Utrecht University

Department of Methodology and Statistics

Survey Data Analysis

Instructor: Dr. Peter Lugtig

Fall Semester 2021

European Social Survey, imputing categorical or not?

Daniel Anadria, Hsuan Lee, Florian van Leeuwen, Katja Sonntag

Word count: 5529

Contents

1 Introduction

The European Social Survey (ESS) is a cross-national survey studying attitudes, behaviours and beliefs of populations in various European countries. The survey first started in 2002 and an increasing number of European countries have joined the project since then. The ESS is conducted once every two years in form of face-to-face interviews (Jowell et al., 2006). The most recent edition of the ESS is round 9 from 2018, which is the focus of this report. As with most large surveys, the ESS round 9 suffers from missing data, especially in the form of item non-response.

In this report, we explore different imputation strategies as a possible solution for item non-response. In particular, we focus on the ESS round 9 for Italy, and our dependent variable *hinctnta* is a measure of income. Income questions are usually seen as sensitive, therefore some respondents may choose to leave this question unanswered. In fact, 44.44% of the Italian sample left this question unanswered – whether through an outspoken refusal to answer, stating that they don't know, or simply providing no answer.

Another consideration central to our report is how to treat the dependent variable, *hinctnta*, which is measured in deciles. Should it be considered continuous or categorical as it enters imputation? As will be shown, the conclusion that we reached is that while deciles make our dependent variable ordinal/categorical, they can be successfully imputed as a continuous variable preserving deciles through predictive mean matching. We perform two imputations, one from the perspective that the dependent variable is categorical, and another for the view that it is continuous. The difference in the average of *hincntna* after imputation is minimal between these two conditions, however, since we think there is a lot of value in preserving deciles in the imputed answer, we prefer to view our dependent variable as categorical.

2 Survey setting: the European Social Survey round 9

Our project is to impute the missing data for one country for variable hincinta. Firstly, we made a bar chart to compare all the countries surveyed by European Social Survey (ESS). Figure 2 shows that Italy among all the participant countries has the most missing data of the variable *hincinta*. For the sake of pursuing challenging and interesting work, we took Italy as our surveyed country.

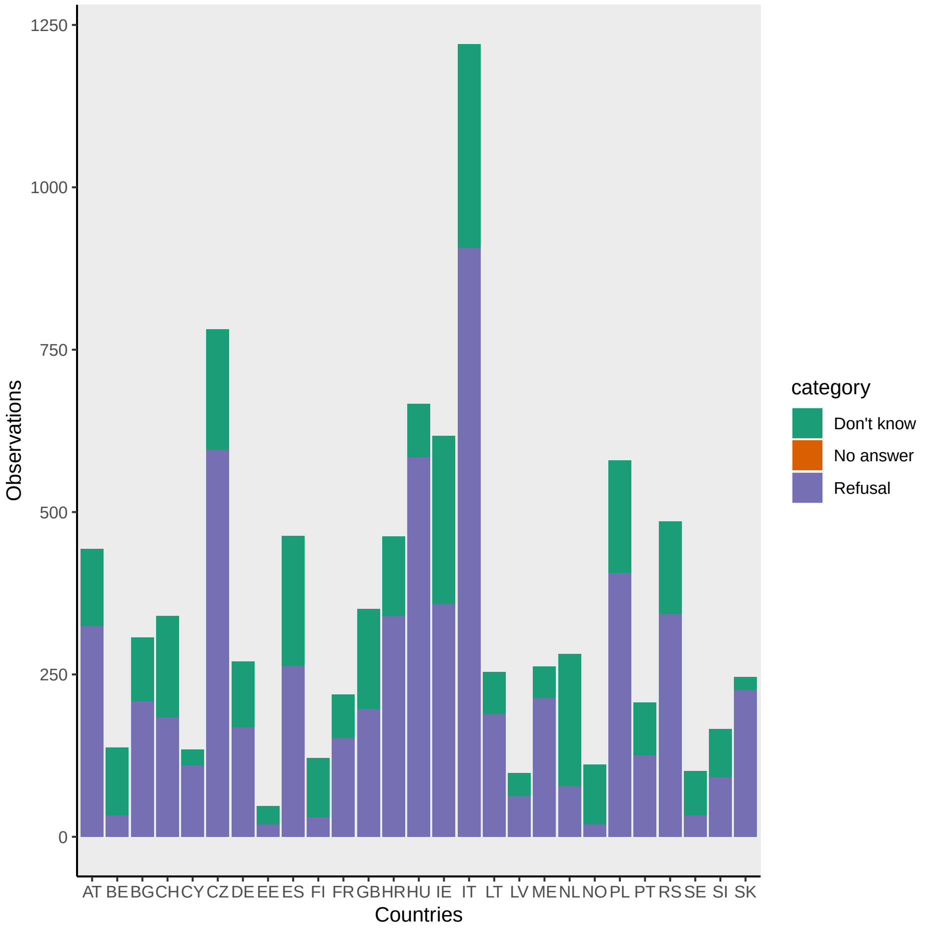


Figure 1: Missing data in the dependent variable (hinctnta) for all countries

2.1 The dependent variable

The dependent variable of our survey is *hinctnta*. It describes total net household income from all sources, measured in deciles. The decile classification of this survey is based on representative income surveys in each country, such as national register data or representative countrywide surveys. On the other hand, regarding the baseline data sources, please highlight the following aspects in particular:

- Deciles should refer to household income for all households, unadjusted for size or type
- Deciles should refer to household's total income, after tax and compulsory deductions, from all sources
- Deciles should reflect household income at the time of fieldwork, and thus be adjusted to 2018 level if necessary

In addition, Italy is our survey country, the deciles of *hincinta* for Italy are:

Number of deciles	**Income deciles**
01	Below 9000
02	9000-14000
03	14001-17500
04	17501-21000
05	21001-25000
06	25001-29500
07	29501-36000
08	36001-43500
09	43501-56000
10	More than 56000

Table 1: The deciles of *hincinta* for Italy in Euro per year

2.2 Sampling Method

Two domain sampling domains are used for the survey of Italy:

The first sampling domain included all the largest Italian cities with approximately 14% of the Italian population. The survey conduct a one-stage sampling design, that is, sampling each individual with a stratified simple random sample. The stratification is based on the information from the Italian Public Register of Individuals and the sample size was assigned proportionally to the target population in each stratum.

The second sampling domain included all cities that were not in the first domain. In this part, a two-stage sampling design is used, with stratification as the sampling method. In the first stage, cities were selected as Primary Sampling Units (PSUs) through stratified sampling. Stratification was performed by crossing two variables, namely geographical area and population size class of residents aged 15 years and older. Besides, the distribution of PSUs to the stratum is proportional to the target population within the stratum. Within strata, the probability of selection of PSUs is proportional to the size of the target population. In the second stage, the survey conducts a simple random sampling from each selected city.

2.3 Survey Weights

The survey provides three main weighting variables, which are *anweight* (analysis weight), *dweight* (design weight), and *pspwght* (post-stratified design weight) respectively.

The details of these three weighting variables are:

- *anweight*: *analysis* weight corrects for differential selection probabilities within each country as specified by sample design, for nonresponse, for noncoverage, and for sampling error related to the four post-stratification variables(gender, age, education and geographical region), and takes into account differences in population size across countries.
- *dweight*: the purpose of the design weights is to correct for unequal probabilities for selection due to the sampling design used.

- *pspwght*: the purpose of the post-stratified design weights is to reduce sampling error, noncoverage, and non-response bias, using auxiliary information. The post-stratification targets use information about age, gender, education and region. For most countries, population targets are weighted estimates from the European Union Labour Force Survey 2018 (Eurostat).

3 Literature review

Missing values embody a serious threat to statistical power by reducing sample size and distorting estimates derived from incomplete datasets. Plenty of datasets consists of missing values, where the cause of the missingness may not be fully observed. An example of such missing not at random (MNAR) data in many real-world applications are survey questions related to income. Usually, respondents' financial income represents the exact cause of that missingness.

Turning a blind eye to missingness mechanisms results in biased imputation and distorted statistical estimates. Besides MNAR data there are two other missing data mechanisms that rarely occur in real-world applications like surveys (Ma and Zhang, 2021; de Leeuw et al., 2018; Van Buuren, 2018). While missing completely at random (MCAR) data is independently missing regardless of observed und observed data and not causing any bias, missing at random (MAR) data is independently missing regardless of unobserved data.

For well over a hundred years social scientists, economists and policymakers strive for a coherent and complete definition of income – so far without success. Income itself is a "constructed idea, inherently driven by policy objectives and pragmatic concerns" (Brooks, 2018, 254). Income concepts regularly consider taxes, student financial grants, measurements of individual wellbeing, perceived job security and for example health care subsidies. With regards to the rather tax-driven concept of income, the following examples illustrate the challenge to defining income:

While homeowners benefit from untaxed returns from their houses in form of housing and avoid paying taxed rents to a third party, those providing child care and housework themselves also "generate imputed income" (Brooks, 2018, 254) by preventing to paying a child care provider (Kleinwächter, 1898). Nowadays dozens of different income definitions emphasize varying aspects of that concept. Piketty and Saez (2003) and Piketty et al. (2018) analyze not only administrative data but also survey and tax data to capture especially the top income shares, which highlights that the real challenge lies in measuring different levels of income. While tax data is on average more suitable to capture income from capital, thus constructing income measures for the wealthiest cohort, it does not capture employer-provided healthcare and government transfers, which constitute important income components for middle-income taxpayers.

As a consequence, the measurement of concepts like income heavily depends on a multitude of indicators. Similarly, psychologists are not able to measure their respondents' level of depression directly but rely on a collection of symptoms like the degree of self-confidence and energy. The extent to which relevant indicators of income are congruent with the concept itself embodies the degree of construct validity. When it comes to complex concepts like income, survey designers are often incentivized to properly explain survey questions to their respondents. The European Social Survey (ESS) comprises the variable "Household's total net income, all sources" which is associated with the following literal question:

> *"Using this card, please tell me which letter describes your household's total income, after tax*

and compulsory deductions, from all sources? If you don't know the exact figure, please give an estimate. Use the part of the card that you know best: weekly, monthly or annual income."

What becomes clear by observing the literal question underlying the household income variable is that the survey designers probably anticipate respondents' hesitation or inability to indicate averaged figures in case of irregular incomes. It is also notable that the survey question relates precisely to the net income after tax and compulsory deduction and specifies income from all sources, which further clarifies respondents' understanding of the whole question.

In the next step, we take a glance at another survey, which addresses income-related questions differently. The European Value Study (EVS) represents a cross-national and repeated cross-sectional survey that investigates the preferences and opinions of citizens in Europe. The main income-related question is close-ended:

"Here is a list of incomes and we would like to know in what group your household is, counting all wages, salaries, pensions and other incomes that come in. Just give the letter of the group your household falls into, after taxes and other deductions."

For Italian respondents, for instance, data for the income groups coding was obtained by the ESS 2016. While the first decile in group A entails the annual income range from 9000 Euro or minder and the fifth decile in group E is associated with incomes ranging from 20,501 to 24,000 Euro, the tenth decile in group 10 comprises households receiving 54,500 Euro or more. In comparison to the question wording of the ESS, respondents are not asked to indicate their income level manually, but by placing themselves based on their income in a pre-defined ranked income scheme. Considering that income-related questions are inherently sensitive and usually associated with non-response rates larger than average, such question wording might presumably hold appeal for respondents to place themselves in higher-income categories than it is the case.

In wave 2017, however, only 16,2 percent responded with "No answer" or "Don't know" (EVS, 2022). Also, none of the other countries is associated with larger non-response rates. Apparently, respondents seem to be more likely to indicate their income based on a pre-defined ranked scheme. With regards to the accuracy of their answers, the income distribution chart below does also not suggest any unusual patterns (EVS, 2021). Comparing the 44,44 percent Italian item non-respondents in ESS's open-ended questions with the 16,2 percent item non-respondents in the ESV's closed-ended question; the fact that open-ended questions, in general, tend to be more prone to non-response and particularly with regards to sensitive income-related topics seems to assert itself. (Reja et al., 2003).

In our discussion, we will later proceed to weigh the pros and cons of various survey design settings and their potential implications for measurement error. We round this review of different approaches on how to survey income off by providing an outlook to the centrepiece of this analysis: Should income as in the ESS (i.e. in deciles) be imputed as a categorical or continuous variable? Income variables that are coded in deciles are usually regarded as categorical, whereas only as

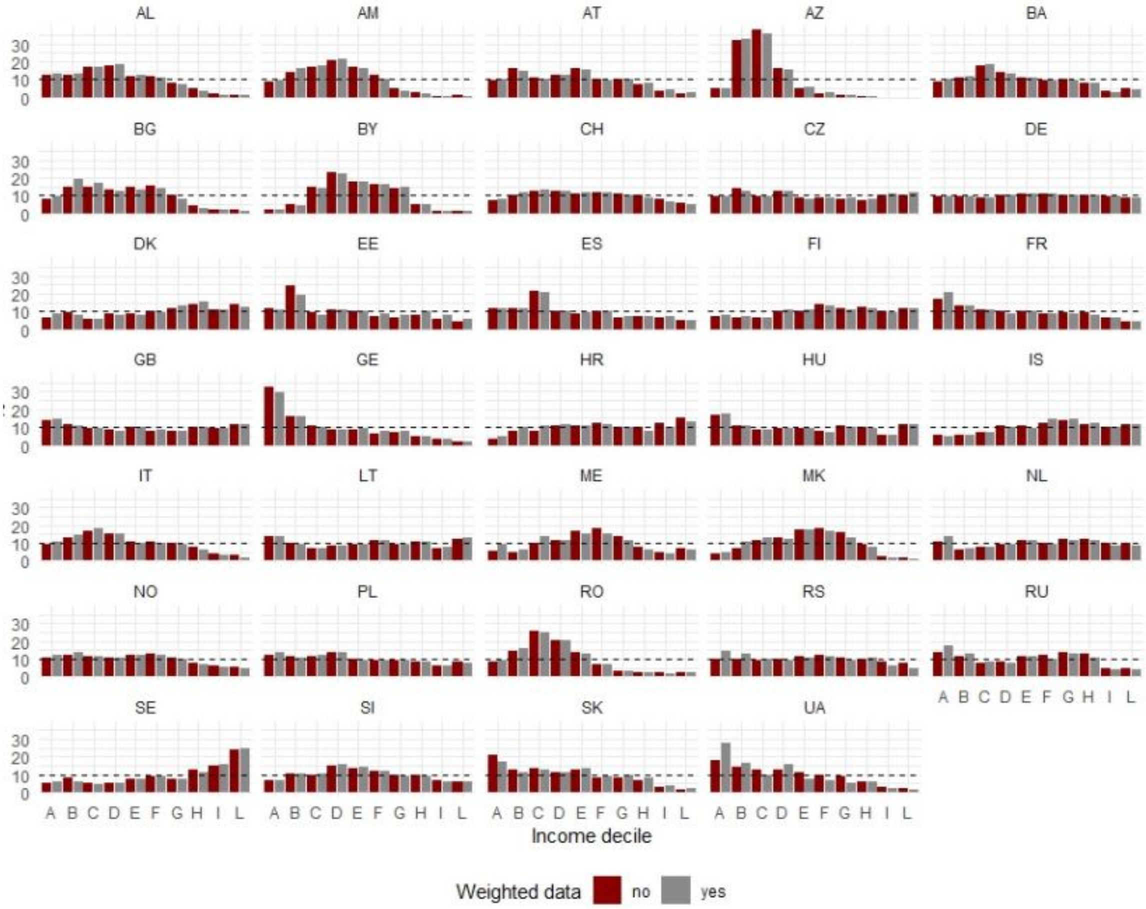


Figure 2: EVS, 2017: Income distribution based on weighted and unweighted data

continuous when they are originally scaled (i.e. in a currency) (Castillo, 2011).

4 Data analysis

The data analysis consists of three parts. The first one is looking at the non-response rate and trying to decide what the mechanism of missing data is. The second part consists of defining the survey design object and the third part discusses different imputation strategies.

4.1 Non-response rate and mechanism of missing data

According to Van Buuren (2018), each data point has a probability of being missing. The process that governs these likelihoods is called the missing data mechanism and it can be modelled using the missing data model. Different missing data mechanisms lead to three different types of missing data. (1) Missing completely at random (MCAR) implies that the probability of being missing is the same for all observations, i.e. the causes of the missing data is unrelated to the data. (2) Missing at random (MAR) means that the probability of being missing is the same only within groups defined by the observed data. As MAR is often more general and realistic than MCAR, modern missing data procedures are often based on the MAR assumption. (3) Missing not at random (MNAR) implies that the probability of being missing varies for unknown reasons. This is the most complex case which is addressed through finding more data about the causes of missingness, or through the performance of what-if analyses to assess how sensitive the results are under various scenarios.

As stated before, 44.44% of responses to *hinctnta* are forms of missing data. It is possible that people on the higher and lower ends of income would be more reluctant to disclose their earnings. That would mean that our pattern of missing data is MNAR. We argue that a fraction of the 44.44% could be distributed on the extremes of the assumed normal distribution of income, however, the majority would be more likely found near the mean of the distribution. According to Turrell (2000), two kinds of people are most likely to not disclose their income: those among higher socio-economic groups, and employed persons who receive their income from a business or a partnership for the fear of the income information being disclosed to tax authorities. In this view, people on the lower ends of the income distribution are just as likely to disclose their income as average earners. The higher-income category would have some missings, but then also earners across the spectrum of income worried about tax authorities learning about their income would be the most represented group of missings. We suspect that the latter is the main source of missing data. Since the missings are spread throughout the distribution, we can regard them as missing at random.

4.2 Survey design object

In our analysis, we choose *pspweight* as our weight in the weighting method. One of the other two options is *dweight*, which only corrected for differential selection probabilities, and thus not as accurate as *pspwght* and *anweight*. On the other hand, even though *pspweight* and *anweight* both corrected the error of non-response, non-coverage, and sampling error, *anweight* also corrects for the population size differences between counties. In our case, we merely extracted Italy as our aim country. Therefore, we should ignore the correction of different population sizes between countries and use *pspweight*.

The survey design before imputation can be specified in R using the variables: *psu*, *stratum* and *pspweight*. The code can be seen in Appendix 7.1.1. The survey design after imputing is a little more difficult. For every imputed dataset the survey design needs to be specified. Then the average values of the mean, SE and deff can be calculated of the different imputed datasets. The function to specify the survey design for the imputed dataset can be found in Appendix 7.1.2.

4.3 Imputation strategies

Before the data can be imputed all the missings need to be converted into NAs. This is because the ESS has different labels for different types of missings which can change depending on the variable. Usually, these labels are: *No answer*, *Don't know*, *Refusal*, and *Not applicable*. The aforementioned labels together constitute all the missings and can thus be changed to NAs in the dataset.

The next question is which variables do we want to use in the imputation. Van Buuren (2018) states that all the variables with >20% missing data should be removed as they are not considered good predictors for the variable we wish to impute, so they are removed. The correlation between a potential predictor and the variable of interest gives an indication of the quality of the predictor. Because of this, we created a correlation matrix between all the variables of the ESS data set and our DV *hinciant*, and we allowed all variables with a correlation of 0.25 or higher to be its predictors. Afterwards, a response indicator is created, which is a dummy variable whose value is 0 in the case of NA and 1 when there is a response. Good predictors will also correlate with the response indicator. We again made a correlation matrix with all variables of the ESS data set and allowed variables with a correlation of 0.125 or higher to be included as predictors. To ensure we do not take the same predictors twice the unique values from the two correlation matrices are used. The predictors that are used in our analysis can be seen in Table 2, where they are color-coded based on the similarity of their nature. While the ESS data set lists all variables as numeric, the documentation indicates that not all variables adhere to this type, which is why we assign the right level of measurement to predictors.

Name	Description	Correlating variable
edulvlb	Highest level of education	hinctnta
eisced	Highest level of education - ES-ISCED	hinctnta
eduyrs	Years of full-time education completed	hinctnta
ifredu	Fair chance achieve level of education I seek	hinctnta
ifrjob	Fair chance get job I seek	hinctnta
edctn	Doing last 7 days: education	response indicator
mnactic	Main activities last 7 days, all respondents	hinctnta
uemp3m	Ever unemployed /seeking work for >3 months	hinctnta
lvpntyr	Paid employment or apprenticeship at least 3 months 20 hours weekly	response indicator
hincsrca	Main source of household income	hinctnta
hincfel	Feeling about household's income nowadays	hinctnta
edulvlfb	Father's highest level of education	hinctnta
eiscedf	Father's highest level of education - ES-ISCED	hinctnta
edlvfeit	Father's highest level of education,ITALY	hinctnta
occf14b	Father's occupation when respondent 14	hinctnta
evpdemp	Year first left parents for living separately for 2 months or more	response indicator
bthcld	Ever given birth to/ fathered a child	response indicator

Table 2: Description of the predictors for Hincinta

Now that the predictors are chosen, we can specify the imputation design with the mice function.

The imputation will be done twice: once imputing *hincinta* as a categorical variable, and once as a continuous variable. The categorical imputation uses predictive mean matching (ppm) and is most commonly used for imputations, the R-code can be seen in Appendix 7.2.1.

The continuous imputation requires a bit more attention. First, *hinctnta* needs to be assigned a numerical level of measurement. Then the mice function can be used with the new dataset using a Bayesian linear regression method (Norn). The values will not be in the range of the original data, so we need to squeeze the values in the range of 1-10 (R-code is shown in 7.2.2). If we would let the variables outside of the range then we obtain values for *hicinta* that make no logical sense. The first decile consists of income from zero up until 9000 euro's, a negative decile could not possibly contain valuable information as income cannot be negative. The same goes for deciles over ten, as the tenth decile consists of incomes of 56000 euros and above. Logically, it is impossible to be above a category that does not have an upper limit. Including these out-of-range predictions could severely bias the mean. By trimming the values, we are able to have interpretable and logical results.

Another decision we had to make was how many imputations and iterations would suffice. We decided to settle on 10 for both. Increasing the number of imputations from the default 5 to 10 will increase standard errors thus more accurately reflecting the uncertainty of our estimates. Increasing the number of imputations stabilizes the results so that the order in which variables were imputed becomes irrelevant (Raghunathan et al., 2002).

5 Results

The results are split into two parts, the first consists of assessing the imputation models and in the second part, the results are discussed of the weights and imputation on *hinicnta*.

5.1 Imputations

As we can see in Figure 3 and Figure 4 the trace lines for *hinctnta*. The plot shows the mean (left) and standard deviation (right) of the imputed values only. Van Buuren (2018) states that we would like the streams to intermingle and be free of any trends at the later iterations, which is the case in Figure 3 and Figure 4.

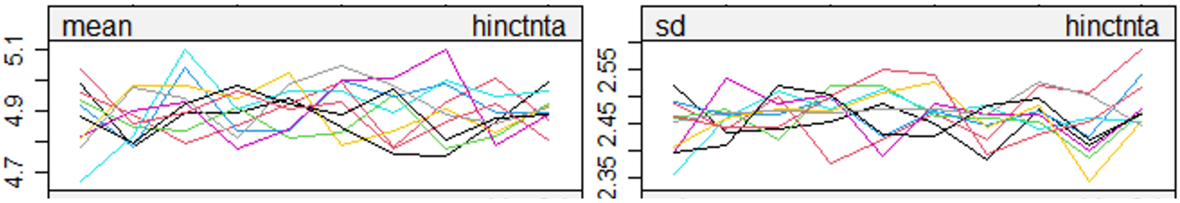


Figure 3: Spaghetti plot categorical imputation

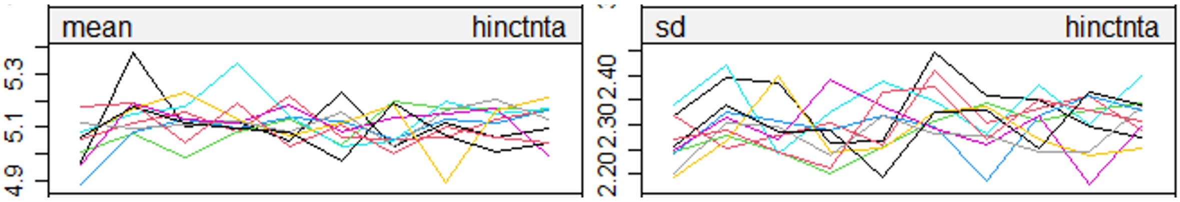


Figure 4: Spaghetti plot continuous imputation

Next, we can take a look at the density plots. Figure 5 show that the density lines of the categorical imputation follow the real data (blue line) closely.

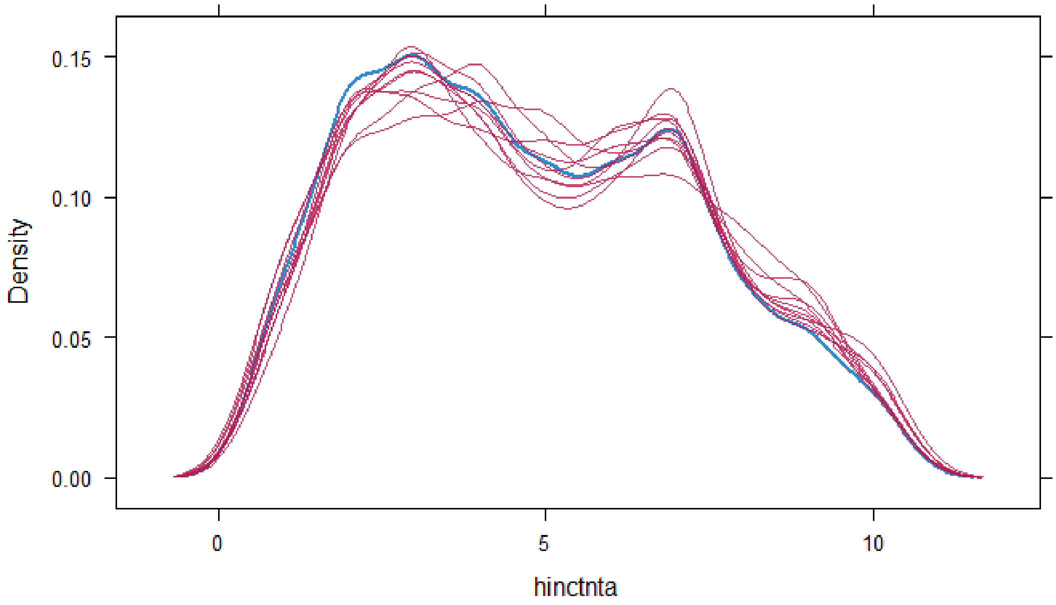


Figure 5: Density plot categorical imputation

In Figure 6 we see that the density lines of the continuous imputation follow the the real data (blue line) less closely between the values of *hinctnta* 2-6. We argue that this is the case because the underlying variable *hinctnta* is a categorical variable due to deciles being categorical. Therefore the densities of imputed values more closely resemble the original data in the categorical imputation than in the continuous one.

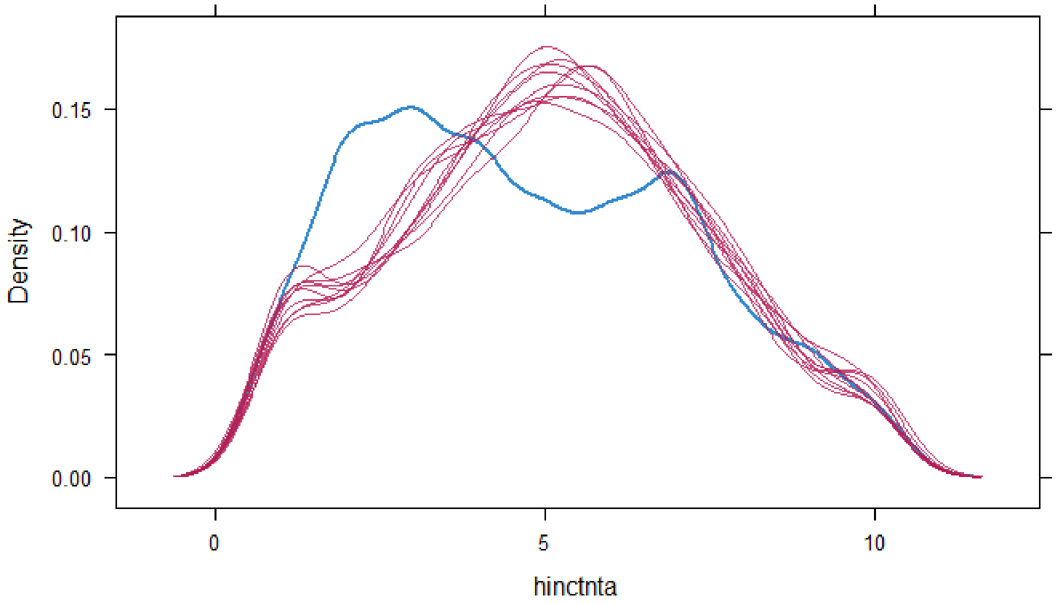
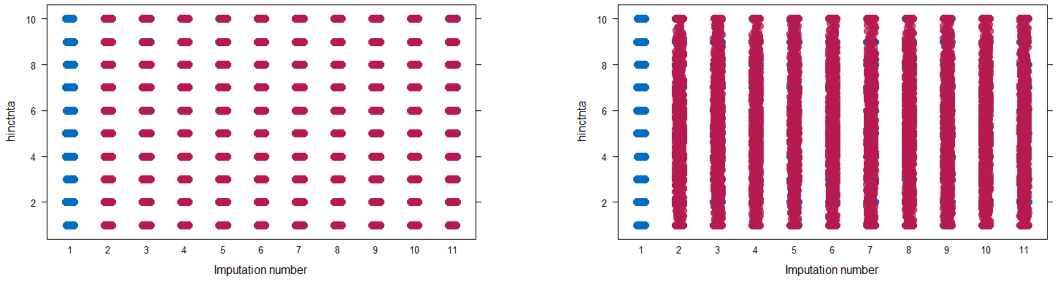


Figure 6: Density plot continuous imputation

Lastly, we can look at the strip plots. Figure 7a shows that the categorical imputation was successful as the imputed values were integers between 1-10. Figure 7a shows that the continuous variable was also successful. The bars are almost filled compared to the real data (imputation number 1), which indicates that the imputed data can take on any value between 1-10.



(a) Categorical imputation

(b) Continuous imputation

Figure 7: Strip plot for the different imputations

5.2 Hinicnta

Finally, we can look at some results of the imputation itself. A good way to compare the observed data with the categorical imputation is by looking at a histogram. Figure 8 reveals that the differences between the two groups are very small, but for both cases the there is not 10% of the data in each deciles (red line). It seems that there are a bit more people from the lower deciles in the original data, indicating that these people are underrepresented.

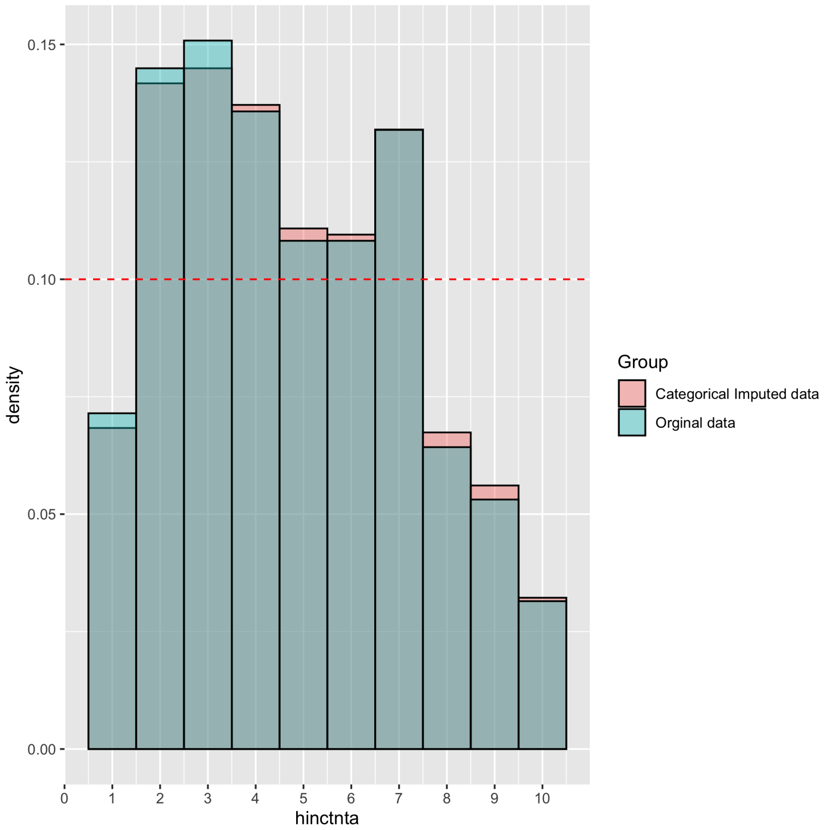


Figure 8: Boxplot hinicnta original data and categorical imputation

It is a bit harder to compare the continuous imputation since it is not a categorical variable anymore. Figure 9 shows the density plot and the general form is similar to that of Figure 8. *Hinicnta* is above the red line from 2 until 7, which is also the case for the observed data and the categorical imputation.

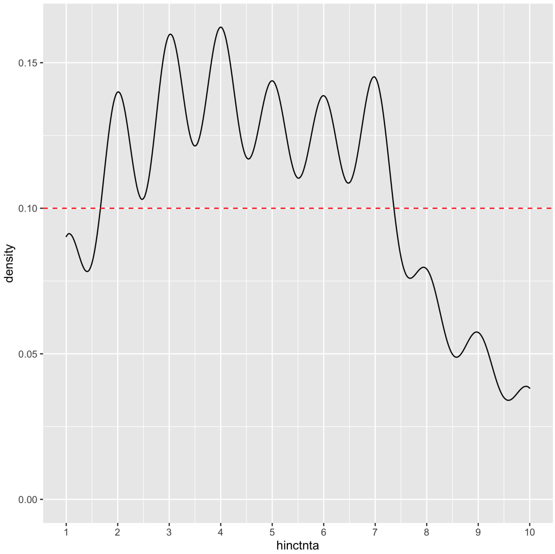


Figure 9: Density plot *hinicnta* with continuous imputation

This can be followed up by comparing the categorical and continuous imputation in a plot, as can be seen from Figure 10. The differences are quite large compared to Figure 8. This means that the distribution of continuous imputation is also quite than from the original data. The continuous imputations seem to have more observatories in the middle deciles (5/6), whereas the categorical imputation has more observations in the first deciles (2/3). Although the continuous imputation does have more observations in the first decile.

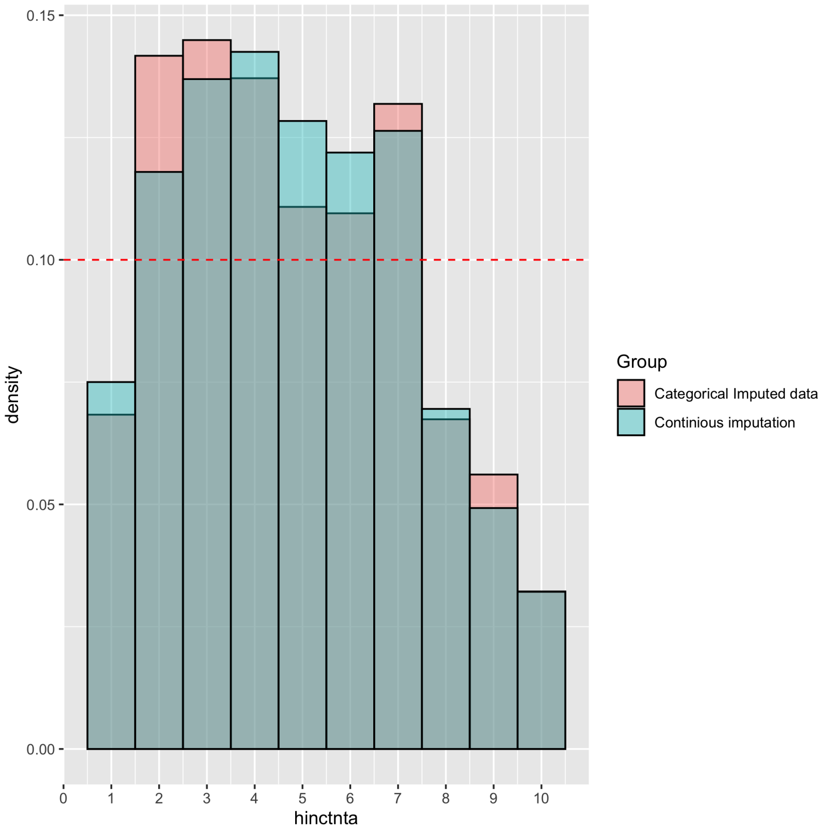


Figure 10: Density plot *hinicnta* categorical and continuous imputation

Lastly, we can take a look at the mean and SE of *hinicnta* for the different scenarios of using weights and imputing the data. In Table 3 we see that the mean of *hinicnta* slightly decreases when using the weights and slightly increases when using both forms of imputation. The combination of weights and imputation decreases the mean a bit, so the effect of the weights are decreasing. There are no major differences in the mean between the different imputation methods, continuous imputations yield a moderate higher mean than categorical imputation. The mean of imputation and weights are almost identical to the mean of the observed data, this indicates that the decreasing effect of the weights is offset by the increasing effect of the imputation.

For the SE we can see that the weights have an increasing effect compared to the observed data. There is no real difference between SE of the two imputation models and the observed data. The combination of weights and imputation yields again similar results for the two imputations models, which is higher than the observed data but lower than the observed data with weights.

Type of data	**Mean**	**SE**
Observed data	4.776	0.0467
Observed data with weights	4.693	0.0966
Observed and Imputed data (categorical)	4.834	0.0471
Observed and Imputed data (continuous)	4.908	0.0459
Observed and Imputed data (categorical) with weights	4.762	0.0706
Observed and Imputed data (continuous) with weights	4.835	0.0748

Table 3: Mean and SE of hinicnta for different analyses

To conclude, we can say that the imputations have been a success as was explained in 5.1. While the means and SEs of the continuous and categorical imputation yielded similar results, we have shown that the imputed values follow different densities in the previous section, with our preference given to categorical imputation. It seems like the weights compensate for too few people from lower deciles and the imputations compensate for too few people from higher deciles.

6 Discussion

The variable "Household's total net income, all sources" consists of 10 categories which are based on deciles of the actual household income range in each country. While using the median income as a reference point, the categories are calculated with that median at the top of the fifth decile. In total four levels of measurement (nominal, ordinal, interval and ratio scale) can be specified. The level of measurement for the income variable clearly is ordinal, because the decile specification offers labels, order and concrete intervals between each of its categories. Unlike a ratio scale, the variable does not indicate information about the value of the true zero and it is a truly numerical scale, for which the ordering and the difference between the categories are known.

Categorical variables possess discrete categories and levels, which are either nominal or ordinal. Nominal variables are not characterized with a specific ordering (e.g. gender, ethnicity) and ordinal variables have two or more categories that can be ranked (e.g. Likert scale). As soon as an ordinal scaled variable has more than five categories, some researchers treat them as continuous. Continuous variables on the contrary are measured mathematically along a continuum, which is either ratio or interval. Ratio variables possess a true zero point (e.g. distance) and interval variables are measured on a continuum between two fixed values and do not have a meaningful zero point (e.g. temperature in Celsius or Fahrenheit). The fact that the income variable's level of measurement (i.e. deciles) is ordinally scaled suggests that it represents a categorical variable.

In a nutshell, we confidently draw the conclusion that based on our findings of the analysis we clearly prefer categorical imputation. To begin with, the density plot of categorical imputation better represents our original data than the continuous imputation. What is more, continuous imputation yields values outside the 0-10 range. These values have no additional meaning and have to be squeezed in, which is not ideal.

We consider the inclusion of visualizations and especially the box plots as beneficial, for instance, to illustrate that categorical imputation has small differences with the deciles of the original data whereas continuous imputation has large differences for the middle deciles. We complete this result section by concluding that our preferred method definitely is observed imputation with weights and what becomes evident in our analysis is that the mean of the imputed and raw data differs only by 0.3 percentage points.

6.1 Total Survey Error (TSE) Components

Bearing in mind that measures and definitions of income are rather dynamic and constructed based on the analysis objective raises several questions with regards to potential error sources.

Survey errors arise from coding, the sampling process and missing data for example, which threatens the accuracy of statistical inference. Such estimates are only reliable as long as the bias is kept small and the impact of Total Survey Error (TSE) is well understood and under control. In the context of our analysis, we focus on non-response as a non-sampling error. Such rather unspecific non-response error occurs either in the shape of the unit or in item non-response. The former entails when for example the whole sampled household does not comply with responding to any item of the survey while the latter only refers to units answering only partially to the survey. Questions related to income are a good example of item-non-response (Biemer, 2010).

There are various reasons why item non-respondents choose to not comply with answering since income-related questions are associated with a relatively high degree of perceived "intrusiveness" into one privacy (Yan and Jans, 2006, 146). Beatty and Herrmann (2002) differentiate respondents' knowledge of a topic, how accurate they perceive that knowledge to be and their willingness to share that knowledge in such a survey situation. While the first two sources of potential risks to response compliance are cognitive or informational, the latter is rather motivational. Applying this to the income-related survey questions predicts that respondents are either not or feel not knowledgeable enough or lack the motivation to indicate their personal income level or even of the whole household.

Underlying factors predicting income non-response have been investigated by Juster and Smith (1997) and Moore et al. (1999) and their models converge to the same conclusion: Income non-response is due to cognitive and motivational factors. Subsequently, we take a closer look at the latter in order to detect proposed solutions. Varying techniques to minimize income non-response are available: Overall, telephone interviews (i.e. Computer-Assisted Personal Interviews (CAPI) are related to higher degrees of item-non-response with regard to income than for example PAPI (Paper- and pencil Personal Interviews), which holds, however, only for at most the first two waves of data collection (Yan and Jans, 2006, 147).

What strikes us, even more, is the fact that interviewer-led modes result in a lower degree of non-response than mail surveys, where respondents are fully self-sufficient regarding their response behavior (De Leeuw, 2001). While the ESS requires all countries to conduct face-to-face data collection (i.e. potentially beneficial to prevent income non-response) the EVS does so too, except that Switzerland, Netherlands, Iceland, Germany, Denmark and Finland have chosen the path to also provide self-administered Web survey.

Previously, we already emphasized the difference between open- and closed-ended questions in our literature review. Most notably, the approach of the EVS with closed-ended questions is more commonly used and more efficient to diminish income non-response, whereas not the open-ended question in the ESS. As soon as item non-respondents possess motivation (i.e. respondents indicating “Don’t know”) but lack cognitive factors to produce valid responses, the “unfolding bracket” technique opens the doors way to further minimize non-response. This specific response format entails a series of follow-up questions, which provides guidance to respondents and supports the researcher to narrow down their income level (Heeringa et al., 1993). This extremely successful technique (i.e. not only the income non-response decreases by up to 50 percent but also the bias) is applied in various surveys.

What we are doing in our analysis, is to overcome income non-response after data collection by means of imputation. This is why we attempt to understand the underlying mechanisms responsible for non-response by finding demographic correlates. In addition to what we find, respondents’ response compliance is more generally affected by their age, sex, education and race. Older, white respondents with relatively high incomes are most likely to produce income non-response (Riphahn and Serfling, 2005).

Aiming to paint a macro picture of the relation between income non-response and survey features, we emphasize that times-series analysis indicates a steady decline of non-response to open-ended income questions and a diminishing beneficial effect of the “unfolding bracket” technique in the last decades, making them fall within the same non-response rate ranges declared by surveys consisting of closed-ended income questions.

Plus, the overall trend of income non-response and respondents’ cooperativeness (i.e. the extent to which they agree to participate in the survey) are closely interconnected. Before investigating how to increase respondents motivation to reply to income-related questions, we consider possible drawbacks associated with re-designing survey features to reduce non-response. After differentiating survey design features that potentially diminish non-response, we proceed to pay attention to the other side of the coin by looking at the costs of additional efforts. Despite those noble intentions, Kreuter et al. (2010) underline that additional efforts do indeed minimize non-response, but may elevate measurement error.

Altogether the total bias (i.e. mean square error (MSE)) still remains lower than with no additional efforts to decrease non-response. As soon as the notion that respondents with higher incomes are less likely to respond and the fact that those respondents tend to overstate their actual income (i.e. low accuracy) fall together, additional non-response adjustments are confronted with major challenges. Reduction of non-response bias is only beneficial for the overall MSE if the measurement error is also effectively tackled. Kaminska et al. (2010, 974) expand the outlook on respondents’ lack of motivation to comply with item response by stressing the extent to which respondents are actually willing to participate in the survey:

> *“These results indicate that reluctant respondents are more likely to provide responses of low quality than cooperative respondents, but this association is present because of differences in cognitive ability rather than differences in motivation.“*

These findings throw a critical light on the would-be distinction between cognitive and motivational factors, suggesting that they rather intermingle. The fact that the effect of respondents' cognitive ability heavily impacts the motivational perspective on the relationship between reluctance and satisficing demands thorough awareness for additional adjustments in answer to non-response (Kaminska et al., 2010, 976).

Peytchev et al. (2010) investigate surveys about abortion experiences, thus an even more sensitive topic. What they find is, that understanding the underlying causes of non-response (i.e. social stigma) as well as incentives for respondents to comply with responding are vital for breaking the link between those two conflicting error sources: non-response and measurement error. By offering a 40 US-Dollar incentive to all respondents and an additional 45 US-Dollar for more reluctant respondents, they only find mixed support that incentives worked out in that specific scenario. The effect of more than doubling incentives in order to gain cooperation from more reluctant respondents on their likelihood to comply turns out to be not significant (p=0.639) and under-reporting of abortion experiences still occurred, probably also related to the delicateness of the topic in question Peytchev et al. (2010).

Providing respondents with (unconditional) incentives finds application in various other surveys, also with much lower cash incentives like 5 or 10 Euro. With regard to face-to-face interviews these unconditional incentives lead in a two-stage survey experiment to significantly higher response rates (+8.9 percent; t=5.14). Moreover, self-administered Web survey respondents are also more likely to finalize their registration after obtaining an unconditional cash incentive (30 percent) than those without obtaining any kind of incentive (13.7 percent). Comparing the amount of cash incentive (i.e. 5 or 10 Euro) shows that respondents receiving 10 Euro are associated with a 3 percentage points higher likelihood to finalize registration than those attributed less stimulus. Unconditional cash incentives are associated with positive effects during the recruitment of respondents and with regard to their response behavior (Blom and Krieger, 2015).

Presumably, the combination of incentives, reminder e-mails or phone calls to remind respondents to participate in case of Web survey situations, an overall personalized interaction between survey organisers and respondents and providing respondents with insight into the research endeavours and results (i.e. providing a homepage with simplified research results) as well as transparency regarding the research team boosts respondents' confidence in the overall project (Blom and Krieger, 2015).

(Non)response rates, in turn, embody at no time a sufficient indicator of the overall data quality – and a fact that is pivotal to us - completeness and accuracy of responses to income-related survey questions are indeed two totally different pairs of shoes, but both pairs must fit a comfortable degree (Groves, 2006).

References

Beatty, P. and Herrmann, D. (2002). To answer or not to answer: Decision processes related to survey item nonresponse. *In Survey Nonresponse, R. Groves, D.A. Dillman, J. Eltinge, and R. Little (eds). New York: Wiley.*

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public opinion quarterly*, 74(5):817–848.

Blom, A. G. andGathmann, C. and Krieger, U. (2015). Setting up an online panel representative of the general population: The german internet panel. *Field methods*, 27(4):391–408.

Brooks, J. R. (2018). International nonresponse trends across countries and years: an analysis of 36 years of labour force survey data. *Tax L. Rev.*, 71:253.

Castillo, J. C. (2011). The legitimacy of economic inequality: An empirical approach to the case of chile. *Universal-Publishers.*

De Leeuw, E. (2001). Reducing missing data in surveys: An overview of methods. *Quality and Quantity*, 35(1):147– 160.

de Leeuw, E., Hox, J., and Luiten, A. (2018). International nonresponse trends across countries and years: an analysis of 36 years of labour force survey data. *Survey Methods: Insights from the Field*, pages 1–11.

EVS (2021). Evs 2017 variable report: Integrated datasets (za7500, za7502); appendix b: Income. gesis-variable reports 2020/16. *GESIS Data Archive*, 2020(16).

EVS (2022). Zacat – gesis online study catalogue. *GESIS Data Archive.*

Groves, R. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70:646 – 675.

Heeringa, S., Hill, D., and Howell, D. (1993). Unfolding brackets for reducing item nonresponse in economic surveys. *AHEAD/HRS Report No. 94-029. Ann Arbor: Institute for Social Research.*

Jowell, R., Roberts, C., Fitzgerald, R., and Eva, G. (2006). *Measuring attitudes cross-nationally: Lessons from the European Social Survey.* Sage.

Juster, T. and Smith, J. (1997). Improving the quality of economic data: Lessons from the hrs and ahead. *Journal of the American Statistical Association*, 92(1):1268 –1278.

Kaminska, O., McCutcheon, A. L., and Billiet, J. (2010). Satisficing among reluctant respondents in a cross-national context. *Public Opinion Quarterly*, 74(5):956–984.

Kleinwächter, F. L. (1898). Das einkommen und seine verteilung (income and its distribution). *unpublished.*

Kreuter, F., Müller, G., and Trappmann, M. (2010). Nonresponse and measurement error in employment research: making use of administrative data. *Public Opinion Quarterly*, 74(5):880–906.

Ma, C. and Zhang, C. (2021). Identifiable generative models for missing not at random data imputation. *Advances in Neural Information Processing Systems*, page 34.

Moore, J., Stinson, L., and Welniak, E. (1999). Income reporting in surveys: Cognitive issues and measurement error. *Cognition and Survey Research.*

Peytchev, A., Peytcheva, E., and Groves, R. M. (2010). Measurement error, unit nonresponse, and self-reports of abortion experiences. *Public Opinion Quarterly*, 74(2):319–327.

Piketty, T. and Saez, E. (2003). Income inequality in the united states, 1913–1998. *The Quarterly journal of economics*, 118(1):1–41.

Piketty, T., Saez, E., and Zucman, G. (2018). Distributional national accounts: methods and estimates for the united states. *The Quarterly Journal of Economics*, 133(2):553–609.

Raghunathan, T. E., Solenberger, P. W., and Van Hoewyk, J. (2002). Iveware: Imputation and variance estimation software. *Ann Arbor, MI: Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan.*

Reja, U., Manfreda, K. L., Hlebec, V., and Vehovar, V. (2003). Open-ended vs. close-ended questions in web questionnaires. *Developments in applied statistics*, 19(1):159 –117.

Riphahn, R. and Serfling, O. (2005). Item nonresponse on income and wealth questions. *Empirical Economics*, 30:521– 538.

Turrell, G. (2000). Income non-reporting: implications for health inequalities research. *Journal of epidemiology & community health*, 54(3):207–214.

Van Buuren, S. (2018). *Flexible imputation of missing data.* CRC press.

Yan, T., C. R. and Jans, M. (2006). Trends in income nonresponse over two decades. *Journal of Official Statistics*, 26(1):145.

7 Appendix

7.1 Survey Design object

7.1.1 Survey Design object before imputation

```
design <- svydesign(ids = ~psu, strata = ~stratum,
weights=~data.italy$pspwght,data = predictors)
```

7.1.2 Survey Design object after imputation

```
#imput should be in the form: complete(dataset, "long)
survey_mice <- function(a) {
  for(i in 1:length(a)){
    means <- c()
    se <- c()
    deff <- c()
    b <- a[[i]]
    surveydesign <-svydesign(ids=~psu, strata = ~stratum,
    weight=~data.italy$pspwght, data=b)
    c<- svymean(~hinctnta, design=surveydesign,
    deff = T, na.rm = T)
    d<- as.data.frame(c)

    means[i]<- d[1]
    se[i]  <- d[2]
    deff[i] <- d[3]
  }
  y <- as.numeric(means)
  Average_means = mean(y) #mean of means
  z <- as.numeric(se)
  vw <-  sum(z^2)*(1/length(a))
  vb <- var(y)
  SE_pooled <- sqrt(vw+vb+vb/length(a))
  def <- as.numeric(deff)
  Average_deff <- mean(def)
  statistics <- cbind(Average_means,SE_pooled,Average_deff)
  return(statistics)
}
```

7.2 Imputation model

7.2.1 For categorical imputation

```
imp <- mice(predictors, m = 10, maxit = 10, mehtod = "ppm",
predictorMatrix =  pred, print = F, seed = 123)
```

7.2.2 For continuous imputation

```
ini <- mice(predictors, maxit = 0)
post <- ini$post
post["hinctnta"] <- "imp[[j]][,␣i]␣<-␣squeeze(imp[[j]][,␣i],␣c(1,␣10))"
imp_continious <- mice(predictors2, m = 10, maxit = 10, method = "norm",
predictorMatrix =  pred, post=post, print = F, seed = 123)
```