Chapter 11

Imputation methods for single variables

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 including preliminary ideas to exploit these. This is concerned those variables particularly that are not in the survey data set, thus such variables that also used for the unit nonrespondents, if they are needed to download in the same sets where are ordinary survey variables.

(iii) Imputation model(s): one or more models are need for each imputation. At this stage, the model should be specified, estimated and outputs saved for the further use.

(iv) Imputation task(s): outputs of the imputation models and other tools are required to impute the desired missing values. It is possible that a new editing is needed if the imputed values do not fit with edit rules.

(v) Estimation: point-estimates so that imputed values are used as well; in addition variance estimation = sampling variance plus imputation variance.

(vi) Creation of the completed data set or several ones. This includes good meta data and information about imputed values (if possible flagging values that are imputed as para data). The documentation of the imputation process and methods should be available as well. All details do not important to include in public files but everything should available inside the survey institution.

Imputation model

There are **two options** to determine the imputation model:

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

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

The imputation model (ii) 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:

either

A. The variable being imputed

or

B. The missingness indicator of the variable being imputed.

Imputation task

The two alternatives in general can be exploited after the imputation model has been estimated:

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

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
A either the variable being imputed itself	Yes	Yes
B the missingness indicator of this variable	No	Yes

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)*.

Nearness metrics of real-donor methods

Main case:

The most common metrics is derived from the predicted values of the binary regression model (then 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 we want to mention a common tool from the Imputation book by Rubin (1987/2004):

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

The simplest, not recommended model-donor method for continuous variables is such in which the imputation model is the linear regression without auxiliary variables. This corresponds to mean imputation in ordinary literature.

If the imputation model includes one categorical auxiliary variable, the method is better but how good it is? You can decide it from the next page graph. The example is for income as others. The auxiliary variable consists of four categories. This method is called cell-based mean imputation. It is not either recommended.

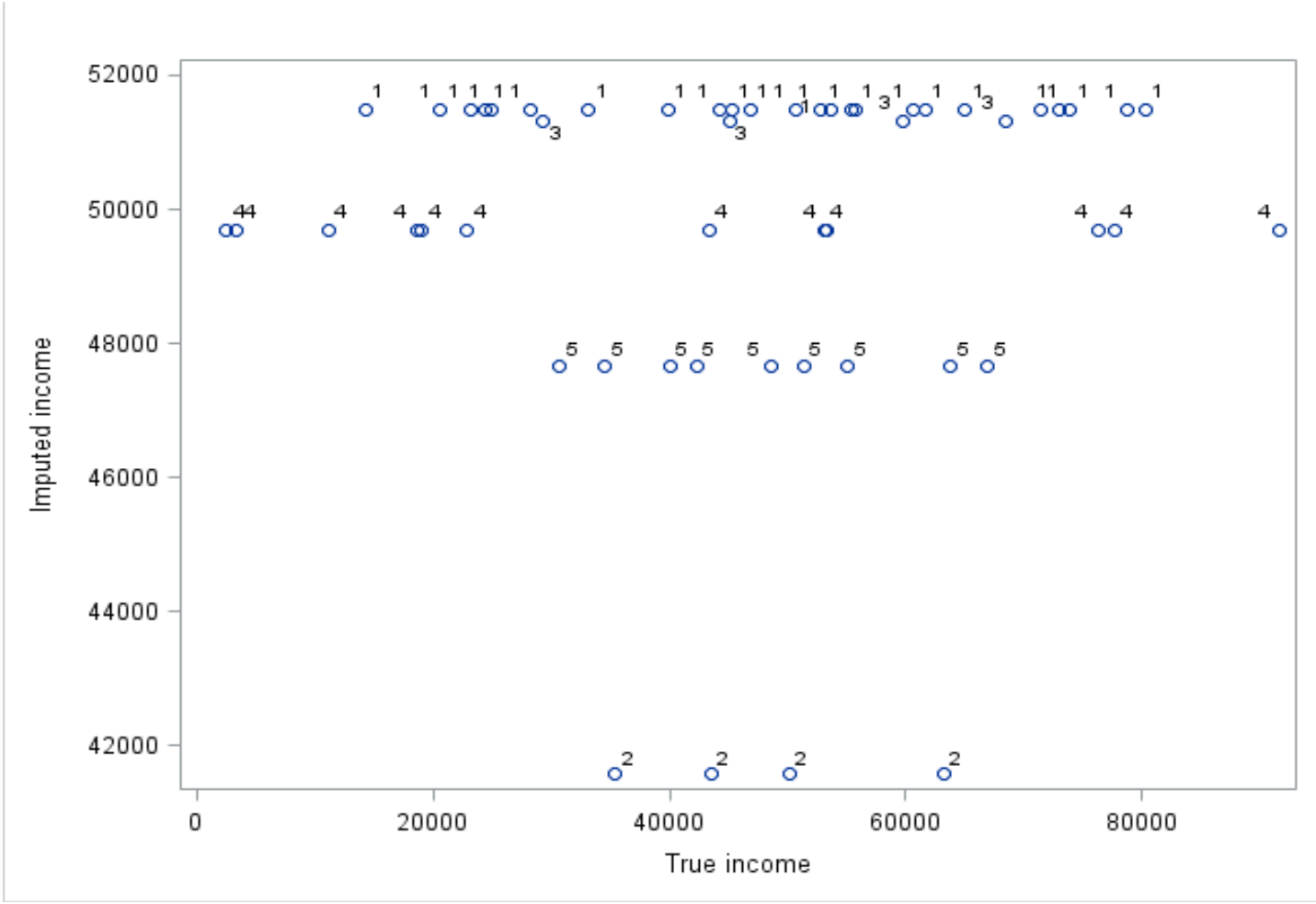
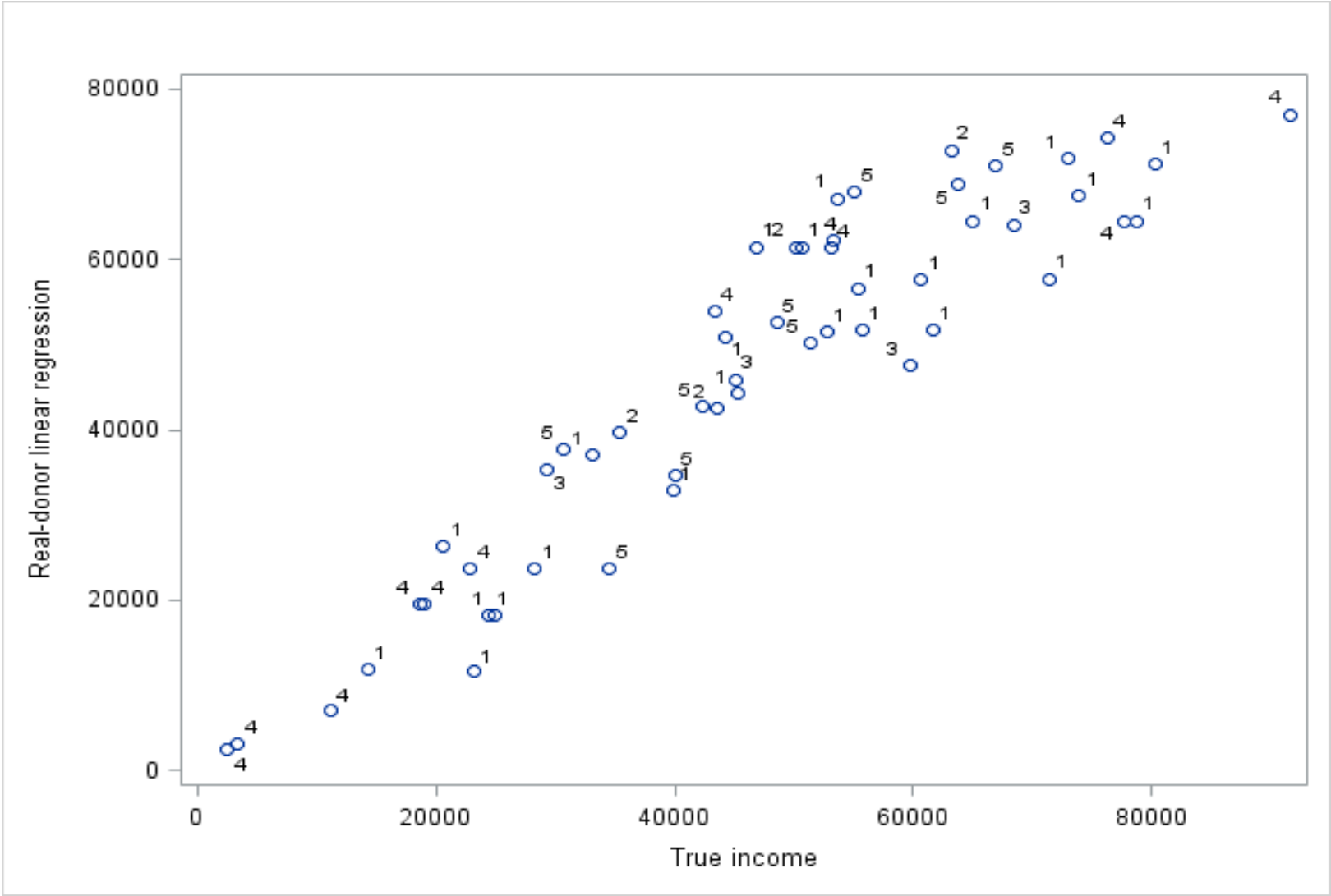


Figure 11.1. Imputed values only with the deterministic model-donor method when one categorical variable (region) is the auxiliary variable of the linear regression (called cell-based mean imputation)

I do not present all possible methods here (see the text) but one real-donor method only. It looks better, but if you carefully you see values that have been borrowed from the respondents.



The full list of the results is here. The bolded values are best. We thus can see the performance since we know true values.

Imputation model and explanatory variables	Imputation task	Imputations	Mean	Minimum	Maximum	CV
TRUE		53	46768	2475	91615	44.8
Linear regression without auxiliary variables = mean imputation	Model-donor	53	49454	49454	49454	0
Linear regression with region the explanatory variable	Model-donor	53	49675	41577	51489	5.5
Linear regression, with region and register income	Model-donor	53	48112	4245	**88032**	41.8
Linear regression with region and register income	Real-donor	53	**47000**	**2360**	76960	**43.4**
Logistic regression with region and register income	Real-donor	53	**46594**	**2360**	81215	47.8
Probit regression with region and register income	Real-donor	53	45634	**2360**	81215	49.3
Logarithmic linear regression with region and register income	Model-donor	53	44938	2962	77375	33.6
Logarithmic linear regression with region and register income	Real-donor	53	47619	**2360**	**97400**	**44.9**

Single Imputation (SI) and Multiple Imputation (MI)

The point estimates of MI are averages of all multiply imputed values (at least 3). (Rubin 1987/2004). It thus is easy if you have such values. The interval estimates are not as easy but we do not consider them in details. Look at the chapter. One example on next page however.

I did not tell this in training. Sorry. I thus gave you one scoring variable that is the average of five. As you see, this falsehood does not mean much for uncertainty. But it is good to tell finally.

PISA 'Multiple imputation'

The literacy scores for reading, science, mathematics and problem solving are unique in the public micro data set, but there are five 'plausible values for each student.' These have been calculated from the results of a number of exam tasks that are not exactly the same for each student. This means that the score includes an additional uncertainty. The PISA group has hence decided to give five different 'plausible values' in the data set.

When calculating the estimates this can be taken into account as an additional uncertainty component. Nevertheless, the means and other point estimates can without problems calculated as the average of those five 'plausible values' as in the case of multiple imputation. The impact of the variation in the means due to those five values is not very big, fortunately, as the following graph shows for some countries. This thus means that it is not catastrophic to omit the component of MI.

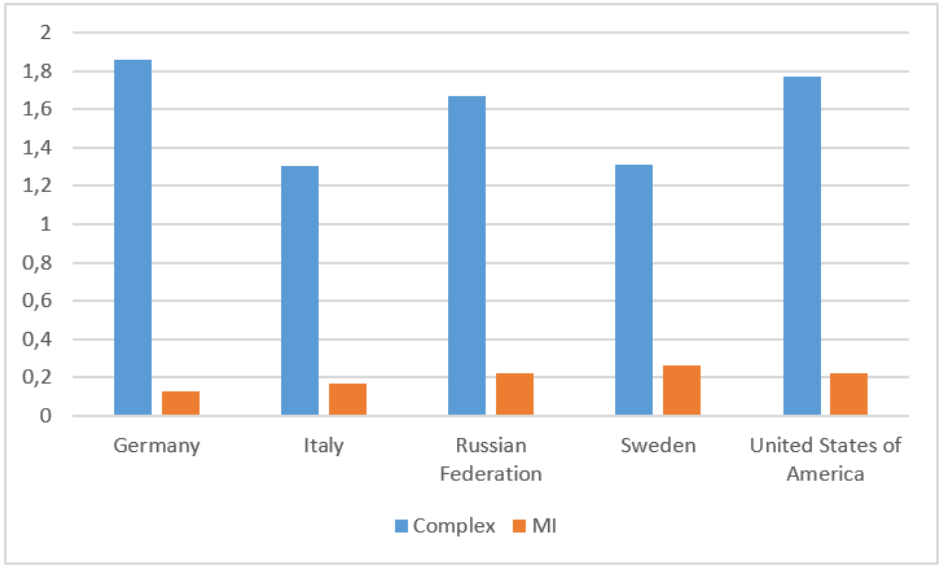


Figure 11.1 The coefficient of variation of the mean estimate for problem solving scores in the 2012 PISA, per cent; Complex = using three survey instruments (Stratum, Cluster and Weights)

General conclusion:

- Imputation thus should be enough successful to be used. It is often difficult know how successful it is. Fortunately, best methods are usually fairly easy to recognize and distinguish from bad methods. The final decision depends much on the research target, and hence it is not automatically clear which criteria are most important.
- There are software's including SPSS that fairly easily can impute all missing values. I have tested both SPPS and SAS, and found that they are not trustable. The big reason is that any technical tool cannot know what are really needed in each task of the study.
- If your team is not good in imputing, do not impute at all and lose your data. Sometimes the results are not still catastrophic.