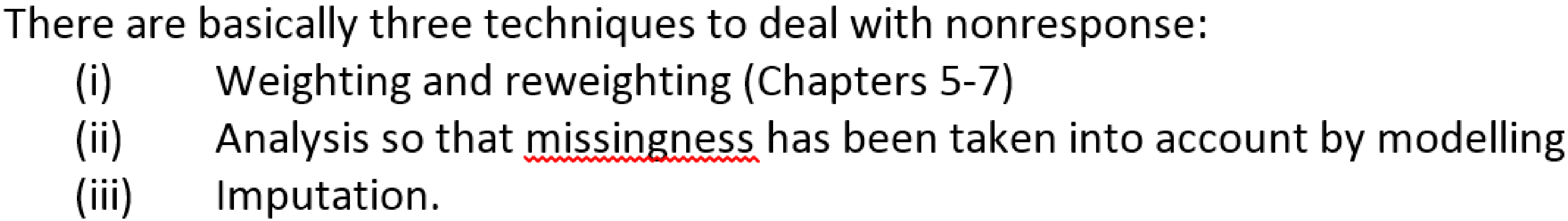
Chapter 10

Introduction to statistical Imputation



There are basically three techniques to deal with nonresponse:

(i) Weighting and reweighting (Chapters 5-7)
(ii) Analysis so that missingness has been taken into account by modelling
(iii) Imputation.

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 are used in analysis. In one-dimensional analysis we drop out only the missing values of this particular variable, but the ordinary multivariate analysis include such statistical units whose variable values are completely observed. Hence the data might reduce dramatically. Data deletion might work only if the response mechanism is MCAR.

We first determine in this chapter the term imputation and the targets of it, then present such data handling tools that do not require any proper imputation. The main aim of the chapter is to present an appropriate framework of the imputation and then applies it in practice in Chapter 11, including concrete real-life examples.

Imputation and its purpose

Imputation is to insert a value into the data in a more or less fabricated way ('best proxy') for following reasons:

- 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 to get a more valid estimate of the distribution (including percentiles, standard deviation and coefficient of variation),

- Since the existing value does not seem to be correct, and consequently, it is desired to get a more reliable value by replacing with a more plausible imputed value.

- Since the current value seems to be too confidential, that is, and this individual unit should be disclosed. Motivation: the fabricated (imputed) value can be considered as less problematic even though when told that it is no true value.

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 after that multiple imputations.

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.

- **to impute or not to impute?**

to give outsiders. The insiders are also more familiar with the data process. However, the most important consideration to impute is

- **The pattern of the imputed values should as good that the estimate using this partially imputed variable will be more valuable than without imputation. Thus if imputation is believed to be 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 and your client that imputation improves something. Note however that all users and clients are happy if the number of missing values have been declined.

Possible reasons 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 might be 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, in this case, imputation would be easier than in the case of nonignorable 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), something 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 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 ordinary high income respondents, it is not fatal if some values are missing. On the other

Targets for imputation should be specified clearly

It is rather clear

(i) That a user is happy if the imputed values are as close as possible to the correct/true values, that is, 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 some-what less required targets are more realistic in practice.

(ii) A user is still fairly happy if the distribution of the imputed values is close to the distribution obtained from true values (Success at distributional level). Of course this is hard to check but however easier than case (i).

(iii) The target to succeed at aggregate level is also satisfactory and specifically in statistical institutes or in other survey institutes where such estimates as average, total, ratio, median, point of decile and standard deviation are typical. These can be checked to some extent in repeated surveys particularly.

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

(v) Finally, success to preserve relationships (like correlations and co-variances) is also important in many studies.

What can be imputed due to missingness?

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. This is done in theoretical or simulation studies.

(iii) Unit non-response, all or some variables. This is called mass imputation. It is competitive to weighting methods. The purpose of mass imputation is to complete the data in order to estimate everything from one data set. This helps in getting all estimates consistent with each other. The success at aggregate level might be good enough.

(iv) **Item non-response. This is the most common case and we only present such examples.**

(v) Deficient and sensitive values. Quite common but we do not present any examples.

‘Aggregate imputation’ - Example

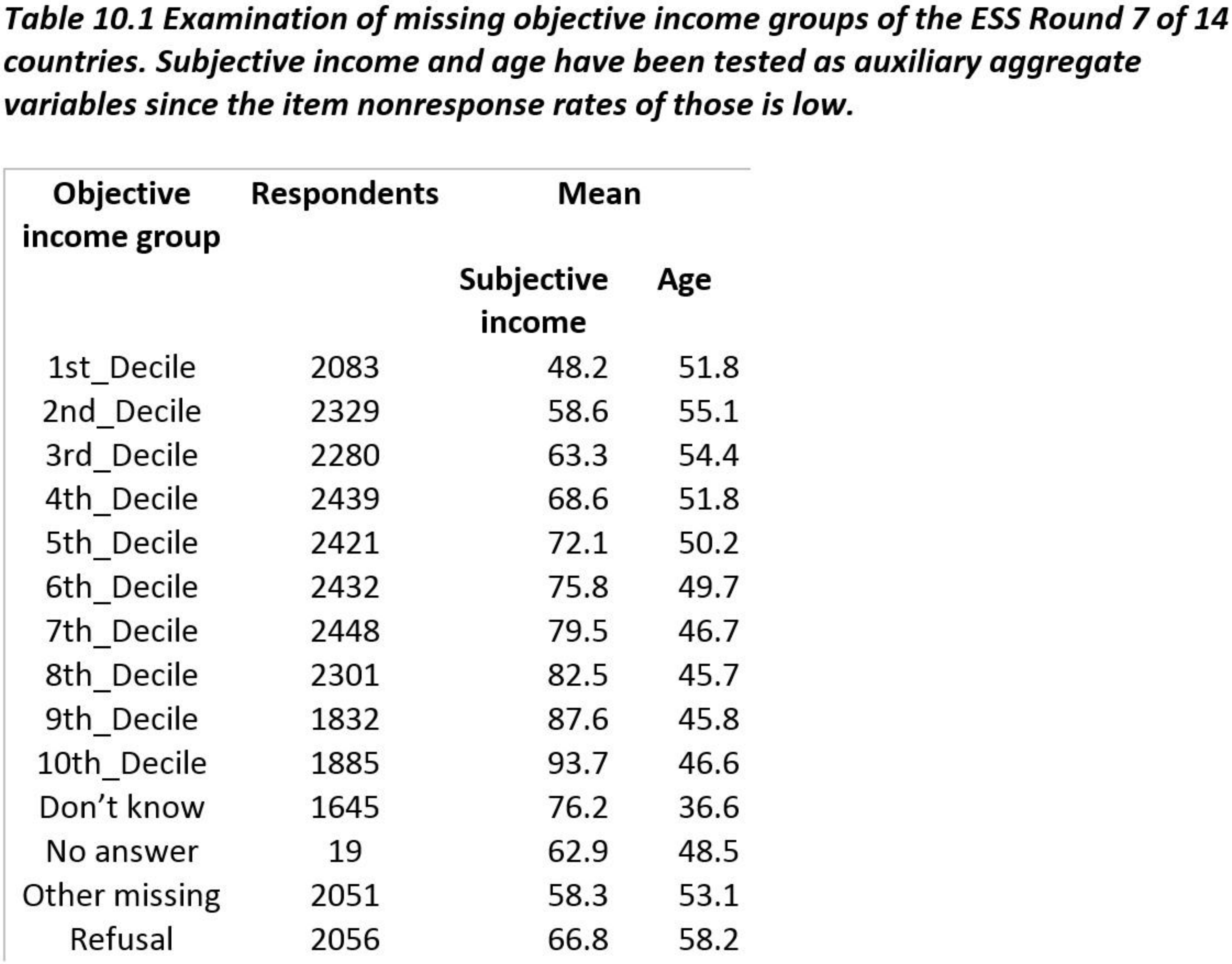
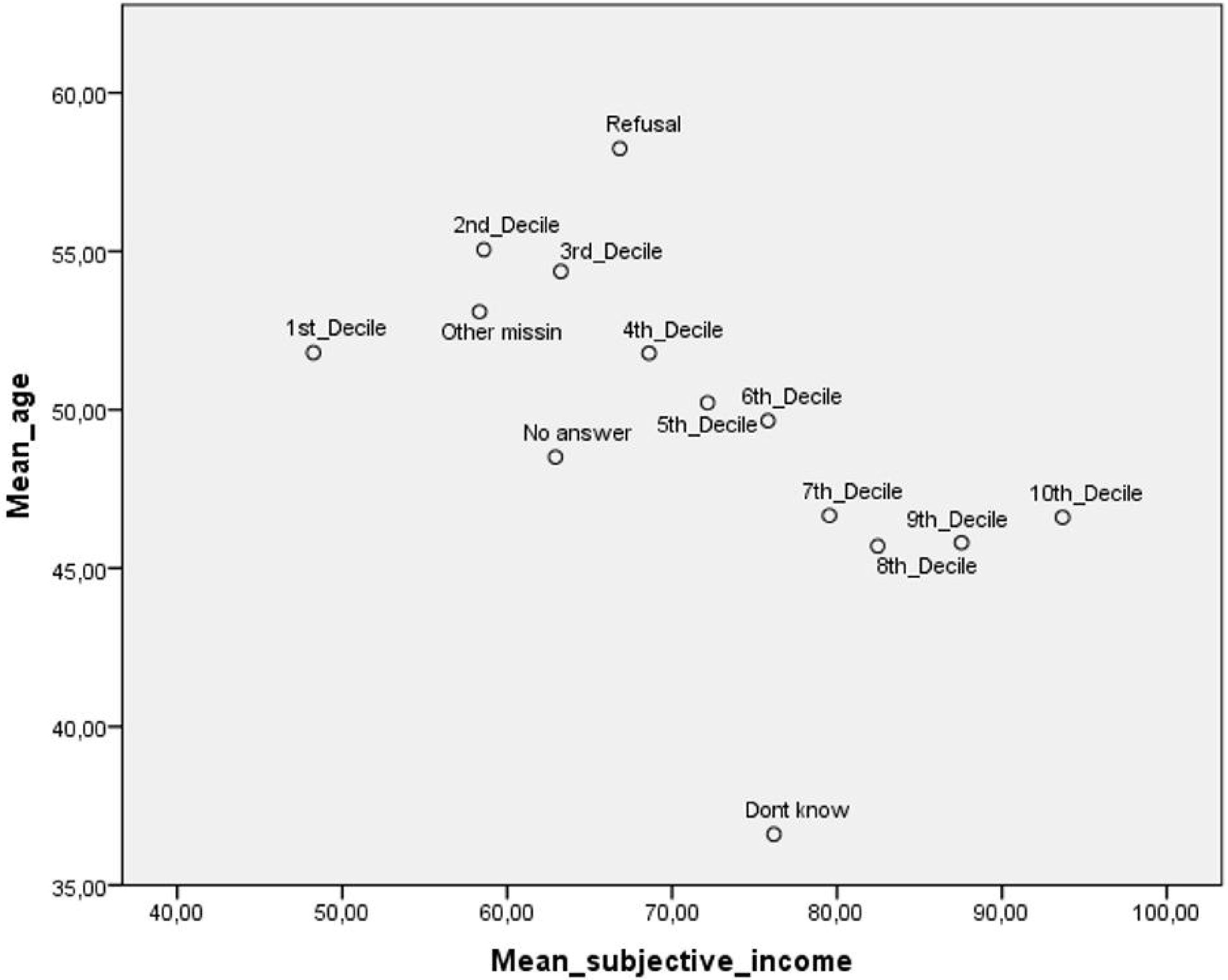


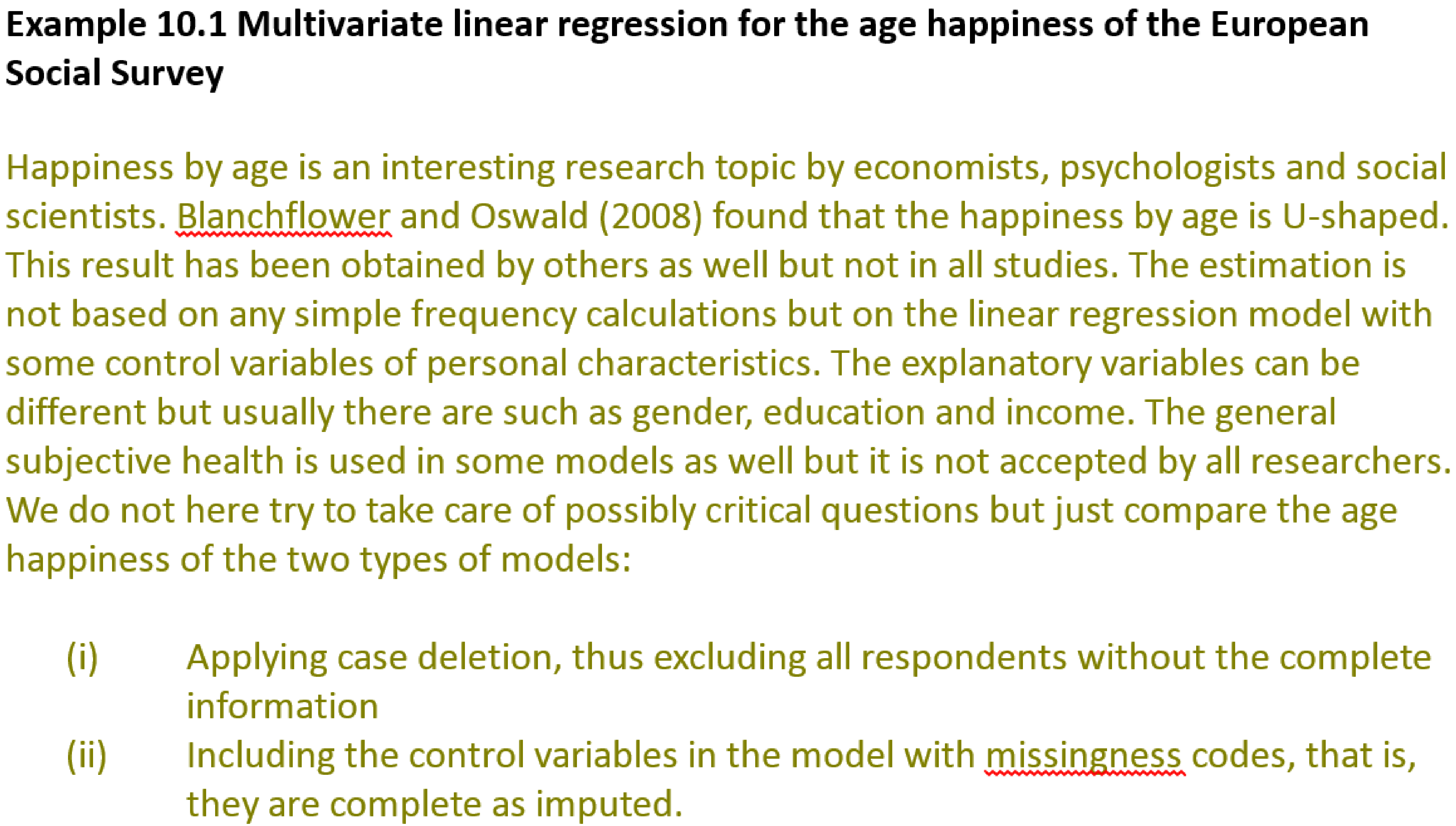
Table 10.1 Examination of missing objective income groups of the ESS Round 7 of 14 countries. Subjective income and age have been tested as auxiliary aggregate variables since the item nonresponse rates of those is low.

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

Figure 10.1 Graphical illustration of Table 10.1



No Imputation



Example 10.1 Multivariate linear regression for the age happiness of the European Social Survey

Happiness by age is an interesting research topic by economists, psychologists and social scientists. Blanchflower and Oswald (2008) found that the happiness by age is U-shaped. This result has been obtained by others as well but not in all studies. The estimation is not based on any simple frequency calculations but on the linear regression model with some control variables of personal characteristics. The explanatory variables can be different but usually there are such as gender, education and income. The general subjective health is used in some models as well but it is not accepted by all researchers. We do not here try to take care of possibly critical questions but just compare the age happiness of the two types of models:

(i) Applying case deletion, thus excluding all respondents without the complete information

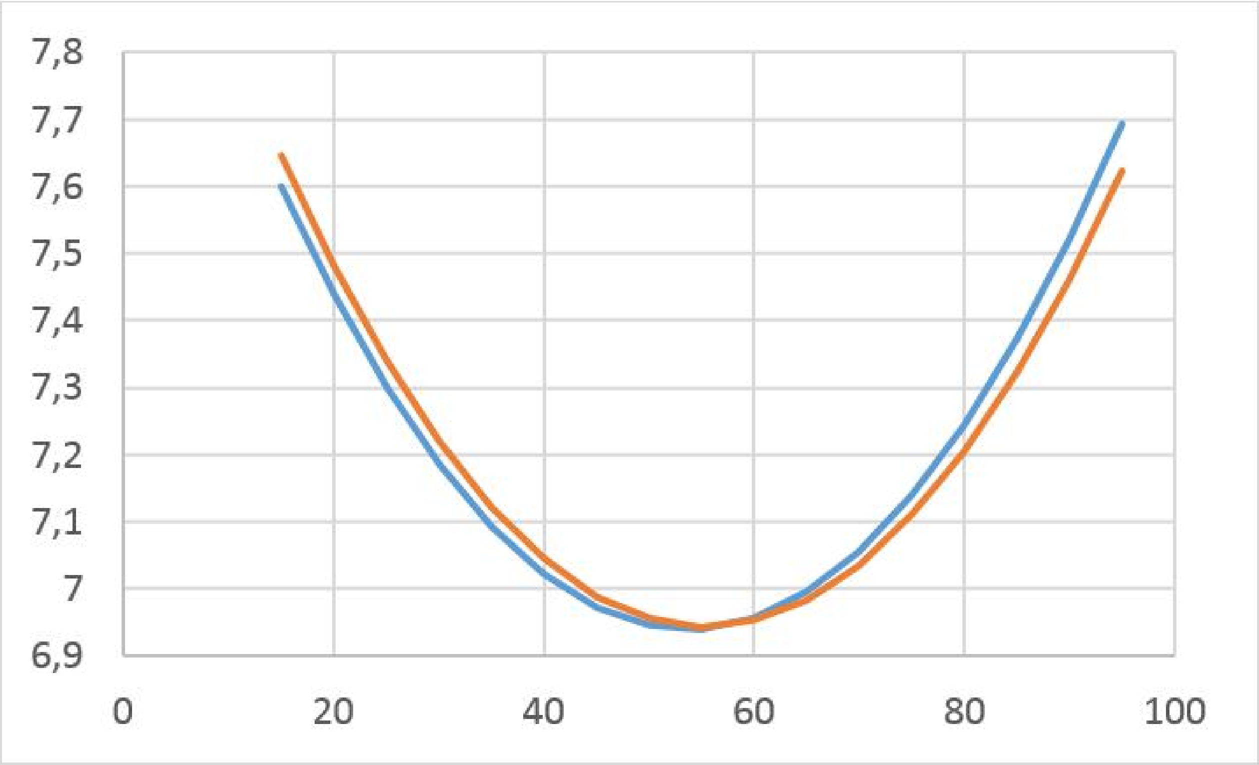
(ii) Including the control variables in the model with missingness codes, that is, they are complete as imputed.

No Imputation

Figure 10.3. Happiness by age in Round 7 of the European Social Survey for 20 countries

Red = Model with missingness categories in education and income

Blue = Model without non-respondents

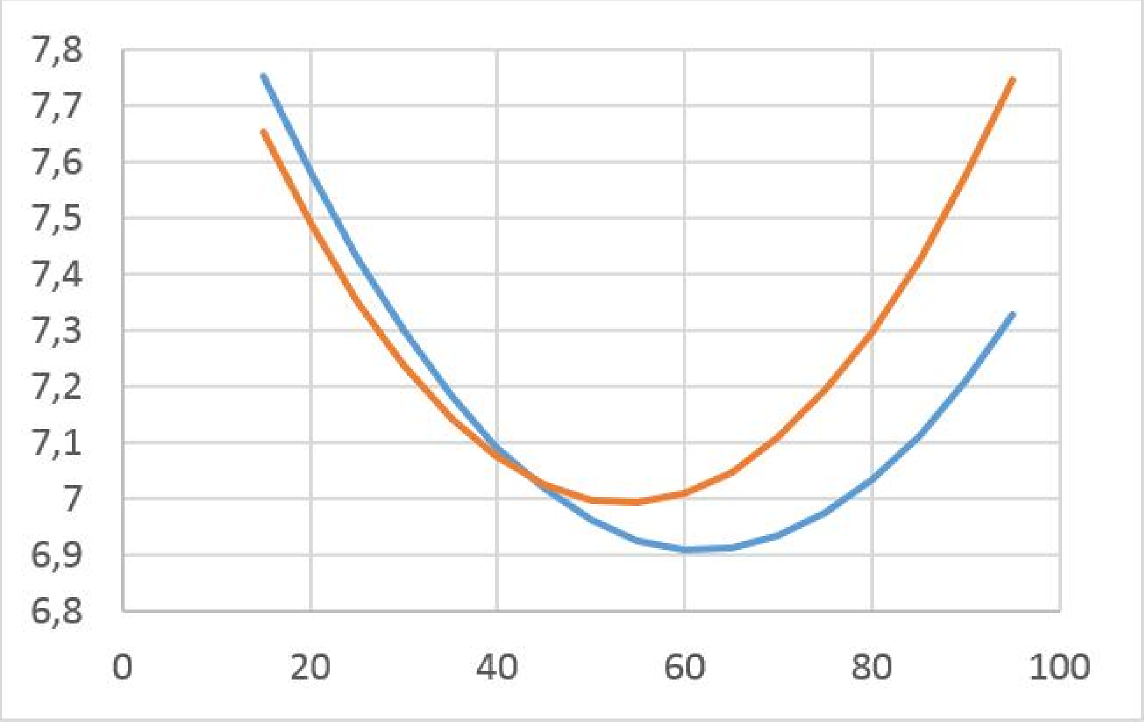


No Impu-tation

Figure 10.4. Happiness by age in Round 7 of the European Social Survey for 20 countries

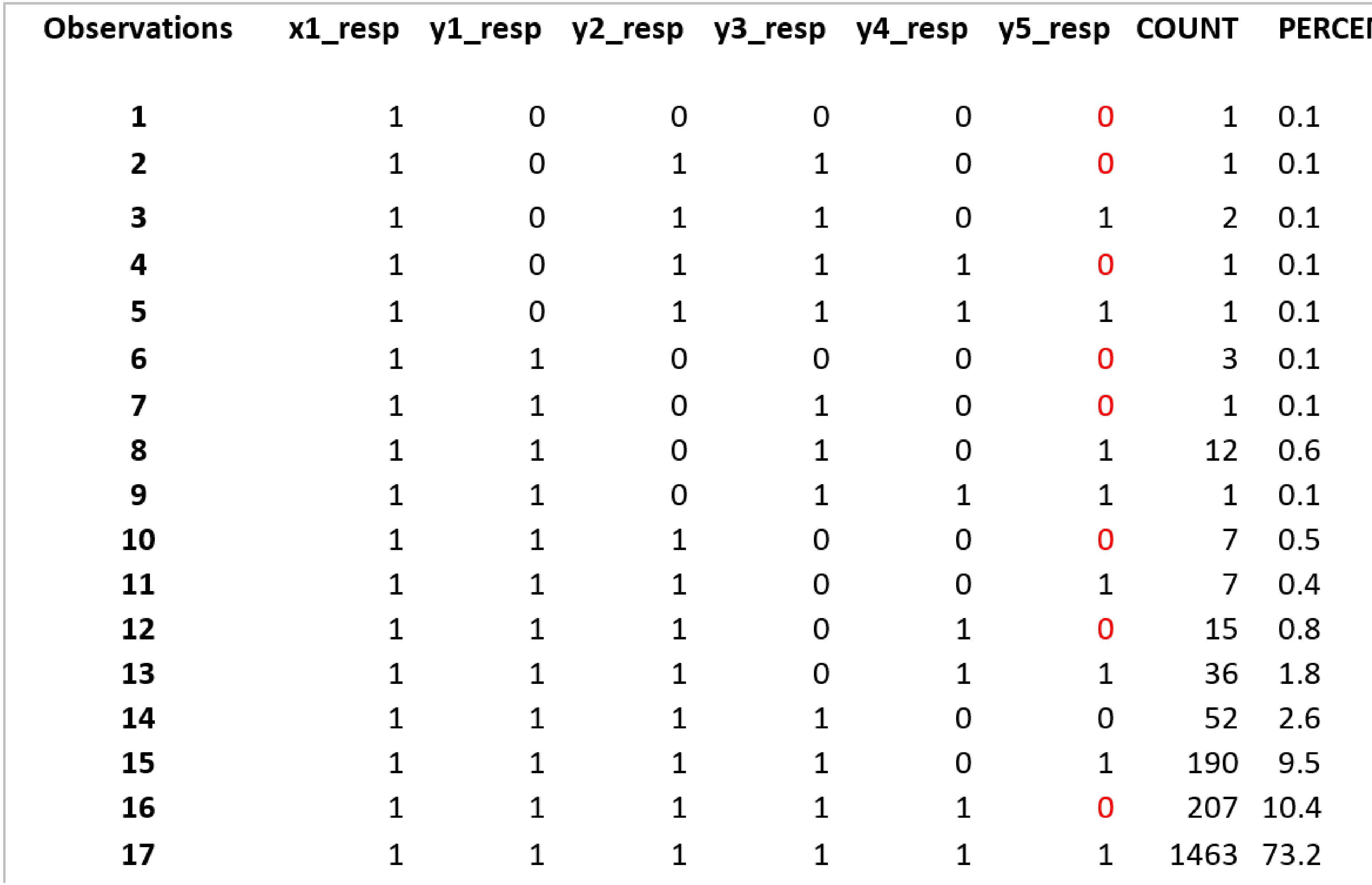
Red = complete respondents

Blue = respondents with missing codes in income and education



What to impute and which order?

Sequential imputation?



Observations	x1_resp	y1_resp	y2_resp	y3_resp	y4_resp	y5_resp	COUNT	PERCE
1	1	0	0	0	0	0	1	0.1
2	1	0	1	1	0	0	1	0.1
3	1	0	1	1	0	1	2	0.1
4	1	0	1	1	1	0	1	0.1
5	1	0	1	1	1	1	1	0.1
6	1	1	0	0	0	0	3	0.1
7	1	1	0	1	0	0	1	0.1
8	1	1	0	1	0	1	12	0.6
9	1	1	0	1	1	1	1	0.1
10	1	1	1	0	0	0	7	0.5
11	1	1	1	0	0	1	7	0.4
12	1	1	1	0	1	0	15	0.8
13	1	1	1	0	1	1	36	1.8
14	1	1	1	1	0	0	52	2.6
15	1	1	1	1	0	1	190	9.5
16	1	1	1	1	1	0	207	10.4
17	1	1	1	1	1	1	1463	73.2