Robust Estimation and Prediction in the Presence of Influential Units in Surveys

Yizhen Teng

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

In surveys, one may face the problem of influential units at the estimation stage. A unit is said to be influential if its inclusion or exclusion from the sample has a drastic impact on the estimates. This is a common situation in business surveys as the distribution of economic variables tends to be highly skewed. We study and examine some commonly used estimators and predictors of a population total and propose a robust estimator and predictor based on an adaptive tuning constant. The proposed tuning constant is based on the concept of conditional bias of a unit, which is a measure of influence. We present the results of a simulation study that compares the performance of several estimators and predictors in terms of bias and efficiency.

Keywords: Robustness; Influential units; Conditional bias; Adaptive tuning constant.

Résumé

Dans les enquêtes, on peut être confronté au problème des unités influentes à l'étape de l'estimation. Une unité est considérée comme influente si son inclusion ou exclusion de l'échantillon a un impact important sur les estimations. C'est une situation courante dans les enquêtes auprès des entreprises car la distribution des variables économiques tend à être très asymétrique. Nous étudions et examinons certains estimateurs et prédicteurs couramment utilisés pour un total de la population et proposons un estimateur et un prédicteur robustes basés sur une constante d'ajustement adaptative. La constante d'ajustement proposée est basée sur le concept de biais conditionnel d'une unité, qui est une mesure de l'influence. Nous présentons les résultats d'une étude par simulation qui compare les performances de plusieurs estimateurs et prédicteurs en termes de biais et d'efficacité.

Mots-clés : Robustesse ; Unités influentes ; Biais conditionnel ; Constante d'ajustement adaptative.

Dedications

I dedicate this work to Guangxin.

Acknowledgement

I am grateful to Professor David Haziza for his invaluable guidance throughout this journey. My heartfelt thanks also go to my parents for their endless support and encouragement. Additionally, I appreciate the help provided by my friend Ziming An and many other friends during my study.

Contents

List of Figures

List of Tables

Introduction

Surveys are widely-used for collecting information about a finite population. Typically, surveys gather data on variables of interest, also known as survey variables or characteristics of interest, with the goal of estimating finite population parameters. These surveys are often referred to as multipurpose surveys, as they aim to serve various research purposes and provide a comprehensive overview of the population under investigation.

At the estimation stage of a survey, the issue of influential units may arise. A unit is influential if its inclusion or exclusion from the sample has a significant impact on the estimates. Influential units are commonly encountered in business surveys due to the highly skewed distribution of economic variables, such as revenues, sales and profits. A small proportion of businesses earn a large proportion of the total revenue, resulting in a highly skewed distribution.

The objective of this work is to examine the properties of estimators/predictors in the presence of influential units. In Chapter 1, we provide an overview of inferential approaches in survey sampling, including the design-based approach, the model-assisted approach, and the model-based approach. We also describe several commonly used estimators and predictors, such as the Horvitz-Thompson estimator, the ratio estimator, the generalized regression (GREG) estimator, and the best unbiased linear predictor (BLUP). In Chapter 2, we introduce a measure of the influence called conditional bias, under all three frameworks. In Chapter 3, we present several robust estimators and predictors such as the naive estimator/predictor, the predictor of Chambers (1986), the robust estimator/predictor based on the concept of conditional bias (2013). We also propose a new robust estimator/predictor based on an adaptive tuning constant. In Chapter 4, we conduct several simulation studies to evaluate the performance of estimators/predictors in the presence of influential units under the model-assisted framework and the model-based framework. We make some final remarks in Chapter 5.

In this thesis, we first provide a literature review on robust estimation methods in survey sampling. We also propose a novel robust method based on an adaptive tuning constant.

Chapter 1

Inferential approaches in survey sampling

1.1 Finite Population and Finite Population Parameter

We consider a finite population of N units, which is denoted by $U = \{1, 2, ..., N\}$. Let y be a survey variable, i.e., a variable collected on the field. Let $\mathbf{y} = (y_1, y_2, ..., y_N)^\top$ be the vector of population y-values, where y_i is the y-value for the ith unit, $i = 1, 2, ..., N$. A finite population parameter is defined as $\theta_N = h(\mathbf{y})$, where $h(\cdot)$ is a given function. Examples of finite population parameters include, a population total, $t_y = \sum_{i \in U} y_i$, and a population mean, $\bar{Y} = t_y/N$. To estimate θ_N, we select a sample S, of size n, from U.

1.2 Sampling Design

We assume that a q-vector of design variables, $\mathbf{x}$, is available for all the population units prior to sampling. Let $\mathbf{X}$ be a $N \times q$ matrix, whose N rows are the q dimensional vectors $\mathbf{x}_1, ..., \mathbf{x}_N$.

The sample S is selected according to a given sampling design $p(S|\mathbf{X})$, which gives the probability of the sample S being drawn from the population, conditional on $\mathbf{X}$. We define the support of a sampling design as $Q = \{s | p(s|\mathbf{X}) > 0\}$, where s denotes a possible sample. A sampling design is a probability distribution, which implies that $\sum_{s \in Q} p(s|\mathbf{X}) = 1$.

We use the vector of sample selection indicators $\mathbf{I} = (I_1, I_2, ..., I_N)^\top$ to indicate which units are selected in the sample and which units are not selected, such that $I_i = 1$, if $i \in S$, and $I_i = 0$, otherwise.

The first-order inclusion probability of unit i is defined as the probability of $I_i = 1$ conditional on $\mathbf{X}$, which is given by $\pi_i = p(I_i = 1|\mathbf{X})$. The design weight of unit i is defined as the inverse of its inclusion probability, i.e., $d_i = 1/\pi_i$. The weight d_i is usually interpreted as the number of units in the population represented by unit i.

When the first-order inclusion probabilities π_i are equal for all i, the design is called an equal probability sampling design. Examples of equal probability sampling designs include simple random sampling without replacement, Bernoulli sampling, systematic sampling and stratified sampling with proportional allocation.

The second-order inclusion probability of units i and j is defined as the probability of both $I_i = 1$ and $I_j = 1$ conditional on $\mathbf{X}$, i.e., $\pi_{ij} = p(I_i = 1, I_j = 1|\mathbf{X})$. Note that $\pi_{ii} = \pi_i$ when $i = j$.

We now present three sampling designs commonly encountered in practice.

1.2.1 Simple random sampling without replacement

Simple random sampling without replacement is a fixed-size sampling design, such that any subset of n units has the same probability of being drawn than any other subset of n units. Since we have $\binom{N}{n}$ possible samples, the probability of selecting a sample S, of size n, is $p(S) = 1/\binom{N}{n}$. The first-order inclusion probability of unit i is given by

$$\pi_i = p(I_i = 1|\mathbf{X}) = \sum_{\substack{s\in Q\\ i\in s}} p(s) = \binom{N-1}{n-1} \Big/ \binom{N}{n} = n/N, \quad i = 1, 2, ..., N,$$

since there are $\binom{N-1}{n-1}$ samples of size n containing unit i. The second-order inclusion probability of units i and j is given by

$$\pi_{ij} = p(I_i = 1, I_j = 1|\mathbf{X}) = \sum_{\substack{s\in Q\\ i,j\in s}} p(s) = \binom{N-2}{n-2} \Big/ \binom{N}{n} = \frac{n(n-1)}{N(N-1)}, \quad i, j = 1, 2, ..., N,$$

since there are $\binom{N-2}{n-2}$ samples of size n containing both i and j.

1.2.2 Poisson sampling

Poisson sampling consists of performing N independent Bernoulli trials with probabilities $\pi_i \in (0, 1)$, $i = 1, 2, ..., N$. If a trial is a success, unit i is selected in the sample (i.e., $I_i = 1$), otherwise, it is not selected (i.e., $I_i = 0$).

The probability of selecting a sample S is given by

$$p(S) = \prod_{i \in S} \pi_i \prod_{i \in U-S} (1 - \pi_i).$$

The sample size n_s is a random variable whose value is between 0 and N. The expected sample size is $\mathbb{E}(n_s) = \sum_{i \in U} \pi_i$, and the variance of n_s is $\mathbb{V}(n_s) = \sum_{i \in U} \pi_i(1 - \pi_i)$.

The first-order inclusion probability of unit i under Poisson sampling is π_i and the second-order inclusion probability of units i and j is $\pi_{ij} = \pi_i \pi_j$, $i \neq j$, since I_i and I_j are independent random variables. When $\pi_i = \pi_0$, for all $i \in U$, Poisson sampling is called Bernoulli sampling.

1.2.3 Stratified sampling

The finite population is partitioned into L strata, $U_1, ..., U_L$, of size $N_1, ..., N_L$, respectively. Note that $\cup_{h=1}^{L} U_h = U$ and $\sum_{h=1}^{L} N_h = N$. From stratum h, we select a random sample S_h, of size n_h, according to a given sampling design $p(S_h|\mathbf{X})$. The selection in one stratum is independent of the selection in any other stratum. The resulting total sample S is given by $S = \cup_{h=1}^{L} S_h$, of size $n = \sum_{h=1}^{L} n_h$.

The probability of selecting a sample S is given by

$$p(S|\mathbf{X}) = \prod_{h=1}^{L} p(S_h|\mathbf{X}).$$

For example, in stratified simple random sampling without replacement, the sample S_h is selected from U_h with probability

$$p(S_h) = 1/\binom{N_h}{n_h}.$$

It follows that

$$p(S|\mathbf{X}) = \prod_{h=1}^{L} 1/\binom{N_h}{n_h}.$$

In the case of stratified simple random sampling without replacement, the first-order inclusion probability of $i \in U_h$ is given by

$$\pi_i = n_h/N_h.$$

The second-order inclusion probability of units i and j is given by

$$\pi_{ij} = \begin{cases} \dfrac{n_h}{N_h}\dfrac{n_l}{N_l} & i \in U_h, j \in U_l, h \neq l \\ \dfrac{n_h(n_h - 1)}{N_h(N_h - 1)} & i \in U_h, j \in U_h \end{cases}$$

1.3 Design-Based Approach

In the design-based approach, a random sample S is selected from the finite population U of size N according to a known probability sampling design $p(S|\mathbf{X})$. When evaluating the properties of estimators (e.g., bias and variance), all the quantities but the vector of selection indicators $\mathbf{I}$ are treated as fixed. Thus, the finite population parameter $\theta_N = h(\mathbf{y})$ is a fixed quantity that we wish to estimate. The design-based approach is nonparametric since we do not need to postulate a model for the survey variable y.

Properties of estimators in the design-based approach are evaluated with respect to the sampling design. Let $\hat{\theta}$ denote an estimator of θ_N. We define $\mathbb{E}_p(\hat{\theta})$ as the design-expectation of the estimator $\hat{\theta}$:

$$\mathbb{E}_p(\hat{\theta}) = \sum_{s \in Q} \hat{\theta}(s) p(s).$$

The design bias of $\hat{\theta}$ is defined as

$$B_p(\hat{\theta}) = \mathbb{E}_p(\hat{\theta}) - \theta_N.$$

The estimator $\hat{\theta}$ is design-unbiased for θ_N if $B_p(\hat{\theta}) = 0$. The design variance of $\hat{\theta}$ is defined as

$$\mathbb{V}_p(\hat{\theta}) = \mathbb{E}_p \left\{ \hat{\theta} - \mathbb{E}_p(\hat{\theta}) \right\}^2.$$

The design variance measures the variability of an estimator across all possible samples from the population.

1.3.1 The Horvitz-Thompson estimator

Horvitz and Thompson (1952) introduced the Horvitz-Thompson estimator of a population total $t_y = \sum_{i \in U} y_i$, which is given by

$$\hat{t}_{y,HT} = \sum_{i \in S} \frac{y_i}{\pi_i}. \tag{1.3.1}$$

Provided that $\pi_i > 0$, for all $i \in U$, the Horvitz-Thompson estimator is design-unbiased for t_y:

$$B_p(\hat{t}_{y,HT}) = \mathbb{E}_p(\hat{t}_{y,HT}) - t_y = 0.$$

The design variance of $\hat{t}_{y,HT}$ is given by

$$\mathbb{V}_p(\hat{t}_{y,HT}) = \sum_{i \in U} \sum_{j \in U} \Delta_{ij} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}, \tag{1.3.2}$$

where $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$. The variance given by (1.3.2) is unknown since we only observe y_i, for $i \in S$. An estimator of $\mathbb{V}_p(\hat{t}_{y,HT})$, called the Horvitz-Thompson variance estimator, is given by

$$\hat{V}_{HT}(\hat{t}_{y,HT}) = \sum_{i \in S} \sum_{j \in S} \frac{\Delta_{ij}}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}.$$

For a fixed-size sampling design, an alternative estimator of $\mathbb{V}_p(\hat{t}_{y,HT})$, called the Sen-Yates-Grundy (SYG) variance estimator (1953), is given by

$$\hat{V}_{SYG}(\hat{t}_{y,HT}) = -\frac{1}{2} \sum_{i \in S} \sum_{j \in S} \frac{\Delta_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2.$$

Both the Horvitz-Thompson variance estimator and the SYG variance estimator are design-unbiased for $\mathbb{V}_p(\hat{t}_{y,HT})$, provided $\pi_{ij} > 0$ for all pairs $(i, j) \in U \times U$. That is,

$$\mathbb{E}_p \left\{ \hat{V}_{HT}(\hat{t}_{y,HT}) \right\} = \mathbb{E}_p \left\{ \hat{V}_{SYG}(\hat{t}_{y,HT}) \right\} = \mathbb{V}_p(\hat{t}_{y,HT}).$$

1.3.2 The Horvitz-Thompson estimator in stratified sampling

In the context of a stratified population, the population total t_y can be expressed as

$$t_y = \sum_{i \in U} y_i = \sum_{h=1}^{L} \sum_{i \in U_h} y_i = \sum_{h=1}^{L} t_h,$$

where $t_h = \sum_{i \in U_h} y_i$ denotes the total in stratum h, $h = 1, .., L$. In stratum h, we estimate t_h by its Horvitz-Thompson estimator:

$$\hat{t}_{h,HT} = \sum_{i \in S_h} \frac{y_i}{\pi_i}.$$

Thus, the Horvitz-Thompson estimator of t_y is given by

$$\hat{t}_{y,stra} = \sum_{h=1}^{L} \hat{t}_{h,HT} = \sum_{h=1}^{L} \sum_{i \in S_h} \frac{y_i}{\pi_i} = \sum_{i \in S} \frac{y_i}{\pi_i}.$$

The design variance of $\hat{t}_{y,stra}$ is given by

$$\mathbb{V}_p \left(\hat{t}_{y,stra} \right) = \mathbb{V}_p \left(\sum_{h=1}^{L} \hat{t}_{h,HT} \right) = \sum_{h=1}^{L} \mathbb{V}_p \left(\hat{t}_{h,HT} \right),$$

and

$$\mathbb{V}_p \left(\hat{t}_{h,HT} \right) = \sum_{i \in U_h} \sum_{j \in U_h} \Delta_{ij} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}.$$

An estimator of $\mathbb{V}_p\left(\hat{t}_{y,stra}\right)$ is given by

$$\hat{V}\left(\hat{t}_{y,stra}\right) = \sum_{h=1}^{L} \hat{V}\left(\hat{t}_{h,HT}\right),$$

where $\hat{V}(\hat{t}_{h,HT})$ is an estimator of $\mathbb{V}_p\left(\hat{t}_{h,HT}\right)$. For instance, we may use the Horvitz-Thompson variance estimator given by

$$\hat{V}(\hat{t}_{h,HT}) = \sum_{i \in S_h} \sum_{j \in S_h} \frac{\Delta_{ij}}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}.$$

1.4 Model-Assisted Approach

In the model-assisted approach, we make use of auxiliary information at the estimation stage. That is, auxiliary information is incorporated in the construction of an estimator according to a working model describing the relationship between the survey variable y and a vector of auxiliary variables $\mathbf{x}$:

$$m : y_i = f(\mathbf{x}_i) + \epsilon_i, \quad i \in U, \tag{1.4.1}$$

where $f(\cdot)$ is an unknown function and the errors ϵ_i are independent random variables such that $\mathbb{E}_m(\epsilon_i|\mathbf{x}_i) = 0$ and $\mathbb{V}_m(\epsilon_i|\mathbf{x}_i) = \sigma_i^2 \equiv \sigma^2\nu(\mathbf{x}_i)$, where σ^2 is an unknown parameter and $\nu(\cdot)$ is a known function.

The unknown function $f(\cdot)$ is estimated by $\hat{f}(\cdot)$ from the sample. A model-assisted estimator of $t_y = \sum_{i\in U} y_i$ is given by

$$\hat{t}_{y,ma} = \sum_{i \in U} \hat{f}(\mathbf{x}_i) + \sum_{i \in S} \frac{y_i - \hat{f}(\mathbf{x}_i)}{\pi_i},$$

where $\hat{f}(\mathbf{x})$ denotes the prediction at $\mathbf{x}$ under the working model (1.4.1).

The estimator $\hat{t}_{y,ma}$ is generally design-biased. However, for a wide class of working models, it can be shown that $\hat{t}_{y,ma}$ is a design-consistent estimator of t_y.

If the sample data is well described by the working model, the population fit residuals $y_i - \hat{f}(\mathbf{x}_i)$ will be small and we expect $\hat{t}_{y,ma}$ to be more efficient than the Horvitz-Thompson estimator. In the model-assisted approach, the properties of $\hat{t}_{y,ma}$ are still evaluated with respect to the sampling design (as in the design-based approach). A nice feature of model-assisted estimator is that they remain design-consistent even if the working model is misspecified. We now present several model-assisted estimators: the ratio estimator and the Generalized REGression (GREG) estimator.

1.4.1 The ratio estimator

We assume that a single auxiliary quantitative variable x is available for all the sample units, and that its population total $t_x = \sum_{i \in U} x_i$ is known. We assume that the relationship between y and x can be described by

$$m : y_i = \beta x_i + \epsilon_i,$$

where β is an unknown parameter. We assume that $\mathbb{E}_m(\epsilon_i|x_i) = 0$, $\mathbb{E}_m(\epsilon_i \epsilon_j|x_i, x_j) = 0$ if $i \neq j$, and $\mathbb{V}_m(\epsilon_i|x_i) = \sigma^2 x_i$. The population total t_y can be written as

$$t_y = \frac{t_y}{t_x} t_x.$$

The ratio estimator of t_y is given by

$$\hat{t}_{y,ra} = \frac{\hat{t}_{y,HT}}{\hat{t}_{x,HT}} t_x, \tag{1.4.2}$$

where $\hat{t}_{x,HT} = \sum_{i \in S} x_i/\pi_i$ is the Horvitz Thompson estimator of t_x. The ratio estimator (1.4.2) can also be expressed as

$$\hat{t}_{y,ra} = \sum_{i \in S} w_i y_i,$$

where $w_i = d_i \times \left(t_x/\hat{t}_{x,HT}\right)$. The ratio estimator is design-biased since

$$\mathbb{E}_p\left(\hat{t}_{y,ra}\right) \neq \frac{\mathbb{E}_p\left(\hat{t}_{y,HT}\right)}{\mathbb{E}_p\left(\hat{t}_{x,HT}\right)} t_x = t_y.$$

However, it can be shown that the ratio estimator is asymptotically design-unbiased in the sense that its square bias is negligible in front of its variance for sufficiently large values of n. The design-variance of $\hat{t}_{y,ra}$ is untractable. Therefore, we apply the first-order Taylor expansion to approximate its design-variance. It is given by

$$AV_p\left(\hat{t}_{y,ra}\right) = \sum_{i \in U} \sum_{j \in U} \Delta_{ij} \frac{E_i}{\pi_i} \frac{E_j}{\pi_j},$$

where $E_i = y_i - R x_i$ with $R = t_y/t_x$. An estimator of $AV_p\left(\hat{t}_{y,ra}\right)$ is given by

$$\hat{V}\left(t_{y,ra}\right) = \sum_{i \in S} \sum_{j \in S} \frac{\Delta_{ij}}{\pi_{ij}} \frac{e_i}{\pi_i} \frac{e_j}{\pi_j},$$

where $e_i = y_i - \hat{R} x_i$ with $\hat{R} = \hat{t}_{y,HT}/\hat{t}_{x,HT}$. Under some regularity conditions, it can be shown that $\hat{V}\left(t_{y,ra}\right)$ is a design-consistent estimator of $AV_p\left(\hat{t}_{y,ra}\right)$. See Särndal et al. (1992, Section 5.6).

1.4.2 Generalized Regression (GREG) Estimator

We assume that a vector of auxiliary variables $\mathbf{x}_i = (x_{1i}, ..., x_{Ji})^\top$ is available for all the sample units, and that the vector of population totals, $\mathbf{t}_\mathbf{x} = (t_{x_1}, ..., t_{x_J})^\top$, is known. We assume that the relationship between y and $\mathbf{x}$ can be described by

$$m : y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \tag{1.4.3}$$

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_J)^\top$ is a J-vector of unknown parameters. We assume that $\mathbb{E}_m(\epsilon_i|\mathbf{x}_i) = 0$, $\mathbb{E}_m(\epsilon_i\epsilon_j|\mathbf{x}_i, \mathbf{x}_j) = 0$ if $i \neq j$ and $\mathbb{V}_m(\epsilon_i|\mathbf{x}_i) = \sigma^2 c_i$, where c_i is a known quantity.

The population total t_y may be decomposed as

$$t_y = \sum_{i\in U} y_i = \sum_{i\in U} \left(\mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i\right) = \sum_{i\in U} \mathbf{x}_i^\top \boldsymbol{\beta} + \sum_{i\in U} \epsilon_i.$$

The GREG estimator is obtained by estimating each total separately, which leads to

$$\hat{t}_{y,GREG} = \sum_{i\in U} \mathbf{x}_i^\top \hat{\boldsymbol{B}} + \sum_{i\in S} d_i e_i, \tag{1.4.4}$$

where $\hat{\boldsymbol{B}}$ is the weighted least squares estimator of $\boldsymbol{\beta}$, given by

$$\hat{\boldsymbol{B}} = \left(\sum_{i\in S} d_i \mathbf{x}_i c_i^{-1} \mathbf{x}_i^\top\right)^{-1} \sum_{i\in S} d_i \mathbf{x}_i c_i^{-1} y_i,$$

and $e_i = y_i - \mathbf{x}_i^\top \hat{\boldsymbol{B}}$ denotes the sample residual for unit i. The GREG estimator can also be written as

$$\hat{t}_{y,GREG} = \sum_{i\in S} w_i y_i,$$

where

$$w_i = d_i \times \left\{1 + c_i^{-1} \left(\mathbf{t}_\mathbf{x} - \hat{\mathbf{t}}_{\mathbf{x},HT}\right)^\top \hat{\mathbf{T}}^{-1} \mathbf{x}_i\right\}$$

with

$$\hat{\mathbf{t}}_{\mathbf{x},HT} = \sum_{i\in S} d_i \mathbf{x}_i,$$

and

$$\hat{\mathbf{T}} = \sum_{i\in S} d_i \mathbf{x}_i c_i^{-1} \mathbf{x}_i^\top.$$

The GREG estimator can also be viewed as the sum of the Horvitz-Thompson estimator and an adjustment term:

$$\hat{t}_{y,GREG} = \hat{t}_{y,HT} + \left(\mathbf{t}_\mathbf{x} - \hat{\mathbf{t}}_{\mathbf{x},HT}\right)^\top \hat{\boldsymbol{B}}. \tag{1.4.5}$$

We can rewrite (1.4.5) as

$$\hat{t}_{y,GREG} = \hat{t}_{y,HT} + \left(\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x},HT}\right)^{\top} \boldsymbol{B} + (\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x},HT})^{\top}(\hat{\boldsymbol{B}} - \boldsymbol{B}), \tag{1.4.6}$$

where

$$\boldsymbol{B} = \left(\sum_{i\in U} \mathbf{x}_i c_i^{-1} \mathbf{x}_i^{\top}\right)^{-1} \sum_{i\in U} \mathbf{x}_i c_i^{-1} y_i.$$

is the weighted least squares estimator of $\boldsymbol{\beta}$ if the model was fitted at the population level. The third term on the right hand-side of (1.4.6) is negligible in front of the first two when the sample size n is sufficiently large. It follows that

$$\begin{aligned}\hat{t}_{y,GREG} &\approx \hat{t}_{y,HT} + \left(\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x},HT}\right)^{\top} \boldsymbol{B} \\ &= \sum_{i\in U} \mathbf{x}_i^{\top} \boldsymbol{B} + \sum_{i\in S} d_i E_i \\ &= \sum_{i\in S} d_i y_i + \left(\sum_{i\in U} \mathbf{x}_i - \sum_{i\in S} d_i \mathbf{x}_i\right)^{\top} \boldsymbol{B}\end{aligned} \tag{1.4.7}$$

where $E_i = y_i - \mathbf{x}_i^{\top}\boldsymbol{B}$. The second term of the right hand-side of (1.4.7) is, on average, equal to 0. This suggests that $\hat{t}_{y,GREG}$ is asymptotically design-unbiased for t_y. The design-variance of the GREG estimator is intractable, so we use first-order Taylor expansion to approximate its design-variance. This leads to

$$AV_p(\hat{t}_{y,GREG}) = \sum_{i\in U} \sum_{j\in U} \Delta_{ij} \frac{E_i}{\pi_i} \frac{E_j}{\pi_j}.$$

If the working model is misspecified, some of the residuals E_i may be large, which may lead to a large variance. An estimator of $AV_p(\hat{t}_{y,GREG})$ is given by

$$\hat{V}(\hat{t}_{y,GREG}) = \sum_{i\in S} \sum_{j\in S} \frac{\Delta_{ij}}{\pi_{ij}} \frac{e_i}{\pi_i} \frac{e_j}{\pi_j}.$$

Proposition 1.4.1. *If there exists a J-vector of constants $\boldsymbol{\lambda}$ such that, for all $i \in U$, $c_i = \boldsymbol{\lambda}^{\top}\mathbf{x}_i$, then $\sum_{i\in S} d_i e_i = 0$ and the GREG estimator* (1.4.4) *reduces to its projection form:*

$$\hat{t}_{y,GREG} = \sum_{i\in U} \mathbf{x}_i^{\top} \hat{\boldsymbol{B}}. \tag{1.4.8}$$

Proof: Given that $c_i = \boldsymbol{\lambda}^\top \mathbf{x}_i$, we have $\boldsymbol{\lambda}^\top \mathbf{x}_i c_i^{-1} = 1$. It follows that

$$\begin{aligned}
\sum_{i \in S} d_i e_i &= \sum_{i \in S} d_i (y_i - \mathbf{x}_i^\top \hat{\boldsymbol{B}}) \\
&= \sum_{i \in S} d_i \boldsymbol{\lambda}^\top \mathbf{x}_i c_i^{-1} \left(y_i - \mathbf{x}_i^\top \left(\sum_{i \in S} d_i \mathbf{x}_i c_i^{-1} \mathbf{x}_i^\top \right)^{-1} \sum_{i \in S} d_i \mathbf{x}_i c_i^{-1} y_i \right) \\
&= \sum_{i \in S} d_i \boldsymbol{\lambda}^\top \mathbf{x}_i c_i^{-1} y_i - \sum_{i \in S} d_i \boldsymbol{\lambda}^\top \mathbf{x}_i c_i^{-1} \mathbf{x}_i^\top \left(\sum_{i \in S} d_i \mathbf{x}_i c_i^{-1} \mathbf{x}_i^\top \right)^{-1} \sum_{i \in S} d_i \mathbf{x}_i c_i^{-1} y_i \\
&= \boldsymbol{\lambda}^\top \left(\sum_{i \in S} d_i \mathbf{x}_i c_i^{-1} y_i - \sum_{i \in S} d_i \mathbf{x}_i c_i^{-1} \mathbf{x}_i^\top \left(\sum_{i \in S} d_i \mathbf{x}_i c_i^{-1} \mathbf{x}_i^\top \right)^{-1} \sum_{i \in S} d_i \mathbf{x}_i c_i^{-1} y_i \right) \\
&= \boldsymbol{\lambda}^\top \left(\sum_{i \in S} d_i \mathbf{x}_i c_i^{-1} y_i - \sum_{i \in S} d_i \mathbf{x}_i c_i^{-1} y_i \right) \\
&= 0.
\end{aligned} \tag{1.4.9}$$

Using (1.4.9) in (1.4.4), leads to

$$\hat{t}_{y,GREG} = \sum_{i \in U} \mathbf{x}_i^\top \hat{\boldsymbol{B}}.$$

∎

The condition $c_i = \boldsymbol{\lambda}^\top \mathbf{x}_i$ is satisfied, for instance, when the model contains the intercept, i.e., the first component of $\mathbf{x}_i$ is equal to 1 and $\mathbb{V}_m(\epsilon_i | \mathbf{x}_i) = \sigma^2$ (i.e., $c_i = 1$). In this case, we have $\boldsymbol{\lambda} = (1, 0, ..., 0)^\top$. Also, if $c_i = \mathbf{x}_{ji}$ for $j = 1, ..., J$, then we set $\boldsymbol{\lambda} = (0, ..., 1, ..., 0)^\top$, where the component 1 is in the j-th position.

1.5 Model-based Approach

In the model-based approach, we assume that the random variables Y_i are generated from a certain model conditional on the vector of auxiliary variables $\mathbf{x}$. We assume that a noninformative sample s, of size n, is selected from U according to a probability or a non-probability sampling procedure. Sampling is called noninformative if the model that holds in the population also holds in the sample. In the model-based approach, when evaluating the properties of predictors, all quantities but Y are treated as fixed. The population total $t_y = \sum_{i \in U} Y_i$ is a random quantity, which can be expressed as

$$t_y = \sum_{i \in s} Y_i + \sum_{i \in U-s} Y_i$$

and $\sum_{i\in U-s} Y_i$ is the part we wish to predict.

Properties of predictors in the model-based approach are evaluated with respect to the selected model conditional on the realized sample. Let $\hat{\theta}$ denote a predictor of θ_N. Let $\mathbb{E}_m(\hat{\theta})$ be the expectation of the predictor $\hat{\theta}$. The prediction bias of $\hat{\theta}$ is defined as $\mathbb{E}_m(\hat{\theta}-\theta_N|s)$, and its prediction variance is defined as $\mathbb{V}_m(\hat{\theta}-\theta_N|s)$.

1.5.1 Best Linear Unbiased Predictor (BLUP)

Royall (1976) proposed the Best Linear Unbiased Predictor (BLUP) of $t_y = \sum_{i\in U} Y_i$, which is given by

$$\hat{t}_{y,BLUP} = \sum_{i\in S} w_i Y_i, \tag{1.5.1}$$

where the weights w_i's have to be determined. We seek a weighting system $\{w_i; i \in s\}$ that satisfies model-unbiasedness, i.e.,

$$\mathbb{E}_m(\hat{t}_{y,BLUP} - t_y|s) = 0 \tag{1.5.2}$$

and such that the prediction variance $\mathbb{V}_m(\hat{t}_{y,BLUP} - t_y|s)$ is minimized.

Proposition 1.5.1. *The weights w_i that minimize $\mathbb{V}_m(\hat{t}_{y,BLUP} - t_y|s)$ subject to* (1.5.2) *are given by*

$$w_i = 1 + c_i^{-1}\boldsymbol{x}_i^\top \left(\sum_{i\in s} \boldsymbol{x}_i c_i^{-1}\boldsymbol{x}_i^\top\right)^{-1} \left(\sum_{i\in U} \boldsymbol{x}_i - \sum_{i\in s} \boldsymbol{x}_i\right). \tag{1.5.3}$$

Proof: We consider the predictor of the form

$$\begin{aligned}\hat{t}_y &= \sum_{i\in s} w_i Y_i \\ &= \sum_{i\in s} Y_i + \sum_{i\in s}(w_i - 1)Y_i \\ &= t_{y,s} + \sum_{i\in s} u_i Y_i,\end{aligned}$$

where $u_i = w_i - 1$, and $t_{y,s} = \sum_{i\in s} Y_i$. The prediction bias of $\hat{t}_y$ can be expressed as

$$\begin{aligned}\mathbb{E}_m(\hat{t}_y - t_y|s) &= \mathbb{E}_m\left(\sum_{i\in s} u_i Y_i - \sum_{k\in U-s} Y_i|s\right) \\ &= \sum_{i\in s} u_i \mathbb{E}_m(Y_i|\mathbf{x}_i) - \sum_{i\in U-s} \mathbb{E}_m(Y_i|\mathbf{x}_i)\end{aligned}$$

$$= \sum_{i \in s} u_i \mathbf{x}_i^\top \boldsymbol{\beta} - \sum_{i \in U-s} \mathbf{x}_i^\top \boldsymbol{\beta}$$
$$= \left(\sum_{i \in s} u_i \mathbf{x}_i^\top - \sum_{i \in U-s} \mathbf{x}_i^\top \right) \boldsymbol{\beta}.$$

We want $\mathbb{E}_m(\hat{t}_y - t_y | s) = 0$ for every $\boldsymbol{\beta}$, which implies that

$$\left(\sum_{i \in s} u_i \mathbf{x}_i^\top - \sum_{i \in U-s} \mathbf{x}_i^\top \right) = 0.$$

The prediction variance of $\hat{t}_y$ can be expressed as

$$\begin{aligned} \mathbb{V}_m(\hat{t}_y - t_y | s) &= \mathbb{V}_m \left(\sum_{i \in s} u_i Y_i | s \right) + \mathbb{V}_m \left(\sum_{i \in U-s} Y_i | s \right) - 2Cov_m \left(\sum_{i \in s} u_i Y_i, \sum_{i \in U-s} Y_i | s \right) \\ &= \sum_{i \in s} u_i^2 \mathbb{V}_m(Y_i | \mathbf{x}_i) + \sum_{i \in U-s} \mathbb{V}_m(Y_i | \mathbf{x}_i) \\ &= \sum_{i \in s} \sigma^2 c_i u_i^2 + \sum_{i \in U-s} \sigma^2 c_i \\ &= \sigma^2 \left(\sum_{i \in s} c_i u_i^2 + \sum_{i \in U-s} c_i \right), \end{aligned} \tag{1.5.4}$$

noting that $Cov_m \left(\sum_{i \in s} u_i Y_i, \sum_{i \in U-s} Y_i \right) = 0$. We use Lagrange L for the minimization problem. Let

$$L(u_1, ..., u_n, \boldsymbol{\lambda}) = \frac{1}{2} \left(\sum_{i \in s} c_i u_i^2 + \sum_{i \in U-s} c_i \right) - \left(\sum_{i \in s} u_i \mathbf{x}_i^\top - \sum_{i \in U-s} \mathbf{x}_i^\top \right) \boldsymbol{\lambda},$$

where $\boldsymbol{\lambda}$ is a vector of Lagrange multipliers.

Taking the first-order partial derivative of L with respect to u_i, we have

$$\frac{\partial L(u_1, ..., u_n, \boldsymbol{\lambda})}{\partial u_i} = c_i u_i - \mathbf{x}_i^\top \boldsymbol{\lambda}.$$

Let $\frac{\partial L(u_1, ..., u_n, \boldsymbol{\lambda})}{\partial u_i} = 0$, we have

$$u_i = c_i^{-1} \mathbf{x}_i^\top \boldsymbol{\lambda}. \tag{1.5.5}$$

Taking the first-order partial derivative of L with respect to $\boldsymbol{\lambda}$, we have

$$\frac{\partial L(u_1, ..., u_n, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = \sum_{i \in s} u_i \mathbf{x}_i - \sum_{i \in U-s} \mathbf{x}_i. \tag{1.5.6}$$

Let $\frac{\partial L(u_1,\ldots,u_n,\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = 0$, and plug-in (1.5.5) in (1.5.6). It leads to

$$\boldsymbol{\lambda} = \left(\sum_{i\in s} \mathbf{x}_i c_i^{-1} \mathbf{x}_i^\top\right)^{-1} \left(\sum_{i\in U-s} \mathbf{x}_i\right). \tag{1.5.7}$$

Using (1.5.7) in (1.5.5), leads to

$$\begin{aligned} u_i &= c_i^{-1}\mathbf{x}_i^\top \left(\sum_{i\in s} \mathbf{x}_i c_i^{-1} \mathbf{x}_i^\top\right)^{-1} \left(\sum_{k\in U-s} \mathbf{x}_i\right) \\ &= c_i^{-1}\mathbf{x}_i^\top \left(\sum_{i\in s} \mathbf{x}_i c_i^{-1} \mathbf{x}_i^\top\right)^{-1} \left(\sum_{i\in U} \mathbf{x}_i - \sum_{i\in s} \mathbf{x}_i\right). \end{aligned}$$

Thus, the optimal weights are given by

$$\begin{aligned} w_i &= u_i + 1 \\ &= 1 + c_i^{-1}\mathbf{x}_i^\top \left(\sum_{i\in s} \mathbf{x}_i c_i^{-1} \mathbf{x}_i^\top\right)^{-1} \left(\sum_{i\in U} \mathbf{x}_i - \sum_{i\in s} \mathbf{x}_i\right) \end{aligned}$$

The BLUP of t_y is thus given by

$$\hat{t}_{y,BLUP} = \sum_{i\in s} w_i Y_i,$$

where w_i is given by (1.5.3).

∎

The predictor $\hat{t}_{y,BLUP}$ can also be written as

$$\hat{t}_{y,BLUP} = \sum_{i\in s} Y_i + \sum_{i\in U-s} \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_{WLS}, \tag{1.5.8}$$

where $\hat{\boldsymbol{\beta}}_{WLS}$ is the weighted least squares estimator of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_{WLS} = \left(\sum_{i\in s} \mathbf{x}_i c_i^{-1} \mathbf{x}_i^\top\right)^{-1} \left(\sum_{i\in s} \mathbf{x}_i c_i^{-1} y_i\right). \tag{1.5.9}$$

From (1.5.4), the prediction variance of the BLUP is given by

$$\mathbb{V}_m(\hat{t}_{y,BLUP} - t_y|s) = \sigma^2 \left(\sum_{i\in s} (w_i - 1)^2 c_i + \sum_{i\in U-s} c_i\right). \tag{1.5.10}$$

The prediction variance of $\hat{t}_{y,BLUP}$ is small if $\sigma_i^2 = \sigma^2 c_i$ is small for all i, then the model is considered to be highly predictive. However, the prediction variance (1.5.10) is unknown as σ^2 is unknown. A model-unbiased estimator of σ^2 under the model (1.4.3) is given by

$$\hat{\sigma}^2 = \frac{1}{n-J} \sum_{i \in s} c_i^{-1} (Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_{WLS})^2.$$

It follows that a model-unbiased estimator of the prediction variance (1.5.10) is given by

$$\hat{\mathbb{V}}_m(\hat{t}_{y,BLUP} - t_y | s) = \hat{\sigma}^2 \left(\sum_{i \in s} (w_i - 1)^2 c_i + \sum_{i \in U - s} c_i \right).$$

1.5.2 Special Cases of the BLUP

In this section, we consider three special cases of (1.5.1).

(i) The common mean model:

$$Y_i = \beta + \epsilon_i,$$

with $\mathbb{E}_m(\epsilon_i) = 0$, $\mathbb{E}_m(\epsilon_i \epsilon_j) = 0$ and $\mathbb{V}_m(\epsilon_i) = \sigma^2$. Under the common mean model, the weights (1.5.3) reduce to

$$w_i = \frac{N}{n}$$

for all i. As a result, the BLUP (1.5.1) reduces to

$$\hat{t}_{y,BLUP} = N\bar{y},$$

where

$$\bar{y} = \frac{1}{n} \sum_{i \in s} Y_i.$$

This is the Horvitz-Thompson estimator in the case of simple random sampling without replacement.

(ii) The ratio model:

$$Y_i = \beta x_i + \epsilon_i,$$

with $\mathbb{E}_m(\epsilon_i | x_i) = 0$, $\mathbb{E}_m(\epsilon_i \epsilon_j | x_i, x_j) = 0$ and $\mathbb{V}_m(\epsilon_i | x_i) = \sigma^2 x_i$. Under the ratio model, the weights (1.5.3) reduce to

$$w_i = \frac{N}{n} \frac{\bar{X}}{\bar{x}}$$

for all i, where $\bar{X} = \frac{1}{n}\sum_{i\in U} x_i$, and $\bar{x} = \frac{1}{n}\sum_{i\in s} x_i$. As a result, the BLUP (1.5.1) reduces to

$$\hat{t}_{y,BLUP} = \frac{\bar{y}}{\bar{x}} t_x.$$

This is the ratio estimator in the case of simple random sampling without replacement.

(iii) The simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

with $\mathbb{E}_m(\epsilon_i|x_i) = 0$, $\mathbb{E}_m(\epsilon_i\epsilon_j|x_i, x_j) = 0$ and $\mathbb{V}_m(\epsilon_i|x_i) = \sigma^2$. Under the simple linear regression model, the weights (1.5.3) reduce to

$$w_i = \frac{N}{n}\left\{1 + n\frac{(x_i - \bar{x})(\bar{X} - \bar{x})}{\sum_{i\in s}(x_i - \bar{x})^2}\right\}$$

for all i. Thus, the BLUP (1.5.1) reduces to

$$\hat{t}_{y,BLUP} = N\left[\bar{y} + \hat{\beta}_1(\bar{X} - \bar{x})\right],$$

where

$$\hat{\beta}_1 = \frac{\sum_{i\in s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i\in s}(x_i - \bar{x})^2}.$$

This is the simple linear regression estimator in the case of simple random sampling without replacement.

Chapter 2

Influential units and conditional bias

2.1 Influential units

A measurement error occurs when a unit is selected and observed but the recorded value is incorrect. An influential unit is not a measurement error but a legitimate sample unit that may be representing other similar units in the nonsample part of the finite population. An influential unit may have a drastic impact on the estimate if it was to be excluded from the sample.

In classical statistics, where the observations are assumed to be independent and identically distributed (i.i.d), a sample s can be viewed as a sequence of n independent observations which are generated according to a given distribution with mean μ and variance σ^2. A simple estimator of μ is the sample mean $\bar{y} = n^{-1} \sum_{i \in s} y_i$, which is unbiased regardless of the underlying distribution but may be highly unstable in the presence of influential units. A parametric alternative is the maximum likelihood estimator (MLE) of μ, which is asymptotically unbiased and optimal if the underlying distribution is known. However, in practice, the underlying true distribution is unknown and it may be hard to postulate an appropriate model unless the sample size is large since the influential units are usually found in the tails.

In survey sampling, both sample and nonsample units can be influential. In the design-based approach, the Horvitz-Thompson estimator is design-unbiased but vulnerable to the presence of influential units, i.e., it becomes highly unstable (resulting in large variance). In the model-based approach, the BLUP is optimal in the class of linear unbiased estimators even when the errors are not normally distributed. However, the BLUP tends to be unstable (i.e., large variance) when the distribution of errors is highly skewed; see Chambers (1986).

Classical estimators and predictors can be highly impacted by influential units. We would like to measure the influence of each unit so that we can reduce the estimation and prediction errors by downweighting those units with a large influence. In the next section, we present the conditional bias as a measure of influence of a unit.

2.2 Measure of the influence: conditional bias

In the classical i.i.d. setup, Muñoz-Pichardo et al. (1995) proposed to use the conditional bias as a measure of the influence of a unit. Let θ be a parameter and $\hat{\theta}$ be an estimator of θ. In the context of classical statistics, the conditional bias of the unit i is given by

$$B_i(y_i; \theta) = \mathbb{E}_F\left(\hat{\theta}|Y_i = y_i\right) - \theta,$$

where $\mathbb{E}_F(\cdot)$ denotes the expectation evaluated with respect to the distribution F. Beaumont et al. (2013) showed that the conditional bias of unit i can be viewed as its contribution to the error, $\hat{\theta} - \theta$. That is,

$$\hat{\theta} - \theta \simeq \sum_{i=1}^{n} B_i(y_i; \theta).$$

Therefore, the conditional bias of a unit can be viewed as a measure of its contribution to the error $\hat{\theta} - \theta$. Next, we present the concept of conditional bias in the context of survey sampling.

2.2.1 Design-based approach

Let $\hat{t}_y$ be an estimator of the population total $t_y = \sum_{i \in U} y_i$. In the design-based approach, the conditional bias attached to the ith sample unit (i.e., $I_i = 1$) with respect to $\hat{t}_y$ is defined as

$$B_{1i} = \mathbb{E}_p\left(\hat{t}_y|I_i = 1\right) - \mathbb{E}_p(\hat{t}_y). \tag{2.2.1}$$

For a nonsample unit, the conditional bias is defined as

$$B_{0i} = \mathbb{E}_p\left(\hat{t}_y|I_i = 0\right) - \mathbb{E}_p(\hat{t}_y); \tag{2.2.2}$$

see Moreno-Rebollo et al. (1999, 2002). Since $\mathbb{E}_p(\hat{t}_y) = \mathbb{E}\left(\mathbb{E}_p(\hat{t}_y)|I_i\right) = \pi_i \mathbb{E}_p(\hat{t}_y|I_i = 1) + (1 - \pi_i)\mathbb{E}_p(\hat{t}_y|I_i = 0)$, we can rewrite (2.2.2) as

$$\begin{aligned}
B_{0i} &= \mathbb{E}_p\left(\hat{t}_y|I_i = 0\right) - \pi_i \mathbb{E}_p(\hat{t}_y|I_i = 1) - (1 - \pi_i)\mathbb{E}_p(\hat{t}_y|I_i = 0) \\
&= \pi_i\left(\mathbb{E}_p\left(\hat{t}_y|I_i = 0\right) - \mathbb{E}_p(\hat{t}_y|I_i = 1)\right) \\
&= \pi_i\left(\mathbb{E}_p\left(\hat{t}_y|I_i = 0\right) - \mathbb{E}_p(\hat{t}_y) - \mathbb{E}_p(\hat{t}_y|I_i = 1) + \mathbb{E}_p(\hat{t}_y)\right) \\
&= \pi_i\left(B_{0i} - B_{1i}\right).
\end{aligned}$$

It follows that

$$B_{0i} = -\frac{\pi_i}{1 - \pi_i} B_{1i}.$$

Proposition 2.2.1. *When $\hat{t}_y$ is the Horvitz-Thompson estimator $\hat{t}_{y,HT}$, the conditional bias of a sample unit is given by*

$$B_{1i}^{HT} = \sum_{j \in U} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1 \right) y_j. \tag{2.2.3}$$

Proof:

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

∎

At the estimation stage, nothing can be done about the nonsample units and, only the influence of sample units can be reduced. Now, the conditional bias (2.2.3) is generally unknown. Therefore, it must be estimated. Provided that $\pi_{ij} > 0$ for all $j \in U$, a conditionally design-unbiased estimator of B_{1i}^{HT} is given by

$$\hat{B}_{1i}^{HT} = \sum_{j \in S} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_j \pi_{ij}} \right) y_j. \tag{2.2.4}$$

Proposition 2.2.2. *The estimator $\hat{B}_{1i}^{HT}$ in* (2.2.4) *is conditionally design-unbiased in the sense that*

$$\mathbb{E}_p \left(\hat{B}_{1i}^{HT} | I_i = 1 \right) = B_{1i}^{HT}.$$

Proof:

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

∎

For stratified simple random sampling without replacement, the conditional bias (2.2.3) reduces to

$$B_{1i}^{HT} = \frac{N_h}{N_h - 1}\left(\frac{N_h}{n_h} - 1\right)(y_i - \bar{Y}_h), \quad i \in U_h, \tag{2.2.5}$$

where $\bar{Y}_h = N_h^{-1}\sum_{i\in U_h} y_i$. From (2.2.5), a unit in stratum h has a large influence when it is far from the stratum mean $\bar{Y}_h$ or if the sampling fraction n_h/N_h in stratum h is small. It follows from (2.2.4) that

$$\hat{B}_{1i}^{HT} = \frac{n_h}{n_h - 1}\left(\frac{N_h}{n_h} - 1\right)(y_i - \bar{y}_h), \quad i \in U_h,$$

where $\bar{y}_h = n_h^{-1}\sum_{i\in S_h} y_i$. For Poisson sampling, the conditional bias (2.2.4) reduces to

$$B_{1i}^{HT} = (d_i - 1)y_i,$$

where $d_i = 1/\pi_i$. Here, the conditional bias is known for all the sample units, so it does not need to be estimated. Unit i has a large influence if its design weight, d_i is large and/or if its y-value, y_i, is large.

Proposition 2.2.3. *For Poisson sampling, the sampling error of $\hat{t}_{y,HT}$ can be written as*

$$\hat{t}_{y,HT} - t_y = \sum_{i\in S} B_{1i}^{HT} + \sum_{i\in U-S} B_{0i}^{HT},$$

For simple random sampling without replacement, we have

$$\hat{t}_{y,HT} - t_y \approx \sum_{i\in S} B_{1i}^{HT} + \sum_{i\in U-S} B_{0i}^{HT},$$

provided that N is large.

Proof: For Poisson sampling, we have

$$\begin{aligned}\sum_{i\in S} B_{1i}^{HT} + \sum_{i\in U-S} B_{0i}^{HT} &= \sum_{i\in S}(d_i-1)y_i + \sum_{i\in U-S}\left(-(d_i-1)^{-1}(d_i-1)y_i\right)\\ &= \sum_{i\in S} d_i y_i - \sum_{i\in S} y_i - \sum_{i\in U-S} y_i\\ &= \sum_{i\in S} d_i y_i - \sum_{i\in U} y_i\\ &= \hat{t}_{y,HT} - t_y.\end{aligned}$$

For simple random sampling without replacement, we have

$$\begin{aligned}\sum_{i\in S} B_{1i}^{HT} + \sum_{i\in U-S} B_{0i}^{HT} &= \sum_{i\in S}\frac{N}{N-1}\left(\frac{N}{n}-1\right)(y_i-\bar{Y})\\ &+ \sum_{i\in U-S}\left(-(d_i-1)^{-1}\frac{N}{N-1}\left(\frac{N}{n}-1\right)(y_i-\bar{Y})\right)\\ &= \sum_{i\in S}\frac{N}{N-1}(d_i-1)(y_i-\bar{Y}) - \sum_{i\in U-S}\frac{N}{N-1}(y_i-\bar{Y})\\ &= \sum_{i\in S}\frac{N}{N-1}(d_i-1)(y_i-\bar{Y}) - \sum_{i\in U}\frac{N}{N-1}(y_i-\bar{Y})\\ &+ \sum_{i\in S}\frac{N}{N-1}(y_i-\bar{Y})\\ &= \sum_{i\in S}\frac{N}{N-1}d_i(y_i-\bar{Y}) - \sum_{i\in U}\frac{N}{N-1}(y_i-\bar{Y})\\ &= \sum_{i\in S}\frac{N}{N-1}d_i y_i - \sum_{i\in S}\frac{N}{N-1}d_i\bar{Y} - \sum_{i\in U}\frac{N}{N-1}y_i + \sum_{i\in U}\frac{N}{N-1}\bar{Y}\\ &= \frac{N}{N-1}\left(\sum_{i\in S} d_i y_i - \sum_{i\in U} y_i\right).\end{aligned}$$

Provided that the population size N is large enough, we have $N/(N-1) \approx 1$ and the result follows. ∎

The design variance of $\hat{t}_{y,HT}$ can be expressed as a function of the conditional bias:

$$\mathbb{V}_p\left(\hat{t}_{y,HT}\right) = \sum_{i\in U} B_{1i}^{HT} y_i.$$

Therefore, the design variance of $\hat{t}_{y,HT}$ is large if the conditional bias B_{1i}^{HT} is large and/or the value of y_i is large. The design variance of $\hat{t}_{y,HT}$ is equal to 0 if and only if $B_{1i}^{HT} = 0$ for all $i \in U$.

2.2.2 Model-assisted approach

Let $\hat{t}_{y,ma}$ be a model-assisted estimator of t_y. The conditional bias of a sample unit with respect to $\hat{t}_{y,ma}$ is defined as

$$B_{1i}^{ma} = \mathbb{E}_p\left(\hat{t}_{y,ma} - t_y | I_i = 1\right).$$

Proposition 2.2.4. *For the GREG estimator, the conditional bias of a sample unit can be approximated by*

$$B_{1i}^{GREG} \approx \sum_{j\in U}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)E_j, \tag{2.2.6}$$

where

$$E_j = y_j - \mathbf{x}_j^\top \boldsymbol{B}$$

with

$$\boldsymbol{B} = \left(\sum_{i\in U}\mathbf{x}_i c_i^{-1}\mathbf{x}_i^\top\right)^{-1}\sum_{i\in U}\mathbf{x}_i c_i^{-1} y_i.$$

Proof: The conditional bias of GREG estimator attached to sample unit i is

$$B_{1i}^{GREG} = \mathbb{E}_p\left(\hat{t}_{y,GREG} - t_y | I_i = 1\right),$$

which is intractable because $\hat{t}_{y,GREG}$ is a complex function of estimated totals. We use a first-order Taylor expansion to obtain

$$\hat{t}_{y,GREG} - t_y = \sum_{j\in s}\frac{E_j}{\pi_j} - \sum_{j\in U}E_j + O_p\left(\frac{1}{n}\right).$$

Ignoring the higher-order terms, the conditional bias of unit i can be approximated by

$$B_{1i}^{GREG} = \mathbb{E}_p\left(\sum_{j\in s}\frac{E_j}{\pi_j} - \sum_{j\in U}E_j | I_i = 1\right) \approx \sum_{j\in U}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)E_j.$$

∎

Therefore, a unit may have a large influence if it is associated with a large residual. An estimator of B_{1i}^{GREG} is given by

$$\hat{B}_{1i}^{GREG} = \sum_{j\in S}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}\pi_j}e_j, \tag{2.2.7}$$

where

$$e_j = y_j - \mathbf{x}_j^\top \hat{\boldsymbol{B}}$$

with

$$\hat{\boldsymbol{B}} = \left(\sum_{i\in S} d_i \mathbf{x}_i c_i^{-1} \mathbf{x}_i^\top\right)^{-1} \sum_{i\in S} d_i \mathbf{x}_i c_i^{-1} y_i.$$

For stratified simple random sampling without replacement, the conditional bias (2.2.6) reduces to

$$B_{1i}^{GREG} \approx \frac{N_h}{N_h - 1}\left(\frac{N_h}{n_h} - 1\right)(E_i - \bar{E}_h) \quad i \in U_h, \tag{2.2.8}$$

where $\bar{E}_h = N_h^{-1} \sum_{i\in U_h} E_i$. An estimator of (2.2.8) is given by

$$\hat{B}_{1i}^{GREG} = \frac{n_h}{n_h - 1}\left(\frac{N_h}{n_h} - 1\right)(e_i - \bar{e}_h),$$

where $\bar{e}_h = n_h^{-1} \sum_{i\in S_h} e_i$. For Poisson sampling, the conditional bias (2.2.6) reduces to

$$B_{1i}^{GREG} \approx (d_i - 1)E_i. \tag{2.2.9}$$

An estimator of (2.2.9) is given by

$$\hat{B}_{1i}^{GREG} = (d_i - 1)e_i.$$

2.2.3 Model-based approach

Let $\hat{t}_y$ be a predictor of the population total $t_y = \sum_{i\in U} Y_i$. The conditional bias of a sample unit i with respect to $\hat{t}_y$ is defined as

$$B_i(y_i; \boldsymbol{\beta}) = \mathbb{E}_m(\hat{t}_y - t_y | s, Y_i = y_i). \tag{2.2.10}$$

Proposition 2.2.5. *Consider the BLUP of the form*

$$\hat{t}_{y,BLUP} = \sum_{i\in s} w_i Y_i,$$

where w_i is given by (1.5.3)*. The conditional bias* (2.2.10) *of unit i is given by*

$$B_i(y_i; \boldsymbol{\beta}) = \begin{cases} (w_i - 1)(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}) & i \in s \\ -(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}) & i \in U - s \end{cases}. \tag{2.2.11}$$

Proof: Using (1.5.3), we have

$$\begin{aligned}
\sum_{i\in s} w_i\mathbf{x}_i &= \sum_{i\in s}\left\{1 + c_i^{-1}\mathbf{x}_i^\top\left(\sum_{i\in s}\mathbf{x}_i c_i^{-1}\mathbf{x}_i^\top\right)^{-1}\left(\sum_{i\in U}\mathbf{x}_i - \sum_{i\in s}\mathbf{x}_i\right)\right\}\mathbf{x}_i \\
&= \sum_{i\in s}\left\{\mathbf{x}_i + \mathbf{x}_i c_i^{-1}\mathbf{x}_i^\top\left(\sum_{i\in s}\mathbf{x}_i c_i^{-1}\mathbf{x}_i^\top\right)^{-1}\left(\sum_{i\in U-s}\mathbf{x}_i\right)\right\} \\
&= \sum_{i\in s}\mathbf{x}_i + \sum_{i\in s}\mathbf{x}_i c_i^{-1}\mathbf{x}_i^\top\left(\sum_{i\in s}\mathbf{x}_i c_i^{-1}\mathbf{x}_i^\top\right)^{-1}\left(\sum_{i\in U-s}\mathbf{x}_i\right) \\
&= \sum_{i\in s}\mathbf{x}_i + \sum_{i\in U-s}\mathbf{x}_i \\
&= \sum_{i\in U}\mathbf{x}_i.
\end{aligned} \tag{2.2.12}$$

For a sample unit i,

$$\begin{aligned}
B_i(y_i;\boldsymbol{\beta}) &= \mathbb{E}_m(\hat{t}_{y,BLUP} - t_y | s, Y_i = y_i) \\
&= \mathbb{E}_m\left(\sum_{\substack{j\in s\\ j\neq i}} w_j y_j + w_i y_i - \sum_{\substack{j\in U\\ j\neq i}} y_j + y_i | s, Y_i = y_i\right) \\
&= (w_i - 1)y_i + \mathbb{E}_m\left(\sum_{\substack{j\in s\\ j\neq i}} w_j y_j - \sum_{\substack{j\in U\\ j\neq i}} y_j | s, Y_i = y_i\right) \\
&= (w_i - 1)y_i + \sum_{\substack{j\in s\\ j\neq i}} w_j \mathbb{E}_m(y_j|\mathbf{x}_j) - \sum_{\substack{j\in U\\ j\neq i}} \mathbb{E}_m(y_j|\mathbf{x}_j) \\
&= (w_i - 1)y_i + \sum_{\substack{j\in s\\ j\neq i}} w_j \mathbf{x}_j^\top\boldsymbol{\beta} - \sum_{\substack{j\in U\\ j\neq i}} \mathbf{x}_j^\top\boldsymbol{\beta} \\
&= (w_i - 1)y_i + \sum_{\substack{j\in s\\ j\neq i}} (w_j - 1)\mathbf{x}_j^\top\boldsymbol{\beta} - \sum_{j\in U-s} \mathbf{x}_j^\top\boldsymbol{\beta} \\
&= (w_i - 1)y_i + \sum_{\substack{j\in s\\ j\neq i}} (w_j - 1)\mathbf{x}_j^\top\boldsymbol{\beta} - \sum_{j\in U} \mathbf{x}_j^\top\boldsymbol{\beta} + \sum_{j\in s} \mathbf{x}_j^\top\boldsymbol{\beta}.
\end{aligned}$$

Using (2.2.12), we have

$$\begin{aligned}
B_i(y_i;\boldsymbol{\beta}) &= (w_i-1)y_i + \sum_{\substack{j\in s\\ j\neq i}}(w_j-1)\mathbf{x}_j^\top\boldsymbol{\beta} - \sum_{j\in s} w_j\mathbf{x}_j^\top\boldsymbol{\beta} + \sum_{j\in s}\mathbf{x}_j^\top\boldsymbol{\beta} \\
&= (w_i-1)y_i + \sum_{\substack{j\in s\\ j\neq i}}(w_j-1)\mathbf{x}_j^\top\boldsymbol{\beta} - \sum_{j\in s}(w_j-1)\mathbf{x}_j^\top\boldsymbol{\beta} \\
&= (w_i-1)y_i - (w_j-1)\mathbf{x}_j^\top\boldsymbol{\beta} \\
&= (w_i-1)(y_i-\mathbf{x}_i^\top\boldsymbol{\beta}).
\end{aligned}$$

For a nonsample unit,

$$\begin{aligned}
B_i(y_i;\boldsymbol{\beta}) &= \mathbb{E}_m(\hat{t}_{y,BLUP} - t_y|s, Y_i = y_i) \\
&= \mathbb{E}_m\left(\sum_{i\in s} w_i y_i - \sum_{\substack{j\in U\\ j\neq i}} y_j - y_i\right) \\
&= \sum_{i\in s} w_i \mathbb{E}_m(y_i|\mathbf{x}_i) - \sum_{\substack{j\in U\\ j\neq i}}\mathbb{E}_m(y_j|\mathbf{x}_j) - y_i \\
&= \sum_{i\in s} w_i\mathbf{x}_i^\top\boldsymbol{\beta} - \sum_{\substack{j\in U\\ j\neq i}}\mathbf{x}_i^\top\boldsymbol{\beta} - y_i
\end{aligned}$$

Using (2.2.12), we have

$$\begin{aligned}
B_i(y_i;\boldsymbol{\beta}) &= \sum_{i\in U}\mathbf{x}_i^\top\boldsymbol{\beta} - \sum_{\substack{j\in U\\ j\neq i}}\mathbf{x}_i^\top\boldsymbol{\beta} - y_i \\
&= -(y_i-\mathbf{x}_i^\top\boldsymbol{\beta}).
\end{aligned}$$

Thus, we have,

$$B_i(y_i;\boldsymbol{\beta}) = \begin{cases} (w_i-1)(y_i-\mathbf{x}_i^\top\boldsymbol{\beta}) & i\in s \\ -(y_i-\mathbf{x}_i^\top\boldsymbol{\beta}) & i\in U-s \end{cases}.$$

∎

A sample unit may have a large influence if the weight w_i is large and/or the residual $y_i - \mathbf{x}_i^\top\widetilde{\boldsymbol{\beta}}$ is large.

Proposition 2.2.6. *The prediction error of $\hat{t}_{y,BLUP}$ can be written as*

$$\hat{t}_{y,BLUP} - t_y = \sum_{i\in U} B_i(Y_i;\boldsymbol{\beta}).$$

Proof:

$$\begin{aligned}
\sum_{i\in U} B_i(Y_i;\boldsymbol{\beta}) &= \sum_{i\in s} B_i(Y_i;\boldsymbol{\beta}) + \sum_{i\in U-s} B_i(Y_i;\boldsymbol{\beta}) \\
&= \sum_{i\in s}(w_i-1)(Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}) + \sum_{i\in U-s}\left(-(Y_i-\mathbf{x}_i^\top\boldsymbol{\beta})\right) \\
&= \sum_{i\in s}(w_i-1)(Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}) - \sum_{i\in U}(Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}) + \sum_{i\in s}(Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}) \\
&= \sum_{i\in s} w_i(Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}) - \sum_{i\in U}(Y_i-\mathbf{x}_i^\top\boldsymbol{\beta}) \\
&= \sum_{i\in s} w_iY_i - \sum_{i\in U} Y_i - \left(\sum_{i\in s} w_i\mathbf{x}_i^\top\boldsymbol{\beta} - \sum_{i\in U}\mathbf{x}_i^\top\boldsymbol{\beta}\right).
\end{aligned}$$

Using (2.2.12), we have $\sum_{i\in s} w_i\mathbf{x}_i^\top\boldsymbol{\beta} - \sum_{i\in U}\mathbf{x}_i^\top\boldsymbol{\beta} = 0$, so that

$$\begin{aligned}
\sum_{i\in U} B_i(Y_i;\boldsymbol{\beta}) &= \sum_{i\in s} w_iY_i - \sum_{i\in U} Y_i \\
&= \hat{t}_{y,BLUP} - t_y
\end{aligned}$$

∎

Proposition 2.2.7. *The prediction variance of $\hat{t}_{y,BLUP}$ can be written as*

$$\mathbb{V}_m\left(\hat{t}_{y,BLUP} - t_y|s\right) = \mathbb{E}_m\left\{\sum_{i\in U} B_i^2(Y_i;\boldsymbol{\beta})|s\right\}.$$

Proof:

$$\begin{aligned}
\mathbb{E}_m\left\{\sum_{i\in U} B_i^2(Y_i;\boldsymbol{\beta})|s\right\} &= \mathbb{E}_m\left\{\sum_{i\in s} B_i^2(Y_i;\boldsymbol{\beta}) + \sum_{i\in U-s} B_i^2(Y_i;\boldsymbol{\beta})|s\right\} \\
&= \mathbb{E}_m\left\{\sum_{i\in s}(w_i-1)^2(Y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2 + \sum_{i\in U-s}(Y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2|s\right\} \\
&= \sum_{i\in s}(w_i-1)^2\mathbb{E}_m\left\{(Y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2|s\right\} + \sum_{i\in U-s}\mathbb{E}_m\left\{(Y_i-\mathbf{x}_i^\top\boldsymbol{\beta})^2|s\right\} \\
&= \sum_{i\in s}(w_i-1)^2\mathbb{E}_m\left(\epsilon_i^2|\mathbf{x}_i\right) + \sum_{i\in U-s}\mathbb{E}_m\left(\epsilon_i^2|\mathbf{x}_i\right) \\
&= \sum_{i\in s}(w_i-1)^2\left\{\mathbb{V}_m(\epsilon_i|\mathbf{x}_i) + (\mathbb{E}_m(\epsilon_i|\mathbf{x}_i))^2\right\}
\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{i\in U-s} \left\{\mathbb{V}_m(\epsilon_i|\mathbf{x}_i) + (\mathbb{E}_m(\epsilon_i|\mathbf{x}_i))^2\right\} \\
&= \sum_{i\in s}(w_i-1)^2\sigma^2 c_i + \sum_{i\in U-s}\sigma^2 c_i \\
&= \sigma^2\left(\sum_{i\in s}(w_i-1)^2 c_i + \sum_{i\in U-s} c_i\right),
\end{aligned}
$$

where $\epsilon_i = Y_i - \mathbf{x}_i^\top\boldsymbol{\beta}$ with $\mathbb{V}_m(\epsilon_i|\mathbf{x}_i) = \sigma^2 c_i$ and $\mathbb{E}_m(\epsilon_i|\mathbf{x}_i) = 0$. From (1.5.10), the result follows.

∎

The conditional bias $B_i(y_i;\boldsymbol{\beta})$ in (2.2.11) is unknown since $\boldsymbol{\beta}$ is unknown. Therefore, we must estimate it. Note that it is not possible to estimate the conditional bias of a nonsampled unit as its y-value is not observed. An estimator of (2.2.11) for a sample unit i is given by

$$
\hat{B}_i(y_i;\widetilde{\boldsymbol{\beta}}) = (w_i-1)(y_i - \mathbf{x}_i^\top\widetilde{\boldsymbol{\beta}}) \quad i\in s, \tag{2.2.13}
$$

where $\widetilde{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}$.

Proposition 2.2.8. *If $\widetilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{WLS}$, then $\hat{B}_i(y_i;\widetilde{\boldsymbol{\beta}})$ is unbiased for $B_i(y_i;\boldsymbol{\beta})$ in the sense that*

$$
\mathbb{E}_m(\hat{B}_i(y_i;\widetilde{\boldsymbol{\beta}})|s) = B_i(y_i;\boldsymbol{\beta})
$$

Proof:

$$
\begin{aligned}
\mathbb{E}_m(\hat{B}_i(y_i;\widetilde{\boldsymbol{\beta}})|s) &= \mathbb{E}_m\left\{(w_i-1)(y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}}_{WLS}|s\right\} \\
&= (w_i-1)\left(y_i - \mathbf{x}_i^\top\mathbb{E}_m\left\{\hat{\boldsymbol{\beta}}_{WLS}|s\right\}\right) \\
&= (w_i-1)(y_i - \mathbf{x}_i^\top\boldsymbol{\beta}) \\
&= B_i(y_i;\boldsymbol{\beta}).
\end{aligned}
$$

∎

Proposition 2.2.9. *If $\widetilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{WLS}^{(-i)}$, then $\hat{B}_i(y_i;\widetilde{\boldsymbol{\beta}})$ is conditionally unbiased for $B_i(y_i;\boldsymbol{\beta})$ in the sense that*

$$
\mathbb{E}_m(\hat{B}_i(y_i;\widetilde{\boldsymbol{\beta}})|s, Y_i = y_i) = B_i(y_i;\boldsymbol{\beta}),
$$

where $\hat{\boldsymbol{\beta}}_{WLS}^{(-i)}$ is the least squares estimator calculated without unit i.

Proof: Without loss of generality, we assume $c_j = 1$ for all $j \in U$.

$$\hat{\boldsymbol{\beta}}_{WLS}^{(-i)} = \left(\sum_{\substack{j\in s\\ j\neq i}} \mathbf{x}_j\mathbf{x}_j^\top\right)^{-1}\left(\sum_{\substack{j\in s\\ j\neq i}} \mathbf{x}_j Y_j\right)$$
$$= \left(\sum_{j\in s} \mathbf{x}_j\mathbf{x}_j^\top - \mathbf{x}_i\mathbf{x}_i^\top\right)^{-1}\left(\sum_{j\in s} \mathbf{x}_j Y_j - \mathbf{x}_i y_i\right).$$

By Sherman-Morrison formula (1949), we have

$$\left(\sum_{j\in s} \mathbf{x}_j\mathbf{x}_j^\top - \mathbf{x}_i\mathbf{x}_i^\top\right)^{-1} = \left(\sum_{j\in s} \mathbf{x}_j\mathbf{x}_j^\top\right)^{-1} + \frac{\left(\sum_{j\in s}\mathbf{x}_j\mathbf{x}_j^\top\right)^{-1}\mathbf{x}_i\mathbf{x}_i^\top\left(\sum_{j\in s}\mathbf{x}_j\mathbf{x}_j^\top\right)^{-1}}{1-\mathbf{x}_i^\top\left(\sum_{i\in s}\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1}\mathbf{x}_i}$$
$$= \left(\sum_{j\in s} \mathbf{x}_j\mathbf{x}_j^\top\right)^{-1} + \frac{\left(\sum_{j\in s}\mathbf{x}_j\mathbf{x}_j^\top\right)^{-1}\mathbf{x}_i\mathbf{x}_i^\top\left(\sum_{j\in s}\mathbf{x}_j\mathbf{x}_j^\top\right)^{-1}}{1-h_{ii}}, \tag{2.2.14}$$

where $h_{ii} = \mathbf{x}_i^\top\left(\sum_{i\in s}\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1}\mathbf{x}_i$. Using (2.2.14), we have

$$\hat{\boldsymbol{\beta}}_{WLS}^{(-i)} = \left(\left(\sum_{j\in s} \mathbf{x}_j\mathbf{x}_j^\top\right)^{-1} + \frac{\left(\sum_{j\in s}\mathbf{x}_j\mathbf{x}_j^\top\right)^{-1}\mathbf{x}_i\mathbf{x}_i^\top\left(\sum_{j\in s}\mathbf{x}_j\mathbf{x}_j^\top\right)^{-1}}{1-h_{ii}}\right)\left(\sum_{j\in s}\mathbf{x}_j Y_j - \mathbf{x}_i y_i\right)$$
$$= \hat{\boldsymbol{\beta}}_{WLS} - \left(\sum_{j\in s} \mathbf{x}_j\mathbf{x}_j^\top\right)^{-1}\mathbf{x}_i y_i + \frac{\left(\sum_{j\in s}\mathbf{x}_j\mathbf{x}_j^\top\right)^{-1}\mathbf{x}_i\mathbf{x}_i^\top\hat{\boldsymbol{\beta}}_{WLS}}{1-h_{ii}} - \frac{\left(\sum_{j\in s}\mathbf{x}_j\mathbf{x}_j^\top\right)^{-1}\mathbf{x}_i h_{ii} y_i}{1-h_{ii}}$$
$$= \hat{\boldsymbol{\beta}}_{WLS} + \frac{\left(\sum_{j\in s}\mathbf{x}_j\mathbf{x}_j^\top\right)^{-1}\mathbf{x}_i\mathbf{x}_i^\top\hat{\boldsymbol{\beta}}_{WLS}}{1-h_{ii}} - \frac{\left(\sum_{j\in s}\mathbf{x}_j\mathbf{x}_j^\top\right)^{-1}\mathbf{x}_i y_i}{1-h_{ii}}$$
$$= \hat{\boldsymbol{\beta}}_{WLS} - \frac{\left(\sum_{j\in s}\mathbf{x}_j\mathbf{x}_j^\top\right)^{-1}\mathbf{x}_i(y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}}_{WLS})}{1-h_{ii}}.$$

Hence,

$$y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}}_{WLS}^{(-i)} = \frac{y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}}_{WLS}}{1-h_{ii}}.$$

When $\widetilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{WLS}^{(-i)}$, we have

$$\hat{B}_i(y_i;\widetilde{\boldsymbol{\beta}}) = (w_i - 1)\frac{y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}}_{WLS}}{1-h_{ii}}$$

Then,

$$
\begin{aligned}
\mathbb{E}_m\left(\hat{B}_i(y_i;\widetilde{\boldsymbol{\beta}})|s,Y_i=y_i\right) &= (w_i-1)\frac{y_i-\mathbf{x}_i^\top\hat{\boldsymbol{\beta}}_{WLS}}{1-h_{ii}} \\
&= \frac{w_i-1}{1-h_{ii}}\left(y_i-\mathbb{E}_m\left\{\mathbf{x}_i^\top\hat{\boldsymbol{\beta}}_{WLS}|s,Y_i=y_i\right\}\right) \\
&= \frac{w_i-1}{1-h_{ii}}\left(y_i-\mathbb{E}_m\left\{\mathbf{x}_i^\top\left(\sum_{j\in s}\mathbf{x}_j\mathbf{x}_j^\top\right)^{-1}\sum_{j\in s}\mathbf{x}_jY_j|s,Y_i=y_i\right\}\right) \\
&= \frac{w_i-1}{1-h_{ii}}\left(y_i-\mathbf{x}_i^\top\left(\sum_{j\in s}\mathbf{x}_j\mathbf{x}_j^\top\right)^{-1}\mathbf{x}_iy_i\right. \\
&\left.-\mathbb{E}_m\left\{\mathbf{x}_i^\top\left(\sum_{j\in s}\mathbf{x}_j\mathbf{x}_j^\top\right)^{-1}\sum_{\substack{j\in s\\ j\neq i}}\mathbf{x}_jY_j|s,Y_i=y_i\right\}\right) \\
&= \frac{w_i-1}{1-h_{ii}}\left(y_i-h_{ii}y_i-\mathbf{x}_i^\top\left(\sum_{j\in s}\mathbf{x}_j\mathbf{x}_j^\top\right)^{-1}\sum_{\substack{j\in s\\ j\neq i}}\mathbf{x}_j\mathbf{x}_j^\top\boldsymbol{\beta}\right) \\
&= \frac{w_i-1}{1-h_{ii}}\left(y_i-h_{ii}y_i-\mathbf{x}_i^\top\left(\sum_{j\in s}\mathbf{x}_j\mathbf{x}_j^\top\right)^{-1}\sum_{j\in s}\mathbf{x}_j\mathbf{x}_j^\top\boldsymbol{\beta}\right. \\
&\left.+\mathbf{x}_i^\top\left(\sum_{j\in s}\mathbf{x}_j\mathbf{x}_j^\top\right)^{-1}\mathbf{x}_i\mathbf{x}_i^\top\boldsymbol{\beta}\right) \\
&= \frac{w_i-1}{1-h_{ii}}\left(y_i-h_{ii}y_i-\mathbf{x}_i^\top\boldsymbol{\beta}+h_{ii}\mathbf{x}_i^\top\boldsymbol{\beta}\right) \\
&= \frac{w_i-1}{1-h_{ii}}\left(y_i-\mathbf{x}_i^\top\boldsymbol{\beta}\right)(1-h_{ii}) \\
&= (w_i-1)\left(y_i-\mathbf{x}_i^\top\boldsymbol{\beta}\right) \\
&= B_i(y_i;\boldsymbol{\beta}).
\end{aligned}
$$

∎

For the simple linear regression model, the conditional bias of unit i is given by

$$B_i(y_i;\beta_0,\beta_1)=\left(\frac{N}{n}\left\{1+n\frac{(x_i-\bar{x})(\bar{X}-\bar{x})}{\sum_{i\in s}(x_i-\bar{x})^2}\right\}-1\right)(y_i-\beta_0-\beta_1x_i),\quad i\in s,\ (2.2.15)$$

An estimator of (2.2.15) is given by

$$\hat{B}_i(y_i; \hat{\beta}_0, \hat{\beta}_1) = \left(\frac{N}{n} \left\{ 1 + n \frac{(x_i - \bar{x})(\bar{X} - \bar{x})}{\sum_{i \in s}(x_i - \bar{x})^2} \right\} - 1 \right) (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i), \quad i \in s,$$

where $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. A sample unit is influential if its residual $y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ is large and/or its x-value is far from the sample mean $\bar{x}$, which may indicate a high leverage point.

Chapter 3

Robust estimation and prediction

In this chapter, we describe some robust estimators and predictors of a population total in the presence of influential units. The goal is to construct a robust estimator/predictor with smaller mean square error than that of the classical (non-robust) estimator/predictor. This is achieved at the expense of introducing a bias.

3.1 Model-based approach

In the model-based approach, it would be tempting to replace the weighted least squares estimator, $\hat{\boldsymbol{\beta}}_{WLS}$ in (1.5.9) by a robust estimator (e.g. an M-estimator) to predict t_y. As we argue below, this approach should generally be avoided. A naive predictor of t_y is defined as

$$\hat{t}_{y,RPRED}(c) = \sum_{i \in s} Y_i + \sum_{i \in U-s} \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_R(c), \tag{3.1.1}$$

where $\hat{\boldsymbol{\beta}}_R(c)$ is a robust estimator of $\boldsymbol{\beta}$ and c is a tuning constant. For instance, we can use an M-estimator of $\boldsymbol{\beta}$, which is defined as the solution of

$$\sum_{i \in s} \psi \left(\frac{Y_i - \mathbf{x}_i^\top \boldsymbol{\beta}}{\hat{\sigma}_i}; c \right) \mathbf{x}_i = 0,$$

where $\psi(\cdot; c)$ is a function whose role is to reduce the influence of units with a large residual, and $\hat{\sigma}_i$ is a robust estimator of σ_i. The function $\psi(\cdot; c)$ is the first derivative of a function $\rho(\cdot; c)$, called the objective function, whose properties include non-negativity, monotonicity, symmetry and passing through the origin; see, e.g., Maronna, Martin, and Yohai (2019). A common objective function is the so-called Huber function defined as

$$\rho(t; c) = \begin{cases} \frac{1}{2} t^2 & (|t| \le c) \\ c|t| - \frac{1}{2} c^2 & (|t| > c) \end{cases}. \tag{3.1.2}$$

Therefore, the first derivative of (3.1.2) is given by

$$\psi(t;c) = \begin{cases} c & (t > c) \\ t & (|t| \leq c) \\ -c & (t < -c) \end{cases}$$

Figure 3.1 shows the Huber function and the associated $\psi(t;c)$ function with $c = 1.345$.

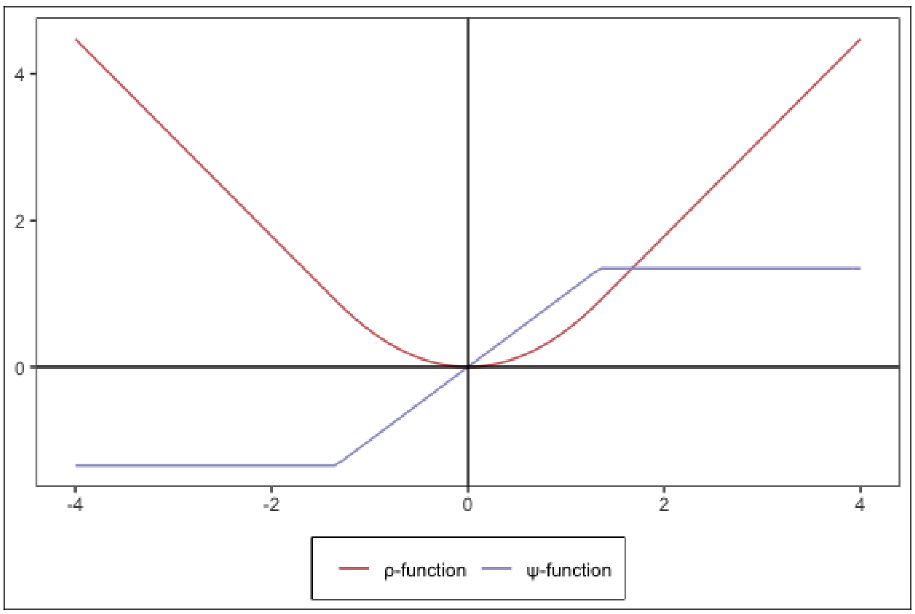

Figure 3.1: Huber function with $c = 1.345$

The naive predictor is expected to perform very well in terms of mean squared error if the sample only contains influential units which are believed to be unique in the population, or if the distribution of the errors is symmetric even if it has a heavy tail (e.g., a t-distribution). However, if the influential units are representative, i.e., there are other similar units in the non-sampled part of the population, or if the distribution of the errors is skewed, the naive predictor may be substantially biased.

As an illustration, we performed the following experiment: We generated a population of size $N = 100$ consisting of a single auxiliary variable x and a survey variable y. We first generated 90 observations, whose x-values were generated from a Gamma distribution with mean equal to 50 and variance equal to 500. The y-values were then generated according to the simple linear regression model, $Y_i = 100 + 2x_i + \epsilon_i$, where the random errors ϵ_i were generated from a normal distribution with mean equal to 0 and variance equal to 25. Then, we generated 10 influential units whose x-values were generated continuously and uniformly between 0 and 25, whereas the y-values were generated from a uniform distribution with mean equal to 379.8 and variance equal to 31.6. We selected a sample of size $n = 25$ from the population according to simple random sampling without replacement. We estimated β_0 and β_1 using both the least squares method and a robust method (M-estimation based on the Huber function with $c = 1.345$).

In Figure 3.2, the green line corresponds to the customary least squares fit. It is clear that the line is highly impacted by the influential units (represented by the red crosses). Predictions based on the least squares line were poor for most of the observations. The red line corresponds to the fit obtained by a robust method (M-estimation using the Huber function with $c = 1.345$). Here, the predictions based on the robust line were good for most observations but for the influential units in the top-left corner, the predictions were too small. If these influential units are representative, we expect the naive predictor $\hat{t}_{y,RPRED}(c)$ to be biased negatively because the predictions will be too small.

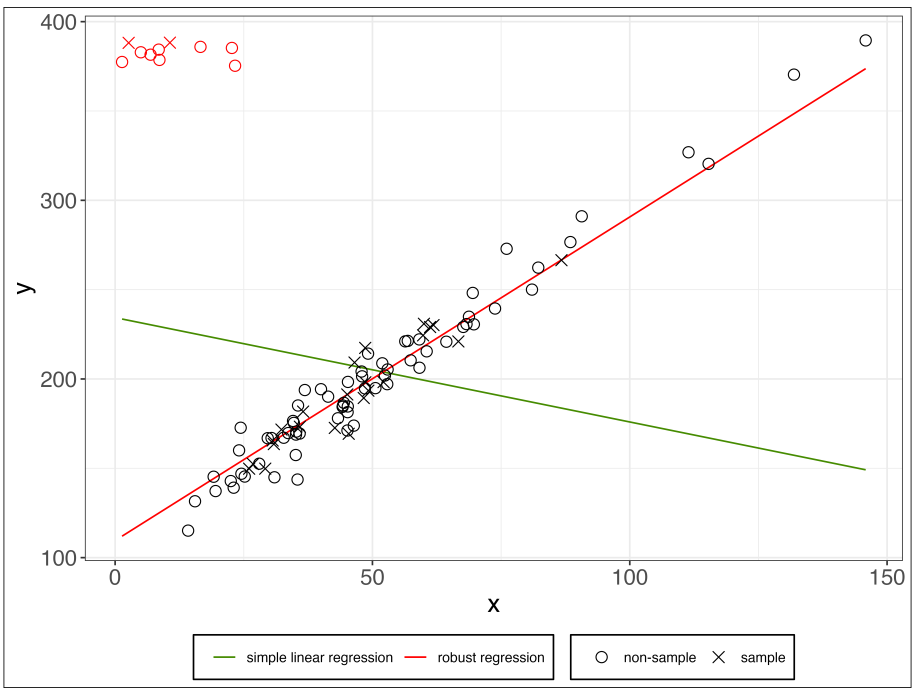

Figure 3.2: Example of simple linear regression (green) and robust regression (red) in presence of influential units

3.1.1 The predictor of Chambers

To cope with the potential bias of the naive predictor, Chambers (1986) proposed the following robust predictor:

$$\hat{t}_{y,CHAM}(k,c) = \hat{t}_{y,RPRED}(k) + \sum_{i \in s}(w_i - 1)\hat{\sigma}_i \psi_2 \left(\frac{\left(Y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_R(k)\right)}{\hat{\sigma}_i}; c \right), \quad (3.1.3)$$

where $\psi_2(\cdot; c)$ is a ψ-function with tuning constant c, and $\hat{\boldsymbol{\beta}}_R(k)$ denotes any robust estimator of $\boldsymbol{\beta}$ based on a ψ-function, denoted by $\psi_1(\cdot; k)$, with tuning constant k. The first term on the right hand-side of (3.1.3) is the naive robust predictor based on $\psi_1(\cdot; k)$ with tuning constant k, whereas the second term can be viewed as a bias correction term. When the tuning constant c is equal to 0, the predictor $\hat{t}_{y,CHAM}(k, 0)$ reduces to $\hat{t}_{y,RPRED}(k)$, which is stable but biased; When both $c = \infty$ and $k = \infty$, the predictor $\hat{t}_{y,CHAM}(\infty, \infty)$ reduces to $\hat{t}_{y,BLUP}$, which is unbiased but unstable. To achieve a trade-off between bias and variance, Chambers (1986) suggested to use $k = 1.345$, whereas c should be large enough, e.g., $4 \leq c \leq 6$.

3.1.2 Approach based on the conditional bias

Beaumont, Haziza and Ruiz-Gazen (2013) used the concept of the conditional bias to construct a robust predictor of t_y. It is defined as

$$\hat{t}_{y,CB}(c) = \hat{t}_{y,BLUP} - \sum_{i \in s} \hat{B}_i(y_i; \tilde{\boldsymbol{\beta}}) + \sum_{i \in s} \psi\left(\hat{B}_i(y_i; \tilde{\boldsymbol{\beta}}); c\right), \quad (3.1.4)$$

where $\hat{B}_i(y_i; \tilde{\boldsymbol{\beta}})$ is an estimator of the conditional bias defined in (2.2.13) and $\psi(\cdot; c)$ is usually the Huber function with tuning constant c. It follows from (2.2.13) that an alternative expression of (3.1.4) is given by

$$\hat{t}_{y,CB}(c) = \left(\sum_{i \in s} Y_i + \sum_{i \in U-s} \mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}} \right) + \sum_{i \in s} \psi \left\{ (w_i - 1)(y_i - \mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}}); c \right\}. \quad (3.1.5)$$

When the tuning constant is set to $c = 0$, $\hat{t}_{y,CB}(c)$ reduces to

$$\hat{t}_{y,CB}(0) = \sum_{i \in s} Y_i + \sum_{i \in U-s} \mathbf{x}_i^\top \widetilde{\boldsymbol{\beta}},$$

which depends on the choice of $\widetilde{\boldsymbol{\beta}}$. If $\widetilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_R(k)$, then

$$\hat{t}_{y,CB}(0) = \hat{t}_{y,RPRED}(k),$$

which is stable but biased. When $c = \infty$, if $\widetilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{WLS}$, then

$$\hat{t}_{y,CB}(\infty) = \hat{t}_{y,BLUP},$$

which is unbiased but unstable.

To derive the robust predictor of Beaumont et al. (2013), we need to evaluate the conditional bias attached to unit i with respect to $\hat{t}_{y,CB}(c)$. It is given by

$$B_i^{CB} = \mathbb{E}_m\left(\hat{t}_{y,CB}(c) - t_y | s, Y_i = y_i\right). \tag{3.1.6}$$

We rewrite (3.1.4) as

$$\hat{t}_{y,CB}(c) = \hat{t}_{y,BLUP} - \Delta(c), \tag{3.1.7}$$

where

$$\Delta(c) = \sum_{i \in s} \hat{B}_i(y_i; \tilde{\boldsymbol{\beta}}) - \sum_{i \in s} \psi\left(\hat{B}_i(y_i; \tilde{\boldsymbol{\beta}}); c\right).$$

Proposition 3.1.1. *The conditional bias* (3.1.6) *can be expressed as*

$$B_i^{CB} = B_i(y_i; \boldsymbol{\beta}) - \mathbb{E}_m\left(\Delta(c) | s, Y_i = y_i\right),$$

where $B_i(y_i; \boldsymbol{\beta})$ is defined in (2.2.11).

Proof:

$$\begin{aligned} B_i^{CB} &= \mathbb{E}_m\left(\hat{t}_{y,CB} - t_y | s, Y_i = y_i\right) \\ &= \mathbb{E}_m\left(\hat{t}_{y,BLUP} - \Delta(c) - t_y | s, Y_i = y_i\right) \\ &= \mathbb{E}_m\left(\hat{t}_{y,BLUP} - t_y | s, Y_i = y_i\right) - \mathbb{E}_m\left(\Delta(c) | s, Y_i = y_i\right) \\ &= B_i(y_i; \boldsymbol{\beta}) - \mathbb{E}_m\left(\Delta(c) | s, Y_i = y_i\right). \end{aligned}$$

∎

Therefore, a conditionally unbiased estimator of $B_i^{CB}(c)$ is given by

$$\hat{B}_i^{CB} = \hat{B}_i(y_i; \widetilde{\boldsymbol{\beta}}) - \Delta(c).$$

To choose the tuning constant c, Beaumont et al. (2013) suggest to determine the value of $\Delta(c)$ that minimizes

$$\max_{i \in s}\left\{|\hat{B}_i^{CB}|\right\}$$

or, equivalently, that minimizes

$$\max_{i \in s}\left\{|\hat{B}_i(y_i; \widetilde{\boldsymbol{\beta}}) - \Delta(c)|\right\}.$$

The resulting value of $\Delta(c)$ is

$$\Delta(c_{opt}) = \frac{1}{2}\left(\hat{B}_{min} + \hat{B}_{max}\right),$$

where

$$\hat{B}_{min} = min\left(\hat{B}_i(y_i;\widetilde{\boldsymbol{\beta}}); i \in s\right)$$

and

$$\hat{B}_{max} = max\left(\hat{B}_i(y_i;\widetilde{\boldsymbol{\beta}}); i \in s\right).$$

This leads to the robust predictor based on the conditional bias:

$$\hat{t}_{y,CB}(c_{opt}) = \hat{t}_{y,BLUP} - \frac{1}{2}\left(\hat{B}_{min} + \hat{B}_{max}\right). \tag{3.1.8}$$

Note that $\hat{t}_{y,CB}(c_{opt})$ can be obtained without actually computing the value c_{opt}. The value of c_{opt} is adaptive in the sense that the tuning constant c is a function of the sample size n so that $\underset{n\to\infty}{lim}\, c_n = \infty$, which is a desirable property. Indeed, when the sample size is large enough, the variance of the predictor is already very small, so there is no need to robustify the predictor.

3.2 Design-based approach

3.2.1 Robust Horvitz-Thompson estimator

Beaumont et al. (2013) proposed a robust alternative to the Horvitz-Thompson estimator based on the concept of conditional bias. It is defined as

$$\hat{t}_{y,RHT}(c) = \hat{t}_{y,HT} - \sum_{i\in S}\hat{B}_{1i}^{HT} + \sum_{i\in S}\psi(\hat{B}_{1i}^{HT};c), \tag{3.2.1}$$

where $\hat{B}_{1i}^{HT}$ is defined in (2.2.4). The conditional bias attached to unit i with respect to $\hat{t}_{y,RHT}$ is given by

$$B_{1i}^{RHT}(c) = \mathbb{E}_p\left(\hat{t}_{y,RHT}(c)|I_i = 1\right) - t_y. \tag{3.2.2}$$

We rewrite (3.2.1) as

$$\hat{t}_{y,RHT}(c) = \hat{t}_{y,HT} - \Delta(c),$$

where

$$\Delta(c) = \sum_{i\in S}\hat{B}_{1i}^{HT} - \sum_{i\in S}\psi(\hat{B}_{1i}^{HT};c).$$

Proposition 3.2.1. *The conditional bias* (3.2.2) *can be expressed as*

$$B_{1i}^{RHT}(c) = B_{1i}^{HT} - \mathbb{E}_p\left(\Delta(c)|I_i = 1\right).$$

Proof:

$$
\begin{aligned}
B_{1i}^{RHT}(c) &= \mathbb{E}_p\left(\hat{t}_{y,RHT}(c)|I_i = 1\right) - t_y \\
&= \mathbb{E}_p\left(\hat{t}_{y,HT} - \Delta(c)|I_i = 1\right) - t_y \\
&= \mathbb{E}_p\left(\hat{t}_{y,HT}|I_i = 1\right) - t_y - \mathbb{E}_p\left(\Delta(c)|I_i = 1\right) \\
&= B_{1i}^{HT} - \mathbb{E}_p\left(\Delta(c)|I_i = 1\right).
\end{aligned}
$$

∎

Therefore, a conditionally unbiased estimator of $B_{1i}^{RHT}(c)$ is given by

$$\hat{B}_{1i}^{RHT}(c) = \hat{B}_{1i}^{HT} - \Delta(c).$$

To choose the tuning constant c for $\hat{t}_{y,RHT}$, Beaumont et al. (2013) suggest to determine the value of $\Delta(c)$ that minimizes

$$\underset{i \in S}{max}\left\{|\hat{B}_{1i}^{RHT}(c)|\right\} = \underset{i \in S}{max}\left\{|\hat{B}_{1i}^{HT} - \Delta(c)|\right\}.$$

The solution is

$$\Delta(c_{opt}) = \frac{1}{2}\left(\hat{B}_{min}^{HT} + \hat{B}_{max}^{HT}\right),$$

where $\hat{B}_{min}^{HT} = min\left(\hat{B}_{1i}^{HT}; i \in S\right)$ and $\hat{B}_{max}^{HT} = max\left(\hat{B}_{1i}^{HT}; i \in S\right)$. This leads to the robust Horvitz-Thompson estimator based on conditional bias:

$$\hat{t}_{y,RHT}(c_{opt}) = \hat{t}_{y,HT} - \frac{1}{2}\left(\hat{B}_{min}^{HT} + \hat{B}_{max}^{HT}\right).$$

Note that $\hat{t}_{y,RHT}(c_{opt})$ can be obtained without actually computing the value c_{opt}.

3.3 Model-assisted approach

3.3.1 Robust version of the GREG estimator

Similar derivations as those presented in Section 3.2.1, lead to a robust version of the GREG estimator based on the concept of conditional bias:

$$\hat{t}_{y,RG}(c) = \hat{t}_{y,GREG} - \Delta(c),$$

where

$$\Delta(c) = \sum_{i \in S} \hat{B}_{1i}^{GREG} - \sum_{i \in S} \psi(\hat{B}_{1i}^{GREG}; c),$$

where $\hat{B}_{1i}^{GREG}$ is defined in (2.2.7). As before, the choice of the tuning constant can be obtained by minimizing the maximum estimated conditional bias of the robust GREG estimator. The resulting robust estimator is

$$\hat{t}_{y,RG} = \hat{t}_{y,GREG} - \frac{1}{2}\left(\hat{B}_{min}^{GREG} + \hat{B}_{max}^{GREG}\right), \tag{3.3.1}$$

where $\hat{B}_{min}^{GREG} = min\left(\hat{B}_{1i}^{GREG}; i \in S\right)$ and $\hat{B}_{max}^{GREG} = max\left(\hat{B}_{1i}^{GREG}; i \in S\right)$.

3.4 A robust predictor/estimator based on an adaptive tuning constant

For robust prediction/estimation, the choice of tuning constant c plays an important role in the trade-off between bias and variance. In the context of Huber method, Huber (1981) suggested a range of 1 to 2 for the tuning constant c. However, when dealing with a normal distribution, the optimal choice for c is ∞. In the presence of influential units, generally, a small c is more suitable for symmetric error distributions, while a larger c is preferable for skewed error distributions. We propose an adaptive tuning constant that can adjust the level of robustness automatically to the specific characteristics of the data in robust estimation and prediction. By using an adaptive tuning constant, we can avoid the need for manual selection of a fixed value, which may not be appropriate for all data sets.

In the model-based approach, the adaptive tuning constant, $\hat{c}$, is defined as

$$\hat{c} = 1.345\left\{1 + \frac{|\hat{B}_{min}^{*} + \hat{B}_{max}^{*}|}{2}\right\} + \frac{n}{N}\alpha, \tag{3.4.1}$$

where α is a specified constant and $\hat{B}_i^{*}(y_i; \tilde{\boldsymbol{\beta}})$ is the standardized version of $\hat{B}_i(y_i; \tilde{\boldsymbol{\beta}})$. More specifically, it is given by

$$\hat{B}_i^{*}(y_i; \tilde{\boldsymbol{\beta}}) = \frac{\hat{B}_i(y_i; \tilde{\boldsymbol{\beta}}) - \bar{\hat{B}}}{\tilde{s}_{\hat{B}}},$$

where $\bar{\hat{B}} = \frac{1}{n}\sum_{i\in s}\hat{B}_i(y_i; \tilde{\boldsymbol{\beta}})$ and $\tilde{s}_{\hat{B}} = \frac{1}{n-1}\sum_{i\in s}(\hat{B}_i(y_i; \tilde{\boldsymbol{\beta}}) - \bar{\hat{B}})^2$. The rationale behind this choice of $\hat{c}$ can be explained as follow: If the sample fraction n/N is small (which is often the case in practice), we can ignore the second term on the right hand-side of (3.4.1). In this case, (i) If the distribution of the conditional bias is symmetric (which would occur if the distribution of the errors is symmetric and the first-order inclusion probabilities are equal), we can expect $|\hat{B}_{min}^{*} + \hat{B}_{max}^{*}|/2$, the average between $\hat{B}_{min}^{*}$ and $\hat{B}_{max}^{*}$, to be close to 0 and, as a result, $\hat{c} \approx 1.345$. (ii) If the distribution of the conditional bias is skewed (which will occur if the distribution of the errors is skewed), we expect to have $|\hat{B}^*_{min} + \hat{B}^*_{max}|/2 > 0$ and $\hat{c} > 1.345$. If the second term $\frac{n}{N}\alpha$ on the right hand-side of (3.4.1) is not negligible, we note that the value of $\hat{c}$ increases as n/N increases. This ensures that the robust predictor approaches the non-robust predictor as the sample size increases, which is a desirable feature. A new robust predictor of t_y based on an adaptive tuning constant in the projection form is

$$\hat{t}_{y,new}(\hat{c}) = \sum_{i \in U} \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_R(\hat{c}), \tag{3.4.2}$$

where $\hat{\boldsymbol{\beta}}_R(\hat{c})$ is obtained by an M-estimation procedure using Huber function with tuning constant equal to $\hat{c}$.

In the model-assisted approach, the new tuning constant is defined as

$$\hat{c} = 1.345 \left\{ 1 + \frac{|\hat{B}^{*GREG}_{min} + \hat{B}^{*GREG}_{max}|}{2} \right\} + \frac{n}{N}\alpha,$$

where $\hat{B}^{*GREG}_i$ is the standardized version of $\hat{B}^{GREG}_i$. Thus, a robust version of GREG estimator based on an adaptive tuning constant in the projection form is

$$\hat{t}^{new}_{y,GREG} = \sum_{i \in U} \mathbf{x}_i^\top \hat{\boldsymbol{B}}_R(\hat{c}), \tag{3.4.3}$$

where $\hat{\boldsymbol{B}}_R(\hat{c})$ is obtained by M-estimation using Huber function with tuning constant equal to $\hat{c}$.

Chapter 4

Simulation study

In this chapter, we present the results from several simulation studies: the first, in the model-based framework, compares several predictors presented in Chapter 3 in the presence of influential units, in terms of relative bias and relative efficiency. The second, in a model-assisted framework, assesses the performance of the GREG estimator and some robust versions in the case of stratified simple random sampling and Poisson sampling. The simulation study was conducted using the R Language. The `MASS` package was used for the call to `rlm()` function; The `VGAM` package was used for the call to `rpareto()` function and `rlaplace()` function; The `Rlab` package was used for the call to `rbern()` function.

4.1 Model-based approach

We generated ten populations of size $N = 5000$, each consisting of a single auxiliary variable x and a survey variable y. In each population, the x-values were first generated from a Gamma distribution with mean equal to 50 and variance equal to 500.

For the first seven populations, the y-values were generated according to the conditional distribution

$$Y_i|x_i \sim \mathcal{D}(\mu_i, \nu_i),$$

where $\mu_i = \beta_0 + \beta_1 x_i$ and $\nu_i = \sigma^2 x_i$. Here, we set $\beta_0 = 100$, $\beta_1 = 2$ and $\sigma = 30$. We used seven different distributions $\mathcal{D}$: normal, Gamma, lognormal, Pareto, Student, Laplace and Weibull. That is, for these seven distributions, the first two moments were identical.

For the last three populations, the y-values were generated according to the mixture model

$$Y_i = \tau_i \mathcal{N}(\mu_1, \nu_1) + (1 - \tau_i)\mathcal{N}(\mu_2, \nu_2),$$

where $\mu_1 = \beta_0 + \beta_1 x_i$, $\nu_1 = \sigma^2 x_i$, $\mu_2 = 10\mu_1$ and $\nu_2 = 10\nu_1$. Here, we set $\beta_0 = 100$, $\beta_1 = 2$ and $\sigma^2 = 2000$. The τ_i's were independently generated from a Bernoulli distribution with probability $\delta = 0.995$, 0.99 and 0.98, respectively. This corresponds to populations that contained approximately 0.5%, 1% and 2% of outliers. Examples of these ten populations can be found in Appendix A; see Figures A.1-A.10. We repeated $10,000$ iterations of the following process:

(a) A finite population of size $N = 5,000$ was generated.

(b) A sample of size n was selected according to simple random sampling without replacement. We used $n = 25; 50; 100$.

(c) In each sample, we computed the following predictors of t_y: (i) the BLUP, $\hat{t}_{y,BLUP}$, given by (1.5.8); (ii) the naive predictor, $\hat{t}_{y,RPRED}(k)$, given by (3.1.1), where $\hat{\beta}_R(k)$ was a Huber M-estimator with $k = 0.1, 0.8, 1.345$ and 2; (iii) the predictor of Chambers, $\hat{t}_{y,CHAM}(k, c)$, given by (3.1.3) with $k = 0.1, 0.8, 1.345, 2$, $c = 2, 4, 6, 8$ and $\hat{\sigma}_i = \hat{\sigma} = \text{med}(|y_i - \text{med}(y_i)|)$, the median of absolute deviations, $i \in s$. Recall that $\hat{t}_{y,CHAM}(k, 0) \equiv \hat{t}_{y,RPRED}(k)$; (iv) the robust predictor based on the concept of the conditional bias, $\hat{t}_{y,CB}(c_{opt})$, given by (3.1.8); (v) the new robust predictor, $\hat{t}_{y,new}(\hat{c})$, based on an adaptive tuning constant given by (3.4.2) with α =3, 6 and $\sqrt{n}$.

As a measure of bias of a predictor, we computed the Monte Carlo percent relative bias:

$$RB_{MC}(\hat{t}_y) = 100 \times \frac{\mathbb{E}_{MC}(\hat{t}_y - t_y)}{\mathbb{E}_{MC}(\hat{t}_y)},$$

where $\mathbb{E}_{MC}(\hat{t}_y) = R^{-1} \sum_{r=1}^{R} \hat{t}_y^{(r)}$ with $\hat{t}_y^{(r)}$ denoting the predictor $\hat{t}_y$ in the rth simulated sample, $r = 1, ..., R$. We also computed the Monte Carlo relative efficiency, using the BLUP, $\hat{t}_{y,BLUP}$, as the reference:

$$RE_{MC}(\hat{t}_y) = 100 \times \frac{MSE_{MC}(\hat{t}_y)}{MSE_{MC}(\hat{t}_{y,BLUP})},$$

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

Tables 4.1-4.10 show the Monte Carlo relative bias and relative efficiency for all the predictors. As expected, $\hat{t}_{y,BLUP}$ showed a negligible bias in all the scenarios. For the normal distribution (i.e., in the absence of influential units), both $\hat{t}_{y,CB}$ and $\hat{t}_{y,new}(\hat{c})$ showed negligible bias and a slight loss of efficiency (between 1% and 3%) with respect to $\hat{t}_{y,BLUP}$; see Table 4.1. This is a desirable property. On the other hand, the predictor $\hat{t}_{y,CHAM}(k, c)$ showed some appreciable loss of efficiency for small values of k and c, but had similar performances of those of $\hat{t}_{y,BLUP}$ for large values of c.

For the symmetric distributions (t-distribution and Laplace distribution), all the robust predictors showed negligible bias in all the scenarios; see Table 4.5 and 4.7. The naive robust predictor, $\hat{t}_{y,CHAM}(k, 0)$ was especially efficient. For instance, for the t-distribution (Table 4.5), the values of relative efficiency ranged between 51% and 59% for $n = 50$. The new predictor, $\hat{t}_{y,new}(\hat{c})$, was more efficient than $\hat{t}_{y,CB}$ with a value of relative efficiency of 62%. We note that the choice of α did not seem to affect the behavior of $\hat{t}_{y,new}(\hat{c})$.

For the skewed distribution (Gamma, Pareto, lognormal and Weibull), we start by noting that the naive predictor, $\hat{t}_{y,CHAM}(k, 0)$, and the predictor of Chambers did not behave well for some combinations of (k, c), especially for small value of k and c. For instance, for the lognormal distribution (Table 4.3) and $n = 50$, the Monte Carlo relative bias of $\hat{t}_{y,CHAM}(k, 0)$ ranged from -40.1% to -11.6% while that of $\hat{t}_{y,CHAM}(k, 2)$, ranged from -19.2% to -15.4%. The value of relative efficiency of $\hat{t}_{y,CHAM}(k, 0)$ ranged from 117% to 419%, and that of $\hat{t}_{y,CHAM}(k, 2)$, ranged from 141% to 174%. In other words, the naive predictor and the predictor of Chambers with small value of k and c were much worse than $\hat{t}_{y,BLUP}$ in many scenarios. Both the naive predictor and the predictor of Chambers showed a similar behavior for the other skewed distributions. The best performances in terms of relative efficiency of $\hat{t}_{y,CHAM}(k, c)$ were obtained for $c = 4$ and $c = 6$, which is consistent with the suggestion of Chambers (1986).

The predictor $\hat{t}_{y,CB}(c_{opt})$ was generally biased, as expected. The bias was moderate in most cases with a maximum absolute relative bias smaller than 5%. We note that the relative bias decreased as the sample size n increased. In terms of relative efficiency, it was better than $\hat{t}_{y,BLUP}$ in most scenarios. For instance for the Pareto distribution, with $n = 50$, the value of relative efficiency was equal to 64%; see Table 4.4. For the lognormal distribution with $n = 50$, the value of relative efficiency was equal to 90%; see Table 4.3. We note that the value of relative efficiency approached 100% as the sample size increased. This suggests that the predictor $\hat{t}_{y,CB}(c_{c_{opt}})$ is consistent for t_y.

The new predictor $\hat{t}_{y,new}(\hat{c})$ performed well in all the scenarios with values of relative efficiency never larger than 100%. In general, $\hat{t}_{y,new}(\hat{c})$ showed larger absolute relative bias than $\hat{t}_{y,CB}(c_{opt})$ but was generally more efficient than $\hat{t}_{y,CB}(c_{opt})$. For example, for the Pareto distribution with $n = 50$, the relative bias for $\hat{t}_{y,CB}(c_{opt})$ and $\hat{t}_{y,new}(\hat{c})$ (with $\alpha = \sqrt{n}$), was respectively equal to -3.4% and -5.5%, but the relative efficiency for $\hat{t}_{y,CB}(c_{opt})$ and $\hat{t}_{y,new}(\hat{c})$ was respectively equal to 64% and 50%. We note that the relative bias of $\hat{t}_{y,new}(\hat{c})$ decreased as the sample size n increased and that its relative efficiency increased as n increased. For a given value of n, the choice of α did not seem to affect the behavior of $\hat{t}_{y,new}(\hat{c})$.

Finally, we turn to the mixture distributions. Again, both $\hat{t}_{y,CB}(c_{opt})$ and $\hat{t}_{y,new}(\hat{c})$ were more efficient than $\hat{t}_{y,BLUP}$ with values of relative efficiency smaller than 100%. Again, $\hat{t}_{y,CB}(c_{opt})$ exhibited smaller values of absolute relative bias than $\hat{t}_{y,new}(\hat{c})$ but the latter was more efficient in all the scenarios. For instance, in the case of a 2% contamination with $n = 50$ (see Table 4.10), the relative bias of $\hat{t}_{y,CB}(c_{opt})$ was equal to -4.7%, whereas it was equal to -10.9% for $\hat{t}_{y,new}(\hat{c})$ (with $\alpha = \sqrt{n}$) but the relative efficiency of the latter was equal to 46% compared to 68% for $\hat{t}_{y,CB}(c_{opt})$.

Normal		$n = 25$					$n = 50$					$n = 100$				
$\hat{t}_{y,BLUP}$		-0.2 (100)					0.0 (100)					-0.1 (100)				
$\hat{t}_{y,CB}(c_{opt})$		-0.2 (101)					0.0 (100)					-0.1 (100)				
$\hat{t}_{y,CHAM}(k,c)$		Huber ψ, c					Huber ψ, c					Huber ψ, c				
		0	2	4	6	8	0	2	4	6	8	0	2	4	6	8
Huber ψ k	0.1	-0.3 (149)	-0.5 (104)	-0.3 (101)	-0.3 (101)	-0.2 (101)	0.1 (148)	-0.1 (104)	0.0 (101)	0.0 (101)	0.0 (101)	0.0 (146)	-0.2 (103)	-0.1 (100)	-0.1 (100)	-0.1 (100)
	0.8	-0.3 (117)	-0.4 (103)	-0.3 (101)	-0.3 (101)	-0.2 (101)	0.0 (116)	-0.1 (103)	0.0 (101)	0.0 (101)	0.0 (101)	-0.1 (115)	-0.2 (103)	-0.1 (100)	-0.1 (100)	-0.1 (100)
	1.345	-0.3 (106)	-0.4 (102)	-0.3 (101)	-0.2 (101)	-0.2 (101)	0.1 (106)	0.0 (103)	0.0 (101)	0.0 (101)	0.0 (101)	-0.1 (105)	-0.2 (102)	-0.1 (100)	-0.1 (100)	-0.1 (100)
	2	-0.2 (102)	-0.4 (102)	-0.3 (101)	-0.2 (101)	-0.2 (101)	0.0 (101)	0.0 (102)	0.0 (101)	0.0 (101)	0.0 (101)	-0.1 (101)	-0.2 (102)	-0.1 (100)	-0.1 (100)	-0.1 (100)
$\hat{t}_{y,new}(\hat{c})$ α	3	-0.3 (102)					0.1 (102)					-0.1 (102)				
	6	-0.3 (102)					0.1 (102)					-0.1 (102)				
	$\sqrt{n}$	-0.3 (102)					0.1 (102)					-0.1 (101)				

Table 4.1: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the BLUP and robust predictors for the normal distribution

Gamma		$n = 25$					$n = 50$					$n = 100$				
$\hat{t}_{y,BLUP}$		-0.1 (100)					0.0 (100)					0.0 (100)				
$\hat{t}_{y,CB}(c_{opt})$		-3.7 (97)					-2.6 (98)					-1.7 (100)				
$\hat{t}_{y,CHAM}(k,c)$		Huber ψ, c					Huber ψ, c					Huber ψ, c				
		0	2	4	6	8	0	2	4	6	8	0	2	4	6	8
Huber ψ k	0.1	-39.8 (273)	-16.3 (138)	-5.3 (104)	-2.1 (100)	-0.9 (99)	-43.5 (499)	-16.5 (182)	-4.9 (107)	-1.7 (100)	-0.6 (100)	-44.7 (951)	-16.3 (272)	-4.6 (116)	-1.6 (101)	-0.6 (100)
	0.8	-30.5 (200)	-15.0 (131)	-4.9 (103)	-1.9 (99)	-0.9 (100)	-32.7 (349)	-15.3 (171)	-4.5 (106)	-1.6 (100)	-0.6 (100)	-33.4 (643)	-15.2 (250)	-4.3 (114)	-1.5 (101)	-0.5 (100)
	1.345	-16.0 (131)	-13.0 (122)	-4.3 (102)	-1.7 (99)	-0.8 (100)	-16.7 (179)	-13.4 (154)	-4.0 (104)	-1.4 (100)	-0.5 (100)	-16.6 (272)	-13.3 (218)	-3.9 (111)	-1.3 (101)	-0.5 (100)
	2	-7.4 (106)	-11.5 (116)	-3.9 (101)	-1.5 (99)	-0.7 (100)	-7.6 (117)	-12.0 (143)	-3.6 (103)	-1.2 (98)	-0.5 (98)	-7.6 (140)	-12.2 (199)	-3.6 (109)	-1.2 (101)	-0.4 (100)
$\hat{t}_{y,new}(\hat{c})$ α	3	-4.0 (98)					-2.6 (99)					-1.6 (100)				
	6	-3.9 (98)					-2.5 (99)					-1.5 (100)				
	$\sqrt{n}$	-3.9 (98)					-2.5 (99)					-1.4 (100)				

Table 4.2: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the BLUP and robust predictors for the Gamma distribution

Lognormal		$n = 25$					$n = 50$					$n = 100$				
$\hat{t}_{y,BLUP}$		0.0					-0.1					-0.3				
		(100)					(100)					(100)				
$\hat{t}_{y,CB}(c_{opt})$		-4.2					-3.3					-2.4				
		(87)					(90)					(94)				
$\hat{t}_{y,CHAM}(k,c)$		Huber ψ c					Huber ψ c					Huber ψ c				
		0	2	4	6	8	0	2	4	6	8	0	2	4	6	8
Huber ψ k	0.1	-37.2	-18.5	-8.2	-4.4	-2.6	-40.1	-19.2	-8.3	-4.4	-2.6	-41.1	-19.3	-8.3	-4.4	-2.6
		(209)	(111)	(82)	(81)	(84)	(419)	(174)	(95)	(84)	(84)	(821)	(296)	(122)	(92)	(88)
	0.8	-30.1	-17.5	-7.9	-4.2	-2.5	-32.2	-18.2	-8.0	-4.2	-2.5	-32.9	-18.4	-8.0	-4.3	-2.5
		(163)	(107)	(82)	(81)	(84)	(313)	(165)	(94)	(84)	(84)	(604)	(278)	(119)	(92)	(88)
	1.345	-19.2	-15.9	-7.3	-3.9	-2.4	-20.2	-16.7	-7.5	-4.0	-2.4	-20.6	-16.9	-7.5	-4.0	-2.4
		(114)	(101)	(82)	(82)	(85)	(185)	(152)	(92)	(84)	(85)	(323)	(250)	(115)	(91)	(88)
	2	-11.0	-14.4	-6.8	-3.7	-2.2	-11.6	-15.4	-7.0	-3.8	-2.3	-11.9	-15.8	-7.1	-3.9	-2.4
		(90)	(96)	(81)	(82)	(85)	(117)	(141)	(91)	(84)	(85)	(171)	(229)	(111)	(91)	(88)
$\hat{t}_{y,new}(\hat{c})$ α	3	-6.4					-4.8					-3.5				
		(85)					(90)					(95)				
	6	-6.3					-4.7					-3.4				
		(85)					(90)					(94)				
	$\sqrt{n}$	-6.3					-4.7					-3.3				
		(85)					(90)					(94)				

Table 4.3: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the BLUP and robust predictors for the lognormal distribution

Pareto		$n = 25$					$n = 50$					$n = 100$				
$\hat{t}_{y,BLUP}$		0.0					-0.1					0.1				
		(100)					(100)					(100)				
$\hat{t}_{y,CB}(c_{opt})$		-3.8					-3.2					-2.3				
		(56)					(64)					(72)				
$\hat{t}_{y,CHAM}(k,c)$		Huber ψ c					Huber ψ c					Huber ψ c				
		0	2	4	6	8	0	2	4	6	8	0	2	4	6	8
Huber ψ k	0.1	-25.6	-14.3	-8.6	-6.0	-4.6	-26.8	-14.5	-8.6	-5.9	-4.5	-27.1	-14.4	-8.5	-5.8	-4.3
		(94)	(44)	(31)	(28)	(29)	(257)	(104)	(58)	(46)	(44)	(450)	(168)	(80)	(55)	(47)
	0.8	-22.2	-13.8	-8.4	-5.9	-4.5	-23.1	-14.0	-8.4	-5.8	-4.4	-23.2	-14.0	-8.3	-5.7	-4.3
		(77)	(43)	(30)	(28)	(29)	(205)	(100)	(57)	(46)	(44)	(353)	(160)	(78)	(55)	(47)
	1.345	-16.0	-13.0	-8.1	-5.7	-4.4	-16.5	-13.2	-8.1	-5.7	-4.3	-16.4	-13.2	-8.0	-5.5	-4.2
		(52)	(41)	(30)	(28)	(29)	(127)	(93)	(55)	(46)	(44)	(208)	(147)	(75)	(54)	(46)
	2	-11.3	-12.2	-7.7	-5.5	-4.3	-11.6	-12.5	-7.8	-5.5	-4.2	-11.5	-12.6	-7.7	-5.4	-4.1
		(39)	(39)	(30)	(28)	(29)	(83)	(87)	(54)	(45)	(44)	(125)	(138)	(72)	(53)	(46)
$\hat{t}_{y,new}(\hat{c})$ α	3	-7.1					-5.6					-4.4				
		(28)					(50)					(65)				
	6	-6.9					-5.5					-4.3				
		(28)					(50)					(65)				
	$\sqrt{n}$	-6.9					-5.5					-4.2				
		(28)					(50)					(64)				

Table 4.4: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the BLUP and robust predictors for the Pareto distribution

t-distribution		$n=25$					$n=50$					$n=100$				
$\hat{t}_{y,BLUP}$		0.0 (100)					0.0 (100)					0.0 (100)				
$\hat{t}_{y,CB}(c_{opt})$		0.0 (76)					0.0 (80)					0.0 (84)				
$\hat{t}_{y,CHAM}(k,c)$		Huber ψ c					Huber ψ c					Huber ψ c				
		0	2	4	6	8	0	2	4	6	8	0	2	4	6	8
Huber ψ k	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		(56)	(65)	(74)	(78)	(80)	(59)	(71)	(82)	(86)	(88)	(61)	(74)	(84)	(89)	(91)
	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		(48)	(65)	(74)	(78)	(80)	(51)	(71)	(82)	(86)	(88)	(52)	(74)	(85)	(89)	(91)
	1.345	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		(50)	(65)	(75)	(78)	(80)	(53)	(71)	(82)	(86)	(88)	(54)	(74)	(85)	(89)	(91)
	2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
		(55)	(66)	(75)	(78)	(80)	(58)	(72)	(82)	(86)	(88)	(59)	(74)	(85)	(89)	(91)
$\hat{t}_{y,new}(\hat{c})$ α	3	0.0 (57)					0.0 (62)					0.0 (65)				
	6	0.0 (58)					0.0 (62)					0.0 (66)				
	$\sqrt{n}$	0.0 (58)					0.0 (62)					0.0 (66)				

Table 4.5: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the BLUP and robust predictors for the t-distribution

Weibull		$n=25$					$n=50$					$n=100$				
$\hat{t}_{y,BLUP}$		0.0 (100)					0.2 (100)					0.0 (100)				
$\hat{t}_{y,CB}(c_{opt})$		-3.8 (96)					-2.4 (98)					-1.7 (99)				
$\hat{t}_{y,CHAM}(k,c)$		Huber ψ c					Huber ψ c					Huber ψ c				
		0	2	4	6	8	0	2	4	6	8	0	2	4	6	8
Huber ψ k	0.1	-40.7	-16.9	-5.6	-2.3	-1.0	-43.4	-16.4	-4.8	-1.5	-0.5	-45.2	-16.6	-4.8	-1.6	-0.6
		(270)	(135)	(102)	(98)	(99)	(484)	(178)	(107)	(100)	(99)	(954)	(275)	(117)	(101)	(100)
	0.8	-31.3	-15.6	-5.3	-2.1	-1.0	-32.7	-15.2	-4.5	-1.4	-0.4	-33.7	-15.4	-4.5	-1.5	-0.6
		(200)	(129)	(101)	(98)	(99)	(338)	(167)	(105)	(99)	(99)	(645)	(253)	(115)	(101)	(100)
	1.345	-16.7	-13.6	-4.7	-1.9	-0.9	-16.7	-13.4	-4.0	-1.3	-0.3	-16.9	-13.6	-4.0	-1.4	-0.5
		(131)	(120)	(100)	(98)	(99)	(176)	(151)	(104)	(99)	(100)	(276)	(221)	(112)	(101)	(100)
	2	-7.8	-12.0	-4.2	-1.7	-0.8	-7.6	-12.0	-3.6	-1.1	-0.3	-7.7	-12.4	-3.7	-1.3	-0.5
		(105)	(114)	(99)	(98)	(99)	(116)	(141)	(103)	(99)	(100)	(142)	(201)	(110)	(100)	(100)
$\hat{t}_{y,new}(\hat{c})$ α	3	-4.1 (97)					-2.5 (98)					-1.7 (100)				
	6	-4.0 (97)					-2.4 (98)					-1.6 (100)				
	$\sqrt{n}$	-4.0 (97)					-2.4 (98)					-1.4 (100)				

Table 4.6: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the BLUP and robust predictors for the Weibull distribution

Laplace		$n = 25$					$n = 50$					$n = 100$				
$\hat{t}_{y,BLUP}$		0.2 (100)					0.2 (100)					0.1 (100)				
$\hat{t}_{y,CB}(c_{opt})$		0.1 (93)					0.2 (95)					0.1 (97)				
$\hat{t}_{y,CHAM}(k,c)$		Huber ψ c					Huber ψ c					Huber ψ c				
		0	2	4	6	8	0	2	4	6	8	0	2	4	6	8
Huber ψ k	0.1	0.0	-0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.2	-0.1	-0.1	0.0	0.0	0.1
		(71)	(78)	(92)	(98)	(100)	(64)	(75)	(92)	(98)	(100)	(59)	(75)	(92)	(98)	(100)
	0.8	0.0	-0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.2	0.2	0.0	-0.1	0.0	0.0	0.1
		(69)	(79)	(93)	(98)	(100)	(66)	(78)	(93)	(98)	(100)	(66)	(78)	(93)	(98)	(100)
	1.345	0.0	0.0	0.1	0.1	0.1	0.2	0.1	0.1	0.2	0.2	0.0	0.0	0.0	0.0	0.1
		(76)	(81)	(94)	(98)	(100)	(74)	(80)	(94)	(98)	(100)	(74)	(80)	(93)	(98)	(100)
	2	0.1	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.0	0.0	0.0	0.0	0.1
		(84)	(83)	(94)	(99)	(100)	(82)	(81)	(94)	(99)	(100)	(82)	(81)	(94)	(98)	(100)
$\hat{t}_{y,new}(\hat{c})$ α	3	0.2 (85)					0.1 (83)					0.0 (83)				
	6	0.1 (85)					0.1 (84)					0.0 (84)				
	$\sqrt{n}$	0.1 (85)					0.1 (84)					0.0 (84)				

Table 4.7: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the BLUP and robust predictors for the Laplace distribution

$\delta = 0.995$		$n = 25$					$n = 50$					$n = 100$				
$\hat{t}_{y,BLUP}$		0.3 (100)					0.1 (100)					0.1 (100)				
$\hat{t}_{y,CB}(c_{opt})$		-1.6 (78)					-1.6 (78)					-1.4 (79)				
$\hat{t}_{y,CHAM}(k,c)$		Huber ψ c					Huber ψ c					Huber ψ c				
		0	2	4	6	8	0	2	4	6	8	0	2	4	6	8
Huber ψ k	0.1	-3.9	-4.2	-3.3	-2.7	-2.2	-4.0	-3.9	-3.2	-2.6	-2.1	-3.8	-3.8	-3.1	-2.5	-2.0
		(88)	(62)	(63)	(67)	(72)	(89)	(65)	(66)	(69)	(73)	(94)	(68)	(67)	(70)	(73)
	0.8	-4.0	-4.1	-3.3	-2.7	-2.2	-4.0	-3.9	-3.2	-2.6	-2.1	-3.9	-3.8	-3.1	-2.5	-2.0
		(69)	(62)	(63)	(67)	(72)	(71)	(65)	(66)	(69)	(73)	(75)	(67)	(67)	(70)	(73)
	1.345	-4.1	-4.0	-3.3	-2.7	-2.2	-3.9	-3.8	-3.2	-2.6	-2.1	-3.9	-3.8	-3.1	-2.5	-2.0
		(63)	(62)	(64)	(67)	(72)	(66)	(65)	(66)	(69)	(73)	(69)	(67)	(67)	(70)	(73)
	2	-3.9	-4.0	-3.3	-2.7	-2.2	-3.8	-3.8	-3.2	-2.6	-2.1	-3.7	-3.8	-3.1	-2.5	-2.0
		(62)	(62)	(64)	(68)	(72)	(64)	(65)	(66)	(69)	(73)	(66)	(67)	(67)	(70)	(73)
$\hat{t}_{y,new}(\hat{c})$ α	3	-3.4 (63)					-2.7 (67)					-2.5 (69)				
	6	-3.4 (63)					-2.7 (67)					-2.5 (69)				
	$\sqrt{n}$	-3.4 (63)					-2.7 (67)					-2.5 (69)				

Table 4.8: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the BLUP and robust predictors for the mixture distribution with $\delta = 0.995$

$\delta = 0.99$		$n = 25$					$n = 50$					$n = 100$				
$\hat{t}_{y,BLUP}$		0.1 (100)					-0.1 (100)					-0.1 (100)				
$\hat{t}_{y,CB}(c_{opt})$		-2.6 (71)					-2.6 (71)					-2.5 (74)				
$\hat{t}_{y,CHAM}(k,c)$		Huber ψ c					Huber ψ c					Huber ψ c				
		0	2	4	6	8	0	2	4	6	8	0	2	4	6	8
Huber ψ k	0.1	-7.7	-7.3	-5.8	-4.6	-3.6	-8.2	-7.7	-6.2	-5.0	-4.0	-7.6	-7.2	-5.9	-4.8	-3.8
		(65)	(48)	(51)	(56)	(62)	(70)	(51)	(52)	(56)	(61)	(77)	(56)	(56)	(59)	(64)
	0.8	-7.6	-7.3	-5.8	-4.6	-3.6	-8.2	-7.6	-6.2	-5.0	-4.0	-7.6	-7.2	-5.9	-4.8	-3.8
		(52)	(48)	(51)	(56)	(62)	(56)	(51)	(52)	(56)	(61)	(62)	(56)	(56)	(59)	(64)
	1.345	-7.2	-7.2	-5.7	-4.5	-3.5	-7.9	-7.6	-6.2	-5.0	-4.0	-7.4	-7.1	-5.9	-4.8	-3.8
		(49)	(48)	(51)	(56)	(63)	(52)	(51)	(52)	(56)	(61)	(57)	(56)	(56)	(59)	(64)
	2	-6.9	-7.1	-5.7	-4.5	-3.5	-7.4	-7.5	-6.2	-5.0	-4.0	-7.1	-7.1	-5.9	-4.8	-3.8
		(48)	(48)	(51)	(57)	(63)	(50)	(51)	(52)	(56)	(61)	(55)	(56)	(56)	(59)	(64)
$\hat{t}_{y,new}(\hat{c})$ α	3	-5.8 (52)					-5.6 (55)					-5.1 (59)				
	6	-5.8 (52)					-5.6 (55)					-5.1 (59)				
	$\sqrt{n}$	-5.8 (52)					-5.6 (55)					-5.1 (59)				

Table 4.9: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the BLUP and robust predictors for the mixture distribution with $\delta = 0.99$

$\delta = 0.98$		$n = 25$					$n = 50$					$n = 100$				
$\hat{t}_{y,BLUP}$		0.4 (100)					-0.2 (100)					-0.2 (100)				
$\hat{t}_{y,CB}(c_{opt})$		-4.5 (65)					-4.7 (68)					-4.0 (74)				
$\hat{t}_{y,CHAM}(k,c)$		Huber ψ c					Huber ψ c					Huber ψ c				
		0	2	4	6	8	0	2	4	6	8	0	2	4	6	8
Huber ψ k	0.1	-15.5	-14.4	-11.3	-8.8	-6.7	-15.5	-14.3	-11.5	-9.1	-7.1	-15.3	-14.2	-11.6	-9.2	-7.2
		(47)	(37)	(40)	(46)	(53)	(57)	(44)	(45)	(49)	(56)	(71)	(56)	(53)	(54)	(58)
	0.8	-15.2	-14.3	-11.3	-8.8	-6.7	-15.3	-14.2	-11.5	-9.1	-7.1	-15.2	-14.2	-11.6	-9.2	-7.2
		(39)	(37)	(40)	(46)	(53)	(48)	(44)	(45)	(49)	(56)	(62)	(56)	(53)	(54)	(58)
	1.345	-14.7	-14.1	-11.2	-8.7	-6.7	-14.8	-14.2	-11.5	-9.1	-7.1	-14.8	-14.2	-11.6	-9.2	-7.2
		(37)	(37)	(40)	(46)	(54)	(45)	(44)	(45)	(50)	(56)	(58)	(56)	(53)	(54)	(58)
	2	-13.7	-13.9	-11.1	-8.7	-6.6	-14.0	-14.1	-11.4	-9.1	-7.1	-14.1	-14.1	-11.6	-9.2	-7.2
		(37)	(37)	(40)	(47)	(54)	(44)	(44)	(45)	(50)	(56)	(55)	(55)	(53)	(54)	(58)
$\hat{t}_{y,new}(\hat{c})$ α	3	-11.3 (41)					-11.0 (46)					-10.1 (54)				
	6	-11.3 (40)					-10.9 (46)					-10.0 (54)				
	$\sqrt{n}$	-11.3 (40)					-10.9 (46)					-9.9 (54)				

Table 4.10: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the BLUP and robust predictors for the mixture distribution with $\delta = 0.98$

4.2 Model-assisted approach

We generated ten populations of size $N = 5000$ consisting of two auxiliary variables x and z, and a survey variable y. In each population, the x-values were generated from a Gamma distribution with mean equal to 50 and variance equal to 500. A size variable z, which was correlated with x, was generated according to the conditional distribution

$$z_i|x_i \sim \Gamma(\mu_i, \nu_i)$$

where Γ denotes the Gamma distribution with mean equal to $\mu_i = x_i + 500$ and variance equal to $\nu_i = 25x_i$, such that the correlation between x and z was approximately equal to 0.5. We used the same models as in Section 4.1 for generating the y-values.

4.2.1 Stratified simple random sampling

From each population, we selected $R = 10,000$ samples of sizes $n = 25, 50, 100, 500$, according to stratified simple random sampling. We sorted the population by the z-values from the smallest to the largest. We then stratified the population into $L = 5$ strata of equal size, $N_h = 1000$, where the first strata included the smallest z-values, while the last stratum included the largest z-values. We used Neyman allocation, which is defined as

$$n_h = n\frac{N_h S_{xh}}{\sum_{n=1}^{L} N_h S_{xh}},$$

where S_{xh} is the standard deviation of the x-values in stratum h and $n = \sum_{h=1}^{L} n_h$. Within each stratum, we selected a sample of size n_h according to simple random sampling without replacement. In each sample, we computed: (i) the standard GREG estimator, $\hat{t}_{y,GREG}$, given by (1.4.8); (ii) the naive robust GREG estimator in the projection form:

$$\hat{t}_{y,GREG}^{naive} = \sum_{i\in U} \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_R(c),$$

where $\hat{\boldsymbol{\beta}}_R(c)$ was a Huber M-estimator with $c = 1.345$; (iii) the robust version of the GREG estimator based on the concept of conditional bias, $\hat{t}_{y,RG}(c_{opt})$, given by (3.3.1); (iv) the new robust GREG estimator based on an adaptive tuning constant given by (3.4.3) with $\alpha = 3, 6$ and $\sqrt{n}$.

As a measure of bias of an estimator, we computed the Monte Carlo percent relative bias:

$$RB_{MC}(\hat{t}_y) = 100 \times \frac{\mathbb{E}_{MC}(\hat{t}_y) - t_y}{t_y}. \tag{4.2.1}$$

We also computed the Monte Carlo relative efficiency, using the GREG estimator,

$\hat{t}_{y,GREG}$, as the reference:

$$RE_{MC}(\hat{t}_y) = 100 \times \frac{MSE_{MC}(\hat{t}_y)}{MSE_{MC}(\hat{t}_{y,GREG})}. \quad (4.2.2)$$

Tables 4.11-4.20 show the Monte Carlo relative bias and relative efficiency for several estimators in the case of stratified simple random sampling. The standard GREG estimator, $\hat{t}_{y,GREG}$, showed a negligible bias in all the scenarios. For the normal distribution (i.e., in the absence of the influential units), $\hat{t}_{y,GREG}$ and $\hat{t}_{y,RG}$ showed almost identical performances in terms of both relative bias and relative efficiency. Both $\hat{t}^{naive}_{y,GREG}$ and $\hat{t}^{new}_{y,GREG}(\hat{c})$ showed negligible bias and a slight loss of efficiency (between 1% and 2% for $\hat{t}^{new}_{y,GREG}(\hat{c})$, between 3% and 5% for $\hat{t}^{naive}_{y,GREG}$) with respect to $\hat{t}_{y,GREG}$; see Table 4.11.

For the symmetric distribution (t-distribution and Laplace distribution), all the robust GREG estimators showed negligible bias in all the scenarios; see Table 4.15 and 4.17. The naive robust GREG estimator, $\hat{t}^{naive}_{y,GREG}$ was especially efficient. For instance, for the t-distribution (Table 4.15), the value of relative efficiency was 57% with $n = 50$. The new GREG estimator, $\hat{t}^{new}_{y,GREG}(\hat{c})$, was more efficient than $\hat{t}_{y,RG}$ with a value of relative efficiency of 67%, compared to 86% for $\hat{t}_{y,RG}$. We note that the value of relative efficiency increased as the value of α increased when the sample size was large. For example, for the t-distribution with $n = 500$, the value of relative efficiency was 73% when $\alpha = 3$ and increased to 83% when $\alpha = \sqrt{n}$.

For the skewed distributions (Gamma, Pareto, lognormal and Weibull), we start by noting that the naive robust GREG estimator, $\hat{t}^{naive}_{y,GREG}$ did not behave well, in general. For instance, for the lognormal distribution (Table 4.13) and $n = 50$, the Monte Carlo relative bias of $\hat{t}^{naive}_{y,GREG}$ was -19.8% and the value of relative efficiency of $\hat{t}^{naive}_{y,GREG}$ was 196%. In other words, the naive robust GREG estimator was much worse than $\hat{t}_{y,BLUP}$ in many scenarios. This was also noted for the other skewed distributions.

The robust GREG estimator $\hat{t}_{y,RG}$ was generally biased as expected. The bias was moderate in most cases with a maximum absolute relative bias smaller than 6%. We note that the relative bias decreased as the sample size n increased. In terms of relative efficiency, it was better than $\hat{t}_{y,BLUP}$ in most scenarios. For instance for the Pareto distribution, with $n = 50$, the value of relative efficiency was equal to 78%; see Table 4.14. For the lognormal distribution with $n = 50$, the value of relative efficiency was equal to 93%; see Table 4.13. We note that the value of relative efficiency approached 100% as the sample size increased. This suggests that the robust GREG estimator $\hat{t}_{y,RG}$ is consistent for t_y.

The new robust GREG estimator $\hat{t}^{new}_{y,GREG}$ performed well in all the scenarios when the sample size was small (e.g. $n = 25$ and 50) with values of relative efficiency never larger than 100%. When the sample size was large (e.g.,$n = 500$), the new robust GREG estimator $\hat{t}_{y,GREG}^{new}$ was slightly less efficient than $\hat{t}_{y,GREG}$ (between 1% and 4%). In general, $\hat{t}_{y,GREG}^{new}$ showed larger absolute relative bias than $\hat{t}_{y,RG}$. In some cases, $\hat{t}_{y,GREG}^{new}$ was slightly less efficient than $\hat{t}_{y,RG}$. For example, for the lognormal distribution with $n = 50$, the relative bias for $\hat{t}_{y,RG}$ and $\hat{t}_{y,new}(\hat{c})$ (with $\alpha = \sqrt{n}$), was respectively equal to -3.3% and -4.2%, and the relative efficiency for $\hat{t}_{y,RG}$ and $\hat{t}_{y,GREG}^{new}$ was respectively equal to 93% and 95%. In the case of Pareto distribution, $\hat{t}_{y,GREG}^{new}$ was slightly more efficient than $\hat{t}_{y,RG}$ with a value of relative efficiency of 73%, compare to 78% for $\hat{t}_{y,RG}$. We note that the relative bias of $\hat{t}_{y,new}(\hat{c})$ decreased as the sample size n increased and that its value of relative efficiency increased as n increased. For a small value of n, the choice of α did not seem to affect the behavior of $\hat{t}_{y,GREG}^{new}$. However, when $n = 500$, the value of relative efficiency decreased as α increased. For example, for the Pareto distribution, the value of relative efficiency was 104% with $\alpha = 3$ and 91% with $\alpha = \sqrt{n}$.

Finally, we turn to the mixture distribution. When the percentage of contamination was small (e.g., 0.5% and 1%) both $\hat{t}_{y,RG}$ and $\hat{t}_{y,GREG}^{new}$ were more efficient than $\hat{t}_{y,GREG}$ with values of relative efficiency smaller than 100. However, when the percentage of contamination increased to 2%, with a large sample size $n = 500$, $\hat{t}_{y,GREG}^{new}$ became less efficient than $\hat{t}_{y,GREG}$. For example, when $\alpha = \sqrt{n}$, the value of relative efficiency of $\hat{t}_{y,GREG}^{new}$ was 116%; see Table 4.20. Except the scenario when $n = 500$ and $\delta = 0.98$, $\hat{t}_{y,RG}$ exhibited smaller values of absolute relative bias than $\hat{t}_{y,GREG}^{new}$ but the latter was more efficient. For instance, in the case of a 2% contamination with $n = 50$ (see Table 4.20), the relative bias of $\hat{t}_{y,RG}$ was equal to -7.6%, whereas it was equal to -17% for $\hat{t}_{y,GREG}^{new}$ (with $\alpha = \sqrt{n}$) but the relative efficiency of the latter was equal to 45% compared to 65% for $\hat{t}_{y,RG}$. Again, for a small value of n, the choice of α did not affect the behavior of $\hat{t}_{y,GREG}^{new}$. However, when $n = 500$, the value of relative efficiency decreased as α increased. For instance, in the case of a 2% contamination, the value of relative efficiency was 145% with $\alpha = 3$ and 116% with $\alpha = \sqrt{n}$.

Normal		n=25	n=50	n=100	n=500
$\hat{t}_{y,GREG}$		-0.3 (100)	-0.2 (100)	0.0 (100)	0.0 (100)
$\hat{t}^{naive}_{y,GREG}$		-0.1 (105)	-0.1 (104)	0.1 (104)	0.1 (103)
$\hat{t}_{y,RG}$		-0.2 (100)	-0.1 (100)	0.1 (100)	0.0 (100)
$\hat{t}^{new}_{y,GREG}$ α	3	-0.2 (102)	-0.1 (102)	0.1 (102)	0.2 (102)
	6	-0.2 (102)	-0.1 (102)	0.2 (102)	0.1 (101)
	$\sqrt{n}$	-0.2 (102)	-0.1 (102)	0.2 (102)	0.0 (100)

Table 4.11: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the GREG estimator and robust estimators for the normal distribution

Gamma		n=25	n=50	n=100	n=500
$\hat{t}_{y,GREG}$		0.1 (100)	-0.1 (100)	0.0 (100)	0.0 (100)
$\hat{t}^{naive}_{y,GREG}$		-15.6 (132)	-16.5 (177)	-16.5 (275)	-16.5 (1130)
$\hat{t}_{y,RG}$		-4.3 (96)	-2.9 (97)	-1.7 (99)	-0.5 (100)
$\hat{t}^{new}_{y,GREG}$ α	3	-3.4 (98)	-2.4 (100)	-1.4 (101)	-0.3 (101)
	6	-3.3 (98)	-2.3 (100)	-1.3 (101)	-0.2 (101)
	$\sqrt{n}$	-3.3 (98)	-2.3 (100)	-1.1 (101)	0.0 (100)

Table 4.12: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the GREG estimator and robust estimators for the Gamma distribution

Lognormal		n=25	n=50	n=100	n=500
$\hat{t}_{y,GREG}$		-0.3	-0.1	0.1	0.0
		(100)	(100)	(100)	(100)
$\hat{t}^{naive}_{y,GREG}$		-18.9	-19.8	-20.0	-20.2
		(130)	(196)	(344)	(1619)
$\hat{t}_{y,RG}$		-5.2	-3.3	-2.0	-0.6
		(90)	(93)	(96)	(99)
$\hat{t}^{new}_{y,GREG}$ α	3	-5.8	-4.3	-2.8	-1.0
		(93)	(96)	(98)	(101)
	6	-5.7	-4.2	-2.7	-0.9
		(93)	(95)	(97)	(99)
	$\sqrt{n}$	-5.7	-4.2	-2.5	-0.3
		(93)	(95)	(97)	(97)

Table 4.13: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the GREG estimator and robust estimators for the lognormal distribution

Pareto		n=25	n=50	n=100	n=500
$\hat{t}_{y,GREG}$		0.0	0.1	0.1	0.0
		(100)	(100)	(100)	(100)
$\hat{t}^{naive}_{y,GREG}$		-14.8	-15.3	-15.4	-15.7
		(103)	(185)	(352)	(1856)
$\hat{t}_{y,RG}$		-4.1	-2.8	-1.8	-0.7
		(73)	(78)	(83)	(93)
$\hat{t}^{new}_{y,GREG}$ α	3	-5.8	-4.7	-3.4	-1.8
		(66)	(73)	(81)	(104)
	6	-5.8	-4.6	-3.4	-1.7
		(66)	(73)	(81)	(100)
	$\sqrt{n}$	-5.8	-4.6	-3.3	-1.2
		(66)	(73)	(80)	(91)

Table 4.14: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the GREG estimator and robust estimators for the Pareto distribution

t		n=25	n=50	n=100	n=500
$\hat{t}_{y,GREG}$		0.0 (100)	-0.1 (100)	0.0 (100)	0.0 (100)
$\hat{t}^{naive}_{y,GREG}$		0.2 (59)	0.2 (57)	0.2 (57)	0.2 (59)
$\hat{t}_{y,RG}$		0.1 (79)	0.0 (83)	0.0 (87)	0.0 (93)
$\hat{t}^{new}_{y,GREG}$ α	3	0.1 (67)	0.1 (66)	0.1 (68)	0.0 (73)
	6	0.1 (68)	0.1 (67)	0.1 (69)	0.0 (75)
	$\sqrt{n}$	0.1 (68)	0.1 (67)	0.1 (69)	0.0 (83)

Table 4.15: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the GREG estimator and robust estimators for the t-distribution

Weibull		n=25	n=50	n=100	n=500
$\hat{t}_{y,GREG}$		-0.1 (100)	0.1 (100)	0.3 (100)	0.0 (100)
$\hat{t}^{naive}_{y,GREG}$		-15.7 (135)	-16.2 (185)	-16.1 (276)	-16.6 (1157)
$\hat{t}_{y,RG}$		-4.3 (99)	-2.6 (100)	-1.2 (100)	-0.4 (101)
$\hat{t}^{new}_{y,GREG}$ α	3	-3.6 (101)	-2.3 (101)	-1.2 (100)	-0.6 (101)
	6	-3.5 (101)	-2.2 (101)	-1.1 (100)	-0.4 (100)
	$\sqrt{n}$	-3.5 (101)	-2.2 (101)	-1.0 (100)	-0.1 (100)

Table 4.16: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the GREG estimator and robust estimators for the Weibull distribution

Laplace		n=25	n=50	n=100	n=500
$\hat{t}_{y,GREG}$		0.1 (100)	0.1 (100)	-0.3 (100)	0.0 (100)
$\hat{t}^{naive}_{y,GREG}$		-0.1 (78)	-0.2 (77)	-0.4 (74)	-0.1 (73)
$\hat{t}_{y,RG}$		-0.1 (93)	0.0 (96)	-0.4 (97)	0.0 (99)
$\hat{t}^{new}_{y,GREG}$ α	3	-0.1 (86)	0.0 (84)	-0.4 (83)	-0.1 (86)
	6	-0.1 (86)	0.0 (85)	-0.4 (84)	-0.1 (89)
	$\sqrt{n}$	-0.1 (86)	0.0 (85)	-0.4 (85)	0.0 (97)

Table 4.17: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the GREG estimator and robust estimators for the Laplace distribution

$\delta = 0.995$		n=25	n=50	n=100	n=500
$\hat{t}_{y,GREG}$		0.1 (100)	0.1 (100)	-0.5 (100)	0.1 (100)
$\hat{t}^{naive}_{y,GREG}$		-4.6 (60)	-4.5 (61)	-5.0 (66)	-4.7 (84)
$\hat{t}_{y,RG}$		-2.4 (71)	-2.3 (74)	-2.7 (77)	-1.4 (89)
$\hat{t}^{new}_{y,GREG}$ α	3	-4.2 (61)	-3.8 (63)	-4.1 (67)	-3.0 (82)
	6	-4.2 (61)	-3.8 (63)	-4.1 (67)	-2.9 (81)
	$\sqrt{n}$	-4.2 (61)	-3.8 (63)	-4.1 (67)	-2.4 (80)

Table 4.18: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the GREG estimator and robust estimators for the mixture distribution with $\delta = 0.995$

$\delta = 0.99$		n=25	n=50	n=100	n=500
$\hat{t}_{y,GREG}$		0.4 (100)	0.2 (100)	-0.4 (100)	0.0 (100)
$\hat{t}^{naive}_{y,GREG}$		-10.0 (45)	-10.0 (49)	-10.5 (59)	-10.2 (129)
$\hat{t}_{y,RG}$		-4.6 (61)	-4.3 (65)	-4.3 (73)	-1.9 (92)
$\hat{t}^{new}_{y,GREG}$ α	3	-8.5 (46)	-8.0 (50)	-8.0 (57)	-5.9 (89)
	6	-8.5 (46)	-8.0 (50)	-8.0 (58)	-5.7 (88)
	$\sqrt{n}$	-8.5 (46)	-8.0 (50)	-7.9 (58)	-4.7 (83)

Table 4.19: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the GREG estimator and robust estimators for the mixture distribution with $\delta = 0.99$

$\delta = 0.98$		n=25	n=50	n=100	n=500
$\hat{t}_{y,GREG}$		0.4 (100)	0.2 (100)	-0.3 (100)	0.1 (100)
$\hat{t}^{naive}_{y,GREG}$		-22.2 (36)	-22.2 (47)	-22.9 (73)	-22.5 (288)
$\hat{t}_{y,RG}$		-9.0 (57)	-7.6 (65)	-6.5 (76)	-1.9 (96)
$\hat{t}^{new}_{y,GREG}$ α	3	-18.1 (37)	-17.0 (45)	-16.3 (60)	-12.8 (145)
	6	-18.0 (37)	-17.0 (45)	-16.2 (60)	-12.4 (140)
	$\sqrt{n}$	-18.0 (37)	-17.0 (45)	-16.1 (60)	-10.0 (116)

Table 4.20: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the GREG estimator and robust estimators for the mixture distribution with $\delta = 0.98$

4.2.2 Poisson sampling

From each population, we selected $R = 10000$ samples according to Poisson sampling with inclusion probabilities, π_i, proportional to the size variable, z_i; i.e., $\pi_i = \widetilde{n} z_i / \sum_{i \in U} z_i$, where $\widetilde{n}$ denotes the expected sample sizes, 25, 50, 100 and 500.

In each sample, we computed the standard GREG estimator and robust GREG estimators described in §4.2.1.

We compared each estimator in terms of Monte Carlo percent relative bias given by (4.2.1) and the Monte Carlo relative efficiency given by (4.2.2).

Tables 4.21-4.30 show the relative bias and relative efficiency of several estimators in the case of Poisson sampling. The results for all the scenarios were quite similar to that obtained in the case of stratified simple random sampling. For this reason, we do not discuss the results further.

Normal		n=25	n=50	n=100	n=500
$\hat{t}_{y,GREG}$		0.2 (100)	-0.1 (100)	0.0 (100)	0.0 (100)
$\hat{t}^{naive}_{y,GREG}$		0.3 (105)	0.0 (105)	0.1 (104)	0.1 (104)
$\hat{t}_{y,RG}$		0.3 (100)	-0.1 (100)	0.0 (100)	0.0 (100)
$\hat{t}^{new}_{y,GREG}$ α	3	0.3 (101)	0.0 (102)	0.0 (102)	0.1 (102)
	6	0.3 (101)	0.0 (102)	0.0 (102)	0.1 (101)
	$\sqrt{n}$	0.3 (101)	0.0 (102)	0.0 (102)	0.1 (101)

Table 4.21: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the GREG estimator and robust estimators for the normal distribution

Gamma		n=25	n=50	n=100	n=500
$\hat{t}_{y,GREG}$		0.0 (100)	0.1 (100)	0.0 (100)	0.0 (100)
$\hat{t}^{naive}_{y,GREG}$		-16.1 (132)	-16.6 (182)	-17.0 (289)	-17.3 (1191)
$\hat{t}_{y,RG}$		-3.5 (98)	-2.4 (98)	-1.5 (99)	-0.5 (100)
$\hat{t}^{new}_{y,GREG}$ α	3	-3.8 (100)	-2.4 (100)	-1.5 (101)	-0.4 (101)
	6	-3.7 (100)	-2.3 (100)	-1.4 (100)	-0.3 (100)
	$\sqrt{n}$	-3.7 (100)	-2.3 (100)	-1.4 (100)	-0.3 (100)

Table 4.22: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the GREG estimator and robust estimators for the Gamma distribution

Lognormal		n=25	n=50	n=100	n=500
$\hat{t}_{y,GREG}$		0.3 (100)	0.0 (100)	0.0 (100)	0.0 (100)
$\hat{t}^{naive}_{y,GREG}$		-18.3 (116)	-19.7 (194)	-20.1 (347)	-20.2 (1672)
$\hat{t}_{y,RG}$		-3.7 (90)	-2.9 (94)	-1.9 (97)	-0.6 (99)
$\hat{t}^{new}_{y,GREG}$ α	3	-5.6 (88)	-4.3 (95)	-3.0 (99)	-1.1 (103)
	6	-5.5 (88)	-4.2 (95)	-2.9 (98)	-0.9 (101)
	$\sqrt{n}$	-5.5 (88)	-4.2 (95)	-2.9 (98)	-0.8 (100)

Table 4.23: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the GREG estimator and robust estimators for the lognormal distribution

Pareto		n=25	n=50	n=100	n=500
$\hat{t}_{y,GREG}$		0.0 (100)	0.1 (100)	0.0 (100)	0.0 (100)
$\hat{t}^{naive}_{y,GREG}$		-15.2 (98)	-15.7 (177)	-16.0 (350)	-16.1 (1822)
$\hat{t}_{y,RG}$		-3.5 (77)	-2.7 (79)	-2.0 (83)	-0.7 (93)
$\hat{t}^{new}_{y,GREG}$ α	3	-6.1 (64)	-4.8 (71)	-3.7 (80)	-1.7 (99)
	6	-6.0 (64)	-4.7 (70)	-3.6 (80)	-1.6 (96)
	$\sqrt{n}$	-6.0 (64)	-4.7 (70)	-3.6 (80)	-1.5 (94)

Table 4.24: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the GREG estimator and robust estimators for the Pareto distribution

t		n=25	n=50	n=100	n=500
$\hat{t}_{y,GREG}$		0.1	0.0	0.0	0.0
		(100)	(100)	(100)	(100)
$\hat{t}^{naive}_{y,GREG}$		0.1	0.1	0.1	0.1
		(58)	(58)	(57)	(60)
$\hat{t}_{y,RG}$		0.1	0.1	0.0	0.0
		(82)	(84)	(86)	(92)
$\hat{t}^{new}_{y,GREG}$ α	3	0.1	0.1	0.1	0.0
		(66)	(67)	(67)	(74)
	6	0.1	0.1	0.1	0.0
		(66)	(67)	(68)	(75)
	$\sqrt{n}$	0.1	0.1	0.1	0.0
		(66)	(67)	(68)	(75)

Table 4.25: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the GREG estimator and robust estimators for the t-distribution

Weibull		n=25	n=50	n=100	n=500
$\hat{t}_{y,GREG}$		-0.1	0.1	0.0	0.0
		(100)	(100)	(100)	(100)
$\hat{t}^{naive}_{y,GREG}$		-16.5	-16.9	-17.1	-17.3
		(131)	(182)	(290)	(1208)
$\hat{t}_{y,RG}$		-3.7	-2.4	-1.6	-0.4
		(98)	(99)	(100)	(100)
$\hat{t}^{new}_{y,GREG}$ α	3	-3.8	-2.3	-1.5	-0.4
		(99)	(100)	(101)	(100)
	6	-3.8	-2.2	-1.4	-0.3
		(99)	(100)	(101)	(100)
	$\sqrt{n}$	-3.8	-2.2	-1.4	-0.3
		(99)	(100)	(101)	(100)

Table 4.26: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the GREG estimator and robust estimators for the Weibull distribution

Laplace		n=25	n=50	n=100	n=500
$\hat{t}_{y,GREG}$		0.1 (100)	-0.1 (100)	0.0 (100)	0.0 (100)
$\hat{t}^{naive}_{y,GREG}$		0.2 (77)	0.1 (74)	0.2 (74)	0.2 (74)
$\hat{t}_{y,RG}$		0.1 (94)	0.0 (96)	0.0 (97)	0.0 (99)
$\hat{t}^{new}_{y,GREG}$ α	3	0.1 (85)	0.0 (83)	0.1 (83)	0.1 (85)
	6	0.1 (85)	0.0 (83)	0.1 (84)	0.0 (88)
	$\sqrt{n}$	0.1 (85)	0.0 (83)	0.1 (84)	0.0 (88)

Table 4.27: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the GREG estimator and robust estimators for the Laplace distribution

$\delta = 0.995$		n=25	n=50	n=100	n=500
$\hat{t}_{y,GREG}$		0.2 (100)	-0.2 (100)	-0.1 (100)	0.0 (100)
$\hat{t}^{naive}_{y,GREG}$		-2.5 (67)	-2.9 (68)	-2.8 (69)	-2.8 (76)
$\hat{t}_{y,RG}$		-0.8 (82)	-1.4 (80)	-1.3 (80)	-0.9 (86)
$\hat{t}^{new}_{y,GREG}$ α	3	-2.2 (68)	-2.5 (69)	-2.2 (71)	-1.8 (81)
	6	-2.2 (68)	-2.5 (69)	-2.2 (71)	-1.8 (81)
	$\sqrt{n}$	-2.2 (68)	-2.5 (69)	-2.2 (71)	-1.8 (81)

Table 4.28: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the GREG estimator and robust estimators for the mixture distribution with $\delta = 0.995$

$\delta = 0.99$		n=25	n=50	n=100	n=500
$\hat{t}_{y,GREG}$		0.6 (100)	-0.2 (100)	0.2 (100)	0.0 (100)
$\hat{t}^{naive}_{y,GREG}$		-8.3 (44)	-8.9 (48)	-8.7 (53)	-8.7 (110)
$\hat{t}_{y,RG}$		-2.7 (70)	-3.5 (69)	-3.0 (71)	-1.8 (89)
$\hat{t}^{new}_{y,GREG}$ α	3	-7.0 (47)	-7.2 (50)	-6.5 (54)	-5.3 (86)
	6	-7.0 (47)	-7.2 (50)	-6.5 (54)	-5.1 (86)
	$\sqrt{n}$	-7.0 (47)	-7.2 (50)	-6.5 (54)	-5.1 (86)

Table 4.29: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the GREG estimator and robust estimators for the mixture distribution with $\delta = 0.99$

$\delta = 0.98$		n=25	n=50	n=100	n=500
$\hat{t}_{y,GREG}$		0.2 (100)	-0.4 (100)	0.0 (100)	0.0 (100)
$\hat{t}^{naive}_{y,GREG}$		-16.0 (36)	-16.6 (44)	-16.5 (58)	-16.4 (193)
$\hat{t}_{y,RG}$		-5.3 (67)	-5.8 (68)	-4.5 (74)	-1.7 (92)
$\hat{t}^{new}_{y,GREG}$ α	3	-12.9 (39)	-12.8 (45)	-11.7 (55)	-9.7 (117)
	6	-12.9 (39)	-12.8 (45)	-11.7 (55)	-9.3 (114)
	$\sqrt{n}$	-12.9 (39)	-12.8 (45)	-11.7 (55)	-9.3 (114)

Table 4.30: Monte Carlo percent relative bias and relative efficiency (in parentheses) of the GREG estimator and robust estimators for the mixture distribution with $\delta = 0.98$

Chapter 5

Conclusion

In this thesis, we studied the problem of influential units in a finite population setting. We considered the model-based framework as well as the design-based framework and its associated model-assisted framework. We proposed a robust estimator and a robust predictor based on an adaptive tuning constant.

In the simulation studies, we showed that the naive estimators/predictors based on a non-adaptive tuning constant (e.g., $c = 1.345$) do not generally perform well in terms of bias and efficiency. In many cases, their mean square error is larger than that of the customary (non-robust) estimator/predictor. One notable exception occurs when the distribution is symmetric (e.g., the t-distribution). The proposed estimator/predictor based on an adaptive tuning constant performed well in term of efficiency in virtually all the scenarios. In comparison with the robust estimator/predictor of Beaumont et al. (2013) based on the concept of conditional bias, we found that the proposed method led generally to a lower mean square error.

In practice, we may be interested in estimating the mean square error of the proposed robust predictor/estimator. This is a challenging problem that is a topic of future investigation. For instance, suppose that we wish to estimate the design mean square error of the robust GREG estimator given by (3.4.3), denoted by $MSE_p(\hat{t}^{new}_{y,GREG})$.

An estimator of $MSE_p(\hat{t}^{new}_{y,GREG})$ is given by

$$\widehat{\text{MSE}} = \widehat{V}\left(\hat{t}^{new}_{y,GREG})\right) + \max\left\{0, (\hat{t}^{new}_{y,GREG} - \hat{t}_{y,GREG})^2 - \widehat{V}\left(\hat{t}^{new}_{y,GREG}) - \hat{t}_{y,GREG}\right)\right\}.$$

The terms $\widehat{V}\left(\hat{t}^{new}_{y,GREG}\right)$ and $\widehat{V}\left(\hat{t}^{new}_{y,GREG} - \hat{t}_{y,GREG}\right)$ may be obtained using a boostrap procedure; see Mashreghi et al. (2016) for a review of bootstrap procedures in finite population sampling.

Appendix A

Graphical illustration

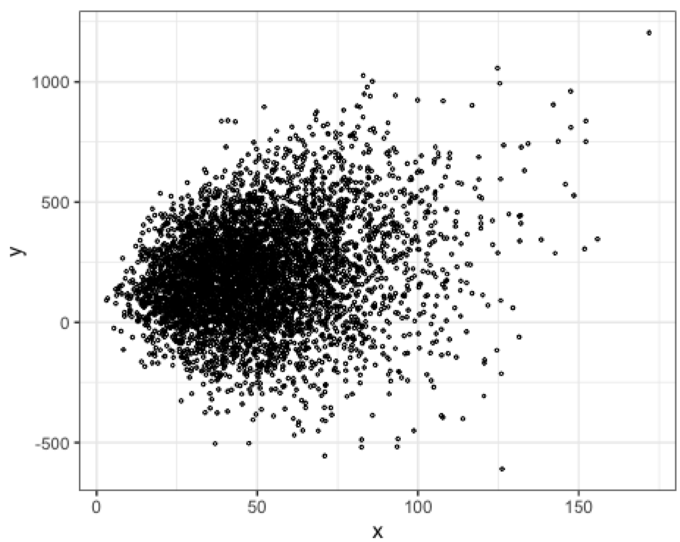

Figure A.1: Example of population generated from the normal distribution

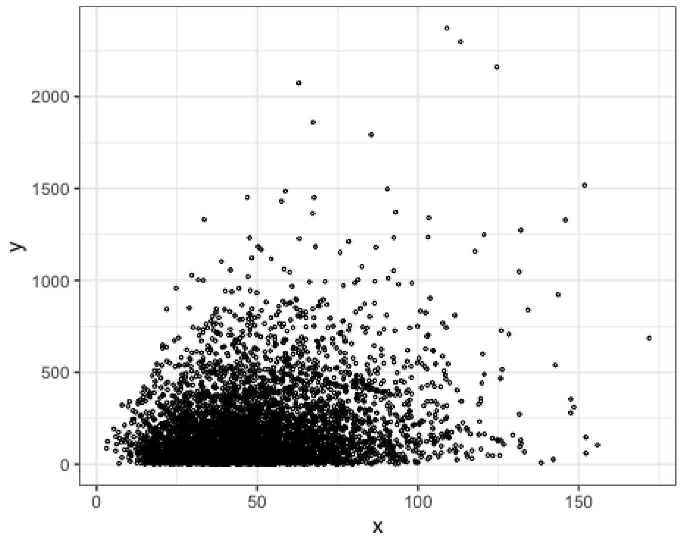

Figure A.2: Example of population generated from the Gamma distribution

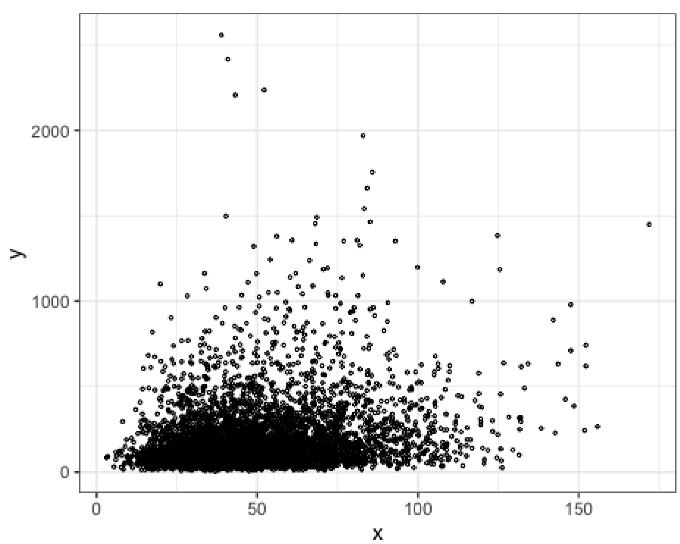

Figure A.3: Example of population generated from the lognormal distribution

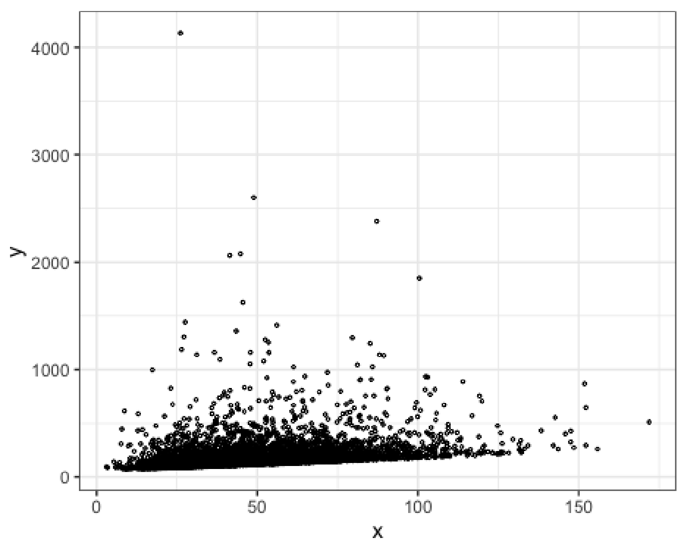

Figure A.4: Example of population generated from the Pareto distribution

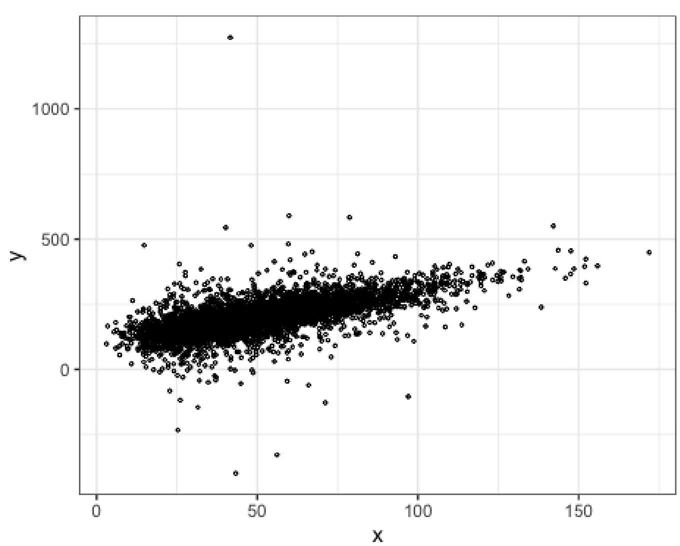

Figure A.5: Example of population generated from the t-distribution

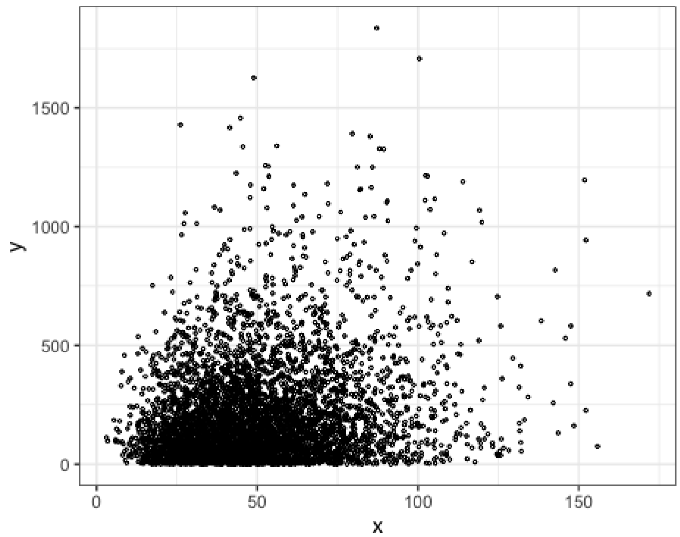

Figure A.6: Example of population generated from the Weibull distribution

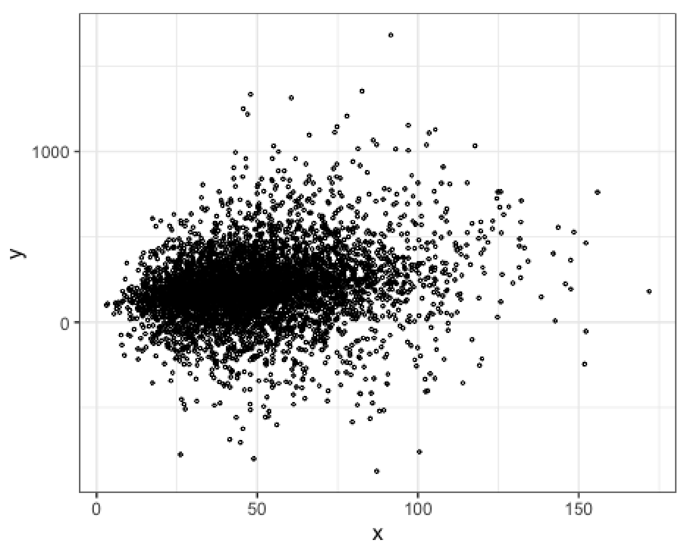

Figure A.7: Example of population generated from the Laplace distribution

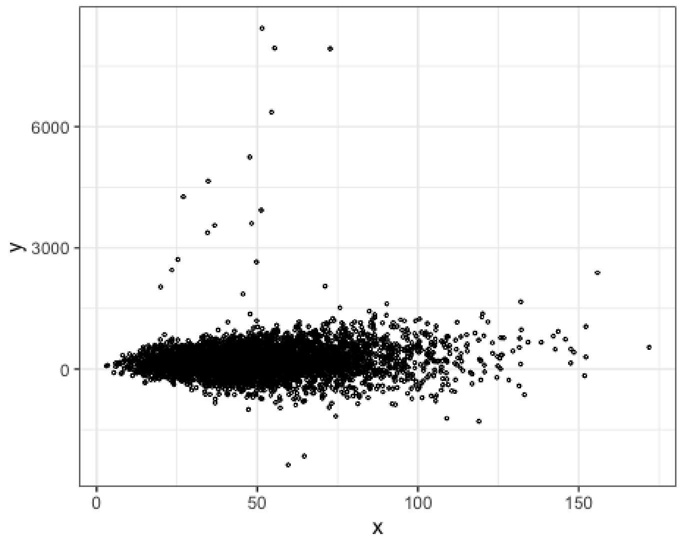

Figure A.8: Example of population generated from the mixture distribution ($\delta = 0.995$)

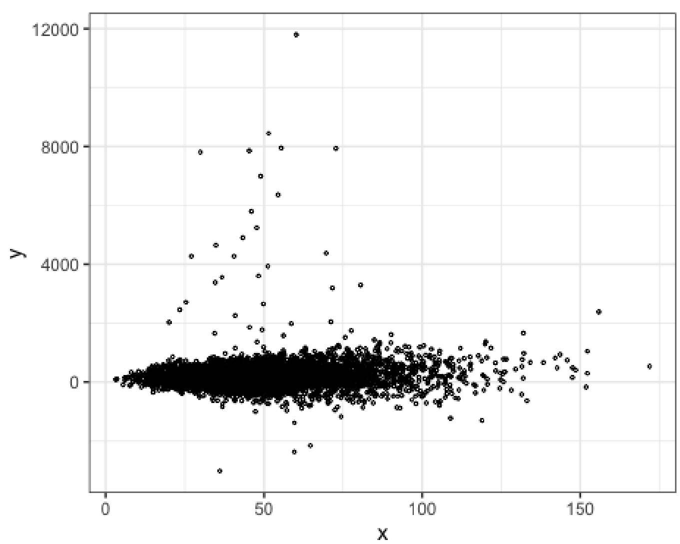

Figure A.9: Example of population generated from the mixture distribution ($\delta = 0.99$)

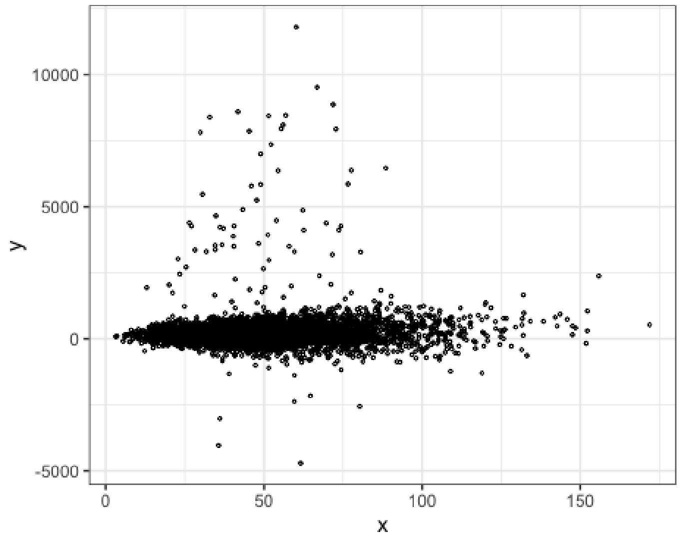

Figure A.10: Example of population generated from the mixture distribution ($\delta = 0.98$)

Bibliography

[1] Andersen, R., (2008). *Modern Methods for Robust Regression*. SAGE Publications, Inc.

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

[3] Chambers, R. L., and Clark, R. G. (2012). *An Introduction to Model-Based Survey Sampling with Applications*. Oxford University Press.

[4] Chambers, R. L. (1986). Outlier Robust Finite Population Estimation. *Journal of the American Statistical Association*, **81**, 1063–1069.

[5] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (2005). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons.

[6] Horvitz, D. G., and D. J. Thompson. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, **47**, 663–685.

[7] Huber, P.J. (1981) *Robust statistics.* John Wiley & Sons.

[8] Lohr, S.L. (2010). *Sampling: Design and Analysis (2nd Edition)*. Brooks/Cole, Boston.

[9] Maronna, R. A., Martin, D. R., and Yohai, V. J. (2019). *Robust Statistics: Theory and Methods*. Wiley Series in Probability and Statistics.

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

[11] Moreno-Rebollo, J. L., Muñoz-Reyez, A. M., and Muñoz-Pichardo, J. M. (1999). Influence diagnostic in survey sampling: conditional bias. *Biometrika*, **86**, 923–928.

[12] Moreno-Rebollo, J. L., Muñoz-Reyez, A. M., Jimenez-Gamero, M. D. and Muñoz-Pichardo, J. M. (2002). Influence diagnostics in survey sampling: Estimating the conditional bias. *Metrika* **55**, 209–214.

[13] Muñoz-Pichardo, J. M., Muñoz-Garcia, J., Moreno-Rebollo, J. L., and Pino-Mejias, R. (1995). A new approach to influence analysis in linear models. *Sankhyā: The Indian Journal of Statistics, Series A*, **57**, 393–409.

[14] Royall, R. M. (1976). The Linear Least-Squares Prediction Approach to Two-State Sampling. *Journal of the American Statistical Association.* **71**, 657–664.

[15] Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model-Assisted Survey Sampling*. Springer-Verlag Publishing.

[16] Sherman, J., and Morrison, W. J. (1949). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, **20**, 621–624.

[17] Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society. Series B (Methodological)*, **15**, 253–261.