Imputation methods for a binary variable

Seppo LAAKSONEN

University of Helsinki, Finland

Seppo.Laaksonen@Helsinki.Fi

Sorry for not speaking French but I understand a little bit and enjoy to visit Paris.

If you wish to read my recent course material on imputation methods, please go my Helsinki University Website where the first material is concerned this course in English: https://wiki.helsinki.fi/display/SocStats/Laaksonen%2C+Seppo.

Is it possible to impute this area in Rome?
A difficult task at least, or impossible?
But many imputations are still done even though not being happy with all quality criteria. But if one criterion is fulfilled, it is a minimum.

Content

What is imputation, its purpose, concepts
Missingness mechanisms
Most common tools for missing item handling without real imputations
Missingness pattern
Targets for imputation
Imputation process

Imputation model
Imputation task

Single and multiple imputation: Bayesian and Non-Bayesian

Test data and results

Conclusion

Imputation process

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

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

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

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

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

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

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

This presentation is focused on (iii), (iv) and (v).

Imputation model

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 (or response) indicator of this variable.

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

Imputation model 2

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 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 3

The model (i) is concerned a continuous variable but this is not included in this presentation.

In this case the most common model is linear regression or its logarithmic version. Recently also mixed models are going to 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.

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 4

The model (ii) is concerned a binary variable (1 = responded, 0 = not) but the same model can used for the model (i) if the dependent variable is binary (e.g. 1 = poor, 0 = non-poor). Both of these are applied here.

You know how to work with the binary model to predict. First you have to choose a link function, that can be:
-logit
-probit
-complementary log-log (cll)
-log-log (ll).
Most imputations using either ll or cll lead to the same results. We use cll in our examples. There are no dramatic differences in explaining models between those link functions but of course with some. Imputation thus requires to use this model for predicting the response propensities for all units (respondents and non-respondents). That is, the first outputs are those values between (0, 1). But we test also a linear link in which case the predicted values can be outside (0, 1).

Imputation task

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

(a) Model-donor approach in which case the imputed values are computed deterministically (or stochastically) from the predicted values (adding noise) of the model.

(b) Real-donor approach in which case the predicted values (or 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).

Nearness metrics is needed for real-donor methods

There are several strategies for searching a nearest or near observed value to replace a missing value. I will give the strategies in the case of a binary variable after next sessions.

This metrics can be deterministic or stochastic. The previous one leads to single imputation (SI) only but the second one gives opportunity to ***multiple*** imputation (MI) that is next explained at general level.

Single and multiple imputation

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 2

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 3

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. I have to say that I am not convinced about all the solutions?

Summary for imputation of a categorical variable

This below illustration is for imputations of categorical variables

	(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 is can not be ideally any linear regression model. These cases are considered in following pages.

That is, use the same nearness metrics in imputing missing values as above.

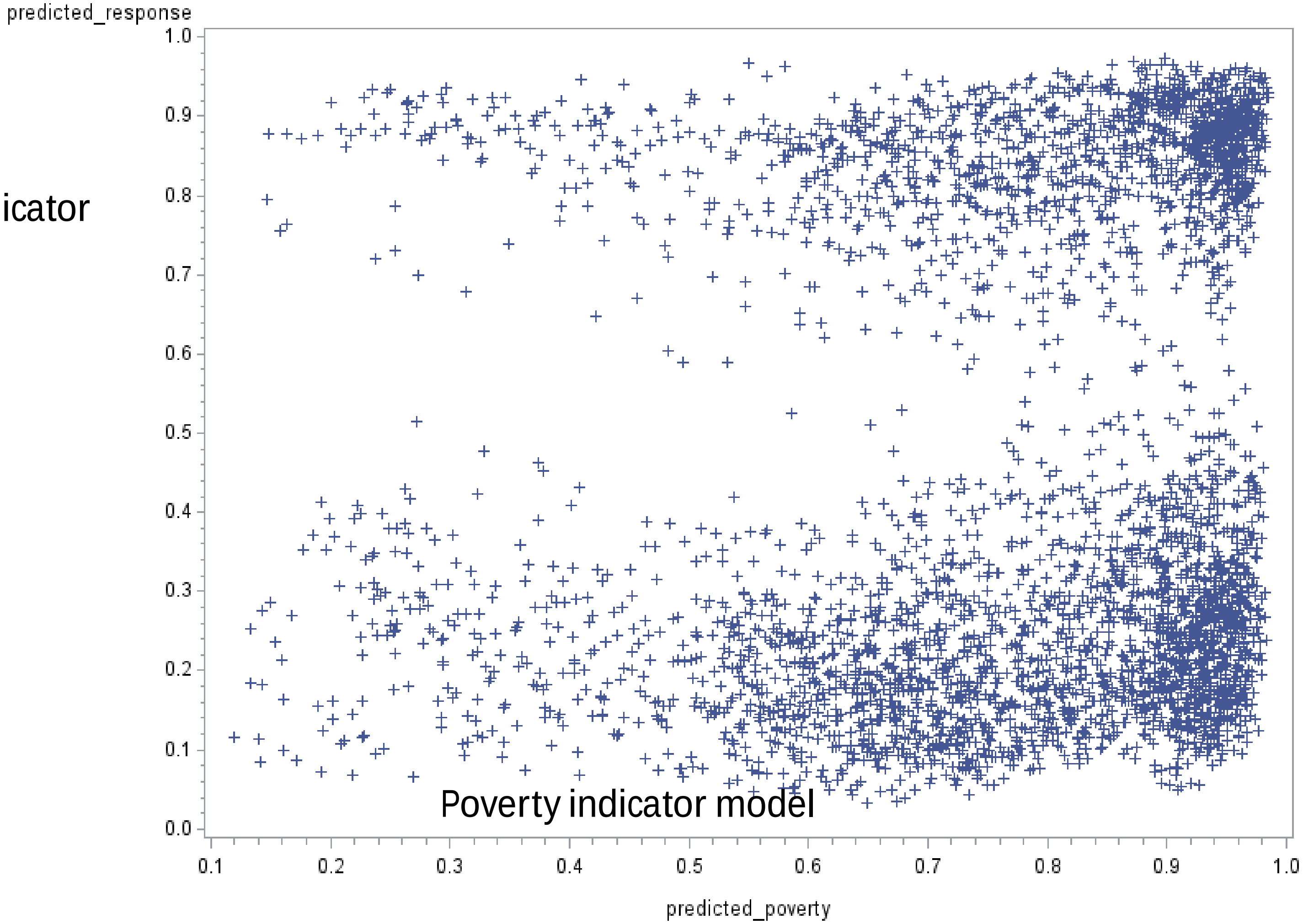
Predictability

has a big role in succeeding imputations. This has been implemented by predicted values. They are used both in model-donor and real-donor methods, so that they should be available both for the respondents (units with observed values) and for the nonrespondents (unit without observed values), and in the same way estimated.

In the case of multiple imputations (and in some single imputations) the initial predicted (deterministic) values are transformed so that a good random noise term are added. This gives opportunity to non-Bayesian multiple imputation without any specific Bayesian rules.

In order to illustrate the deterministic predicted values of some imputation models, following pages include a pattern of such values so that two different variables are in each scatter plot. You see that these values are sometimes close to each other but sometimes not so. As far as real-donor methods are concerned, the most important point is the nearness of these values.

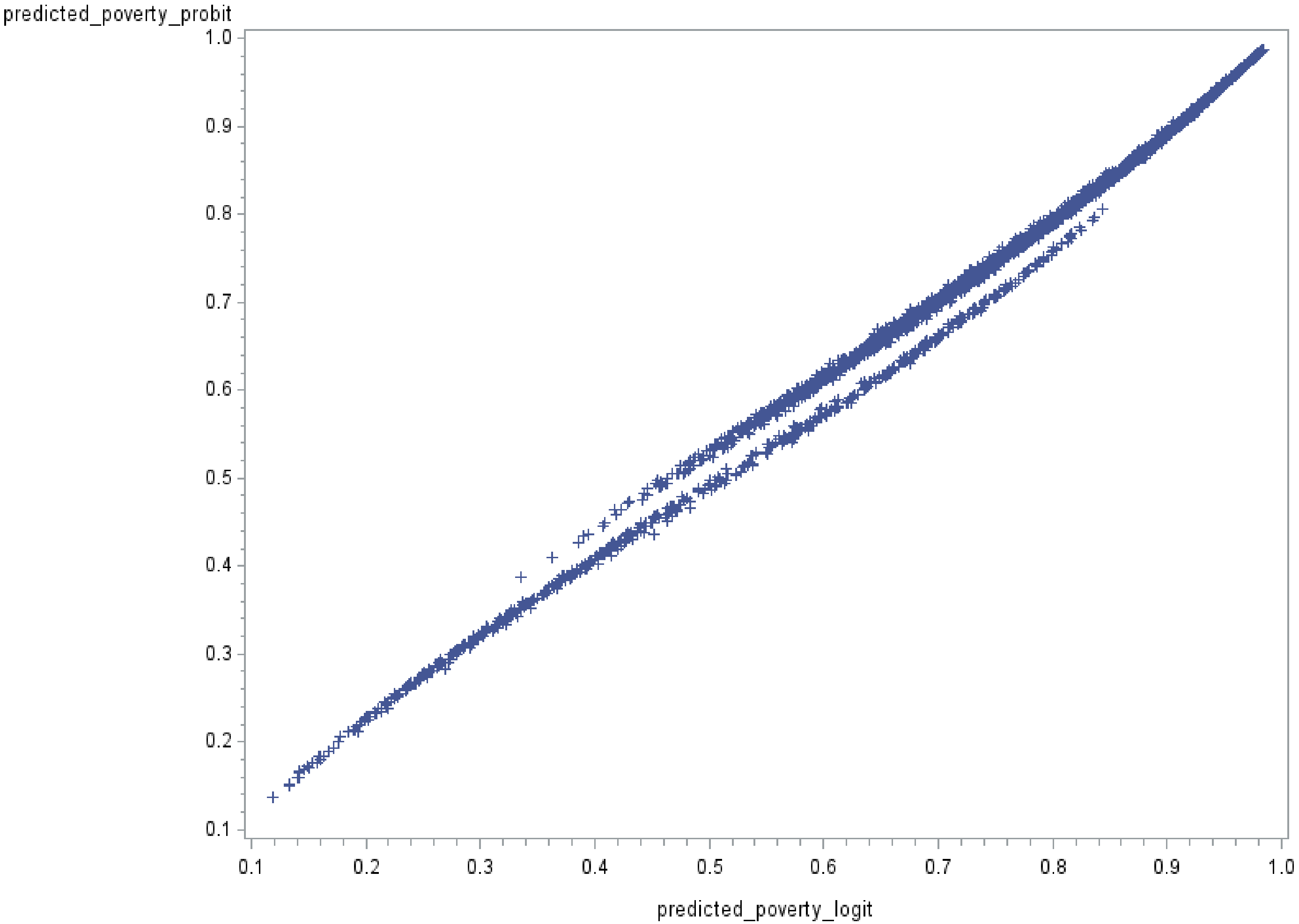
Two different imputation models with the logit link function



Response indicator model

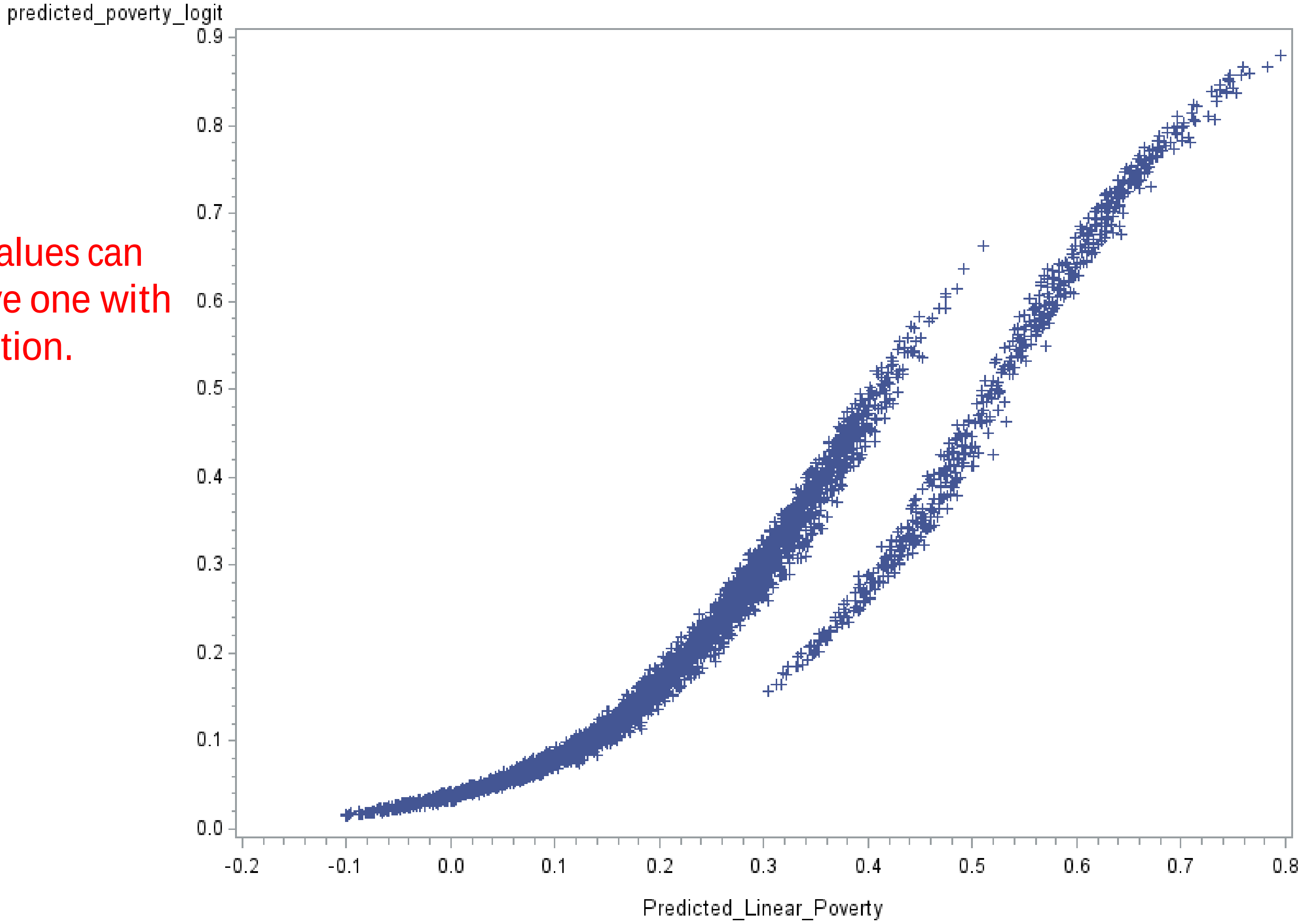
Poverty indicator model

Two poverty indicator models with the two link functions (probit and logit)

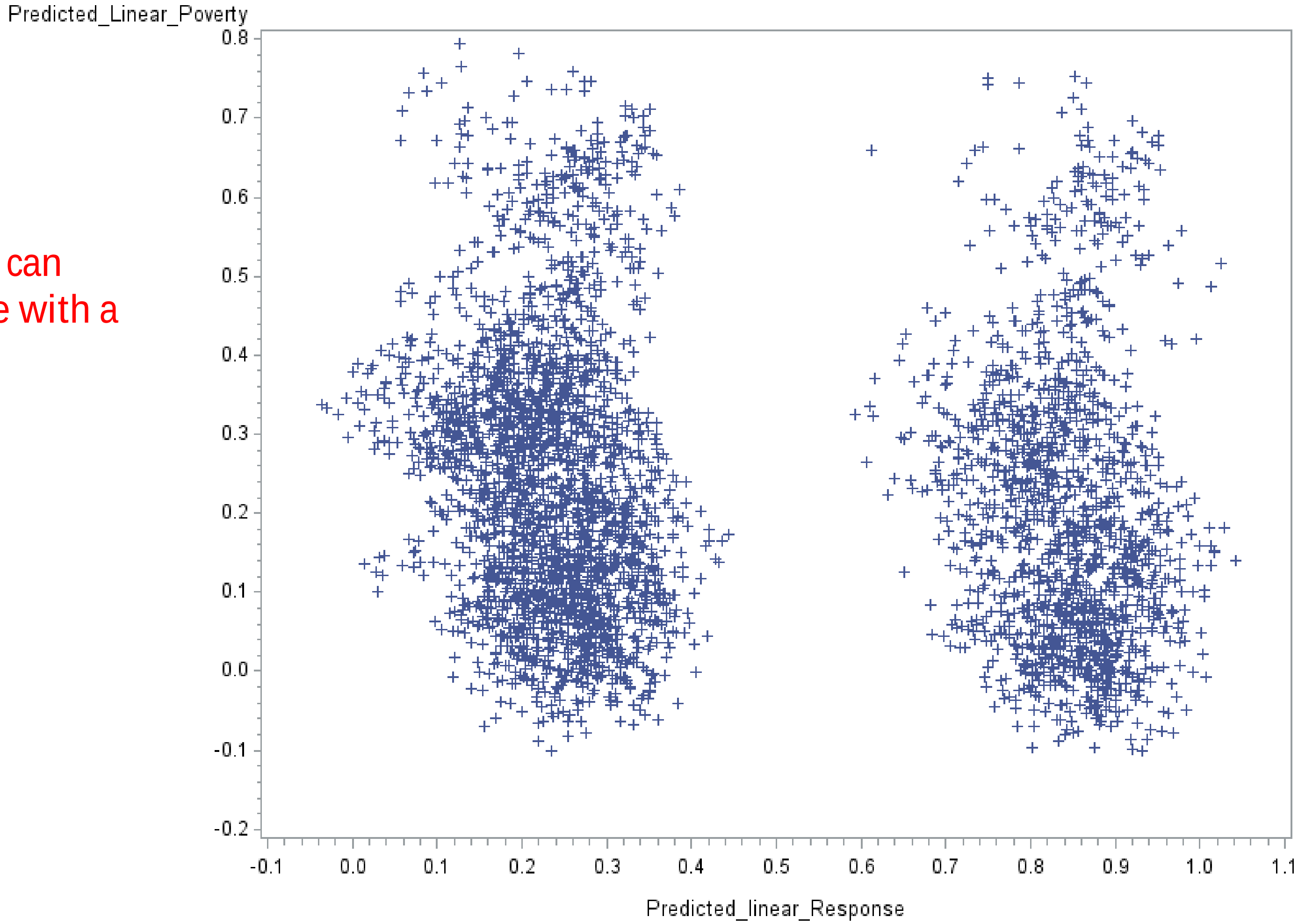


Two poverty indicator models with the two link functions (logit and linear)

Obs. Predicted values can negative or above one with a linear link function.



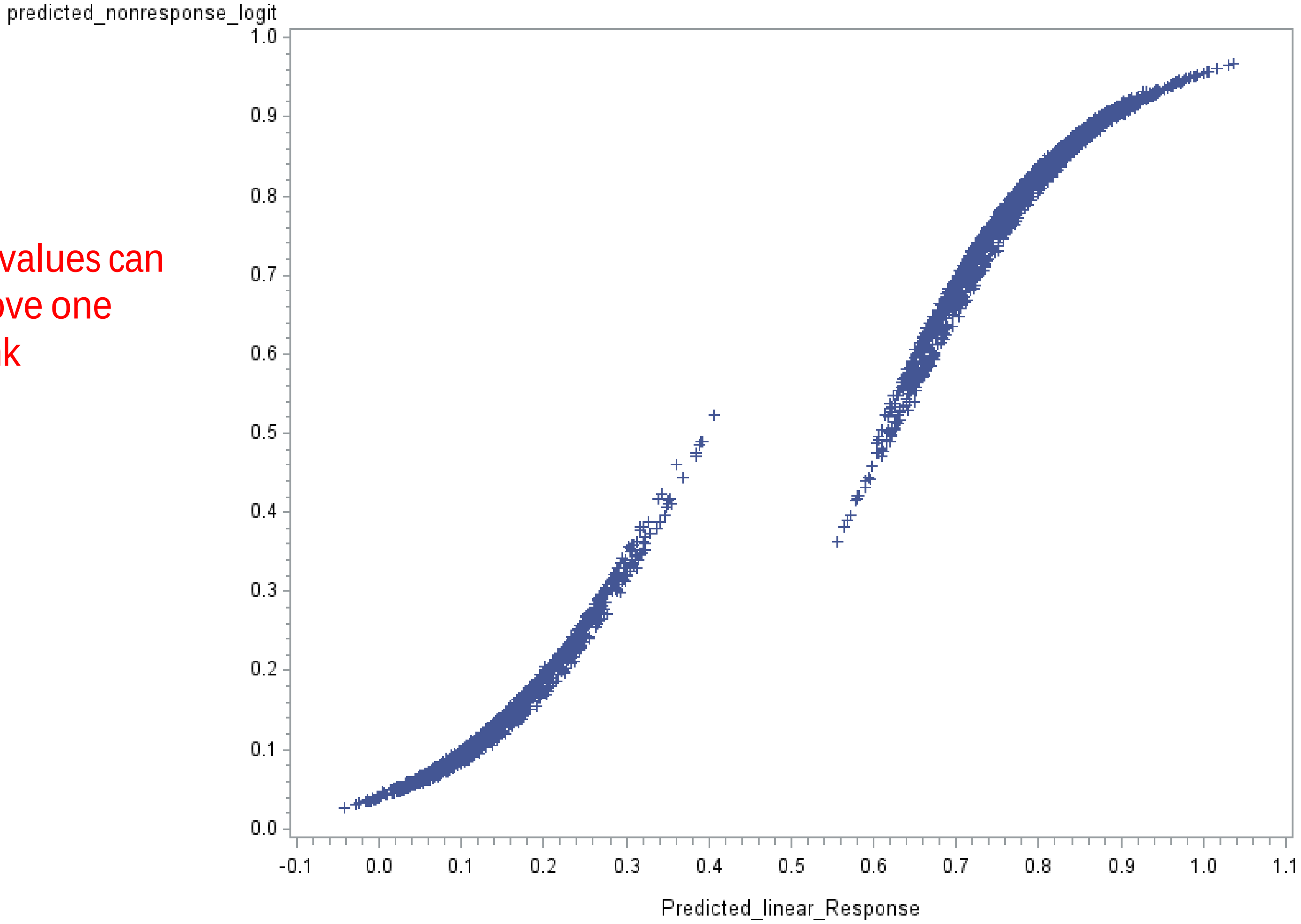
Two models with linear link function for poverty and response



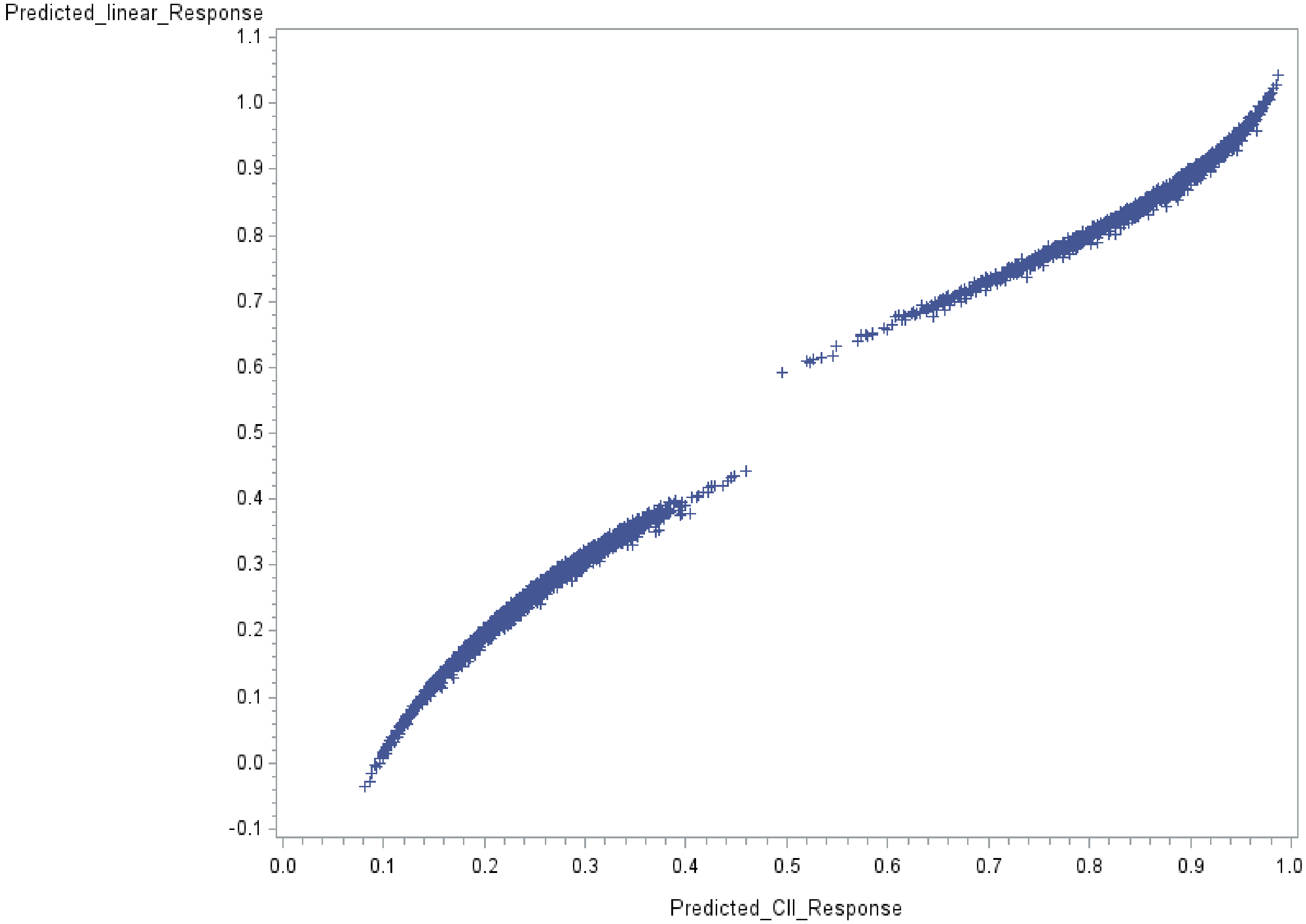
Obs. Predicted values can negative or above one with a linear link function.

Two response indicator models with the two link functions: logit and linear

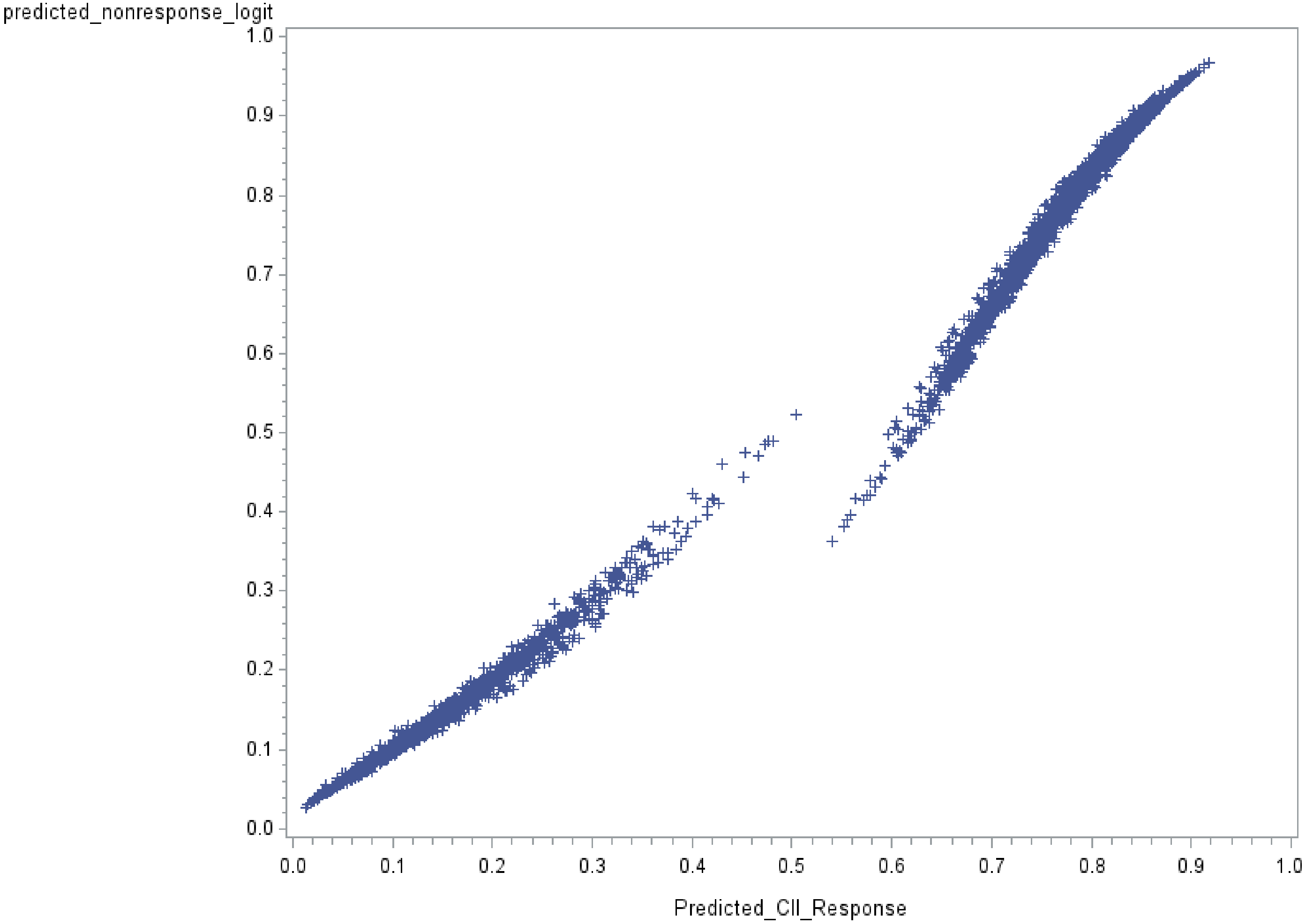
Obs. Predicted values can negative or above one with a linear link function.



Two response models: linear and cll link function



Two response indicator models: logit and cll



Non-Bayesian imputations today

Here the imputation model is the poverty binary model.

Our single model-donor imputation follows a Bernoulli approach so that we first calculate the predicted values of each unit *k* with a missing value, let say p_k. On the other hand, we create a uniformly distributed random number within (0, 1) for the same units, let say u_k. The imputed values are obtained as follows:

- if $u_k > p_k$ then *y_imputed* = 1, otherwise *y_imputed* = 0;

This strategy thus gives model-donor imputed values with desired link function. It is needed to be careful in order to get the correct codes 1 and 0 so that 1 = poor, and 0 = non-poor, for instance. When changing a random seed number 10 times, 10 non-Bayesian model-donor multiply imputed values are obtained.

Non-Bayesian real-donor imputations after either a poverty model or after a response model.

This single real-donor imputation method is completed to multiple imputation using the following procedure:

1. The predicted values p_k are estimated as for single imputation.
2. The standard error of the predicted values is estimated and included as a constant value in the data set, let say *stderr*.
3. The normally distributed random numbers are created with the zero mean and the standard deviation equal to one, let say *u_nor*.
4. The new predicted values for searching for a nearest neighbor are = p_k + *u_nor*stderr*.

When changing a random seed number 10 times, 10 non-Bayesian real-donor multiply imputed values are obtained.

Test data

It is initially derived from a EU project Euredit but here it is re-formulated so that using the poverty indicator is created using the yearly income. The missingness mechanism is the same as in the Euredit that is not known.

We can compare the results easily since the true values are known. The data set consists of a good number of auxiliary variables. We use the two sets of these (a smaller and larger number) but the main results are from the full set. These auxiliary variables are not very well related with poverty (and less well with income), that is a good thing since imputations are not easy. Fortunately, as far as our categorical poverty variable is concerned, all imputation methods are better than the benchmarking (either mean imputation that is the simplest model-donor method or random real-donor method that is called random hot decking in historical literature).

The two tables of the results are on the following pages.

Non-Bayesian Single imputation results for two models (poorer and richer pattern of covariates)

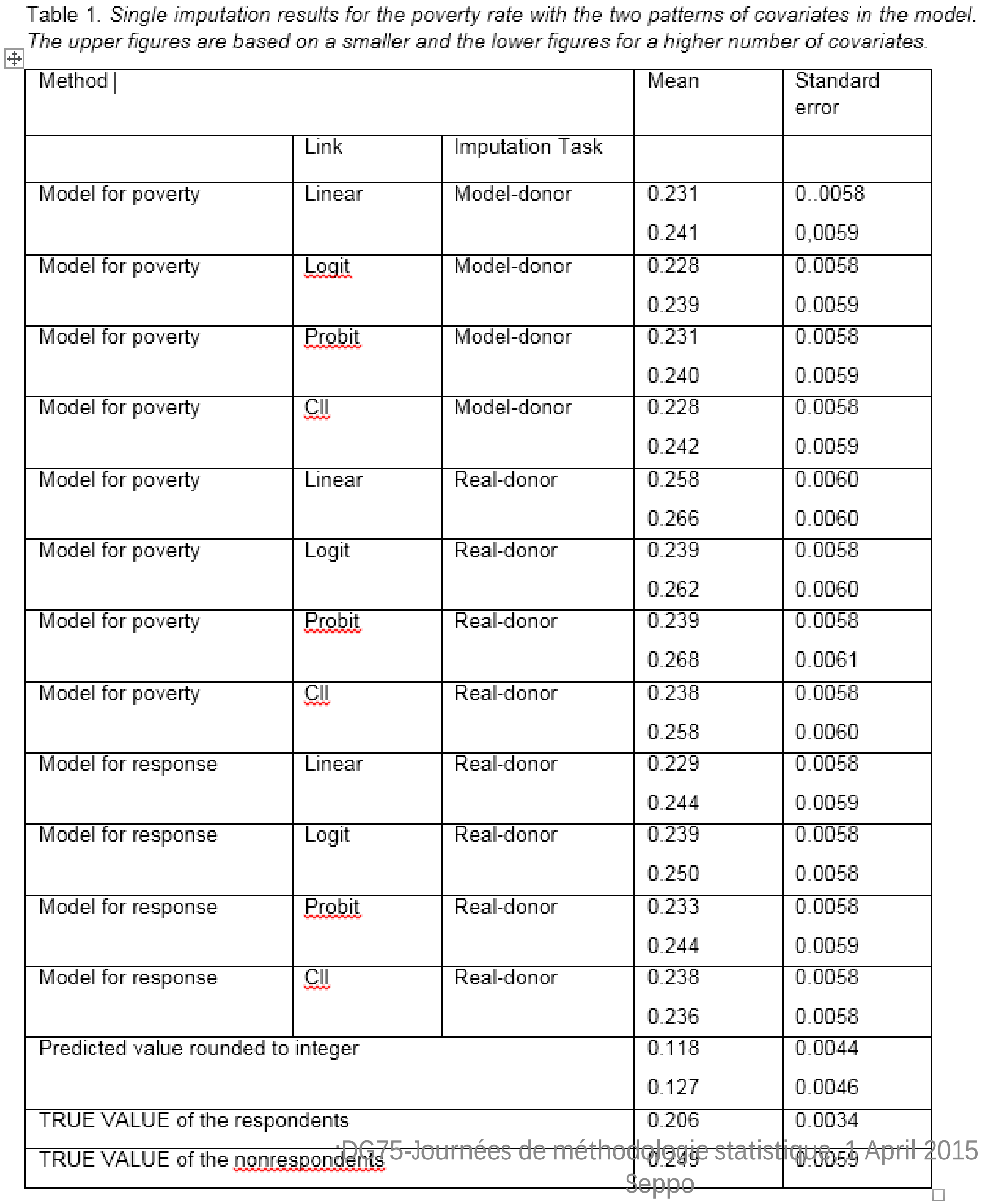


Table 1. *Single imputation results for the poverty rate with the two patterns of covariates in the model. The upper figures are based on a smaller and the lower figures for a higher number of covariates.*

Method			Mean	Standard error
	Link	Imputation Task		
Model for poverty	Linear	Model-donor	0.231 0.241	0..0058 0,0059
Model for poverty	Logit	Model-donor	0.228 0.239	0.0058 0.0059
Model for poverty	Probit	Model-donor	0.231 0.240	0.0058 0.0059
Model for poverty	Cll	Model-donor	0.228 0.242	0.0058 0.0059
Model for poverty	Linear	Real-donor	0.258 0.266	0.0060 0.0060
Model for poverty	Logit	Real-donor	0.239 0.262	0.0058 0.0060
Model for poverty	Probit	Real-donor	0.239 0.268	0.0058 0.0061
Model for poverty	Cll	Real-donor	0.238 0.258	0.0058 0.0060
Model for response	Linear	Real-donor	0.229 0.244	0.0058 0.0059
Model for response	Logit	Real-donor	0.239 0.250	0.0058 0.0058
Model for response	Probit	Real-donor	0.233 0.244	0.0058 0.0059
Model for response	Cll	Real-donor	0.238 0.236	0.0058 0.0058
Predicted value rounded to integer			0.118 0.127	0.0044 0.0046
TRUE VALUE of the respondents			0.206	0.0034
TRUE VALUE of the nonrespondents			0.249	0.0059

The best results are obtained using real-donor methods with response model although the difference is not substantial to model-donor methods that are obtained with the same random number.

The results with different link functions vary but any clear conclusion cannot be given. Interestingly, linear link function works fairly well as well.

Non-Bayesian and Bayesian multiple imputation

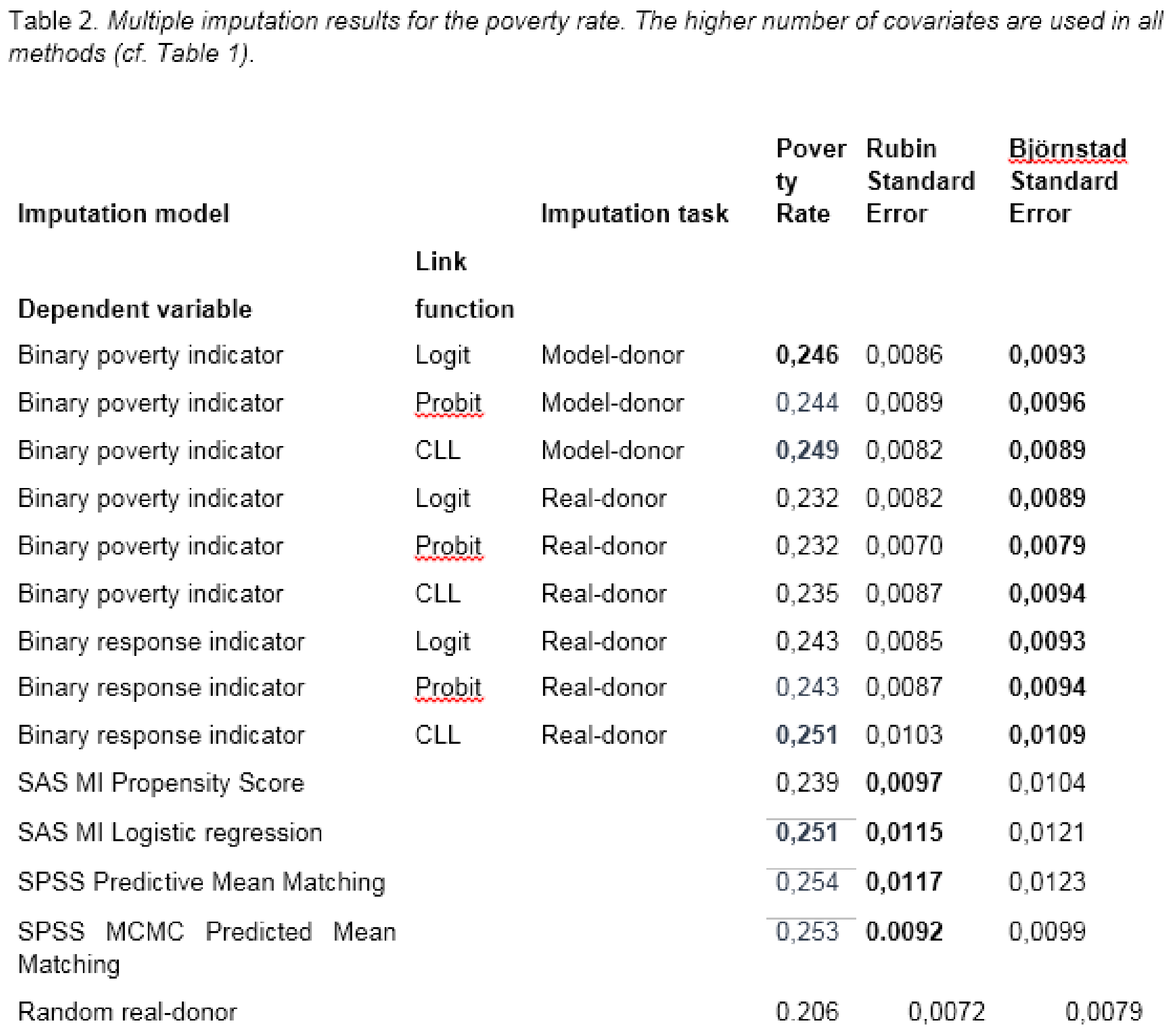


Table 2. *Multiple imputation results for the poverty rate. The higher number of covariates are used in all methods (cf. Table 1).*

Imputation model		Imputation task	Poverty Rate	Rubin Standard Error	Björnstad Standard Error
Dependent variable	Link function				
Binary poverty indicator	Logit	Model-donor	**0,246**	0,0086	**0,0093**
Binary poverty indicator	Probit	Model-donor	0,244	0,0089	**0,0096**
Binary poverty indicator	CLL	Model-donor	**0,249**	0,0082	**0,0089**
Binary poverty indicator	Logit	Real-donor	0,232	0,0082	**0,0089**
Binary poverty indicator	Probit	Real-donor	0,232	0,0070	**0,0079**
Binary poverty indicator	CLL	Real-donor	0,235	0,0087	**0,0094**
Binary response indicator	Logit	Real-donor	0,243	0,0085	**0,0093**
Binary response indicator	Probit	Real-donor	0,243	0,0087	**0,0094**
Binary response indicator	CLL	Real-donor	**0,251**	0,0103	**0,0109**
SAS MI Propensity Score			0,239	**0,0097**	0,0104
SAS MI Logistic regression			**0,251**	**0,0115**	0,0121
SPSS Predictive Mean Matching			0,254	**0,0117**	0,0123
SPSS MCMC Predicted Mean Matching			0,253	**0.0092**	0,0099
Random real-donor			0.206	0,0072	0,0079

Most non-Bayesian methods work a bit better than Bayesian ones in point estimates. Standard errors of Bayesian methods are slightly higher than those of non-Bayesian methods. We could discuss whether the standard errors of non-Bayesian methods should be calculated using Björnstad's formula whereas these of Bayesian methods using Rubin's formula. If so, the standard errors would be closer to each other, but this is not any suggestion, it is a discussion point only. On the other hand, the standard errors of model-donor methods are lowest nevertheless, and this could be a value of imputation methods as well. Our study thus suggests to use this methodology for binary variables whenever it is possible.

Conclusion

Single imputations are much used in survey institutes but they are often very simple and deterministic. The theory behind them can be called non-Bayesian or frequentist. Their specific advantage is to get relatively easily standard errors if some data are imputed. Since the imputation is an additional factor of uncertainty, multiple imputation (MI) can be considered to be a useful tool for statisticians.

The initial theory behind multiple imputations is Bayesian. Consequently, MI methods using standard software packages like SAS and SPPS are implemented much following Rubin’s framework. A big question is whether MI methods could work under a non-Bayesian framework as well. The focus of this study has been to examine non-Bayesian techniques and tools for multiple or repeated imputations when a binary variable is attempted to impute. We have found several competitive strategies respectively so that the imputation consists of the two main stages, an imputation model and an imputation task respectively.

Conclusion 2

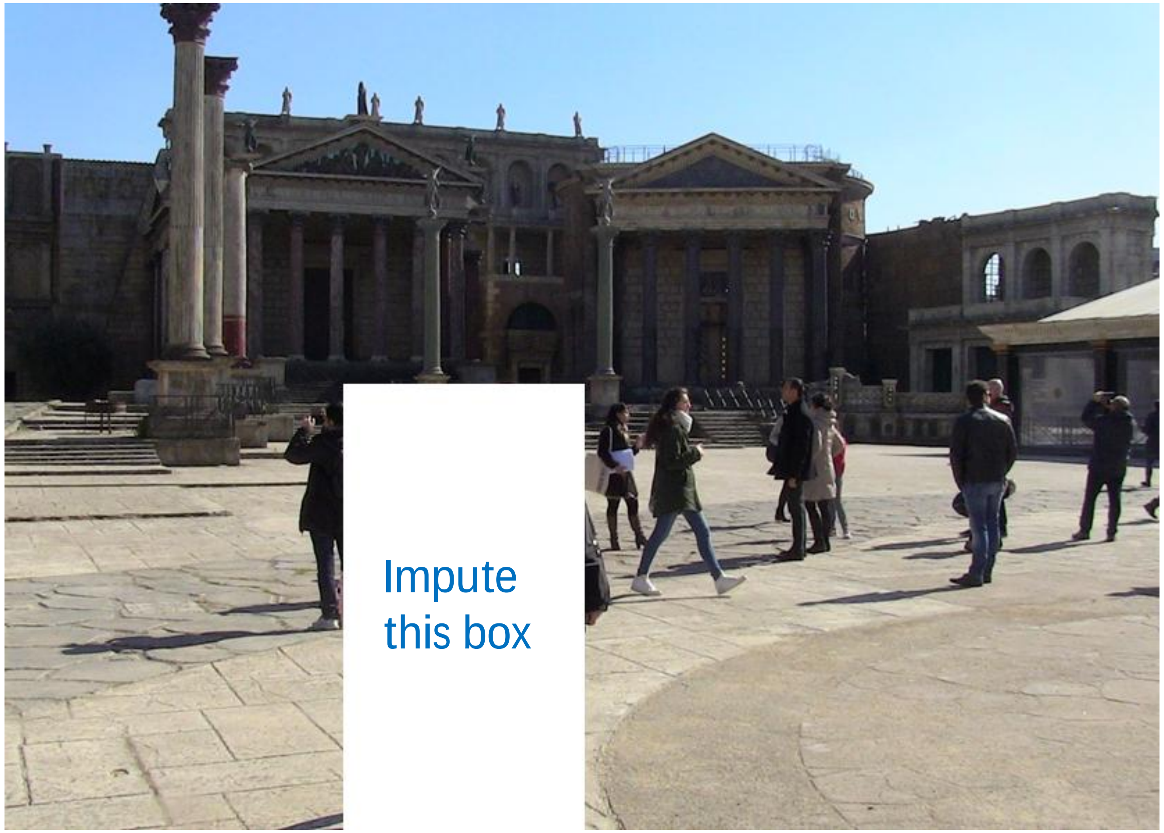
The comparisons with ordinary Bayesian methods of SPSS and SAS suggest that these much used software package methods are not superior to non-Bayesian alternatives. Our results even suggest that some non-Bayesian methods are better than Bayesians. We cannot say surely why this is the case. It seems that there are in Bayesian tools some additional technical elements that do not improve anything. Hence an ordinary user have to apply them as a black box.

When following appropriate non-Bayesian tools available to each situation in a better strategy, a user can see easily how each method works and revise its implementation if needed. This is a big advantage in imputations in general since the missing data replacements should be tailored to each particular case, not done automatically unless the quality of the method has not been checked well in advance.

This is one attempt to impute a part of the Old Rome

Merci pour votre attention

Cinecittà, Rome 2015



Impute this box