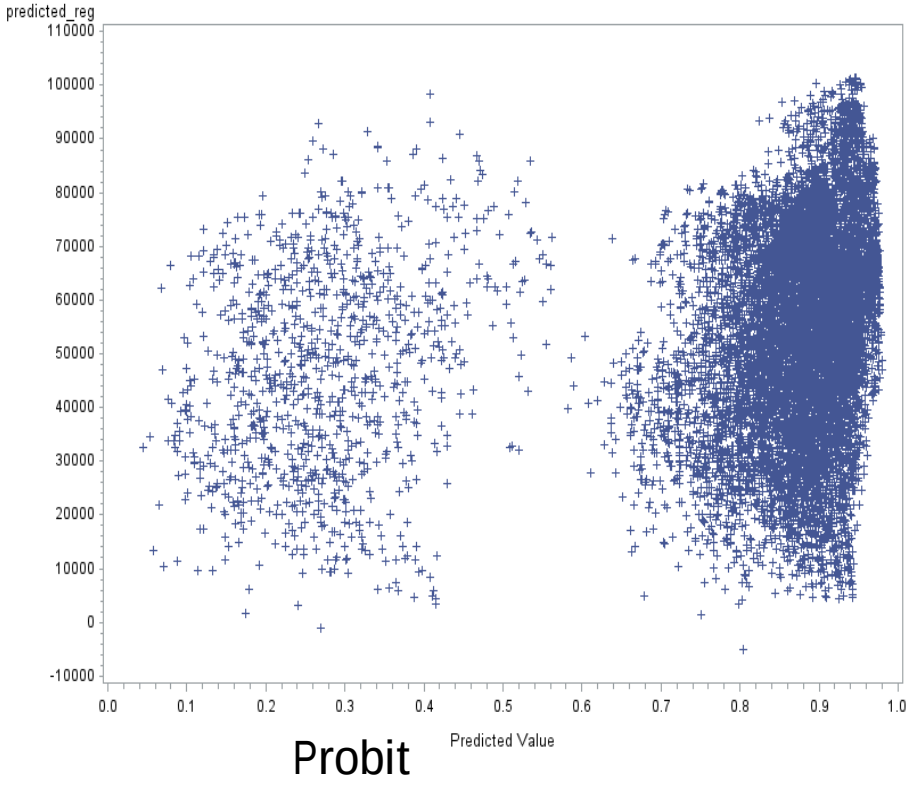


Probit

Scatter plots by predicted values:
For the respondent

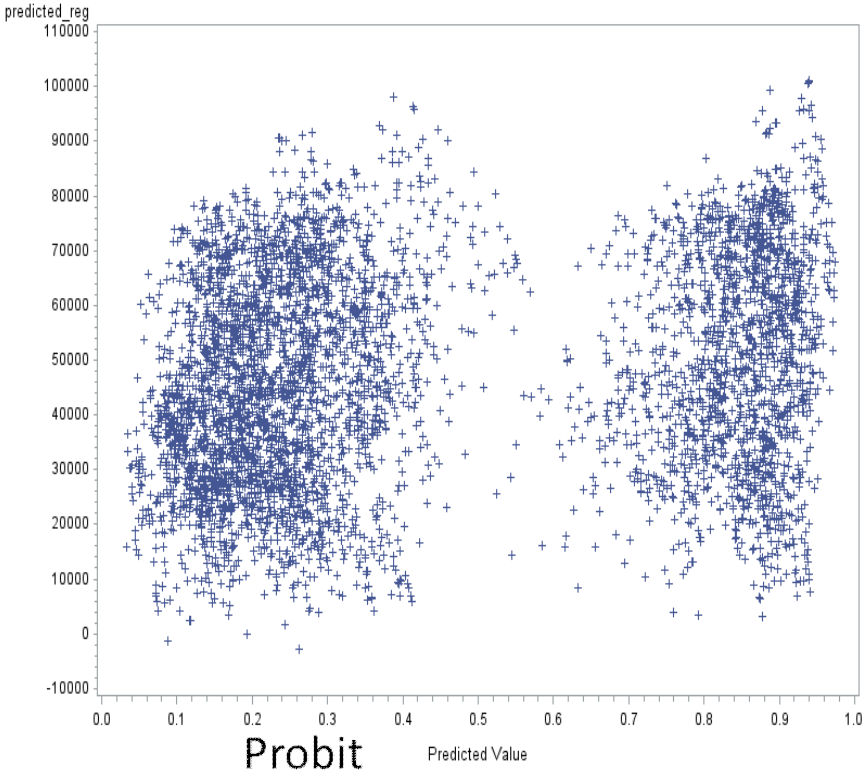
The nonrespondents

REG



Probit

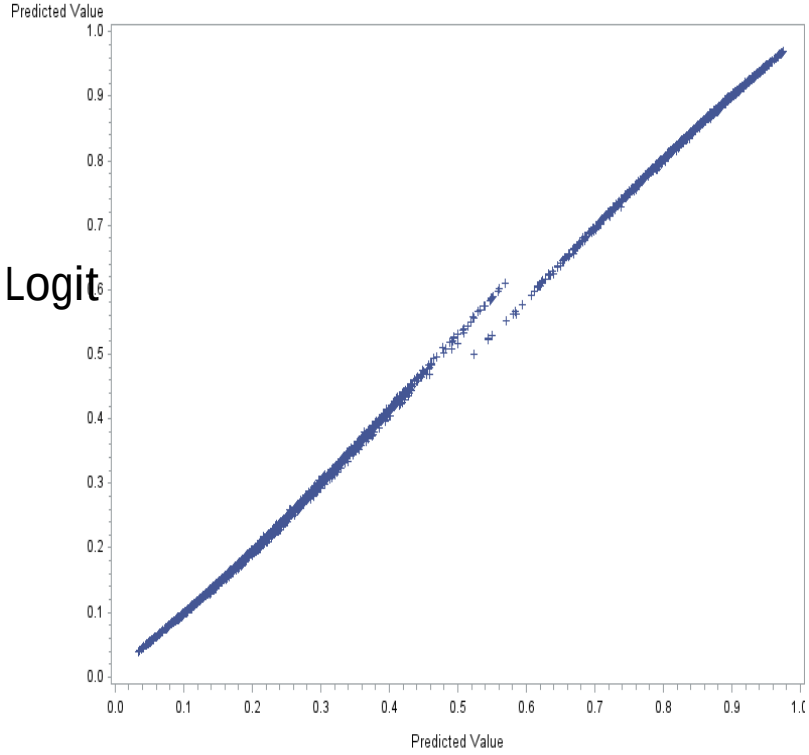
REG



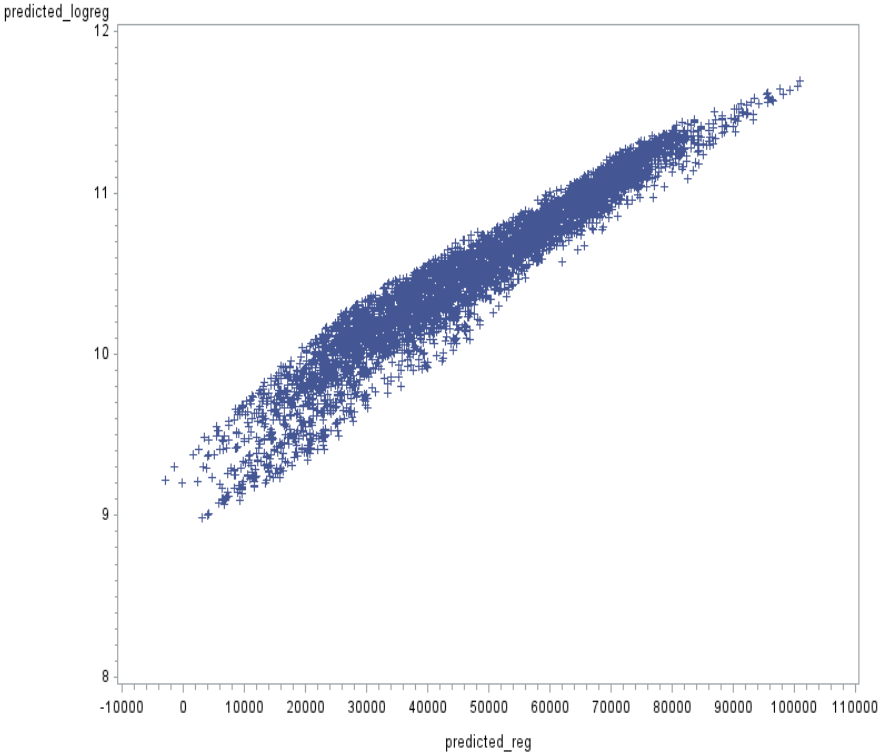
Probit

Scatter plots by predicted values: The nonrespondents
The similar models but with different ‘scales’
Predicting the response probability Predicting the income

LOG_REG



Logit

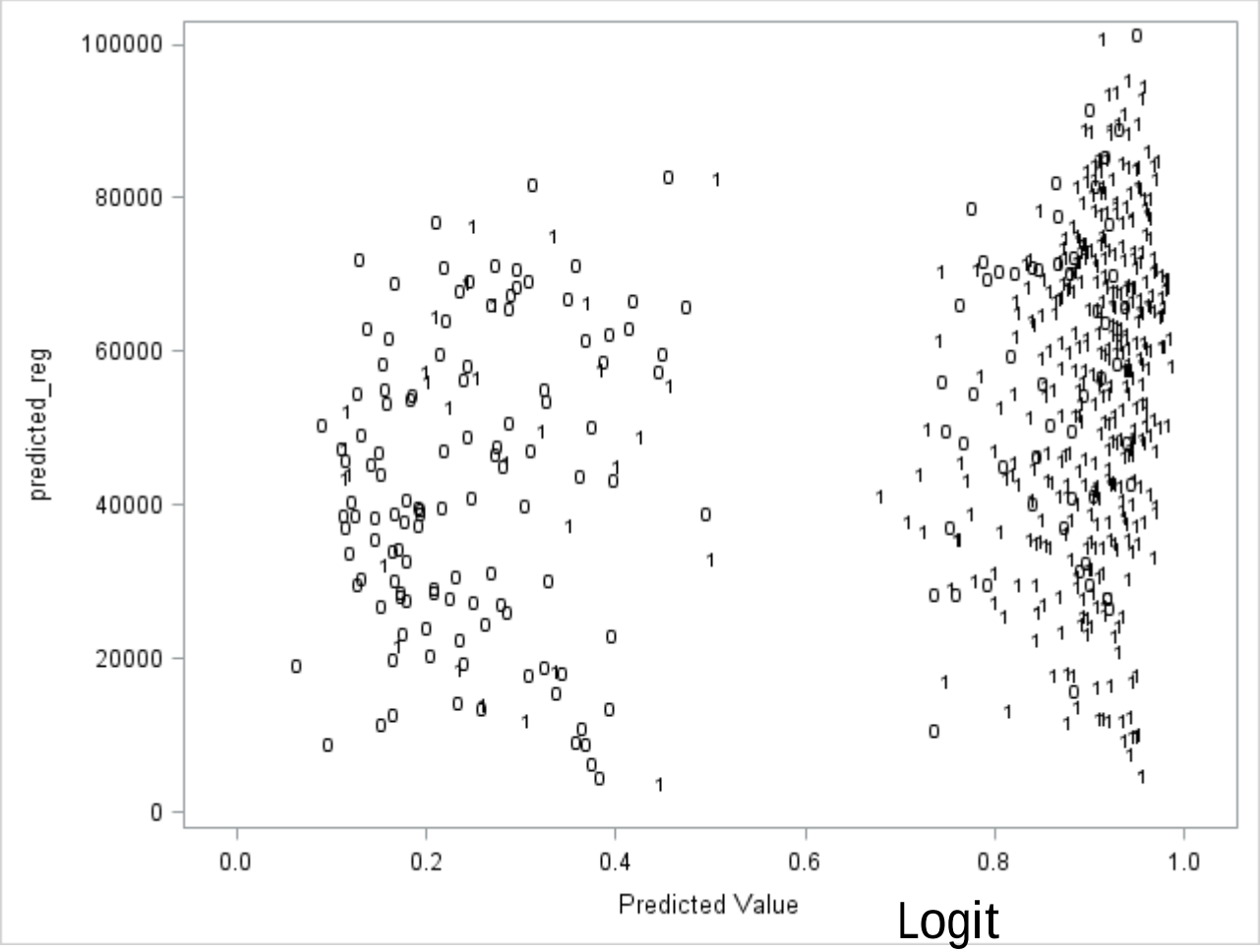


Probit

REG

Scatter plot by two predicted values: A 3% random sample of the complete data so that the response indicator is marked: 1=observed, 0=not observed

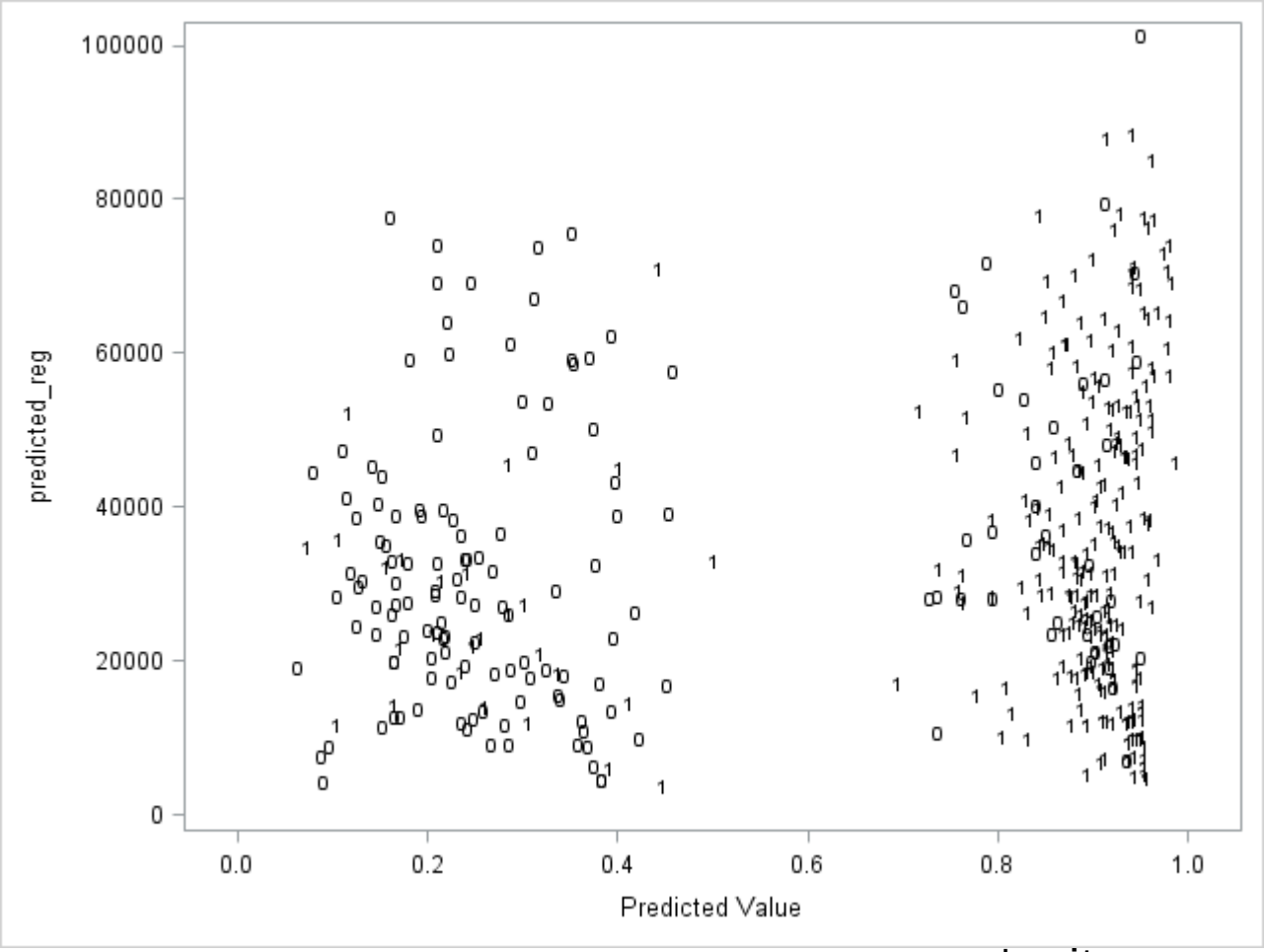
Reg



Logit

Scatter plot by two predicted values as the previous but for poor people, 10% random sample

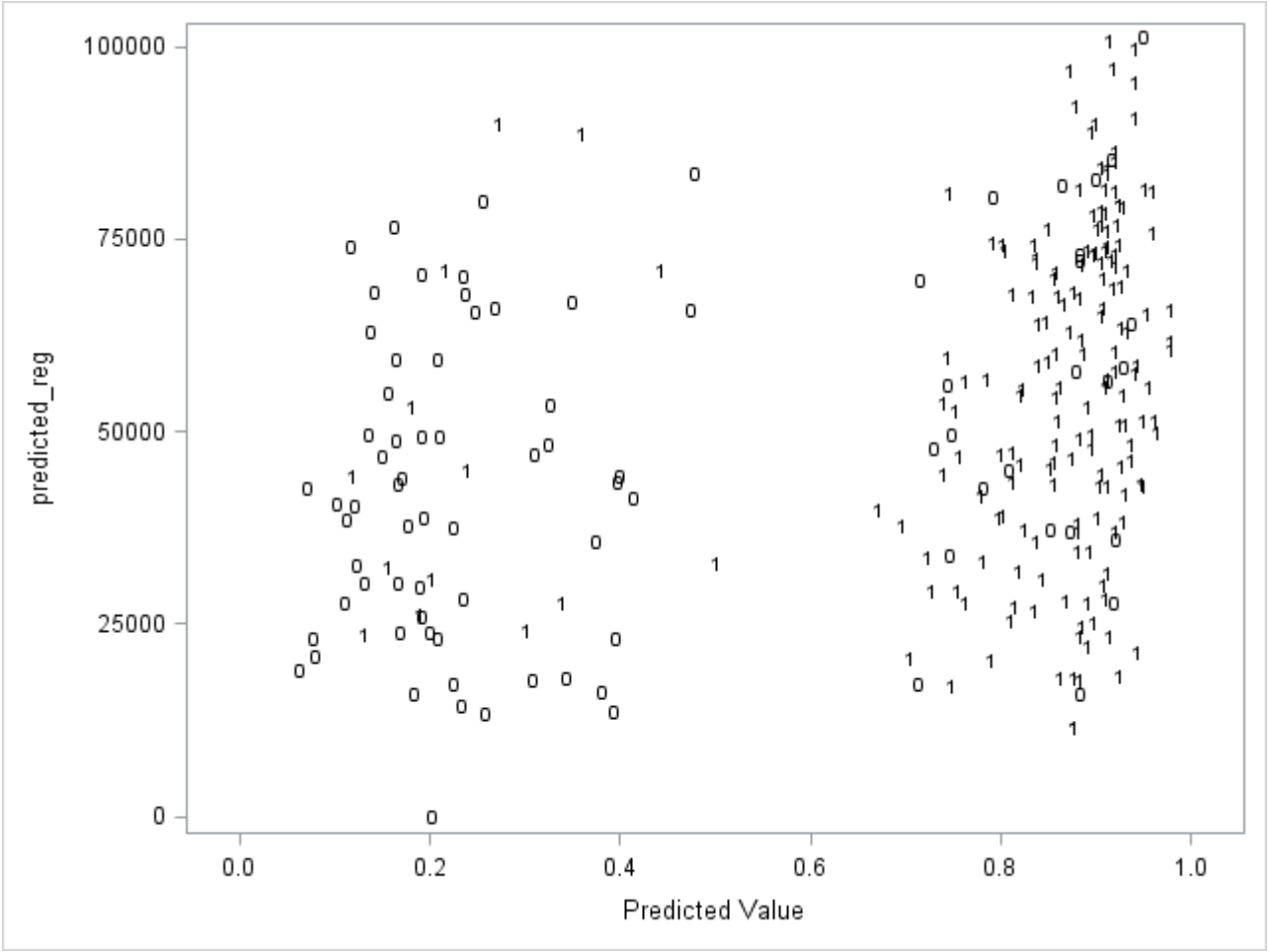
Reg



Logit

Scatter plot by two predicted values as the previous but for whose happiness is below 7, 10% random sample

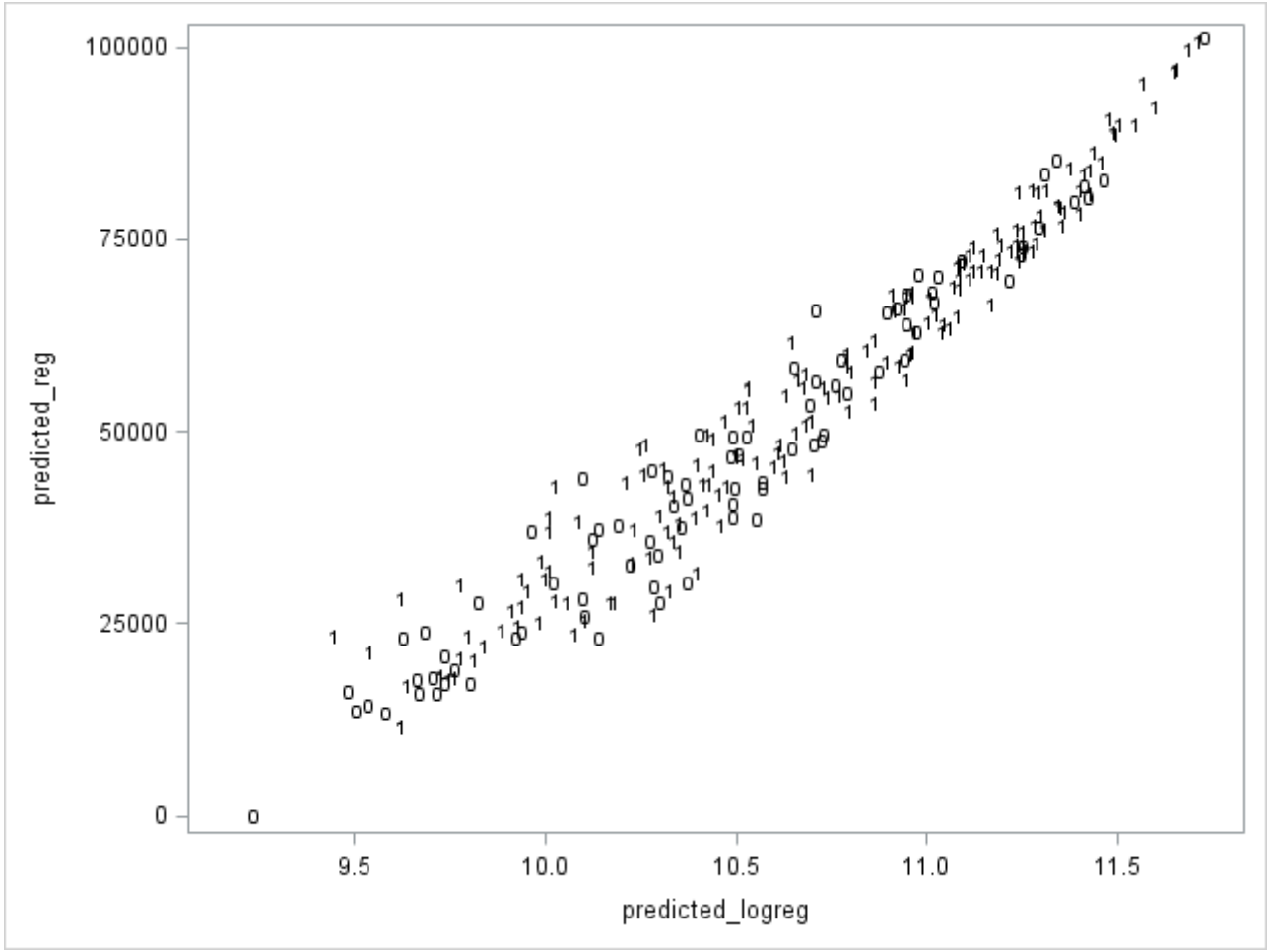
Reg



Logit

Scatter plot by two regression-based predicted values as the previous but for whose happiness is below 7, 10% random sample

Reg



Log_Reg

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 as in our trainings. All the previous predictions can be attempted. We have observed that there are many similarities but also essential differences and we cannot say definitely 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.

Concluding points about imputation models 2

However, it is expected that some methods are not good although used in real life. Our training data set is not easy that is good to keep in mind for understanding difficulties of imputations. 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. I hope that the examples of this course give some understanding about appropriate imputation methods, including both the model and the task that are connected to each other.

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.

).

Imputation task 2

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 and imputation model

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:
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 (deterministic) 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.

Imputation task 4

Comment continues:

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

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

Imputation task 5

Comment 2:
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 as we will do too. 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.’

Imputation task 6

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

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 7

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)*.

Imputation task 8

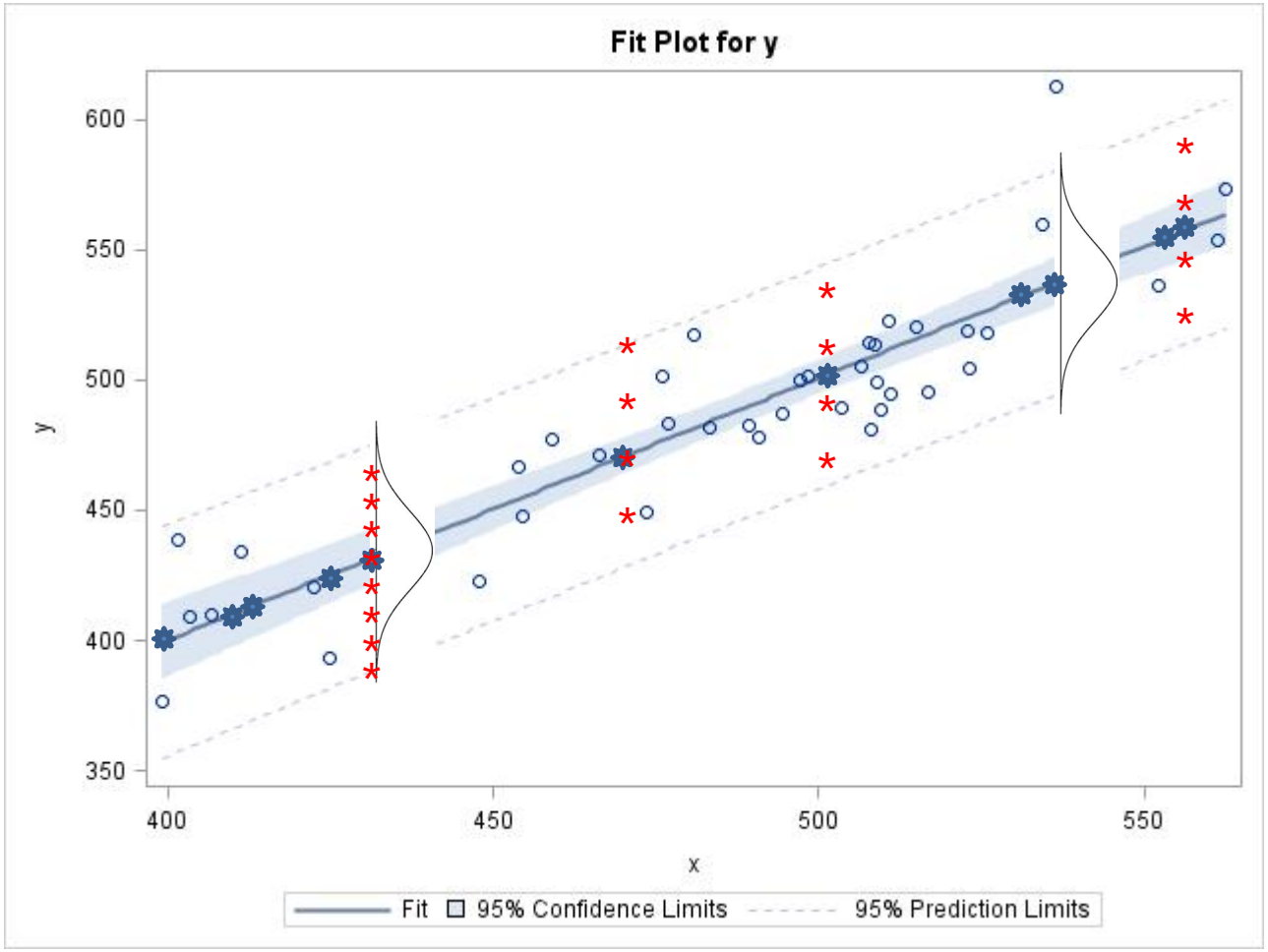
I do not here go to details of the noise term but when using regression model it is often assumed its distribution to be normal with the mean = zero and the standard deviation = root mean square error (standard deviation of the residuals). A problem is that there can be outliers in random values and 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.

Example, why and how to add a noise into the linear regression model assuming the noise variable to be normally distributed 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 a normal distribution, 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.

The SAS codes in this case you will find after the next page illustration.

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

SAS codes for adding the noise with N(0, rmse) Continues for a next page

```
proc glm data=a.impucomplete; class z1 z2 z3 z4 ;
model income_resp=z1 z2 z3 z3*z4 x1 x1*x1 /solution ;
output out=reg p=predicted_reg; r=residuals_reg; run;

/* It is needed to include those residuals and their minimum
and the maximum in the merged file. This can be made in
varous ways but this is my way: I create a new variable i and
give the simplest constant value. This same variable is needed
in the initial output file = reg in order to merge them together.
This requires the sorting by this variable. It looks maybe
strange but it works. Next we thus merge these files and we
have constant values root mean square error, and its
minimum and maximum that are used to robust the random
number based residuals with the normal distribution that we
will get by the operator rr.*/
```

```
proc summary data=reg nway; var predicted_reg residuals_reg; output
out=rmse std(residuals_reg)=rmse min(residuals_reg)=min
max(residuals_reg)=max
mean(predicted_reg)=mean;
data rmse; set rmse;
i=1; proc sort; by i;
data reg; set reg;
i=1; proc sort; by i;
/* Next we calculate these imputed values
This is one strategy for robusting imputes, that is, avoiding extreme values.*/
data reg2; merge reg rmse; by i;
rn=rannor(1); if rn<-min/mean then rn=min/mean; if rn>max/mean then
rn=max/mean;
if resp=1 then income_imp=income_resp;
else income_imp=predicted_reg+rn*rmse;
mae=abs(income_imp-income);
/* And we get out results including true values);*/
proc means data=reg2 n mean cv min p1 p5 p25 p75 p95 max; where resp=0;
var predicted_reg income_imp income mae; run;
```

Post-Editing after the model-donor method

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

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:

- Our second variable in training is 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.

Post-Editing after the model-donor method 2

In SAS codes this can be done as follows:

```
data new2; set new;
if happy_resp ne . then happy_imp_reg=happy_resp;
else happy_imp_reg=round(predicted_reg, 1);
mae=abs(happy- happy_imp_reg);
proc means data=new2 n mean cv min p1 p5 p25 p75 p95 max;
where income_resp=0;
var happy happy_imp_reg mae; run;
```

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.

Post-Editing after the model-donor method possibly 3
However, most clients do not like e.g. incomes with several decimals as we have obtained. Such values also indicate clearly for an expert that these are imputed. Thus: if the confidentiality is important as it is often, a rounding is a good solution but what is the best rounding?
My answer: the same as in the observed values. I looked at our data and found that the income values are in five euro's. Hence the rounding due to confidentiality and esthetic reasons can be as follows:

```
data new2; set new;
if income_resp ne . then income_imp_reg=income_resp;
else income_imp_reg=round(predicted_reg, 5);
mae=abs(income- income_imp_reg);
proc means data=new2 n mean cv min p1 p5 p25 p75 p95 max;
where income_res=0;var income income_imp_reg mae; run;
```

Nearness metrics of real-donor methods

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. The Mahalonobis distance is one such metrics used. Most typically, it is assumed that certain units (overall or within each imputation cell) are as close to each other. This means that a donor has been selected randomly (within the entire data or within an imputation cell). It is thus stochastic. This method is just the initial random hot deck method from 1950's.

Another 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.

Nearness metrics of real-donor methods 2

The third and most common metrics as guessed from the previous graphs is the metrics 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.

Nearness metrics of real-donor methods 3

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 4

The cell-based random hot deck or 'real-donor method using response propensity cells' does not give any nearest neighbor but a near neighbor. There are literature e.g. such term as '*k*-Nearest Neighbors algorithm' that is close to this idea so that this gives *k* nearest for each unit selected. The same idea was used in the Euredit project by the York University team.

See the paper abstract in which the CMM is mentioned.
This method was not bad in the Euredit examinations.

A Binary Correlation Matrix Memory *k*-NN Classifier

Ping Zhou and Jim Austin
Department of Computer Science, University of York, York YO10 5DD, UK
Email: zhoup@cs.york.ac.uk, austin@cs.york.ac.uk, Fax: +44 (0) 1904 432767

Abstract

In this work we investigate the use of a binary CMM (Correlation Matrix Memory) neural network for pattern classification. It is known that a *k*-NN rule is applicable to a wide range of classification problems but it is slow, and that the CMM is simple and quick to train, and has highly flexible and fast search ability. We combine the two techniques to obtain a generic and fast classifier which uses a CMM for storing and matching a large amount of patterns efficiently, and the *k*-NN rule for classification. To meet requirements of the CMM, a robust encoder has been developed to convert numerical inputs into binary ones with the maximally achievable uniformity. Experimental results on several benchmarks show our method can be over 4 times faster than the simple

Nearness metrics of real-donor methods 5a

The York University method worked but was not best. The better solution is to try to find a nearer neighbor than one randomly from a possibly large group.
I have used (and we will use it in our training) such a method that

- (i) sorts all the units in the data by the predicted values from the largest to the smallest (or opposite)
- (ii) creates the lagged variables as many as needed (maybe even 10-20 lag variables), nearest= lag1, 2nd nearest= lag3, 3rd nearest=lag5, ...
- (iii) sorts this sorted data set to the opposite order, that is, from the smallest to the largest (or opposite)
- (iv) creates the lag variables similarly to (ii), 1st=lag2, 2nd=lag4, 3rd=lag6, ...
- .

Nearness metrics of real-donor methods 5b

- (v) begins the imputation so that if a missing value is observed then it is first looked whether lag1 is non-missing; if ‘yes’, this value has been chosen as an imputed value; if ‘no’, then it is checked lag2 and so on as long as all the values are imputed.

This method works well except if going too far from a nearest value to find an observed value. It is possible that a big number of lagged variables is needed to impute all missing values. It means that the same real-donor will be used more than once as a real-donor. It is possible to choose a model-donor method in such cases instead of real-donor method.

In general, a real-donor method is a kind of a weighting method but much more flexible than its ordinary form: in weighting, one weight is for each respondent, in this case the values of respondents can be used as imputed values or not, some several times.

Nearness metrics of real-donor methods 6

The fourth good and rational strategy (like my regression-based nearest neighbor hot decking) 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 and we made it already in graphs below. 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.

Nearness metrics of real-donor methods 7

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 8

To make the previous point "Thus e.g. our nearness metrics can be the previous model-donor output" clearer:

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.

It is also possible to choose a model-donor imputed value for those units whose nearest neighbor is too far and thus not be plausible. In this case the final imputation is a mixed real&model-donor method. It is allowed. It is possible to take the mean of both imputed values too.

Nearness metrics of real-donor methods 9

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). The SAS codes are thus similar in both cases but the values are not. 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. These will be considered later. 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.

Strategy in our SAS training for real-donor methods

Imputation model:

(i) The same options as in model-donor methods

and

(ii) Binary regression models of the response indicator

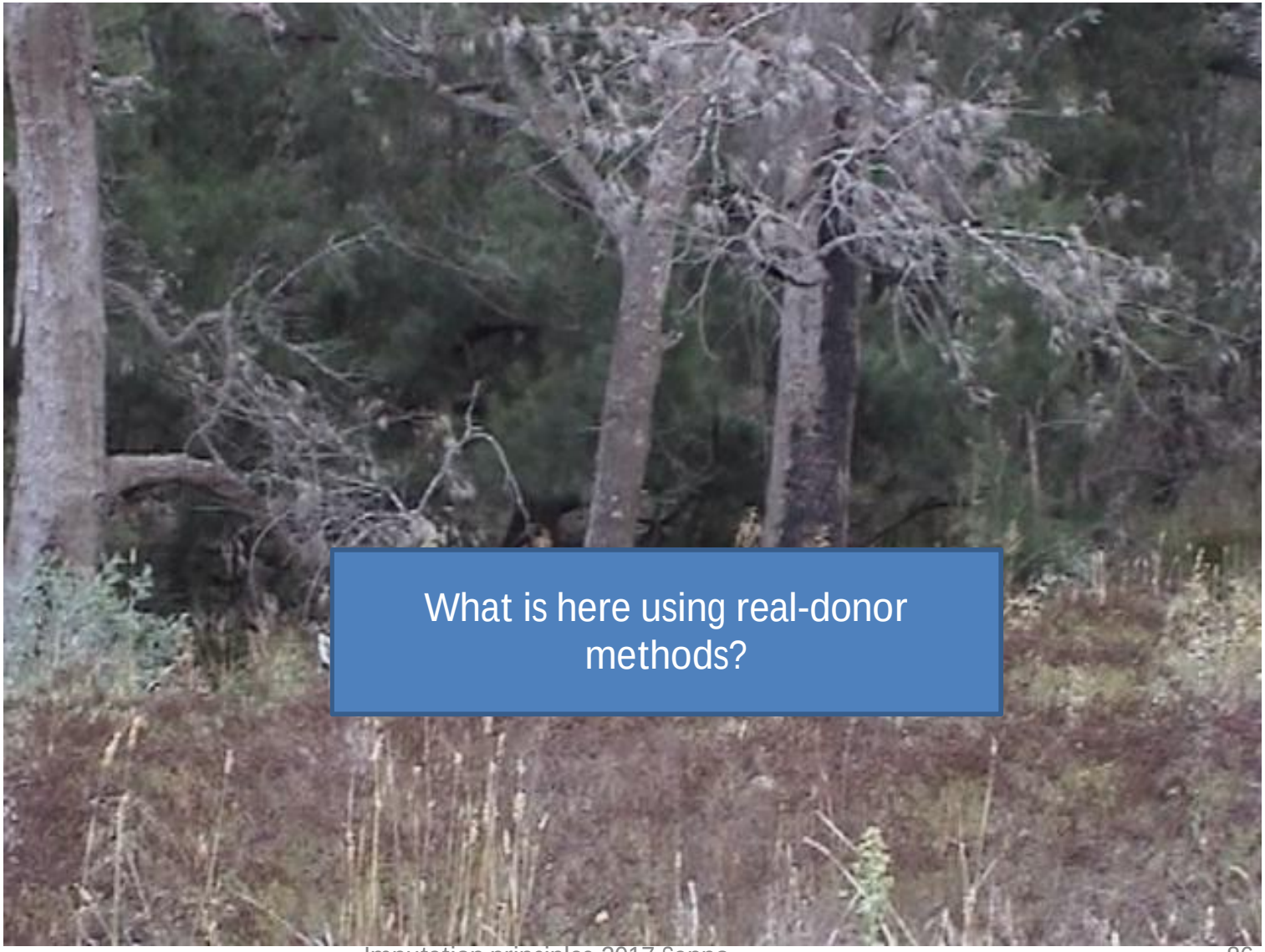
but the predicted values are for both the observed and unobserved units

Imputation task:

Find a near neighbor using the output file of the imputation model

You need to run one SAS program to get the predicted values for all units, and then go to another SAS program to run the imputation task. The required file and the nearness metrics respectively are needed to use as inputs of the second part.

A break task: Try to guess what is missing and tell me if you want



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 often criticized.

Single and multiple imputation 2

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(ald) 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.

Summary: Imputation model plus Imputation task
In the case of the response indicator model

Model-Donor	Deterministic Single	Stochastic Single Multiple
Real-Donor	Nothing	Nothing
	E. Logit, probit or Comple-mentary log-log (CLL) reg-ression model with the respective explanatory variables as above. Those predicted values (response propensities) are computed both for the respondents and for the non-respondents that used as a nearness metrics.	F. Adding 'noise' in which different strategies can be used, always uniformly distributed random numbers, but I do not go now to details

Single and multiple imputation 2
Technics

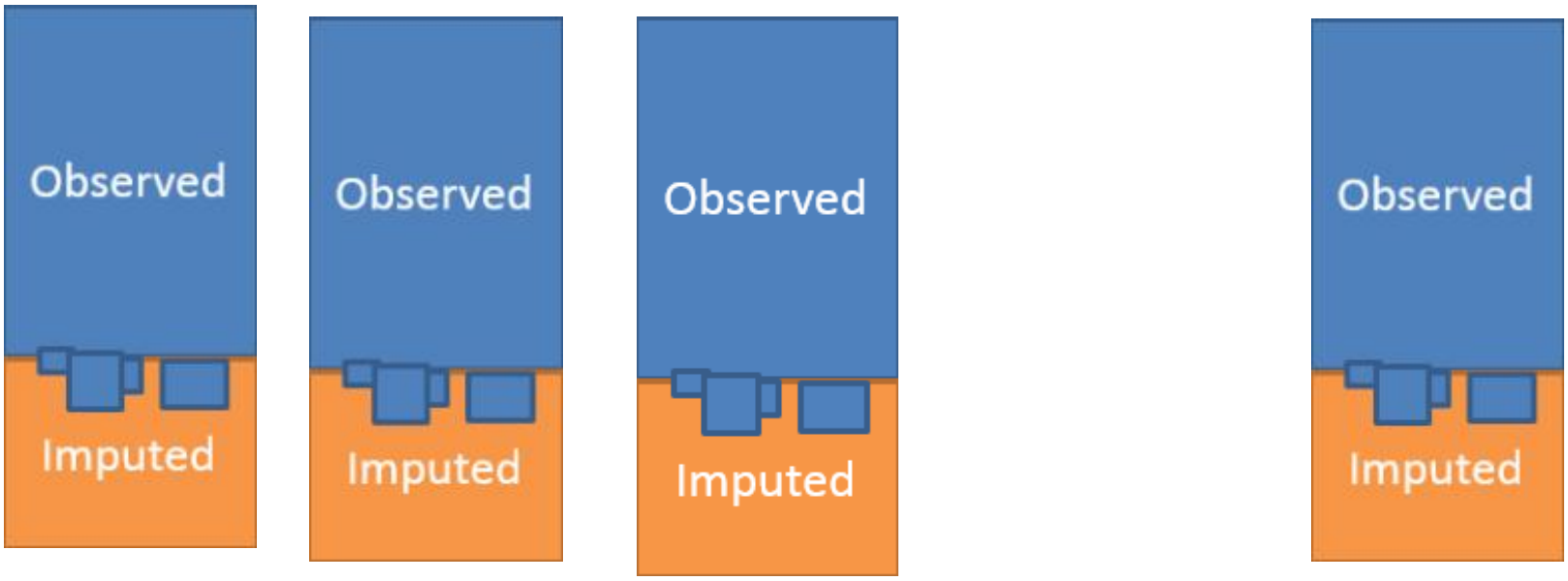
Let
L = number of imputations *u*,
Θ = parameter being estimated,
and its point-estimate = *Q* (e.g. mean income and CV)
and variance estimate, respectively, = *B*
And then standard error of the mean = square root of the variance.

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 remembers how inefficient the computers were in 1980's.

Single and multiple imputation 3

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



Point and interval estimates from each data set as usually

Single and multiple imputation 4

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 5

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 is 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.

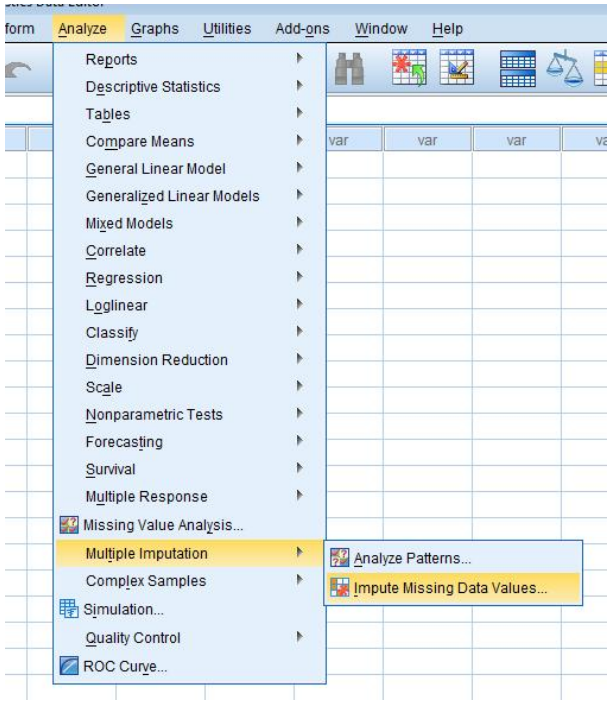
Single and multiple imputation 6

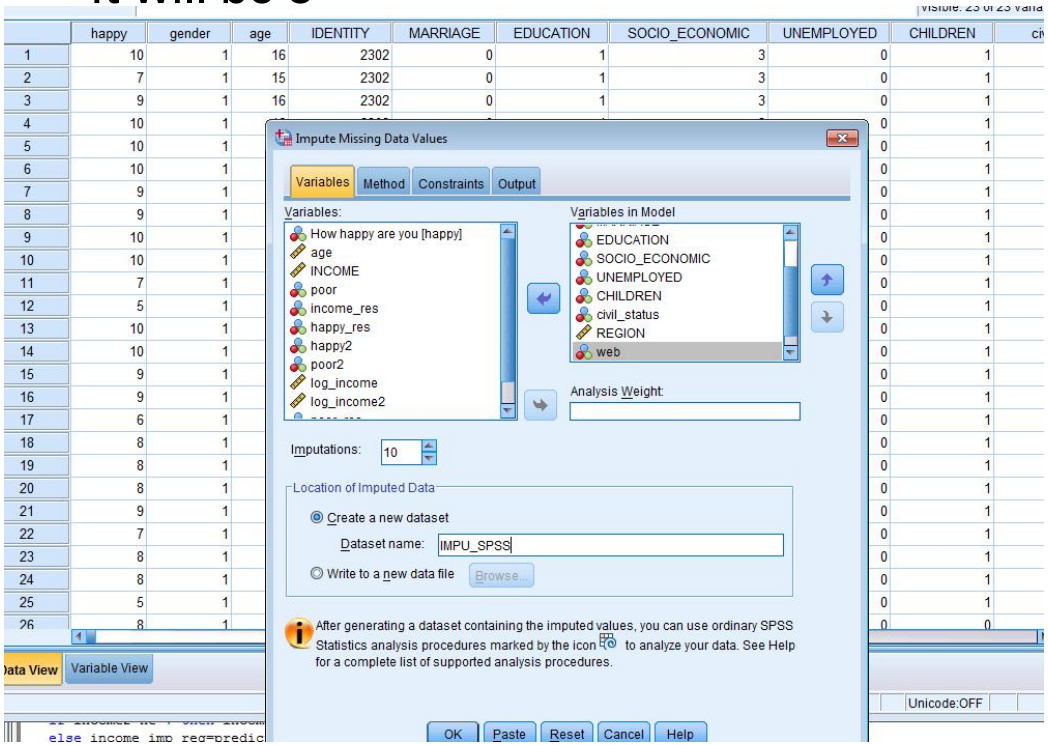
The initial multiple imputation was developed by Donald Rubin. It was based on the Bayesian theory. This theory thus was reformulated by the Norwegian Jan Björnstad. A reason was that Rubin's strategy is not well working in many practical situations like in statistical offices. Hence he uses the term 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. Note that these rules are developed by Rubin and a user thus have to trust in him or his specifications. I have to say that I am not convinced about all the solutions?

SPSS Imputation

First push here right.
Next you have to select your SAS
data that SPSS understands
and select the variables (the same as earlier),
no income and happy
Number of imputations and the output file name
If you do not select the number of imputation,
it will be 5

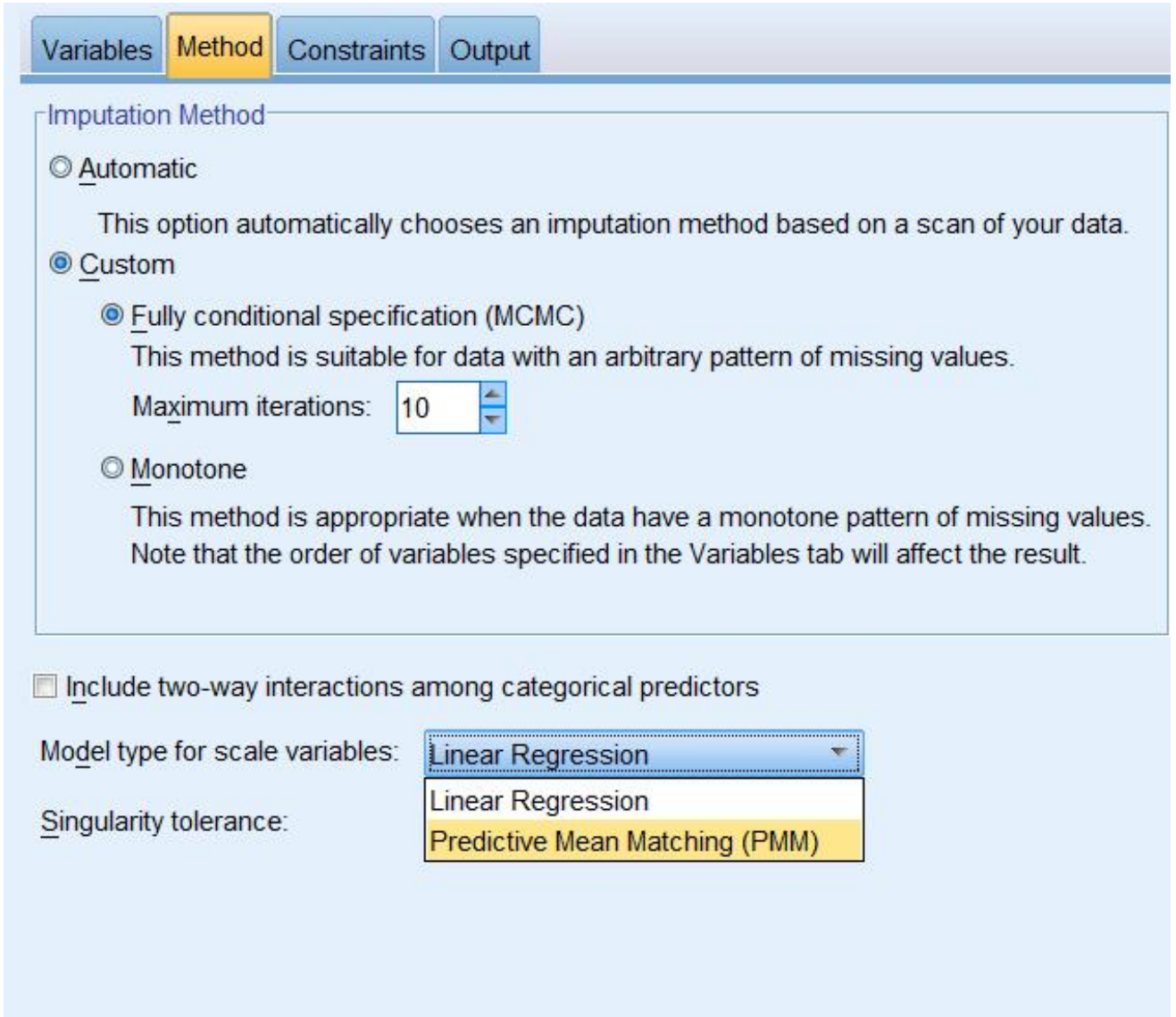




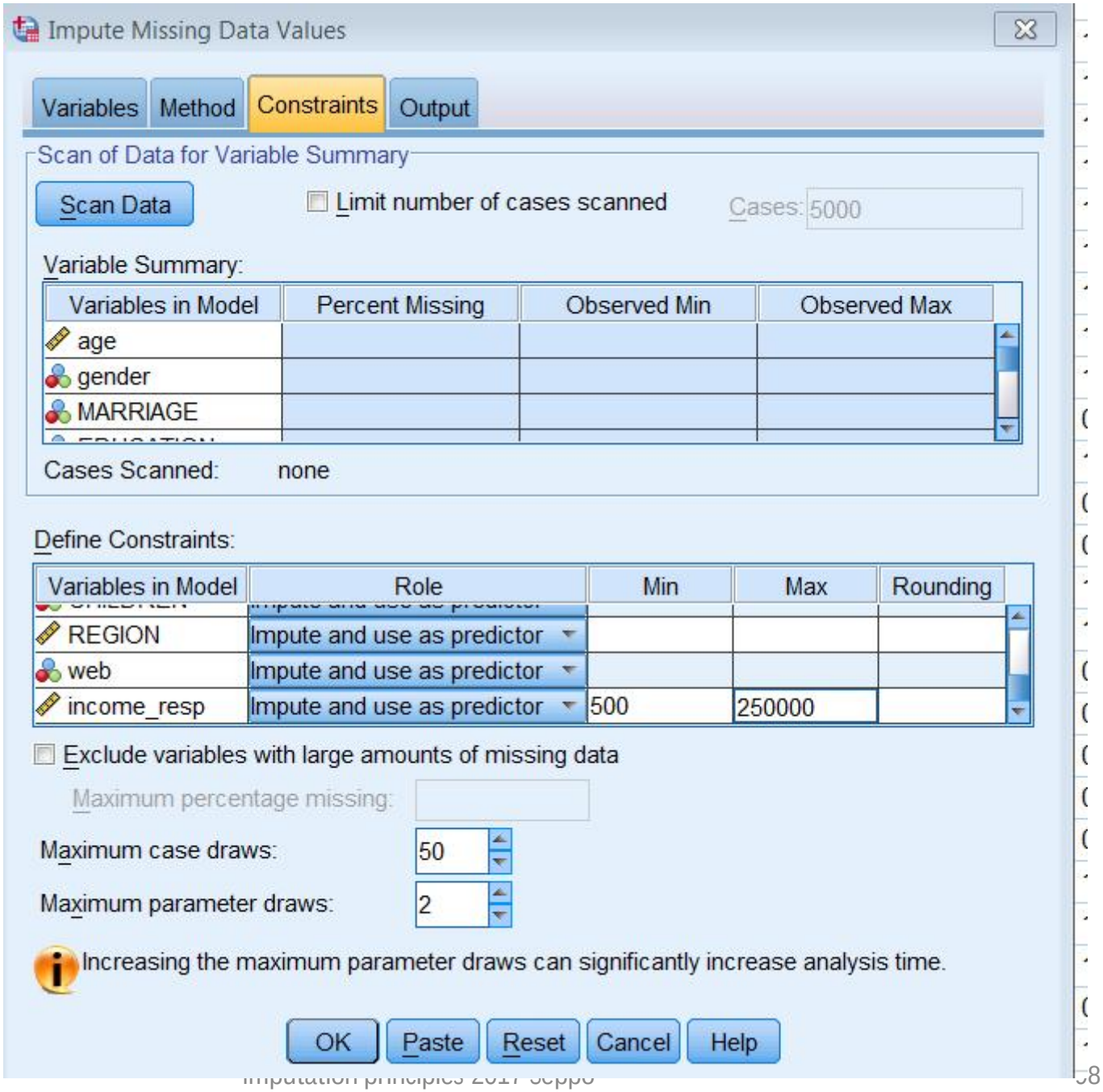
As you see, next you can choose
a method and also constraints
for the variable being imputed

It is purpose to test both methods
in our training. One option in each
at least but more is better.

This copy tells which methods there are. The options are possible using 'Constraints.'



This is a possible constraint



Linear regression is a Bayesian model-donor method with the linear regression model as the imputation model

Predictive Mean Matching (PMM) Method is a Bayesian real-donor method using the linear regression. We do not explain its theory in details but it thus gives observed values that is a good point. There are different algorithms for PMM but any is not the same as my real-donor algorithm. You do not need to know the algorithm exactly.

The next page shows a part of the two output files where you see which values are imputed but the variable name is the same. They are linear regression imputation without and with constraints.

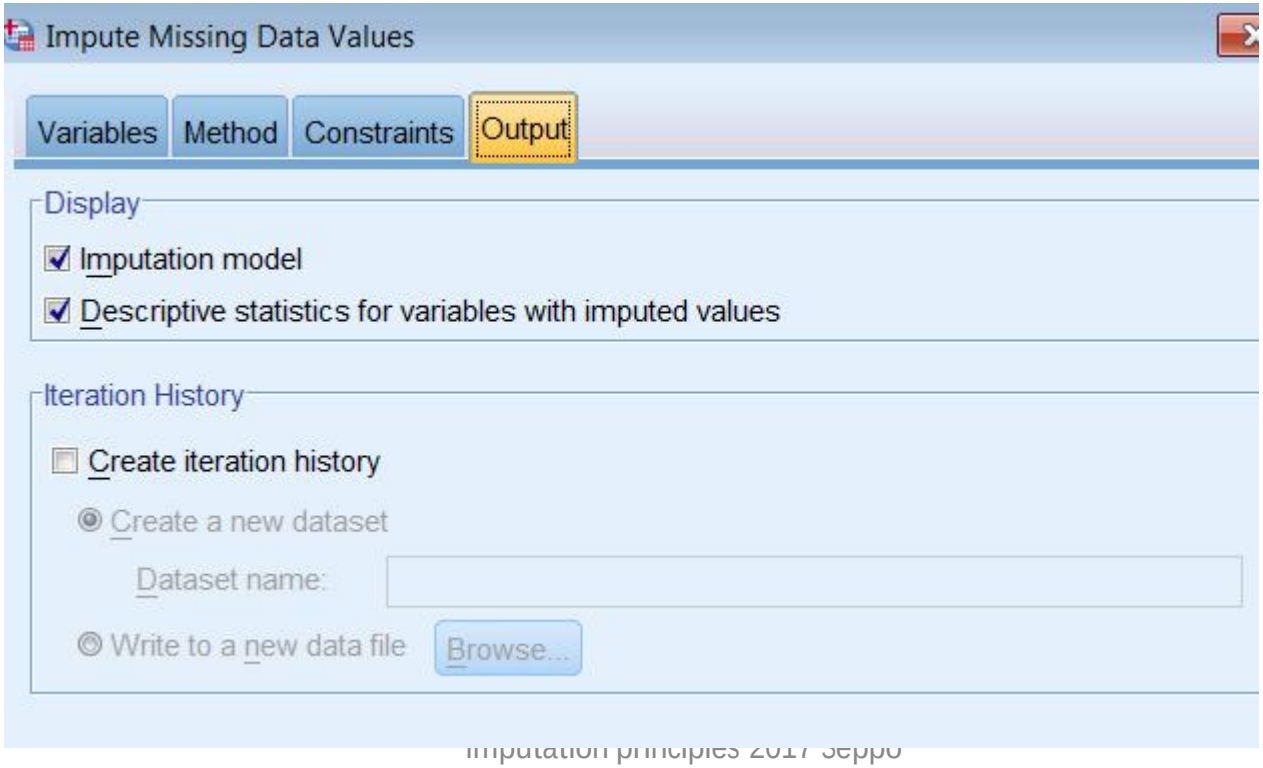
Comments on Bayesian vs Non-Bayesian MI

My approach thus is non-Bayesian so that MI imputations are made 'straightforwardly' in some sense. Bayesian MI always includes one further step to get an additional random draw (a random draw from the posterior predictive distribution). This step thus adds variability (standard error); Björnstad suggests to use k>1 ($k = \frac{1}{1 - f}$) to do the same. It is difficult for me as non-Bayesian to convince about this additional step. In the case of the linear regression model this step consists e.g. of random chi-squared numbers. There are in general many algorithms for Bayesian MI, and they are criticized as well (e.g. Allison 2015: http://statisticalhorizons.com/predictive-mean-matching).

Output data, no constraints above, positive constraints below, IMPUTATION_=1

	civil_status	REGION	INCOME	web	poor	resp	agegroup	age2	happy_resp	income_resp	poor_resp	HOMESIZE
19844	1	31	18990	1	1	1	60	3136	10	18990	1	2
19845	1	14	12840	1	1	1	60	3136	10	12840	1	2
19846	1	32	1035	1	1	1	60	3025	10	1035	1	2
19847	0	11	1420	1	1	1	60	3136	10	1420	1	1
19848	0	23	14925	1	1	1	60	3025	10	14925	1	1
19849	1	24	25230	1	0	0	60	3136	.	47104	.	3
19850	0	11	22365	1	1	1	60	3025	10	22365	1	1
19851	1	33	21845	1	1	1	60	3025	10	21845	1	3
19852	1	32	18100	1	1	0	60	3025	.	28978	.	3
19853	0	21	25870	0	0	0	60	3025	.	2688	.	1
19854	1	12	22110	0	1	1	60	3025	10	22110	1	2
19855	1	31	25215	0	0	0	60	3025	.	-12138	.	2
19856	0	12	26055	0	0	1	60	3025	10	26055	0	1
19857	1	23	19065	0	1	0	60	3025	.	105036	.	3
19858	1	34	26385	0	0	0	60	3025	.	51929	.	2
19859	1	13	16440	0	1	0	60	3136	.	13851	.	2
19860	0	24	25480	0	0	0	60	3025	.	77684	.	1
19861	1	32	2130	1	1	0	60	3364	.	36134	.	3
19862	1	32	28170	1	0	1	60	3364	2	28170	0	2
19863	1	32	22110	1	1	1	60	3249	8	22110	1	3
19864	0	31	12015	1	1	1	60	3364	8	12015	1	1
19844	1	31	18990	1	1	1	60	3136	10	18990	1	2
19845	1	14	12840	1	1	1	60	3136	10	12840	1	2
19846	1	32	1035	1	1	1	60	3025	10	1035	1	2
19847	0	11	1420	1	1	1	60	3136	10	1420	1	1
19848	0	23	14925	1	1	1	60	3025	10	14925	1	1
19849	1	24	25230	1	0	0	60	3136	.	100360	.	3
19850	0	11	22365	1	1	1	60	3025	10	22365	1	1
19851	1	33	21845	1	1	1	60	3025	10	21845	1	3
19852	1	32	18100	1	1	0	60	3025	.	35212	.	3
19853	0	21	25870	0	0	0	60	3025	.	31252	.	1
19854	1	12	22110	0	1	1	60	3025	10	22110	1	2
19855	1	31	25215	0	0	0	60	3025	.	29104	.	2
19856	0	12	26055	0	0	1	60	3025	10	26055	0	1
19857	1	23	19065	0	1	0	60	3025	.	56790	.	3
19858	1	34	26385	0	0	0	60	3025	.	30029	.	2
19859	1	13	16440	0	1	0	60	3136	.	70695	.	2
19860	0	24	25480	0	0	0	60	3025	.	50765	.	1
19861	1	32	2130	1	1	0	60	3364	.	65227	.	3
19862	1	32	28170	1	0	1	60	3364	2	28170	0	2
19863	1	32	22110	1	1	1	60	3249	8	22110	1	3
19864	0	31	12015	1	1	1	60	3364	8	12015	1	1

The previous page is from the first imputed data set. If variable_=0, it is the initial non-imputed data set. Now it is possible to calculate the results. One alternative is to select descriptive statistics as below.
It does give the same we have with SAS. Hence I also use SAS so that this is first saved as a SAS file and then the operations of the next page are used. This gives opportunity to compare these results with our earlier ones.



SPSS Multiple Imputation

The SAS codes for getting 10 multiply imputed results.
I thus saved the former file with the name SPSS in my library ‘a.’

```
data spss; set a.spss;
if Imputation_ ne 0;
if resp=0;
mae=abs(income-income_resp);
run;
proc means data=spss n mean cv min p1 p5 p25 p75 p95
max ;
class Imputation_; var income income_resp mae; run;
```

It is possible to calculate e.g. the average of all 10 imputations and get one estimate. This is not included in this material but if you can wish to do it, you can follow the formulas above (Rubin and Björnstad). In that stage it is good to use PROC SUMMARY.

Bayesian MI estimates. This gives an option for additional credits if done. Two options: with or without AND RESP=0.

```
data impu_SPSS; set a.spss;
mae=abs(income_resp-income);
if imputation_>0 /*AND RESP=0 */; run;
proc summary nway DATA=IMPU_SPSS; class
imputation_; var income_resp;
output out=impu_MI stderr(income_resp)=stderr
mean(income_resp MAE)=mean MAE; run;
data impu_MI2; set impu_mi;
drop _type_ _freq_;
Bi=stderr*stderr; run;
proc summary data=impu_MI2 nway; var Bi mean MAE;
output out=impu_aggre mean(Bi)=B
var(mean)=var_between mean(MEAN MAE)=MEAN MAE;
data impu_aggre2; set impu_aggre;
drop _type_ _freq_;
B_MI = B+(1+1/10)*var_between;
std_B=sqrt(B);
stderr_mi= sqrt(B_MI); run;
proc print; run;
```

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. These cases are considered in following pages.

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

Model-donor imputation of a categorical variable

It is easiest to use this case so that each category has been imputed separately but this takes more time of course when comparing the real-donor imputation. This can be made using multinomial distribution and an available link function (logit, probit, cll). We do not concretize this methodology in this course, but apply in imputing a binary variable poor vs not poor.

In this case, the imputation model is binary with an available link function as for the real-donor imputation above but this dependent variable is this binary variable being imputed, not any indicator.

The imputation model (link: logit, probit or cll) is estimated and the predicted values respectively. When going to imputation tasks on next pages, the variable of these predicted values is 'predicted_md' in which 'md' refers to 'model-donor.'
You remember that these values are within the interval (0, 1).

Model-donor imputation of a categorical variable 2

Alternatives:

(i) If the prediction is working well that is not guaranteed, it is easiest
- to give an imputed value = 1 if a predicted_md > 0.5
- an imputed value = 0, otherwise.

Thus if the binary model is concerned the variable 'poor' so that 1 = poor and 0 = not poor, this basically works but I have not seen any good empirical result on this.

(ii) Calculate the average of this binary indicator for the respondents, and assume that this works for the respondents as well. If the poverty rate is 0.1174 then the SAS codes are:

```
data new3; set new2;
ran=ranuni(1);
if predicted_md> 0.1174 then poor_imp_md1=1; else poor_imp_md1=0;
proc means n mean cv min p1 p10 p75 p90 p95 p99 max ; where resp=0;
var poor_imp_md1 predicted_md poor; run;
```

.

Model-donor imputation of a categorical variable 2

Alternatives:

(iii) Create an uniformly distributed random variable within the same interval as the predicted values are i.e. (0, 1). This is below the variable ‘ran’.

```
data new3; set new2;
ran=ranuni(1);
if predicted_md>ran then poor_imp_md2=1; else poor_imp_md2=0;

proc means n mean cv min p1 p10 p75 p90 p95 p99 max ; where resp=0;
var poor_imp_md2 predicted_md poor; run;
```

This method is usually best (but which is the best link function, it is not known) but do your test as well.

Model-donor imputation of a categorical variable 3

As you see, the first two alternatives are deterministic but the third is stochastic. Hence the third can be used for non-Bayesian multiple imputation just changing the seed number that is now = 1. If changing this number the results vary to some extent but not in most cases much. Note that the third approach follows the Bernoulli distribution.

There are other deterministic solutions like learning about the observed results but I am not convinced about any. However, when calculating the average of the predicted_md this 'aggregate imputation' for the mean is fairly good but you cannot know any correct individual values. Thus if this aggregate only is needed, you can use this aggregate.

Solution to the break page imputation task.

Did you impute correctly or how close? Not difficult due to good auxiliary data.

What is this animal? In the aboriginal's language it means

'I do not understand.' They answered so the question of the explorers.

‘Aggregate imputation’

It is not always possible to impute well enough, but it is good to know something about the missingness categories. One possibility is to analyze missingness at aggregate level. We give here an example from the European Social Survey in which the objective income has been one demanding variable. Its quality in the first three rounds (2002-2006) was fairly bad or the values were even completely missing in some countries but after that the quality has been improved much. The strategy since Round 4 has been to use 10 categories by deciles of each country. Such income categories are enough for most analysis although the ESS documents indicate the deciles in currencies as well.

‘Aggregate imputation’ 2

Nevertheless, there are four missing categories: ‘Refusal’, ‘Don’t know’, ’Other missing’ and ‘No answer’. The first three ones are in the questionnaire but the last was added later since some missing values were still found. Table 10.1 gives the counts of the respondents in each category for 14 countries in the ESS Round 7. We see that the ‘No answer’ group is very small but the other three are about as big as the proper income decile groups.

The missingness rate of the objective income is fairly high, 14.7 per cent. This thus means that we will lose this number at minimum in all multivariate analysis in which objective income is included. Hence it would be nice to know something about those missingness categories. One strategy is to use such auxiliary variables without missingness or the missingness rate is low. We test here one variable without missingness, that is, age, and the other with a low missingness, that is, subjective income. Its missingness rate is 0.8 per cent. This latter variable could be considered to be close to objective income at the same time, and hence it is a good auxiliary variable.

‘Aggregate imputation’ 3

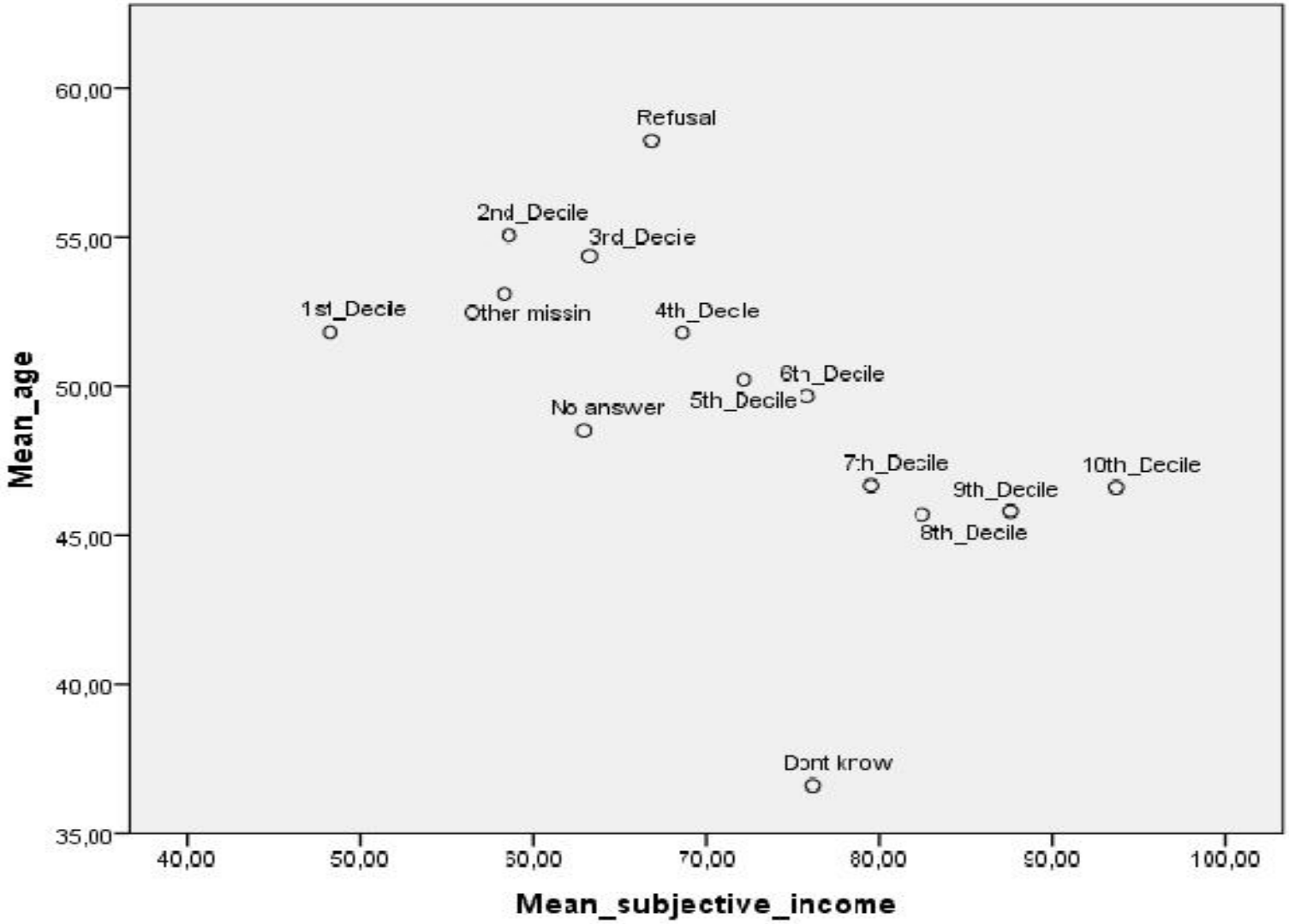
The subjective income of the ESS is computed from the variable:
“Which of the descriptions on this card comes closest to how you feel about your household’s income nowadays?”
Living comfortably on present income = 1
Coping on present income = 2
Finding it difficult on present income = 3
Finding it very difficult on present income = 4
(Don’t know) = 8.

We rescaled this variable linearly so that it varies from 0 (very difficult) to 100 (comfortably). Table includes also the averages of this variable and age by objective income groups. Now it is possible to see what types of groups are those missingness categories. The figure facilitates this comparison.

Examination of missing objective income groups of the ESS Round 7 of 14 countries. Subjective income and age have been tested as auxiliary aggregate variables

Objective income group	Respondents	Mean	
		Subjective income	Age
1st_Decile	2083	48.2	51.8
2nd_Decile	2329	58.6	55.1
3rd_Decile	2280	63.3	54.4
4th_Decile	2439	68.6	51.8
5th_Decile	2421	72.1	50.2
6th_Decile	2432	75.8	49.7
7th_Decile	2448	79.5	46.7
8th_Decile	2301	82.5	45.7
9th_Decile	1832	87.6	45.8
10th_Decile	1885	93.7	46.6
Don’t know	1645	76.2	36.6
No answer	19	62.9	48.5
Other missing	2051	58.3	53.1
Refusal	2056	66.8	58.2

Graphical illustration of Table 10.1



Estimates of the happiness model for Income without meta data

income 1	-0.975014250	B	0.03288081	-29.65	<.0001
income 2	-0.550124177	B	0.03241294	-16.97	<.0001
income 3	-0.301514727	B	0.03204129	-9.41	<.0001
income 4	-0.156128051	B	0.03198495	-4.88	<.0001
income 5	-0.062633503	B	0.03206008	-1.95	0.0507
income 6	0.107356628	B	0.03239465	3.31	0.0009
income 7	0.160667760	B	0.03253089	4.94	<.0001
income 8	0.253324182	B	0.03278936	7.73	<.0001
income 9	0.372888113	B	0.03362707	11.09	<.0001
income 10	0.513224969	B	0.03355944	15.29	<.0001
income 77	0.037282517	B	0.03002931	1.24	0.2144
income 88	-0.049621112	B	0.03157858	-1.57	0.1161
income 99	0.133865231	B	0.09192052	1.46	0.1453
income 9999	0.000000000	B			

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 (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.

Preserving associations in the case of missing data

(iv) Apply a real-donor methodology so that the whole (or essential) pattern of the variable values has been borrowed 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. The correlations might be reasonable any way.

End comments

The ‘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’ (as 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 complete data.

Thank you Kiitos