Robust imputation methods in the presence of influential units in surveys

Jia Ning Zhang

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science Mathematics and Statistics[1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

[1]The M.Sc. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Abstract

Item nonresponse is typically addressed by using single imputation techniques. When influential units are present in a sample, the classical imputed estimator of a population total is approximately unbiased provided that the first moment of the imputation model is correctly specified but may be very unstable. Thus, it is desirable to develop robust imputation methods that produce biased but more stable imputed estimators, that is an estimator whose mean square error is smaller than that of the corresponding non-robust imputed estimator. To achieve this, we propose using robust regression based on the Huber function with an adaptive tuning constant. In this thesis, we study three robust imputed estimators in the presence of influential units. We conduct a simulation study to compare the empirical performance of the proposed methods in terms of bias and relative efficiency for a wide range of distributions. Finally, we study the problem of mean square error estimation using both first-order Taylor procedures and bootstrap.

Résumé

La non-réponse partielle est souvent traitée au moyen de méthodes d'imputation simples. En présence d'unités influentes dans l'échantillon, l'estimateur imputé classique d'un total est approximativement sans biais si le premier moment du modèle d'imputation est correctement spécifié mais il peut être très instable. Il est donc désirable de développer des méthodes d'imputation robustes qui produisent des estimateurs imputés biaisés mais plus stables, c'est-à-dire des estimateurs dont l'erreur quadratique moyenne est inférieure à celle de l'estimateur imputé classique. Afin d'atteindre cet objectif, nous proposons d'utiliser une régression robuste basée sur la fonction de Huber avec une constante de réglage adaptatif. Dans cette thèse, nous étudions trois estimateurs imputés robustes en présence d'unités influentes. Nous menons une étude par simulation pour comparer les performances empiriques des méthodes proposées en termes de biais et d'efficacité relative pour une grande classe de distributions. Finalement, nous étudions le problème de l'estimation de l'erreur quadratique moyenne en utilisant à la fois les procédures de Taylor du premier ordre et le bootstrap.

Dedications

I dedicate this work to my parents.

Acknowledgement

I would like to thank my supervisor, Professor David Haziza, for his consistent support throughout my master's degree. I am grateful for his patience, guidance, invaluable assistance, and encouragement. His expertise in this field, along with his insightful comments, greatly contributed to the realization of this thesis. Additionally, I wish to acknowledge Professor Sixia Chen for providing valuable advice and insights during the development of this thesis.

I extend my heartfelt thanks to my parents, family, and friends for their continuous support and encouragement from the very beginning.

Thank you all for accompanying me on this journey. This thesis would not have been possible without each and every one of you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction to sampling and nonresponse

1.1 Survey sampling vs. classical statistics

In classical statistics, we are in the presence of a conceptual, infinite population. For instance, consider the following model:

$$y_i \overset{i.i.d}{\sim} N(\mu, \sigma^2), \quad i = 1, \ldots, n. \tag{1.1.1}$$

In (1.1.1), the values $y_1, \ldots, y_n$, are assumed to be generated independently from a normal distribution with mean μ and variance σ^2. In other words, in classical statistics, we assume that a random sample was generated from a given model, here the normal distribution. In the case of an infinite and conceptual population, it is impossible to determine the true values of the model parameters μ and σ^2 as we cannot obtain data on every population unit. Thus, our interest lies in making inferences about them. The main objective is to find the optimal (in a certain sense) estimator(s) of the parameter(s) of interest under the postulated model. For instance, we may want to determine the estimator that displays the smallest variance in the class of unbiased estimators. However, if the model is misspecified, the estimator may be biased and/or inefficient, which in turn, may lead to invalid inferences. For example, under the model (1.1.1), the optimal estimator of the population mean, μ, is the sample mean

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i. \tag{1.1.2}$$

However, if data were generated from a lognormal distribution, the estimator (1.1.2) may be very inefficient due to model misspecification. To determine an optimal estimator under the lognormal distribution, the Maximum Likelihood estimation (MLE)

approach is commonly used. Under some mild regularity conditions, MLE estimators have some desirable properties: (i) they are consistent; (ii) their asymptotic distribution is normal with mean at the true parameter value(s); and (iii) they are asymptotically efficient.

Survey sampling deals with a real and finite population of size N, which differs from classical statistics. In survey sampling, we are interested in estimating finite population parameters, which are those that describe some aspects of the finite population under study. Commonly encountered finite population parameters include population totals, population means, population proportions and population quantiles. If we conduct a census, it is possible to determine the true value of finite population parameters.

We distinguish between two types of survey: (i) censuses, whereby all the population units are surveyed and (ii) sample surveys, whereby a (usually small) fraction of the population is surveyed. Although sampling errors are not present in a census, non-sampling errors, such as nonresponse errors, coverage errors and measurement errors are almost always present in both a census and a sample survey. In fact, a census may result in larger non-sampling errors. The presence of non-sampling errors poses significant challenges in obtaining high-quality estimates, thereby affecting the accuracy and reliability of data analysis results. Reducing non-sampling errors is therefore important to ensure the quality of estimates. In this thesis, we focus on the nonresponse errors. Some examples of measures to reduce the nonresponse errors include offering incentives (rewards) to survey participants, providing training to interviewers and assigning them a reasonable workload, sending a notification letter in advance to participants, giving message reminders to them as needed, and ensuring a well-designed questionnaire. Those measures may be effective in enhancing the response rate. However, there will inevitably be some non-responding units, questionnaires partially answered by some sampling units, and the presence of inconsistent or invalid answers.

1.2 The setup for probability sampling

Let $U = \{1, \ldots, i, \ldots, N\}$ be a finite population of size N and S, a random sample, of size n, selected from U according to a probability sampling design. Since S is random, we use s to denote a realization of S. Let Ω denote the set of all possible samples that can be selected from U. A sampling design is a probability distribution on Ω. It is a function that assigns a probability of selection to each sample $s \in \Omega$ such that

i) $p(s) \geq 0$ for all $s \in \Omega$;

ii) $\sum_{s\in\Omega} p(s) = 1$,

where $p(s)$ denotes the probability of selecting a sample s, $s \in \Omega$. For simplicity, we use the notation $p(s)$ for $p(S = s)$.

Let I_i be a sample selection indicator attached to unit i, such that $I_i = 1$ if $i \in S$, and $I_i = 0$, otherwise. Each unit i, $i = 1, \ldots, N$, has a known probability of being included in the selected sample s, which is called the first-order inclusion probability. It is defined as

$$\pi_i = p\left(I_i = 1\right) = \sum_{\substack{s\in\Omega \\ s\ni i}} p(s).$$

The value of π_i, which is known prior to sampling, corresponds to the (weighted) proportion of samples that contain unit i, $i \in U$. Each sample unit i is assigned a sampling weight equal to $w_i = 1/\pi_i$. This weight can be interpreted as the number of units that unit i represents in the population.

The second-order inclusion probability for units i and j is defined as

$$\pi_{ij} = p\left(I_i = 1, I_j = 1\right) = \sum_{\substack{s\in\Omega \\ s\ni(i,j)}} p(s).$$

We can interpret π_{ij} as the (weighted) proportion of samples that include both units i and j. Furthermore, note that $\pi_{ij} = \pi_{ji}$ and $\pi_{ii} = \pi_i$.

1.3 Some basic sampling designs

In this section, we present some basic sampling designs: simple random sampling without replacement (SRSWOR) and Bernoulli sampling (BE).

1.3.1 Simple random sampling without replacement

In SRSWOR, any sample s that contains a set of n distinct units has an equal chance of being selected from the population. Since there are $\binom{N}{n}$ possible samples of size n, the probability of selecting any sample s, $s \in \Omega$, is equal to

$$p(s) = \frac{1}{\binom{N}{n}}.$$

The first-order inclusion probability for unit i and the second-order inclusion probability for units i and j are given by $\pi_i = n/N$ for all i and $\pi_{ij} = n(n-1)/N(N-1)$ for all (i, j), $i \neq j$, respectively. SRSWOR is a fixed-size sampling design since the sample size n is fixed before the selection of the sample.

1.3.2 Bernoulli sampling

Bernoulli consists in carrying out N independent Bernoulli trials with first-order inclusion probability $\pi_i = \pi \in (0,1)$. When trial i is a success, the population unit i is included in the sample, otherwise the unit is not included. For BE sampling, the second-order inclusion probability for units i and j is given by

$$\pi_{ij} = P\left(I_i = 1, I_j = 1\right) = P(I_i = 1)P(I_j = 1) = \pi_i\pi_j = \pi^2,$$

since the I_i's are mutually independent random variables. Therefore, the sample size n_s is random as it is not possible to predict the sample size with certainty. The number of possible samples is 2^N and the probability of drawing a sample s with size n_s is

$$p(s) = \pi^{n_s}\left(1-\pi\right)^{N-n_s}.$$

1.4 The Horvitz-Thompson estimator

Let y denote a survey variable (i.e., a variable that we collect for all units in the sample). The Horvitz-Thompson estimator of $t_y = \sum_{i\in U} y_i$, also known as the expansion estimator, Narain-Horvitz-Thompson estimator, or π estimator, was proposed by Narain (1951) and Horvitz and Thompson (1952). It is given by

$$\widehat{t}_{y,\pi} = \sum_{i\in S}\frac{y_i}{\pi_i} = \sum_{i\in S} w_i y_i = \sum_{i\in U} w_i y_i I_i. \tag{1.4.1}$$

Below, we will present some properties of the Horvitz-Thompson estimator. We use the notation $\mathbb{E}_p\left(\cdot\right)$, $\mathbb{V}_p\left(\cdot\right)$ and $Cov_p\left(\cdot\right)$ to denote the expectation, variance and covariance with respect to the sampling design. In this case, all quantities in (1.4.1), but the sample selection indicators, $I_1, \ldots, I_N$, are treated as fixed.

Proposition 1.1. *If $\pi_i > 0$ for all $i \in U$, the Horvitz-Thompson estimator (1.4.1) is design-unbiased for t_y.*
Proof:

$$\begin{aligned}
\mathbb{E}_p\left(\widehat{t}_{y,\pi}\right) &= \mathbb{E}_p\left(\sum_{i\in S}\frac{y_i}{\pi_i}\right) \\
&= \mathbb{E}_p\left(\sum_{i\in U}\frac{y_i I_i}{\pi_i}\right) \\
&= \sum_{i\in U}\frac{y_i}{\pi_i}\mathbb{E}_p\left(I_i\right) \\
&= \sum_{i\in U}\frac{y_i\pi_i}{\pi_i}
\end{aligned}$$

$$= \sum_{i \in U} y_i$$
$$= t_y.$$

∎

The Horvitz-Thompson estimator is unbiased for t_y for any sampling design satisfying $\pi_i > 0$ for all $i \in U$, and any survey variable y.

Proposition 1.2. *Provided that $\pi_i > 0$ for all $i \in U$, the design variance of $\widehat{t}_{y,\pi}$ for either a fixed-size or a random-size sampling design, is given by*

$$\mathbb{V}_p\left(\widehat{t}_{y,\pi}\right) = \sum_{i \in U} \sum_{j \in U} \left(\pi_{ij} - \pi_i \pi_j\right) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}. \tag{1.4.2}$$

Proof:

$$\begin{aligned}
\mathbb{V}_p\left(\widehat{t}_{y,\pi}\right) &= \mathbb{V}_p\left(\sum_{i \in S} \frac{y_i}{\pi_i}\right) \\
&= \mathbb{V}_p\left(\sum_{i \in U} \frac{y_i I_i}{\pi_i}\right) \\
&= \sum_{i \in U} \left(\frac{y_i}{\pi_i}\right)^2 \mathbb{V}_p\left(I_i\right) + \sum_{\substack{i \in U \\ i \neq j}} \sum_{j \in U} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} Cov_p\left(I_i, I_j\right) \\
&= \sum_{i \in U} \frac{y_i^2}{\pi_i^2} \pi_i \left(1 - \pi_i\right) + \sum_{\substack{i \in U \\ i \neq j}} \sum_{j \in U} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \left(\pi_{ij} - \pi_i \pi_j\right) \\
&= \sum_{i \in U} \sum_{j \in U} \left(\pi_{ij} - \pi_i \pi_j\right) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}.
\end{aligned}$$

The fifth equality follows from the fact that

$$Cov_p\left(I_i, I_j\right) = E_p\left(I_i I_j\right) - E_p\left(I_i\right) E_p\left(I_j\right) = \pi_{ij} - \pi_i \pi_j.$$

∎

Expression (1.4.2) is a measure of the volatility of $\widehat{t}_{y,\pi}$ that we would observe if we were able to select all the possible samples from the population. The variance in (1.4.2) is unknown since we can only observe the y-values for the sample units. Therefore, we need to estimate the variance.

Proposition 1.3. *Provided that $\pi_{ij} > 0$ for all $(i,j) \in U \times U$, a design-unbiased estimator of $\mathbb{V}_p\left(\widehat{t}_{y,\pi}\right)$ is given by*

$$\widehat{V}_{HT}\left(\widehat{t}_{y,\pi}\right) = \sum_{i\in S}\sum_{j\in S}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}. \tag{1.4.3}$$

Proof:

$$\begin{aligned}
\mathbb{E}_p\left\{\widehat{V}_{HT}(\widehat{t}_{y,\pi})\right\} &= \mathbb{E}_p\left\{\sum_{i\in S}\sum_{j\in S}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\right\} \\
&= \mathbb{E}_p\left\{\sum_{i\in U}\sum_{j\in U}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\frac{y_iI_i}{\pi_i}\frac{y_jI_j}{\pi_j}\right\} \\
&= \sum_{i\in U}\sum_{j\in U}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\mathbb{E}_p(I_iI_j) \\
&= \sum_{i\in U}\sum_{j\in U}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\pi_{ij} \\
&= \sum_{i\in U}\sum_{j\in U}(\pi_{ij}-\pi_i\pi_j)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j} \\
&= \mathbb{V}_p\left(\widehat{t}_{y,\pi}\right).
\end{aligned}$$

∎

In the case of SRSWOR, Expression (1.4.3) reduces to the following variance estimator

$$\widehat{V}_{HT}\left(\widehat{t}_{y,\pi}\right) = N^2\left(1-\frac{n}{N}\right)\frac{s_y^2}{n},$$

where $s_y^2 = (n-1)^{-1}\sum_{i\in S}(y_i-\bar{y}_s)^2$, with $\bar{y}_s = n^{-1}\sum_{i\in S}y_i$. For Bernoulli sampling, Expression (1.4.3) reduces to

$$\widehat{V}_{HT}\left(\widehat{t}_{y,\pi}\right) = \sum_{i\in S}\frac{(1-\pi)}{\pi^2}y_i^2,$$

noting that $\pi_{ij}-\pi_i\pi_j = 0$ for $i \neq j$.

1.5 Nonresponse and imputation methods

We distinguish between two types of nonresponse: (i) Unit nonresponse and (ii) item nonresponse. In the case of unit nonresponse, no usable information is collected on a sample unit. Some common reasons for unit nonresponse include the inability to establish contact with the sampled unit or the refusal of the sampled unit to participate in the survey. To deal with unit nonresponse, weight adjustment procedures methods are usually employed. It consists of removing non-responding units from the data file and increasing the sampling weight of the responding units in order to compensate for non-responding units that were eliminated. Item nonresponse occurs when some survey variables, but not all, have a missing value. This situation can arise when a sample unit responds to a subset of the questions, rather than completing the entire survey questionnaire (especially when there are sensitive questions such as income), or when the responses to certain questions on the questionnaire are inconsistent or invalid. Item nonresponse is usually treated using some form of single imputation, which consists of constructing a single replacement value for each of the missing values, which leads to the creation of a complete data file.

Table 1.1 below shows a typical data file after data collection. We have n sampled units and p survey variables, y_1, y_2, ..., y_p.

	y_1	y_2	y_3	...	y_p	
1	✓	✓	✓	...	✓	Full response
2	✓	✓	✓	...	✓	
⋮	✓	×	×	...	✓	Item nonresponse
⋮	×	✓	✓	...	×	
⋮	×	×	×	...	×	Unit nonresponse
n	×	×	×	...	×	

Table 1.1: Levels of nonresponse

From now on, we use y_i to denote the survey variable y for the ith unit, $i = 1, \ldots, N$. The main issue with nonresponse is the potential for nonresponse bias, which may be significant if the characteristics of respondents differ from those of the nonrespondents. For instance, let y be the income of individuals. If the nonrespondents have, on average, a lower income compared to the respondents, estimates based on the responding units only, called complete case estimates, may result in an overestimation of the average income in the population. Complete case estimators are thus subject to nonresponse bias, which would be especially appreciable if the nonresponse

rate is high. Moreover, due to the presence of non-responding units in the sample, the effective sample size is smaller than n. Hence, the variance of the estimators will suffer from an additional source of variance, called the nonresponse variance. The main purpose of any nonresponse treatment method is to reduce nonresponse bias as much as possible and, if possible, control the variance due to nonresponse. Achieving an efficient bias reduction relies on the availability of fully observed variables, which are those observed for both the respondents and nonrespondents.

1.5.1 Nonresponse mechanism

Consider a variable of interest, y, which is subject to missingness. Let r_i be the response indicator attached to unit i such that $r_i = 1$ if y_i is observed, and $r_i = 0$ if y_i is missing. Let $S_r = \{i \in S; r_i = 1\}$ be the set of respondents to item y and let $S_m = \{i \in S; r_i = 0\}$ be the set of nonrespondents. Let $\mathbf{R} = (r_1, \ldots, r_N)^\top$ be the vector of response indicators and $\mathbf{y} = (y_1, \ldots, y_N)^\top$ be the vector of the population y-values. The nonresponse mechanism is defined as the distribution $\mathcal{F}(\mathbf{R}|\mathbf{y}, \boldsymbol{V})$, where $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N)^\top$ denotes the matrix of fully observed auxiliary variables $\boldsymbol{v}_i$, $i = 1, \ldots, N$, as rows.

The first moment of $\mathcal{F}(\mathbf{R}|\mathbf{y}, \boldsymbol{V})$ is $\mathbb{E}\{\mathbf{R}|\mathbf{y}, \boldsymbol{V}\} = (p_1, \ldots, p_i, \ldots, p_N)^\top$, where the response probability $p_i = P(r_i = 1|y_i, \boldsymbol{v}_i)$. Since the nonresponse mechanism is unknown, it is necessary to make some assumptions: (i) We assume that the response indicators r_i are mutually independent random variables. (ii) We assume that $0 < p_i \leqslant 1$ for all i. This is referred to as the positivity assumption.

Missing data mechanisms are categorized into three types: *Missing Completely At Random* (MCAR), *Missing At Random* (MAR) and *Not Missing At Random* (NMAR).

For *Missing completely at random* (MCAR) nonresponse mechanism, the set of respondents S_r can be considered as a random sample selected from the population. In other words, there are no systematic differences between respondents and nonrespondents. As a result, the mean of the respondents is approximately unbiased for the population mean. An example of MCAR occurs when $p_i = p_0$ for all $i \in U$. This is called a uniform response mechanism. Another example of MCAR is when the response probability p_i depends on a fully observed auxiliary variable v_i, but y and v_i are not related. Since the assumption of MCAR is often considered too strong and unrealistic, a weaker condition, known as *Missing at Random* (MAR), is usually assumed in practice. Note that MCAR is a special case of MAR.

Under the MAR assumption (Rubin, 1976), we have

$$p_i = p(r_i = 1|y_i, \widetilde{\boldsymbol{v}}_i) = p(r_i = 1|\widetilde{\boldsymbol{v}}_i), \tag{1.5.1}$$

where $\widetilde{\boldsymbol{v}}$ is a vector that consists of the components of $\boldsymbol{v}$ that are related to both the y-variable and the probability of response. This assumption implies that the distribution of the variable of interest y among respondents is the same as the distribution of y among nonrespondents after accounting for the fully observed variables $\widetilde{\boldsymbol{v}}$. That is,

$$f\left(y|\widetilde{\boldsymbol{v}}_i, r_i = 1\right) = f\left(y|\widetilde{\boldsymbol{v}}_i, r_i = 0\right).$$

In this case, the imputed values can be generated from $f\left(y|\widetilde{\boldsymbol{v}}_i, r_i = 1\right)$, which can be estimated from the responding units.

Finally, for *Not Missing at Random* (NMAR) nonresponse mechanism, the distribution of the variable of interest y among respondents is not the same as the distribution of y among nonrespondents even after conditioning on $\widetilde{\boldsymbol{v}}$. Therefore,

$$f\left(y|\widetilde{\boldsymbol{v}}_i, r_i = 1\right) \neq f\left(y|\widetilde{\boldsymbol{v}}_i, r_i = 0\right).$$

Under NMAR assumption, the survey variable y that is subject to nonresponse causes nonresponse. It is not possible to refute or support this nonresponse mechanism with the data at hand. In addition, NMAR may also occur when there are other variables that are related to both the response probability and the survey variable y but not used in the nonresponse treatment process.

We conducted a simulation study to illustrate the differences between three nonresponse mechanisms: MCAR, MAR and NMAR. To that end, we generated a finite population U of size $N = 50,000$. Let y be a survey variable that corresponds to the annual salary (in thousand dollars) and v_1 and v_2 be two categorical variables that correspond to the post-secondary degree and the gender, respectively. The variable v_1 was equal to 1 if the population unit had a post-secondary degree and was equal to 0, otherwise. The variable v_2 was equal to 1 if the population unit was a male and was equal to 0 if it was a female. The survey variable y was generated according to the model

$$y_i = 50 + 13v_{1i} + 10v_{2i} + \epsilon_i,$$

where $\epsilon_i \sim N\left(0, 16\right)$, $i = 1, \ldots, N$. In the population, note that the average salary for an individual with a post-secondary degree was \$13,000 higher than an individual who did not have one. Also, the average salary was \$10,000 higher for males. The overall average salary in the population was around \$61,500.

A sample of size $n = 4,000$ was selected from U according to simple random sampling without replacement. In the sample, we generated nonresponse according to three nonresponse mechanisms:

i) *Missing Completely at Random* (MCAR): we used a uniform nonresponse mechanism with $p_i = 50\%$ for all $i \in U$;

ii) *Missing at Random* (MAR): we set $p_i = 0.30 + 0.20v_{1i} + 0.25v_{2i}$. In the population, the overall response rate was around 52.5%. For units in different categories, the response rate was different. The response rate was 75% for males with a post-secondary degree, 55% for males without a post-secondary degree, 50% for females with a post-secondary degree, and 30% for females without a post-secondary degree.

iii) *Not Missing At Random* (NMAR): we generated nonresponse using a logistic function, such that the response rate was a function of the annual salary:

$$p_i = \frac{\exp(-18.5 + 0.3y_i)}{1 + \exp(-18.5 + 0.3y_i)}.$$

The overall response rate in the population was around 50%. The response rate was approximately 94% for males with a post-secondary degree, 40% for males without a post-secondary degree, 58% for females with a post-secondary degree, and 5% for females without a post-secondary degree.

For MCAR, the distribution of salary among respondents and nonrespondents are approximately the same; see Figure (1.1).

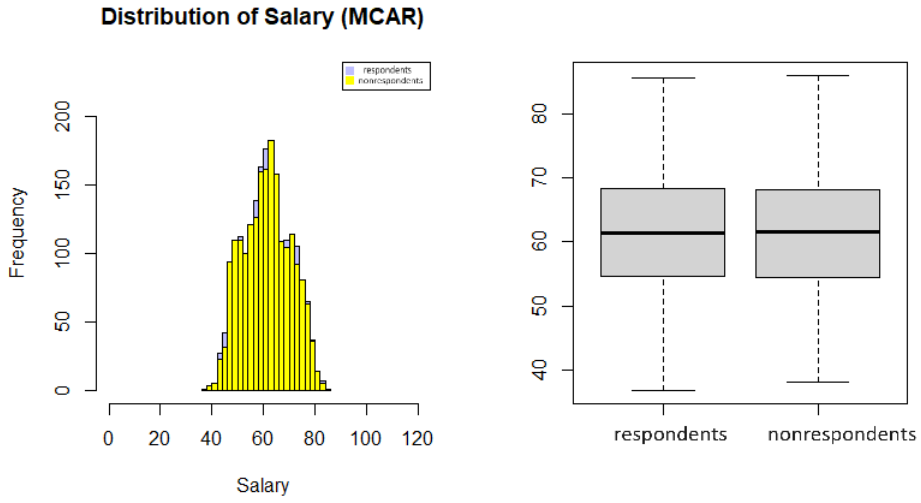


Figure 1.1: Distribution of respondents and nonrespondents under MCAR

For MAR, the distribution of salary is not the same for responding and non-responding units in the selected sample. The average salary of responding units is clearly higher than that of the non-responding units as we can observed (see Figure 1.2).

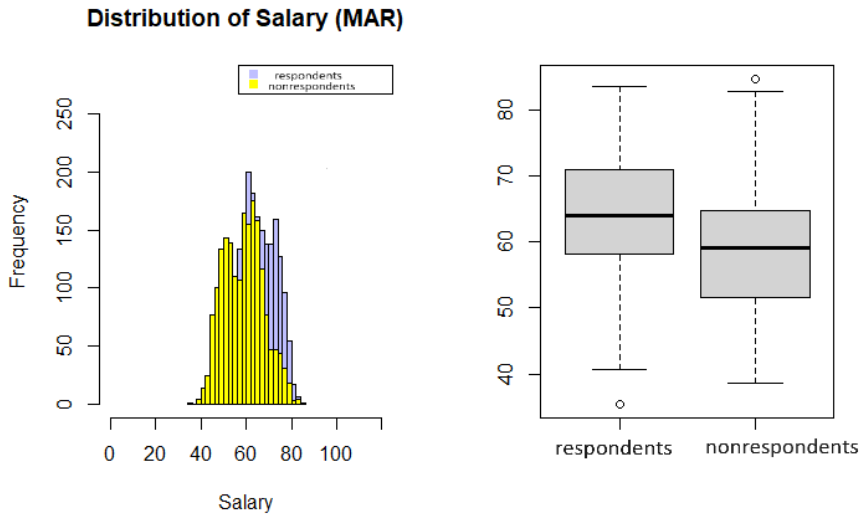
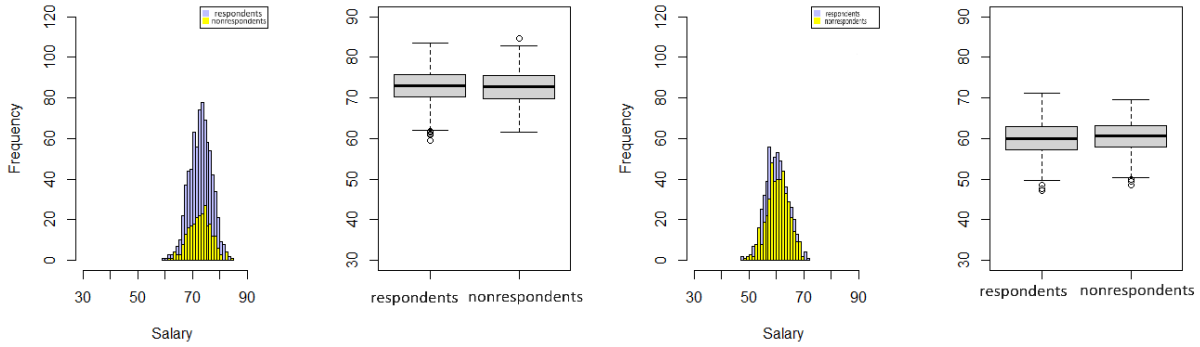


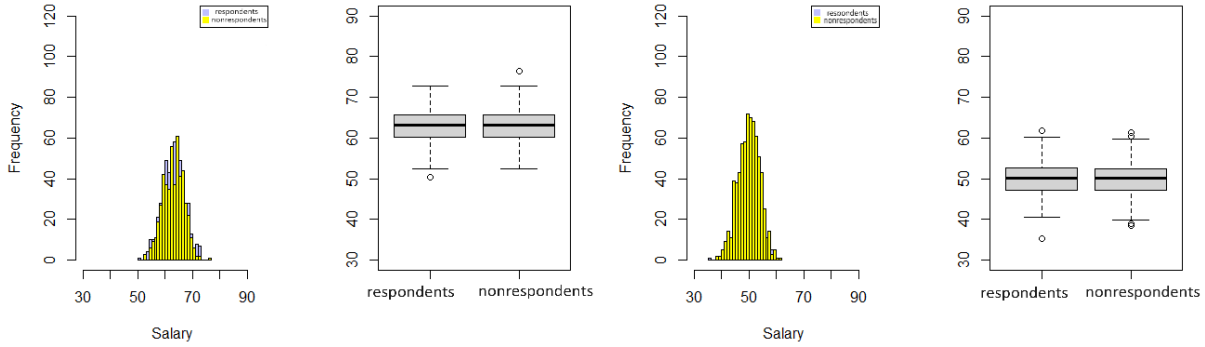
Figure 1.2: Distribution of respondents and nonrespondents under MAR

Since v_1 (post-secondary degree) and v_2 (gender) are binary variable, we obtain four cells: post-secondary degree and male ($v_1 = 1$, $v_2 = 1$), post-secondary degree and female ($v_1 = 1$, $v_2 = 0$), no post-secondary degree and male ($v_1 = 0$, $v_2 = 1$), and finally no post-secondary degree and female ($v_1 = 0$, $v_2 = 0$). However, after conditioning on appropriate set auxiliary variables, i.e., v_1 and v_2, we can observe that the distribution of the y-variable is approximately the same among respondents and nonrespondents.



(a) Distribution of y in cell $v_1 = 1$, $v_2 = 1$

(b) Distribution of y in cell $v_1 = 0$, $v_2 = 1$



(c) Distribution of y in cell $v_1 = 1$, $v_2 = 0$

(d) Distribution of y in cell $v_1 = 0$, $v_2 = 0$

Figure 1.3: Distribution of respondents and nonrespondents under MAR after conditioning on v_1 and v_2

For NMAR, the distribution of salary is not the same for responding and nonresponding units in the selected sample (see Figure 1.4).

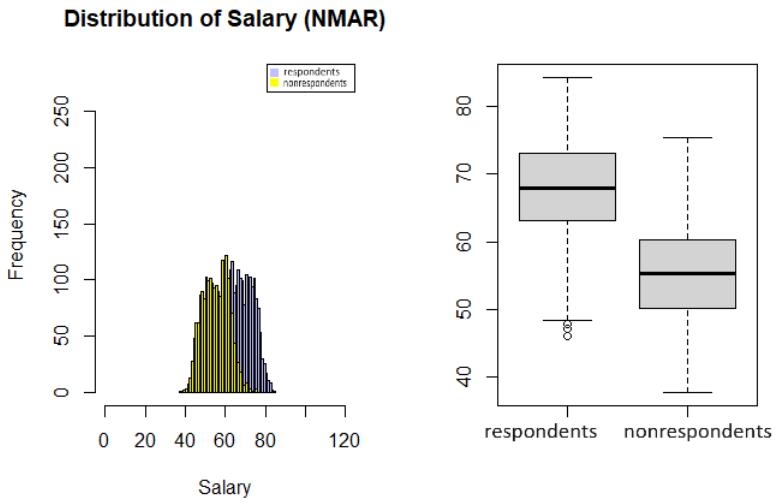
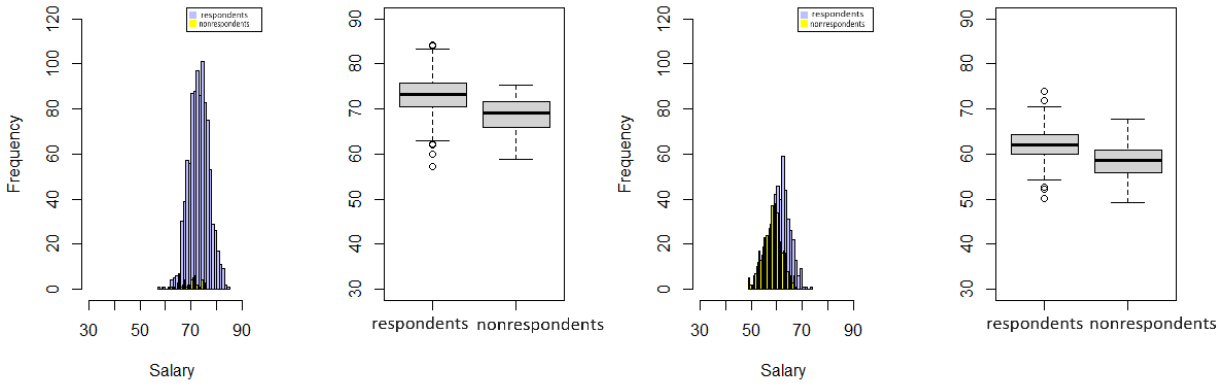


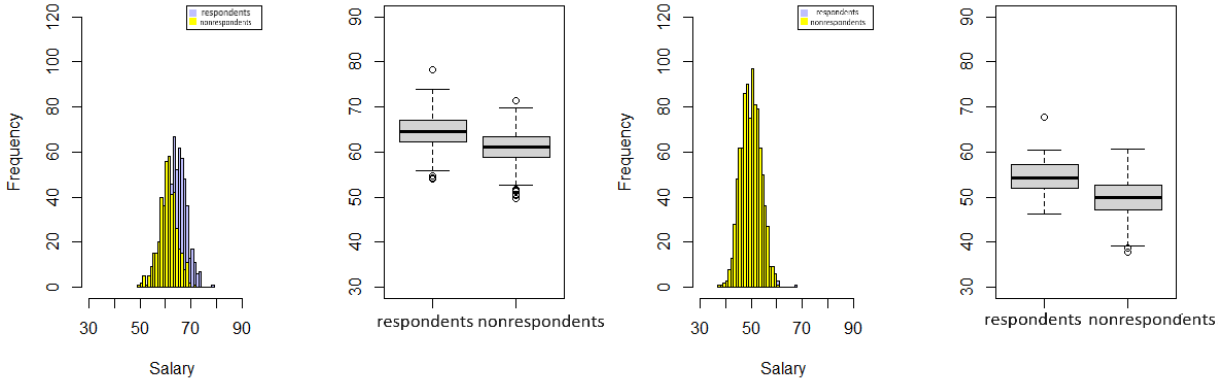
Figure 1.4: Distribution of respondents and nonrespondents under NMAR

Under NMAR, the survey variable y that is subject to nonrepsonse causes nonresponse. We observe that under this nonresponse mechanism the distribution of salary is not the same among respondents and nonrespondents even after conditioning on v_1 and v_2 (see Figure 1.5).



(a) Distribution of y in cell $v_1 = 1$, $v_2 = 1$

(b) Distribution of y in cell $v_1 = 0$, $v_2 = 1$



(c) Distribution of y in cell $v_1 = 1$, $v_2 = 0$

(d) Distribution of y in cell $v_1 = 0$, $v_2 = 0$

Figure 1.5: Distribution of respondents and nonrespondents under NMAR after conditioning on v_1 and v_2

1.5.2 Imputation methods

In the presence of missing values to item y, an imputed estimator of the population total $t_y = \sum_{i \in U} y_i$, is defined as

$$\widehat{t}_I = \sum_{i \in S} w_i r_i y_i + \sum_{i \in S} w_i (1 - r_i) y_i^*, \tag{1.5.2}$$

where y_i^* denotes the imputed value for the missing y_i. The estimator $\widehat{t}_I$ can be written as

$$\widehat{t}_I = \sum_{i \in S} w_i \widetilde{y}_i,$$

where $\widetilde{y}_i = r_i y_i + (1 - r_i) y_i^*$.

How to obtain the imputed values y_i^*? We introduce several imputation methods that are commonly employed in practice.

Imputation methods can be classified into two groups: deterministic imputation procedures and random imputation procedures. For deterministic imputation procedures, if different people perform the same imputation procedure on the same data set, they would obtain the same completed data file and, thus, the same estimate. In contrast, random imputation methods would lead to different completed data sets because of added random noise. Alternatively, we may also distinguish donor imputation procedures from predicted value imputation procedures. In the case of donor imputation procedures, a missing value associated with a recipient is substituted with the value of a donor. Therefore, the missing values are replaced by actual values observed among the respondents. In contrast, for predicted value imputation procedures, a missing value is replaced by a predicted value, not an actual value from the set of respondents.

1.5.2.1 Semi-parametric imputation

Many imputation methods can be motivated by the following general model, also known as the imputation model or outcome regression model:

$$y_i = m\left(\boldsymbol{v}_i; \boldsymbol{\beta}\right) + \epsilon_i, \tag{1.5.3}$$

such that

$$\mathbb{E}_m\left(\epsilon_i | \boldsymbol{v}_i\right) = 0, \mathbb{E}_m\left(\epsilon_i \epsilon_j | \boldsymbol{v}_i, \boldsymbol{v}_j\right) = 0, i \neq j, \text{ and } \mathbb{V}_m\left(\epsilon_i | \boldsymbol{v}_i\right) = \sigma^2 c_i,$$

where $c_i > 0$ is a known factor attached to unit i. In (1.5.3), the function $m(.; \boldsymbol{\beta})$ is a prespecified function, $\boldsymbol{v}_i$ is a vector of fully observed variables associated with unit i and ϵ_i is a random error associated with unit i. The model (1.5.3) is not fully parametric since we do not make any distributional assumptions about the distribution of the $\epsilon_i's$. For this reason, Model (1.5.3) is referred to as semi-parametric.

Deterministic semi-parametric imputation procedures use the following imputed values y_i^*, $i \in S_m$:

$$y_i^* = m\left(\boldsymbol{v}_i; \widehat{\boldsymbol{\beta}}\right), \tag{1.5.4}$$

where $\widehat{\boldsymbol{\beta}}$ is a suitable estimator of $\boldsymbol{\beta}$ based on the responding units; e.g., least squares estimator, maximum likelihood estimator, etc.

Deterministic linear regression imputation is a special case of (1.5.4). In this case,

$$y_i^* = m\left(\boldsymbol{v}_i; \widehat{\boldsymbol{B}}_{WLS}\right) = \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{WLS}, \tag{1.5.5}$$

where

$$\widehat{\boldsymbol{B}}_{WLS} = \left(\sum_{i \in S_r} \phi_i \boldsymbol{v}_i c_i^{-1} \boldsymbol{v}_i^\top\right)^{-1} \sum_{i \in S_r} \phi_i \boldsymbol{v}_i c_i^{-1} y_i, \tag{1.5.6}$$

where $\phi_i = 1$ for unweighted imputation and equal to w_i for survey weighted imputation.

In the case of random imputation procedures, Expression (1.5.4) is modified as follows:

$$y_i^* = m\left(\boldsymbol{v}_i; \widehat{\boldsymbol{\beta}}\right) + \widehat{\sigma}\sqrt{c_i} e_i^*, \tag{1.5.7}$$

with $\mathbb{E}_I(e_i^*) = 0$. The expectation $\mathbb{E}_I(\cdot)$ refers to the expectation with respect to the imputation mechanism used to generate the random residuals e_i^*. If the distribution of the errors ϵ_i in the model (1.5.3) is assumed to be normally distributed, then the $e_i^{*\prime}s$ can simply be generated from a normal distribution with mean 0 and variance equal to $\widehat{\sigma}^2 c_i$, where $\widehat{\sigma}^2$ is an estimator of σ. In practice, the $e_i^{*\prime}s$ are usually generated in a nonparametric way as the normal assumption for the error term may not always be valid. Next, we present some special cases of deterministic and random imputation procedures.

Simple linear regression imputation

Simple linear regression imputation is a special case of deterministic linear regression imputation in (1.5.5), whereby $\boldsymbol{v}_i = (1, v_i)^\top$ and $\widehat{\boldsymbol{B}}_{WLS} = (\widehat{B}_{0,WLS}, \widehat{B}_{1,WLS})^\top$:

$$y_i^* = \widehat{B}_{0,WLS} + \widehat{B}_{1,WLS} v_i, \quad i \in S_m, \tag{1.5.8}$$

where

$$\widehat{B}_{0,WLS} = \bar{y}_r - \widehat{B}_{1,WLS} \bar{v}_r,$$

and

$$\widehat{B}_{1,WLS} = \frac{\sum_{i \in S_r} \phi_i (v_i - \bar{v}_r)(y_i - \bar{y}_r)}{\sum_{i \in S_r} \phi_i (v_i - \bar{v}_r)^2},$$

with $\bar{y}_r = \frac{\sum_{i \in S_r} \phi_i y_i}{\sum_{i \in S_r} \phi_i}$ and $\bar{v}_r = \frac{\sum_{i \in S_r} \phi_i v_i}{\sum_{i \in S_r} \phi_i}$.

Mean imputation

Mean imputation is a special case of deterministic linear regression imputation with $\boldsymbol{v}_i = 1$ and $c_i = 1$ for all i. This is a poor model since it is essentially a regression model that contains only the intercept. Setting $\boldsymbol{v}_i = 1$ and $c_i = 1$ in (1.5.5) leads to

$$y_i^* = \overline{y}_r, \;\; i \in S_m. \tag{1.5.9}$$

Ratio imputation

Ratio imputation is an appropriate method to use when there is a linear relationship between the y-variable and v, with the relationship passing through the origin. It is a special case of deterministic linear regression imputation with $\boldsymbol{v}_i = v_i$ and $c_i = v_i$. This leads to

$$\widehat{B}_{WLS} = \frac{\bar{y}_r}{\bar{v}_r}, \tag{1.5.10}$$

and

$$y_i^* = \widehat{B}_{WLS} v_i = \frac{\bar{y}_r}{\bar{v}_r} v_i, \;\; i \in S_m. \tag{1.5.11}$$

Historical imputation

Historical imputation is an imputation procedure that consists of using the value observed in the previous wave of the survey to impute for each missing y value. Let v_i be the y-variable observed at a previous time. The imputed values under historical imputation are given by

$$y_i^* = v_i, \;\; i \in S_m. \tag{1.5.12}$$

It is a special case of deterministic linear regression imputation with $\boldsymbol{v}_i = v_i$ and $\widehat{\boldsymbol{B}}_{WLS} = 1$.

1.5.2.2 A nonparametric procedure: Nearest-neighbour imputation

Nearest-neighbour imputation is a nonparametric imputation method in the class of deterministic imputation procedures. We assume the following model:

$$y_i = m\left(\boldsymbol{v}_i\right) + \epsilon_i, \tag{1.5.13}$$

where $m\left(\boldsymbol{v}_i\right)$ is an unspecified function and the variance structure $V_m\left(\epsilon_i | \boldsymbol{v}_i\right) = \sigma^2 c_i$ is also left unspecified. The imputed values are given by

$$y_i^* = y_j, \quad i \in S_m, \tag{1.5.14}$$

where the index j is such that

$$D(\boldsymbol{v}_j, \boldsymbol{v}_i) \leq D(\boldsymbol{v}_k, \boldsymbol{v}_i), \quad \text{for all } k \in S_r.$$

Therefore, the index $j \in S_r$ is selected so that it minimizes $D(\boldsymbol{v}_j, \boldsymbol{v}_i)$, where $D(\cdot, \cdot)$ denotes a distance function and $\boldsymbol{v}$ is a vector of auxiliary variables of dimension G.

A general distance measure is given by

$$D(\boldsymbol{v}_j, \boldsymbol{v}_i) = \left(\sum_{\ell=1}^{G} \alpha_\ell \left| v_{\ell j} - v_{\ell i} \right|^b \right)^{1/b}, \tag{1.5.15}$$

where $b \geq 1$ and α_ℓ, $\ell = 1, \ldots, G$, is a weight assigned by the imputer to the auxiliary variables v_ℓ, $\ell = 1, \ldots, G$, which are also referred to as matching variables. When $b = 1$, we obtain the L_1-norm, whereas $b = 2$ corresponds to the L_2-norm (i.e., the euclidean distance).

Furthermore, the nearest-neighbour imputation is a donor imputation procedure, which is especially useful for imputing categorical variables. In addition, it tends to preserve the distribution function of the survey variables being imputed, which is a desirable feature.

However, this imputation procedure suffers from the curse of dimensionality. As the number of matching variables increases, the bias of $\widehat{t}_I$ tends to increase. We conducted a simulation study to illustrate this issue.

We generated a population U of size $N = 5000$. The survey variable y was generated according to 20 models of the form:

$$y_i = 2 + \sum_{g=1}^{G} v_{gi} + \epsilon_i, \tag{1.5.16}$$

where $\epsilon_i \sim N(0, 0.25)$, $i = 1, \ldots, N$. The first model (M1) included the first explanatory variable v_1 only, the second model (M2) included the first two explanatory variable v_1 and v_2, and so on. The matching variables $v_1, \ldots, v_G$ were generated from three different distributions. We have $v_g \sim Gamma(2, 10)$ for $g = 1, 2, 3, 10, 11, 12, 16, 17, 18$, $v_g \sim Unif(0, 1)$ for $g = 4, 5, 6, 13, 14, 15$ and $v_g \sim Chisq(1)$ for $g = 7, 8, 9, 19, 20$. In our simulation, we used the euclidean distance, and for simplicity, we set $\alpha_\ell = 1$, $\ell = 1, \ldots, G$, in (1.5.15).

Nonresponse to the survey variable y was generated according to 20 MAR nonresponse mechanism with probability

$$p_i = \frac{\exp(-2.2 + \mathbf{c}^\top \boldsymbol{v}_i)}{1 + \exp(-2.2 + \mathbf{c}^\top \boldsymbol{v}_i)},$$

where $\boldsymbol{v} = (v_1, v_2, \ldots, v_G)$ is a vector with G matching variables and $\mathbf{c}$ is a vector of constants chosen such that the overall response probability was approximately equal to 50%. The first nonresponse mechanism (NR1) included the first explanatory variable v_1 only, the second nonresponse mechanism (NR2) included the first two explanatory variable v_1 and v_2, and so on.

We repeated $R = 1,000$ iterations of the following process:

i) From the population, select a sample of size $n = 200$ according to simple random sampling without replacement;

ii) Compute the full sample estimator of population total $\widehat{t}_{y,\pi}$;

iii) Compute the imputed estimator of population total $\widehat{t}_I$ using nearest-neighbour imputation based on the euclidean distance for each pair of the imputation/nonresponse mechanism (Mg)/(NRg) for $g = 1, \ldots, 20$.

We computed the Monte Carlo percent relative bias of the imputed estimator $\widehat{t}_I$:

$$RB_{MC}\left(\widehat{t}_I\right) = \frac{\mathbb{E}_{MC}\left(\widehat{t}_I\right) - t_y}{t_y} \times 100, \tag{1.5.17}$$

where $\mathbb{E}_{MC}(\widehat{t}_I) = \frac{1}{R}\sum_{r=1}^{R} \widehat{t}_I^{(r)}$, and $\widehat{t}_I^{(r)}$ denotes the value of $\widehat{t}_I$ obtained in the rth sample, $r = 1, \ldots, R$. We also computed the Monte Carlo relative efficiency

$$RE\left(\widehat{t}_I\right) = \frac{MSE_{MC}\left(\widehat{t}_I\right)}{MSE_{MC}\left(\widehat{t}_{y,\pi}\right)} \times 100, \tag{1.5.18}$$

where $MSE_{MC}(\widehat{t}_I) = \frac{1}{R}\sum_{r=1}^{R}\left(\widehat{t}_I^{(r)} - t_y\right)^2$.

Table (1.2) shows the Monte Carlo percent relative bias (RB) and Monte Carlo relative efficiency (RE) of $\widehat{t}_I$ for different values of G.

G	RB	RE
1	0.0	359
2	0.1	344
5	0.5	218
8	3.2	320
10	4.3	468
15	4.4	660
20	4.9	755

Table 1.2: Nearest-neighbour imputation based on different number of matching variables

The results in Table (1.2) suggest that the bias of $\widehat{t}_I$ increases as the number of matching variables increases. Indeed, the bias was negligible for $G \leq 5$ but was no longer negligible for $G \geq 8$. As a result, the value increased as G increased. As a result, the efficiency of $\widehat{t}_I$ deteriorated as G increased. This is an illustration of the curse of dimensionality in the case of nearest-neighbour imputation.

1.5.2.3 Random regression imputation

Random regression imputation can be viewed as deterministic regression imputation with an added random residual. The imputed value y_i^* is given by

$$y_i^* = \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{WLS} + \widehat{\sigma}\sqrt{c_i}e_i^*, \tag{1.5.19}$$

where $\widehat{\boldsymbol{B}}_{WLS}$ is given by the expression (1.5.6) and e_i^* is generated by randomly selecting, with replacement, a value from the set of standardized residuals $\{e_j; j \in S_r\}$ such that $P\left(e_i^* = e_j\right) = \phi_j \left(\sum_{l \in S_r} \phi_l\right)^{-1}$, $j \in S_r$, where $e_j = \widetilde{e}_j - \sum_{l \in S_r} \phi_l \widetilde{e}_l \left(\sum_{l \in S_r} \phi_l\right)^{-1}$ with $\widetilde{e}_j = \left(\widehat{\sigma}\sqrt{c_j}\right)^{-1} \left\{y_j - \boldsymbol{v}_j^\top \widehat{\boldsymbol{B}}_{WLS}\right\}$.

Random hot-deck imputation

A special case of random regression imputation is random hot-deck imputation. It is a donor imputation method, whereby the imputed values y_i^*, $i \in S_m$, are selected at random from the set of responding units (the donors). Random hot-deck imputation can be viewed as mean imputation plus an added random residual. In other words, mean imputation is the deterministic counterpart of random hot-deck imputation.

Recall that in the case of mean imputation, we have the following imputation model $y_i = \beta + \varepsilon_i$ and we have found that $\widehat{\beta} = \bar{y}_r$. Therefore, the standardized

residuals are given by $e_i = y_i - \bar{y}_r$, $i \in S_r$. Thus, the imputed values y_i^*, $i \in S_m$, are given by

$$\begin{aligned} y_i^* &= \overline{y}_r + e_i^* \\ &= \overline{y}_r + (y_j - \overline{y}_r) \\ &= y_j, \end{aligned}$$

where e_i^* is a residual selected at random from the set $\{e_j, j \in S_r\}$ with probability $\frac{\phi_j}{\sum_{l \in S_r} \phi_l}$.

1.5.3 Properties of imputed estimators

In this section, we examine the properties of the imputed estimator $\widehat{t}_I$.

1.5.3.1 Decomposition of total error

The total error of $\widehat{t}_I$ in the case of deterministic imputation procedures can be expressed as

$$\widehat{t}_I - t_y = \underbrace{\left(\widehat{t}_{y,\pi} - t_y\right)}_{\text{Sampling error}} + \underbrace{\left(\widehat{t}_I - \widehat{t}_{y,\pi}\right)}_{\text{Nonresponse error}}, \tag{1.5.20}$$

where t_y denotes the true population total and $\widehat{t}_{y,\pi}$ (1.4.1) is the full sample estimator, i.e., the estimator that would have been used in the absence of nonresponse. The Horvitz-Thompson estimator $\widehat{t}_{y,\pi}$ has the property of being design-unbiased (see Section 1.4).

In the case of random imputation, the total error of $\widehat{t}_I$ can be expressed as

$$\widehat{t}_I - t_y = \underbrace{\left(\widehat{t}_{y,\pi} - t_y\right)}_{\text{Sampling error}} + \underbrace{\left(\breve{t}_I - \widehat{t}_{y,\pi}\right)}_{\text{Nonresponse error}} + \underbrace{\left(\widehat{t}_I - \breve{t}_I\right)}_{\text{Imputation error}}, \tag{1.5.21}$$

where $\breve{t}_I$ denotes the imputed estimator obtained from the corresponding deterministic imputation counterpart.

1.5.3.2 Nonresponse bias and variance

Under the so-called *mpq*-inferential framework, the bias of the imputed estimator $\widehat{t}_I$ is defined as

$$\mathbb{E}_{mpq}\left(\widehat{t}_I - t_y\right) = \mathbb{E}_m \mathbb{E}_p \mathbb{E}_q \left(\widehat{t}_I - t_y\right), \tag{1.5.22}$$

where $\mathbb{E}_m(\cdot)$, $\mathbb{E}_p(\cdot)$, and $\mathbb{E}_q(\cdot)$ denote the expectation with respect to the imputation model, the sampling design and the nonresponse mechanism, respectively.

Note that when taking the expectation and variance with respect to the imputation model, all quantities are treated as fixed except for $\mathbf{y}$. When taking the expectation and variance with respect to the sampling design, all quantities are treated as fixed except for $\mathbf{I}$. Finally when taking expectation and variance with respect to the nonresponse mechanism, all quantities are treated as fixed except for $\mathbf{R}$.

Using (1.5.20), the bias of $\widehat{t}_I$ is given by

$$\begin{aligned} Bias\left(\widehat{t}_I\right) &= \mathbb{E}_m\,\mathbb{E}_p\,\mathbb{E}_q\left(\widehat{t}_I - t_y\right) \\ &= \mathbb{E}_m\,\mathbb{E}_p\left(\widehat{t}_{y,\pi} - t_y\right) + \mathbb{E}_m\,\mathbb{E}_p\,\mathbb{E}_q\left(\widehat{t}_I - \widehat{t}_{y,\pi}\right) \\ &= \mathbb{E}_q\,\mathbb{E}_p\,\mathbb{E}_m\left(\widehat{t}_I - \widehat{t}_{y,\pi}\right), \end{aligned}$$

since $\mathbb{E}_p\left(\widehat{t}_{y,\pi} - t_y\right) = 0$. Above, we have used the fact that

$$\mathbb{E}_m\,\mathbb{E}_p\,\mathbb{E}_q\left(\widehat{t}_I - \widehat{t}_{y,\pi}\right) = \mathbb{E}_q\,\mathbb{E}_p\,\mathbb{E}_m\left(\widehat{t}_I - \widehat{t}_{y,\pi}\right).$$

Interchanging the order of expectations is justified because we are assuming that the data are MAR and the sampling design is non-informative. We say that the sampling is non-informative if $\mathcal{F}(\mathbf{y} \mid \mathbf{I}, \boldsymbol{V}) = \mathcal{F}(\mathbf{y} \mid \boldsymbol{V})$, i.e., sampling process has not modified the relationship between $\mathbf{y}$ and $\boldsymbol{V}$. The model that holds in the population still holds in the sample.

Example 1.5.1. Under deterministic linear regression imputation (1.5.5), the imputed estimator of t_y is given by

$$\widehat{t}_I = \sum_{i \in S_r} w_i y_i + \sum_{i \in S_m} w_i \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{WLS}.$$

The nonresponse error of $\widehat{t}_I$ can be written as

$$\widehat{t}_I - \widehat{t}_{y,\pi} = -\sum_{i \in S_m} w_i \left(y_i - \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{WLS}\right).$$

Therefore,

$$\begin{aligned} \mathbb{E}_m\left(\widehat{t}_I - \widehat{t}_{y,\pi}\right) &= \mathbb{E}_m\left\{-\sum_{i \in S_m} w_i \left(y_i - \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{WLS}\right)\right\} \\ &= -\sum_{i \in S_m} w_i\,\mathbb{E}_m\left(y_i - \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{WLS}\right) \end{aligned}$$

$$
\begin{aligned}
&= -\sum_{i \in S_m} w_i \left\{ \mathbb{E}_m \left(y_i \right) - \boldsymbol{v}_i^\top \mathbb{E}_m \left(\widehat{\boldsymbol{B}}_{WLS} \right) \right\} \\
&= -\sum_{i \in S_m} w_i \left(\boldsymbol{v}_i^\top \boldsymbol{\beta} - \boldsymbol{v}_i^\top \boldsymbol{\beta} \right) \\
&= 0.
\end{aligned}
$$

Therefore, under MAR and the non-informativeness assumption, we have

$$
\begin{aligned}
Bias \left(\widehat{t}_I \right) &= \mathbb{E}_m \, \mathbb{E}_p \, \mathbb{E}_q \left(\widehat{t}_I - t_y \right) \\
&= 0.
\end{aligned}
$$

Hence, under the *mpq*-inferential framework, the deterministic linear regression imputation leads to an *mpq*-unbiased estimator of population total provided that the first moment $\mathbb{E} \left(y_i \mid \boldsymbol{v}_i \right)$ of the imputation model is correctly specified.

Example 1.5.2. Under random linear regression imputation, the imputed estimator of t_y is given by

$$
\widehat{t}_I = \sum_{i \in S_r} w_i y_i + \sum_{i \in S_m} w_i \left(\boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{WLS} + \widehat{\sigma} \sqrt{c_i} e_i^* \right).
$$

The nonresponse error of $\widehat{t}_I$ can be expressed as (see example 1.5.1)

$$
\breve{t}_I - \widehat{t}_{y,\pi} = -\sum_{i \in S_m} w_i \left(y_i - \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{WLS} \right),
$$

whereas the imputation error can be written as

$$
\begin{aligned}
\widehat{t}_I - \breve{t}_I &= \left\{ \sum_{i \in S_r} w_i y_i + \sum_{i \in S_m} w_i \left(\boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{WLS} + \widehat{\sigma} \sqrt{c_i} e_i^* \right) \right\} - \left(\sum_{i \in S_r} w_i y_i + \sum_{i \in S_m} w_i \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{WLS} \right) \\
&= \sum_{i \in S_m} \widehat{\sigma} w_i \sqrt{c_i} e_i^*.
\end{aligned}
$$

Recall that $\mathbb{E}_I \left(\cdot \right)$ denote the expectation with respect to the imputation mechanism that is used for random selection of the residuals e_i^*. When taking the expectation with respect to the imputation mechanism, except for the residuals e_i^*, all the other quantities are treated as fixed. Under MAR (see section 1.5.1) and non-informative sampling, we have

$$
\begin{aligned}
Bias \left(\widehat{t}_I \right) &= \mathbb{E}_m \, \mathbb{E}_p \, \mathbb{E}_q \, \mathbb{E}_I \left(\widehat{t}_I - t_y \right) \\
&= \mathbb{E}_m \, \mathbb{E}_p \left(\widehat{t}_{y,\pi} - t_y \right) + \mathbb{E}_m \, \mathbb{E}_p \, \mathbb{E}_q \left(\breve{t}_I - \widehat{t}_{y,\pi} \right) + \mathbb{E}_m \, \mathbb{E}_p \, \mathbb{E}_q \, \mathbb{E}_I \left(\widehat{t}_I - \breve{t}_I \right)
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}_q\, \mathbb{E}_p\, \mathbb{E}_m \left(\breve{t}_I - \widehat{t}_{y,\pi}\right) + \mathbb{E}_m\, \mathbb{E}_p\, \mathbb{E}_q\, \mathbb{E}_I \left(\sum_{i \in S_m} \widehat{\sigma} w_i \sqrt{c_i} e_i^* \right) \\
&= \mathbb{E}_m\, \mathbb{E}_p\, \mathbb{E}_q \left\{ \sum_{i \in S_m} \widehat{\sigma} w_i \sqrt{c_i}\, \mathbb{E}_I \left(e_i^*\right) \right\} \\
&= 0,
\end{aligned}
$$

since $\mathbb{E}_I\left(e_i^*\right) = 0$. We used the decomposition of total error (1.5.21). The penultimate equality follows from the fact that deterministic linear regression imputation leads to an *mpq*-unbiased estimator of t_y (see Example 1.5.1).

Next, we turn to the variance of $\widehat{t}_I$. Assuming that $\widehat{t}_I$ is an unbiased estimator for t_y, i.e., $Bias\left(\widehat{t}_I\right) = 0$, the overall variance of the imputed estimator $\widehat{t}_I$ can be expressed as (Särndal, 1992):

$$
\begin{aligned}
\mathbb{V}_{tot} &= \mathbb{V}\left(\widehat{t}_I - t_y\right) \\
&= \mathbb{E}_{mpq}\left(\widehat{t}_I - t_y\right)^2 - \left\{\mathbb{E}_{mpq}\left(\widehat{t}_I - t_y\right)\right\}^2 \\
&= \mathbb{E}_m\, \mathbb{E}_p\, \mathbb{E}_q \left(\widehat{t}_I - t_y\right)^2 \\
&= \mathbb{E}_m\, \mathbb{E}_p\, \mathbb{E}_q \left\{\left(\widehat{t}_{y,\pi} - t_y\right) + \left(\widehat{t}_I - \widehat{t}_{y,\pi}\right)\right\}^2 \\
&= \mathbb{E}_m\, \mathbb{E}_p\, \mathbb{E}_q \left(\widehat{t}_{y,\pi} - t_y\right)^2 + \mathbb{E}_m\, \mathbb{E}_p\, \mathbb{E}_q \left(\widehat{t}_I - \widehat{t}_{y,\pi}\right)^2 \\
&\quad + 2\, \mathbb{E}_m\, \mathbb{E}_p\, \mathbb{E}_q \left\{\left(\widehat{t}_{y,\pi} - t_y\right)\left(\widehat{t}_I - \widehat{t}_{y,\pi}\right)\right\} \\
&= \mathbb{E}_m\, \mathbb{V}_p \left(\widehat{t}_{y,\pi}\right) + \mathbb{E}_p\, \mathbb{E}_q\, \mathbb{E}_m \left\{\left(\widehat{t}_I - \widehat{t}_{y,\pi}\right)^2\right\} \\
&\quad + 2\, \mathbb{E}_p\, \mathbb{E}_q\, \mathbb{E}_m \left\{\left(\widehat{t}_{y,\pi} - t_y\right)\left(\widehat{t}_I - \widehat{t}_{y,\pi}\right)\right\} \\
&= \mathbb{E}_m\, \mathbb{V}_p \left(\widehat{t}_{y,\pi}\right) + \mathbb{E}_p\, \mathbb{E}_q\, \mathbb{V}_m \left(\widehat{t}_I - \widehat{t}_{y,\pi}\right) \\
&\quad + 2\, \mathbb{E}_p\, \mathbb{E}_q\, Cov_m \left\{\left(\widehat{t}_{y,\pi} - t_y\right), \left(\widehat{t}_I - \widehat{t}_{y,\pi}\right)\right\} \\
&= \mathbb{V}_{sam} + \mathbb{V}_{nr} + 2\, \mathbb{V}_{mix}, \qquad (1.5.23)
\end{aligned}
$$

where

$$\mathbb{V}_{sam} = \mathbb{E}_m\, \mathbb{V}_p \left(\widehat{t}_{y,\pi}\right), \qquad (1.5.24)$$

$$\mathbb{V}_{nr} = \mathbb{E}_p\, \mathbb{E}_q\, \mathbb{V}_m \left(\widehat{t}_I - \widehat{t}_{y,\pi}\right), \qquad (1.5.25)$$

and

$$\mathbb{V}_{mix} = \mathbb{E}_p\, \mathbb{E}_q\, Cov_m \left\{\left(\widehat{t}_{y,\pi} - t_y\right), \left(\widehat{t}_I - \widehat{t}_{y,\pi}\right)\right\}. \qquad (1.5.26)$$

The total variance of $\widehat{t}_I$ consists of the sum of three components: the sampling variance, the nonresponse variance and a mixed term that corresponds to the covariance between the sampling and the nonresponse error. An estimate of the total

variance $\widehat{V}_{tot}$ is obtained by estimating each component individually, which leads to

$$\widehat{V}_{tot} = \widehat{V}_{sam} + \widehat{V}_{nr} + 2\widehat{V}_{mix}. \tag{1.5.27}$$

To estimate the sampling variance given by (1.5.24), there exists several approaches. We introduce the method proposed by Särndal (1992) and Deville and Särndal (1994). We start with the naive variance estimator that treats the imputed values y_i^* as observed values:

$$\widehat{V}_{naive} = \sum_{i\in S}\sum_{j\in S} \frac{(\pi_{ij} - \pi_i\pi_j)}{\pi_{ij}} \frac{\tilde{y}_i}{\pi_i}\frac{\tilde{y}_j}{\pi_j}. \tag{1.5.28}$$

The estimator $\widehat{V}_{naive}$ often underestimates the true sampling variance. Therefore, (Särndal, 1992) suggested a bias-adjusted estimator of $\mathbb{V}_{sam}$

$$\widehat{V}_{sam} = \widehat{V}_{naive} + \widehat{V}_{diff},$$

where $\widehat{V}_{diff}$ is an estimator of

$$\mathbb{V}_{diff} = \mathbb{E}_m\left(\widehat{V}_{HT} - \widehat{V}_{naive}\right).$$

Estimating the other two variance terms $\mathbb{V}_{nr}$ and $\mathbb{V}_{mix}$ is more straightforward (see Haziza and Vallée (2020)).

We illustrate the method of (Särndal, 1992) in the case of SRSWOR and mean imputation (see Appendix A for more details). Recall that, mean imputation is a special case of regression imputation with $\boldsymbol{v}_i = 1$ and $c_i = 1$ for all i.

First, to obtain the sampling variance estimator $\widehat{V}_{sam}$, we need to determine $\mathbb{V}_{diff}$. From (1.4.3) and (1.5.28), we have

$$\widehat{V}_{HT} - \widehat{V}_{naive} = \sum_{i\in S}\sum_{j\in S} \frac{(\pi_{ij} - \pi_i\pi_j)}{\pi_{ij}} \frac{1}{\pi_i}\frac{1}{\pi_j}\left(y_iy_j - \tilde{y}_i\tilde{y}_j\right),$$

where $\widehat{V}_{HT}$ is given by (1.4.3). In the case of SRSWOR and mean imputation, straightforward algebra leads to

$$\widehat{V}_{HT} - \widehat{V}_{naive} = N^2\left(1 - \frac{n}{N}\right)\frac{1}{n}\left(s_y^2 - \frac{n_r - 1}{n-1}s_{yr}^2\right),$$

where $s_y^2 = \frac{1}{n-1}\sum_{i\in S}(y_i - \bar{y})^2$ and $s_{yr}^2 = \frac{1}{n_r-1}\sum_{i\in S_r}(y_i - \bar{y}_r)^2$. It follows that

$$\mathbb{V}_{diff} = N^2\left(1 - \frac{n}{N}\right)\frac{n - n_r}{n-1}\frac{\sigma^2}{n}$$

$$= N^2 \left(1 - \frac{n}{N}\right) (1 - \widehat{p}) \frac{n}{n-1} \frac{\sigma^2}{n},$$

where $\widehat{p} = n_r/n$ denotes the response rate. Since σ^2 in the above formula for $\mathbb{V}_{diff}$ is unknown, we replace it by an m-unbiased estimator s^2_{yr}, i.e., $\mathbb{E}_m\left(s^2_{yr}\right) = \sigma^2$, which leads to

$$\begin{aligned} \widehat{V}_{sam} &= \widehat{V}_{naive} + \widehat{V}_{diff} \\ &= N^2 \left(1 - \frac{n}{N}\right) \frac{s^2_{yr}}{n}. \end{aligned} \tag{1.5.29}$$

Since $\mathbb{E}_m\left(s^2_{yr}\right) = \sigma^2$, we have $\mathbb{E}_m\left(\widehat{V}_{sam}\right) = V_{sam}$.

Second, it can be shown that the nonresponse variance $\mathbb{V}_{nr}$ is given by

$$\mathbb{V}_{nr} = \sigma^2 \frac{N^2}{n^2} \sum_{i \in S} \left(r_i \frac{n}{n_r} - 1\right)^2,$$

which can be estimated by

$$\widehat{V}_{nr} = \frac{N^2}{n_r} \left(1 - \frac{n_r}{n}\right) s^2_{yr}. \tag{1.5.30}$$

Finally, for the $\mathbb{V}_{mix}$ term, we have

$$\mathbb{V}_{mix} = \sigma^2 \frac{N(N-n)}{n^2} \sum_{i \in S} \left(r_i \frac{n}{n_r} - 1\right) = 0. \tag{1.5.31}$$

An estimator of the total variance $\mathbb{V}_{tot}$ under mean imputation and simple random sampling without replacement is obtained by summing (1.5.29) and (1.5.30), which leads to

$$\widehat{V}_{tot} = N^2 \left(1 - \frac{n_r}{N}\right) \frac{s^2_{yr}}{n_r}.$$

Finally, in the case of random imputation, the imputed estimator $\widehat{t}_I$ suffers from an additional variability due to random draws of the e_i^*'s. This additional variability is referred to as the imputation variance and is added to the overall variance. Therefore, using the decomposition of total error (1.5.21), the overall variance can be expressed as

$$\mathbb{V}_{tot} = \mathbb{V}_{sam} + \mathbb{V}_{nr} + 2\,\mathbb{V}_{mix} + \mathbb{V}_{imp}, \tag{1.5.32}$$

where $\mathbb{V}_{imp} = \mathbb{E}_p \mathbb{E}_q \mathbb{E}_m \mathbb{V}_I \left(\widehat{t}_I\right)$ is an additional component that corresponds to the imputation variance. As before, we have a sampling variance term $\mathbb{V}_{sam} = \mathbb{E}_m \mathbb{V}_p \left(\widehat{t}_{y,\pi}\right)$, a nonresponse variance term $\mathbb{V}_{nr} = \mathbb{E}_p \mathbb{E}_q \mathbb{V}_m \mathbb{E}_I \left(\widehat{t}_I - \widehat{t}_{y,\pi}\right)$ and a mixed term that corresponds to the covariance between the sampling and the nonresponse error $\mathbb{V}_{mix} = \mathbb{E}_p \mathbb{E}_q Cov_m \mathbb{E}_I \left\{\left(\widehat{t}_{y,\pi} - t_y\right), \left(\widehat{t}_I - \widehat{t}_{y,\pi}\right)\right\}$.

An estimator of the total variance $\widehat{V}_{tot}$ can be obtained by estimating each of the four variance components individually. For instance, for SRSWOR and random hot-deck imputation (see Section 1.5.2.3), the imputed estimator of t_y is given by

$$\widehat{t}_I = \sum_{i \in S_r} w_i y_i + \sum_{i \in S_m} w_i \bar{y}_r + \sum_{i \in S_m} w_i e_i^*.$$

We have

$$\mathbb{E}_I \left(e_i^*\right) = 0,$$

and

$$\mathbb{V}_I \left(e_i^*\right) = \frac{\sum_{i \in S_r} e_i^2}{n_r}, \tag{1.5.33}$$

where $e_i = y_i - \bar{y}_r$, $i \in S_r$. Therefore, using (1.5.33), we have

$$\begin{aligned}
\mathbb{V}_{imp} &= \mathbb{E}_p \mathbb{E}_q \mathbb{E}_m \mathbb{V}_I \left(\widehat{t}_I\right) \\
&= \mathbb{E}_p \mathbb{E}_q \mathbb{E}_m \left\{ \left(\frac{N^2}{n^2}\right) \sum_{i \in S_m} \mathbb{V}_I \left(e_i^*\right) \right\} \\
&= \mathbb{E}_p \mathbb{E}_q \mathbb{E}_m \left\{ \frac{N^2}{n} (1 - \widehat{p}) \frac{\sum_{i \in S_r} e_i^2}{n_r} \right\}.
\end{aligned}$$

Thus, an estimator of the imputation variance is given by

$$\widehat{V}_{imp} = \frac{N^2}{n} (1 - \widehat{p}) \frac{\sum_{i \in S_r} e_i^2}{n_r}.$$

The terms $\widehat{V}_{sam}$, $\widehat{V}_{nr}$ are given by (1.5.29) and (1.5.30), respectively, and we have $\mathbb{V}_{mix} = 0$ as in (1.5.31). This is because random hot-deck imputation can be viewed as mean imputation method plus an added random residual and $\mathbb{E}_I \left(e_i^*\right) = 0$.

Chapter 2

Common treatment of influential units

In surveys, we often encounter the problem of influential values in the selected sample. This problem is especially common in business surveys that collect economic variables whose distributions are highly skewed. By influential values, we refer to an accurately recorded value provided by the respondent, and not a measurement error that is usually addressed during the data editing stage. Thus, influential units are legitimate observations that can represent similar non-responding units or other population units that were not selected in the sample. A unit is considered to be influential when its presence or absence in the selected sample can significantly impact the magnitude of the estimates. Classical estimators such as the Horvitz-Thompson estimator (see Section 1.4) may be unstable in the presence of these influential units.

In this thesis, we consider deterministic linear regression imputation to fill in for the missing values. The imputed estimator of a population total $\widehat{t}_I$ is valid provided that the first moment of the imputation model is correctly specified; see Example 1.5.1. However, in the presence of influential units, the imputed estimator $\widehat{t}_I$ may be very unstable, which is undesirable. To address this issue, the main idea is to reduce the influence of units that have a large influence, resulting in a biased but more stable estimator. This involves a trade-off between bias and variance. Indeed, a robust imputed estimator, despite its bias, is expected to achieve a smaller mean square error compared to the non-robust imputed estimator.

In this chapter, we introduce two commonly employed methods for the treatment of influential values at the imputation stage. Through a simulation study, we show that these methods are generally not satisfactory.

2.1 Two naive methods commonly used in practice

Recall that the classical imputed estimator of a population total (see section 1.5.2) is given by

$$\widehat{t}_I = \sum_{i \in S_r} w_i y_i + \sum_{i \in S_m} w_i y_i^*, \tag{2.1.1}$$

where y_i^* is the imputed value for the missing y_i. Under deterministic linear regression imputation (see section 1.5.2.1), recall that the imputation model is given by

$$y_i = \boldsymbol{v}_i^\top \boldsymbol{\beta} + \epsilon_i,$$

such that

$$\mathbb{E}_m\left(\epsilon_i | \boldsymbol{v}_i\right) = 0, \quad \mathbb{E}_m\left(\epsilon_i \epsilon_j | \boldsymbol{v}_i, \boldsymbol{v}_j\right) = 0, \; i \neq j, \quad \mathbb{V}_m\left(\epsilon_i | \boldsymbol{v}_i\right) = \sigma^2 c_i.$$

The missing values y_i, $i \in S_m$, are replaced with $y_i^* = \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{WLS}$, where

$$\widehat{\boldsymbol{B}}_{WLS} = \left(\sum_{i \in S_r} \phi_i \boldsymbol{v}_i c_i^{-1} \boldsymbol{v}_i^\top \right)^{-1} \sum_{i \in S_r} \phi_i \boldsymbol{v}_i c_i^{-1} y_i \tag{2.1.2}$$

denotes the weighted least squares (WLS) estimator of $\boldsymbol{\beta}$ based on the responding units.

Below we present two naive methods usually employed in practice. The first method consists of replacing the estimator $\widehat{\boldsymbol{B}}_{WLS}$ by a robust version, for instance, an M-estimator based on the Huber function with the usual tuning constant $c = 1.345$. The second method consists of finding an estimator of $\boldsymbol{\beta}$ based on the responding units after removing outliers. A more detailed discussion of the two methods is provided next.

2.1.1 Robust regression

Robust regression methods are often employed to limit the influence of unusual observations on the values of estimated regression coefficients. There exists various types of robust estimators in the context of a linear regression model. Here, we consider *M-estimators*, where M stands for maximum likelihood-type (Huber, 1981).

Definition 2.1. *The M-estimator of $\boldsymbol{\beta}$ is an estimator that satisfies*

$$\widehat{\boldsymbol{\beta}}_M = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left(\frac{y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}}{\widehat{s}} \right),$$

where the function ρ is an objective function of the M-estimator and $\widehat{s}$ denotes a robust estimator for the residual standard deviation, such as the median absolute deviation (MAD), to ensure scale-invariance for the M-estimator.

The goal is to replace the estimator $\widehat{\boldsymbol{B}}_{WLS}$ by a robust version $\widehat{\boldsymbol{B}}_R(c)$. We consider an M-estimator based on the Huber function. Using the set of responding units, the minimization problem is given by

$$\widehat{\boldsymbol{B}}_R(c) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i \in S_r} \rho\left(\frac{y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}}{\sqrt{c_i}\widehat{\sigma}}\right), \tag{2.1.3}$$

where $\sqrt{c_i}\widehat{\sigma}$ is an estimator of $\sqrt{\mathbb{V}(\epsilon_i|\boldsymbol{v}_i)} = \sqrt{\sigma^2 c_i}$, and ρ is the Huber objective function (see Table 2.1 for a list of commonly encountered objective functions).

Taking the derivative of (2.1.3) with respect to $\boldsymbol{\beta}$, the estimator $\widehat{\boldsymbol{B}}_R(c)$ is then the solution of the following estimating equation of M-estimator:

$$\sum_{i \in S_r} \psi_c\left(\frac{y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}}{\sqrt{c_i}\widehat{\sigma}}\right) \frac{\boldsymbol{v}_i}{\sqrt{c_i}} = \mathbf{0}, \tag{2.1.4}$$

where $\psi_c(\cdot)$ is the Huber influence function (see Table 2.1) and c is a tuning constant. The tuning constant c determines how resistant the estimate is to outliers. The smaller the value of the tuning constant c, the more resistant the estimate to outliers. It eventually approaches the estimate obtained using weighted least squares as c gets larger. In classical statistics, the value of c is often set to 1.345 because this value offers 95 % efficiency for normally distributed data. Thus, the imputed values are given by

$$y_i^* = \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_R(1.345), \quad i \in S_m.$$

There is, in general, no closed-form solution for the expression of an M-estimator. Define the weight function as $w(x) = \psi(x)/x$ and the weights $w_i = w\left(\frac{y_i - \boldsymbol{v}_i^\top \widehat{\boldsymbol{\beta}}}{\hat{s}}\right)$. Then, the M-estimator can be viewed as the solution of a weighted least-squares problem:

$$\widehat{\boldsymbol{\beta}}_R = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} w_i (y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta})^2.$$

Let $\boldsymbol{V}$ be a matrix of independent variables, including the constant term for intercept, $\boldsymbol{W}$ be a diagonal matrix, whose ith diagonal element is w_i, $i = 1, \ldots, n$, and $\boldsymbol{y}$ be an n-vector of observed values of the dependent variable. The M-estimator can be expressed as the solution of a weighted least squares problem

$$\widehat{\boldsymbol{\beta}}_R = \left(\boldsymbol{V}^\top \boldsymbol{W} \boldsymbol{V}\right)^{-1} \boldsymbol{V}^\top \boldsymbol{y}. \tag{2.1.5}$$

However, we cannot directly compute $\widehat{\boldsymbol{\beta}}_R$ using (2.1.5) since the matrix $\boldsymbol{W}$ depends on the estimator $\widehat{\boldsymbol{\beta}}_R$. Therefore, the estimate is usually solved numerically using methods such as the *iterative reweighted least squares (IRLS) procedure*. Starting

with an initial estimate $\widehat{\boldsymbol{\beta}}_R$, this procedure iteratively refines this estimate of the M-estimator by updating the matrix $\boldsymbol{W}$ at each iteration.

In Table (2.1), we give some examples of the objective function $\rho(x)$ for the M-estimator, along with the corresponding derivative, called the influence function $\psi(x)$, and the weight function $w(x)$.

Type	Objective function $\rho(x)$	Influence function $\psi(x)$	Weight function $w(x)$
Least squares	$\frac{1}{2}x^2$	x	1
Huber $(c>0)$	$\begin{cases} \frac{1}{2}x^2 & \text{if } \lvert x\rvert < c \\ c\lvert x\rvert - \frac{1}{2}c^2 & \text{if } \lvert x\rvert \geq c \end{cases}$	$\begin{cases} x & \text{if } \lvert x\rvert < c \\ csign(x) & \text{if } \lvert x\rvert \geq c \end{cases}$	$\begin{cases} 1 & \text{if } \lvert x\rvert < c \\ \frac{c}{\lvert x\rvert} & \text{if } \lvert x\rvert \geq c \end{cases}$
Tukey bisquare $(c>0)$	$\begin{cases} \frac{c^2}{6}\left(1-\left[1-\left(\frac{x}{c}\right)^2\right]^3\right) & \text{if } \lvert x\rvert \leq c \\ \frac{c^2}{6} & \text{if } \lvert x\rvert > c \end{cases}$	$\begin{cases} x\left(1-\left(\frac{x}{c}\right)^2\right)^2 & \text{if } \lvert x\rvert \leq c \\ 0 & \text{if } \lvert x\rvert > c \end{cases}$	$\begin{cases} \left(1-\left(\frac{x}{c}\right)^2\right)^2 & \text{if } \lvert x\rvert \leq c \\ 0 & \text{if } \lvert x\rvert > c \end{cases}$

Table 2.1: Some examples of functions for M-estimator

As previously mentioned, in addition to M-estimators, there are also other types of robust estimators, such as *S-estimators, MM-estimators, GM-estimators, LTS estimators*, etc. See (Andersen, 2008) for more details.

2.1.2 Exclusion of outliers

The second method consists of identifying the influential units (usually by an outlier detection method), removing these units, and fitting the customary linear regression model based on the remaining responding units. This leads to

$$y_i^* = \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{WLS}^*, \quad i \in S_m,$$

where

$$\widehat{\boldsymbol{B}}_{WLS}^* = \left(\sum_{i\in S_r} w_i a_i \boldsymbol{v}_i c_i^{-1} \boldsymbol{v}_i^\top\right)^{-1} \sum_{i\in S_r} w_i a_i \boldsymbol{v}_i c_i^{-1} y_i, \tag{2.1.6}$$

with $a_i = 1$ if unit i is not identified as an influential unit and $a_i = 0$, otherwise.

The underlying assumption is that the discarded respondent y-values are unique, i.e., they do not represent similar nonrespondents. These units are called the nonrepresentative respondents. In practice, a respondent may represent other similar units in the set of nonrespondents or in the non-sampled part of the population.

Next, we present two measures of influence of a unit often used in practice to identify outliers: studentized residuals and Cook's distance.

2.1.2.1 Studentized residuals

The studentized residual attached to the ith observation is defined as

$$t_i = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}},$$

where the residuals $e_i = y_i - \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{WLS}$ and h_{ii} is the ith diagonal element of the "hat" matrix, $\mathbf{H} = \boldsymbol{V}\left(\boldsymbol{V}^\top \boldsymbol{V}\right)^{-1}\boldsymbol{V}^\top$, where $\boldsymbol{V}$ is a $n \times (p+1)$ design matrix. The term $\hat{\sigma}_{(i)} = \sum_{\substack{j=1 \\ j\neq i}}^{n} e_j^2 / (n-p-1)$ is the mean square error of the fitted model with the ith observation deleted. The studentized residual t_i follows a t-distribution with $n-p-2$ degrees of freedom, where n is the number of observations and p, the number of variables in the model.

An observation is influential if

$$|t_i| > t_{\alpha,n-p-2},$$

where $t_{\alpha,n-p-2}$ is the critical value for a two-tailed test with a level of significance α. Here, we set $\alpha = 0.05$. For a large number of observations n, we expect around 5% of observations to fall outside of the range $|t_i| \leq 2$. These observations may be considered as influential.

2.1.2.2 Cook's distance

Cook's distance (Cook, 1977) is a measure of the influence on regression coefficients and therefore fitted response values when the ith observation is deleted. The Cook's distance for the ith observation is given by

$$\begin{aligned} D_i &= \frac{e_i^2}{(p+1)\hat{\sigma}^2} \times \frac{h_{ii}}{(1-h_{ii})^2} \\ &= \frac{1}{(p+1)} \left(\frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}\right)^2 \times \frac{h_{ii}}{1-h_{ii}} \\ &= \frac{\tilde{e}_i^2}{(p+1)} \times \frac{h_{ii}}{1-h_{ii}}, \end{aligned} \tag{2.1.7}$$

where $\hat{\sigma}^2 = \sum_{j=1}^{n} e_j^2 / (n-p-1)$ is the mean square error of the fitted model and

$$\tilde{e}_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

denotes the standardized residual of the ith observation. A large Cook's distance D_i is an indication that the ith observation has a significant influence on the fitted response values, which is the case when the vertical residual e_i and/or the leverage h_{ii} are large.

To determine which observations are influential, we consider the cutoff

$$D_i > \frac{4}{n-p-1},$$

which was advocated by Chatterjee and Hadi (1988).

2.1.3 Simulation study: assessing the performance of naive methods

We conducted a simulation study to examine the performance of the methods presented in Sections 2.1.1 and 2.1.2.

We repeated 10, 000 iterations of the following process:

(1) A population U of size $N = 10,000$ was generated, with one survey variable y and one covariate v, using a mixture of normal distribution with a proportion of outliers equal to 5%. First, we generated $v \sim Gamma(1, 10)$. In the case of asymmetric outliers, given the v-values, the y-values were generated according to the following mixture model

$$Y_i = \alpha_i \times \{1,000 + 5v_i + \mathcal{N}(0; 19,600v_i)\} + (1-\alpha_i) \times \{9,000 + 20v_i + \mathcal{N}(0; 640,000v_i)\},$$

where $P(\alpha_i = 1) = 0.95$ for a population containing approximately 5% of outliers; see Figure 2.1. In the case of symmetric outliers, we used

$$Y_i = \alpha_i \times \{100 + 8v_i + \mathcal{N}(0; 64v_i)\} + (1-\alpha_i) \times \{100 + 8v_i + \mathcal{N}(0; 14,400v_i)\},$$

where $P(\alpha_i = 1) = 0.95$; see Figure 2.2.

(2) A sample S of size $n = 100; 200; 500$ was selected from U according to simple random sampling without replacement;

(3) Nonresponse to the y-variable was generated according to a uniform nonresponse mechanism with $p_i = 50\%$ for all i;

(4) Missing values were imputed using three imputation procedures, which led to three imputed estimators:

(i) Non-robust imputed estimator:

$$\widehat{t}_{I,WLS} = \sum_{i \in S_r} w_i y_i + \sum_{i \in S_m} w_i \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{WLS}, \tag{2.1.8}$$

where $\widehat{\boldsymbol{B}}_{WLS}$ is given by (2.1.2).

(ii) Imputed estimator based on robust regression:

$$\widehat{t}_I(c) = \sum_{i \in S_r} w_i y_i + \sum_{i \in S_m} w_i \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_R(c), \tag{2.1.9}$$

where $\widehat{\boldsymbol{B}}_R(c)$ is given by (2.1.3). We used the Huber function with $c = 0.1; 1.345; 2.5$.

(iii) Imputed estimator based on the exclusion of outliers:

$$\widehat{t}^*_{I,WLS} = \sum_{i \in S_r} w_i y_i + \sum_{i \in S_m} w_i \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}^*_{WLS}, \tag{2.1.10}$$

where $\widehat{\boldsymbol{B}}^*_{WLS}$ is given by (2.1.6). To identify the outliers, we used the Cook distance with threshold $c = 4/(n-3)$ and the studentized residuals with $c = 2; 2.5; 3$.

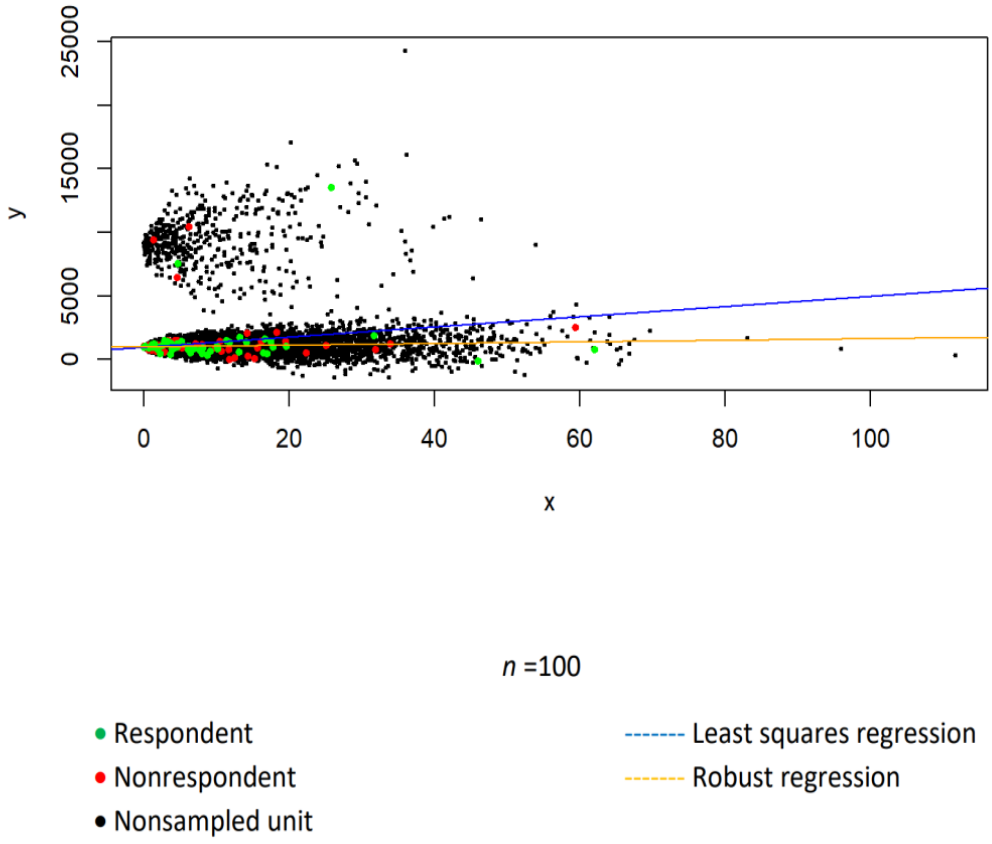


Figure 2.1: Data generated from a mixture distribution with 5% asymmetric outliers

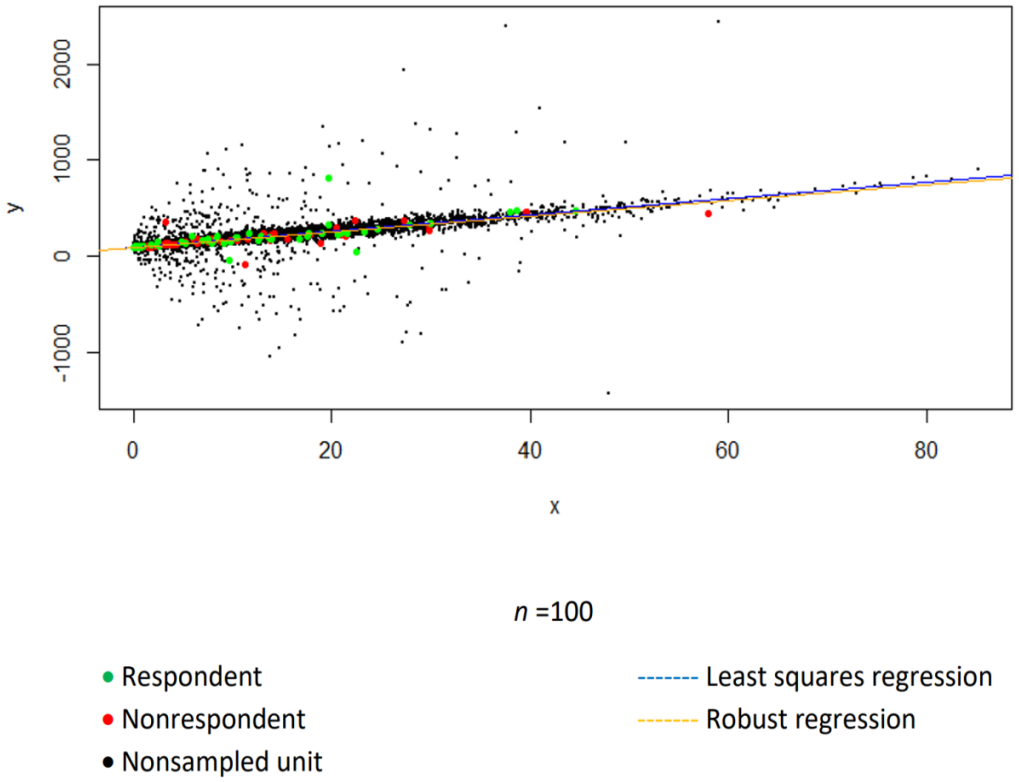


Figure 2.2: Data generated from a mixture distribution with 5% symmetric outliers

We computed the Monte Carlo percent relative bias and the Monte Carlo relative efficiency of the imputed estimators.

(1) Monte Carlo percent relative bias :

$$\mathrm{RB}(\widehat{t}_I) = \frac{\mathbb{E}_{MC}\left(\widehat{t}_I - t_y\right)}{\mathbb{E}_{MC}\left(\widehat{t}_I\right)} \times 100;$$

(2) Monte Carlo relative efficiency, using $\widehat{t}_{I,WLS}$ as the reference :

$$\mathrm{RE}(\widehat{t}_I) = 100 \times \frac{\mathrm{MSE}_{MC}(\widehat{t}_I)}{\mathrm{MSE}_{MC}(\widehat{t}_{I,WLS})}.$$

Table (2.2) displays the empirical performance of the imputed estimators in terms of bias and efficiency for different sample sizes n in the case of asymmetric outliers.

	WLS	Robust regression			WLS (Exclude outliers)			
n		$c = 0.1$	$c = 1.345$	$c = 2.5$	Studentized $c = 2$	Studentized $c = 2.5$	Studentized $c = 3$	Cook distance
100	0.0	-13.0	-12.7	-11.9	-10.7	-9.6	-8.6	-8.6
	(100)	(77)	(75)	(74)	(87)	(89)	(92)	(92)
200	-0.0	-13.0	-12.7	-12.0	-10.2	-8.9	-7.7	-7.9
	(100)	(122)	(118)	(112)	(117)	(114)	(111)	(114)
500	0.0	-13.1	-12.8	-12.1	-9.7	-8.1	-6.6	-6.9
	(100)	(267)	(256)	(235)	(204)	(177)	(156)	(167)

Table 2.2: Monte Carlo percent relative bias and Monte Carlo relative efficiency of several estimators in the case of asymmetric outliers

Table (2.3) displays the empirical performance of the imputed estimators in terms of bias and efficiency for different sample sizes n in the case of symmetric outliers.

	WLS	Robust regression			WLS (Exclude outliers)			
n		$c = 0.1$	$c = 1.345$	$c = 2.5$	Studentized $c = 2$	Studentized $c = 2.5$	Studentized $c = 3$	Cook distance
100	0.0	0.0	0.0	0.0	0.0	0.0	-0.0	-0.0
	(100)	(54)	(55)	(57)	(54)	(55)	(57)	(56)
200	-0.1	-0.0	-0.0	-0.1	-0.0	-0.1	-0.1	-0.1
	(100)	(54)	(55)	(57)	(53)	(55)	(57)	(55)
500	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0
	(100)	(54)	(54)	(56)	(53)	(54)	(56)	(54)

Table 2.3: Monte Carlo percent relative bias and Monte Carlo relative efficiency of several estimators in the case of symmetric outliers

In the case of symmetric outliers, robust regression and weighted least squares regression after removing outliers behaved very well in terms of bias and efficiency; see Table 2.3. For all sample sizes, the Monte Carlo percent relative bias was close to 0 and the Monte Carlo relative efficiency was much smaller than 100 with values ranging from 53 to 57.

However, in the case of asymmetric outliers, both methods worked well in some scenarios but broke down as the sample size increased. From Table (2.2), we observe that, for $n = 100$, the Monte Carlo relative efficiency was smaller than 100 for both methods. Robust regression displayed values of RE ranging from 74 to 77, and weighted least squares regression after removing outliers had values ranging from 87 to 92. For $n = 200$ and $n = 500$, the Monte Carlo relative efficiency obtained using the two methods was above 100, with values ranging from 111 to 122 for $n = 200$ and with values ranging from 156 to 267 for $n = 500$.

Therefore, the imputed estimators based on a naive treatment were much less efficient than the imputed estimator based on weighted least squares for large sample sizes, which is not desirable. Based on our observations in Table (2.2), we notice that the Monte Carlo relative efficiency varied depending on the chosen tuning constant c. For the approach based on robust regression, the bad performance of the imputed estimator can be explained by the fact that the tuning constant $c = 1.345$ was fixed and not adaptative. This approach is appropriate in the classical setup, whereby the interest lies in describing the behavior of the inliers. In survey sampling, the goal is different, as the interest lies in estimating the overall population total that consists of a mix of outliers and inliers. As a result, the tuning constant c should be adaptative in the sense that c should increase as n increases. Finally, for the approach based on weighted least squares regression after removing outliers, the bad performance of the imputed estimator can be explained by the fact it relies on the assumption that the discarded respondent y-values are unique, i.e., they do not represent similar units in the non-responding set. In general, this assumption is not tenable.

Chapter 3

Robust estimators based on an adaptative tuning constant

In this chapter, we describe three robust estimators in the presence of influential units that are all based on an adaptative tuning constant. We also introduce the concept of conditional bias, which is a measure of influence in a finite population setting. We consider two robust estimators based on the concept of conditional bias and a robust estimator that involves finding the optimal tuning constant c that minimizes the estimated mean square error of the robust estimator.

3.1 Conditional bias

3.1.1 Conditional bias in the complete data case

The concept of conditional bias is an appropriate measure of the influence (or impact) of a unit in a finite population setting. A unit is said to be influential if its presence or absence in the sample has a significant impact on the estimate. Let θ_N be a finite population parameter and $\widehat{\theta}$ be an estimator of θ_N. The conditional bias of a sampled unit i is defined as

$$B_{1i} = \mathbb{E}_p\left(\widehat{\theta} - \theta_N | I_i = 1\right). \tag{3.1.1}$$

The conditional bias of a sampled unit, B_{1i}, is equal to the average of the sampling error, $\widehat{\theta} - \theta_N$, calculated over all samples that contain unit i. In a finite population setting, non-sampled units may also be influential. The conditional bias of an unsampled unit i is defined as

$$B_{0i} = \mathbb{E}_p\left(\widehat{\theta} - \theta_N | I_i = 0\right). \tag{3.1.2}$$

The conditional bias B_{0i} corresponds to the average difference between the estimator $\widehat{\theta}$ and the population parameter θ_N calculated over all samples that do not contain unit i; see Moreno-Rebollo et al. (1999). Here, we are interested in estimating the population total, i.e., $\theta_N \equiv t_y$.

Proposition 3.1. *Let $\widehat{\theta} = \widehat{t}_{y,\pi}$ be the Horvitz-Thompson estimator of t_y given by (1.4.1). In the absence of missing values, the conditional bias of a sampled unit i is given by*

$$B_{1i} = \sum_{j \in U} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_j. \tag{3.1.3}$$

Proof:

$$\begin{aligned}
B_{1i} &= \mathbb{E}_p \left(\widehat{t}_{y,\pi} - t_y | I_i = 1 \right) \\
&= \mathbb{E}_p \left(\widehat{t}_{y,\pi} | I_i = 1 \right) - t_y \\
&= \mathbb{E}_p \left(\sum_{j \in S} \frac{y_j}{\pi_j} | I_i = 1 \right) - t_y \\
&= \mathbb{E}_p \left(\sum_{j \in U} I_j \frac{y_j}{\pi_j} | I_i = 1 \right) - t_y \\
&= \sum_{j \in U} \frac{y_j}{\pi_j} \mathbb{E}_p \left(I_j | I_i = 1 \right) - t_y \\
&= \sum_{j \in U} \frac{y_j}{\pi_j} \frac{P\left(I_j = 1, I_i = 1 \right)}{P\left(I_i = 1 \right)} - t_y \\
&= \sum_{j \in U} \frac{y_j}{\pi_j} \frac{\pi_{ij}}{\pi_i} - \sum_{j \in U} y_j \\
&= \sum_{j \in U} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_j.
\end{aligned}$$

∎

The conditional bias B_{1i} can also be expressed as

$$B_{1i} = \sum_{j \in U} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_j = \left(\frac{1}{\pi_i} - 1 \right) y_i + \sum_{\substack{j \in U \\ j \neq i}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_j. \tag{3.1.4}$$

If $\pi_i = 1$, we have $B_{1i} = 0$ and the unit i has no influence.

The design variance of $\widehat{t}_{y,\pi}$ given by (1.4.2) can be expressed as

$$\mathbb{V}_p\left(\widehat{t}_{y,\pi}\right) = \sum_{i \in U} B_{1i} y_i. \tag{3.1.5}$$

For (3.1.5), the design variance of $\widehat{t}_{y,\pi}$ is large if either B_{1i} or y_i is large for some $i \in U$; see Beaumont et al. (2013).

Since B_{1i} given by (3.1.3) is unknown, we need to estimate it.

Proposition 3.2. *Provided that $\pi_{ij} > 0$ for all $i \neq j$, a conditionally design-unbiased estimator of B_{1i} is given by*

$$\widehat{B}_{1i} = \sum_{j \in S} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_j \pi_{ij}}\right) y_j. \tag{3.1.6}$$

Proof:

$$\begin{aligned}
\mathbb{E}_p\left(\widehat{B}_{1i} | I_i = 1\right) &= \mathbb{E}_p\left\{\sum_{j \in S}\left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_j \pi_{ij}}\right) y_j | I_i = 1\right\} \\
&= \mathbb{E}_p\left\{\sum_{j \in U}\left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_j \pi_{ij}}\right) I_j y_j | I_i = 1\right\} \\
&= \sum_{j \in U}\left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_j \pi_{ij}}\right) y_j \, \mathbb{E}_p\left(I_j | I_i = 1\right) \\
&= \sum_{j \in U}\left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_j \pi_{ij}}\right) y_j \frac{P\left(I_i = 1, I_j = 1\right)}{P\left(I_i = 1\right)} \\
&= \sum_{j \in U}\left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_j \pi_{ij}}\right) y_j \frac{\pi_{ij}}{\pi_i} \\
&= \sum_{j \in U}\left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}\right) y_j \\
&= \sum_{j \in U}\left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right) y_j \\
&= B_{1i}
\end{aligned}$$

∎

Expressions (3.1.3) and (3.1.6) depend on the first-order inclusion probabilities, π_i, and the second-order inclusion probabilities, π_{ij}. In the case of simple random sampling without replacement, Expression (3.1.3) reduces to

$$B_{1i} = \frac{N}{N-1}\left(\frac{N}{n} - 1\right)\left(y_i - \overline{Y}\right),$$

where $\overline{Y} = N^{-1}\sum_{j\in U} y_j$. The influence of unit i is large if y_i is far from the population mean $\overline{Y}$ and/or the sampling fraction n/N is small. In the case of simple random sampling without replacement, Expression (3.1.6) reduces to

$$\widehat{B}_{1i} = \frac{n}{n-1}\left(\frac{N}{n} - 1\right)\left(y_i - \overline{y}\right),$$

where $\overline{y} = n^{-1}\sum_{j\in S} y_j$ denotes the sample mean of y-values.

For Bernoulli sampling, Expression (3.1.3) reduces to

$$B_{1i} = \left(\frac{1}{\pi} - 1\right) y_i. \tag{3.1.7}$$

For this sampling design, the conditional bias given by (3.1.7) is known for all the sample units and, therefore, does not need to be estimated. The influence of unit i is large if the first-order inclusion probability π is small and/or y_i is large.

3.1.2 Conditional bias in the context of linear regression imputation

In this section, we extend the concept of conditional bias in the context of nonresponse and linear regression imputation. In the presence of nonresponse, the expression for the conditional bias of the responding unit i can be approximated by (see Appendix B):

$$\begin{aligned} B_i^I &= \mathbb{E}_m\,\mathbb{E}_p\,\mathbb{E}_q\left(\widehat{t}_{I,WLS} - t_y | Y_i = y_i, I_i = 1, r_i = 1\right) \\ &\approx \sum_{j\in U}\left(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\right) y_j + w_i \mathbf{C}\boldsymbol{v}_i c_i^{-1}\left(y_i - \boldsymbol{v}_i^\top\boldsymbol{\beta}\right), \end{aligned} \tag{3.1.8}$$

where

$$\mathbf{C} = \left\{\sum_{i\in U}(1-p_i)\boldsymbol{v}_i^\top\right\}\left\{\sum_{i\in U} p_i\boldsymbol{v}_i c_i^{-1}\boldsymbol{v}_i^\top\right\}^{-1}, \tag{3.1.9}$$

p_i in (3.1.9) denotes the response probability of unit i in the population, and $\widehat{t}_{I,WLS}$ is given by (2.1.8). The first term on the right hand-side of (3.1.8) measures the influence of unit i on the sampling error, $\widehat{t}_{y,\pi} - t_y$ (see Expression 3.1.3), whereas the second term measures the influence of unit i on the nonresponse error, $\widehat{t}_{I,WLS} - \widehat{t}_{y,\pi}$.

Under simple random sampling without replacement, the conditional bias (3.1.8) can be written as

$$B_i^I \approx \frac{N}{N-1}\left(\frac{N}{n}-1\right)\left(y_i - \bar{Y}\right) + \left(\frac{N}{n}\right)\mathbf{C}\boldsymbol{v}_i c_i^{-1}\left(y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}\right). \tag{3.1.10}$$

Since the y-values are only available for the respondents, it is not possible to calculate the conditional bias given by (3.1.10). An estimator of the conditional bias is given by

$$\widehat{B}_i^I = \frac{n}{n-1}\left(\frac{N}{n}-1\right)\left(y_i - \bar{y}_{I,WLS}\right) + \left(\frac{N}{n}\right)\widehat{\mathbf{C}}\boldsymbol{v}_i c_i^{-1}(y_i - \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{WLS}), i \in S_r, \tag{3.1.11}$$

where

$$\widehat{\mathbf{C}} = \left\{\sum_{i\in S}(1-r_i)\boldsymbol{v}_i^\top\right\}\left\{\sum_{i\in S} r_i \boldsymbol{v}_i c_i^{-1}\boldsymbol{v}_i^\top\right\}^{-1},$$

and $\bar{y}_{I,WLS} = \widehat{t}_{I,WLS}/N$ denotes the imputed estimator of the population mean, t_y/N, under linear regression imputation.

Example 3.1.1. Under simple linear regression imputation (i.e., $\boldsymbol{v}_i = (1, v_i)^\top$ and $c_i = 1$), Expression (3.1.11) reduces to

$$\widehat{B}_i^I = \left(\frac{N}{n}-1\right)(y_i - \overline{y}_I) + \left(\frac{N}{n}\right)\frac{1}{\widehat{p}}\left\{(1-\widehat{p}) + \frac{(v_i - \overline{v}_r)(\overline{v} - \overline{v}_r)}{s_{vr}^2}\right\}\left(y_i - \widehat{B}_{0,WLS} - \widehat{B}_{1,WLS} v_i\right), \tag{3.1.12}$$

where $\widehat{p} = n_r/n$ denotes the response rate, $\overline{v} = n^{-1}\sum_{i\in S} v_i$ denotes the sample mean of v-values, $\overline{v}_r = n_r^{-1}\sum_{i\in S_r} v_i$ denotes the mean v-values of respondents, and

$$s_{vr}^2 = (n_r - 1)^{-1}\sum_{i\in S_r}(v_i - \overline{v}_r)^2;$$

see Appendix C.

It follows from (3.1.12) that the responding unit i has a large influence if

1) the sampling fraction n/N is small;

2) its y-value is far from the overall estimated mean $\overline{y}_{I,WLS}$;

3) the response rate $\widehat{p}$ is low;

4) its v-value is far from the mean of respondents $\overline{v}_r$;

5) it has a large vertical residual $e_i = y_i - \widehat{B}_{0,WLS} - \widehat{B}_{1,WLS} v_i$.

3.2 Robust estimators based on adaptative tuning constant

3.2.1 A proposal based on conditional bias

We consider a special case of the robust estimator studied by Chen et al. (2020). The estimator we consider is a robust version of

$$\widehat{t}_{I,WLS} = \sum_{i \in S_r} w_i y_i + \sum_{i \in S_m} w_i \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{WLS} \tag{3.2.1}$$

based on the concept of conditional bias:

$$\widehat{t}_{I,CB}(c) = \widehat{t}_{I,WLS} - \sum_{i \in S_r} \widehat{B}_i^I + \sum_{i \in S_r} \psi_c \left\{ \widehat{B}_i^I \right\}, \tag{3.2.2}$$

where $\widehat{B}_i^I$ is an estimator of the conditional bias given by (3.1.11) and $\psi_c(\cdot)$ denotes the Huber influence function with tuning constant c (see Table 2.1).

We can rewrite the estimator $\widehat{t}_{I,CB}(c)$ (3.2.2) as

$$\widehat{t}_{I,CB}(c) = \widehat{t}_{I,WLS} + \Delta(c), \tag{3.2.3}$$

where $\Delta(c) = -\sum_{i \in S_r} \widehat{B}_i^I + \sum_{i \in S_r} \psi_c \left\{ \widehat{B}_i^I \right\}$.

As in Beaumont et al. (2013) and Chen et al. (2020), we search for the value of $\Delta(c)$ that minimizes

$$\max_{i \in S_r} \left| \widehat{B}_i^R \right|,$$

where $\widehat{B}_i^R$ is the estimated conditional bias (influence) of unit i with respect to the robust estimator $\widehat{t}_{I,CB}(c)$. The conditional bias with respect to the robust estimator $\widehat{t}_{I,CB}(c)$ is defined as

$$
\begin{aligned}
B_i^R &= E_m E_p E_q \left(\widehat{t}_{I,CB}(c) - t_y | Y_i = y_i, I_i = 1, r_i = 1\right) \\
&= E_m E_p E_q \left(\widehat{t}_{I,WLS} + \Delta(c) - t_y | Y_i = y_i, I_i = 1, r_i = 1\right) \\
&= E_m E_p E_q \left(\widehat{t}_{I,WLS} - t_y | Y_i = y_i, I_i = 1, r_i = 1\right) + E_m E_p E_q \left(\Delta(c) | Y_i = y_i, I_i = 1, r_i = 1\right) \\
&= B_i^I + E_m E_p E_q \left(\Delta(c) | Y_i = y_i, I_i = 1, r_i = 1\right),
\end{aligned}
$$

where B_i^I is given by (3.1.10). An estimator of B_i^R is thus given by

$$\widehat{B}_i^R = \widehat{B}_i^I + \Delta(c),$$

where $\widehat{B}_i^I$ is an estimator of B_i^I. The value of $\Delta(c)$ that minimizes $\max_{i \in S_r} \left|\widehat{B}_i^R\right|$ is given by

$$\Delta(c_{opt}) = -\frac{1}{2}\left[\min_{i \in S_r}\left\{\widehat{B}_i^I\right\} + \max_{i \in S_r}\left\{\widehat{B}_i^I\right\}\right].$$

The resulting robust estimator is given by

$$
\begin{aligned}
\widehat{t}_{I,CB}(c_{opt}) &= \widehat{t}_{I,WLS} + \Delta(c_{opt}) \\
&= \widehat{t}_{I,WLS} - \frac{1}{2}\left[\min_{i \in S_r}\left\{\widehat{B}_i^I\right\} + \max_{i \in S_r}\left\{\widehat{B}_i^I\right\}\right]. \qquad (3.2.4)
\end{aligned}
$$

From Expression (3.2.4), we note that it is not necessary to determine the constant c_{opt}. Indeed, the estimator (3.2.4) can be computed explicitly without computing c_{opt}. It can be shown that c_{opt} increases as n increases, i.e., it is adaptative. Under mild regularity conditions, the robust estimator $\widehat{t}_{I,CB}(c_{opt})$ is $\sqrt{n}$-design-consistent for t_y in the sense that $\widehat{t}_{I,CB}(c_{opt}) - t_y = O_p\left(N/\sqrt{n}\right)$, which is a desirable property; see Chen et al. (2020).

3.2.2 Estimator based on a new adaptative tuning constant

The results from the simulation in Section (2.1.3) suggest that, in the case of distribution with symmetric outliers, robust regression with a fixed tuning constant $c = 1.345$ behaved very well in terms of bias and efficiency, but broke down in the case of asymmetric outliers for large sample sizes. Therefore, the idea is to propose an adaptative tuning constant c, called c_{new}, and use a robust regression method (e.g., M-estimation) with this new constant. For asymmetric outliers, using an adaptive tuning constant that adjusts automatically based on sample size and data characteristics, is expected to give better results in terms of efficiency compared to the robust estimator based on the naive tuning constant $c = 1.345$.

The value c_{new} will be displayed below. Once the value of c_{new} is determined, we can find $\widehat{\boldsymbol{B}}_{\text{R}}(c_{\text{new}})$, which is the solution of the following estimating equation of the M-estimator based on the Huber function with tuning constant c_{new} (see 2.1.4):

$$\sum_{i \in S_r} \psi_{c_{\text{new}}} \left(\frac{y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}}{\widehat{\sigma}\sqrt{c_i}} \right) \frac{\boldsymbol{v}_i}{\sqrt{c_i}} = \mathbf{0},$$

where $\psi(\cdot)$ denotes the Huber influence function (see Table 2.1). An imputed robust estimator is given by

$$\widehat{t}_{I,R}(c_{\text{new}}) = \sum_{i \in S_r} w_i y_i + \sum_{i \in S_m} w_i \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{\text{R}}(c_{\text{new}}). \tag{3.2.5}$$

Should we use (3.2.5)? It may not be a good idea because, in (3.2.5), we are only "taking care" of the missing values. However, some y-values associated with the respondents may also be influential. Therefore, we also need to reduce the influence of those units.

To cope with this issue, we write (3.2.1) in the so-called projection form

$$\widehat{t}_{I,WLS} = \sum_{i \in S} w_i \boldsymbol{v}_i^\top \widehat{\mathbf{B}}_{WLS}. \tag{3.2.6}$$

Indeed, the estimator (3.2.1) can be written in the form (3.2.6) provided that $c_i = \boldsymbol{\lambda}^\top \boldsymbol{v}_i$, for a vector of constant $\boldsymbol{\lambda}$ as we now show:

$$\begin{aligned}
\widehat{t}_{I,WLS} &= \sum_{i \in S_r} w_i y_i + \sum_{i \in S_m} w_i \boldsymbol{v}_i^\top \widehat{\mathbf{B}}_{WLS} \\
&= \sum_{i \in S_r} w_i y_i + \sum_{i \in S} w_i \boldsymbol{v}_i^\top \widehat{\mathbf{B}}_{WLS} - \sum_{i \in S_r} w_i \boldsymbol{v}_i^\top \widehat{\mathbf{B}}_{WLS} \\
&= \sum_{i \in S} w_i \boldsymbol{v}_i^\top \widehat{\mathbf{B}}_{WLS} + \sum_{i \in S_r} w_i y_i - \sum_{i \in S_r} w_i \boldsymbol{v}_i^\top \left(\sum_{i \in S_r} w_i \boldsymbol{v}_i c_i^{-1} \boldsymbol{v}_i^\top \right)^{-1} \sum_{i \in S_r} w_i \boldsymbol{v}_i c_i^{-1} y_i \\
&= \sum_{i \in S} w_i \boldsymbol{v}_i^\top \widehat{\mathbf{B}}_{WLS} + \sum_{i \in S_r} w_i y_i - \left\{ \sum_{i \in S_r} w_i \left(\frac{\boldsymbol{\lambda}^\top \boldsymbol{v}_i}{c_i} \right) \boldsymbol{v}_i^\top \right\} \left(\sum_{i \in S_r} w_i \boldsymbol{v}_i c_i^{-1} \boldsymbol{v}_i^\top \right)^{-1} \sum_{i \in S_r} w_i \boldsymbol{v}_i c_i^{-1} y_i \\
&= \sum_{i \in S} w_i \boldsymbol{v}_i^\top \widehat{\mathbf{B}}_{WLS} + \sum_{i \in S_r} w_i y_i - \boldsymbol{\lambda}^\top \left(\sum_{i \in S_r} w_i \boldsymbol{v}_i c_i^{-1} \boldsymbol{v}_i^\top \right) \left(\sum_{i \in S_r} w_i \boldsymbol{v}_i c_i^{-1} \boldsymbol{v}_i^\top \right)^{-1} \sum_{i \in S_r} w_i \boldsymbol{v}_i c_i^{-1} y_i \\
&= \sum_{i \in S} w_i \boldsymbol{v}_i^\top \widehat{\mathbf{B}}_{WLS} + \sum_{i \in S_r} w_i y_i - \boldsymbol{\lambda}^\top \sum_{i \in S_r} w_i \boldsymbol{v}_i c_i^{-1} y_i \\
&= \sum_{i \in S} w_i \boldsymbol{v}_i^\top \widehat{\mathbf{B}}_{WLS} + \sum_{i \in S_r} w_i y_i - \sum_{i \in S_r} w_i \frac{\boldsymbol{\lambda}^\top \boldsymbol{v}_i}{c_i} y_i
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i\in S} w_i \boldsymbol{v}_i^\top \widehat{\mathbf{B}}_{WLS} + \sum_{i\in S_r} w_i y_i - \sum_{i\in S_r} w_i y_i \\
&= \sum_{i\in S} w_i \boldsymbol{v}_i^\top \widehat{\mathbf{B}}_{WLS}. \qquad (3.2.7)
\end{aligned}
$$

Our proposal is obtained from (3.2.6) by replacing $\widehat{\mathbf{B}}_{WLS}$ with $\widehat{\boldsymbol{B}}_{\mathrm{R}}(c_{\mathrm{new}})$:

$$\widehat{t}_{I,R}(c_{\mathrm{new}}) = \sum_{i\in S} w_i \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{\mathrm{R}}(c_{\mathrm{new}}). \qquad (3.2.8)$$

We propose the following value for c_{new}:

$$c_{\mathrm{new}} = 1.345 \left\{ 1 + \frac{\left| \min_{i\in S_r} \left\{ \widehat{B}_i^* \right\} + \max_{i\in S_r} \left\{ \widehat{B}_i^* \right\} \right|}{2} \right\} + \frac{n}{N}\sqrt{n}, \qquad (3.2.9)$$

where

$$\widehat{B}_i^* = \frac{\widehat{B}_i^I - \overline{\widehat{B}_r^I}}{\widehat{\sigma}_{\widehat{B}}}$$

denotes the standardized version of $\widehat{B}_i^I$, with $\overline{\widehat{B}_r^I} = \frac{1}{n_r}\sum_{i\in S_r} \widehat{B}_i^I$ and

$$\widehat{\sigma}^2_{\widehat{B}} = \frac{1}{n_r - 1} \sum_{i\in S_r} \left(\widehat{B}_i^I - \overline{\widehat{B}_r^I} \right)^2.$$

We provide some justification for our choice (3.2.9). First, if the sampling fraction n/N is negligible, the second term on the right hand-side of (3.2.9) is small and we can focus on the first term. If the distribution has symmetric outliers and the weights w_i are constant, we expect $\left| \min_{i\in S_r} \left\{ \widehat{B}_i^* \right\} + \max_{i\in S_r} \left\{ \widehat{B}_i^* \right\} \right| /2$ to be close to 0. As a result, c_{new} will be slightly larger than 1.345. If the distribution has asymmetric outliers (say to the right), the term $\left| \min_{i\in S_r} \left\{ \widehat{B}_i^* \right\} + \max_{i\in S_r} \left\{ \widehat{B}_i^* \right\} \right| /2$ will be larger than 0 and c_{new} will be significantly larger than 1.345, which is desirable. As the sample size n gets larger, the sampling fraction n/N is no longer negligible. The second term on the right hand-side of (3.2.9) gets larger as n increases and $\widehat{\boldsymbol{B}}_{\mathrm{R}}(c_{\mathrm{new}})$ becomes closer and closer to $\widehat{\boldsymbol{B}}_{\mathrm{WLS}}$.

3.2.3 Optimal tuning constant

In this section, we determine the optimal constant c, denoted by c^*, that minimizes the estimated mean square error of the robust imputed estimator in the projection form:

$$\widehat{t}_{I,R}(c) = \sum_{i \in S} w_i \boldsymbol{v}_i^\top \widehat{\mathbf{B}}_R(c), \tag{3.2.10}$$

where $\widehat{\mathbf{B}}_R(c)$ is the solution of the following robust estimating equation of the M-estimator based on the Huber function with tuning constant c (see 2.1.4):

$$\sum_{i \in S_r} \psi_c \left(\frac{y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}}{\sqrt{c_i}\widehat{\sigma}} \right) \frac{\boldsymbol{v}_i}{\sqrt{c_i}} = \mathbf{0},$$

and $\psi_c(t)$ denotes the Huber influence function (see Table 2.1).

In practice, the optimal tuning constant, c^*, is determined by computing the estimated mean square error of (3.2.10) for a grid of c-values. The value corresponding to the minimum estimated mean square error is selected as c^*.

An estimator of the mean squared error of (3.2.10) is given by

$$\begin{aligned} \widehat{MSE}(\widehat{t}_{I,R}(c)) &= \max\left\{ \left(\widehat{t}_{I,R}(c) - \widehat{t}_{I,WLS}\right)^2 - \widehat{\mathbb{V}}\left(\widehat{t}_{I,R}(c) - \widehat{t}_{I,WLS}\right), 0 \right\} \\ &+ \widehat{\mathbb{V}}\left\{\widehat{t}_{I,R}(c)\right\}. \end{aligned} \tag{3.2.11}$$

The explicit expressions and derivations for the terms on the right-hand side of (3.2.11) will be given in Chapter 5. The proposed robust estimator is given by

$$\widehat{t}_{I,R}(c^*) = \sum_{i \in S} w_i \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{\mathrm{R}}(c^*). \tag{3.2.12}$$

Chapter 4

Simulation Study

In this chapter, we present the results from a simulation study whose goal was to assess the empirical performance of several imputed estimators in terms of bias and efficiency.

4.1 Simulation Setup

We repeated $R = 10,\ 000$ iterations of the following process:

(1) A finite population U, of size $N = 5,000$ was generated, with a survey variable y and a single auxiliary variable v. To that end, we generated $v \sim Gamma(5, 10)$. Given the v-values, the y-values were generated according to

$$y_i \mid v_i \sim \mathcal{D}(\mu_i; \sigma_i^2),$$

where $\mu_i = E\left(y_i|v_i\right) = \beta_0 + \beta_1 v_i$ and $\sigma_i^2 = V\left(y_i|v_i\right) = \sigma^2$.

We considered the following distributions $\mathcal{D}$: Normal, Lognormal, Pareto, Frechet, Student and Double exponential (Laplace). The values of β_0 and β_1 were set to 50 and 12, respectively. The value of σ was set to 600. For these six distributions, the first two moments of the distribution $\mathcal{D}(\mu_i; \sigma_i^2)$ were the same. We also considered mixture distributions, including mixtures of normal and lognormal distributions. For mixtures of normal distributions, given the v-values, the y-values were generated according to

$$y_i = \alpha_i \times \mathcal{N}(\mu_{1i}; \sigma_{1i}^2) + (1 - \alpha_i) \times \mathcal{N}(\mu_{2i}; \sigma_{2i}^2),$$

where α_i is a binary variable such that $P\left(\alpha_i = 1\right) = 0.99$ or 0.97, leading to 1% and 3% of outliers, respectively. We set $\mu_{1i} = 150 + 20v_i$, $\mu_{2i} = 2000 + 85v_i$, $\sigma_{1i} = 400$ and $\sigma_{2i} = 2000$.

Similarly, for mixtures of lognormal distributions, we have

$$y_i = \alpha_i \times \mathcal{LN}(\mu_{1i}; \sigma^2_{1i}) + (1 - \alpha_i) \times \mathcal{LN}(\mu_{2i}; \sigma^2_{2i}),$$

where $P(\alpha_i = 1) = 0.99$ and 0.97. We set $\mu_{1i} = 150 + 8v_i$, $\mu_{2i} = 1200 + 60v_i$, $\sigma_{1i} = 150$ and $\sigma_{2i} = 1500$; see Appendix D for plots of the relationship between y and v for each of the distributions.

(2) From the finite population generated in the previous step, we selected a sample, of size $n \in \{50, 100, 200\}$, according to simple random sampling without replacement.

(3) In each sample, nonresponse to the y-variable was generated with probability

$$p_i = 0.1 + 0.9\frac{\exp(4 - 0.09v_i)}{1 + \exp(4 - 0.09v_i)}.$$

This led to a response rate approximately equal to 50%.

(4) In each sample, we computed six estimators of t_y:

(i) The non-robust estimator described in Section 2.1:

$$\widehat{t}_{I,WLS} = \sum_{i \in S_r} w_i y_i + \sum_{i \in S_m} w_i \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{WLS}.$$

(ii) The naive estimator described in Section 2.1:

$$\widehat{t}_{I,R}(1.345) = \sum_{i \in S_r} w_i y_i + \sum_{i \in S_m} w_i \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_R(1.345).$$

(iii) The robust estimator based on the conditional bias described in Section 3.2.1:

$$\widehat{t}_{I,CB}(c_{opt}) = \widehat{t}_{I,WLS} - \frac{1}{2}\left[\min_{i \in S_r}\left\{\widehat{B}_i^I\right\} + \max_{i \in S_r}\left\{\widehat{B}_i^I\right\}\right].$$

(iv) The robust estimator based on c_{new} described in Section 3.2.2:

$$\widehat{t}_{I,R}(c_{\text{new}}) = \sum_{i \in S} w_i \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{\text{R}}(c_{\text{new}}).$$

(v) The robust estimator based on c^* described in Section 3.2.3:

$$\widehat{t}_{I,R}(c^*) = \sum_{i \in S} w_i \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{\text{R}}(c^*).$$

(vi) The (unfeasible) robust estimator based on the tuning constant $\widetilde{c}$ that minimizes the Monte Carlo mean square error of the imputed estimator $\widehat{t}_{I,R}(c)$ given by (3.2.10) :

$$\widehat{t}_{I,R}(\widetilde{c}) = \sum_{i \in S} w_i \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{\mathrm{R}}(\widetilde{c}).$$

This robust estimator can be viewed as a "gold standard". However, it is unfeasible in practice since the Monte Carlo mean square error cannot be computed using the available sample data.

To assess the bias of an estimator, we computed its Monte Carlo percent relative bias, defined as

$$RB_{MC} = \frac{1}{R} \sum_{r=1}^{R} \frac{\widehat{t}_I^{(r)} - t_y}{t_y} \times 100,$$

where $\widehat{t}_I$ is a generic notation for an imputed estimator of t_y and $\widehat{t}_I^{(r)}$ is the imputed estimator $\widehat{t}_I$ at the rth iteration, $r = 1, \ldots, R$. We also computed Monte Carlo percent relative efficiency (RE), using the non-robust estimator $\widehat{t}_{I,WLS}$, as the reference:

$$RE = 100 \times \frac{\mathrm{MSE}_{MC}(\widehat{t}_I)}{\mathrm{MSE}_{MC}(\widehat{t}_{I,WLS})},$$

where

$$\mathrm{MSE}_{MC}(\widehat{t}_I) = \frac{1}{R} \sum_{r=1}^{R} \left(\widehat{t}_I^{(r)} - t_y \right)^2.$$

4.2 Simulation Results

Tables 4.1 to 4.3 display the Monte Carlo relative bias (in %) and Monte Carlo relative efficiency (in %) for six estimators. In the case of the symmetric distributions (Normal, Student and Laplace), all the estimators exhibited a negligible bias in all the scenarios; see Table 4.1. For the normal distribution, all the robust estimators suffer from a slight loss of efficiency with values ranging from 100 to 106, which is a desirable feature. For the t-distribution and the Laplace distribution, the robust estimators were much more efficient than $\widehat{t}_{I,WLS}$. The estimator $\widehat{t}_{I,R}(c^*)$ was the best but, as expected, incurred some loss of efficiency with respect to the gold standard estimator $\widehat{t}_{I,R}(\widetilde{c})$. The estimator $\widehat{t}_{I,CB}(c_{opt})$ was outperformed by the other robust estimators.

In the case of asymmetric distributions (Pareto, Frechet, Lognormal, and Weibull), all the robust estimators exhibited some bias, as expected; see Table 4.2. In virtually all the scenarios, the estimator $\widehat{t}_{I,CB}(c_{opt})$ was less biased than its competitors. The naive estimator $\widehat{t}_{I,R}(1.345)$ performed well in some scenarios but performed poorly in others, especially for larger sample sizes. For instance, for the Lognormal distribution, the estimator $\widehat{t}_{I,R}(1.345)$ exhibited a value of RE equal to 134% for $n = 200$. For highly skewed distributions such as Pareto and Frechet, the proposed robust estimators showed substantial improvement in terms of relative efficiency with respect to $\widehat{t}_{I,WLS}$. In particular, $\widehat{t}_{I,R}(c_{\text{new}})$ was the best estimator with a value of RE close to that of the gold standard $\widehat{t}_{I,R}(\widetilde{c})$. For the Lognormal distribution, all the proposed estimators were more efficient than the non-robust estimator for all the sample sizes. The robust estimator $\widehat{t}_{I,R}(c_{\text{new}})$ was the best with a value of RE close to that of the gold standard estimator $\widehat{t}_{I,R}(\widetilde{c})$. For the Weibull distribution, all the estimators showed a value of RE close to that of $\widehat{t}_{I,WLS}$.

Finally, in the case of the mixture distributions (see Table 4.3), the robust estimators exhibited substantial improvement over $\widehat{t}_{I,WLS}$. Both $\widehat{t}_{I,R}(1.345)$ and $\widehat{t}_{I,R}(c_{\text{new}})$ performed well, and outperformed $\widehat{t}_{I,CB}(c_{opt})$ by a significant margin. Again, in terms of efficiency, the robust estimator $\widehat{t}_{I,R}(c_{\text{new}})$ showed values of RE comparable to those obtained with the gold standard $\widehat{t}_{I,R}(\widetilde{c})$.

Estimator	n	Distribution		
		Normal	Student's t	Double Exponential (Laplace)
$\widehat{t}_{I,WLS}$	50	0.2 (100)	0.1 (100)	-0.2 (100)
	100	-0.2 (100)	-0.0 (100)	-0.0 (100)
	200	0.1 (100)	-0.1 (100)	0.1 (100)
$\widehat{t}_{I,R}(1.345)$	50	0.2 (103)	0.0 (69)	-0.2 (85)
	100	-0.2 (103)	-0.1 (68)	-0.0 (82)
	200	0.1 (102)	-0.1 (66)	0.1 (81)
$\widehat{t}_{I,CB}(c_{opt})$	50	-0.1 (101)	-0.1 (83)	-0.4 (93)
	100	-0.4 (100)	-0.2 (82)	-0.2 (93)
	200	-0.1 (100)	-0.2 (83)	-0.1 (95)
$\widehat{t}_{I,R}(c_{\text{new}})$	50	0.4 (102)	0.4 (75)	0.2 (90)
	100	-0.1 (101)	0.3 (73)	0.3 (90)
	200	0.1 (100)	0.1 (74)	0.2 (92)
$\widehat{t}_{I,R}(c^*)$	50	0.1 (106)	-0.0 (68)	-0.3 (82)
	100	-0.3 (105)	-0.1 (65)	-0.1 (80)
	200	0.0 (103)	-0.1 (64)	0.1 (80)
$\widehat{t}_{I,R}(\widetilde{c})$	50	0.2 (100)	0.1 (60)	-0.2 (76)
	100	-0.2 (100)	-0.1 (57)	0.0 (68)
	200	0.1 (100)	-0.0 (55)	0.2 (65)

Table 4.1: Monte Carlo percent relative bias and Monte Carlo relative efficiency (values in parentheses) of several estimators for symmetric distributions

Estimator	n	Distribution			
		Pareto	Frechet	Lognormal	Weibull
$\widehat{t}_{I,WLS}$	50	0.2 (100)	-0.3 (100)	-0.2 (100)	-0.3 (100)
	100	-0.1 (100)	0.1 (100)	0.2 (100)	0.1 (100)
	200	0.1 (100)	0.0 (100)	0.0 (100)	0.1 (100)
$\widehat{t}_{I,R}(1.345)$	50	-7.9 (50)	-7.8 (55)	-8.4 (79)	-7.0 (92)
	100	-8.6 (58)	-8.2 (69)	-8.9 (95)	-7.4 (111)
	200	-8.7 (93)	-8.5 (102)	-9.3 (134)	-7.7 (141)
$\widehat{t}_{I,CB}(c_{opt})$	50	-3.4 (70)	-3.8 (70)	-3.8 (88)	-3.3 (95)
	100	-3.1 (64)	-2.9 (73)	-2.7 (90)	-2.0 (98)
	200	-2.3 (69)	-2.2 (76)	-1.9 (94)	-1.2 (98)
$\widehat{t}_{I,R}(c_{\text{new}})$	50	-6.3 (50)	-6.6 (52)	-6.6 (83)	-5.9 (91)
	100	-5.5 (45)	-5.4 (60)	-5.1 (88)	-4.6 (103)
	200	-4.1 (54)	-4.1 (67)	-3.4 (94)	-3.0 (102)
$\widehat{t}_{I,R}(c^*)$	50	-6.9 (63)	-7.8 (59)	-7.8 (90)	-6.6 (95)
	100	-5.4 (52)	-5.4 (67)	-5.0 (92)	-3.8 (101)
	200	-4.0 (58)	-4.0 (70)	-3.3 (96)	-2.2 (100)
$\widehat{t}_{I,R}(\widetilde{c})$	50	-7.1 (50)	-6.5 (51)	-6.9 (83)	-2.8 (91)
	100	-4.8 (45)	-4.7 (58)	-3.9 (86)	-1.3 (98)
	200	-3.1 (53)	-3.0 (63)	-2.4 (91)	-0.5 (97)

Table 4.2: Monte Carlo percent relative bias and Monte Carlo relative efficiency (values in parentheses) of several estimators for asymmetric distributions

Estimator	n	Distribution			
		Mixture of Normal		Mixture of Lognormal	
		1% outliers	3% outliers	1% outliers	3% outliers
$\widehat{t}_{I,WLS}$	50	0.2 (100)	0.0 (100)	0.0 (100)	-0.1 (100)
	100	-0.1 (100)	-0.1 (100)	0.0 (100)	0.3 (100)
	200	0.0 (100)	0.1 (100)	-0.1 (100)	0.0 (100)
$\widehat{t}_{I,R}(1.345)$	50	-1.9 (50)	-5.8 (38)	-3.9 (30)	-9.4 (27)
	100	-2.2 (51)	-5.9 (45)	-4.1 (33)	-9.5 (36)
	200	-2.1 (53)	-5.9 (58)	-4.2 (43)	-9.6 (60)
$\widehat{t}_{I,CB}(c_{opt})$	50	-1.3 (74)	-3.2 (72)	-2.4 (61)	-5.0 (67)
	100	-1.5 (71)	-3.1 (73)	-2.4 (60)	-4.4 (69)
	200	-1.3 (73)	-2.5 (79)	-2.2 (64)	-3.9 (76)
$\widehat{t}_{I,R}(c_{\text{new}})$	50	-2.4 (53)	-7.2 (50)	-5.2 (29)	-12.6 (38)
	100	-2.6 (53)	-7.0 (55)	-4.9 (32)	-12.2 (48)
	200	-2.2 (58)	-5.9 (70)	-4.5 (42)	-11.1 (74)
$\widehat{t}_{I,R}(c^*)$	50	-3.0 (58)	-7.7 (66)	-5.9 (43)	-11.5 (67)
	100	-3.2 (57)	-7.0 (73)	-5.6 (39)	-10.9 (72)
	200	-2.9 (60)	-5.6 (83)	-4.9 (48)	-10.1 (78)
$\widehat{t}_{I,R}(\widetilde{c})$	50	-3.3 (46)	-10.0 (41)	-5.6 (26)	-15.0 (33)
	100	-3.4 (50)	-7.6 (54)	-5.0 (30)	-12.6 (50)
	200	-2.7 (55)	-5.4 (67)	4.1 (41)	-9.4 (71)

Table 4.3: Monte Carlo percent relative bias and Monte Carlo relative efficiency (values in parentheses) of several estimators for mixture distributions

Chapter 5

Estimation of the mean square error

In this chapter, using a first-order Taylor expansion, we derive an explicit expression of the mean square error of the robust estimator (3.2.10). We assume that both the tuning constant c and the standard deviation σ are fixed. The expression of the mean square error will then be used to estimate the mean square error of the proposed robust estimators $\widehat{t}_{I,R}(c_{\text{new}})$ and $\widehat{t}_{I,R}(c^*)$; see Sections 3.2.2 and 3.2.3, respectively. As an alternative to a Taylor expansion procedure, we also proposed a pseudo-population bootstrap procedure in Section 5.2.1 to estimate the mean square error of $\widehat{t}_{I,R}(c_{\text{new}})$ given by (5.1.13). In both Sections 5.1 and 5.2, we assume that the sampling fraction n/N is negligible. This assumption is not too restrictive as the sample size is often small compared to the population size in most practical situations.

5.1 First-order Taylor expansion

In Section 5.1.1, we first derive an estimator of the mean square error of $\widehat{t}_{I,R}(c_{\text{new}})$ and $\widehat{t}_{I,R}(c^*)$. In Section 5.1.2, we assess the performance of the proposed estimator of the mean square error in terms of bias.

5.1.1 Derivations

We start with the robust imputed estimator in the projection form:

$$\widehat{t}(\widehat{\mathbf{B}}_R, \sigma, c) = \sum_{i \in S} w_i \boldsymbol{v}_i^\top \widehat{\mathbf{B}}_R(c), \tag{5.1.1}$$

where $\widehat{\mathbf{B}}_R(c)$ is the solution of the following robust estimating equations

$$\widehat{U}(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i \in S_r} \psi_c \left(\frac{y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}}{\sigma \phi_i^{1/2}} \right) \frac{w_i \boldsymbol{v}_i}{\phi_i^{1/2}} = \mathbf{0}, \tag{5.1.2}$$

with $\psi_c(t)$ denoting the Huber function such that $\psi_c(t) = cI(t \geq c) + tI(-c \leq t \leq c) + (-c)I(t \leq -c)$. Let $\boldsymbol{\beta}^*$ be the probability limit of $\widehat{\mathbf{B}}_R(c)$. By using a first-order Taylor expansion, we obtain

$$\begin{aligned} \mathbf{0} &= \widehat{U}(\widehat{\mathbf{B}}_R(c)) \\ &= \widehat{U}(\boldsymbol{\beta}^*) + \frac{\partial \mathbb{E}(\widehat{U}(\boldsymbol{\beta}^*))}{\partial \boldsymbol{\beta}} \left(\widehat{\mathbf{B}}_R(c) - \boldsymbol{\beta}^*\right) + o_p(n^{-1/2}). \end{aligned} \tag{5.1.3}$$

In addition, it can be shown that

$$\frac{\partial \mathbb{E}(\widehat{U}(\boldsymbol{\beta}^*))}{\partial \boldsymbol{\beta}} = M_1 + M_2 + M_3, \tag{5.1.4}$$

where

$$M_1 = -\mathbb{E}\left\{ p(\boldsymbol{v}_i) c \frac{\boldsymbol{v}_i \boldsymbol{v}_i^\top}{\phi_i^{1/2}} f_{y|\boldsymbol{v}}(\boldsymbol{v}_i^\top \boldsymbol{\beta}^* + c\sigma\phi_i^{1/2}) \right\}, \tag{5.1.5}$$

$$\begin{aligned} M_2 &= \mathbb{E}\left\{ p(\boldsymbol{v}_i) \frac{\boldsymbol{v}_i \boldsymbol{v}_i^\top}{\phi_i^{1/2}} c f_{y|\boldsymbol{v}}(\boldsymbol{v}_i^\top \boldsymbol{\beta}^* + c\sigma\phi_i^{1/2}) \right\} + \mathbb{E}\left\{ p(\boldsymbol{v}_i) \frac{\boldsymbol{v}_i \boldsymbol{v}_i^\top}{\phi_i^{1/2}} c f_{y|\boldsymbol{v}}(\boldsymbol{v}_i^\top \boldsymbol{\beta}^* - c\sigma\phi_i^{1/2}) \right\} \\ &- \mathbb{E}\left\{ p(\boldsymbol{v}_i) \frac{\boldsymbol{v}_i \boldsymbol{v}_i^\top}{\sigma\phi_i} I(-c \leq \frac{y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}^*}{\sigma\phi^{1/2}} \leq c) \right\}, \end{aligned} \tag{5.1.6}$$

and

$$M_3 = -\mathbb{E}\left\{ p(\boldsymbol{v}_i) \frac{\boldsymbol{v}_i \boldsymbol{v}_i^\top}{\phi_i^{1/2}} c f_{y|\boldsymbol{v}}(\boldsymbol{v}_i^\top \boldsymbol{\beta}^* - c\sigma\phi_i^{1/2}) \right\}. \tag{5.1.7}$$

According to (5.1.4)-(5.1.7), we have

$$\frac{\partial \mathbb{E}(\widehat{U}(\boldsymbol{\beta}^*))}{\partial \boldsymbol{\beta}} = -\mathbb{E}\left\{ p(\boldsymbol{v}_i) \frac{\boldsymbol{v}_i \boldsymbol{v}_i^\top}{\sigma\phi_i} I(-c \leq \frac{y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}^*}{\sigma\phi_i^{1/2}} \leq c) \right\}. \tag{5.1.8}$$

According to (5.1.3), we have

$$\widehat{\mathbf{B}}_R(c) - \boldsymbol{\beta}^* = -\left\{ \frac{\partial \mathbb{E}(\widehat{U}(\boldsymbol{\beta}^*))}{\partial \boldsymbol{\beta}} \right\}^{-1} \widehat{U}(\boldsymbol{\beta}^*) + o_p(n^{-1/2}). \tag{5.1.9}$$

It follows that

$$\begin{aligned} \widehat{t}(\widehat{\mathbf{B}}_R, \sigma, c) &= \sum_{i \in S} w_i \boldsymbol{v}_i^\top \widehat{\mathbf{B}}_R(c) \\ &= \sum_{i \in S} w_i \boldsymbol{v}_i^\top (\boldsymbol{\beta}^* + \widehat{\mathbf{B}}_R(c) - \boldsymbol{\beta}^*) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i\in S} w_i \boldsymbol{v}_i^\top \boldsymbol{\beta}^* + \sum_{i\in S} w_i \boldsymbol{v}_i^\top (\widehat{\mathbf{B}}_R(c) - \boldsymbol{\beta}^*) \\
&= \sum_{i\in S} w_i \boldsymbol{v}_i^\top \boldsymbol{\beta}^* - \sum_{i\in S} w_i \boldsymbol{v}_i^\top \left\{ \frac{\partial\, \mathbb{E}(\widehat{U}(\boldsymbol{\beta}^*))}{\partial \boldsymbol{\beta}} \right\}^{-1} \widehat{U}(\boldsymbol{\beta}^*) + o_p(Nn^{-1/2}) \\
&= \sum_{i\in S} w_i \eta_i + o_p(Nn^{-1/2}), \qquad (5.1.10)
\end{aligned}
$$

where

$$
\eta_i = \boldsymbol{v}_i^\top \boldsymbol{\beta}^* - \left(\frac{1}{N} \sum_{i\in S} w_i \boldsymbol{v}_i^\top \right) \left\{ \frac{\partial\, \mathbb{E}(\widehat{U}(\boldsymbol{\beta}^*))}{\partial \boldsymbol{\beta}} \right\}^{-1} r_i \psi_c \left(\frac{y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}^*}{\sigma \phi_i^{1/2}} \right) \frac{\boldsymbol{v}_i}{\phi_i^{1/2}}. \qquad (5.1.11)
$$

The mean square error of $\widehat{t}(\widehat{\mathbf{B}}_R, \sigma, c)$ can thus be written as

$$
\begin{aligned}
MSE(\widehat{t}(\widehat{\mathbf{B}}_R, \sigma, c)) &= \left\{ \mathbb{E}(\widehat{t}(\widehat{\mathbf{B}}_R, \sigma, c)) - t_y \right\}^2 + \mathbb{V}\left\{ \widehat{t}(\widehat{\mathbf{B}}_R, \sigma, c) \right\} \\
&= \left(\sum_{i\in U} \eta_i - t_y \right)^2 + \sum_{i\in U} \sum_{j\in U} (\pi_{ij} - \pi_i \pi_j) \frac{\eta_i}{\pi_i} \frac{\eta_j}{\pi_j} \\
&+ o(N^2/n). \qquad (5.1.12)
\end{aligned}
$$

It follows that an estimator of the estimated mean square error is given by

$$
\begin{aligned}
\widehat{MSE}(\widehat{t}(\widehat{\mathbf{B}}_R, \sigma, c)) &= \max\left\{ \left(\widehat{t}(\widehat{\mathbf{B}}_R, \sigma, c) - \widehat{t}_{I,WLS} \right)^2 - \widehat{\mathbb{V}}\left(\widehat{t}(\widehat{\mathbf{B}}_R, \sigma, c) - \widehat{t}_{I,WLS} \right), 0 \right\} \\
&+ \widehat{\mathbb{V}}\left\{ \widehat{t}(\widehat{\mathbf{B}}_R, \sigma, c) \right\}. \qquad (5.1.13)
\end{aligned}
$$

In addition, it can be shown that

$$
\begin{aligned}
\widehat{t}_{I,WLS} &= \sum_{i\in S_r} w_i y_i + \sum_{i\in S_m} w_i \boldsymbol{v}_i^\top \widehat{\mathbf{B}}_{WLS} \\
&= \sum_{i\in S_r} w_i y_i + \sum_{i\in S_m} w_i \boldsymbol{v}_i^\top \boldsymbol{\beta}^*_{WLS} + \sum_{i\in S_m} w_i \boldsymbol{v}_i^\top \left(\widehat{\mathbf{B}}_{WLS} - \boldsymbol{\beta}^*_{WLS} \right) \\
&= \sum_{i\in S_r} w_i y_i + \sum_{i\in S_m} w_i \boldsymbol{v}_i^\top \boldsymbol{\beta}^*_{WLS} + \sum_{i\in S_m} w_i \boldsymbol{v}_i^\top \left(\sum_{i\in S_r} w_i \boldsymbol{v}_i \phi_i^{-1} \boldsymbol{v}_i^\top \right)^{-1} \sum_{i\in S_r} w_i \boldsymbol{v}_i \phi_i^{-1} \left(y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}^*_{WLS} \right) \\
&= \sum_{i\in S} w_i \tau_i, \qquad (5.1.14)
\end{aligned}
$$

where $\boldsymbol{\beta}^*_{WLS}$ is the probability limit of $\boldsymbol{\beta}^*_{WLS}$,

$$
\tau_i = r_i y_i + (1 - r_i) \boldsymbol{v}_i^\top \boldsymbol{\beta}^*_{WLS} + A r_i \boldsymbol{v}_i \phi_i^{-1} \left(y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}^*_{WLS} \right), \qquad (5.1.15)
$$

and

$$A = \sum_{i \in S_m} w_i \boldsymbol{v}_i^\top \left(\sum_{i \in S_r} w_i \boldsymbol{v}_i \phi_i^{-1} \boldsymbol{v}_i^\top \right)^{-1}.$$

Therefore, we have

$$\widehat{V}\left(\widehat{t}(\widehat{\mathbf{B}}_R, \sigma, c) - \widehat{t}_{I,WLS}\right) = \sum_{i \in S} \sum_{j \in S} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{\widehat{\eta}_i - \widehat{\tau}_i}{\pi_i} \frac{\widehat{\eta}_j - \widehat{\tau}_j}{\pi_j}, \quad (5.1.16)$$

where

$$\widehat{\eta}_i = \boldsymbol{v}_i^\top \widehat{\mathbf{B}}_R(c) - \left(\sum_{i \in S} w_i \boldsymbol{v}_i^\top \right) \left\{ N \frac{\partial \widehat{\mathbb{E}}(\widehat{U}(\boldsymbol{\beta}^*))}{\partial \boldsymbol{\beta}} \right\}^{-1} r_i \psi_c \left(\frac{y_i - \boldsymbol{v}_i^\top \widehat{\mathbf{B}}_R(c)}{\widehat{\sigma} \phi_i^{1/2}} \right) \frac{\boldsymbol{v}_i}{\phi_i^{1/2}}, \quad (5.1.17)$$

$$\frac{\partial \widehat{\mathbb{E}}(\widehat{U}(\boldsymbol{\beta}^*))}{\partial \boldsymbol{\beta}} = -\frac{1}{N} \sum_{i \in S_r} w_i \frac{\boldsymbol{v}_i \boldsymbol{v}_i^\top}{\widehat{\sigma} \phi_i} I(-c \leq \frac{y_i - \boldsymbol{v}_i^\top \widehat{\mathbf{B}}_R(c)}{\widehat{\sigma} \phi_i^{1/2}} \leq c), \quad (5.1.18)$$

and

$$\widehat{\tau}_i = r_i y_i + (1 - r_i) \boldsymbol{v}_i^\top \widehat{\mathbf{B}}_{WLS} + A r_i \boldsymbol{v}_i \phi_i^{-1} \left(y_i - \boldsymbol{v}_i^\top \widehat{\mathbf{B}}_{WLS} \right). \quad (5.1.19)$$

In addition, we have

$$\widehat{V}\left\{\widehat{t}(\widehat{\mathbf{B}}_R, \sigma, c)\right\} = \sum_{i \in S} \sum_{j \in S} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{\widehat{\eta}_i}{\pi_i} \frac{\widehat{\eta}_j}{\pi_j}. \quad (5.1.20)$$

As a result, the estimator of mean square error, $\widehat{MSE}(\widehat{t}(\widehat{\mathbf{B}}_R, \sigma, c))$, is given by (5.1.13) with $\widehat{\mathbb{V}}\left\{\widehat{t}(\widehat{\mathbf{B}}_R, \sigma, c)\right\}$ given by (5.1.20) and $\widehat{\mathbb{V}}\left(\widehat{t}(\widehat{\mathbf{B}}_R, \sigma, c) - \widehat{t}_{I,WLS}\right)$ given by (5.1.16).

5.1.2 Performance of the estimator of the mean square error based on a first-order Taylor expansion

In this section, we present the results of a simulation study whose goal was to assess the performance of the estimator of (5.1.13) in terms of bias. We used (5.1.13) to estimate the mean square error of the two proposed robust estimators $\widehat{t}_{I,R}(c_{\text{new}})$ and $\widehat{t}_{I,R}(c^*)$.

We repeated $R = 10,000$ iterations of the following process:

The steps (1), (2), and (3) are the same as those described in Chapter 4; see Section 4.1. The next steps are described next.

(4) In each sample, we computed: (i) the non-robust estimator $\widehat{t}_{I,WLS}$ given by (3.2.1); (ii) the robust estimator $\widehat{t}_{I,R}(c_{\text{new}})$ given by (3.2.8); and (iii) the robust estimator $\widehat{t}_{I,R}(c^*)$ given by (3.2.12).

(5) In each sample, we computed the estimator of the mean square error given by (5.1.13) to estimate the mean square error of both $\widehat{t}_{I,R}(c_{\text{new}})$ and $\widehat{t}_{I,R}(c^*)$.

We computed the Monte Carlo percent relative bias of (5.1.13) defined as

$$RB_{MC}\{\widehat{MSE}(\widehat{t}_I)\} = \frac{\mathbb{E}_{MC}\{\widehat{MSE}(\widehat{t}_I)\} - \text{MSE}_{MC}(\widehat{t}_I)}{\text{MSE}_{MC}(\widehat{t}_I)} \times 100, \tag{5.1.21}$$

where

$$\mathbb{E}_{MC}(\widehat{MSE}(\widehat{t}_I)) = \frac{1}{R}\sum_{r=1}^{R} \widehat{MSE}(\widehat{t}_I^{(r)}),$$

and

$$\text{MSE}_{MC}(\widehat{t}_I) = \frac{1}{R}\sum_{r=1}^{R} \left(\widehat{t}_I^{(r)} - t_y\right)^2.$$

with $\widehat{t}_I$ denotes either $\widehat{t}_{I,R}(c_{\text{new}})$ and $\widehat{t}_{I,R}(c^*)$.

Table 5.1 displays the Monte Carlo percent relative bias of the mean square error estimator for the two robust estimators $\widehat{t}_{I,R}(c_{\text{new}})$ and $\widehat{t}_{I,R}(c^*)$. We first discuss the results pertaining to $\widehat{t}_{I,R}(c_{\text{new}})$. For symmetric distributions, the results suggest that the estimator of mean square error performed well for all sample sizes n, with values of absolute RB ranging from 0.4% to 7.4%. For highly asymmetric distributions such as Pareto and Frechet, the estimator of MSE exhibited relatively large biases for small values of n. The bias decreased as n increased. For instance, for the Pareto distribution, the RB was equal to 62.6% for $n = 50$, 21.4% for $n = 100$, and 9.0% for $n = 200$. For the other two asymmetric distributions, Lognormal and Weibull, the estimator of MSE showed relatively small values of absolute RB ranging from 0.1% to 10.4%. Finally, for mixture distributions, the estimator of MSE exhibited large biases, in general, especially for the Mixture of Lognormals with 1% and 3% outliers. In this case, the values of absolute RB ranged from 39.1 % to 118.7 %. Again, the bias decreased as the sample size increased.

We now discuss the results pertaining to $\widehat{t}_{I,R}(c^*)$. We note that the bias was negative for all cases except for the Mixture of Lognormals with 1% and 3% outliers. This may be explained by the fact that the estimated mean square error obtained through a first-order Taylor expansion was derived by assuming that the tuning constant c was fixed. In practice, the tuning constant is sample-dependent; i.e., it varies from one sample to another. Ignoring the variability associated with the estimation

of the tuning constant is most likely responsible for the underestimation of the true mean square error.

	Distribution	n	$\widehat{MSE}(\widehat{t}_{I,R}(c_{\text{new}}))$ RB	$\widehat{MSE}(\widehat{t}_{I,R}(c^*))$ RB
Symmetric	Normal	50	-7.4	-27.6
		100	-3.8	-18.6
		200	-6.6	-13.2
	Student's t	50	4.0	-25.2
		100	0.4	-18.2
		200	-1.2	-20.3
	Double Exponential (Laplace)	50	0.7	-29.0
		100	-0.5	-25.1
		200	-1.1	-21.4
Asymmetric	Pareto	50	62.6	-18.3
		100	21.4	-9.0
		200	9.0	-14.5
	Frechet	50	39.8	-25.3
		100	11.9	-15.4
		200	2.8	-15.3
	Lognormal	50	0.1	-29.0
		100	-6.9	-22.1
		200	-10.4	-17.4
	Weibull	50	-5.3	-28.2
		100	-9.5	-20.0
		200	-7.7	-17.3
Mixture	Mixture of Normals (1 %)	50	25.2	-17.1
		100	24.0	-15.0
		200	10.2	-14.6
	Mixture of Normals (3 %)	50	43.5	-34.5
		100	38.4	-30.3
		200	10.0	-22.5
	Mixture of Lognormals (1%)	50	118.7	-16.9
		100	87.7	13.6
		200	58.1	10.3
	Mixture of Lognormals (3%)	50	87.5	-35.2
		100	77.7	-20.0
		200	39.1	6.2

Table 5.1: Monte Carlo percent relative bias of the estimator of mean square error of $\widehat{t}_{I,R}(c_{\text{new}})$ and $\widehat{t}_{I,R}(c^*)$ for several distributions

5.2 Bootstrap

In this section, we propose a bootstrap procedure for estimating the mean square error of $\widehat{t}_{I,R}(c_{\text{new}})$. In finite population sampling, several bootstrap approaches have been proposed in the last four decades. Essentially, the bootstrap procedures may be classified into three broad categories: (i) the bootstrap weight procedures; (ii) the direct bootstrap procedures; and (iii) the pseudo-population bootstrap procedures; see Mashreghi et al. (2016) for a comprehensive overview of bootstrap procedures in survey sampling. In this section, we adopt the third approach, namely the pseudo-population bootstrap approach. To illustrate the rationale behind pseudo-population bootstrap procedures, consider the case of a simple random sample without replacement of size $n = 100$ selected from a population U of size $N = 1,000$ and let us assume that the sample observations are all recorded (i.e., there are no missing values). In this case, the sampling weight of unit i, defined as the inverse of its first-order inclusion probability $\pi_i = 100/1,000$, $i = 1, \ldots, N$, is equal to $N/n = 10$. A pseudo-population is created by replicating each sample unit 10 times. Then, from the pseudo-population, B bootstrap samples are selected using the same sampling design that was utilized to select the original sample (here, simple random sampling without replacement). If N/n is not an integer, we replicate each unit $\lfloor N/n \rfloor$ times and then complete the pseudo-population by selecting a simple random sample without replacement sample of size $N - \{n \times \lfloor N/n \rfloor\}$, from the original sample S in order to obtain a pseudo-population of size N. This procedure can be extended to unequal probability sampling designs in a relatively straightforward fashion; e.g., see Mashreghi et al. (2016).

In the presence of imputed data, applying a complete data pseudo-population bootstrap will typically lead to an underestimation of the true variance of the imputed estimator as it amounts to treating the imputed values as observed values. Shao and Sitter (1996) proposed a bootstrap procedure that involves reimputing the missing values in each bootstrap sample using the same imputation that was utilized to impute the missing values in the original sample. As argued in Haziza and Vallée (2020), the Shao-Sitter bootstrap variance estimator can be justified through the so-called reverse approach of Fay (1991) and Shao and Steel (1999). In Section 5.2.1, we propose a pseudo-population bootstrap procedures that account for both the missing and influential values. The performance of the proposed procedure is assessed through a limited simulation study in Section 5.2.2.

5.2.1 Reweighted Pseudo-Population Bootstrap

In this section, we describe a Reweighted Pseudo-Population Bootstrap (RPPB) procedure. The proposed procedure consists of creating a pseudo-population by repli-

cating the sample units as described above. However, instead of replicating each unit N/n times (assuming N/n is an integer), we first assign a lower sampling weight to units that were identified as influential using a method described below. Indeed, replicating influential units using the original weight N/n times, may lead to a pseudo-population that contains too many influential units, which would potentially lead to estimates of the mean square error that are too large.

The proposed RPPB algorithm is described below. For simplicity, we confine to the case of simple random sampling without replacement but the extension to unequal probability sampling designs is relatively straightforward.

RPPB Algorithm:

1. Using the set of responding units, S_r, in the original sample S, find $\widehat{\boldsymbol{B}}_R(c)$ by solving (2.1.3), and compute the set of scaled residuals $\{\widetilde{e}_i\}_{i\in S_r}$, where $\widetilde{e}_i = e_i/\widehat{s}$. Here, $e_i = y_i - \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_R(c)$ and $\widehat{s}$ is a robust estimator for the residual standard deviation, such as the median absolute deviation (MAD).

2. Construct the pseudo-population as follows: For $i \in S_r$, compute initial weights $\widetilde{w}_i$ given by

$$\widetilde{w}_i = d_i \times \frac{\psi_c(\widetilde{e}_i)}{\widetilde{e}_i} = \begin{cases} d_i & \text{if } |\widetilde{e}_i| < c \\ d_i \frac{c}{|\widetilde{e}_i|} & \text{if } |\widetilde{e}_i| \geq c \end{cases}$$

where $d_i = N/n$ and $\psi_c(\cdot)$ denotes the Huber influence function with tuning constant c (see Table 2.1). For a non-responding unit $i \in S_m$, we do not modify the original weight so that $\widetilde{w}_i = N/n$. We then define the weights w_i^* as

$$w_i^* = \delta_i \widetilde{w}_i + (1-\delta_i)\widehat{\gamma}\widetilde{w}_i, \quad i \in S, \tag{5.2.1}$$

where $\delta_i = 1$ if unit i was trimmed, i.e., $\psi_c(\widetilde{e}_i) < \widetilde{e}_i$, and $\delta_i = 0$, otherwise. The scaling factor $\widehat{\gamma}$ in (5.2.1) ensures that $\sum_{i\in S} w_i^* = N$. In other words, the sum of the weights matches the population size, which is a desirable property. It can be shown that the scaling factor $\widehat{\gamma}$ is given by

$$\widehat{\gamma} = \frac{N - \sum_{i\in S}\delta_i \widetilde{w}_i}{\sum_{i\in S}(1-\delta_i)\,\widetilde{w}_i}.$$

Once we find the weights w_i^* for all $i \in S$, make $\lfloor w_i^* \rfloor$ copies of unit i to construct a partial pseudo-population $\widetilde{U}^f$. To complete the pseudo-population, we take a

sample $\widetilde{U}^{c*}$ by applying Poisson sampling with first-order inclusion probabilities, $w_i^* - \lfloor w_i^* \rfloor$, for $i \in S$. By doing this, we obtain a reweighted pseudo-population $\widetilde{U}^* = \widetilde{U}^f \cup \widetilde{U}^{c*} = \{(y_i^*, v_i^*, w_i^*, r_i^*)\}_{i=1}^{\widetilde{N}^*}$ of size $\widetilde{N}^*$.

3. Take a bootstrap sample $\widetilde{S}^* = \{(y_i^*, v_i^*, w_i^*, r_i^*)\}_{i=1}^{\widetilde{n}^*}$ from $\widetilde{U}^*$ using the original sampling design, here simple random sampling without replacement. The set of respondents $\widetilde{S}_r^*$ are those units with $r_i^* = 1$ and the set of nonrespondents $\widetilde{S}_m^*$ are those with $r_i^* = 0$.

4. Impute the missing values in the bootstrap sample $\widetilde{S}^*$ using the same imputation procedure utilized for imputing missing values in the original sample S. Compute the imputed estimator $\widehat{t}_I^*$.

5. Repeat steps 3 and 4 B times to get the imputed estimates $\widehat{t}_I^{*(1)}, \ldots, \widehat{t}_I^{*(B)}$. This leads to bootstrap variance estimator

$$\widehat{V}_B^* = \frac{1}{B-1} \sum_{b=1}^{B} \left(\widehat{t}_I^{*(b)} - \widehat{t}_I^{*(\cdot)} \right)^2,$$

where $\widehat{t}_I^{*(\cdot)} = B^{-1} \sum_{b=1}^{B} \widehat{t}_I^{*(b)}$.

5.2.2 Performance of the estimator of the mean square error based on the RPPB procedure

In this section, we present the results of a simulation study to assess the empirical performance of the estimator of the mean square error (5.1.13), in terms of bias, in the case of $\widehat{t}_{I,R}(c_{\text{new}})$, when the variance components $V\left(\widehat{t}_{I,R}(c_{\text{new}}) - \widehat{t}_{I,WLS}\right)$ and $V\left(\widehat{t}_{I,R}(c_{\text{new}})\right)$ are estimated using the the RPPB algorithm presented in Section 5.2.1.

First, a finite population U of size N was generated in the same way as described in Step (1) of the simulation study presented in Chapter 4. Then, we repeated $A = 5,000$ iterations of the following process:

1) From the finite population U, select a sample, of size $n \in \{50, 100, 200\}$, according to simple random sampling without replacement (SRSWOR). Nonresponse to the y-variable was generated in the same way as described in Step (3) of the simulation study presented in Chapter 4.

2) Apply the Steps 1-5 of the *RPPB Algorithm* described in Section (5.2.1) to obtain the variance components in (5.1.13). We were interested in estimating the population total, i.e., $t_y = \sum_{i \in U} y_i$. In step 5 of the *RPPB Algorithm*, we set $B = 500$.

3) Compute the mean square error estimate, $\widehat{MSE^*}(\widehat{t}_{I,R}(c_{\text{new}}))$ using the two estimated variance terms $\widehat{V}_B^*\left(\widehat{t}_{I,R}(c_{\text{new}})\right)$ and $\widehat{V}_B^*\left(\widehat{t}_{I,R}(c_{\text{new}}) - \widehat{t}_{I,WLS}\right)$ obtained in step 5 of the *SRSWOR RPPB Algorithm.*

To assess the bias of the estimator of mean square error of $\widehat{t}_{I,R}(c_{\text{new}})$, we computed its Monte Carlo percent relative bias, defined as

$$RB_{MC}\{\widehat{MSE^*}(\widehat{t}_{I,R}(c_{\text{new}}))\} = \frac{\mathbb{E}_{MC}\{\widehat{MSE^*}(\widehat{t}_{I,R}(c_{\text{new}}))\} - \text{MSE}_{MC}(\widehat{t}_{I,R}(c_{\text{new}}))}{\text{MSE}_{MC}(\widehat{t}_{I,R}(c_{\text{new}}))} \times 100.$$

Table 5.2 displays the Monte Carlo relative bias (in %) for the mean square error estimator of $\widehat{t}_{I,R}(c_{\text{new}})$. For symmetric distributions, the proposed bootstrap procedure performed relatively well with values of absolute RB ranging from 0.7% to 12.6%. These results were close to those obtained through a first-order Taylor expansion procedure (see Table 5.1). For highly asymmetric distributions such as Pareto and Frechet, the estimator of MSE showed values of RB under 20% in absolute values for any sample size n, which was an improvement over the results obtained through a first-order Taylor expansion procedure. For the other two asymmetric distributions, Lognormal and Weibull, the estimator of MSE showed some biases ranging from 0.4% to 26.7% in absolute value. Finally, for mixture distributions, the estimator of MSE showed much better results compared to those obtained through a first-order Taylor expansion procedure (see Table 5.1), especially in the case of the mixture of Lognormals with 1% and 3% outliers.

	Distribution	$\widehat{MSE}(\widehat{t}_{I,R}(c_{\text{new}}))$ n		
		50	100	200
Symmetric	Normal	9.6	2.6	-5.1
	Student's t	0.7	-10.6	-12.6
	Double Exponential (Laplace)	1.1	-7.5	-11.4
Asymmetric	Pareto	-15.1	-15.1	4.8
	Frechet	-13.6	-19.5	-17.5
	Lognormal	-12.3	-24.6	-20.2
	Weibull	0.4	-13.4	26.7
Mixture	Mixture of Normals (1 %)	11.5	-1.8	-12.8
	Mixture of Normals (3%)	1.1	-3.3	-16.8
	Mixture of Lognormals (1 %)	34.0	27.7	19.2
	Mixture of Lognormals (3 %)	23.5	23.7	4.1

Table 5.2: Monte Carlo percent relative bias (RB) of the estimator of mean square error of $\widehat{t}_{I,R}(c_{\text{new}})$ for several distributions

Chapter 6

Conclusion

In this thesis, we considered the problem of item nonresponse in the presence of influential units in surveys. The classical non-robust imputed estimator of a population total is (approximately) unbiased provided that the first moment of the imputation model is correctly specified, but it may suffer from a large variance. We examined the performance of two commonly employed methods for the treatment of influential values at the imputation stage. Although these methods behaved very well in terms of bias and efficiency in the case of symmetric distributions, they tend to perform poorly in the case of asymmetric distributions. To address this problem, we introduced three robust imputed estimators that are all based on an adaptative tuning constant. The first two rely on the concept of conditional bias, and the third is based on an optimal tuning constant c that minimizes the estimated mean square error of the robust imputed estimator (3.2.10).

We conducted simulation studies to assess the empirical performance of the proposed imputed estimators and compared them to the non-robust and the naive estimators. The results showed that the performance of naive estimators, based on a non-adaptative tuning constant $c = 1.345$, deteriorated as the sample size increased in the case of asymmetric distributions. In general, the proposed imputed estimators performed well in almost all scenarios, showing some significant gains compared to non-robust estimator. Among the three robust estimators, $\widehat{t}_{I,R}(c^*)$ performed the best in terms of relative efficiency for symmetric outliers, whereas the estimator $\widehat{t}_{I,R}(c_{\text{new}})$ was generally the best for asymmetric distributions.

In this thesis, we focused on linear regression imputation, which implicitly assumed that the survey variable y was continuous. In practice, the y-variable may be binary or may correspond to a count. In this case, generalized linear models are generally more appropriate; e.g., Logistic regression model, and Poisson model. Therefore, it would be useful to extend our results to the case of GLM. Also, if the

relationship between y and v is complex, it may be wise to employ a non-parametric model (e.g., spline regression, local polynomial regression, etc.). Again, it would be useful to extend our results to the case of nonparametric models.

Finally, we studied the problem of mean square error estimation, which is a challenging problem, even in the absence of missing values. The procedure based on a Taylor expansion generally led to some underestimation in the case of $\widehat{t}_{I,R}(c^*)$. The pseudo-population bootstrap performed much better than the procedure based on a Taylor expansion, but remained unsatisfactory for some scenarios (e.g., mixture of Lognormal distributions). More work is needed to improve on the proposed bootstrap procedure.

References

Andersen, R. (2008). *Modern methods for robust regression*. Number 152. Sage.

Ayinde, K., Lukman, A. F., and Arowolo, O. (2015). Robust regression diagnostics of influential observations in linear regression model. *Open Journal of Statistics*, 5:273–283.

Beaumont, J.-F., Haziza, D., and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100:555–569.

Chatterjee, S. and Hadi, A. S. (1988). Impact of simultaneous omission of a variable and an observation on a linear regression equation. *Computational Statistics & Data Analysis*, 6:129–144.

Chen, S., Haziza, D., and Michal, V. (2020). Efficient multiply robust imputation in the presence of influential units in surveys. *Canadian Journal of Statistics*.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19:15–18.

Deville, J.-C. and Särndal, C.-E. (1994). Variance estimation for the regression imputed horvitz-thompson estimator. *Journal of Official Statistics*, 10:381–394.

Fay, R. E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference, US Census Bureau*, pages 429–440.

Haziza, D. and Vallée, A.-A. (2020). Variance estimation procedures in the presence of singly imputed survey data: A critical review. *Japanese Journal of Statistics and Data Science*, 3:583–623.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47:663–685.

Huber, P. (1981). *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley.

Mashreghi, Z., Haziza, D., and Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, 10:1–52.

Moreno-Rebollo, J., Muñoz-Reyes, A., and Muñoz-Pichardo, J. (1999). Miscellanea. influence diagnostic in survey sampling: conditional bias. *Biometrika*, 86:923–928.

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3:169–175.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63:581–592.

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey methodology*, 18:241–252.

Shao, J. and Sitter, R. R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91:1278–1288.

Shao, J. and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94:254–265.

Appendix A

Total variance estimator $\widehat{V}_{tot}$ under mean imputation procedure and SRSWOR

Under Särndal's method, the total variance estimate $\widehat{V}_{tot}$ is given by

$$\widehat{V}_{tot} = \widehat{V}_{sam} + \widehat{V}_{nr} + 2\widehat{V}_{mix}. \tag{A.0.1}$$

To obtain the sampling variance estimator $\widehat{V}_{sam}$, we have

$$\begin{aligned}
\widehat{V}_{HT} &= \sum_{i\in S}\sum_{j\in S}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j} \\
&= \sum_{i\in S}\frac{\pi_i(1-\pi_i)}{\pi_i}\frac{y_i^2}{\pi_i^2} + \sum_{\substack{i\in S\\ i\neq j}}\sum_{j\in S}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j} \\
&= \left(\frac{N^2}{n^2}-\frac{N}{n}\right)\sum_{i\in S}y_i^2 + \left(\frac{N^2}{n^2}-\frac{N(N-1)}{n(n-1)}\right)\sum_{\substack{i\in S\\ i\neq j}}\sum_{j\in S}y_iy_j \\
&= N^2\left(\frac{1}{n}-\frac{1}{N}\right)\left(\frac{1}{n}\sum_{i\in S}y_i^2 - \frac{1}{n(n-1)}\sum_{\substack{i\in S\\ i\neq j}}\sum_{j\in S}y_iy_j\right) \\
&= N^2\left(\frac{1}{n}-\frac{1}{N}\right)\left(\frac{(n-1)}{n(n-1)}\sum_{i\in S}y_i^2 - \frac{1}{n(n-1)}\sum_{\substack{i\in S\\ i\neq j}}\sum_{j\in S}y_iy_j\right)
\end{aligned}$$

$$
\begin{aligned}
&= N^2\left(\frac{1}{n}-\frac{1}{N}\right)\left\{\frac{1}{n-1}\sum_{i\in S}y_i^2-\frac{1}{n(n-1)}\left(\sum_{i\in S}y_i^2+\sum_{\substack{i\in S\\ i\neq j}}\sum_{j\in S}y_iy_j\right)\right\}\\
&= N^2\left(\frac{1}{n}-\frac{1}{N}\right)\left\{\frac{1}{n-1}\left(\sum_{i\in S}y_i^2-\frac{1}{n}\left(\sum_{i\in S}y_i\right)^2\right)\right\}\\
&= N^2\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{n-1}\left(\sum_{i\in S}y_i^2-n\bar{y}^2\right)\\
&= N^2\left(\frac{1}{n}-\frac{1}{N}\right)s_y^2,
\end{aligned}
$$

where $s_y^2=\frac{1}{n-1}\sum_{i\in S}\left(y_i-\bar{y}\right)^2$.

Then,

$$
\begin{aligned}
\mathbb{E}_m\left(\widehat{V}_{HT}\right) &= \mathbb{E}_m\left\{N^2\left(\frac{1}{n}-\frac{1}{N}\right)s_y^2\right\}\\
&= N^2\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{n-1}\,\mathbb{E}_m\left\{\sum_{i\in S}y_i^2-\frac{1}{n}\left(\sum_{i\in S}y_i^2+\sum_{\substack{i\in S\\ i\neq j}}\sum_{j\in S}y_iy_j\right)\right\}\\
&= N^2\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{n-1}\left\{\sum_{i\in S}\mathbb{E}_m\left(y_i^2\right)-\frac{1}{n}\sum_{i\in S}\mathbb{E}_m\left(y_i^2\right)-\frac{1}{n}\sum_{\substack{i\in S\\ i\neq j}}\sum_{j\in S}\mathbb{E}_m\left(y_iy_j\right)\right\}\\
&= N^2\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{n-1}\left\{\frac{n-1}{n}\sum_{i\in S}\left(\beta^2+\sigma^2\right)-\frac{1}{n}\sum_{\substack{i\in S\\ i\neq j}}\sum_{j\in S}\beta^2\right\}\\
&= N^2\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{n-1}\left\{(n-1)\beta^2+(n-1)\sigma^2-(n-1)\beta^2\right\}\\
&= N^2\left(\frac{1}{n}-\frac{1}{N}\right)\sigma^2.
\end{aligned}
$$

The fourth equality follows from the fact that

$$
\begin{aligned}
\mathbb{E}_m\left(y_i^2\right) &= \mathbb{E}_m\left\{(\beta+\epsilon_i)^2\right\}\\
&= \beta^2+2\beta\,\mathbb{E}_m\left(\epsilon_i\right)+\mathbb{E}_m\left(\epsilon_i^2\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \beta^2 + \left\{\mathbb{V}_m\left(\epsilon_i\right) + \mathbb{E}_m\left(\epsilon_i\right)^2\right\} \\
&= \beta^2 + \sigma^2
\end{aligned}
$$

and for $i \neq j$,

$$
\begin{aligned}
\mathbb{E}_m\left(y_i y_j\right) &= Cov_m\left(y_i, y_j\right) + \mathbb{E}_m\left(y_i\right)\mathbb{E}_m\left(y_j\right) \\
&= Cov_m\left(\beta + \epsilon_i, \beta + \epsilon_j\right) + \mathbb{E}_m\left(\beta + \epsilon_i\right)\mathbb{E}_m\left(\beta + \epsilon_j\right) \\
&= Cov_m\left(\epsilon_i, \epsilon_j\right) + \left\{\beta + \mathbb{E}_m\left(\epsilon_i\right)\right\}\left\{\beta + \mathbb{E}_m\left(\epsilon_j\right)\right\} \\
&= \mathbb{E}_m\left(\epsilon_i \epsilon_j\right) - \mathbb{E}_m\left(\epsilon_i\right)\mathbb{E}_m\left(\epsilon_j\right) + \beta^2 \\
&= \beta^2.
\end{aligned}
$$

Also, we have

$$
\begin{aligned}
\widehat{V}_{naive} &= \sum_{i\in S}\sum_{j\in S}\frac{\left(\pi_{ij} - \pi_i\pi_j\right)}{\pi_{ij}}\frac{\tilde{y}_i}{\pi_i}\frac{\tilde{y}_j}{\pi_j} \\
&= \sum_{i\in S}\frac{\pi_i(1-\pi_i)}{\pi_i}\frac{\tilde{y}_i^2}{\pi_i^2} + \sum_{\substack{i\in S\\ i\neq j}}\sum_{j\in S}\frac{\left(\pi_{ij} - \pi_i\pi_j\right)}{\pi_{ij}}\frac{\tilde{y}_i}{\pi_i}\frac{\tilde{y}_j}{\pi_j} \\
&= \left(\frac{N^2}{n^2} - \frac{N}{n}\right)\sum_{i\in S}\tilde{y}_i^2 + \left(\frac{N^2}{n^2} - \frac{N(N-1)}{n(n-1)}\right)\sum_{\substack{i\in S\\ i\neq j}}\sum_{j\in S}\tilde{y}_i\tilde{y}_j \\
&= N^2\left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n-1}\left\{\sum_{i\in S}\tilde{y}_i^2 - \frac{1}{n}\left(\sum_{i\in S}\tilde{y}_i^2 + \sum_{\substack{i\in S\\ i\neq j}}\sum_{j\in S}\tilde{y}_i\tilde{y}_j\right)\right\} \\
&= N^2\left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n-1}\left\{\sum_{i\in S}\tilde{y}_i^2 - \frac{1}{n}\left(\sum_{i\in S}\tilde{y}_i\right)^2\right\} \\
&= N^2\left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n-1}\left\{\sum_{i\in S} r_i y_i^2 + \sum_{i\in S}(1-r_i)\bar{y}_r^2 - \frac{1}{n}\left(\sum_{i\in S} r_i y_i\right)^2\right. \\
&\quad \left. - \frac{2}{n}\sum_{i\in S} r_i y_i \sum_{i\in S}(1-r_i)\bar{y}_r - \frac{1}{n}\left(\sum_{i\in S}(1-r_i)\right)^2\bar{y}_r^2\right\} \\
&= N^2\left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n-1}\left\{\sum_{i\in S} r_i y_i^2 + n\bar{y}_r^2 - n_r\bar{y}_r^2 - \frac{\bar{y}_r^2}{n}\left(n_r^2 + 2(n-n_r)n_r + (n-n_r)^2\right)\right\} \\
&= N^2\left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n-1}\left\{\sum_{i\in S} r_i y_i^2 + n\bar{y}_r^2 - n_r\bar{y}_r^2 - \frac{\bar{y}_r^2}{n}\left(n^2\right)\right\}
\end{aligned}
$$

$$= N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{n-1} \left\{ \sum_{i \in S} r_i y_i^2 - n_r \bar{y}_r^2 \right\}$$
$$= N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \frac{n_r - 1}{n-1} s_{yr}^2,$$

where $s_{yr}^2 = \frac{1}{n_r - 1} \sum_{i \in S_r} (y_i - \bar{y}_r)^2$.

Then,

$$\begin{aligned}
\mathbb{E}_m \left(\widehat{V}_{naive}\right) &= \mathbb{E}_m \left\{ N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \frac{n_r - 1}{n-1} s_{yr}^2 \right\} \\
&= N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{n-1} \mathbb{E}_m \left(\sum_{i \in S} r_i y_i^2 - n_r \bar{y}_r^2 \right) \\
&= N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{n-1} \mathbb{E}_m \left\{ \sum_{i \in S_r} y_i^2 - \frac{1}{n_r} \left(\sum_{i \in S_r} y_i^2 + \sum_{\substack{i \in S_r \\ i \neq j}} \sum_{j \in S_r} y_i y_j \right) \right\} \\
&= N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{n-1} \left\{ \frac{n_r - 1}{n_r} \sum_{i \in S_r} \mathbb{E}_m \left(y_i^2\right) - \frac{1}{n_r} \sum_{\substack{i \in S_r \\ i \neq j}} \sum_{j \in S_r} \mathbb{E}_m \left(y_i y_j\right) \right\} \\
&= N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{n-1} \left\{ \frac{n_r - 1}{n_r} \sum_{i \in S_r} \left(\beta^2 + \sigma^2\right) - \frac{1}{n_r} \sum_{\substack{i \in S_r \\ i \neq j}} \sum_{j \in S_r} \beta^2 \right\} \\
&= N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{n-1} \left\{ \left(n_r - 1\right) \beta^2 + \left(n_r - 1\right) \sigma^2 - \left(n_r - 1\right) \beta^2 \right\} \\
&= N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \frac{n_r - 1}{n-1} \sigma^2
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\mathbb{V}_{diff} &= \mathbb{E}_m \left(\widehat{V}_{HT} - \widehat{V}_{naive}\right) \\
&= \mathbb{E}_m \left(\widehat{V}_{HT}\right) - \mathbb{E}_m \left(\widehat{V}_{naive}\right) \\
&= N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \sigma^2 - N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \frac{n_r - 1}{n-1} \sigma^2 \\
&= N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \left\{ 1 - \frac{n_r - 1}{n-1} \right\} \sigma^2
\end{aligned}$$

$$= N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \frac{n - n_r}{n-1} \sigma^2.$$

Since σ^2 in the above formula for $\mathbb{V}_{diff}$ is unknown, we can replace it by an m-unbiased estimator s^2_{yr}. Therefore, we have

$$\begin{aligned}
\widehat{V}_{sam} &= \widehat{V}_{naive} + \widehat{V}_{diff} \\
&= N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \frac{n_r - 1}{n-1} s^2_{yr} + N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \frac{n - n_r}{n-1} s^2_{yr} \\
&= N^2 \left(\frac{1}{n} - \frac{1}{N}\right) \left\{\frac{n_r - 1}{n-1} + \frac{n - n_r}{n-1}\right\} s^2_{yr} \\
&= \frac{N^2}{n} \left(1 - \frac{n}{N}\right) s^2_{yr}.
\end{aligned}$$

Next, since we can write the nonresponse error term as

$$\begin{aligned}
\widehat{t}_I - \widehat{t}_{y,\pi} &= \sum_{i \in S} w_i r_i y_i + \sum_{j \in S} w_j \left(1 - r_j\right) y^*_j - \sum_{i \in S} w_i y_i \\
&= \frac{N}{n} \sum_{i \in S} r_i y_i + \frac{N}{n} \sum_{i \in S} r_i \sum_{j \in S} \left(1 - r_j\right) \frac{1}{n_r} y_i - \frac{N}{n} \sum_{i \in S} y_i \\
&= \frac{N}{n} \sum_{i \in S} \left\{ r_i \left(1 + \sum_{j \in S} (1 - r_j) \frac{1}{n_r}\right) - 1 \right\} y_i \\
&= \frac{N}{n} \sum_{i \in S} \left\{ r_i \left(1 + \frac{n - n_r}{n_r}\right) - 1 \right\} y_i \\
&= \frac{N}{n} \sum_{i \in S} \left(r_i \frac{n}{n_r} - 1 \right) y_i,
\end{aligned}$$

the nonresponse variance term is given by

$$\begin{aligned}
\mathbb{V}_{nr} &= \mathbb{E}_p \, \mathbb{E}_q \, \mathbb{V}_m \left(\widehat{t}_I - \widehat{t}_{y,\pi}\right) \\
&= \mathbb{E}_p \, \mathbb{E}_q \, \mathbb{V}_m \left(\frac{N}{n} \sum_{i \in S} \left(r_i \frac{n}{n_r} - 1 \right) y_i \right) \\
&= \mathbb{E}_p \, \mathbb{E}_q \left\{ \frac{N^2}{n^2} \sum_{i \in S} \left(r_i \frac{n}{n_r} - 1 \right)^2 \mathbb{V}_m \left(y_i\right) \right\} \\
&= \sigma^2 \frac{N^2}{n^2} \sum_{i \in S} \left(r_i \frac{n}{n_r} - 1 \right)^2.
\end{aligned}$$

Then, by replacing the unknown σ^2 in the above formula for $\mathbb{V}_{nr}$ by the m-unbiased estimator s^2_{yr}, we obtain the nonresponse variance estimator

$$\begin{aligned}
\widehat{V}_{nr} &= \hat{\sigma}^2 \frac{N^2}{n^2} \sum_{i \in S} \left(r_i \frac{n}{n_r} - 1 \right)^2 \\
&= s^2_{yr} \sum_{i \in S} \left(r_i \frac{N^2}{n_r^2} - 2 r_i \frac{N^2}{n_r n} + \frac{N^2}{n^2} \right) \\
&= s^2_{yr} \left\{ n_r \left(\frac{N^2}{n_r^2} \right) - 2 n_r \left(\frac{N^2}{n_r n} \right) + n \left(\frac{N^2}{n^2} \right) \right\} \\
&= s^2_{yr} \left(\frac{N^2}{n_r} - \frac{N^2}{n} \right) \\
&= \frac{N^2}{n_r} \left(1 - \frac{n_r}{n} \right) s^2_{yr}.
\end{aligned}$$

Also, for the $\mathbb{V}_{mix}$ term, we have

$$\begin{aligned}
\mathbb{V}_{mix} &= \mathbb{E}_p \, \mathbb{E}_q \, Cov_m \left\{ \left(\widehat{t}_{y,\pi} - t_y \right), \left(\widehat{t}_I - \widehat{t}_{y,\pi} \right) \right\} \\
&= \mathbb{E}_p \, \mathbb{E}_q \, Cov_m \left\{ \left(\frac{N}{n} \sum_{i \in S} y_i - \sum_{i \in U} y_i \right), \frac{N}{n} \sum_{i \in S} \left(r_i \frac{n}{n_r} - 1 \right) y_i \right\} \\
&= \mathbb{E}_p \, \mathbb{E}_q \left\{ \frac{N^2}{n^2} \sum_{i \in S} \sum_{j \in S} \left(r_j \frac{n}{n_r} - 1 \right) Cov_m \left(y_i, y_j \right) \right. \\
&\quad \left. - \frac{N}{n} \sum_{i \in U} \sum_{j \in S} \left(r_j \frac{n}{n_r} - 1 \right) Cov_m \left(y_i, y_j \right) \right\} \\
&= \mathbb{E}_p \, \mathbb{E}_q \left\{ \frac{N^2}{n^2} \sum_{i \in S} \left(r_i \frac{n}{n_r} - 1 \right) \mathbb{V}_m \left(y_i \right) + \frac{N^2}{n^2} \sum_{\substack{i \in S \\ i \neq j}} \sum_{j \in S} \left(r_j \frac{n}{n_r} - 1 \right) Cov_m \left(y_i, y_j \right) \right. \\
&\quad \left. - \frac{N}{n} \sum_{i \in S} \left(r_i \frac{n}{n_r} - 1 \right) \mathbb{V}_m \left(y_i \right) - \frac{N}{n} \sum_{\substack{i \in U \\ i \neq j}} \sum_{j \in S} \left(r_j \frac{n}{n_r} - 1 \right) Cov_m \left(y_i, y_j \right) \right\} \\
&= \mathbb{E}_p \, \mathbb{E}_q \left\{ \frac{N^2}{n^2} \sum_{i \in S} \left(r_i \frac{n}{n_r} - 1 \right) \sigma^2 - \frac{N}{n} \sum_{i \in S} \left(r_i \frac{n}{n_r} - 1 \right) \sigma^2 \right\} \\
&= \mathbb{E}_p \, \mathbb{E}_q \left\{ \sigma^2 \left(\frac{N^2}{n^2} - \frac{N}{n} \right) \sum_{i \in S} \left(r_i \frac{n}{n_r} - 1 \right) \right\}
\end{aligned}$$

$$= \sigma^2 \frac{N(N-n)}{n^2} \sum_{i \in S} \left(r_i \frac{n}{n_r} - 1 \right),$$

where the fifth equality follows from the fact that

$$\mathbb{V}_m(y_i) = \mathbb{V}_m(\beta + \epsilon_i) = \mathbb{V}_m(\epsilon_i) = \sigma^2$$

and

$$Cov_m(y_i, y_j) = Cov_m(\beta + \epsilon_i, \beta + \epsilon_j) = Cov_m(\epsilon_i, \epsilon_j) = 0.$$

Again, by replacing the unknown σ^2 in the above formula for $\mathbb{V}_{mix}$ by the m-unbiased estimator s^2_{yr}, we obtain the estimator $\widehat{V}_{mix}$ which is given by

$$\begin{aligned}
\widehat{V}_{mix} &= \hat{\sigma}^2 \frac{N(N-n)}{n^2} \sum_{i \in S} \left(r_i \frac{n}{n_r} - 1 \right) \\
&= \hat{\sigma}^2 \sum_{i \in S} r_i \frac{n}{n_r} \left(\frac{N(N-n)}{n^2} \right) - \hat{\sigma}^2 \left(\frac{N(N-n)}{n} \right) \\
&= \hat{\sigma}^2 \frac{N(N-n)}{n} - \hat{\sigma}^2 \frac{N(N-n)}{n} \\
&= 0
\end{aligned}$$

Therefore, the total variance estimator $\widehat{V}_{tot}$ under mean imputation and SRSWOR sampling design is given by

$$\begin{aligned}
\widehat{V}_{tot} &= \widehat{V}_{sam} + \widehat{V}_{nr} + 2\widehat{V}_{mix} \\
&= \frac{N^2}{n} \left(1 - \frac{n}{N} \right) s^2_{yr} + \frac{N^2}{n_r} \left(1 - \frac{n_r}{n} \right) s^2_{yr} + 0 \\
&= s^2_{yr} \left(\frac{N^2}{n} - N + \frac{N^2}{n_r} - \frac{N^2}{n} \right) \\
&= s^2_{yr} \left(\frac{N^2}{n_r} - N \right) \\
&= \frac{N^2}{n_r} \left(1 - \frac{n_r}{N} \right) s^2_{yr}
\end{aligned}$$

Appendix B

Conditional bias in the presence of nonresponse

We consider the imputation model:

$$y_i = \boldsymbol{v}_i^\top \boldsymbol{\beta} + \epsilon_i,$$

such that

$$\mathbb{E}_m\left(\epsilon_i|\boldsymbol{v}_i\right) = 0, \quad \mathbb{E}_m\left(\epsilon_i\epsilon_j|\boldsymbol{v}_i,\boldsymbol{v}_j\right) = 0,\ i \neq j, \quad \mathbb{V}_m\left(\epsilon_i|\boldsymbol{v}_i\right) = \sigma^2 c_i.$$

We have the following non-robust imputed estimator

$$\widehat{t}_I = \sum_{i\in S} w_i r_i y_i + \sum_{i\in S} w_i\left(1-r_i\right)\boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{WLS}.$$

Let $\boldsymbol{\beta}^*$ be the probability limit of $\widehat{\boldsymbol{B}}_{WLS}$. Since we have

$$\begin{aligned}\widehat{\boldsymbol{B}}_{WLS} - \boldsymbol{\beta}^* &= \left(\sum_{i\in S} w_i r_i \boldsymbol{v}_i c_i^{-1}\boldsymbol{v}_i^\top\right)^{-1}\sum_{i\in S} w_i r_i \boldsymbol{v}_i c_i^{-1} y_i - \boldsymbol{\beta}^* \\ &= \left(\sum_{i\in S} w_i r_i \boldsymbol{v}_i c_i^{-1}\boldsymbol{v}_i^\top\right)^{-1}\sum_{i\in S} w_i r_i \boldsymbol{v}_i c_i^{-1}\left(y_i - \boldsymbol{v}_i^\top\boldsymbol{\beta}^*\right),\end{aligned}$$

we can write

$$\begin{aligned}\widehat{t}_I - t_y &= \sum_{i\in S} w_i r_i y_i + \sum_{i\in S} w_i\left(1-r_i\right)\boldsymbol{v}_i^\top\widehat{\boldsymbol{B}}_{WLS} - t_y \\ &= \sum_{i\in S} w_i r_i y_i + \sum_{i\in S} w_i\left(1-r_i\right)\boldsymbol{v}_i^\top\left\{\widehat{\boldsymbol{B}}_{WLS} - \boldsymbol{\beta}^* + \boldsymbol{\beta}^*\right\} - t_y\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i\in S} w_i r_i y_i + \sum_{i\in S} w_i \left(1-r_i\right) \boldsymbol{v}_i^\top \boldsymbol{\beta}^* - t_y + \sum_{i\in S} w_i \left(1-r_i\right) \boldsymbol{v}_i^\top \left(\widehat{\boldsymbol{B}}_{WLS} - \boldsymbol{\beta}^*\right) \\
&= \sum_{i\in S} w_i r_i y_i + \sum_{i\in S} w_i \left(1-r_i\right) \boldsymbol{v}_i^\top \boldsymbol{\beta}^* - t_y \\
&\quad + \left(\sum_{i\in S} w_i \left(1-r_i\right) \boldsymbol{v}_i^\top\right) \left(\sum_{i\in S} w_i r_i \boldsymbol{v}_i c_i^{-1} \boldsymbol{v}_i^\top\right)^{-1} \sum_{i\in S} w_i r_i \boldsymbol{v}_i c_i^{-1} \left(y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}^*\right) \\
&= \sum_{i\in S} w_i r_i y_i + \sum_{i\in S} w_i \left(1-r_i\right) \boldsymbol{v}_i^\top \boldsymbol{\beta}^* - t_y \\
&\quad + \left(\sum_{i\in U} (1-p_i) \boldsymbol{v}_i^\top\right) \left(\sum_{i\in U} p_i \boldsymbol{v}_i c_i^{-1} \boldsymbol{v}_i^\top\right)^{-1} \sum_{i\in S} w_i r_i \boldsymbol{v}_i c_i^{-1} \left(y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}^*\right) + o_p\left(Nn^{-1/2}\right) \\
&= \sum_{i\in S} w_i \left\{r_i y_i + \left(1-r_i\right) \boldsymbol{v}_i^\top \boldsymbol{\beta}^* + \boldsymbol{C} r_i \boldsymbol{v}_i c_i^{-1} \left(y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}^*\right)\right\} - t_y + o_p\left(Nn^{-1/2}\right),
\end{aligned}
$$

where

$$
\boldsymbol{C} = \left(\sum_{i\in U} (1-p_i) \boldsymbol{v}_i^\top\right) \left(\sum_{i\in U} p_i \boldsymbol{v}_i c_i^{-1} \boldsymbol{v}_i^\top\right)^{-1}.
$$

Therefore, we have

$$
\widehat{t}_I - t_y = \sum_{j\in S} w_j \psi_j - t_y + o_p\left(Nn^{-1/2}\right),
$$

where,

$$
\psi_j = r_j y_j + \left(1-r_j\right) \boldsymbol{v}_j^\top \boldsymbol{\beta}^* + \boldsymbol{C} r_j \boldsymbol{v}_j c_j^{-1} \left(y_j - \boldsymbol{v}_j^\top \boldsymbol{\beta}^*\right).
$$

We can also write

$$
\widehat{t}_I - t_y = \sum_{j\in U} w_j \psi_j I_j - t_y + o_p\left(Nn^{-1/2}\right).
$$

Now,

$$
\begin{aligned}
\mathbb{E}_q\left(\widehat{t}_I - t_y | Y_i = y_i, I_i = 1, r_i = 1\right) &= w_i y_j + w_i \boldsymbol{C} \boldsymbol{v}_i c_i^{-1} \left(y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}^*\right) \\
&\quad + \mathbb{E}_q\left(\sum_{\substack{j\in U \\ j\neq i}} w_j \psi_j I_j | Y_i = y_i, I_i = 1, r_i = 1\right) - t_y
\end{aligned}
$$

$$
\begin{aligned}
&= w_i y_j + w_i \boldsymbol{C}\boldsymbol{v}_i c_i^{-1} \left(y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}^*\right) \\
&\quad + \sum_{\substack{j \in U \\ j \neq i}} w_j I_j \, \mathbb{E}_q \left(\psi_j | Y_i = y_i, I_i = 1, r_i = 1\right) - t_y \\
&= w_i y_j + w_i \boldsymbol{C}\boldsymbol{v}_i c_i^{-1} \left(y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}^*\right) + \sum_{\substack{j \in U \\ j \neq i}} w_j I_j \widetilde{\psi}_j - t_y,
\end{aligned}
$$

where

$$
\widetilde{\psi}_j = p_j y_j + \left(1 - p_j\right) \boldsymbol{v}_j^\top \boldsymbol{\beta}^* + \boldsymbol{C} p_j \boldsymbol{v}_j c_j^{-1} \left(y_j - \boldsymbol{v}_j^\top \boldsymbol{\beta}^*\right).
$$

Here, we used the fact that $\mathbb{E}_q \left(r_j | r_i = 1\right) = \mathbb{E}_q \left(r_j\right) = p_j, \quad j \neq i.$

Note that we can also write $\widetilde{\psi}_j$ as:

$$
\begin{aligned}
\widetilde{\psi}_j &= y_j - \left(1 - p_j\right) \left(y_j - \boldsymbol{v}_j^\top \boldsymbol{\beta}^*\right) + \boldsymbol{C} p_j \boldsymbol{v}_j c_j^{-1} \left(y_j - \boldsymbol{v}_j^\top \boldsymbol{\beta}^*\right) \\
&= y_j + \left\{\boldsymbol{C} p_j \boldsymbol{v}_j c_j^{-1} - \left(1 - p_j\right)\right\} \left(y_j - \boldsymbol{v}_j^\top \boldsymbol{\beta}^*\right).
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
\mathbb{E}_q \left(\widehat{t}_I - t_y | Y_i = y_i, I_i = 1, r_i = 1\right) &= w_i y_j + w_i \boldsymbol{C}\boldsymbol{v}_i c_i^{-1} \left(y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}^*\right) \\
&\quad + \sum_{\substack{j \in U \\ j \neq i}} w_j I_j y_j + \sum_{\substack{j \in U \\ j \neq i}} w_j I_j \psi_j^* - t_y,
\end{aligned}
$$

where

$$
\psi_j^* = \left\{\boldsymbol{C} p_j \boldsymbol{v}_j c_j^{-1} - \left(1 - p_j\right)\right\} \left(y_j - \boldsymbol{v}_j^\top \boldsymbol{\beta}^*\right).
$$

This implies that

$$
\begin{aligned}
\mathbb{E}_p \, \mathbb{E}_q \left(\widehat{t}_I - t_y | Y_i = y_i, I_i = 1, r_i = 1\right) &= \left(w_i - 1\right) y_j + w_i \boldsymbol{C}\boldsymbol{v}_i c_i^{-1} \left(y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}^*\right) \\
&\quad + \sum_{\substack{j \in U \\ j \neq i}} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}\right) y_j + \sum_{\substack{j \in U \\ j \neq i}} w_j \psi_j^* \, \mathbb{E}_p \left(I_j | I_i = 1\right)
\end{aligned}
$$

$$= B_i^s + w_i \boldsymbol{C} \boldsymbol{v}_i c_i^{-1} \left(y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}^* \right) + \sum_{\substack{j \in U \\ j \neq i}} \frac{\pi_{ij}}{\pi_i \pi_j} \psi_j^*,$$

where

$$B_i^s = \sum_{j \in U} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_j.$$

Finally, we have

$$\begin{aligned} B_i^I &= \mathbb{E}_m \, \mathbb{E}_p \, \mathbb{E}_q \left(\widehat{t}_I - t_y | Y_i = y_i, I_i = 1, r_i = 1 \right) \\ &\approx \mathbb{E}_m \left(B_i^s | Y_i = y_i \right) + w_i \boldsymbol{C} \boldsymbol{v}_i c_i^{-1} \left(y_i - \boldsymbol{v}_i^\top \boldsymbol{\beta}^* \right) + 0, \end{aligned}$$

using the fact that

$$\mathbb{E}_m \left(y_j | Y_i = y_i \right) = \boldsymbol{v}_i^\top \boldsymbol{\beta}.$$

An estimator of B_i^I is given by

$$\widehat{B}_i^I = \widehat{B}_i^s + w_i \widehat{\boldsymbol{C}} \boldsymbol{v}_i c_i^{-1} \left(y_i - \boldsymbol{v}_i^\top \widehat{\boldsymbol{B}}_{WLS} \right),$$

where

$$\widehat{\boldsymbol{C}} = \left(\sum_{i \in S} w_i \left(1 - r_i \right) \boldsymbol{v}_i^\top \right) \left(\sum_{i \in S} w_i r_i \boldsymbol{v}_i c_i^{-1} \boldsymbol{v}_i^\top \right)^{-1},$$

and

$$\widehat{B}_i^s = \sum_{j \in S} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_j \pi_{ij}} \right) y_j.$$

Appendix C

Estimator of conditional bias under simple linear regression imputation

An estimator of conditional bias under SRSWOR is given by (3.1.11).

Recall that under simple linear regression imputation, we have $\boldsymbol{v}_i = (1, v_i)^\top$ and $c_i = 1$.

Therefore, we can write

$$
\begin{aligned}
\widehat{\mathbf{C}}\boldsymbol{v}_i &= \left\{\sum_{i\in S}\frac{N}{n}(1-r_i)\boldsymbol{v}_i^\top\right\}\left\{\sum_{i\in S}\frac{N}{n}r_i\boldsymbol{v}_i c_i^{-1}\boldsymbol{v}_i^\top\right\}^{-1}\boldsymbol{v}_i \\
&= \left\{\sum_{i\in S}\frac{N}{n}(1-r_i)(1,v_i)\right\}\left\{\sum_{i\in S}\frac{N}{n}r_i\begin{pmatrix}1\\v_i\end{pmatrix}(1,v_i)\right\}^{-1}\begin{pmatrix}1\\v_i\end{pmatrix} \\
&= \begin{pmatrix}\sum_{i\in S_m}1 & \sum_{i\in S_m}v_i\end{pmatrix}\begin{pmatrix}\sum_{i\in S_r}1 & \sum_{i\in S_r}v_i\\ \sum_{i\in S_r}v_i & \sum_{i\in S_r}v_i^2\end{pmatrix}^{-1}\begin{pmatrix}1\\v_i\end{pmatrix} \\
&= \begin{pmatrix}\sum_{i\in S_m}1 & \sum_{i\in S_m}v_i\end{pmatrix}\begin{pmatrix}\sum_{i\in S_r}1 & \sum_{i\in S_r}v_i\\ \sum_{i\in S_r}v_i & \sum_{i\in S_r}v_i^2\end{pmatrix}^{-1}\begin{pmatrix}1\\v_i\end{pmatrix} \\
&= \begin{pmatrix}n_m & n_m\overline{v}_{nr}\end{pmatrix}\begin{pmatrix}n_r & n_r\overline{v}_r\\ n_r\overline{v}_r & \sum_{i\in S_r}v_i^2\end{pmatrix}^{-1}\begin{pmatrix}1\\v_i\end{pmatrix} \\
&= \frac{n_m\sum_{i\in S_r}v_i^2 - n_m n_r\overline{v}_{nr}\overline{v}_r - n_m n_r\overline{v}_r v_i + n_m n_r\overline{v}_{nr}v_i}{n_r\sum_{i\in S_r}v_i^2 - (n_r\overline{v}_r)^2} \\
&= \frac{n_m\sum_{i\in S_r}v_i^2 - n_m n_r\overline{v}_r^2 + n_r n\overline{v}v_i - n_r n v_i\overline{v}_r - n_r\overline{v}_r n\overline{v} + n n_r\overline{v}_r^2}{n_r\sum_{i\in S_r}v_i^2 - (n_r\overline{v}_r)^2}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{n_m \left(\sum_{i\in S_r} v_i^2 - n_r \overline{v}_r^2\right) + n_r n \left(v_i \overline{v} - v_i \overline{v}_r - \overline{v}_r \overline{v} + \overline{v}_r^2\right)}{n_r \sum_{i\in S_r} v_i^2 - \left(n_r \overline{v}_r\right)^2} \\
&= \frac{n_m \sum_{i\in S_r} \left(v_i - \overline{v}_r\right)^2 + n_r n \left(v_i - \overline{v}_r\right)\left(\overline{v} - \overline{v}_r\right)}{n_r \sum_{i\in S_r} \left(v_i - \overline{v}_r\right)^2} \\
&= \frac{n_m}{n_r} + \frac{n \left(v_i - \overline{v}_r\right)\left(\overline{v} - \overline{v}_r\right)}{\sum_{i\in S_r} \left(v_i - \overline{v}_r\right)^2} \\
&= \frac{n}{n_r}\left(\frac{n_m}{n} + \frac{\left(v_i - \overline{v}_r\right)\left(\overline{v} - \overline{v}_r\right)}{\frac{1}{n_r}\sum_{i\in S_r} \left(v_i - \overline{v}_r\right)^2}\right).
\end{aligned}
$$

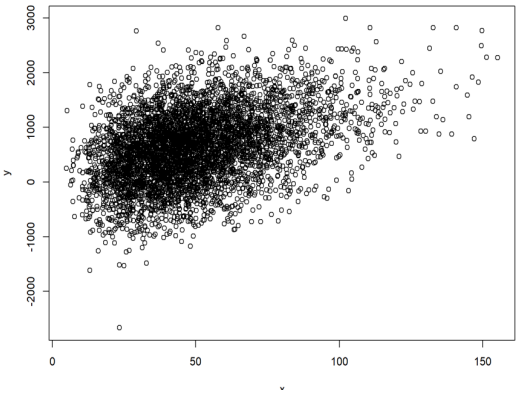
Then, an estimator of conditional bias is given by

$$
\begin{aligned}
\hat{B}_i^I &= \frac{n}{n-1}\left(\frac{N}{n} - 1\right)\left(y_i - \bar{y}_I\right) + \left(\frac{N}{n}\right)\frac{n}{n_r}\left(\frac{n_m}{n} + \frac{\left(v_i - \overline{v}_r\right)\left(\overline{v} - \overline{v}_r\right)}{\frac{1}{n_r}\sum_{i\in S_r} \left(v_i - \overline{v}_r\right)^2}\right)\left(y_i - \widehat{B}_{0,WLS} - \widehat{B}_{1,WLS} v_i\right) \\
&\approx \left(\frac{N}{n} - 1\right)\left(y_i - \bar{y}_I\right) + \left(\frac{N}{n}\right)\frac{1}{\widehat{p}}\left\{(1 - \widehat{p}) + \frac{(v_i - \overline{v}_r)(\overline{v} - \overline{v}_r)}{s_{vr}^2}\right\}\left(y_i - \widehat{B}_{0,WLS} - \widehat{B}_{1,WLS} v_i\right),
\end{aligned}
$$

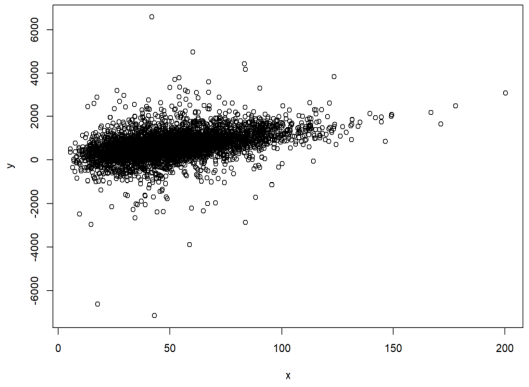
where $\bar{y}_I = \widehat{t}_I / N$, $\widehat{p} = n_r / n$, and $s_{vr}^2 = (n_r - 1)^{-1} \sum_{i\in S_r} (v_i - \overline{v}_r)^2$. The approximation follows from the fact that $\frac{n}{n-1} \approx 1$ and $n_r - 1 \approx n_r$ for large n and n_r.

Appendix D

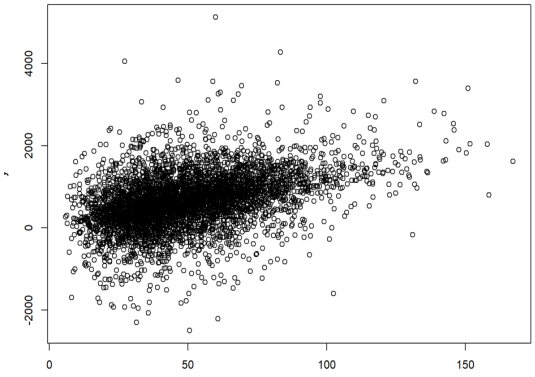
Graphical representation of distributions



(a) Normal Distribution

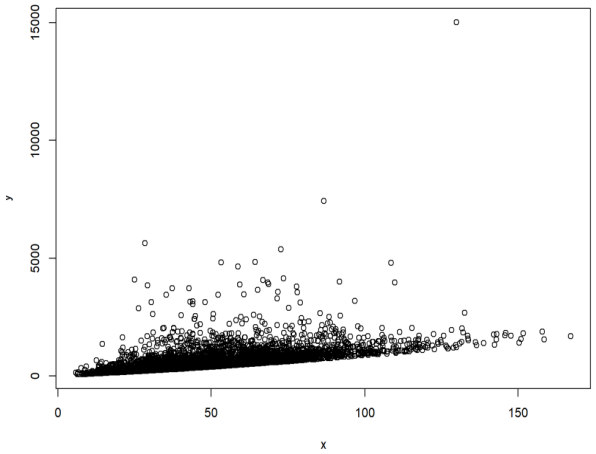


(b) Student's t-Distribution

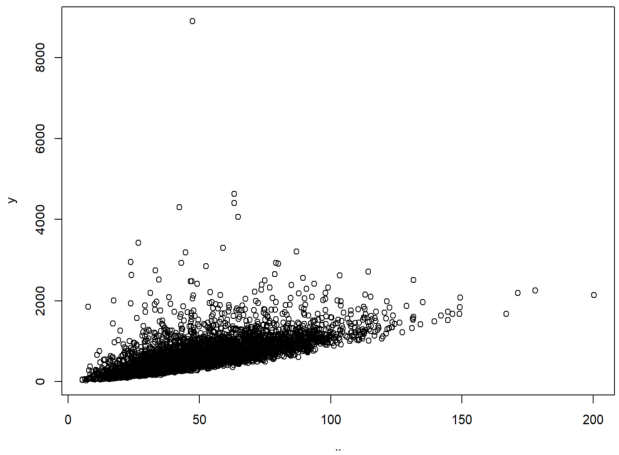


(c) Double Exponential (Laplace) Distribution

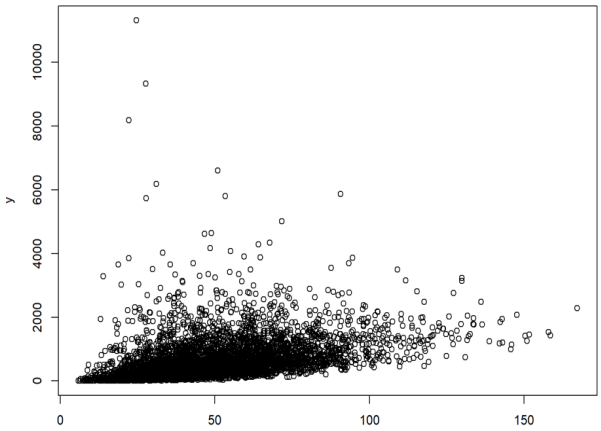
Figure D.1: Symmetric Distributions



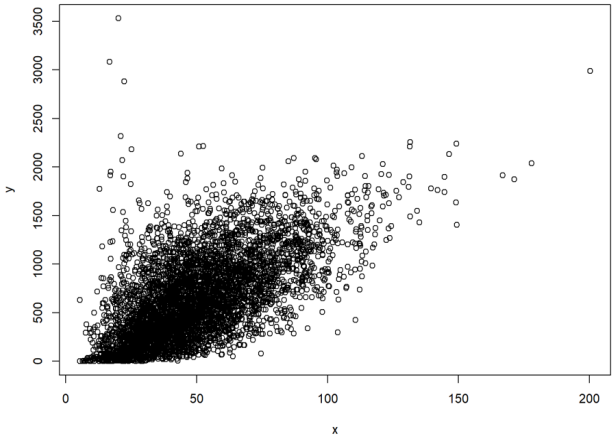
(a) Pareto Distribution



(b) Frechet Distribution

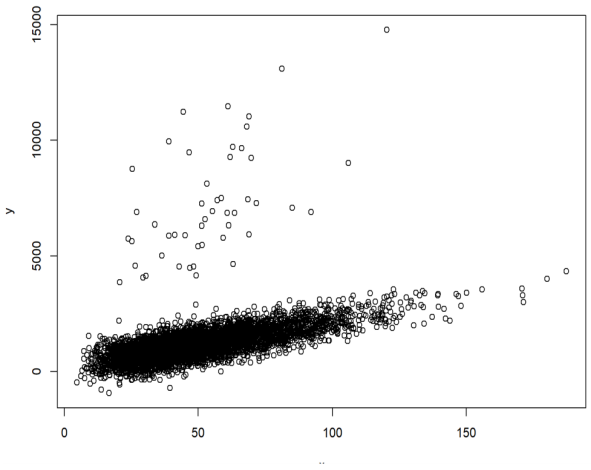


(c) Lognormal Distribution

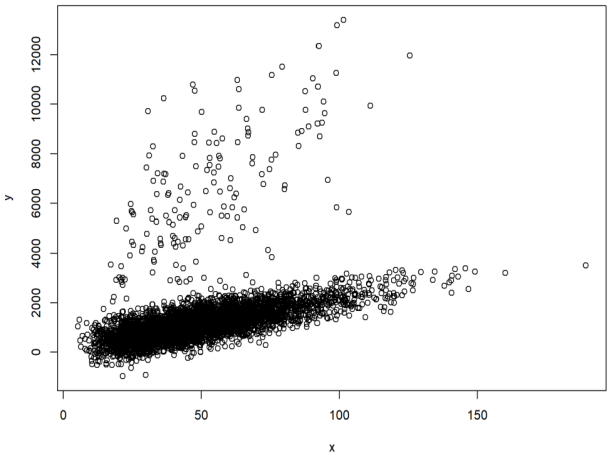


(d) Weibull Distribution

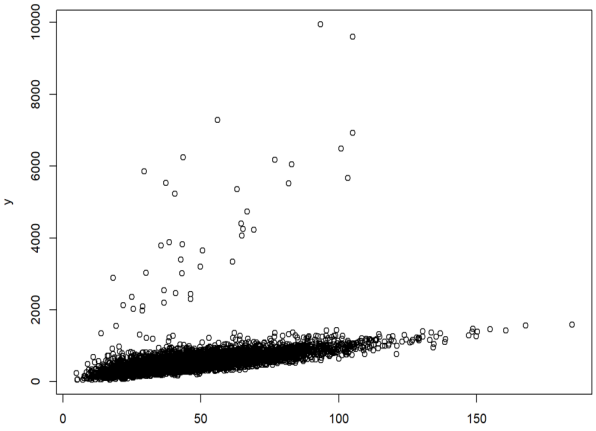
Figure D.2: Asymmetric Distributions



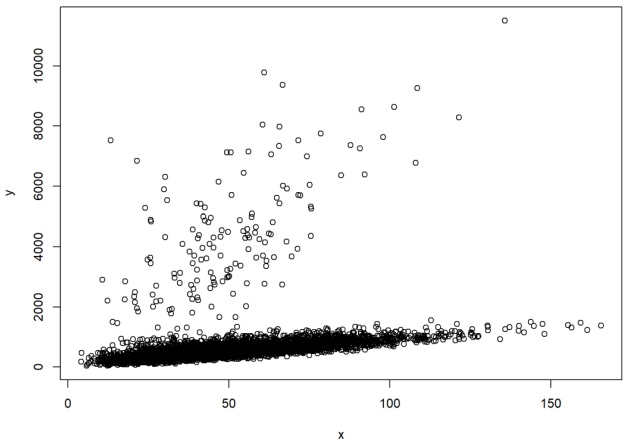
(a) Mixture of Normal (1% outliers)



(b) Mixture of Normal (3% outliers)



(c) Mixture of Lognormal (1% outliers)



(d) Mixture of Lognormal (3% outliers)

Figure D.3: Mixture Distributions