POZNAŃ UNIVERSITY
OF ECONOMICS
AND BUSINESS

Institute of Informatics and
Quantitative Economy

Kyrylo Mordan

Porównanie wybranych metod redukcji obciążenia spowodowanego brakami odpowiedzi w badaniach zmiennych jakościowych

Comparison of selected methods to reduce bias due to non-ignorable non-response for categorical variables

Bachelor's thesis

Thesis Supervisor: dr Maciej Beręsewicz
Date of submission:

Supervisor's signature

Field of study: Informatics and Econometrics
Specialisation: Economic Analytics

Poznań 2020

Contents

Introduction

Declining response rates is a growing concern for any research, which data comes from the sources like surveys and censuses. Non-response creates precedent for all sorts of errors related to the missingness of the data. These errors cause estimates, made with such data, to be biased. Different methods of dealing with the problem of non-ignorable non-response like preventive and reductive techniques involve actions during data collection phase, while corrective techniques, with a use of statistical adjustment, face the problem afterwards.

The aim of the thesis is a comparison of the two different corrective, model-based approaches. One of the methods is based on a parametric model from (Lee i Marsh 2000), while the other is based on non-parametric model from (Zhang 2001). The point is to compare the ability of the select methods in minimizing bias of the estimates that rely on a categorical variable subject to non-response.

In order to make the comparison, these methods were implemented from scratch in R. The data was generated with a use of the distinct parametric models, in order to test the performance in different conditions. Mean absolute bias in estimated proportion of the categorical variable subject to non-response, as well as, mean absolute bias in estimated parameters were the metrics used for the comparison of the methods.

First chapter of the thesis introduces some basic concepts related to the non-ignorable non-response. There, data sources in statistics are described and possible errors that can potentially take place, are presented. Different response mechanisms and the negative effect of the bias on the quality of the estimates, also appear in the chapter. In the last sections of the chapter, possible approaches to deal with non-ignorable non-response are classified. Second chapter of the thesis describes select corrective, model-based methods, in depth. The third chapter starts with details on the data simulation procedure. The empirical comparison between the methods appears later in the chapter. Each method and its variants is first compared to the estimates based on only the observed data, and only then the best variants are going to be compared to each other. At the end of the thesis is the conclusion.

Chapter 1

Non-ignorable non-response in sample surveys

1.1 Data sources in statistics

1.1.1 Traditional classification of data sources

Data can come in different shapes and forms. In general, a distinction could be made between primary and secondary data. Primary data is data that has been collected for the purpose of the research, while secondary, on the contrary, for some other purposes. The value of the primary data is greater since a researcher has more control over the data collection process or at least does not need to worry about any statistical treatment that the secondary data might have undergone (Sobczyk 2007).

Data for statistical purposes can come from different sources. The traditional classification of the data sources would be differentiating between statistical (primary) and non-statistical (secondary) sources. Census and sample surveys are examples of statistical data sources. Non-statistical sources, on the other hand, are the sources created for purposes other than statistics. These sources can be a result of data being collected by governments or private organizations. Administrative records are an example of a non-statistical data source that was collected for the government. Registers are used to construct sampling frames and as a source of auxiliary variables for a model-based approach (Beręsewicz 2016).

1.1.2 Modern classification of data sources

In today's world, administrative records are not the only alternative data source a researcher can hope for. Widespread adoption of the Internet has presented opportunities for more and more new data being collected in the private sector. This new data is associated with the term "Big data" (BD).

BD is a non-statistical term that describes data that is very large in size. BD as a secondary data source presents a researcher with some challenges. It tends to be messy and needs to be "cleaned up" before the data can be used. BD encompasses all kinds of data, like social media posts, traffic camera feeds, locations of the smartphone users, search terms used on the Web, and so on.

In this "brave new world", data is being collected "everywhere" and by "everything", that's why it is important to classify these data in accordance with the source. There are multiple ways to make such a classification, but in this thesis, the choice has fallen on the one proposed by Citro (2014). This classification includes both traditional data sources and the new ones in four distinct categories.

1. **Surveys and censuses** are certain statistical data collection methods that ensure generalizability for a defined population. These methods are designed according to principals of survey research by the data collectors (government or commercial survey organizations) to query sampled individuals on one or more select topics.
2. **Administrative records** are the data source that was designed by the administrative body according to law, regulation, or policy for the purposes of the institution, not statics. Administrative records can be also run by non-government organizations and are usually ongoing.
3. **Commercial transaction records** are the data initiated by the buyer about the goods that were sold, their prices, bar-codes, etc, which were collected with electronic capture of the purchase.
4. **Interactions of individuals with the World Wide Web** include all sorts of data collected on the Internet with the use of tools, like browsers and social media sites.

1.2 Classification of errors in statistics

1.2.1 Single-source approach

Quality of the survey outcomes is greatly impacted by the choice of a sampling frame, a sampling design, and an estimation procedure. But apart from that, there are things outside of the researcher's control impacting the quality of survey outcomes in a negative way. Quality of the estimates can be measured by the distance of the values that have been estimated from the true values, that have been observed in the population. This distance is called the *total error* of the estimate. Since estimates will always differ from the population characteristics, there will always be some error in the estimates. Errors can have different origins, and therefore can be classified into different categories (Biffignandi i Bethlehem 2011).

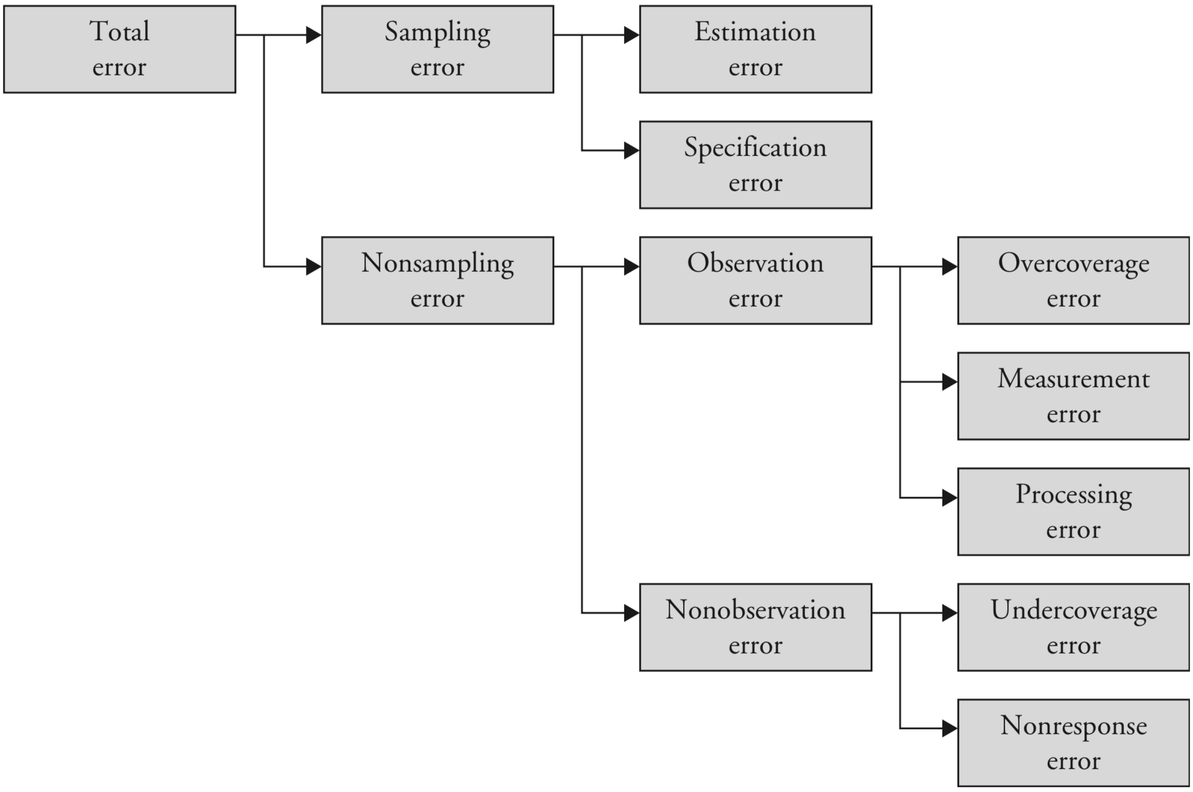

Figure 1.1. A classification of survey errors

Source : Based on Biffignandi and Bethlehem (2011)

Sampling errors are the errors caused by the sampling nature of the data, so when data from the whole of the population is present, there cannot be any sampling errors. Sampling errors include *an estimation error* and *a specification error*.

An *estimation error* caused by a deviation of estimated value from the true population value when each new sample is selected. Changes in sampling design, like increasing the sample size or using an auxiliary variable to adjust selection probability, can control for an estimation error.

A *specification error* is primarily an error caused by the differences in *true selection probabilities* and *anticipated selection probabilities*, due to *problems in the sampling frame*. Unbiased estimator, like Horvitz–Thompson, can be constructed if selection probabilities are known and correct, but when specification error is present, the estimator may be biased.

Nonsampling errors are different from the sampling errors in a way that measuring the whole of the population will not solve them. These errors are committed during the data collection procedure and can be divided into *observation errors* and *non- observation errors*.

Observation errors are errors made during data acquisition from respondents and their *further processing*. There are three types of observation errors: *an overcoverage error, a measurement error*, and *a processing error*.

An *overcoverage error* manifests itself if the target population does not include some elements that have been observed. The possibility of the error can be fixed by excluding these pieces of the data as if they were not there in the first place.

A *measurement error* is a problem caused by an inability to measure the true answers of a respondent. This error can be a result of inaccurate measurements due to *incorrect calibration of the instrument* used for collecting measurements, which in the case of survey questionnaires may be a clarity of the question or even the ability of a respondent to give true answers to those questions.

A *processing error* is an error that happens in the *phase of recording and processing*, somewhere between respondent that is giving an answer and researcher that receives it.

Nonobservation errors are the kind of errors that occur because of the inability to take indented measurements. *An undercoverage error* and *a non-response error* are both non-response errors.

An *undercoverage error* is an error caused by the sampling frame deficiency of some elements that are present in a target population, effectively giving them a zero probability of selection for the survey.

A *non-response error* is an error caused by the lack of response among some sampled units. If there is a disproportional response between different groups from the target population, some groups will appear larger based on estimates from the sample, therefore overrepresented in a survey while others due to their lower levels of response, will be underrepresented.

1.2.2 Multiple sources approach

As discussed earlier, a researcher has plenty of data sources to choose from for statistical research. Some of these sources like censuses or sample surveys could be feasible on their own for statistical purposes, so the researcher would not be forced to use additional data sources. If the researcher chooses to combine multiple data sources, for example, census data with administrative records, additional errors can emerge.

Throughout the "life cycle" of the data, from its collection to processing errors can occur and this is also true for integrated data sources. To illustrate the emergence of these errors, *a two-phase life-cycle model of integrated statistical micro data*, which is presented in Figure 1.2, was proposed by (Zhang 2012).

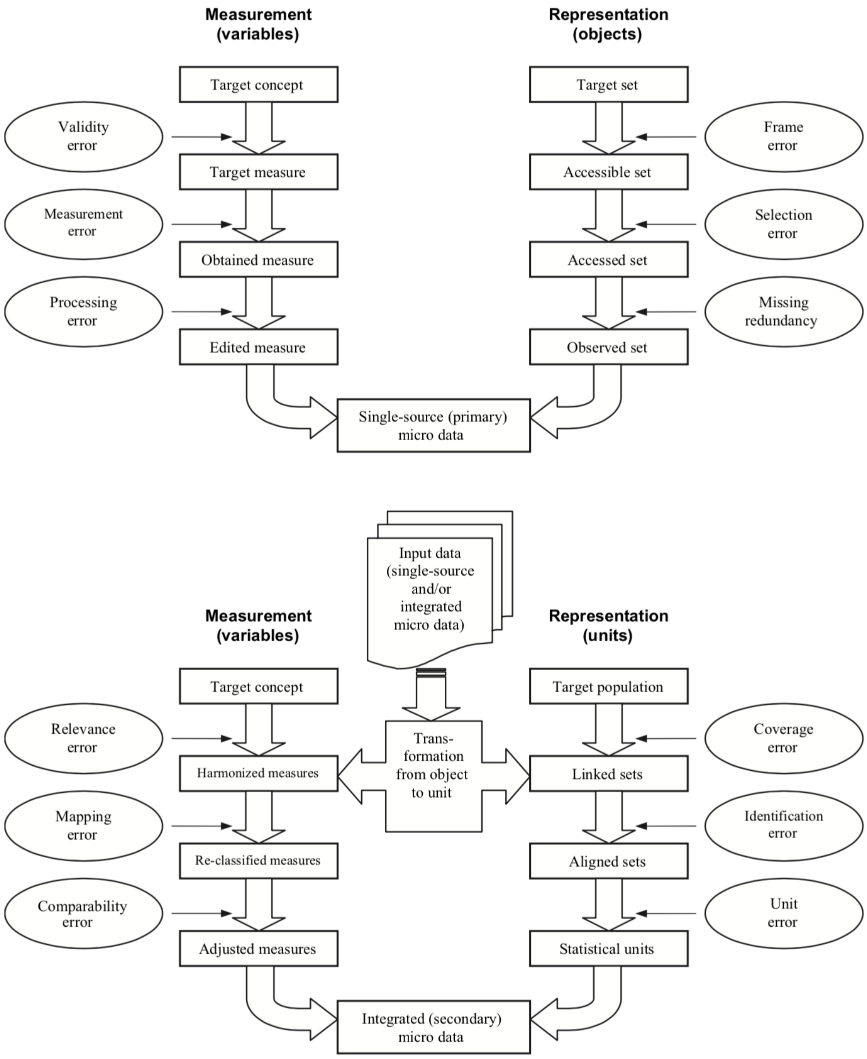

Figure 1.2. Two-phase (primary and secondary) life cycle of integrated statistical micro data from a quality perspective. Data concept (square); Error Type (oval). Source of error indicated by narrow arrow; Input, flow and/or processing of data indicated by broad arrow.

Source : Based on Zhang (2012).

The rectangles describe various states the data encounter along the lines of "measurement" and "representation" respectively. The broad arrows indicate the flow and/or processing between two successive states. The source of error is located between any two states, corresponding to an error type (the ovals).The first phase concerns the data from each single source, the second phase concerns the possible integration of data from different sources, often involving necessary transformations of the initial input data (Zhang 2012).

The illustration is applicable to different data sources, and for that reason, the term "object" contrasts "unit" in the "representation" lines of the phases. In phase two, the integration of the data sources requires a transformation from objects to units. This happens *because many* *administrative registers are initially based on events*, as objects in phase one do not necessarily represent people. Target statistical units need to be available in all the input data, that why sets should be first linked across relevant data sources. In phase two, various data sets need to be linked with some key, and that key is a unit, which represents a person. Although, additional linkages could be made with the use of other keys, *for example, using addresses to obtain housing information from the dwelling register*. When there is a difference between the target population and the statistical units in linked data, the *(under-/over-) coverage error* occurs. The *coverage error* may also be caused by the error in the linkage procedure.

After linking sets, *alignment* is needed to clarify relationships between different units in linked data, so that statistical units could be created. An error that happens during alignments is called an *identification error*.

Since some linked data was created for some other purposes that statistics, statistical units may not exist in the available data source and need to be created by the statistician. An error that may arise in this process is the *unit error*.

There can be a whole range of categories a measure can take, so the conceptual alignment of multiple measures is needed. This is called *harmonization* and it is used to arrive at a common standard. *The relevance error* is the extent of disagreement between the target concept and the harmonized measures.

When primary input-source measures are turned into harmonized measures, *re-classification measures* are obtained. If input categories do not have a well-defined mapping, *mapping error* occurs.

Between the *reclassification measures* and *adjusted measures* happens editing and imputation of the data. Editing in the second phase is different from the one in the first phase, because of *inconsistency across the data sources, which may not necessarily imply a quality problem with the input data source from the register owner's perspective*. This may be caused by the time delay between an event happening and it being recorded in different data sources. The inconsistency needs to be edited and is a source for *comparability error*.

1.3 The problem of missing data

1.3.1 Taxonomy of missing data

Missing data is a common and rather expected occurrence. Most of it is due to non-response which happens when complete measurements on the survey sample unit have not been successfully obtained. A distinction could be made between two types of non-response: Unit non-response and Item non-response.

Unit non-response occurs if there is no information provided by the selected unit. Failure to make contact with a sample unit or straight-out refusal of a sample unit to participate in the survey are just a few of the numerous examples of unit non-response. Unit non-response basically translates into a lack of measurements for the unit.

Item non-response, on the other hand, happens when some of the measurements have been obtained, but not all of them. Refusals caused by the sensitive nature of some questions or even loss of data could be the root of the item non-response (Biffignandi i Bethlehem 2011).

Crucial in understanding non-response are, so-called, *Response mechanisms* that cause non-response (Brick 2013):

- Missing Completely at Random (MCAR) – the probability of non-response is independent of the target variable with missing data and auxiliary variable for which data is complete. The response is not selective.
- Missing at Random (MAR) – the probability of non-response is independent of the target variable with missing data but not from auxiliary variables for which data is complete. The response will be selective, but this can be cured by applying a weighting technique using the auxiliary variable.
- Not Missing at Random (NMAR) – the probability of non-reposnse is dependant on the target variable with missing data and this relationship cannot be accounted for by an auxiliary variable. Correction techniques based on the use of auxiliary variables will be able to partially weaken this dependency. This type of missing data is also called non-ignorable non-response when referring to the non-response problem.

Missing data, less for MAR, and more so for NMAR response mechanisms, can lay the ground for sample bias. Sample bias can be defined as a possibility that sampled units differ systematically from those of the population to which estimates are to be generalized (Blair i Zinkhan 2006).

1.3.2 Impact of non-response on bias

After selecting a target variable, from those available in a sample data, collected values can be used to estimate population characteristics with an estimator. An estimator is a recipe to compute a selected characteristic of the population, based on sample data. Bias is one of the things that are undesirable for any good estimator.

For example, if $\overline{y}_E$ is an unbiased estimator of the population mean $\overline{Y}$, then average value over all possible outcomes must be equal to the population mean to be estimated:

$$E(\overline{y}_E) = \overline{Y}. \tag{1.1}$$

But if this is not the case, then the bias of the estimator $\overline{y}_E$ is:

$$B(\overline{y}_E) = E(\overline{y}_E) - \overline{Y}. \tag{1.2}$$

Errors that take place in survey data can result in sample bias. There are three general instances of sample bias:

1. Coverage bias – parts of the population are not considered in the sample or due to research method are not accessible. Internet surveys are an example of such a method that can exclude parts of the population that, for some reason, do not use the internet.
2. Selection bias – very high or low chances of selection occur for some groups of the population.
3. Non-response bias – uneven respond deficit across population groups.

The *random response model* can be used to illustrate the effect of non-response on estimators. Suppose there are N elements in the population and every element k has an (unknown) response probability ρ_k. If element k is selected it responses with probability ρ_k and with $1-\rho_k$ non-response is observed. The selected sample is denoted by two sets of indicators R_k to indicate response, and a_k to indicate whether element k is selected in the sample. The response happens when $a_k = 1$ and $R_k = 1$ and the number of available cases :

$$n_R = \sum_{k=1}^{N} a_k R_k. \tag{1.3}$$

The mean of the n_R responding elements :

$$\overline{y}_k = \frac{1}{n_R} \sum_{k=1}^{N} a_k R_k Y_k. \tag{1.4}$$

Expected value of the response mean is approximately:

$$E(\overline{y}_k) = \widetilde{Y}, \tag{1.5}$$

where

$$\widetilde{Y} = \frac{1}{N}\sum_{k=1}^{N}\frac{\rho_k}{\overline{\rho}}Y_k \tag{1.6}$$

and

$$\overline{\rho} = \frac{1}{N}\sum_{k=1}^{N}\rho_k. \tag{1.7}$$

Since the expected value of the response mean is generally unequal to the population mean, the estimator is biased and approximately is:

$$B(\overline{y}_R) = \widetilde{Y} - \overline{Y} = \frac{S_{\rho Y}}{\overline{\rho}} = \frac{R_{\rho Y}S_\rho S_Y}{\overline{\rho}}, \tag{1.8}$$

where $S_{\rho Y}$ is covariance between the values of the target variable and the response probabilities, $R_{\rho Y}$ is the corresponding correlation coefficient, S_Y is the standard deviation of the variable Y, and S_ρ is the standard deviation of the response probabilities. Following conclusions can be drawn from the expression of the bias (Biffignandi i Bethlehem 2011):

- *The bias vanishes if there is no relationship between the target variable and the response behavior. This implies $R_{\rho Y} = 0$. The stronger the relationship between the target variable and the response behavior, the larger the bias will be.*
- *The bias vanishes if all response probabilities are equal. Then $S_\rho = 0$. In this situation, the non-response just reduces the sample size.*
- *The magnitude of the bias increases as the mean of the response probabilities decreases, which means that lower response rates will lead to larger biases.*

1.4 Dealing with non-response

Missing data due to non-response is a problem that can be approached from different angles. There are three ways a researcher can deal with missing data (Szymkowiak 2019):

- **Preventive techniques** focus on avoiding the emergence of the unit and item non-response in the survey. They are to reduce skepticism and the reluctance of the respondent to participate in the study, as well as promote a positive attitude to it. Preventive measures may also include appropriate training of interviewers or proper preparation of questionnaires and the sampling frame.

- **Reductive techniques** involve the use of financial and material incentives (an encouraging factor for individuals to take part in the study), sending reminders, re-calling by phone or email, replacing units that do not participate in the study with others that possess similar characteristics (for example, from a substituted sample).
- **Corrective techniques** are statistical adjustment methods, meant to compensate for bias, caused by missing responses in the data.

Preventive and reductive techniques are exclusive for researchers that have the opportunity to intervene at the data collection stage, while others can only choose among corrective techniques to combat non-response bias. Correction techniques could be divided into *imputation*, *weighting*, *model-based*, and their combination.

Imputation is the method of ascribing some values in place of the missing data. These values should not increase bias, the variation of the estimator, nor should they lead to changes in the distribution of the feature set. Imputed values are estimated based in some way on available sample data. Imputation is usually used in the case of item non-response (Szymkowiak 2009). Some of the imputation methods include (Szymkowiak 2009):

- *Deductive imputation* – involves deducing missing value based on available information. For example, if sex of the respondent is missing but the respondent has a female name, it can be deduced that the respondent is a female.
- *Mean imputation* – involves replacing missing values with the feature mean of all the respondents or some respondents from the imputation classed.
- *Regression imputation* – involves using values from the appropriately selected regression model to replace missing values.

Weighting is the method of adjusting weights of the respondents to compensate for non-respondents. This process can be made possible with auxiliary information that is available about both respondents and non-respondents. Weighting is usually used in case of unit non-response. Some of the weighting methods include (Kalton i Flores-Cervantes 2003) :

- *Cell weighting* - adjusts sample weights based on the ratio between the sample totals and the population totals on a cell-by-cell basis. It is assumed that respondents within the cell represent non-respondents and the response mechanism is MAR. In contrary to other methods, cell weighting makes no assumptions about the structure of the response probabilities across the cells.
- *Raking* operates on marginal distributions of the sample, adjusting them iteratively to get to the population marginal distribution. During the iteration procedure, sample row and

column totals are conformed to population row and column totals in turns until convergence point is reached. Raking makes the same assumption about the response mechanism as cell weighting but also that *the response probability for cell (h,k) of the form $\phi_{hk} = \alpha_h \beta_k$, where α_h and β_k denote row and column effects.*

- *Logistic regression weighting* - uses a logistic regression model to adjust weights for non-response. Each respondent's weight is corrected with the inverse of their predicted response probability that comes from the logistic regression model. It has better flexibility then raking as it can work with uncategorized continuous predictors but cannot give weighting adjustments of less than 1.

A *combined approach* is the method that utilizes imputation and weighting methods at the same time to compensate for both item and unit non-response (Szymkowiak 2019).

Model-based are the kind of methods that have either likelihood or Bayesian analysis at their core, model observed data, and even response mechanism. These kinds of approaches are best suited for the case of the non-ignorable non-response, but can also be used when missing data is ignorable. (Sikov 2018) Methods that are going to be compared in this thesis are related to this particular group of methods.

1.4.1 Estimation under non-ignorble non-response

In earlier sections, missing data and its causes were discussed. This section will briefly discuss some statistical approaches, mainly based on maximum likelihood method for this kind of a problem. More detailed and formal explanation of the topic can be found in Kim and Shao (2013), of which this section was based on.

Missing data aside, if z is a variable observed and present throughout the sample with a distribution function $f(z;\theta)$, then population parameter θ can be estimated using maximum likelihood method. Maximum likelihood estimator (MLE) $\hat{\theta}$ of θ_0 has to maximize likelihood function $L(\theta) = f(z,\theta)$, so that $L(\hat{\theta}) = \max_{\theta \in \Omega} L(\theta)$. MLE has to be unique, therefore family of densities has to be identifiable $f(z;\theta_1) \neq f(z;\theta_2)$ for every $\theta_1 \neq \theta_2$ and for all z. This assumption about MLE is important, because if it is not met, $\hat{\theta}$ may not converge (in probability) to a single point.

Missing data can pose a real challenge for MLE. One way of dealing with missing data would be computing MLE based on *observed likelihood*. To be able to do that, data has to be separated with δ, a *response indicator* :

$$\delta_{ij} = \begin{cases} 1 & \text{if } z_{ij} \text{ is observed,} \\ 0 & \text{otherwise} \end{cases} \tag{1.9}$$

where z_{ij} is $p-$dimensional random vector with probability distribution function $f(z;\theta)$. With introduction of missing data, apart from a sample distribution model $f(z,\theta)$, *response mechanism* $Pr(\delta|z)$ also has to be modeled. Modeling *missing mechanism* is problematic, as it *is unknown and it often depends on some unknown parameter* ϕ

$$L_{obs}(\theta,\phi) = \prod_{\delta_i=1}[f(y_i;\theta)\pi(y_i;\phi)] \times \prod_{\delta_i=0}[\int f(y;\theta)\{1-\pi(y;\phi)\}d\mu(y)], \tag{1.10}$$

where $\pi(y;\phi) = P(\delta = 1|y;\phi)$. So, to deal with this complication, an assumptions about response mechanism can be used to simplify the observed likelihood. Assuming the response mechanism is MAR or in a special case MCAR, is assuming that response mechanism depends only on observed portion of the data $P(\delta,z) = P(\delta,z_{obs})$, which effectively allows for ignoring missing portion of the data while obtaining MLE.

Unfortunately, if the assumption about randomness of missing data does not hold, identifiability assumption under which MLE is obtainable, may not hold either. This is a case of non-ignorable missing data. Nonignorable missing data *occurs when the probability of response depends on the variable that is not always observed*. In other words, if missing data were to be ignored in this situation, obtained MLE cannot be trusted to be unbiased.

Nonresponse instrument are additional variables that can be helpful in some cases with a lack of parameter identifiability, caused by non-ignorable missing data. *Nonresponse instrumental variables* identify parameters, so that consistent estimates can be made inspite non-ignorable missing data. Unique MLE can be obtained in the case of non-ignorable missing data to maximize the observed likelihood, after non-response instrumental variables ensure identifibility. But as response mechanism cannot be ignored, full observed likelihood function has to be maximized and integrals that are present there *are not easy to handle*.

Outside of observed likelihood approach, there are also more sophisticated methods of estimation to deal this non-ignorable missing data. For example, *conditional likelihood approach*. *Conditional likelihood approach* avoids complicated computations, that plague observed likelihood approach, and instead uses only part of the observed sample data for parameter estimation, with non-ignorable missing data. Method also has an advantage over observed likelihood likelihood, as it can be used *even when x_i's associated with the missing y-values are not observed*. As response probability is generally unknown, this method can use an estimator of the response property, obtained with *Generalized method of moments (GMM) approach*.

1.5 Conclusions

Data sources in statistics are traditionally surveys and censuses that can be augmented with a non-statistical source, administrative records. Today, many more non-statistical data sources, like commercial transaction records and social media data are available. Data for the research can come from one or multiple sources, and errors that occur in the estimates are different for these scenarios. If some of the sampled units fail to respond, a non-response error may occur, which leads to a bias in the estimates if the non-response is non-ignorable. When non-response is non-ignorable, a response mechanism is assumed to be NMAR (not missing at random), which means, that probability of the non-response depends on the target variable and not the explanatory, auxiliary variables. Non-ignorable non-response is a problem that can be attended with preventive, reductive, and corrective techniques. Corrective techniques are the methods that with the use of a statistical adjustment, should minimize the bias after the data was collected, contrary to other techniques. The corrective methods include imputation, weighting, model-based and combined approaches. In the next chapter, selected corrective, model-based approaches that suppose to deal with non-ignorable non-response are shown.

Chapter 2

Selected methods for reducing non-ignorable non-response for categorical variables

2.1 Basic settings

In the survey data, a discrete variable of interest may have missing values due to non-response. As discussed earlier, if data is not missing at random, it cannot be simply ignored. Non-ignorable non-response, in the variable of interest, is a deficiency that affects its distribution. A researcher can use available auxiliary variables, to minimize bias that can arise from this deficiency.

Assume that the categorical variable of interest has J categories. Table 2.1 shows frequencies of that categorical variable for a sample with missing data and table 2.2 shows how the sample would look like as if there were no deficiencies

Table 2.1. Frequencies of observed information

Category	1	2	..	J	Missing	Total
Number	n_1	n_2	...	n_J	m	$n+m$

Source: Own elaboration based on Lee and Marsh (2000) and Zhang (2001)

where n_j for $j = 1,2,...,J$ is a number of individuals that belong to j^{th} category, m_j for $j = 1,2,...,J$ is a number of missing values for the j^{th} category, and m is a total number of missing values.

Table 2.2. Whole sample frequencies

Category	1	2	..	J	Total
Number	$n_1 + m_1$	$n_2 + m_2$	...	$n_J + m_J$	$n + m$

Source: Own elaboration based on Lee and Marsh (2000) and Zhang (2001)

If t^{th} individual responds to j^{th} category, the joint probability of response P_{tj0} can be calculated for each category, and for the non-respondents, the marginal probability of missing response P_{tm}, as in table 2.3.

Table 2.3. Probability Distribution of J Categories

Category	1	2	..	J	Missing	Total
Probability	P_{t1o}	P_{t2o}	...	P_{tJo}	P_{tm}	1.0

Source: Lee and Marsh (2000)

Table 2.4. Joint Probability Distribution in Terms of Observed Probabilities

Category	1	2	...	n	Marginal Prob.
Observed	P_{t1o}	P_{t2o}	...	P_{tJo}	P_{to}
Missing	$\alpha_1 P_{t1}$	$\alpha_2 P_{t2}$	...	$\alpha_J P_{tJ}$	P_{tm}
Marginal Prob.	P_{t1}	P_{t2}	..	P_{tJ}	1.0

Source: Lee and Marsh (2000)

Table 2.4 shows the probability of missing response P_{tm}, decomposed into J joint probabilities, so that

$$\alpha_1 P_{t1o} + \alpha_2 P_{t2o} + \ldots + \alpha_J P_{tJo} = P_{tm} \tag{2.1}$$

and

$$(1 + \alpha_j) P_{tjo} = P_{tj}. \tag{2.2}$$

If the data is missing at random, then each of the missing joint probabilities would be proportional to the observed, and weights $\alpha_j, j = 1, 2, \ldots, J$ would be equal. If this is not the case, then each category has its own weight and missing data is non-ignorable. This means, that structural parameters of the model cannot be estimated only using the observed responses without encountering a selection bias. In this case, missing data and auxiliary variables have to be taken into consideration, as well as, observed responses. Later in this section, some of the approaches that deal with non-ignorable missing data are going to be described. Both of these approaches develop weights, to correct bias caused by non-response. These weights are used to estimate the total number of individuals that belong to each category of the variable of interest.

2.2 Selected methods

2.2.1 Class weighting technique

This technique of weighting classes was proposed by Zhang (2001). It takes advantage of the inclusion probabilities of the units in a given sample and auxiliary variables to compensate for missing data. The approach can be used in case of non-ignorable non-response and is an extension of the standard procedure that assumes ignorable response model.

Under the ignorable model response depends on auxiliary variables, but not on the response variable:

$$P[R_t = 1|x_t = x, y_t = y] = P[R_t = 1|x_t = x], \tag{2.3}$$

where $R_t = 1$ if response variable is observed for t^{th} unit and $R_t = 0$ otherwise. y_t is a response class indicator for $1, ..., J$ classes and x_t is an auxiliary variable available for all $t = 1, ..., T$ units, with possible values ranging from 1 to K. From the response model, since response does not depend on a response variable y but only on auxiliary variable x, estimation can be based only on the observed portion data.

The standard weighting technique goes through the following steps:

1. The sample is grouped into c_{xy} cells, where $c_{xy} = n_{xy}$ is a number of observed units with $x_t = x$ and $y_t = y$.
2. The weights are obtained in the following way:

$$w_t = c_x^{-1} N_x \quad \text{for} \quad t \in s_x, \tag{2.4}$$

where s_x is a subsample where $x_t = x$, and N_x is number of units within the subpopulation where $x_t = x$. Assuming all units among the subgroups $x_t = x$ have the same inclusion probability in the survey sample design:

$$c_x^{-1} c_{xy} \quad \text{where} \quad c_x = \sum_{j=1}^{J} c_{xj} \tag{2.5}$$

is an estimate of the unit selection propapability from the subpopulation where $x_t = x$ and $y_t = y$.

3. A total number of units within response categories $y_t = y$ is estimated with:

$$\hat{T}_y = \sum_x \sum_{t \in s_x} w_t I_{y_t = y} \tag{2.6}$$

where $I_{y_t=y} = 1$ if $y_t = y$ and $I_{y_t=y} = 0$ otherwise.

Under the non-ignorable model response depends on response variable, but is independant of auxiliary variable :

$$P[R_t = 1 | x_t = x, y_t = y] = P[R_t = 1 | y_t = y] \tag{2.7}$$

With non-ignorable non-response model estimates cannot be made with only observed portion of the data, and the technique is a bit more complicated from the standard approach because of that. The same as in standard approach, the sample is grouped into c_{xy} cells. Unlike the standard approach, $c_{xy} = n_{xy} + m_{xy}$ is a sum of observed and missing number of units with $x_t = x$ and $y_t = y$. Number of non-respondents m_{xy} for $x_t = x$ and $y_t = y$ is unavailable, except from the marginal total $m_x = \sum_y m_{xy}$. To calculate weights, estimates $\hat{m}_{xy}$ of the unobserved part of the sample can be used. To get this estimates, a EM algorithm for simple multinomial sampling data can be used and it goes as follows:

$$\hat{P}[R_t = 0 | y_t = y] = \frac{\sum_x \hat{m}_{xy}}{\sum_x n_{xy} + \sum_x \hat{m}_{xy}}, \tag{2.8}$$

and

$$\hat{E}[m_{xy} | n_{xy} + \hat{m}_{xy}] = (n_{xy} + \hat{m}_{xy} \hat{P}[R_t = 0 | y_t = y]) \tag{2.9}$$

Estimate $\hat{m}_{xy}$ is updated conditioned to observed $m_x = \sum_y \hat{m}_{xy}$ by

$$\hat{m}_{xy} = \frac{m_x \hat{E}[m_{xy} | n_{xy} + \hat{m}_{xy}]}{\sum_{j=1}^{J} \hat{E}[m_{xj} | n_{xj} + \hat{m}_{xj}]}, \tag{2.10}$$

and iterate, after setting adequate sizes of J and K if needed, to avoid dealing with empty or the ones that have small values.

Estimates $\hat{m}_{xy}$ for $y = 1, ..., J$ should satisfy the following:

$$\frac{\hat{m}_{1y}}{n_{1y} + \hat{m}_{1y}} = \frac{\hat{m}_{2y}}{n_{2y} + \hat{m}_{2y}} = ... = \frac{\hat{m}_{Ky}}{n_{Ky} + \hat{m}_{Ky}}. \tag{2.11}$$

Weights can be calculated for the imputed sample, where $\hat{c}_{xy} = n_{xy} + \hat{m}_{xy}$, in the following way:

$$w_t = N \frac{(\pi_t)^{-1} a_t}{\sum_{t \in s; r_t = 1} (\pi_t)^{-1} a_t}, \tag{2.12}$$

where π_t is the inclusion probability of unit t, $N = \sum_{t \in s} (\pi_t)^{-1} = \sum_{t \in s; r_t = 1} w_t$ is the size of the population.

The adjustment weight a_t of a respondent t is

$$a_t = \frac{\sum_x n_{xy} + \sum_x \hat{m}_{xy}}{\sum_x n_{xy}}. \tag{2.13}$$

After all weights have been calculated, the population total can be estimated with:

$$\hat{T} = \sum_{t \in s; r_t = 1} w_t y_t = \sum_{t \in s} r_t w_t y_t, \tag{2.14}$$

and a total number of units within response categories $y_t = y$ is:

$$\hat{T}_y = \sum_{t \in s; r_t = 1} w_t I_{y_t = y} = \sum_{t \in s} r_t w_t I_{y_t = y}. \tag{2.15}$$

2.2.2 Selection bias modeling

A technique for recovering the underlying probability distribution, that is going to be described here, was proposed by Lee and Marsh (2000). The approach assumes that non-response is not related to the lack of the correct option. But rather that each individual participating in the survey, does fall into one of the suggested categories and, for some reason, fails to respond. In other words, if there are J categories in a survey that an individual can respond to, whether the individual responds or not, s\he would still belong to either one of those J categories.

This approach can be employed to correct estimates effected by non-response by estimating parameter of the following multinomial response model with a use of auxiliary information.

$$y^*_{tj} = x'_{tj} \beta_j + e_{tj}, \quad t = 1, ..., T \ and \ j = 1, ..., J \tag{2.16}$$

In this model y^*_{tj} is a latent variable that represents a categorical variable and is a dummy variable. Response variable y_{tj} is observed where $y_{tj} = 1$ and $y_k = 0$ for all $k \neq j$ if the t^{th} individual belongs to the j^{th} category. Individual response variable y_{tj} is observed if the following selection criteria is met:

$$s_t^* = x^{'}\gamma + \delta_t \quad and \quad s_t^* \geq 0, \tag{2.17}$$

where s_t equals 1 if $s_t^* \geq 0$ and 0 otherwise, which means that in this model the t^{th} individual responds if the threshold of 0 in selection criteria is surpassed.

Residual term ε_{tj} is assumed to have a multinomial logit distribution, and in turn, marginal probability P_{tj} of each category j has a multinomial logit probability

$$P_{tj} = \frac{e^{x_{tj}^{\prime}\beta_j}}{\sum_{i=1}^{J} e^{x_{ti}^{\prime}\beta_i}}. \tag{2.18}$$

The observed (joint) probability of j^{th} category is

$$P_{tjo} = \frac{e^{x_{tj}^{\prime}\beta_j}}{(1+\alpha_j)\sum_{i=1}^{J} e^{x_{ti}^{\prime}\beta_i}}. \tag{2.19}$$

The marginal probability of observed is

$$P_{to} = \sum_{j=1}^{J} P_{tjo} = \sum_{j=1}^{J} \frac{e^{x_{tj}^{\prime}\beta_j}}{(1+\alpha_j)\sum_{i=1}^{J} e^{x_{ti}^{\prime}\beta_i}}, \tag{2.20}$$

and the marginal probability of missing is

$$P_{tm} = \sum_{j=1}^{J} \alpha_j \cdot P_{tjo} = \sum_{j=1}^{J} \alpha_j \cdot \frac{e^{x_{tj}^{\prime}\beta_j}}{(1+\alpha_j)\sum_{i=1}^{J} e^{x_{ti}^{\prime}\beta_i}} \tag{2.21}$$

Explanatory variable $x_{tj}^{'}$ is a realization of transposed matrix, where each column is a vector of an auxiliary information and β, α are parameters to be estimated. This approach doesn't need to estimate parameters of selection criteria. Missing probabilities, expressed as the weighted sum of observed joint probabilities, can be used to estimate model parameters.

Estimation of parameters of the model can be done with maximum likelihood method. The log-likelihood function for this multinomial response model with missing response observations is

$$lnL(\beta,\alpha) = \sum_{t=1}^{T}\left(\sum_{j=1}^{J} y_{tj} \cdot lnP_{tjo}\right) + y_{tm} \cdot lnP_{tm} = \sum_{t=1}^{T}\left(\sum_{j=1}^{J} y_{tj} \cdot lnP_{tjo}\right) + y_{tm} \cdot ln\left(\sum_{j=1}^{J} \alpha_j \cdot P_{tjo}\right) \tag{2.22}$$

The log-likelihood function consists of two parts. First part is the observed joint probability of each category and the second part is the marginal probability of missing observations expressed as the weighted sum of the observed joint probabilities.

The first order conditions for the log-likelihood function are:

$$\frac{\partial lnL_3(\beta,\alpha)}{\partial\beta_j}=\sum_{t=1}^{T}\{\sum_{j=1}^{5}x'_{tj}(y_{tj}-\sum_{j=1}^{5}\frac{e^{x'_{tj}\beta_j}}{\sum_{i=1}^{5}e^{x'_{ti}\beta_i}}+y_{tm}(\sum_{j=1}^{5}\frac{\frac{\alpha_j e^{x'_{tj}\beta_j}}{1+\alpha_j}}{\sum_{i=1}^{5}\frac{\alpha_i e^{x'_{tj}\beta_i}}{1+\alpha_i}}-\sum_{j=1}^{5}\frac{e^{x'_{tj}\beta_j}}{\sum_{i=1}^{5}e^{x'_{tj}\beta_i}}))\} \tag{2.23}$$

$$\frac{\partial lnL_3(\beta,\alpha)}{\partial\alpha_j}=\sum_{t=1}^{T}\{\sum_{j=1}^{5}y_{tj}\times(\frac{-1}{1+\alpha_j})+y_{tm}\times(\frac{\frac{e^{x'_{tj}\beta_j}}{(1+\alpha_j)^2}}{\sum_{j=1}^{5}\frac{\alpha_j e^{x'_{tj}\beta_j}}{1+\alpha_j}})\} \tag{2.24}$$

2.3 Conclusions

Two methods that were presented in earlier sections are both corrective methods that make use of available auxiliary information to mitigate the effect of missing data that are not missing at random. Both of these methods have some simplified versions that assume MAR missing data mechanism. The main difference is that the method presented under the name *class weighting technique* is based on a non-parametric model, and the one named *selection bias modeling* based on a parametric model.

In the next chapter, the simulation will be conducted, where the generated data will be used to compare both methods. For the data with 3 response categories and 2 auxiliary variables there are 10 parameters in total, that the *selection bias modeling* method is trying to estimate, of which only 6 are comparable between variations of the method. *Class weighting technique,* instead of parameters, estimates matrix that contains a number of non-respondents for each response category and auxiliary variable. For the data with 3 response categories and 2 auxiliary variables, this is the 3x4 matrix, with 12 individual elements to estimate. Estimation of the matrix is exclusive to the NMAR variant, which is why it cannot be the object of comparison.

Among other things, in practical terms, this means that the bias of estimated parameters will be assessed only for the *selection bias modeling* method, in addition to comparing bias of the response category proportions estimates. In the next chapter, the simulation will be conducted, where the generated data will be used to compare both methods. The methods will be first compared to their simplified versions to uncover obvious weak points and check if there are any improvements in the NMAR mechanism variants. Then the best ones from both camps will be compared based on their ability to estimate response category proportions.

Chapter 3

Empirical comparison of selected methods for non-ignorable non-response

3.1 Data simulation procedure

The simulation study has been conducted to compare methods for dealing with non-response described in Chapter 2. In this limited study we solely focus on bias in the estimated proportions for categorical variable with three levels. For parametric model we also report bias in estimated model parameters.

The pseudo-random data was generated in a way, that it consists of auxiliary variables, a response variable, and a response indicator that is equal to 1 if the unit responds and 0 otherwise. The response variable was generated with the use of a multinomial logistic regression:

$$\Pr(Y_t = j) = \frac{e^{\beta_{0j}+x'_{tk}\beta_{kj}}}{\sum_{j=1}^{J} e^{\beta_{0j}+x'_{tk}\beta_{kj}}}, \tag{3.1}$$

where Y_t is the response variable, for the t^{th} unit, with J categories. β's are the parameters of the model and x_{tk} is the explanatory variable k for t^{th} unit. Each explanatory variable was generated from the multinomial distribution with probability p defined as above. Next, the response indicator R_t for the t^{th} unit was generated assuming non-ignorable non-response, i.e.

$$R_t = \frac{e^{\gamma_0+\gamma_1 y_t}}{1+e^{\gamma_0+\gamma_1 y_t}}, \tag{3.2}$$

where $\gamma_i, i \in \{0,1\}$ is the parameters of the selection model and y_t are observed values of Y_t.

In the simulation study we assume that Y variable is generated by two categorical variables X_1 and X_2 and collection of (Y, X_1, X_2) we treat as population (i.e. we know the true proportions). Then, the procedure consists of two steps: 1) simple random sample from the population and 2) generating non-ignorable non-response for the sampled units. The following pseudo-code presents the simulation

```
FUNCTION        Population (Number of rows, parameters)
    $x_1 \sim Bernoulli(p_1)$
    $x_2 \sim Bernoulli(p_2)$
    $\eta_1 \leftarrow \beta_{01} + \beta_{11}x_1 + \beta_{21}x_2$
    $\eta_2 \leftarrow \beta_{02} + \beta_{12}x_1 + \beta_{22}x_2$
    $\eta_3 \leftarrow \beta_{03} + \beta_{13}x_1 + \beta_{23}x_2$
    $ex1 \leftarrow \frac{\eta_1}{\eta_1 + \eta_2 + \eta_3}$
    $ex2 \leftarrow \frac{\eta_2}{\eta_1 + \eta_2 + \eta_3}$
    $ex3 \leftarrow \frac{\eta_3}{\eta_1 + \eta_2 + \eta_3}$
    $y \sim Multinomial(ex1, ex2, ex3)$
END FUNCTION
FUNCTION        Data (population,number of rows in sample, iterations)
    FOR iterations
    sampled data ← sample(population)
    $prob \leftarrow \frac{\exp\{\gamma_0 + \gamma_1 y\}}{1 + \exp\{\gamma_0 + \gamma_1 y\}}$
    $r \sim Bernoulli(prob)$
    END FOR
END FUNCTION
```

In the simulation, we define population size as $N = 100,000$ units with data generated from each parametric model with a response variable that has 3 distinct categories and 2 auxiliary variables. Then $B = 100$ samples, of $n = 1000$ units, were drawn from that population using simple random sampling.

The simulation study compares performance of the selected methods for five different distributions of the response variable and levels of non-response. The response mechanism was chosen to recreate the case of non-ignorable non-response in simulated data, so that missing values are not missing at random and the response indicator is directly correlated with the response variable but not with auxiliary (explanatory) variables. To ensure the non-ignorable nature of the missing data, each parametric model, for all the response levels, had greater non-response in either the first category or the third category of the response variable.

The non-response rates that were chosen for the simulation study are: 10%, 20% and 30%. The parameters used in each parametric model (PM) and non-response level (NL) can be seen in the table 3.1

Table 3.1. Parameters of the parametric models for target variables Y (PM) and response patterns (R)

	PM_1	PM_2	PM_3	PM_4	PM_5
β_{01}	0.5	0.4	0.3	0.4	0.2
β_{11}	0.8	0.8	0.4	0.5	0.3
β_{21}	0.6	0.6	0.2	0.3	0.1
β_{02}	0.4	0.4	0.2	0.4	0.3
β_{12}	0.5	0.5	0.4	0.8	0.4
β_{22}	0.35	0.3	0.1	0.6	0.3
β_{03}	0.1	0.2	0.25	0.2	0.4
β_{13}	0.1	0.2	0.19	0.35	0.8
β_{23}	0.1	0.1	0.09	0.45	0.6

	R_1	R_2	R_3
	90%	80%	70%
γ_0	0.59	0.513	0.307
γ_1	1	0.5	0.3

Figure 3.1 shows distributions of a response variable subject to non-response, generated with a use of five distinct parametric models, where for the parametric model number:

1. the distribution of the response variable is right skewed and the non-response unequally affects first category more then other categories,
2. the distribution of the response variable is right skewed and the non-response unequally affects third category more then other categories,
3. the distribution of the response variable is right skewed, although somewhat flattened and the non-response unequally affects first category more then other categories, so that it becomes even more flattened,
4. distribution of the response variable is roughly symmetrical and the non-response unequally affects first category more then other categories, so that it becomes more symmetrical,

5. the distribution of the response variable is left skewed and the non-response unequally affects third category more then other categories.

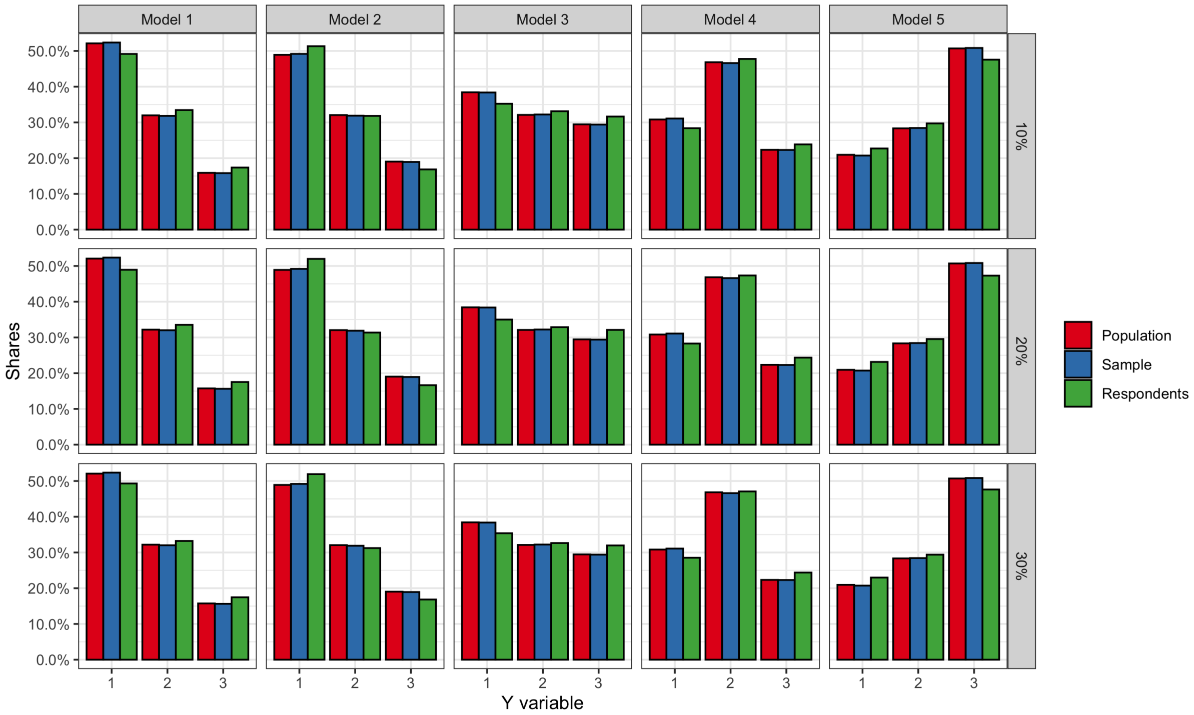

Figure 3.1. Distributions of the response variable Y, generated from the parametric models

Source : own elaboration

Each of these parametric models aims to investigate five situations, to test the performance of selected correction methods on. Although these distributions may be just a glimpse of all possible cases, but they may be enough to uncover some of the obvious weaknesses of the selected correction methods.

Table 3.2 shows correlation between response category variable and auxiliary variables (XY), response category variable and auxiliary variables for respondents (XYr), response category variable and response indicator (RY), auxiliary variables, and response indicator (RX). As is can be seen from the table, non-response in generated data is non-ignorable. There is some correlation between response category variable and auxiliary variables, but what is more important, the correlation between response indicator and auxiliary variables is a lot smaller, then correlation between response indicator and response category variable. In generated data, with higher non-response level, correlation between response indicator and response category variable weakens.

NL		PM1	PM2	PM3	PM4	PM5
	XY	0.145	0.130	0.047	0.012	-0.130
10%	XYr	0.150	0.126	0.048	0.018	-0.131
	RY	0.177	-0.202	0.201	0.183	-0.190
	RX	0.018	-0.029	0.004	-0.005	0.017
	XY	0.145	0.130	0.047	0.012	-0.130
20%	XYr	0.151	0.128	0.046	0.018	-0.132
	RY	0.136	-0.148	0.158	0.138	-0.150
	RX	0.014	-0.024	0.002	-0.006	0.013
	XY	0.145	0.130	0.047	0.012	-0.130
30%	XYr	0.147	0.130	0.046	0.017	-0.130
	RY	0.097	-0.105	0.112	0.100	-0.111
	RX	0.009	-0.021	-0.002	-0.006	0.010

Table 3.2. Correlations in generated data

3.2 Comparison of missing response correction methods performance on simulated data

The described above generated data was used to estimate mean absolute bias for two variants of *class weighting technique* and *selection bias modeling*. The case where missing values were ignored, was used as a baseline. Estimated biases are presented in the tables below, where the column **bias_ignore** represents the case where missing values were ignored, **bias_mar** response mechanism is assumed to be MAR, and **bias_nmar** response mechanism is assumed to be NMAR (i.e. non-ignorable). The absolute bias based on B replications is calculated in the following way:

$$\begin{aligned}
\textbf{bias\_ignore}_{cml} &= \sum_{b=1}^{B} \left| \theta^b_{cml} - \theta^b_{\text{cml,ignore}} \right| / B, \\
\textbf{bias\_mar}_{cml} &= \sum_{b=1}^{B} \left| \theta^b_{cml} - \theta^b_{\text{cml,mar}} \right| / B, \\
\textbf{bias\_nmar}_{cml} &= \sum_{b=1}^{B} \left| \theta^b_{cml} - \theta^b_{\text{cml,nmar}} \right| / B,
\end{aligned} \tag{3.3}$$

where θ represents actual proportions in the generated data, θ_{ignore}, θ_{mar} and θ_{nmar} represent estimated proportions based on B-times generated data with missing values for each method within each variation of category c, parametric model m and level of non-response l.

3.2.1 Results for the class weighting technique

3.2.1.1 Results for proportions

Table 3.3 shows the mean of bias estimared according to (3.3).

Table 3.3. Mean bias for each response category (cwt)

category	bias_ignore	bias_mar	bias_nmar
1	0.028053	0.027766	0.014528
2	0.008692	0.008569	0.008822
3	0.023780	0.023547	0.016100

The results of the **bias_mar** are very close to those in **bias_ignore**, as expected, but the ones in **bias_nmar** are generally better for all the categories except the second one. The performance of *class weighting technique* is varied across presented parametric models.

Table 3.4. Mean bias for each parametric model (cwt)

pmodel	bias_ignore	bias_mar	bias_nmar
1	0.021170	0.020866	0.006483
2	0.017273	0.016876	0.004848
3	0.021526	0.021518	0.018492
4	0.018000	0.017966	0.018649
5	0.022904	0.022578	0.017275

There it can be seen that for the generated data, selected non-parametric model was better, for ordered categories, like the ones generated with parametric models 1,2,5 and less so 3. Figure 3.2 shows how good did the method perform for 1,2,5 parametric models.

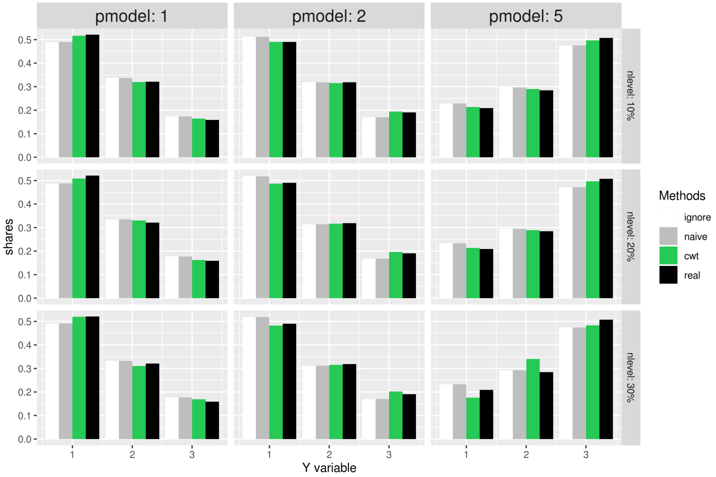

Figure 3.2. Distributions of the response variable Y, generated from the selected parametric models (cwt)

Source : own elaboration

Figure 3.3 shows the performance of the method for the parametric model 4, where **bias_nmar** was slightly greater then **bias_mar**. Parametric model 4, where improvement was less then for other parametric models, is also shown in the figure 3.3.

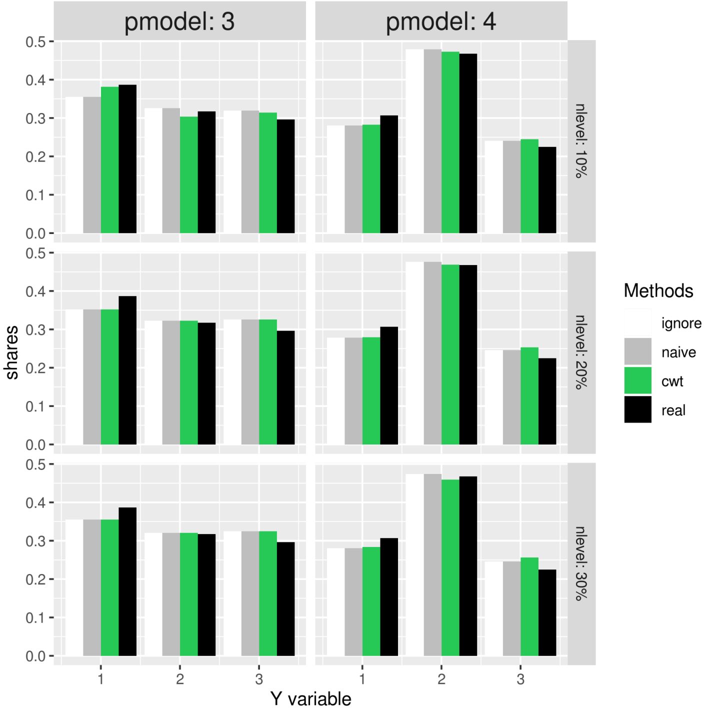

Figure 3.3. Distributions of the response variable Y, generated from the selected parametric models (cwt)

Source : own elaboration

The increase of non-response level has some detrimental effect on *class weighting technique*'s ability to minimize bias. As it can be seen in the table 3.5, even though for 10% and 20% improvement was more or less the same, for the 30% level the bias even increased. The table 3.5 shows the mean bias for each level of non-reponse

Table 3.5. Mean bias for each level of non-response (cwt)

nlevel	bias_ignore	bias_mar	bias_nmar
10%	0.019168	0.018979	0.008226
20%	0.021424	0.021198	0.012192
30%	0.019933	0.019705	0.019031

This dynamic can also be seen within each parametric model in the table 3.6 that shows mean bias for each level of non-response within a parametric model.

Table 3.6. Mean bias for each parametric model within a level of non-response (cwt)

nlevel	pmodel	bias_ignore	bias_mar	bias_nmar
10%	1	0.021148	0.020844	0.003857
	2	0.014124	0.013828	0.002323
	3	0.020780	0.020764	0.011686
	4	0.017634	0.017612	0.016340
	5	0.022152	0.021847	0.006925
20%	1	0.022419	0.022092	0.008150
	2	0.018643	0.018231	0.004132
	3	0.022937	0.022920	0.022936
	4	0.018921	0.018895	0.018519
	5	0.024199	0.023851	0.007222
30%	1	0.019945	0.019661	0.007443
	2	0.019053	0.018568	0.008091
	3	0.020862	0.020870	0.020854
	4	0.017447	0.017392	0.021088
	5	0.022361	0.022036	0.037678

3.2.2 Results for the sample selection modeling method

The *selection bias modeling* method, being based on a parametric model, has a number of parameters as an output of log-likelihood estimation. Although, the number of parameters used for the data generation is the same as in the model assumed by the method, these parameters cannot be directly compared. The parameters in the output come in the J sets of βs for each explanatory variable, where J is the number of response categories, just like the ones used for the data generation. The difference is that, the first set of parameters is set to zero as a normalization, so that in the output there are $J-1$ sets of parameters instead of J. To overcome this issue, the simplified variant of the *selection bias modeling* method was fed a complete data, to estimate true parameters, to which the rest will be compared to. Apart from that, bias in estimated proportions was compared, like for the *class weighting technique*.

3.2.2.1 Results for model parameters

The results of the *selection bias modeling* of the parameter estimation are varied for the generated data. For the majority of the parameters, **bias_nmar** is less then **bias_mar**, although not by much. Mean bias for each parametric model is not less ambigous.

Table 3.7. Mean bias for each parameter (sbm)

param	bias_mar	bias_nmar
1	0.087680	0.125978
2	0.007499	0.007739
3	0.006650	0.006193
4	0.169018	0.107320
5	0.007923	0.007898
6	0.006711	0.006496

Table 3.8. Mean bias for each parametric model (sbm.p)

pmodel	bias_mar	bias_nmar
1	0.053837	0.044373
2	0.041511	0.037073
3	0.049149	0.032235
4	0.048721	0.050868
5	0.044682	0.053472

Although, for three out of five parametric models **bias_nmar** is better then **bias_mar**. For parametric models 4 and 5, which distributions seem to have nothing in common, it is not the case.

The higher non-response levels do worsen the results of *selection bias modeling* NMAR method, but it's for some reason, under the 20% non-response level, **bias_nmar** has worse results then **bias_mar**, while for the 30% non-response level, **bias_nmar** is somewhat better then **bias_mar**. The table 3.9 shows the mean bias for each level of non-reponse.

Table 3.9. Mean bias for each level of non-response (sbm.p)

nlevel	bias_mar	bias_nmar
10%	0.044144	0.030278
20%	0.051017	0.055238
30%	0.047578	0.045297

3.2.2.2 Results for proportions

Table 3.10 shows the mean bias for each response category. There it can be seen that **bias_nmar** is better, for all categories except the second one, where the change is negative.

category	bias_ignore	bias_mar	bias_nmar
1	0.028053	0.027767	0.020570
2	0.008692	0.008569	0.024846
3	0.023780	0.023549	0.019274

Table 3.10. Mean bias for each response category (sbm)

Table 3.11 shows the mean bias for each parametric model. Results of the *selection bias modeling* NMAR method are very unpromissing. The only positive change can be observed for the parametric models 3 and 5, where the improvement is minor. For the rest of the parametric models, bias has increased from those in **bias_ignore** and **bias_mar**.

Table 3.11. Mean bias for each parametric model (sbm)

pmodel	bias_ignore	bias_mar	bias_nmar
1	0.021170	0.020866	0.026789
2	0.017273	0.016877	0.022353
3	0.021526	0.021519	0.020386
4	0.018000	0.017966	0.019051
5	0.022904	0.022580	0.019236

Some of that can be contributed to the negative effect of the higher non-response levels on the estimates. Table 3.12 shows the mean bias for each level of non-response. Similar to the results of the parameter estimation, for the 20% non-response rate, **bias_nmar** was the worst, except that here **bias_nmar** was slightly bigger then **bias_mar** for the 30% non-response level.

Table 3.12. Mean bias for each level of non-response (sbm)

nlevel	bias_ignore	bias_mar	bias_nmar
10%	0.019168	0.018979	0.014539
20%	0.021424	0.021200	0.029311
30%	0.019933	0.019705	0.020839

Table 3.13 shows the mean bias for each parametric model within a level of non-response non-response. This is the table, where a sensitivity to the non-response level is even more clear. The *selection bias modeling* NMAR method, seem to be quite effective for the 10% non-response. Only one out of five parametric models have greater **bias_nmar** then **bias_mar**. For the 20% and 30% non-response, on the other hand, only two out of five have lesser **bias_nmar** then **bias_mar**.

Under the *selection bias modeling*, **bias_nmar** saw improvement for eight out of fifteen variations of the parametric models, which can be seen in the figure 3.4. Figure 3.5 portraits the results of the rest seven variation, that saw no improvement, and have even worsen.

Table 3.13. Mean bias for each parametric model within a level of non-response (sbm)

nlevel	pmodel	bias_ignore	bias_mar	bias_nmar
10%	1	0.021148	0.020842	0.013292
	2	0.014124	0.013827	0.011049
	3	0.020780	0.020765	0.012517
	4	0.017634	0.017612	0.023741
	5	0.022152	0.021850	0.012096
20%	1	0.022419	0.022095	0.042987
	2	0.018643	0.018236	0.048284
	3	0.022937	0.022921	0.016551
	4	0.018921	0.018895	0.006014
	5	0.024199	0.023854	0.032720
30%	1	0.019945	0.019661	0.024088
	2	0.019053	0.018568	0.007725
	3	0.020862	0.020870	0.032091
	4	0.017447	0.017391	0.027400
	5	0.022361	0.022036	0.012893

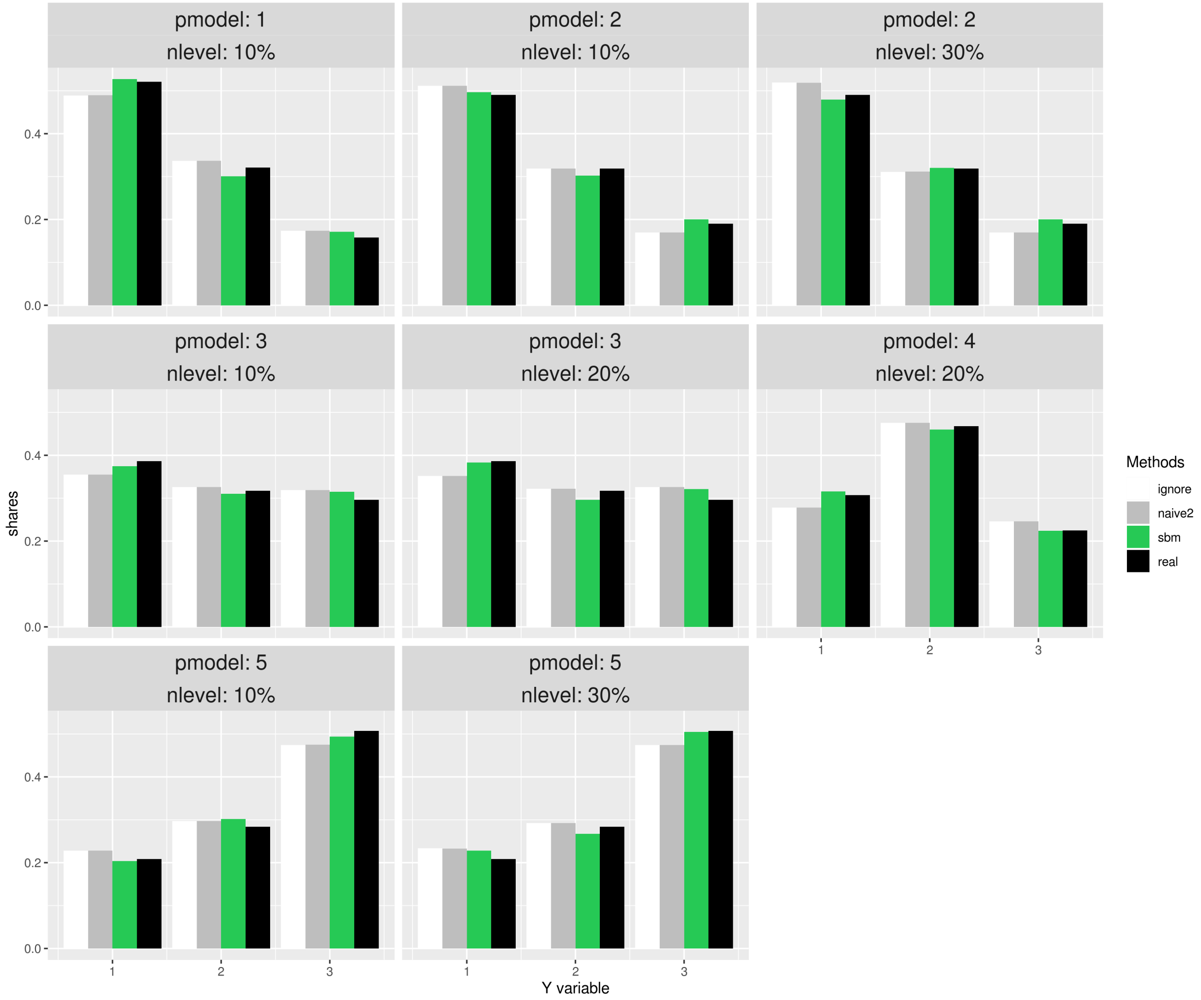

Figure 3.4. Distributions of the response variable Y, generated from the selected parametric models (sbm)

Source : own elaboration

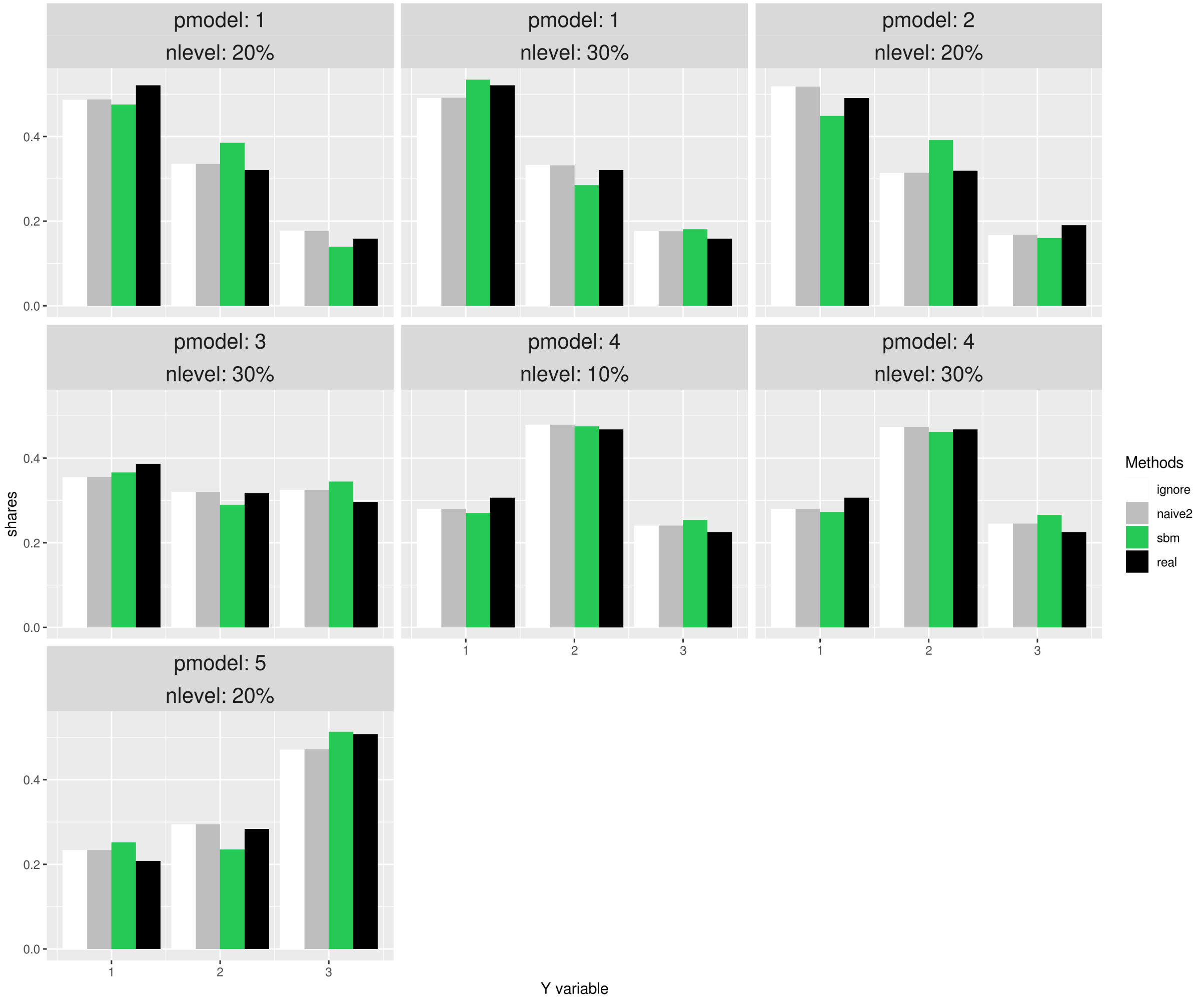

Figure 3.5. Distributions of the response variable Y, generated from the selected parametric models (sbm)

Source : own elaboration

3.3 Summary of the results from both methods

In the earlier sections, *class weighting technique* (cwt) and *selection bias modeling* (sbm) were compared to their MAR variants. In both cases, the variant that assumed NMAR was better then the one with MAR, although in the case of *selection bias modeling* - results were ambiguous. This section is devoted to comparison of the mean bias for both methods and summary of the differences between them. Tables used in the section, are very similar to the ones presented in earlier sections where the same equations are used to calculate bias.

From the table 3.14, it can be seen, that mean bias is better, for all categories except for the second one, where the change is negative, for either method. But improvements made by **cwt** is more drastic. Problem with the second category might be cause by the fact that the second

category	bias_ignore	bias_sbm	bias_cwt
1	0.028053	0.020570	0.014528
2	0.008692	0.024846	0.008822
3	0.023780	0.019274	0.016100

Table 3.14. Mean bias for each response category

category was not affected by the non-response much, which could have led to some confusion. Even though, mean bias for the **cwt** is slightly bigger then the for the ignore method, for the **sbm** the difference is noticeably greater for the second category.

pmodel	bias_ignore	bias_sbm	bias_cwt
1	0.021170	0.026789	0.006483
2	0.017273	0.022353	0.004848
3	0.021526	0.020386	0.018492
4	0.018000	0.019051	0.018649
5	0.022904	0.019236	0.017275

Table 3.15. Mean bias for each parametric model

Even larger divide between two methods, can be seen in the table 3.15. For the 3 out of 5 parameter models, **cwt** managed to make notable positive change in mean bias estimates, while **sbm** managed to very slightly improve only 2 out of 5. For the **cwt**, this were the same parametric models, for which stronger correlation between response category variable (Y) and auxiliary variables (X) was observed in generated data, as it can be seen in table 3.2.

For the rest of the parametric models, **cwt** had very low increase of the mean bias, which is not the case for **sbm**.

nlevel	bias_ignore	bias_sbm	bias_cwt
10%	0.019168	0.014539	0.008226
20%	0.021424	0.029311	0.012192
30%	0.019933	0.020839	0.019031

Table 3.16. Mean bias for each level of non-response

Detrimental effect of the larger non-response levels on performance of the estimates is present for both methods, and show in 3.16. However, mean bias estimates for **cwt** is lower then **bias_ignore** for all three non-response levels, while for **sbm** only mean bias for the 10% non-respones level is under **bias_ignore**. From table 3.16, can be concluded that **sbm** is only useful for the non-response levels around 10% or lower, while **cwt** remains useful up to the 30% non-response. For the given generated data, **cwt** appears to have way better results then **sbm**, even for the 10% non-response rate, which can be seen in the table 3.17 and in the figure 3.6.

nlevel	pmodel	bias_ignore	bias_sbm	bias_cwt
10%	1	0.021148	0.013292	0.003857
	2	0.014124	0.011049	0.002323
	3	0.020780	0.012517	0.011686
	4	0.017634	0.023741	0.016340
	5	0.022152	0.012096	0.006925
20%	1	0.022419	0.042987	0.008150
	2	0.018643	0.048284	0.004132
	3	0.022937	0.016551	0.022936
	4	0.018921	0.006014	0.018519
	5	0.024199	0.032720	0.007222
30%	1	0.019945	0.024088	0.007443
	2	0.019053	0.007725	0.008091
	3	0.020862	0.032091	0.020854
	4	0.017447	0.027400	0.021088
	5	0.022361	0.012893	0.037678

Table 3.17. Mean bias for each parametric model within a level of non-response

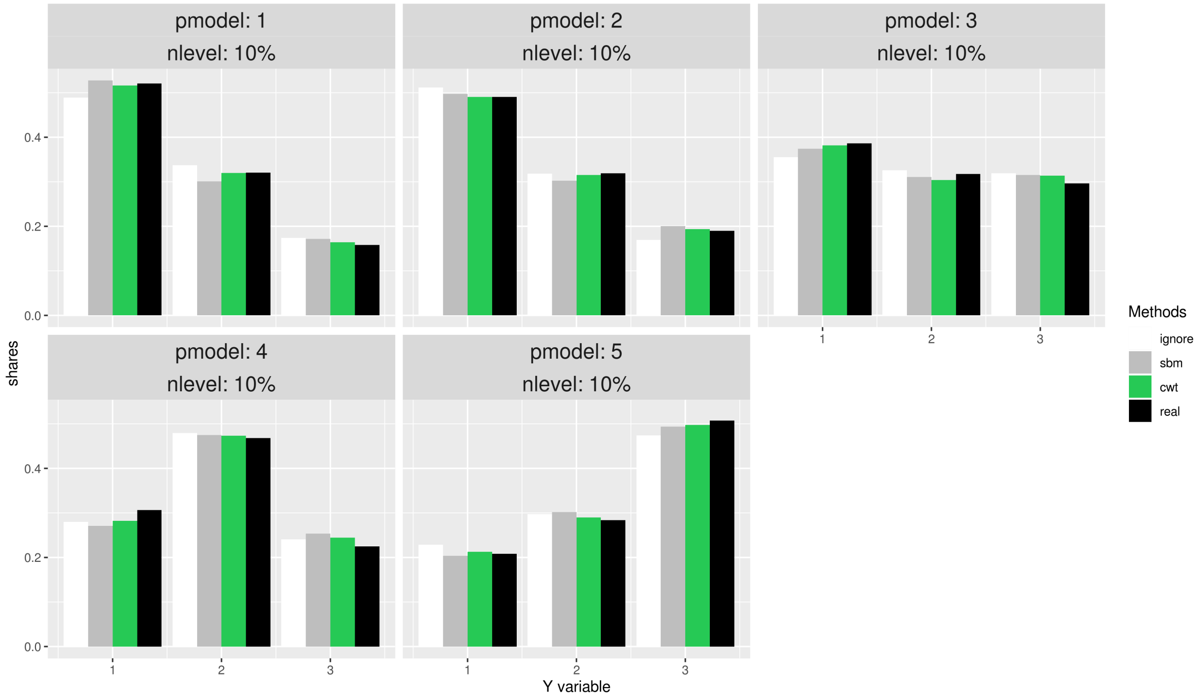

Figure 3.6. Distributions of the response variable Y, generated from the selected parametric models for 10% non-response rate

Source : own elaboration

The main differences between the compared method, can be summarized below, based on all the presented results.

Selection bias modeling is less sensitive to the shape of the response category variable distribution then, to the level of non-response. It remains effective for up to 10% non-response, but has problems with categories, that were not affected by the non-response. In the table 3.17, it can be seen, that the method is able to decrease bias in most estimates of the parameters for the 10% non-response.

Class weighting technique seem to be more sensitive to the shape of the distribution, as it can be seen in the table 3.6. The method achieves good results for the parametric models, that have either descending or ascending order in proportion of the generated response categories, which illustrates figure 3.2. The method remains effective for up to 30% non-response, and is has minimal negative effect on the categories, that were not affected by the non-response.

Conclutions

Modern-day alternative data sources are of great variety since more and more data is collected by public and private entities. But with more data sources comes also the problem of non-response. When missing data is not missing at random, it cannot be ignored, or else, the bias in estimates can occur. This problem can be prevented during the data collection phase, or mitigated afterward. Correction methods need auxiliary variables, which are not subject to non-response and in some way are related to the variable of interest, that suffers from the non-response. These variables can come from the survey itself, or from alternative data sources.

The goal of the thesis was to review two different methods of dealing with non-ignorable non-response and compare their performance on simulated data. Both of the methods claim the ability to deal with non-ignorable missing data and can be classified as model-based approaches, since they model response in one way or another. The method named *class weighting technique (cwt)* is based on some weighting method, and the one named *selection bias modeling (sbm)* - on the multinomial response model. Modeling of the response and assumption about NMAR response mechanism makes both methods the model-based approaches, able to deal with non-ignorable non-response.

To compare both methods empirically, the simulation data was generated in the way, so that there will be 3 response categories and 2 auxiliary variables for each of the 15 different populations. These populations were generated from 5 distinct parametric models and 3 levels of non-response. The data simulation, all the calculations, and presentation of the results was done with the use of self-produced R codes. The results for *cwt* were more consistent and generally better from those produced for *sbm*. Both methods, nevertheless, start to lose effectiveness for higher non-response levels, although *sbm* is more sensitive to that. The obvious weak point of the *sbm*, in addition to high non-response sensitivity, is that it can also make significant negative changes, unlike *cwt*. The *cwt* is not perfect. It struggles to make any positive changes for unordered distributions of the response category, or the flattened ones.

As a final remark, it should be noted that, the detrimental effect of the particular nonresponse levels on the performance of the methods, presented in the thesis, may very much vary for the real data or even data simulated in a slightly different fashion.

Bibliography

Beręsewicz, M. (2016). Internet data sources for real estate market statistics, 9–28. online.

Biffignandi, S., & Bethlehem, J. G. (2011). Handbook of Web Surveys, 59–140.

Blair, E., & Zinkhan, G. M. (2006). Nonresponse and generalizability in academic research. *Journal of the Academy of Marketing Science*, *34*(1), 4–7. https://doi.org/10.1177/0092070305283778

Brick, J. M. (2013). Unit Nonresponse and Weighting Adjustments: A Critical Review. *Journal of Official Statistics*, *29*(3), 335–336. https://content.sciendo.com/view/journals/jos/29/3/article-p329.xml

Citro, C. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, *40*, 147–152.

Kalton, G., & Flores-Cervantes, I. (2003). Weighting Methods. *http://lst-iiep.iiep-unesco.org/cgi-bin/wwwi32.exe/[in=epidoc1.in]/?t2000=019622/(100)*, *19*, 81–97.

Kim, J., & Shao, J. (2013). Statistical methods for handling incomplete data. *Statistical Methods for Handling Incomplete Data*, 1–141. https://doi.org/10.1201/b13981

Lee, B.-J., & Marsh, L. (2000). Sample Selection Bias Correction for Missing Response Observations. *Oxford Bulletin of Economics and Statistics*, *62*, 305–22.

Sikov, A. (2018). A Brief Review of Approaches to Non-ignorable Non-response. *International Statistical Review*, *86*(3), https://onlinelibrary.wiley.com/doi/pdf/10.1111/insr.12264, 417–418. https://doi.org/10.1111/insr.12264

Sobczyk, M. (2007). Statystyka / (Wyd. 5 uzup.), 20–21.

Szymkowiak, M. (2009). Estymatory kalibracyjne w badaniu budżetów gospodarstw domowych, 3–19. online.

Szymkowiak, M. (2019). Podejście kalibracyjne w badaniach społeczno-ekonomicznych, 135–141.

Zhang, L.-C. (2001). A method of weighting adjustment for survey data subject to nonignorable nonresponse.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, *66*(1), 41–63. https://doi.org/10.1111/j.1467-9574.2011.00508.x

List of Tables

List of Figures